6
Multivariate Models

Financial risk models, whether for market or credit risks, are inherently multivariate. The value change of a portfolio of traded instruments over a fixed time horizon depends on a random vector of risk-factor changes or returns. The loss incurred by a credit portfolio depends on a random vector of losses for the individual counterparties in the portfolio.

This chapter is the first of two successive ones that focus on models for random vectors. The emphasis in this chapter is on tractable models that describe both the individual behaviour of the components of a random vector and their joint behaviour or dependence structure. We consider a number of distributions that extend the multivariate normal but provide more realistic models for many kinds of financial data.

In Chapter 7 we focus explicitly on modelling the dependence structure of a random vector and largely ignore marginal behaviour. We introduce copula models of dependence and study a number of dependence measures and concepts related to copulas. Both Chapter 6 and Chapter 7 take a static, distributional view of multivariate modelling; for multivariate time-series models, see Chapter 14.

Section 6.1 reviews basic ideas in multivariate statistics and discusses the multivariate normal (or Gaussian) distribution and its deficiencies as a model for empirical return data.

In Section 6.2 we consider a generalization of the multivariate normal distribution known as a multivariate normal mixture distribution, which shares much of the structure of the multivariate normal and retains many of its properties. We treat both variance mixtures, which belong to the wider class of elliptical distributions, and mean–variance mixtures, which allow asymmetry. Concrete examples include t distributions and generalized hyperbolic distributions, and we show in empirical examples that these models provide a better fit than a Gaussian distribution to asset return data. In some cases, multivariate return data are not strongly asymmetric and models from the class of elliptical distributions are good enough; in Section 6.3 we investigate the elegant properties of these distributions.

In Section 6.4 we discuss the important issue of dimension-reduction techniques for reducing large sets of risk factors to smaller subsets of essential risk drivers. The key idea here is that of a factor model, and we also review the principal components method of constructing factors.

6.1 Basics of Multivariate Modelling

This first section reviews important basic material from multivariate statistics, which will be known to many readers. The main topic of the section is the multivariate normal distribution and its properties; this distribution is central to much of classical multivariate analysis and was the starting point for attempts to model market risk (the variance–covariance method of Section 9.2.2).

6.1.1 Random Vectors and Their Distributions

Joint and marginal distributions. Consider a general d-dimensional random vector of risk-factor changes (or so-called returns) $\boldsymbol{X} = (X_1, \ldots, X_d)'$. The distribution of $\boldsymbol{X}$ is completely described by the joint distribution function (df)

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = F_{\boldsymbol{X}}(x_1, \ldots, x_d) = P(\boldsymbol{X} \leqslant \boldsymbol{x}) = P(X_1 \leqslant x_1, \ldots, X_d \leqslant x_d).$$

Where no ambiguity arises we simply write F, omitting the subscript.

The *marginal* df of X_i, written F_{X_i} or often simply F_i, is the df of that risk factor considered individually and is easily calculated from the joint df. For all i we have

$$F_i(x_i) = P(X_i \leqslant x_i) = F(\infty, \ldots, \infty, x_i, \infty, \ldots, \infty). \tag{6.1}$$

If the marginal df $F_i(x)$ is absolutely continuous, then we refer to its derivative $f_i(x)$ as the marginal density of X_i. It is also possible to define k-dimensional marginal distributions of $\boldsymbol{X}$ for $2 \leqslant k \leqslant d-1$. Suppose we partition $\boldsymbol{X}$ into $(\boldsymbol{X}_1', \boldsymbol{X}_2')'$, where $\boldsymbol{X}_1 = (X_1, \ldots, X_k)'$ and $\boldsymbol{X}_2 = (X_{k+1}, \ldots, X_d)'$, then the marginal df of $\boldsymbol{X}_1$ is

$$F_{\boldsymbol{X}_1}(\boldsymbol{x}_1) = P(\boldsymbol{X}_1 \leqslant \boldsymbol{x}_1) = F(x_1, \ldots, x_k, \infty, \ldots, \infty).$$

For bivariate and other low-dimensional margins it is convenient to have a simpler alternative notation in which, for example, $F_{ij}(x_i, x_j)$ stands for the marginal distribution of the components X_i and X_j.

The df of a random vector $\boldsymbol{X}$ is said to be absolutely continuous if

$$F(x_1, \ldots, x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f(u_1, \ldots, u_d)\, \mathrm{d}u_1 \cdots \mathrm{d}u_d$$

for some non-negative function f, which is then known as the *joint density* of $\boldsymbol{X}$. Note that the existence of a joint density implies the existence of marginal densities for all k-dimensional marginals. However, the existence of a joint density is not necessarily implied by the existence of marginal densities (counterexamples can be found in Chapter 7 on copulas).

