Linear Correlation Estimation

Filip Lindskog
RiskLab
D-MATH
ETH-Zentrum
CH-8092 Zürich
Switzerland
lindskog@math.ethz.ch
http://www.math.ethz.ch/~lindskog

December 11, 2000

Abstract. Most financial models for modelling dependent risks are based on the assumption of multivariate normality and linear correlation is used as a measure of dependence. However, observed financial data are rarely normally distributed and tend to have marginal distributions with heavier tails. Furthermore, the observed synchronized extreme falls in financial markets can not be modelled by multivariate normal distributions. However, there are other elliptical distributions with these properties and the assumption of multivariate normality can often be replaced by the assumption of ellipticality. A useful property of elliptical distributions is that these distributions support the standard approaches of risk management. Value-at-Risk fulfills the desired properties of a risk measure and the mean-variance (Markowitz) approach can be used for portfolio optimization.

For elliptical distributions linear correlation is still a natural measure of dependence. However, a linear correlation estimator such as the Pearson product-moment correlation estimator (the standard estimator) being suitable for data from uncontaminated multivariate normal distributions has a very bad performance for heavier tailed or contaminated data. Therefore robust estimators are needed, robust in the sense of being insensitive to contamination and yet maintaining a high efficiency for heavier tailed elliptical distributions as well as for multivariate normal distributions.

In this paper an overview of techniques for robust linear correlation estimation is given and the various techniques are compared for contaminated and uncontaminated elliptical distributions. In particular the relation

$$\tau = \frac{2}{\pi} \arcsin \rho$$

between Kendall's tau and the linear correlation coefficient ρ is shown to hold for (essentially) all elliptical distributions and the estimator of linear correlation provided by this relation is studied. This non-parametric estimator inherits the robustness properties of the Kendall's tau estimator and is an efficient (low variance) estimator for all elliptical distributions.

Key words and phrases. Robust estimation; linear correlation; Kendall's tau; elliptical distributions.
The research was supported by Credit Suisse, Swiss Re and UBS through RiskLab, Switzerland.

Contents

1. Introduction 1
2. Elliptical Distributions 3
3. Bivariate Correlation Estimators 11
4. Transformations of non-PSD correlation matrices 15
5. Multivariate Correlation Estimators 21
6. Comparisons 25
References 31

1 Introduction

Empirical studies suggest that the multivariate normal distribution is in general not suitable for modelling financial data.

- Observed financial data are rarely normally distributed and tend to have marginal distributions with heavier tails.
- Extreme, synchronized rises and falls in financial markets occur infrequently, but more often than what can be derived from a model based on a multivariate normal distribution.

While these properties of financial data will never be explained by a multivariate normal distribution, there are other elliptical distributions with these properties.

A useful property of elliptical distribution is that these distributions support the standard approaches of risk management. Value-at-Risk fulfills the desired properties of a risk measure and the mean-variance (Markowitz) approach can be used for portfolio optimization.

Linear correlation is a natural measure of dependence for elliptical distributions, having the well known multivariate normal distribution as one of its members. However, a linear correlation estimator such as the Pearson product-moment correlation estimator (the standard estimator) being suitable for uncontaminated (multivariate) normally distributed data has a very bad performance for heavier tailed or contaminated data.

Furthermore, it seems reasonable to expect some portion of the data to be contaminated due to wrong measurements, wrong decimal points or some other form of errors. What is needed are robust estimators, robust in the sense of being insensitive to contamination and yet maintaining a high efficiency for heavier tailed elliptical distributions as well as for multivariate normal distributions.

We might ask whether robust estimators are needed at all. Perhaps the following approach would suffice:

1. Clean the data according to some outlier rejection scheme.
2. Apply classical estimators on the cleaned data.

Unfortunately this approach is not good. For multivariate data it is rarely possible to separate contamination from the uncontaminated underlying distribution since in general contamination can not be observed as just observations “sticking out”. Even if “extreme” outliers are sorted out, the data is probably still contaminated and a large amount of remaining contamination may disturb the estimate just as much. Furthermore, under only the assumption of ellipticality, no outlier rejection procedure seem to outperform the best robust estimators.

In following chapters we will discuss and compare linear correlation matrix estimators under the assumption that data are approximately elliptically distributed which can be interpreted as the point clouds of p-dimensional observations roughly having the form of ellipsoids with some deviations due to contamination.

It should be noted that the linear correlation estimators are used for estimating the standardized shape parameter of an elliptical distribution, which equals the linear correlation matrix (coefficient) when the necessary second order moments are finite. To avoid complicated terminology we will however call this standardized shape parameter a linear correlation matrix (coefficient) even when it is not *the* linear correlation matrix (coefficient) of the distribution.

2 Elliptical Distributions

In this chapter the broad class of elliptical distributions is presented. They provide useful alternatives to the multivariate normal distribution and many of the nice properties of the multivariate normal distribution holds more generally for elliptical distributions.

The main focus will be on presenting basic properties of elliptical distributions and showing that the relation between Kendall's tau and the linear correlation coefficient for a bivariate normal distribution holds for (essentially) all elliptical distributions. A useful application of this result is robust and efficient estimation of linear correlation for elliptical distributions. This will be explored in the following chapters.

2.1 Definitions and Properties. A spherical distribution is an extension of the standard multivariate normal distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$ and an elliptical distribution is an extension of $\mathcal{N}_p(\mu, \Sigma)$. Recall that $\mathcal{N}_p(\mu, \Sigma)$ can be defined from $\mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$ via

$$\mathbf{X} =_d \mu + A\mathbf{Y},$$

where $\mathbf{X} \sim \mathcal{N}_p(\mu, \Sigma), \mathbf{Y} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$ and $\Sigma = AA^t$. With this in mind it seems natural to start with spherical distributions and then define elliptical distributions from the spherical in the way indicated above.

For the results regarding basic properties of spherical and elliptical distributions, the proofs are omitted. The proofs can be found in Fang, Kotz and Ng (1987) [6].

Definition 1 A random p-vector $\mathbf{X}$ is said to have a spherical distribution if for every $\Gamma \in \mathcal{O}(p)$,

$$\Gamma\mathbf{X} =_d \mathbf{X},$$

where $\mathcal{O}(p)$ denotes the set of $p \times p$ orthogonal matrices, i.e. matrices Γ such that $\Gamma^t\Gamma = \Gamma\Gamma^t = \mathbf{I}_p$.

Theorem 2.1 *A p-vector $\mathbf{X}$ has a spherical distribution if and only if its characteristic function $\Psi(t)$ satisfies one of the following equivalent conditions:*

1. $\Psi(\Gamma^t\mathbf{t}) = \Psi(\mathbf{t})$ *for any* $\Gamma \in \mathcal{O}(p)$*;*
2. *There exists a function $\phi(\cdot)$ of a scalar variable such that $\Psi(\mathbf{t}) = \phi(\mathbf{t}^t\mathbf{t})$.*

We write $\mathbf{X} \sim S_p(\phi)$ to mean that $\mathbf{X}$ is a p-vector with a spherical distribution and characteristic function $\Psi(\mathbf{t}) = \phi(\mathbf{t}^t\mathbf{t})$. Furthermore, we write $\mathbf{X} \sim S_p^+(\phi)$ to mean that $\mathbf{X} \sim S_p(\phi)$ and $\mathbb{P}\{\mathbf{X} = \mathbf{0}\} = 0$.

Example 2.1 Let $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$. Since the components $X_i \sim \mathcal{N}(0,1), i = 1, \ldots, p$ are independent and the characteristic function of X_i is $\exp(-t_i^2/2)$, the characteristic function of $\mathbf{X}$ is

$$\exp\{-\frac{1}{2}(t_1^2 + \ldots + t_p^2)\} = \exp\{-\frac{1}{2}\mathbf{t}^t\mathbf{t}\}.$$

From Theorem 2.1 it then follows that $\mathbf{X} \sim S_p(\phi)$ where $\phi(u) = \exp(-u/2)$. Furthermore, since $\mathbb{P}\{\mathbf{X} = \mathbf{0}\} = 0$, $\mathbf{X} \sim S_p^+(\phi)$.

Spherical distributions are distributions of uncorrelated random variables. However, among the spherical distributions only $\mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$ is the distribution of independent random variables.

Theorem 2.2 *Let $\mathbf{X}$ be a random p-vector. The following statements are equivalent.*

1. $\Gamma\mathbf{X} =_d \mathbf{X}$ *for every* $\Gamma \in \mathcal{O}(p)$*;*
2. $\mathbf{X}$ *has a stochastic representation* $\mathbf{X} =_d R\,\mathbf{U}^{(p)}$ *for some* $R \geq 0$ *which is independent of* $\mathbf{U}^{(p)}$*, a random vector uniformly distributed on the unit hyper sphere* $\{\mathbf{z} \in \mathbb{R}^p | \mathbf{z}^t\mathbf{z} = 1\}$.

Theorem 2.3 *Suppose* $\mathbf{X} =_d R\,\mathbf{U}^{(p)} \sim S_p^+(\phi)$, *then*

$$\|\mathbf{X}\| =_d R, \quad \mathbf{X}/\|\mathbf{X}\| =_d \mathbf{U}^{(p)}.$$

Moreover, $\|\mathbf{X}\|$ *and* $\mathbf{X}/\|\mathbf{X}\|$ *are independent.*

Note that if $R\,\mathbf{U}^{(p)} \sim S_p^+(\phi)$, then $\mathbb{P}\{R = 0\} = 0$.

Theorem 2.4 *If* $\mathbf{X} =_d R\,\mathbf{U}^{(p)} \sim S_p^+(\phi)$, *then all marginal distributions of* $\mathbf{X}$ *possess densities.*

We now continue with extending the spherical distributions to the elliptical distributions. From the extension many properties of elliptical distributions can be derived from properties of spherical distributions.

Definition 2 A random p-vector $\mathbf{X}$ is said to have an elliptical distribution with parameters μ and Σ if

$$\mathbf{X} =_d \mu + A\mathbf{Y}, \quad \mathbf{Y} \sim S_k(\phi)$$

where $A : p \times k$ and $AA^t = \Sigma$ with $\text{rank}(\Sigma) = k$. We write $\mathbf{X} \sim E_p(\mu, \Sigma, \phi)$.

As a direct consequence of Theorem 2.2 $\mathbf{X}$ has the stochastic representation

$$\mathbf{X} =_d \mu + RA\mathbf{U}^{(k)},$$

where $R \geq 0$ is independent of $\mathbf{U}^{(k)}$ and $AA^t = \Sigma$.
Furthermore, if $\mathbb{P}\{\mathbf{X} = \mu\} = 0$ (or equivalently $\mathbb{P}\{R = 0\} = 0$) and Σ is positive definite then all marginal distributions of $\mathbf{X}$ possess densities.
Remark 1: If the condition $\mathbb{P}\{R = 0\} = 0$ is replaced by the condition that R has a density, then also $\mathbf{X}$ has a density.
Remark 2: If Σ is positive definite, then

$$\mathbf{X} =_d \mu + A\mathbf{Y}, \quad \mathbf{Y} \sim S_p(\phi)$$

where $A : p \times p$ and $AA^t = \Sigma$ with $\text{rank}(\Sigma) = p$.
The following corollary will prove useful for proving the main result of this chapter.

Corollary 2.1 *Let* $\mathbf{X} \sim E_p(\mu, \Sigma, \phi)$ *and* $\mathbf{X}' \sim E_p(\mu', c\Sigma, \phi')$ *for* $c > 0$ *be independent. Then for* $a, b \in \mathbb{R}$, $a\mathbf{X} + b\mathbf{X}' \sim E_p(a\mu + b\mu', \Sigma, \tilde{\phi})$ *with* $\tilde{\phi}(u) = \phi(a^2u)\phi'(b^2cu)$.

Proof From the stochastic representation $\mathbf{X} =_d \mu + A\mathbf{Y}$ and $\mathbf{X}' =_d \mu' + \sqrt{c}A\mathbf{Y}'$, where $\mathbf{Y} \sim S_k(\phi)$, $\mathbf{Y}' \sim S_k(\phi')$ and $AA^t = \Sigma$ with $A : p \times k$ it follows that $a\mathbf{X} + b\mathbf{X}' =_d a\mu + b\mu' + A(a\mathbf{Y} + b\sqrt{c}\mathbf{Y}')$. Let $\Gamma \sim \mathcal{O}(k)$.

$$\begin{aligned}
\Psi_{a\mathbf{Y}+b\sqrt{c}\mathbf{Y}'}(\Gamma^t\mathbf{t}) &= \Psi_{a\mathbf{Y}}(\Gamma^t\mathbf{t})\Psi_{b\sqrt{c}\mathbf{Y}'}(\Gamma^t\mathbf{t}) \\
&= \Psi_{\mathbf{Y}}(\Gamma^t a\mathbf{t})\Psi_{\mathbf{Y}'}(\Gamma^t b\sqrt{c}\mathbf{t}) \\
&= \phi((a\mathbf{t})^t(a\mathbf{t}))\phi'((b\sqrt{c}\mathbf{t})^t(b\sqrt{c}\mathbf{t})) \\
&= \Psi_{\mathbf{Y}}(a\mathbf{t})\Psi_{\mathbf{Y}'}(b\sqrt{c}\mathbf{t}) \\
&= \Psi_{a\mathbf{Y}}(\mathbf{t})\Psi_{b\sqrt{c}\mathbf{Y}'}(\mathbf{t}) \\
&= \Psi_{a\mathbf{Y}+b\sqrt{c}\mathbf{Y}'}(\mathbf{t})
\end{aligned}$$

From Theorem 2.1 it follows that $a\mathbf{Y} + b\sqrt{c}\mathbf{Y}' \sim S_k(\tilde{\phi})$ with $\tilde{\phi}(u) = \phi(a^2u)\phi'(b^2cu)$ and then from Definition 2 that $a\mathbf{X} + b\mathbf{X}' \sim E_p(a\mu + b\mu', \Sigma, \tilde{\phi})$. □

Example 2.2 Let $(X', Y')^t$ be an independent copy of $(X, Y)^t$. Kendall's tau between X and Y is defined

$$\tau(X, Y) = \mathbb{P}\{(X - X')(Y - Y') > 0\} - \mathbb{P}\{(X - X')(Y - Y') < 0\}.$$

Let $(X, Y)^t \sim \mathcal{N}_2(\mu, \Sigma)$, where

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}$$

with $\sigma_1, \sigma_2 > 0$. If $\rho = 1$ then $\tau(X, Y) = 1 - 0 = 1$ and if $\rho = -1$ then $\tau(X, Y) = 0 - 1 = -1$. If $|\rho| < 1$ then Σ is positive definite so that $(X, Y)^t$ has a density. Since

$$\begin{pmatrix} X - X' \\ Y - Y' \end{pmatrix} = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} X \\ Y \\ X' \\ Y' \end{pmatrix},$$

where $(X, Y, X', Y')^t$ has a 4-variate normal distribution it follows immediately that

$$(X - X', Y - Y')^t \sim \mathcal{N}_2(\mathbf{0}, 2\Sigma).$$

Set $\tilde{X} = X - X'$ and $\tilde{Y} = Y - Y'$. Then

$$\begin{aligned}
\tau(X, Y) &= \mathbb{P}\{\tilde{X}\tilde{Y} > 0\} - \mathbb{P}\{\tilde{X}\tilde{Y} < 0\} \\
&\quad \Big\{\tilde{X}, \tilde{Y} \text{ are continuous}\Big\} \\
&= 2\mathbb{P}\{\tilde{X}\tilde{Y} > 0\} - 1 \\
&= 2\mathbb{P}\{\tilde{X} > 0, \tilde{Y} > 0\} + 2\mathbb{P}\{\tilde{X} < 0, \tilde{Y} < 0\} - 1 \\
&\quad \Big\{(\tilde{X}, \tilde{Y}) \text{ is symmetric about } (0, 0)\Big\} \\
&= 4\mathbb{P}\{\tilde{X} > 0, \tilde{Y} > 0\} - 1 \\
&= \frac{1}{\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \int_0^\infty \int_0^\infty \exp\{-\frac{1}{4(1-\rho^2)}(\frac{u^2}{\sigma_1^2} - \frac{2\rho uv}{\sigma_1\sigma_2} + \frac{v^2}{\sigma_2^2})\}\,\mathrm{d}u\,\mathrm{d}v - 1 \\
&\quad \Big\{s = (u/\sigma_1 - \rho v/\sigma_2)/\sqrt{1-\rho^2}, t = v/\sigma_2, \mathrm{d}u\,\mathrm{d}v = \sigma_1\sigma_2\sqrt{1-\rho^2}\,\mathrm{d}s\,\mathrm{d}t\Big\} \\
&= \frac{1}{\pi} \int_0^\infty \int_{-\rho t/\sqrt{1-\rho^2}}^\infty \exp\{-\frac{1}{4}(s^2 + t^2)\}\,\mathrm{d}s\,\mathrm{d}t - 1 \\
&\quad \Big\{s = r\cos\theta, t = r\sin\theta, \mathrm{d}s\,\mathrm{d}t = r\,\mathrm{d}r\,\mathrm{d}\theta\Big\} \\
&= \frac{1}{\pi} \int_0^{\pi/2 + \arctan(\rho/\sqrt{1-\rho^2})} \int_0^\infty r\exp\{-r^2/4\}\,\mathrm{d}r\,\mathrm{d}\theta - 1 \\
&= \frac{2}{\pi}(\frac{\pi}{2} + \arctan(\rho/\sqrt{1-\rho^2})) - 1 \\
&= \frac{2}{\pi}\arctan(\rho/\sqrt{1-\rho^2}) = \frac{2}{\pi}\arcsin(\rho).
\end{aligned}$$

Thus for $\mathbf{X} \sim \mathcal{N}_p(\mu, \Sigma)$ with $\Sigma_{ii} > 0$ for $i = 1, \dots, p$

$$\tau(X_i, X_j) = \frac{2}{\pi}\arcsin(\rho_{ij}),$$

where $i, j \in \{1, \dots, p\}$ and $\rho_{ij} = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$.

Elliptical distributions share many of the attractive properties of the multivariate normal distributions.

- Any linear combination of an elliptically distributed random vector is elliptical. If $\mathbf{X} \sim E_k(\mu, \Sigma, \phi)$, B is an $p \times k$ matrix and $\mathbf{b}$ is a $p \times 1$ vector, then

$$\mathbf{b} + B\mathbf{X} \sim E_k(\mathbf{b} + B\mu, B\Sigma B^t, \phi).$$

- All marginal distributions of elliptical distributions are elliptical and have the same generator. If $\mathbf{X} \sim E_p(\mu, \Sigma, \phi)$ and if $\mathbf{X} = \binom{\mathbf{X}_1}{\mathbf{X}_2}$, where $\mathbf{X}_1 \in \mathbb{R}^q$ and $\mathbf{X}_2 \in \mathbb{R}^r$ $(q + r = p)$, where $\mu = \left(\begin{array}{c} \mu_\mathbf{1} \\ \mu_\mathbf{2} \end{array} \right)$ ($\mu_\mathbf{1} \in \mathbb{R}^q$ and $\mu_\mathbf{2} \in \mathbb{R}^r$) and $\Sigma = \left(\begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{array} \right)$, then

$$\mathbf{X}_1 \sim E_q(\mu_1, \Sigma_{11}, \phi), \mathbf{X}_2 \sim E_r(\mu_2, \Sigma_{22}, \phi).$$

- If Σ is positive definite, then the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2$ is elliptical, however in general with a different generator.

Since the type of all marginal distributions is the same, it follows that an elliptical distribution is uniquely determined by its mean, its covariance matrix and knowledge of the type of its marginal distributions.

Knowledge of the distribution of $\mathbf{X}$ does not completely determine the elliptical representation $E_p(\mu, \Sigma, \phi)$. It uniquely determines μ but Σ and ϕ are only determined up to a positive constant. However, Σ can be chosen so that it is interpretable as the covariance matrix of $\mathbf{X}$ if $\mathbb{E}[R^2] < \infty$.

Even when $\mathrm{Cov}(\mathbf{X})$ is not defined, Σ will be called a covariance matrix if it is positive semi definite. If $\Sigma_{ii} > 0, i = 1, \ldots, p$ then $D\Sigma D$, where D is a diagonal matrix with diagonal elements $1/\sqrt{\Sigma_{ii}}$, will be called the linear correlation matrix corresponding to Σ.

Example 2.3 If $\mathbf{X}$ has the stochastic representation

$$\mathbf{X} =_d \mu + \frac{\sqrt{\nu}}{\sqrt{S}}\mathbf{Z},$$

where $\mu \in \mathbb{R}^p$, $S \in \chi^2_\nu$ and $\mathbf{Z} \in \mathcal{N}_p(\mathbf{0}, \Sigma)$, then $\mathbf{X}$ has a p-variate t_ν-distribution with mean μ (for $\nu > 1$) and covariance matrix $\frac{\nu}{\nu-2}\Sigma$ (for $\nu > 2$). If $\nu \leq 2$ then $\mathrm{Cov}(\mathbf{X})$ is not defined. In this case we just interpret Σ as being the shape parameter of the distribution of $\mathbf{X}$.

If a random vector $\mathbf{X}$ is such that

$$\mathbf{X} =_d \mu + A\mathbf{Y},$$

where $\mu \in \mathbb{R}^p$, $A : p \times k$ and $\mathbf{Y} \sim \mathcal{N}_k(\mathbf{0}, \mathbf{I}_k)$, then $\mathbf{X}$ has a multivariate normal distribution $\mathcal{N}_p(\mu, \Sigma)$ with $\Sigma = AA^t$.
Since $\mathbf{Y} \sim S_k^+(\phi)$ with $\phi(u) = \exp(-u/2)$ we have that $\mathbf{X} \sim E_p(\mu, \Sigma, \phi)$ and

$$\mathbf{X} =_d \mu + RA\mathbf{U}^{(k)},$$

with $R^2 =_d \|\mathbf{Y}\|^2 \sim \chi^2_k$. Here R has a density and hence $\mathbf{X} \sim \mathcal{N}_p(\mu, \Sigma)$ has a density if Σ is positive definite.

In the previous section we saw that calculating the Kendall's tau rank correlation matrix for a multivariate normal distribution is straightforward. For other elliptical distributions these straight-on calculations become much more difficult or impossible. This calls for a different approach.

Now consider $\mathbf{Z} \sim E_p(\mathbf{0}, \Sigma, \phi)$ with $\mathbb{P}\{\mathbf{Z} = \mathbf{0}\} = 0$. $\mathbf{Z}$ has the stochastic representation

$$\mathbf{Z} =_d RA\mathbf{U}^{(k)},$$

where R (> 0 a.s.) is independent of $\mathbf{U}^{(k)}$ and $AA^t = \Sigma$. By using the fact that $\mathbf{Y} =_d \|\mathbf{Y}\|\mathbf{U}^{(k)}$ for $\mathbf{Y} \sim S_k^+$ we get

$$\mathbf{Z} =_d \frac{R}{\|\mathbf{Y}\|} A\|\mathbf{Y}\|\mathbf{U}^{(k)},$$

for $R/\|\mathbf{Y}\| > 0$ and $\mathbf{Y} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$ with $\Sigma = AA^t$.
This gives the following theorem.

Theorem 2.5 *If* $\mathbf{Z} \sim E_p(\mu, \Sigma, \phi)$ *with* $\mathbb{P}\{\mathbf{Z} = \mu\} = 0$, *then* $\mathbf{Z}$ *has the stochastic representation*

$$\mathbf{Z} =_d \mu + R\mathbf{X},$$

where $R > 0$ *a.s. and* $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$.

Note that R and $\mathbf{X}$ are not necessarily independent. If R and $\mathbf{X}$ are independent, then $\mathbf{Z}$ is called a mixture of normal distributions.