In some situations it is convenient to work with the *survival function* of $\boldsymbol{X}$, defined by

$$\bar{F}_{\boldsymbol{X}}(\boldsymbol{x}) = \bar{F}_{\boldsymbol{X}}(x_1, \ldots, x_d) = P(\boldsymbol{X} > \boldsymbol{x}) = P(X_1 > x_1, \ldots, X_d > x_d)$$

and written simply as $\bar{F}$ when no ambiguity arises. The marginal survival function of X_i, written $\bar{F}_{X_i}$ or often simply $\bar{F}_i$, is given by

$$\bar{F}_i(x_i) = P(X_i > x_i) = \bar{F}(-\infty, \ldots, -\infty, x_i, -\infty, \ldots, -\infty).$$

Conditional distributions and independence. If we have a multivariate model for risks in the form of a joint df, survival function or density, then we have implicitly described the *dependence structure* of the risks. We can make conditional probability statements about the probability that certain components take certain values given that other components take other values. For example, consider again our partition of $\boldsymbol{X}$ into $(\boldsymbol{X}_1', \boldsymbol{X}_2')'$ and assume absolute continuity of the df of $\boldsymbol{X}$. Let $f_{\boldsymbol{X}_1}$ denote the joint density of the k-dimensional marginal distribution $F_{\boldsymbol{X}_1}$. Then the conditional distribution of $\boldsymbol{X}_2$ given $\boldsymbol{X}_1 = \boldsymbol{x}_1$ has density

$$f_{\boldsymbol{X}_2|\boldsymbol{X}_1}(\boldsymbol{x}_2 \mid \boldsymbol{x}_1) = \frac{f(\boldsymbol{x}_1, \boldsymbol{x}_2)}{f_{\boldsymbol{X}_1}(\boldsymbol{x}_1)}, \tag{6.2}$$

and the corresponding df is

$$F_{\boldsymbol{X}_2|\boldsymbol{X}_1}(\boldsymbol{x}_2 \mid \boldsymbol{x}_1) = \int_{u_{k+1}=-\infty}^{x_{k+1}} \cdots \int_{u_d=-\infty}^{x_d} \frac{f(x_1, \ldots, x_k, u_{k+1}, \ldots, u_d)}{f_{\boldsymbol{X}_1}(\boldsymbol{x}_1)} \, \mathrm{d}u_{k+1} \cdots \mathrm{d}u_d.$$

If the joint density of $\boldsymbol{X}$ factorizes into $f(\boldsymbol{x}) = f_{\boldsymbol{X}_1}(\boldsymbol{x}_1) f_{\boldsymbol{X}_2}(\boldsymbol{x}_2)$, then the conditional distribution and density of $\boldsymbol{X}_2$ given $\boldsymbol{X}_1 = \boldsymbol{x}_1$ are identical to the marginal distribution and density of $\boldsymbol{X}_2$: in other words, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent. We recall that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent if and only if

$$F(\boldsymbol{x}) = F_{\boldsymbol{X}_1}(\boldsymbol{x}_1) F_{\boldsymbol{X}_2}(\boldsymbol{x}_2), \quad \forall \boldsymbol{x},$$

or, in the case where $\boldsymbol{X}$ possesses a joint density, $f(\boldsymbol{x}) = f_{\boldsymbol{X}_1}(\boldsymbol{x}_1) f_{\boldsymbol{X}_2}(\boldsymbol{x}_2)$.

The components of $\boldsymbol{X}$ are *mutually* independent if and only if $F(\boldsymbol{x}) = \prod_{i=1}^d F_i(x_i)$ for all $\boldsymbol{x} \in \mathbb{R}^d$ or, in the case where $\boldsymbol{X}$ possesses a density, $f(\boldsymbol{x}) = \prod_{i=1}^d f_i(x_i)$.

Moments and characteristic function. The *mean vector* of $\boldsymbol{X}$, when it exists, is given by

$$E(\boldsymbol{X}) := (E(X_1), \ldots, E(X_d))'.$$

The *covariance matrix*, when it exists, is the matrix $\mathrm{cov}(\boldsymbol{X})$ defined by

$$\mathrm{cov}(\boldsymbol{X}) := E((\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))'),$$

where the expectation operator acts componentwise on matrices. If we write Σ for $\mathrm{cov}(\boldsymbol{X})$, then the (i, j)th element of this matrix is

$$\sigma_{ij} = \mathrm{cov}(X_i, X_j) = E(X_i X_j) - E(X_i) E(X_j),$$

the ordinary pairwise covariance between X_i and X_j. The diagonal elements $\sigma_{11}, \ldots, \sigma_{dd}$ are the variances of the components of $\boldsymbol{X}$.