Example 2.4 Let $(X', Y')^t$ be an independent copy of $(X, Y)^t$, where $(X, Y)^t \in E_2(\mu, \Sigma, \phi)$ with $\mathbb{P}\{(X, Y) = (\mu_1, \mu_2)\} = 0$ and Σ is positive definite. From Theorem 2.5 it follows that $(X, Y)^t =_d (\mu_1, \mu_2)^t + R(Z, W)^t$ and $(X', Y')^t =_d (\mu_1, \mu_2)^t + R'(Z', W')^t$, where $R, R' > 0$ are independent and $(Z, W)^t, (Z', W')^t \sim \mathcal{N}_2(\mathbf{0}, \Sigma)$ are independent. Furthermore, if R and $(Z, W)^t$ are independent then an expression for Kendall's tau can be obtained as follows:
Set $\tilde{X} = X - X'$ and $\tilde{Y} = Y - Y'$.

$$\begin{aligned}\mathbb{P}\{\tilde{X} > 0, \tilde{Y} > 0\} &= \mathbb{E}[\mathbf{1}_{\{\tilde{X}>0,\tilde{Y}>0\}}] = \mathbb{E}[\mathbb{E}[\mathbf{1}_{\{\tilde{X}>0,\tilde{Y}>0\}} \mid (R, R')]] \\ &= \int_0^\infty \int_0^\infty \mathbb{P}\{rZ - r'Z' > 0, rW - r'W' > 0\}\, \mathrm{d}F_R(r)\, \mathrm{d}F_{R'}(r')\end{aligned}$$

Furthermore,

$$\begin{pmatrix} rZ - r'Z' \\ rW - r'W' \end{pmatrix} = \begin{pmatrix} r & 0 & -r' & 0 \\ 0 & r & 0 & -r' \end{pmatrix} \begin{pmatrix} Z \\ W \\ Z' \\ W' \end{pmatrix} \sim \mathcal{N}_2(\mathbf{0}, (r^2 + r'^2)\Sigma),$$

and hence

$$\mathbb{P}\{rZ - r'Z' > 0, rW - r'W' > 0\} = \frac{1}{4} + \frac{1}{2\pi}\arcsin(\rho)$$

which does not depend on r and r'. Thus

$$\begin{aligned}&\mathbb{P}\{\tilde{X} > 0, \tilde{Y} > 0\} \\ =\ & \int_0^\infty \int_0^\infty \mathbb{P}\{rZ - r'Z' > 0, rW - r'W' > 0\}\, \mathrm{d}F_R(r)\, \mathrm{d}F_{R'}(r') \\ =\ & (\frac{1}{4} + \frac{1}{2\pi}\arcsin(\rho)) \int_0^\infty \mathrm{d}F_R(r) \int_0^\infty \mathrm{d}F_{R'}(r') \\ =\ & \frac{1}{4} + \frac{1}{2\pi}\arcsin(\rho).\end{aligned}$$

Hence we get the following result

$$\begin{aligned}\tau(X, Y) &= \mathbb{P}\{(X - X')(Y - Y') > 0\} - \mathbb{P}\{(X - X')(Y - Y') < 0\} \\ &= 4\,\mathbb{P}\{X - X' > 0, Y - Y' > 0\} - 1 = \frac{2}{\pi}\arcsin(\rho).\end{aligned}$$

As shown in the example above if $\mathbf{Z} \sim E_p(\mu, \Sigma, \phi)$ is a mixture of normal distributions with $\mathbb{P}\{\mathbf{Z} = \mu\} = 0$ and $\Sigma_{ii} > 0$ for $i = 1, \dots, p$ then

$$\tau(Z_i, Z_j) = \frac{2}{\pi} \arcsin(\rho_{ij}),$$

where $i, j \in \{1, \dots, p\}$ and $\rho_{ij} = \Sigma_{ij} / \sqrt{\Sigma_{ii}\Sigma_{jj}}$.

In fact this is just a special case of a much more general result.

Theorem 2.6 *Let $\mathbf{Z} \sim E_p(\mu, \Sigma, \phi)$, where $\mathbb{P}\{\mathbf{Z} = \mu\} = 0$ and Σ has non-zero diagonal elements. Then*

$$\tau(Z_i, Z_j) = \frac{2}{\pi} \arcsin(\rho_{ij}),$$

where $i, j \in \{1, \dots, p\}$ and $\rho_{ij} = \Sigma_{ij} / \sqrt{\Sigma_{ii}\Sigma_{jj}}$.

Proof Let $\mathbf{Z}'$ be an independent copy of $\mathbf{Z} =_d \mu + AR\mathbf{U}$, where $R\mathbf{U} \sim S_p(\phi)$ and $AA^t = \Sigma$. Set $\tilde{\mathbf{Z}} := \mathbf{Z} - \mathbf{Z}' =_d A(R\mathbf{U} - R'\mathbf{U}')$. From Corollary 2.1 it follows that $\tilde{\mathbf{Z}} =_d A\tilde{R}\tilde{\mathbf{U}} \sim E_p(\mathbf{0}, \Sigma, \tilde{\phi})$ and hence

$$\mathbb{P}\{\tilde{Z}_i > 0, \tilde{Z}_j > 0\} = \mathbb{P}\{c\tilde{Z}_i > 0, c\tilde{Z}_j > 0\}$$

for all $c > 0$. Since $\mathbb{P}\{\mathbf{Z} = \mu\} = 0$ it follows that $\mathbb{P}\{\tilde{R} = 0\} = 0$. Set $\tilde{\mathbf{W}} = A\tilde{\mathbf{U}}$. Then

$$\begin{aligned}
\mathbb{P}\{\tilde{Z}_i > 0, \tilde{Z}_j > 0\} &= \mathbb{P}\{\tilde{R}\tilde{W}_i > 0, \tilde{R}\tilde{W}_j > 0\} \\
&= \int_0^\infty \mathbb{P}\{r\tilde{W}_i > 0, r\tilde{W}_j > 0 | \tilde{R} = r\} \, \mathrm{d}F_{\tilde{R}}(r) \\
&= \int_0^\infty \mathbb{P}\{r\tilde{W}_i > 0, r\tilde{W}_j > 0\} \, \mathrm{d}F_{\tilde{R}}(r) \\
&= \int_0^\infty \mathbb{P}\{\tilde{W}_i > 0, \tilde{W}_j > 0\} \, \mathrm{d}F_{\tilde{R}}(r) = \mathbb{P}\{\tilde{W}_i > 0, \tilde{W}_j > 0\}.
\end{aligned}$$

The distribution of $\tilde{\mathbf{W}}$ does not depend on the particular elliptical family of $\mathbf{Y}$. As a consequence

$$\tau(Z_i, Z_j) = 4\mathbb{P}\{\tilde{Z}_i > 0, \tilde{Z}_j > 0\} - 1 = 4\mathbb{P}\{X_i > 0, X_j > 0\} - 1,$$

where $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$. Hence $\tau(Z_i, Z_j) = \frac{2}{\pi} \arcsin(\rho_{ij})$. □

2.2 Contaminating Elliptical Distributions. The aim of this paper is finding linear correlation estimators having a high efficiency for heavier tailed elliptical distributions as well as for multivariate normal distributions which are insensitive to mild contamination of the data. With mild contamination is meant that from observations it should at least be reasonable to assume that the uncontaminated underlying distribution belongs to the class of elliptical distributions. A few suggestions for distributions which can be thought of as contaminated elliptical distributions are presented below. The samples from these distributions resemble observed financial data such as stationary asset return time series.

Example 2.5 Let

$$\mathbf{Z} =_d \mu + (R + (1 - R)\lambda)\mathbf{X},$$

where $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$ is independent of $R \sim \mathrm{Be}(q)$ (Be denotes the Bernoulli distribution) and $\lambda > 0$. $\mathbf{Z}$, elliptically distributed, is a two point normal mixture. This is a mixture of $100q\%$ $\mathcal{N}_p(\mu, \Sigma)$ and $100(1 - q)\%$ $\mathcal{N}_p(\mu, \lambda^2\Sigma)$.

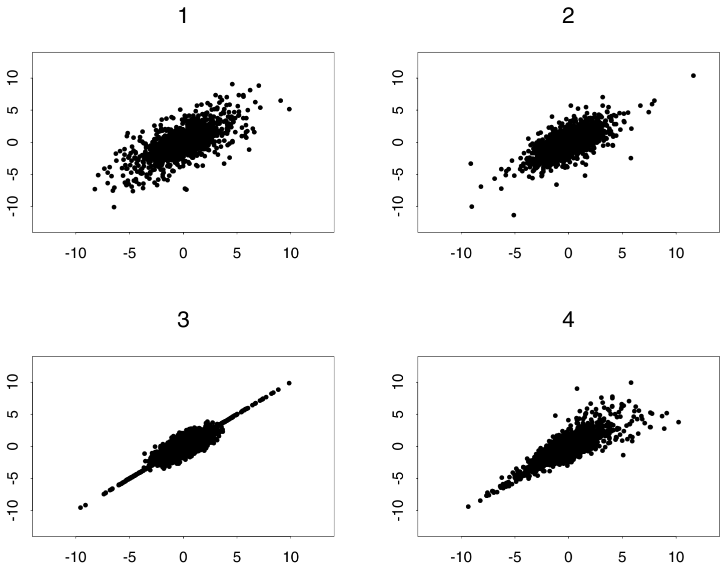

Figure 2.1 Bivariate samples of size 6000 from the distributions described in example 2.5, 2.6, 2.7 and 2.8 respectively with $(q, \lambda) = (0.9, 9)$ and for 2, $(\nu_1, \nu_2) = (8, 2)$.

Example 2.6 Let

$$\mathbf{Z} =_d \mu + \left(R\frac{\sqrt{\nu_1}}{\sqrt{S_1}} + (1-R)\frac{\sqrt{\nu_2}}{\sqrt{S_2}} \right) \mathbf{X},$$

where $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$ and independent of $R \sim \mathrm{Be}(q)$, $S_1 \sim \chi^2_{\nu_1}$ and $S_2 \sim \chi^2_{\nu_2}$. $\mathbf{Z}$, elliptically distributed, is a two point mixture of t_{ν_1}- and t_{ν_2}-distributions. This is a mixture of $100q\%$ $\mathrm{t}_{\nu_1}(\mu, \Sigma)$ and $100(1-q)\%$ $\mathrm{t}_{\nu_2}(\mu, \Sigma)$.

Example 2.7 Let

$$\mathbf{Z} =_d \mu + R\mathbf{X} + (1-R)\lambda(Y, \ldots, Y)^t,$$

where $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$, $Y \sim \mathcal{N}(0, 1)$ and $R \sim \mathrm{Be}(q)$ are pairwise independent. $\mathbf{Z}$, not elliptically distributed, will be said to have a comonotonic contaminated multivariate normal distribution. That is, with probability $100(1-q)\%$ an observation is a normally distributed random vector where all pairwise correlations are one.

Example 2.8 Let

$$\mathbf{Z} =_d \mu + R\mathbf{X} + (1-R)\lambda\mathbf{Y},$$

where $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$, $\mathbf{Y}$ and $R \sim \mathrm{Be}(q)$ are pairwise independent. The distribution function of $\mathbf{Y}$ is $C(\Phi(x_1), \ldots, \Phi(x_p))$, where Φ is the standard normal distribution function and C is the Clayton copula

$$C(u_1, \ldots, u_p) = \varphi^{-1}(\varphi(u_1) + \cdots + \varphi(u_p)),$$

where $\varphi(u) = \left(u^{-\theta} - 1\right)/\theta$ and $\theta = 2\tau/(1-\tau)$. τ denotes the common Kendall's tau rank correlation coefficient for (Z_i, Z_j) $(i \neq j)$. The distribution of $\mathbf{Z}$ will be referred to as a Clayton-normal contaminated normal distribution. This is not an elliptical distribution.

3 Bivariate Correlation Estimators

There are a variety of methods for calculating a robust correlation matrix. One approach is to estimate each off-diagonal element separately by a robust correlation coefficient. The disadvantage with this approach is that the resulting matrix need not be positive semi definite (PSD) and hence may need adjustment to achieve this property. One advantage with this element-wise approach is that in case of missing data it uses the available data better than multivariate methods. One other advantage is its relatively low computational cost.

3.1 The Kendall's tau Transform. From the first chapter we know that the linear correlation coefficient for bivariate elliptical distributions is given by

$$\rho(X,Y) = \sin\left(\frac{\pi}{2}\tau(X,Y)\right), \tag{3.1}$$

where $\tau(X,Y)$ denotes Kendall's tau between X and Y. Since $\tau(X,Y)$ is the probability of concordance minus the probability of discordance for (X,Y) the sample version of Kendall's tau is given by

$$\hat{\tau} = \frac{c-d}{c+d},$$

where c denotes the number of concordant pairs and d the number of discordant pairs. More explicitly, consider a sample of n points (x_i, y_i). There is a total of $n(n-1)/2$ pairs of points $((x_i,y_i),(x_j,y_j))$, where a point cannot be paired with itself and two points in either order count as one pair. Recall that $((x_i,y_i),(x_j,y_j))$ is concordant if $(x_i - x_j)(y_i - y_j) > 0$ and discordant if $(x_i - x_j)(y_i - y_j) < 0$. If the (x,y)'s are observations from a continuous distribution, then the probability of ties in the x's or y's is zero. However, in practice it might occur that $x_i = x_j$ and/or $y_i = y_j$ in which case the pair is neither concordant nor discordant. If there is a tie in the x's the pair will be called an "extra y pair" and an "extra x pair" if there is a tie in the y's. The adjusted sample version of Kendall's tau is then given by

$$\hat{\tau} = \frac{c-d}{\sqrt{c+d+e_y}\sqrt{c+d+e_x}},$$

where e_x and e_y denotes the number of extra x and extra y pairs. Since

$$\mathbb{E}[\hat{\tau}] = \tau,$$

$\hat{\tau}$ is an unbiased estimator of τ and since (3.1) holds for all elliptical distributions

$$r_\tau = \sin\left(\frac{\pi}{2}\hat{\tau}\right)$$

is a natural estimator of linear correlation for elliptical distributions. It is however *not* an unbiased estimator of ρ since we would require that $\mathbb{E}[\sin\left(\frac{\pi}{2}\hat{\tau}\right)] = \rho$. Results on the behaviour of r_τ for multivariate normal distributions can be found in Kendall (1990) [14].

It should be noted that the algorithm for calculating Kendall's tau is an $O(n^2)$ algorithm since all $n(n-1)/2$ pairs $((x_i,y_i),(x_j,y_j))$ have to be checked. A C-implementation of the algorithm calculates $\hat{\tau}$ for data sets of a few thousand points in a few seconds, but calculating $\hat{\tau}$ for data sets of some hundred thousand points is not recommended.

3.1.1 *Bivariate Comparisons.* The table 3.1 below shows mean squared errors (MSEs) for the standard linear correlation estimator (the Pearson product-moment correlation)

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

and the linear correlation estimator provided by the Kendall's tau transform

$$r_\tau = \sin\left(\frac{\pi}{2}\hat{\tau}\right)$$

Distribution		MSE(r)			MSE(r_τ)		
		$n = 30$	$n = 90$	$n = 300$	$n = 30$	$n = 90$	$n = 300$
t_2	$\rho = 0.1$	0.14919	0.11590	0.08299	0.05367	0.01736	0.00529
	$\rho = 0.3$	0.13679	0.10109	0.07901	0.04741	0.01560	0.00441
	$\rho = 0.5$	0.12154	0.07659	0.06713	0.03671	0.01108	0.00332
	$\rho = 0.7$	0.06416	0.04546	0.03853	0.01905	0.00567	0.00175
	$\rho = 0.9$	0.01810	0.01204	0.01267	0.00398	0.00102	0.00026
t_4	$\rho = 0.1$	0.06903	0.03016	0.01157	0.04825	0.01492	0.00419
	$\rho = 0.3$	0.05849	0.02601	0.01029	0.04105	0.01314	0.00399
	$\rho = 0.5$	0.04024	0.01852	0.00699	0.02993	0.00932	0.00284
	$\rho = 0.7$	0.01973	0.00824	0.00354	0.01545	0.00449	0.00150
	$\rho = 0.9$	0.00421	0.00140	0.00046	0.00288	0.00076	0.00020
Normal	$\rho = 0.1$	0.03414	0.01120	0.00322	0.03911	0.01268	0.00365
	$\rho = 0.3$	0.02980	0.00971	0.00297	0.03580	0.01111	0.00329
	$\rho = 0.5$	0.02001	0.00640	0.00195	0.02533	0.00749	0.00218
	$\rho = 0.7$	0.01004	0.00310	0.00087	0.01311	0.00380	0.00099
	$\rho = 0.9$	0.00152	0.00041	0.00012	0.00234	0.00055	0.00015

Table 3.1 Mean squared errors for the r and r_τ for different sample sizes and distributions. The results were obtained using Monte Carlo simulation; 3000 independent samples of sample size $n = 30$, $n = 90$ and $n = 300$.

for different elliptical distributions, sample sizes and linear correlation coefficients. Recall that the mean squared error for an estimator $\hat{\rho}$ is

$$\mathrm{MSE}(\hat{\rho}) := \mathbb{E}[(\hat{\rho} - \rho)^2] = \mathrm{Var}(\hat{\rho}) + \big(\mathbb{E}[\hat{\rho} - \rho]\big)^2.$$

The results were obtained using Monte Carlo simulation, where the MSEs were calculated on 3000 independent samples of the respective sample sizes.

Even though r is more biased for heavier tailed distributions, its large MSE is caused by its high variance as seen in figure 3.1. Figure 3.2 shows the performance of r and r_τ for samples from the two point normal mixture with common correlation matrix, which is a mixture of 90% $\mathcal{N}(\mathbf{0}, R)$ and 10% $\mathcal{N}(\mathbf{0}, 9R)$, where R is a 2×2 correlation matrix with off diagonal elements 0.5.

The standard estimator applied on the data presented in figure 3.1 gives a bias -0.012, a variance 0.031 and a mean square error 0.031. The Kendall's tau transform estimator applied on this data gives a bias -0.004, a variance 0.010 and a mean square error 0.010. Hence the three times bigger MSE for the standard estimator is caused by its big variance. Furthermore, for the standard estimator the worst overestimate is 0.961 and the worst underestimate is -0.765!

3.2 Estimation Through Robust Variances. For two random variables X and Y with finite variances

$$\begin{aligned}\mathrm{Var}(X+Y) - \mathrm{Var}(X-Y) &= 4\mathrm{Cov}(X,Y)\\ \mathrm{Var}(X+Y) + \mathrm{Var}(X-Y) &= 2\mathrm{Var}(X) + 2\mathrm{Var}(Y).\end{aligned}$$

If X and Y are scaled to have unit variances, that is

$$\tilde{X} = X/\sqrt{\mathrm{Var}(X)} \text{ and } \tilde{Y} = Y/\sqrt{\mathrm{Var}(Y)},$$

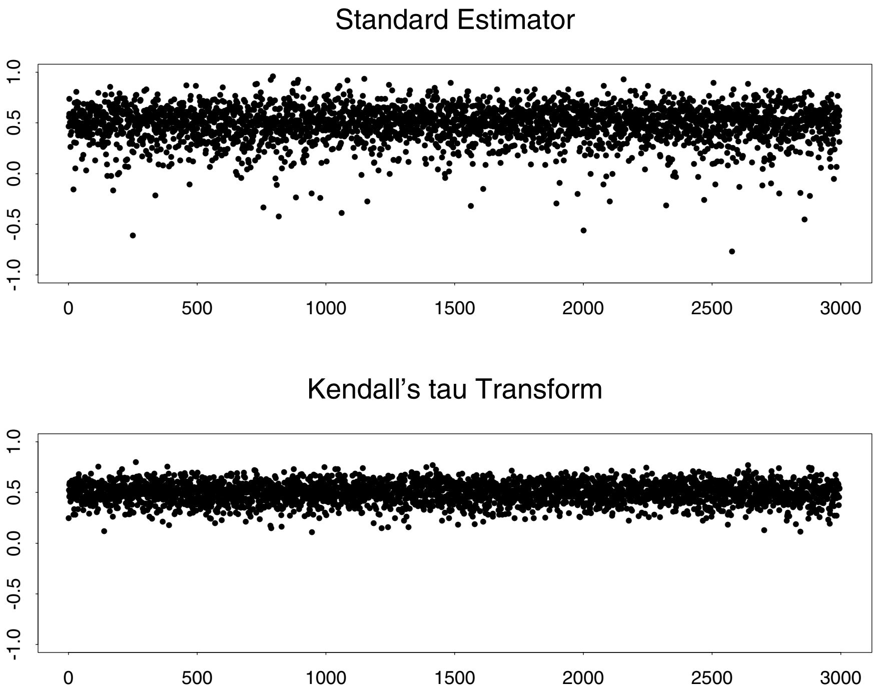

Figure 3.1 Linear correlation estimates for 3000 independent samples of size 90 from a bivariate t_3-distribution with linear correlation 0.5.

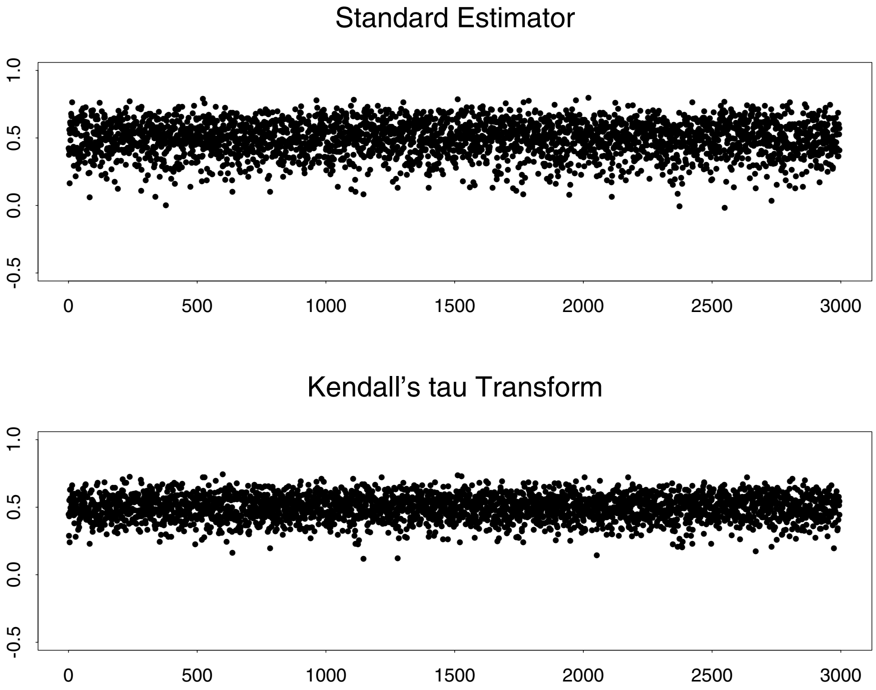

Figure 3.2 Linear correlation estimates for 3000 independent samples of size 90 from a two point bivariate normal mixture with common linear correlation 0.5.

then

$$\frac{\mathrm{Var}(\tilde{X}+\tilde{Y})-\mathrm{Var}(\tilde{X}-\tilde{Y})}{\mathrm{Var}(\tilde{X}+\tilde{Y})+\mathrm{Var}(\tilde{X}-\tilde{Y})}=\rho(X,Y).$$

Hence given a robust variance estimator v

$$r_{\mathrm{SSD}}(X,Y)=\frac{v(\tilde{X}+\tilde{Y})-v(\tilde{X}-\tilde{Y})}{v(\tilde{X}+\tilde{Y})+v(\tilde{X}-\tilde{Y})},$$

is a robust estimator of $\rho(X,Y)$, where

$$v(\tilde{X}\pm\tilde{Y})=v(X/\sqrt{\mathrm{Var}(X)}\pm Y/\sqrt{\mathrm{Var}(Y)}).$$

The variances of X and Y are estimated by some robust estimator, for example v. The estimator is denoted r_{SSD} since it is based on standardized variables and variances of their sums and differences. It was proposed by Gnanadeskian and Kettenring (1972) [7]. One interesting property of r_{SSD} is that the multiplicative constant depending on the sample size and the underlying distribution, which is required for removing the bias in some commonly used variance estimators, cancels out by appearing in both the numerator and denominator in the expression for r_{SSD}. Furthermore, $r_{\mathrm{SSD}}(X,Y)\in[-1,1]$ by construction.

We will now continue by presenting one possible robust variance estimator based on symmetrically trimming of the data.

From a univariate sample of size n, if $k=n-2\lceil\beta n\rceil$ observations, relabeled $x_{(1)}\leq x_{(2)}\leq\cdots\leq x_{(k)}$ ($x_{(i)}$ is the $\lceil\beta n\rceil+i$ smallest observation), are left after trimming $\lceil\beta n\rceil$ observations from each end, the β-trimmed estimate of the population variance σ^2 is given by

$$\tilde{\sigma}^2=c(n,k)\frac{1}{k-1}\sum_{i=1}^{k}(x_{(i)}-\tilde{\mu})^2,$$

where $\tilde{\mu}$ is a robust estimate of the population mean, such as an α-trimmed mean, and $c(n,k)$ is a correction for the bias. Using this variance estimator for the correlation estimator above means that $c(n,k)$ cancels out by appearing in both the numerator and denominator. Hence we can set $c(n,k)=1$.

We now consider the so called α-trimmed mean. Since the number $(1-2\alpha)n$ of observations to be trimmed might not be an integer, we set $\alpha n=m+f$, where m is an integer and $0\leq f<1$. The α-trimmed mean is then

$$\tilde{\mu}=\frac{(1-f)x_{(m+1)}+x_{(m+2)}+\cdots+x_{(n-m-1)}+(1-f)x_{(n-m)}}{n(1-2\alpha)}.$$

Unless otherwise stated, we will take v to be an α-trimmed variance with $\tilde{\mu}$ being an α-trimmed mean.

For the variance, and to a less extent for the mean, there is a possible conflict between the desire to protect the estimate from outliers and the fact that the information for estimating the variance relies heavily on the tails. Therefore it might be advisable to use a smaller proportion of trimming for the variance estimation than for the mean estimation.

4 Transformations of non-PSD correlation matrices

Since some robust correlation estimators do not extend to the multivariate case, the correlation matrix has to be estimated by estimating pairwise correlations. In this case, the resulting matrix of estimated pairwise correlations is not necessarily positive semi-definite (PSD). Hence we are interested in techniques for transforming non-PSD "correlation matrices" to PSD correlation matrices such that the transformed matrix is "close" to the original matrix.