The *correlation matrix* of $\boldsymbol{X}$, denoted by $\rho(\boldsymbol{X})$, can be defined by introducing a standardized vector $\boldsymbol{Y}$ such that $Y_i = X_i / \sqrt{\mathrm{var}(X_i)}$ for all i and taking $\rho(\boldsymbol{X}) := \mathrm{cov}(\boldsymbol{Y})$. If we write P for $\rho(\boldsymbol{X})$, then the (i, j)th element of this matrix is

$$\rho_{ij} = \rho(X_i, X_j) = \frac{\mathrm{cov}(X_i, X_j)}{\sqrt{\mathrm{var}(X_i)\,\mathrm{var}(X_j)}}, \tag{6.3}$$

the ordinary pairwise linear correlation of X_i and X_j. To express the relationship between correlation and covariance matrices in matrix form, it is useful to introduce operators on a covariance matrix Σ as follows:

$$\Delta(\Sigma) := \operatorname{diag}(\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{dd}}), \tag{6.4}$$

$$\wp(\Sigma) := (\Delta(\Sigma))^{-1}\Sigma(\Delta(\Sigma))^{-1}. \tag{6.5}$$

Thus $\Delta(\Sigma)$ extracts from Σ a diagonal matrix of standard deviations, and $\wp(\Sigma)$ extracts a correlation matrix. The covariance and correlation matrices Σ and P of $\boldsymbol{X}$ are related by

$$P = \wp(\Sigma). \tag{6.6}$$

Mean vectors and covariance matrices are manipulated extremely easily under linear operations on the vector $\boldsymbol{X}$. For any matrix $B \in \mathbb{R}^{k\times d}$ and vector $\boldsymbol{b} \in \mathbb{R}^k$ we have

$$E(B\boldsymbol{X} + \boldsymbol{b}) = BE(\boldsymbol{X}) + \boldsymbol{b}, \tag{6.7}$$

$$\operatorname{cov}(B\boldsymbol{X} + \boldsymbol{b}) = B\operatorname{cov}(\boldsymbol{X})B'. \tag{6.8}$$

Covariance matrices (and hence correlation matrices) are therefore *positive semi-definite*; writing Σ for $\operatorname{cov}(\boldsymbol{X})$ we see that (6.8) implies that $\operatorname{var}(\boldsymbol{a}'\boldsymbol{X}) = \boldsymbol{a}'\Sigma\boldsymbol{a} \geqslant 0$ for any $\boldsymbol{a} \in \mathbb{R}^d$. If we have that $\boldsymbol{a}'\Sigma\boldsymbol{a} > 0$ for any $\boldsymbol{a} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$, we say that Σ is *positive definite*; in this case the matrix is invertible. We will make use of the well-known *Cholesky factorization* of positive-definite covariance matrices at many points; it is well known that such a matrix can be written as $\Sigma = AA'$ for a lower triangular matrix A with positive diagonal elements. The matrix A is known as the Cholesky factor. It will be convenient to denote this factor by $\Sigma^{1/2}$ and its inverse by $\Sigma^{-1/2}$. Note that there are other ways of defining the "square root" of a symmetric positive-definite matrix (such as the symmetric decomposition), but we will always use $\Sigma^{1/2}$ to denote the Cholesky factor.

In this chapter many properties of the multivariate distribution of a vector $\boldsymbol{X}$ are demonstrated using the characteristic function, which is given by

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}}), \quad \boldsymbol{t} \in \mathbb{R}^d.$$

6.1.2 Standard Estimators of Covariance and Correlation

Suppose we have n observations of a d-dimensional risk-factor return vector denoted by $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Typically, these would be daily, weekly, monthly or yearly observations forming a multivariate time series. We will assume throughout this chapter that the observations are *identically distributed* in the window of observation and either independent or at least serially *uncorrelated* (also known as multivariate white noise). As discussed in Chapter 3, the assumption of independence may be roughly tenable for longer time intervals such as months or years. For shorter time intervals independence may be a less appropriate assumption (due to a phenomenon known as volatility clustering, discussed in Section 3.1.1), but serial correlation of returns is often quite weak.

We assume that the observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ come from a distribution with mean vector $\boldsymbol{\mu}$, finite covariance matrix Σ and correlation matrix P. We now briefly review the standard estimators of these vector and matrix parameters.

Standard method-of-moments estimators of $\boldsymbol{\mu}$ and Σ are given by the *sample mean vector* $\bar{\boldsymbol{X}}$ and the *sample covariance matrix* S. These are defined by

$$\bar{\boldsymbol{X}} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i, \qquad S := \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})', \tag{6.9}$$

where arithmetic operations on vectors and matrices are performed componentwise. $\bar{\boldsymbol{X}}$ is an unbiased estimator but S is biased; an unbiased version may be obtained by taking $S_u := nS/(n-1)$, as may be seen by calculating

$$\begin{aligned} nE(S) &= E\bigg(\sum_{i=1}^{n} (\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})' - n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\bigg) \\ &= \sum_{i=1}^{n} \operatorname{cov}(\boldsymbol{X}_i) - n\operatorname{cov}(\bar{\boldsymbol{X}}) = n\Sigma - \Sigma, \end{aligned}$$

since $\operatorname{cov}(\bar{\boldsymbol{X}}) = n^{-1}\Sigma$ when the data vectors are iid, or identically distributed and uncorrelated.