A $p \times p$ matrix $R \in \mathbb{R}^{n \times n}$ is called a *pseudo-correlation* matrix if

1. R is symmetric;
2. $r_{ii} = 1$ for $i = 1, \ldots, p$;
3. $|r_{ij}| \leq 1$ for $i, j = 1, \ldots, p$.

A correlation matrix is a PSD pseudo-correlation matrix.

One possible measure of distance between two matrices $R, \tilde{R} \in \mathbb{R}^{p \times p}$ is the sum of the squared Euclidean distances between the entries of R and $\tilde{R}$,

$$d^2(R, \tilde{R}) = \sum_{i=1}^{p} \sum_{j=1}^{p} (r_{ij} - \tilde{r}_{ij})^2.$$

Clearly,

$$d^2(R, \tilde{R}) > 0 \text{ if } R \neq \tilde{R}, d^2(R, R) = 0 \text{ and } d^2(R, \tilde{R}) = d^2(\tilde{R}, R).$$

However, since d^2 does not satisfy the triangle inequality d^2 is not a metric on the set of real $p \times p$ matrices.

Another measure of distance between two matrices R and $\tilde{R}$ is

$$d_a(R, \tilde{R}) = \sum_{i=1}^{p} \sum_{j=1}^{p} |r_{ij} - \tilde{r}_{ij}|.$$

Since d_a also satisfies $d_a(R, \tilde{R}) \leq d_a(R, R') + d_a(R', \tilde{R})$ for any $R' \in \mathbb{R}^{p \times p}$, d_a is a metric. Although d^2 is not a metric we will use it as our measure of distance and regard $\tilde{R}$ as closer to R than R' if $d^2(R, \tilde{R}) < d^2(R, R')$.

4.1 Shrinking Methods. We begin with discussing a method, for transforming an non-PSD pseudo-correlation matrix to a correlation matrix, called *linear shrinking*.

Let R be a non-PSD pseudo-correlation matrix and let R^* be an arbitrary correlation matrix. Define the matrix $\tilde{R}$ as

$$\tilde{R} = R + \epsilon(R^* - R),$$

where $\epsilon \in [0, 1]$. Clearly $\tilde{R}$ is a pseudo-correlation matrix for all $\epsilon \in [0, 1]$ and also a correlation matrix for some $\epsilon \in [0, 1]$ since $\tilde{R} = R^*$ for $\epsilon = 1$. The idea is to find the smallest $\epsilon \in [0, 1]$ so that $\tilde{R}$ is a correlation matrix.

Since R is not PSD the smallest eigenvalue of R, which we denote by λ, is negative. If we chose ϵ so that $\tilde{\lambda} \geq 0$ (the smallest eigenvalue of $\tilde{R}$) then $\tilde{R}$ is a correlation matrix. Furthermore, we want to chose ϵ so that $d^2(R, \tilde{R})$ is as small as possible.

$$\tilde{\lambda} = \min_{\mathbf{x}^t\mathbf{x}=1} \mathbf{x}^t \tilde{R} \mathbf{x} = \min_{\mathbf{x}^t\mathbf{x}=1} \Big((1 - \epsilon)\mathbf{x}^t R \mathbf{x} + \epsilon \mathbf{x}^t R^* \mathbf{x} \Big) = (1 - \epsilon)\lambda + \epsilon \lambda^*.$$

Hence $\tilde{\lambda} \geq 0$ if $\epsilon \geq -\lambda/(\lambda^* - \lambda)$. Furthermore, since

$$d^2(R, \tilde{R}) = \sum_{i=1}^{p} \sum_{j=1}^{p} (r_{ij} - r_{ij} - \epsilon(r^*_{ij} - r_{ij}))^2 = \epsilon^2 \sum_{i=1}^{p} \sum_{j=1}^{p} (r^*_{ij} - r_{ij})^2 = \epsilon^2 d^2(R, R^*),$$

the smallest possible ϵ should be chosen.

If $R^* = I_p$ then using this method means shrinking R linearly towards the p-dimensional identity matrix. Other possible choices for R^* are robustly estimated correlation matrices.

One reason why linear shrinking with $R^* = I_p$ should be avoided is that since all elements of R are shrunk by the same relative amount, large elements (of either sign) are shrunk much more than small ones. That is

$$\frac{r_{ij} - \tilde{r}_{ij}}{r_{ij}} = \frac{r_{kl} - \tilde{r}_{kl}}{r_{kl}},$$

for $i \neq j$ and $k \neq l$. This is in general not what we want. To avoid this problem R can instead be shrunk *non linearly* towards I_p, such that large element are shrunk by a less relative amount than small elements. That is

$$\frac{r_{ij} - \tilde{r}_{ij}}{r_{ij}} < \frac{r_{kl} - \tilde{r}_{kl}}{r_{kl}},$$

if $|r_{ij}| > |r_{kl}|$.

Consider the element wise transformation procedure operating on the off diagonal elements of R given by

$$g(r_{ij}) = \begin{cases} f^{-1}(f(r_{ij}) + \Delta), & \text{if } r_{ij} < -f^{-1}(\Delta), \\ 0, & \text{if } |r_{ij}| \leq f^{-1}(\Delta), \\ f^{-1}(f(r_{ij}) - \Delta), & \text{if } r_{ij} > f^{-1}(\Delta), \end{cases} \tag{4.1}$$

where f is strictly increasing odd function with $f(0) = 0$ and $\Delta > 0$. This transformation procedure with $f = f_2$ (see below) was suggested by Gnanadeskian and Kettenring (1972) [7]. The non linear shrinking of R terminates if $g(R)$ is PSD, then $\tilde{R} = g(R)$ is the correlation matrix obtained from the shrinking procedure. If not, then g is simply applied to $g(R)$ and so on. That the procedure terminates is clear (if we assume an arbitrary good numerical precision) since $g^k(R) = I_p$ for some $k \in \mathbb{N}$. Clearly a small Δ, e.g. $\Delta = 0.001$, is likely to give a bigger k but a smaller $d^2(R, g^k(R))$. Four possible choices of f are

$$\begin{aligned} f_1(x) &= \tanh(x), \quad f_2(x) = \tanh^{-1}(x), \\ f_3(x) &= \frac{2}{\pi}\arctan(x), \quad f_4(x) = \tan(\pi x/2) \end{aligned}$$

and for each choice of f

$$\frac{r_{ij} - g(r_{ij})}{r_{ij}} < \frac{r_{kl} - g(r_{kl})}{r_{kl}},$$

if $|r_{ij}| > |r_{kl}|$.

Suppose that $\tilde{R} = g_i^N(R)$, where $N = \min\{k \in \mathbb{N} : g_i^k(R) \text{ is PSD}\}$. Which shrinking function f_i should be chosen so that $g_i^N(R)$ is the best PSD approximation of R?

In the worst case $\tilde{R} = g_i^N(R) = I_p$, where N is such that for the biggest off diagonal element r of R, $g_i^N(r) = 0$. Without loss of generality r can be assumed to be positive. What size on N can be expected in this worst case for the different choices of shrinking functions f?
For $i = 2$ and $i = 4$ we have

$$g_i^k(r) - g_i^{k+1}(r) < g_i^{k+1}(r) - g_i^{k+2}(r),$$

which means that small elements are shrunk with a bigger absolute amount than larger. Furthermore, $r - g_i(r)$ is a decreasing function of r.

$$\begin{aligned} \lim_{r \nearrow 1} (r - g_2(r)) &= \lim_{r \nearrow 1} \left(r - \tanh(\tanh^{-1}(r) - \Delta) \right) \\ &= \lim_{r \nearrow 1} \left(r - \frac{e^{-\Delta}\sqrt{(1+r)/(1-r)} - e^{\Delta}\sqrt{(1-r)/(1+r)}}{e^{-\Delta}\sqrt{(1+r)/(1-r)} + e^{\Delta}\sqrt{(1-r)/(1+r)}} \right) = 0, \end{aligned}$$

for all Δ and since $g_2^k(r) \leq g_4^k(r)$ for all k also

$$\lim_{r \nearrow 1} (r - g_4(r)) = 0$$

for all Δ. If ξ denotes the numerical precision of whatever computer is used, there are $r < 1 - \xi$ such that $r - g_i(r) < \xi$ for all choices of Δ. Thus for f_2 and f_4 applying non linear shrinking on random samples may lead to $g^k(R)$ being non-PSD for all k and hence the procedure might never terminate. If (which is more likely) $N < \infty$ but still quite big $g_i^N(R)$ will consist of the large elements of R virtually unchanged but the smaller ones replaced by zeros. The conclusion is that leaving large elements virtually unchanged has a high price in terms of the smaller elements being drastically shrunk towards zero. This disadvantage is more pronounced for f_4 than for f_2. Therefore there is no good reason for choosing f_4.

The choice of f_1 or f_3 is unproblematic but will result in the large elements being shrunk by a bigger absolute amount than the smaller ones (still however by a smaller relative amount). That is for $i = 1$ and $i = 3$

$$g_i^k(r) - g_i^{k+1}(r) > g_i^{k+1}(r) - g_i^{k+2}(r).$$

Since

$$\begin{aligned} \lim_{r \nearrow 1} (r - g_1(r)) &= \lim_{r \nearrow 1} \left(r - \tanh^{-1}(\tanh(r) - \Delta) \right) \\ &= \lim_{r \nearrow 1} \left(r - \frac{1}{2} \ln \frac{2e^r - \Delta(e^r + e^{-r})}{2e^{-r} + \Delta(e^r + e^{-r})} \right) > 0, \end{aligned}$$

and for any reasonable choice of Δ also bigger than the numerical precision, and

$$\begin{aligned} \lim_{r \searrow f_1^{-1}(\Delta)} (r - g_1(r)) &= f_1^{-1}(\Delta) - \lim_{r \searrow f_1^{-1}(\Delta)} \tanh^{-1}(\tanh(r) - \Delta) \\ &= f_1^{-1}(\Delta) - \lim_{r \searrow f_1^{-1}(\Delta)} \frac{1}{2} \ln \frac{2e^r - \Delta(e^r + e^{-r})}{2e^{-r} + \Delta(e^r + e^{-r})} > 0, \end{aligned}$$

which is also bigger than the numerical precision for any reasonable choice of Δ, $g_1^k(R)$ is PSD for some $k < \infty$. The same results hold for g_3 but since g_3 shrinks big elements more than g_1 and small elements less than g_1 it is hard to see any good reason for using f_3 instead of f_1.

The advantage of the linear and non linear shrinking techniques is that they are easy to apply, because they operate on each matrix element separately. A drawback with this is that the relation between the matrix elements is never used (except when testing whether the result is PSD). The non linear shrinking techniques preserve the correlation structure more than linear shrinking *if* the procedure terminates in a few iterations.

4.2 The Eigenvalue Method. Let R be a $p \times p$ pseudo-correlation matrix of correlation estimates. Since R is symmetric there exists a orthogonal matrix P such that

$$R = PDP^t,$$

where D is a diagonal matrix with the eigenvalues of R on the diagonal. If R is not PSD, then one or some of the elements in D is negative. Typically, the negative eigenvalues will have small

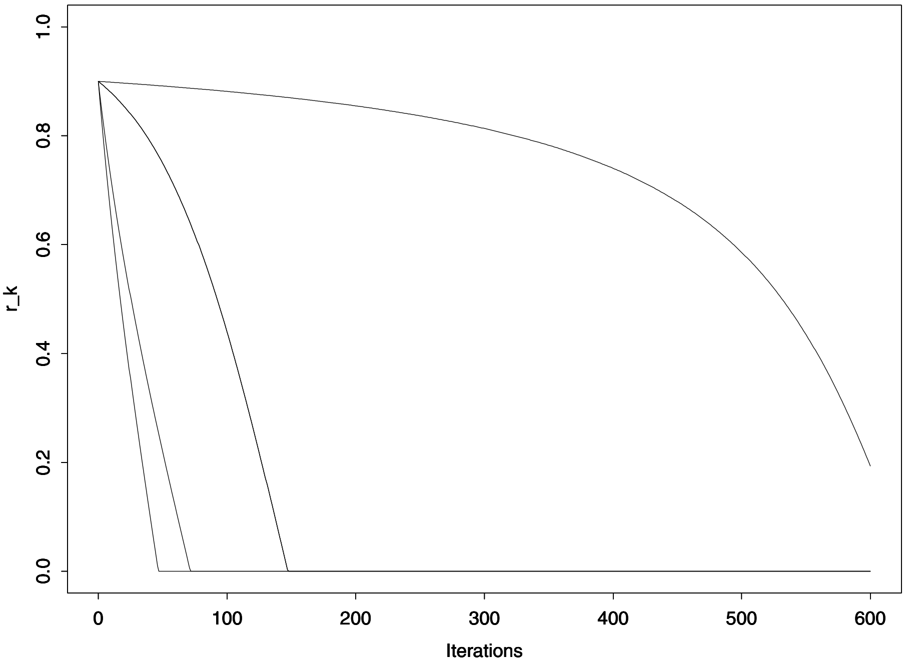

Figure 4.1 Non linear shrinking of $r = 0.9$ using different shrinking functions and $\Delta = 0.01$. From left to right the figure shows $g_3^k(r), g_1^k(r), g_2^k(r)$ and $g_4^k(r)$ for $k = 0, 1, \dots, 600$.

absolute values. A natural approach is to replace the negative eigenvalues by some constant $\delta \geq 0$. Set

$$R' = PD'P^t,$$

where D' is diagonal matrix D with negative elements replaced by $\delta \geq 0$. Clearly R' is symmetric and since

$$PD'P^tP_i = d'_{ii}P_i, i = 1, \dots, p,$$

where P_i denotes the ith column of P, R' has eigenvalues $d'_{11}, \dots, d'_{pp} \geq \delta$. Hence R' is PSD (PD for $\delta > 0$). However, the diagonal elements of R' will not necessarily equal 1. Therefor set

$$\tilde{R} = \tilde{D}R'\tilde{D},$$

where $\tilde{D}$ is a diagonal matrix with diagonal elements $1/\sqrt{r'_{ii}}, i = 1, \dots, p$. $\tilde{R}$ is the correlation matrix obtained from R by using the eigenvalue method. This method was described by Rousseeuw and Molenberghs (1993) [17].