The *sample correlation matrix* R may be easily calculated from the sample covariance matrix; its (j,k)th element is given by $r_{jk} = s_{jk}/\sqrt{s_{jj}s_{kk}}$, where s_{jk} denotes the (j,k)th element of S. Or, using the notation introduced in (6.5), we have

$$R = \wp(S),$$

which is the analogous equation to (6.6) for estimators.

Further properties of the estimators $\bar{\boldsymbol{X}}$, S and R will very much depend on the *true multivariate distribution* of the observations. These quantities are not necessarily the best estimators of the corresponding theoretical quantities in all situations. This point is often forgotten in financial risk management, where sample covariance and correlation matrices are routinely calculated and interpreted with little critical consideration of underlying models.

If our data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are iid multivariate normal, then $\bar{\boldsymbol{X}}$ and S are the *maximum likelihood estimators* (MLEs) of the mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Their behaviour as estimators is well understood, and statistical inference for the model parameters is described in all standard texts on multivariate analysis.

However, the multivariate normal is certainly not a good description of financial risk-factor returns over short time intervals, such as daily data, and is often not good over longer time intervals either. Under these circumstances the behaviour of the standard estimators in (6.9) is often less well understood, and other estimators of the true mean vector $\boldsymbol{\mu}$ and covariance matrix Σ may perform better in terms of *efficiency* and *robustness*. Roughly speaking, by a more efficient estimator we mean an estimator with a smaller expected estimation error; by a more robust estimator we mean an estimator whose performance is not so susceptible to the presence of outlying data values.

6.1.3 The Multivariate Normal Distribution

Definition 6.1. $\boldsymbol{X} = (X_1, \ldots, X_d)'$ has a multivariate normal or *Gaussian* distribution if

$$\boldsymbol{X} \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + A\boldsymbol{Z},$$

where $\boldsymbol{Z} = (Z_1, \ldots, Z_k)'$ is a vector of iid univariate *standard* normal rvs (mean 0 and variance 1), and $A \in \mathbb{R}^{d\times k}$ and $\boldsymbol{\mu} \in \mathbb{R}^d$ are a matrix and a vector of constants, respectively.

It is easy to verify, using (6.7) and (6.8), that the mean vector of this distribution is $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and the covariance matrix is $\mathrm{cov}(\boldsymbol{X}) = \Sigma$, where $\Sigma = AA'$ is a positive-semidefinite matrix. Moreover, using the fact that the characteristic function of a standard univariate normal variate Z is $\phi_Z(t) = \mathrm{e}^{-t^2/2}$, the characteristic function of $\boldsymbol{X}$ may be calculated to be

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}}) = \exp(\mathrm{i}\boldsymbol{t}'\boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{t}'\Sigma\boldsymbol{t}), \quad \boldsymbol{t} \in \mathbb{R}^d. \tag{6.10}$$

Clearly, the distribution is characterized by its mean vector and covariance matrix, and hence a standard notation is $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$. Note that the components of $\boldsymbol{X}$ are mutually independent if and only if Σ is diagonal. For example, $\boldsymbol{X} \sim N_d(\boldsymbol{0}, I_d)$ if and only if $X_1, \ldots, X_d$ are iid $N(0, 1)$, the standard univariate normal distribution.

We concentrate on the *non-singular case* of the multivariate normal when $\mathrm{rank}(A) = d \leqslant k$. In this case the covariance matrix Σ has full rank d and is therefore invertible (non-singular) and positive definite. Moreover, $\boldsymbol{X}$ has an absolutely continuous distribution function with joint density given by

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\{-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\}, \quad \boldsymbol{x} \in \mathbb{R}^d, \tag{6.11}$$

where $|\Sigma|$ denotes the determinant of Σ.

The form of the density clearly shows that points with equal density lie on *ellipsoids* determined by equations of the form $(\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = c$, for constants $c > 0$. In two dimensions the contours of equal density are ellipses, as illustrated in Figure 6.1. Whenever a multivariate density $f(\boldsymbol{x})$ depends on $\boldsymbol{x}$ only through the quadratic form $(\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$, it is the density of a so-called elliptical distribution, as discussed in more detail in Section 6.3.

Definition 6.1 is essentially a simulation recipe for the multivariate normal distribution. To be explicit, if we wished to generate a vector $\boldsymbol{X}$ with distribution $N_d(\boldsymbol{\mu}, \Sigma)$, where Σ is positive definite, we would use the following algorithm.

Algorithm 6.2 (simulation of multivariate normal distribution).

(1) Perform a Cholesky decomposition of Σ (see, for example, Press et al. 1992) to obtain the Cholesky factor $\Sigma^{1/2}$.

(2) Generate a vector $\boldsymbol{Z} = (Z_1, \ldots, Z_d)'$ of independent standard normal variates.

(3) Set $\boldsymbol{X} = \boldsymbol{\mu} + \Sigma^{1/2}\boldsymbol{Z}$.