Since the probability that two or more of the eigenvalues of R is equal is zero and eigenvectors corresponding to different eigenvalues are orthogonal, P is simply a matrix whose columns are the normalized eigenvectors of R. If (e.g. due to truncation) two eigenvalues are tied, then the matrix P is obtained by using the Gram-Schmidt procedure.

4.3 Comparisons Between Methods. Let R be a $p \times p$ random pseudo-correlation matrix with iid off-diagonal elements $U_1, \dots, U_l \sim U(-1, 1)$ $(l = \binom{p}{2})$. Let $\tilde{R}$ be the corresponding correlation matrix obtained from R using some method for transforming a non-PSD pseudo-correlation matrix to a correlation matrix. Note that $\tilde{R} = R$ if R is PSD. One possible measure

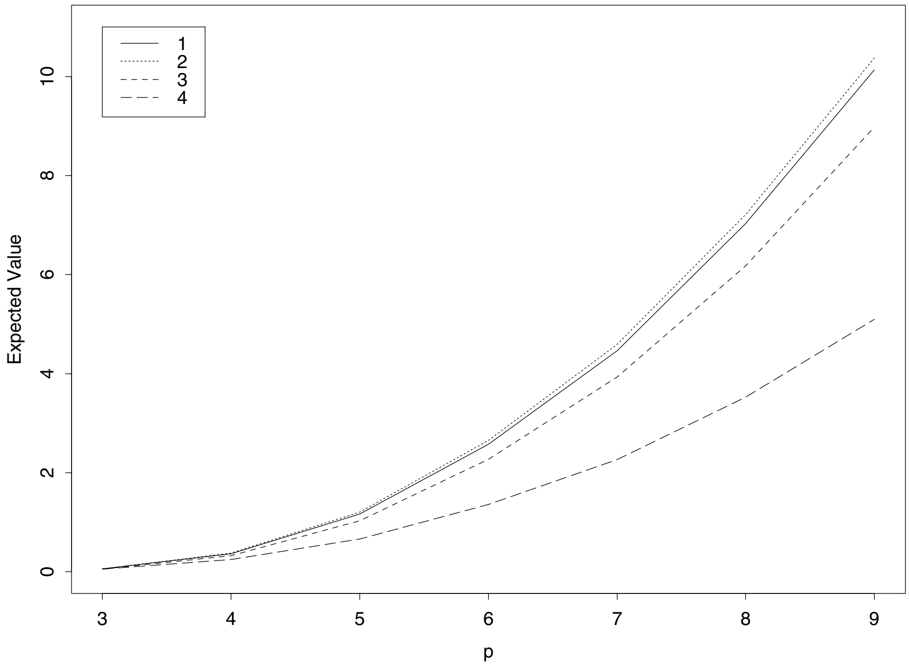

Figure 4.2 The performances of techniques 1 - 4 applied on 4000 random $p \times p$ pseudo-correlation matrices and measured in terms of $\mathbb{E}[d^2(R, \tilde{R})]$.

of the performance of the transformation methods is

$$\mathbb{E}[d^2(R, \tilde{R})].$$

The following techniques for transforming a non-PSD pseudo-correlation matrix to a PSD correlation matrix were compared

1. Non linear shrinking towards the identity matrix, with $f(x) = \tanh(x)$ and step size $\Delta = 0.01$.
2. Non linear shrinking towards the identity matrix, with $f(x) = \frac{2}{\pi} \arctan(x)$ and step size $\Delta = 0.01$.
3. Linear shrinking towards the identity matrix.
4. The eigenvalue method.

The results for $\mathbb{E}[d^2(R, \tilde{R}_i)]$, where i indicates which transformation method was used, were obtained using Monte Carlo simulation of 4000 independent $p \times p$ pseudo-correlation matrices R for $p = 3, \ldots, 9$.

From figure 4.2 it is clear that among the four methods the eigenvalue method is the best. Among the non linear shrinking techniques we should choose $f(x) = \tanh(x)$ since it on average gives the smallest d^2-distance, requires not too many iterations and preserves the correlation structure best in the sense that the transformed matrix looks more similar to the original matrix than for the other three presented shrinking functions. Bigger step size Δ gives on average fewer iterations but less accuracy in terms of the distance between the original and transformed matrix.

The non linear shrinking with step size $\Delta = 0.01$ and shrinking function $f(x) = \tanh(x)$ applied on pseudo-correlation matrices with iid $U(-1, 1)$ off diagonal elements needed on average

3.2, 10.4, 18.0, 23.6 and 28.0 iterations for $p = 3, \ldots, 7$ (this includes the cases where no iterations were needed).

It should be noted that the pseudo-correlation matrices obtained from bivariate linear correlation estimators applied on multivariate data are not of the form used for the comparison. The bias introduced by the non-PSD to PSD transformation is in general much less.

5 Multivariate Correlation Estimators

The so far presented techniques for robust estimation of linear correlation are bivariate techniques. For multivariate data, the pairwise correlations are patched together to form a robust estimate R. The problem of R not necessarily being PSD is solved by various techniques for transforming R to some PSD matrix $\tilde{R}$ close to R. However, there are techniques for estimating a linear correlation matrix which are truly multivariate. An overview of such techniques is given in this chapter.

5.1 The Multivariate Standard Estimator. Given a data set of n p-dimensional observations $\mathbf{y}_k = (y_{k1}, \ldots, y_{kp})^t$ for $k = 1, \ldots, n$, the sample product-moment covariance matrix is given by

$$C = \frac{1}{n-1} \sum_{k=1}^{n} (\mathbf{y}_k - \overline{\mathbf{y}})(\mathbf{y}_k - \overline{\mathbf{y}})^t,$$

where $\overline{\mathbf{y}}$ denotes the sample mean vector. This matrix is always PSD and when $n > p$ it is PD almost surely. The sample product-moment correlation matrix corresponding to C is then given by

$$R = DCD,$$

where D is a diagonal matrix with diagonal elements $d_{ii} = 1/\sqrt{c_{ii}}$ for $i = 1, \ldots, p$. The off-diagonal elements r_{ij} of R are

$$r_{ij} = \frac{\sum_{k=1}^{n}(y_{ki} - \overline{y_i})(y_{kj} - \overline{y_j})}{\sqrt{\sum_{k=1}^{n}(y_{ki} - \overline{y_i})^2}\sqrt{\sum_{k=1}^{n}(y_{kj} - \overline{y_j})^2}}.$$

which is recognized as the standard linear correlation estimator.

The (multivariate) standard estimator is extremely sensitive to outliers and should never be used if the data is not from a multivariate normal distribution.
This estimator is denoted R_{SE}.

5.2 Multivariate Trimming. We will now discuss robust methods for estimation of the linear correlation matrix on p-dimensional data based on multivariate trimming, described by Gnanadeskian and Kettenring (1972) [7]. Again we consider a data set of n p-dimensional observations $\mathbf{y}_k = (y_{k1}, \ldots, y_{kp})^t$ for $k = 1, \ldots, n$.

Method 1:

1. Rank the observations $\mathbf{y}_k$ in a decreasing order in terms of the squared Euclidean distance $(\mathbf{y}_k - \tilde{\mu})^t(\mathbf{y}_k - \tilde{\mu})$ from some robust estimate of location, $\tilde{\mu}$ (e.g. a γ-trimmed mean).
2. Set $l = \lceil \alpha n \rceil$ and
$$\tilde{C} = \frac{1}{n-l} \sum_{k=l+1}^{n} (\mathbf{y}_{(k)} - \tilde{\mu})(\mathbf{y}_{(k)} - \tilde{\mu})^t,$$
where $\mathbf{y}_{(k)}$ is the observation with the kth biggest Euclidean distance from $\tilde{\mu}$. If $n - l > p$ then $\tilde{C}$ is PD almost surely.
3. Rank the observations $\mathbf{y}_k$ in a decreasing order in terms of the squared Mahalanobis distance $(\mathbf{y}_k - \tilde{\mu})^t\tilde{C}^{-1}(\mathbf{y}_k - \tilde{\mu})$ from $\tilde{\mu}$ with respect to $\tilde{C}$. Let $\mathbf{y}_{(k)}$ denote the observation with the kth biggest Mahalanobis distance.
4. Set $m = \lceil \beta n \rceil$ and
$$\tilde{C} = \frac{1}{n-m} \sum_{k=m+1}^{n} (\mathbf{y}_{(k)} - \tilde{\mu})(\mathbf{y}_{(k)} - \tilde{\mu})^t.$$

If $n-m>p$ then $\tilde{C}$ is PD almost surely.

5. Set $\tilde{R}=D\tilde{C}D$, where D is a diagonal matrix with $d_{ii}=1/\sqrt{\tilde{c}_{ii}}$ for $i=1,\dots,p$. If $\tilde{R}$ is sufficiently stable, stop. Otherwise go to step 3.

There are many possible criteria for determining whether the estimate is sufficiently stable. One might consider the estimate stable if after k steps $d^2(\tilde{R}^{k-1},\tilde{R}^k)<0.001$ or if $\max_{i,j}|\tilde{r}_{ij}^{k-1}-\tilde{r}_{ij}^k|<0.001$. When $\alpha=\beta$ and $\tilde{\mu}$ is taken to be an α-trimmed mean this estimator is denoted $R_{\mathrm{MT1}}^{\alpha}$.

Method 2:

1. Take $\tilde{C}$ to be some estimate of the covariance matrix and $\tilde{\mu}$ to be some estimate of location.
2. Rank the observations $\mathbf{y}_k$ in a decreasing order in terms of the squared Mahalanobis distance $(\mathbf{y}_k-\tilde{\mu})^t\tilde{C}^{-1}(\mathbf{y}_k-\tilde{\mu})$ from $\tilde{\mu}$ with respect to $\tilde{C}$. Let $\mathbf{y}_{(k)}$ denote the observation with the kth biggest Mahalanobis distance.
3. Set $l=\lceil\alpha n\rceil$,

$$\tilde{\mu}=\frac{1}{n-l}\sum_{k=l+1}^{n}\mathbf{y}_{(k)}\text{ and }\tilde{C}=\frac{1}{n-l}\sum_{k=l+1}^{n}(\mathbf{y}_{(k)}-\tilde{\mu})(\mathbf{y}_{(k)}-\tilde{\mu})^t.$$

   If $n-l>p$ then $\tilde{C}$ is PD almost surely.

4. Set $\tilde{R}=D\tilde{C}D$, where D is a diagonal matrix with $d_{ii}=1/\sqrt{\tilde{c}_{ii}}$ for $i=1,\dots,p$. If $\tilde{R}$ is sufficiently stable, stop. Otherwise go to step 2.

The sample mean vector and the sample product-moment covariance matrix can be used as a starting point for the iterations. This estimator is denoted $R_{\mathrm{MT2}}^{\alpha}$.

However, for increased robustness, it is sometimes advisable to use robust alternatives as a starting point.