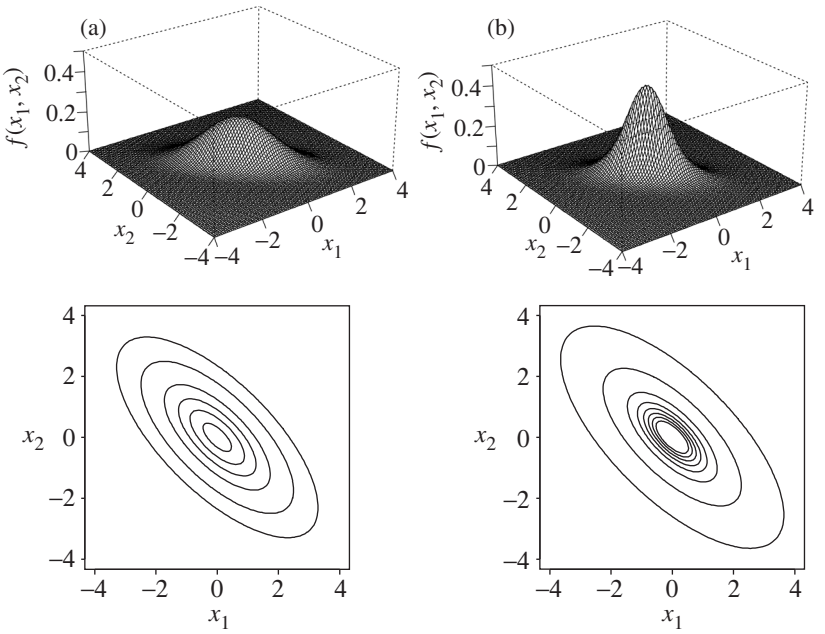

Figure 6.1. (a) Perspective and contour plots for the density of a bivariate normal distribution with standard normal margins and correlation −70%. (b) Corresponding plots for a bivariate t density with four degrees of freedom (see Example 6.7 for details) and the *same mean vector and covariance matrix* as the normal distribution. Contour lines are plotted at the same heights for both densities.

We now summarize further useful properties of the multivariate normal. These properties underline the attractiveness of the multivariate normal for computational work in risk management. Note, however, that many of them are in fact shared by the broader classes of normal mixture distributions and elliptical distributions (see Section 6.3.3 for properties of the latter).

Linear combinations. If we take linear combinations of multivariate normal random vectors, then these remain multivariate normal. Let $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ and take any $B \in \mathbb{R}^{k \times d}$ and $\boldsymbol{b} \in \mathbb{R}^k$. Then it is easily shown (e.g. using the characteristic function (6.10)) that

$$B\boldsymbol{X} + \boldsymbol{b} \sim N_k(B\boldsymbol{\mu} + \boldsymbol{b}, B\Sigma B'). \tag{6.12}$$

As a special case, if $\boldsymbol{a} \in \mathbb{R}^d$, then

$$\boldsymbol{a}'\boldsymbol{X} \sim N(\boldsymbol{a}'\boldsymbol{\mu}, \boldsymbol{a}'\Sigma\boldsymbol{a}), \tag{6.13}$$

and this fact is used routinely in the variance–covariance approach to risk management, as discussed in Section 9.2.2.

In this context it is interesting to note the following elegant characterization of multivariate normality. It is easily shown using characteristic functions that $\boldsymbol{X}$ is multivariate normal *if and only if* $\boldsymbol{a}'\boldsymbol{X}$ is univariate normal for all vectors $\boldsymbol{a} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$.

Marginal distributions. It is clear from (6.13) that the univariate marginal distributions of $\boldsymbol{X}$ must be univariate normal. More generally, using the $\boldsymbol{X} = (\boldsymbol{X}_1', \boldsymbol{X}_2')'$ notation from Section 6.1.1 and extending this notation naturally to $\boldsymbol{\mu}$ and Σ,

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

property (6.12) implies that the marginal distributions of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are also multivariate normal and are given by $\boldsymbol{X}_1 \sim N_k(\boldsymbol{\mu}_1, \Sigma_{11})$ and $\boldsymbol{X}_2 \sim N_{d-k}(\boldsymbol{\mu}_2, \Sigma_{22})$.

Conditional distributions. Assuming that Σ is positive definite, the conditional distributions of $\boldsymbol{X}_2$ given $\boldsymbol{X}_1$ and of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2$ may also be shown to be multivariate normal. For example, $\boldsymbol{X}_2 \mid \boldsymbol{X}_1 = \boldsymbol{x}_1 \sim N_{d-k}(\boldsymbol{\mu}_{2.1}, \Sigma_{22.1})$, where

$$\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1) \quad \text{and} \quad \Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

are the conditional mean vector and covariance matrix.

Quadratic forms. If $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ with Σ positive definite, then

$$(\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi_d^2, \tag{6.14}$$

a chi-squared distribution with d degrees of freedom. This is seen by observing that $\boldsymbol{Z} = \Sigma^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}) \sim N_d(\boldsymbol{0}, I_d)$ and $(\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{Z}'\boldsymbol{Z} \sim \chi_d^2$. This property (6.14) is useful for checking multivariate normality (see Section 6.1.4).