5.3 M-estimators. Instead of omitting observations with large Mahalanobis distances as in the multivariate trimming methods described above, the observations can be weighted using weights that depend on the respective distances. This is the basis of the multivariate M-estimators proposed by Maronna (1976) [15].

1. Take $\tilde{C}$ to be the some estimate of the covariance matrix and $\tilde{\mu}$ to be some estimate of location.
2. Set $\mathcal{M}_k^2=\mathcal{M}^2(\mathbf{y}_k,\tilde{\mu},\tilde{C})=(\mathbf{y}_k-\tilde{\mu})^t\tilde{C}^{-1}(\mathbf{y}_k-\tilde{\mu})$, for $k=1,\dots,n$.
3. Set

$$\tilde{\mu}=\left\{\sum_{k=1}^{n}w_1(\mathcal{M}_k)\mathbf{y}_i\right\}\Big/\left\{\sum_{k=1}^{n}w_1(\mathcal{M}_k)\right\}$$

   and

$$\tilde{C}=\frac{1}{n}\sum_{k=1}^{n}w_2(\mathcal{M}_k^2)(\mathbf{y}_k-\tilde{\mu})(\mathbf{y}_k-\tilde{\mu})^t.$$

4. Set $\tilde{R}=D\tilde{C}D$, where D is a diagonal matrix with $d_{ii}=1/\sqrt{\tilde{c}_{ii}}$ for $i=1,\dots,p$. If $\tilde{R}$ is sufficiently stable, stop. Otherwise go to step 2.

Two weight functions w_1 and w_2 resulting in maximum likelihood estimators for p-variate t_ν-distributions are

$$w_1(d)=\frac{p+\nu}{\nu+d^2}=w_2(d^2).$$

Using these weight functions, this estimator is denoted R_{MLT}^{ν} for maximum likelihood t.

Another choice of weight functions is based on giving observations in the "central part" of the data (with small Mahalanobis distances) full weight, while extreme observations (with large Mahalanobis distances) are down weighted. The weight functions are

$$w_1(d) = \begin{cases} 1, & d \leq a, \\ a/d, & d > a \end{cases}$$

and

$$w_2(d^2) = (w_1(d))^2 / b,$$

where b is a constant required to make the covariance estimator asymptotically unbiased. Since the squared Mahalanobis distance is asymptotically χ^2_p under normality a^2 is often taken to be the 90% quantile of χ^2_p.

5.4 The Minimum Volume Ellipsoid Estimator. The minimum volume ellipsoid (MVE) is an important robust estimator of the covariance (correlation) matrix in multivariate data analysis. The MVE covariance estimate is the covariance matrix that is defined by the ellipsoid with minimum volume covering h points of a data set of n points. Here h is taken to be $\lfloor (n+p+1)/2 \rfloor$. The ellipsoid is given by

$$(\mathbf{y} - \mathbf{c})^t \Gamma^{-1} (\mathbf{y} - \mathbf{c}) = p,$$

where $\mathbf{c}$ is the location estimate, Γ is a scatter matrix and p is the dimension of the data. $\mathbf{c}$ and Γ are defined by

$$\mathbf{c} = \sum_{k=1}^{h} v_k \mathbf{y}_{j_k}$$

and

$$\Gamma = \sum_{k=1}^{h} v_k (\mathbf{y}_{j_k} - \mathbf{c})(\mathbf{y}_{j_k} - \mathbf{c})^t,$$

where $j_1, \ldots, j_h$ are the covered points.

The first step of the algorithm is identifying the subset of points to be used in the estimate. There are several proposed algorithms for the subset selection. Once the subset is selected, the weights v_k are determined.

To find the actual MVE estimate is too computationally intense, so an approximation is used. The approximation algorithm is a combination of a subset selection and weight finding algorithm.

The MVE estimator is very robust, but has a bad accuracy. Several suggestions for retaining the robustness but improving the accuracy have been proposed.

One idea for improving the accuracy in the MVE correlation estimation under the assumption that the "good part" of the data is from a multivariate normal distribution was proposed in Rousseeuw and Zomeren (1990) [18] (this is the S-plus linear correlation estimator cov.mve).

1. Rank the observations in a decreasing order in terms of the squared Mahalobinis distance $\mathcal{M}^2(\mathbf{y}_k, \mathbf{c}, \Gamma) = (\mathbf{y}_k - \mathbf{c})^t \Gamma^{-1} (\mathbf{y}_k - \mathbf{c})$ from the MVE location estimate $\mathbf{c}$ with respect to Γ.
2. Find l such that $\mathcal{M}(\mathbf{y}_{(l+1)}, \mathbf{c}, \Gamma) < F^{-1}(0.975) < \mathcal{M}(\mathbf{y}_{(l)}, \mathbf{c}, \Gamma)$, where F is the distribution function of a χ^2_p-distributed random variable and $\mathbf{y}_{(l)}$ is the point with lth biggest Mahalobinis distance.
3. Set $\tilde{\mu} = \frac{1}{n-l} \sum_{k=l+1}^{n} \mathbf{y}_{(k)}$. That is, $\tilde{\mu}$ is the sample mean of the "good" part of the data.
4. Set $\tilde{C} = \frac{1}{n-l} \sum_{k=l+1}^{n} (\mathbf{y}_{(k)} - \tilde{\mu})(\mathbf{y}_{(k)} - \tilde{\mu})^t$. That is, $\tilde{C}$ is the standard covariance matrix estimate of the "good" part of the data.

5. Set $\tilde{R} = D\tilde{C}D$, where D is a diagonal matrix with $d_{ii} = 1/\sqrt{\tilde{c}_{ii}}$ for $i = 1, \ldots, p$. This estimator is denoted R_{MVE}.

6 Comparisons

In this chapter the performances of the estimators described in the previous chapter are compared using Monte Carlo simulations for different contaminated and uncontaminated elliptical distributions. The performance is measured in terms of the expected d^2-distance between the true linear correlation matrix R of the underlying uncontaminated distribution and the estimate $\tilde{R}$.

6.1 Simulation Conditions. For each type of underlying distribution the three different correlation matrices R_1, R_2 and R_3 are estimated, where

$$R_1 = \begin{pmatrix} 1 & 0.7 & 0.7 & 0.7 \\ 0.7 & 1 & 0.7 & 0.7 \\ 0.7 & 0.7 & 1 & 0.7 \\ 0.7 & 0.7 & 0.7 & 1 \end{pmatrix} \quad R_2 = \begin{pmatrix} 1 & 0.5 & 0.7 & 0.8 \\ 0.5 & 1 & 0.2 & 0.4 \\ 0.7 & 0.2 & 1 & 0.2 \\ 0.8 & 0.4 & 0.2 & 1 \end{pmatrix}$$

$$R_3 = \begin{pmatrix} 1 & 0.2 & -0.5 & -0.3 \\ 0.2 & 1 & 0.2 & 0.6 \\ -0.5 & 0.2 & 1 & 0.4 \\ -0.3 & 0.6 & 0.4 & 1 \end{pmatrix}.$$

An efficient correlation estimator ought not to be overly influenced by the contaminated portion of the data. Therefore it is reasonable to take R_1, R_2 and R_3 as the appropriate targets. The following estimators are compared:

R_τ: The bivariate method based on the Kendall's tau transform. The eigenvalue method is used for the non-PSD to PSD transform.

R_{SE}: The multivariate standard estimator.

R^α_{SSD}: The bivariate method based on variance estimation on symmetrically trimmed data, with equal variance and mean trimming factors α. The eigenvalue method is used for the non-PSD to PSD transform.

R_{MVE}: The S-plus minimum volume ellipsoid linear correlation estimator.

R^α_{MT1}: The multivariate trimming estimator (1) with equal trimming factors α. Two iterations are used.

R^α_{MT2}: The multivariate trimming estimator (2) with equal trimming factors α. The iterations stop when the d^2-distance between two successive correlation matrix estimates is less than 0.001.

R^ν_{MLT}: The maximum likelihood linear correlation estimator for a multivariate t_ν-distribution. The iterations stop when the d^2-distance between two successive correlation matrix estimates is less than 0.001.

The estimators are compared for the following underlying distributions:

a: The multivariate t_4-distribution with covariance matrix R and mean zero.

b: The multivariate normal distribution with covariance matrix R and mean zero.

c: The two point normal mixture, which is a mixture of 90% $\mathcal{N}(\mathbf{0}, R)$ and 10% $\mathcal{N}(\mathbf{0}, 9R)$.

d: The two point t_2, t_8 mixture, which is a mixture of 90% $\mathrm{t}_8(\mathbf{0}, R)$ and 10% $\mathrm{t}_2(\mathbf{0}, R)$.

e: The symmetric comonotonic-contaminated normal distribution, which is a mixture of 90% $\mathcal{N}(\mathbf{0}, R)$ and 10% $\mathcal{N}(\mathbf{0}, \mathbf{9})$.

f: The Clayton-normal contaminated normal distribution, which is a mixture of 90% $\mathcal{N}(\mathbf{0}, R)$ and 10% the distribution with a clayton copula, Kendall's tau rank correlation matrix R_1 and normal margins with mean zero and variance 4.

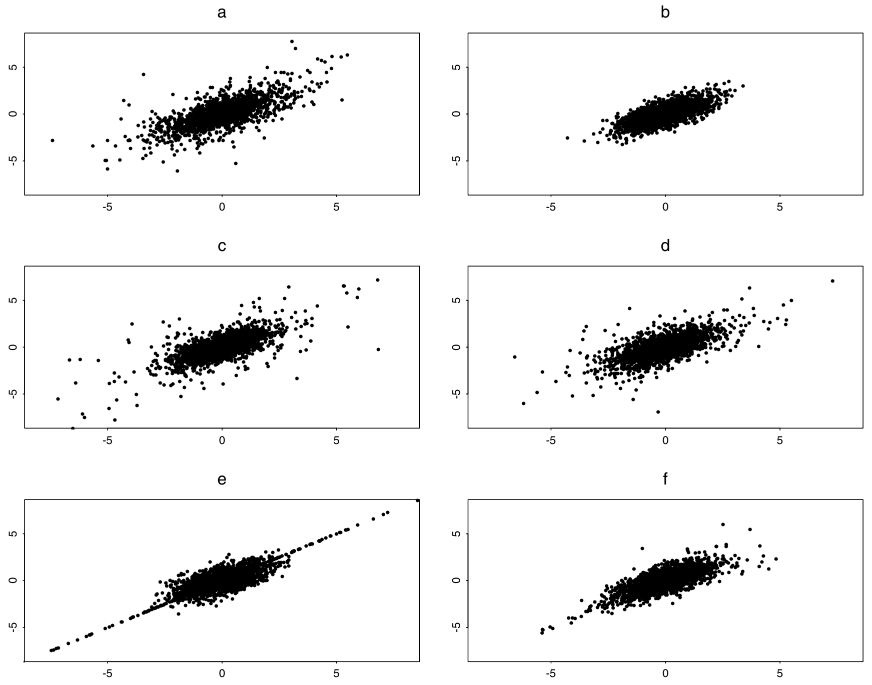

Figure 6.1 Samples of size 2000 from bivariate versions of the distributions a, b, c, d, e, f with R being a 2×2 matrix with off diagonal elements 0.7.

6.2 Simulation Results. The simulation results are summarized in table 6.1 showing the number of non-PSD estimates using the bivariate methods and in tables 6.2 and 6.3 showing estimates of $\mathbb{E}[d^2(R, \tilde{R})]$ for the tested linear correlation matrix estimators for sample size $n = 90$ and $n = 30$ respectively.

It should be noted that a, b, c and d are elliptical distributions, while e and f are not. Furthermore (e, R_1) and (f, R_1) are not that far from elliptically distributions as the orientation of the contamination is close to (for e) or the same (for f) as for the underlying uncontaminated elliptical distribution. (e, R_3) and (f, R_3) are quite far from elliptical distributed and are probably not realistic as distributions for observed data.

It should again be stressed that for clearly non elliptical distributions linear correlation is not a good measure of dependence. Therefore, if the data is far from elliptical and if it is not clear whether this is due to contamination of elliptically distributed data or simply that the data does not have an elliptical distribution, then a linear correlation matrix estimate might give very little information about dependence.

6.3 Conclusions. *Estimator* R_τ. Overall this estimator performs impressively for all elliptical distributions (a, b, c and d) combining the robustness properties of the Kendall's tau estimator with a high efficiency. It is also appealing because it is non-parametric and hence the problems of trimming or down weighting parts of the data properly are avoided. It should however be noted that if a contamination causes the distribution to be far from elliptical, which can be thought of in geometric terms, then R_τ is likely to have a bad performance. In such a case an outlier detecting scheme is called for, enabling enough critical outliers to be removed.