Convolutions. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent d-dimensional random vectors satisfying $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ and $\boldsymbol{Y} \sim N_d(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma})$, then we may take the product of characteristic functions to show that $\boldsymbol{X} + \boldsymbol{Y} \sim N_d(\boldsymbol{\mu} + \tilde{\boldsymbol{\mu}}, \Sigma + \tilde{\Sigma})$.

6.1.4 Testing Multivariate Normality

We now consider the issue of testing whether the data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are observations from a multivariate normal distribution.

Univariate tests. If $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are iid multivariate normal, then for $1 \leqslant j \leqslant d$ the univariate sample $X_{1,j}, \ldots, X_{n,j}$ consisting of the observations of the jth component must be iid univariate normal; in fact, any univariate sample constructed from a linear combination of the data of the form $\boldsymbol{a}'\boldsymbol{X}_1, \ldots, \boldsymbol{a}'\boldsymbol{X}_n$ must be iid univariate normal. This can be assessed graphically with a Q–Q plot against a standard normal reference distribution, or it can be tested formally using one of the many numerical tests of normality (see Section 3.1.2 for more details of univariate tests of normality).

Multivariate tests. To test for multivariate normality it is not sufficient to test that the univariate margins of the distribution are normal. We will see in Chapter 7 that it is possible to have multivariate distributions with normal margins that are not themselves multivariate normal distributions. Thus we also need to be able to test *joint normality*, and a simple way of doing this is to exploit the fact that the quadratic

form in (6.14) has a chi-squared distribution. Suppose we estimate $\boldsymbol{\mu}$ and Σ using the standard estimators in (6.9) and construct the data

$$\{D_i^2 = (\boldsymbol{X}_i - \bar{\boldsymbol{X}})' S^{-1} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}) : i = 1, \ldots, n\}. \tag{6.15}$$

Because the estimates of the mean vector and the covariance matrix are used in the construction of each D_i^2, these data are not independent, even if the original $\boldsymbol{X}_i$ data were. Moreover, the marginal distribution of D_i^2 under the null hypothesis is not exactly chi-squared; in fact, we have that $n(n-1)^{-2} D_i^2 \sim \text{Beta}(\frac{1}{2}d, \frac{1}{2}(n-d-1))$, so that the true distribution is a scaled beta distribution, although it turns out to be very close to chi-squared for large n. We expect $D_1^2, \ldots, D_n^2$ to behave roughly like an iid sample from a χ_d^2 distribution, and for simplicity we construct Q–Q plots against this distribution. (It is also possible to make Q–Q plots against the beta reference distribution, and these look very similar.)

Numerical tests of multivariate normality based on multivariate measures of skewness and kurtosis are also possible. Suppose we define, in analogy to (3.1),

$$b_d = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij}^3, \qquad k_d = \frac{1}{n} \sum_{i=1}^{n} D_i^4, \tag{6.16}$$

where D_i is given in (6.15) and is known as the *Mahalanobis distance* between $\boldsymbol{X}_i$ and $\bar{\boldsymbol{X}}$, and $D_{ij} = (\boldsymbol{X}_i - \bar{\boldsymbol{X}}) S^{-1} (\boldsymbol{X}_j - \bar{\boldsymbol{X}})$ is known as the *Mahalanobis angle* between $\boldsymbol{X}_i - \bar{\boldsymbol{X}}$ and $\boldsymbol{X}_j - \bar{\boldsymbol{X}}$. Under the null hypothesis of multivariate normality the asymptotic distributions of these statistics as $n \to \infty$ are

$$\tfrac{1}{6} n b_d \sim \chi^2_{d(d+1)(d+2)/6}, \qquad \frac{k_d - d(d+2)}{\sqrt{8d(d+2)/n}} \sim N(0, 1). \tag{6.17}$$

Mardia's test of multinormality involves comparing the skewness and kurtosis statistics with the above theoretical reference distributions. Since large values of the statistics cast doubt on the multivariate normal model, one-sided tests are generally performed. Usually, the tests of kurtosis and skewness are performed separately, although there are also a number of joint (or so-called omnibus) tests (see Notes and Comments).

Example 6.3 (on the normality of returns on Dow Jones 30 stocks). In Section 3.1.2 we applied univariate tests of normality to an arbitrary subgroup of ten stocks from the Dow Jones index. We took eight years of data spanning the period 1993–2000 and formed daily, weekly, monthly and quarterly logarithmic returns. In this example we apply Mardia's tests of multinormality based on both multivariate skewness and kurtosis to the multivariate data for all ten stocks. The results are shown in Table 6.1. We also compare the D_i^2 data (6.15) to a χ_{10}^2 distribution using a Q–Q plot (see Figure 6.2).