Estimator R_{SE}. This estimator is optimal if the data is from an uncontaminated multivariate normal distribution. However, mild contamination or heavier tailed data results in R_{SE} having a very poor performance. In fact one single wild observation can cause the estimator to be arbitrarily bad. That is, the breakdown point (the fraction of outliers that can be tolerated

	$n = 90$				$n = 30$			
	R_τ	$R^{0.025}_{\text{SSD}}$	$R^{0.05}_{\text{SSD}}$	$R^{0.075}_{\text{SSD}}$	R_τ	$R^{0.025}_{\text{SSD}}$	$R^{0.05}_{\text{SSD}}$	$R^{0.075}_{\text{SSD}}$
a, R_1	0	0	0	1	0	0	3	23
a, R_2	17	396	664	904	535	827	1081	1292
a, R_3	0	0	0	13	2	26	132	268
b, R_1	0	0	0	0	0	0	2	12
b, R_2	0	154	496	809	376	499	974	1151
b, R_3	0	0	0	11	1	7	77	249
c, R_1	0	0	0	0	0	2	4	22
c, R_2	6	415	625	825	531	792	1138	1250
c, R_3	0	1	0	13	0	32	109	224
d, R_1	0	0	0	0	0	1	5	10
d, R_2	9	301	608	879	548	745	1118	1313
d, R_3	0	0	2	13	2	16	116	260
e, R_1	0	0	0	0	0	0	7	37
e, R_2	0	609	1124	1365	278	738	1356	1544
e, R_3	0	1	31	130	2	28	313	544
f, R_1	0	0	0	0	0	0	2	6
f, R_2	0	31	163	357	112	240	659	910
f, R_3	0	0	2	15	1	5	134	308

Table 6.1 The number of non-PSD estimates out of 4000 estimates. Sample size $n = 90$ and $n = 30$.

	R_τ	R_{SE}	$R^{0.025}_{\text{SSD}}$	$R^{0.05}_{\text{SSD}}$	R_{MVE}	$R^{0.05}_{\text{MT1}}$	$R^{0.1}_{\text{MT1}}$	$R^{0.1}_{\text{MT2}}$	R^{4}_{MLT}
a, R_1	0.055	0.106	0.053	0.061	0.093	0.072	0.080	0.063	0.047
a, R_2	0.111	0.218	0.115	0.130	0.193	0.140	0.149	0.138	0.100
a, R_3	0.138	0.218	0.140	0.161	0.236	0.164	0.175	0.172	0.125
b, R_1	0.045	0.038	0.047	0.059	0.069	0.056	0.073	0.058	0.040
b, R_2	0.090	0.079	0.099	0.125	0.144	0.114	0.138	0.125	0.087
b, R_3	0.114	0.099	0.124	0.155	0.183	0.135	0.165	0.155	0.110
c, R_1	0.053	0.094	0.049	0.059	0.068	0.072	0.056	0.051	0.042
c, R_2	0.109	0.205	0.105	0.126	0.149	0.130	0.113	0.107	0.093
c, R_3	0.131	0.249	0.129	0.153	0.180	0.151	0.134	0.137	0.119
d, R_1	0.054	0.197	0.052	0.062	0.084	0.062	0.073	0.059	0.045
d, R_2	0.107	0.372	0.109	0.128	0.180	0.123	0.137	0.128	0.098
d, R_3	0.130	0.439	0.134	0.158	0.217	0.144	0.162	0.152	0.120
e, R_1	0.092	0.266	0.114	0.112	0.129	0.087	0.077	0.123	0.185
e, R_2	0.284	0.979	0.336	0.320	0.169	0.153	0.122	0.152	0.333
e, R_3	0.600	2.752	0.597	0.519	0.245	0.311	0.190	0.266	0.777
f, R_1	0.049	0.059	0.049	0.057	0.073	0.049	0.059	0.058	0.048
f, R_2	0.168	0.285	0.164	0.173	0.144	0.117	0.115	0.120	0.130
f, R_3	0.356	0.885	0.295	0.287	0.210	0.184	0.173	0.192	0.316

Table 6.2 $\mathbb{E}[d^2(R, \tilde{R})]$ for different estimators, distributions and $R = R_1, R_2, R_3$. The results are obtained from 4000 independent samples of size 90.

	R_τ	R_{SE}	$R^{0.025}_{\mathrm{SSD}}$	$R^{0.05}_{\mathrm{SSD}}$	R_{MVE}	$R^{0.05}_{\mathrm{MT1}}$	$R^{0.1}_{\mathrm{MT1}}$	$R^{0.1}_{\mathrm{MT2}}$	R^{4}_{MLT}
a, R_1	0.195	0.270	0.181	0.216	0.382	0.276	0.300	0.201	0.153
a, R_2	0.357	0.495	0.345	0.396	0.679	0.417	0.431	0.408	0.308
a, R_3	0.439	0.641	0.432	0.487	0.820	0.474	0.496	0.512	0.384
b, R_1	0.156	0.120	0.147	0.198	0.329	0.207	0.262	0.183	0.134
b, R_2	0.303	0.247	0.296	0.379	0.630	0.351	0.409	0.363	0.280
b, R_3	0.380	0.315	0.374	0.475	0.768	0.405	0.459	0.448	0.350
c, R_1	0.188	0.269	0.177	0.205	0.330	0.274	0.262	0.179	0.144
c, R_2	0.350	0.529	0.341	0.382	0.608	0.393	0.383	0.352	0.301
c, R_3	0.420	0.649	0.406	0.453	0.728	0.445	0.425	0.443	0.373
d, R_1	0.183	0.309	0.167	0.209	0.358	0.243	0.286	0.181	0.145
d, R_2	0.341	0.533	0.328	0.391	0.674	0.373	0.411	0.377	0.299
d, R_3	0.426	0.659	0.411	0.485	0.820	0.436	0.459	0.463	0.365
e, R_1	0.169	0.279	0.188	0.190	0.334	0.177	0.198	0.219	0.229
e, R_2	0.458	0.974	0.559	0.499	0.649	0.378	0.368	0.507	0.545
e, R_3	0.875	2.665	1.169	0.833	0.928	0.682	0.534	1.049	1.201
f, R_1	0.141	0.132	0.132	0.169	0.305	0.165	0.204	0.163	0.122
f, R_2	0.350	0.431	0.348	0.391	0.602	0.316	0.340	0.367	0.314
f, R_3	0.629	1.122	0.629	0.611	0.811	0.477	0.470	0.593	0.627

Table 6.3 $\mathbb{E}[d^2(R,\tilde{R})]$ for different estimators, distributions and $R = R_1, R_2, R_3$. The results are obtained from 4000 independent samples of size 30.

without the estimator potentially breaking down) is zero. R_{SE} should never be used unless the data is from an uncontaminated multivariate normal distribution.

Estimator $R^{\alpha}_{\mathrm{SSD}}$. Like R_τ this estimator performs impressively for all elliptical distributions, provided that the trimming parameters for the mean and more so for the variance are well chosen. However, because of the trimming the pseudo-correlation matrix consisting of the bivariate estimates is much more likely to be non-PSD than for R_τ. Since a transform from non-PSD to PSD tends to introduce bias, this disadvantage will probably effect the performance in higher dimensions. The breakdown point of $R^{\alpha}_{\mathrm{SSD}}$ is equal to trimming percentage α.

Estimator R_{MVE}. The minimum volume ellipsoid estimator has received much attention because it is extremely robust, the breakdown point is approximately 1/2. The efficiency is however bad. R_{MVE} uses the MVE covariance estimate but is designed to give a better efficiency under the assumption that the "good" part of the data is from a multivariate normal distribution. Overall R_{MVE} has the worst performance of the studied estimators, except for R_{SE}. For data from comonotonic contaminated multivariate normal data, for which it is designed to have a good efficiency, it performs quite well. However, it is still outperformed by $R^{\alpha}_{\mathrm{MT2}}$. Furthermore, for small sample sizes of about six times or less the dimension, R_{MVE} has the worst performance of the studied estimators except for comonotonic contaminated multivariate normal data.

Estimator $R^{\alpha}_{\mathrm{MT1}}$. Overall this estimator performs well, provided that the trimming parameters are chosen reasonably well. In terms of efficiency it can not match R_τ and $R^{\alpha}_{\mathrm{SSD}}$ for the elliptical distributions. However, it performs impressively for comonotonic contaminated multivariate normal data and more generally it retains a fairly high efficiency for contamination causing the data to be quite far from elliptically distributed. It should be noted that in our comparison only one repetition for stabilizing the estimate were used. More repetitions could in some cases improve the efficiency. This is certainly true for bigger trimming parameters. $R^{\alpha}_{\mathrm{MT1}}$ can handle a fraction α of symmetrically distributed contamination (outliers).

Estimator $R^{\alpha}_{\mathrm{MT2}}$. As expected, this estimator has a performance similar to $R^{\alpha}_{\mathrm{MT1}}$. For the tested elliptical distributions it has a better performance, but a worse performance for the non elliptical distributions e and f. Compared to all tested estimators the overall performance is not impressive but also not bad. Like $R^{\alpha}_{\mathrm{MT1}}$ it can handle a fraction α of symmetrically distributed contamination (outliers).

Estimator R^{ν}_{MLT}. Among the presented estimators, R^{ν}_{MLT} has the best performance for elliptical distributions for a reasonable choice of ν (in our case $\nu = 4$). This is seen from the simulation results, with nearly constant behavior across the elliptical distributions tested (a, b, c and d). One weakness is that in the presence of contamination causing the distribution of the data to be non elliptical R^{ν}_{MLT} performs badly. As contamination causes the distribution to be further away from ellipticality this is more pronounced as seen from the simulation results for e (f). One serious disadvantage of R^{ν}_{MLT} is its low breakdown point, approximately $1/p$. This means that in higher dimensions R^{ν}_{MLT} is not very robust. Therefore, $R^{\alpha}_{\mathrm{MT1}}$ and R_{τ} seems to be more natural choices for high dimensional data.

References

[1] Campbell N.A. (1980) Robust procedures in multivariate analysis I: Robust covariance estimation, *Applied Statistics*, **29**, 231-237.

[2] Devlin S.J., Gnanadeskian R. and Kettenring J.R. (1975) Robust estimation and outlier detection with correlation coefficients, *Biometrika* **62**(3), 531-545.

[3] Devlin S.J., Gnanadeskian R. and Kettenring J.R. (1981) Robust estimation of dispersion matrices and principal components, *Journal of American statistical association*, **76**, 354-362.

[4] Embrechts P., McNeil A.J. and Straumann D. (1999) Correlation: Pitfalls and Alternatives, *RISK* **12**(5), 69-71.

[5] Embrechts P., McNeil A.J. and Straumann D. (1999) Correlation and Dependence in Risk Management: Properties and Pitfalls, *Preprint ETH Zürich*, available from http://www.math.ethz.ch/ ∼embrechts.

[6] Fang K.-T., Kotz S. and Ng K.-W. (1987) *Symmetric Multivariate and Related Distributions*, Chapman & Hall, London.

[7] Gnanadeskian R. and Kettenring J.R. (1972) Robust estimates, residuals, and outlier detection with multiresponse data, *Biometrika* **28**, 81-124.

[8] He X. and Wang G. (1992) On properties and applicability of the minimum volume estimators. Technical Report, University of Illinois.

[9] He X. and Wang G. (1996) Cross-checking using the minimum volume ellipsoid estimator, *Statistica Sinica*, **6**, 367-374.

[10] Huber P.J. (1981) *Robust Statistics*, John Wiley & Sons, New York.

[11] Joe H. (1997) *Multivariate Models and Dependence Concepts*, Chapman & Hall, London.

[12] Johnson M.E. (1987) *Multivariate Statistical Simulations*, John Wiley & Sons, New York.

[13] Johnson N.L. and Kotz S. (1972) *Distributions in Statistics: Continuous Multivariate Distributions*, John Wiley & Sons, New York.

[14] Kendall M.G. and Gibbons J.D. (1990) *Rank Correlation Methods*, 5th edition, Griffin, London.

[15] Maronna R.A. (1976) Robust M-estimators of multivariate location and scatter, *The annals of statistics* **4**, 51-67.

[16] Ripley B.D. (1987) *Stochastic Simulation*, John Wiley & Sons, New York.

[17] Rousseeuw P.J. and Molenberghs G. (1993) Transformation of Non Positive Semidefinite Correlation Matrices, *Communications in Statistics - Theory and Methods* **22**(4), 965-984.

[18] Rousseeuw P.J. and van Zomeren B.C. (1990) Unmasking multivariate outliers and leverage points, *Journal of American statistics association* **85**, 633-639.

[19] Seber G.A.F. (1984) *Multivariate Observations*, John Wiley & Sons, New York.