The daily, weekly and monthly return data fail the multivariate tests of normality. For quarterly return data the multivariate kurtosis test does not reject the null hypothesis but the skewness test does; the Q–Q plot in Figure 6.2 (d) looks slightly

Table 6.1. Mardia's tests of multivariate normality based on the multivariate measures of skewness and kurtosis in (6.16) and the asymptotic distributions in (6.17) (see Example 6.3 for details).

		Daily	Weekly	Monthly	Quarterly
	n	2020	416	96	32
	b_{10}	9.31	9.91	21.10	50.10
	p-value	0.00	0.00	0.00	0.02
	k_{10}	242.45	177.04	142.65	120.83
	p-value	0.00	0.00	0.00	0.44

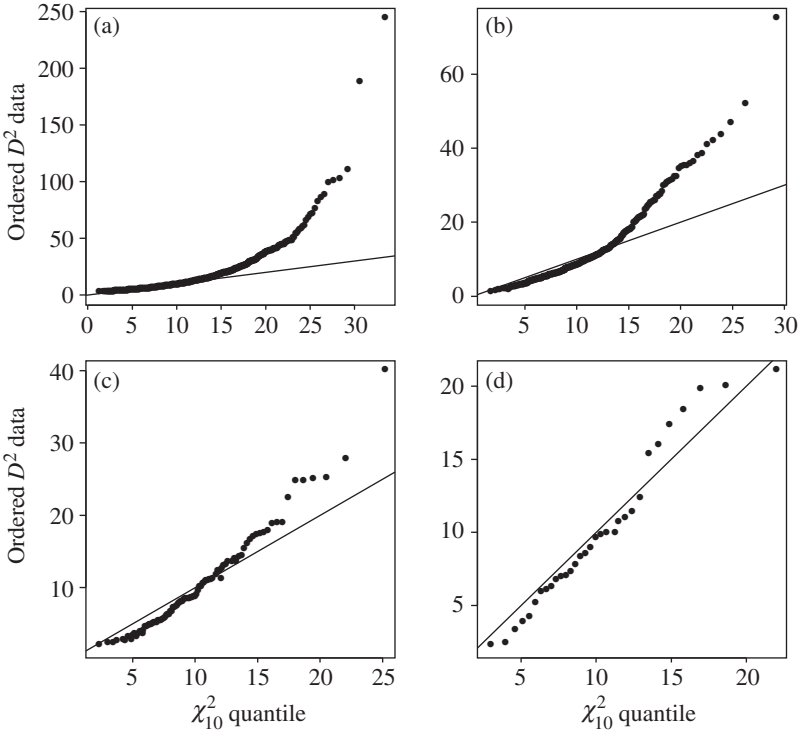

Figure 6.2. Q–Q plot of the D_i^2 data in (6.15) against a χ^2_{10} distribution for the data sets of Example 6.3: (a) daily analysis, (b) weekly analysis, (c) monthly analysis and (d) quarterly analysis. Under the null hypothesis of *multivariate* normality these should be roughly linear.

more linear. There is therefore some evidence that returns over a quarter year are close to being normally distributed, which might indicate a central limit theorem effect taking place, although the sample size is too small to reach any more reliable conclusion. The evidence against the multivariate normal distribution is certainly overwhelming for daily, weekly and monthly data.

The results in Example 6.3 are fairly typical for financial return data. This suggests that in many risk-management applications the multivariate normal distribution is

not a good description of reality. It has three main defects, all of which are discussed at various points in this book.

(1) The tails of its univariate marginal distributions are too thin; they do not assign enough weight to *extreme* events (see also Section 3.1.2).

(2) The joint tails of the distribution do not assign enough weight to *joint extreme* outcomes (see also Section 7.3.1).

(3) The distribution has a strong form of symmetry, known as elliptical symmetry.

In the next section we look at models that address some of these defects. We consider normal variance mixture models, which share the elliptical symmetry of the multivariate normal but have the flexibility to address (1) and (2) above; we also look at normal mean–variance mixture models, which introduce some asymmetry and thus address (3).

Notes and Comments

Much of the material covered briefly in Section 6.1 can be found in greater detail in standard texts on multivariate statistical analysis such as Mardia, Kent and Bibby (1979), Seber (1984), Giri (1996) and Johnson and Wichern (2002).

The true distribution of $D_i^2 = (\boldsymbol{X}_i - \bar{\boldsymbol{X}})S^{-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})$ for iid Gaussian data was shown by Gnanadesikan and Kettenring (1972) to be a scaled beta distribution (see also Gnanadesikan 1997). The implications of this fact for the construction of Q–Q plots in small samples are considered by Small (1978). References for multivariate measures of skewness and kurtosis and Mardia's test of multinormality are Mardia (1970, 1974, 1975). See also Mardia, Kent and Bibby (1979), the entry on "multivariate normality, testing for" in Volume 6 of the *Encyclopedia of Statistical Sciences* (Kotz, Johnson and Read 1985), and the entry on "Mardia's test of multinormality" in Volume 5 of the same publication. A paper that compares the performance of different goodness-of-fit tests for the multivariate normal distribution implemented in R is Joenssen and Vogel (2014).

6.2 Normal Mixture Distributions

In this section we generalize the multivariate normal to obtain multivariate normal mixture distributions. The crucial idea is the introduction of randomness into first the covariance matrix and then the mean vector of a multivariate normal distribution via a positive mixing variable, which will be known throughout as W.

6.2.1 Normal Variance Mixtures

Definition 6.4. The random vector $\boldsymbol{X}$ is said to have a (multivariate) normal variance mixture distribution if

$$\boldsymbol{X} \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + \sqrt{W}A\boldsymbol{Z}, \tag{6.18}$$

where

(i) $\boldsymbol{Z} \sim N_k(\mathbf{0}, I_k)$;

(ii) $W \geqslant 0$ is a non-negative, scalar-valued rv that is independent of $\boldsymbol{Z}$, and

(iii) $A \in \mathbb{R}^{d\times k}$ and $\boldsymbol{\mu} \in \mathbb{R}^d$ are a matrix and a vector of constants, respectively.

Such distributions are known as variance mixtures, since if we condition on the rv W, we observe that $\boldsymbol{X} \mid W = w \sim N_d(\boldsymbol{\mu}, w\Sigma)$, where $\Sigma = AA'$. The distribution of $\boldsymbol{X}$ can be thought of as a composite distribution constructed by taking a set of multivariate normal distributions with the same mean vector and with the same covariance matrix up to a multiplicative constant w. The mixture distribution is constructed by drawing randomly from this set of component multivariate normals according to a set of "weights" determined by the distribution of W; the resulting mixture is not itself a multivariate normal distribution. In the context of modelling risk-factor returns, the mixing variable W could be interpreted as a *shock* that arises from new information and impacts the volatilities of all risk factors.

As for the multivariate normal, we are most interested in the case where $\operatorname{rank}(A) = d \leqslant k$ and Σ is a full-rank, positive-definite matrix; this will give us a non-singular normal variance mixture.

Provided that W has a finite expectation, we may easily calculate that

$$E(\boldsymbol{X}) = E(\boldsymbol{\mu} + \sqrt{W}A\boldsymbol{Z}) = \boldsymbol{\mu} + E(\sqrt{W})AE(\boldsymbol{Z}) = \boldsymbol{\mu}$$

and that

$$\operatorname{cov}(\boldsymbol{X}) = E((\sqrt{W}A\boldsymbol{Z})(\sqrt{W}A\boldsymbol{Z})') = E(W)AE(\boldsymbol{Z}\boldsymbol{Z}')A' = E(W)\Sigma. \quad (6.19)$$

We generally refer to $\boldsymbol{\mu}$ and Σ as the *location vector* and the *dispersion matrix* of the distribution. Note that Σ (the covariance matrix of $A\boldsymbol{Z}$) is only the covariance matrix of $\boldsymbol{X}$ if $E(W) = 1$, and that $\boldsymbol{\mu}$ is only the mean vector when $E(\boldsymbol{X})$ is defined, which requires $E(W^{1/2}) < \infty$. The correlation matrices of $\boldsymbol{X}$ and $A\boldsymbol{Z}$ are the same when $E(W) < \infty$. Note also that these distributions provide good examples of models where a lack of correlation does not necessarily imply independence of the components of $\boldsymbol{X}$; indeed, we have the following simple result.

Lemma 6.5. *Let (X_1, X_2) have a normal mixture distribution with $A = I_2$ and $E(W) < \infty$ so that $\operatorname{cov}(X_1, X_2) = 0$. Then X_1 and X_2 are independent if and only if W is almost surely constant, i.e. (X_1, X_2) are normally distributed.*

Proof. If W is almost surely a constant, then (X_1, X_2) have a bivariate normal distribution and are independent. Conversely, if (X_1, X_2) are independent, then we must have $E(|X_1||X_2|) = E(|X_1|)E(|X_2|)$. We calculate that

$$\begin{aligned} E(|X_1||X_2|) &= E(W|Z_1||Z_2|) = E(W)E(|Z_1|)E(|Z_2|) \\ &\geqslant (E(\sqrt{W}))^2 E(|Z_1|)E(|Z_2|) = E(|X_1|)E(|X_2|), \end{aligned}$$

and we can only have equality throughout when W is a constant. □

Using (6.10), we can calculate that the characteristic function of a normal variance mixture is given by

$$\begin{aligned}\phi_{\boldsymbol{X}}(\boldsymbol{t}) = E(E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}} \mid W)) &= E(\exp(\mathrm{i}\boldsymbol{t}'\boldsymbol{\mu} - \tfrac{1}{2}W\boldsymbol{t}'\Sigma\boldsymbol{t})) \\ &= \mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{\mu}}\hat{H}(\tfrac{1}{2}\boldsymbol{t}'\Sigma\boldsymbol{t}), \end{aligned} \tag{6.20}$$

where $\hat{H}(\theta) = \int_0^\infty \mathrm{e}^{-\theta v}\,\mathrm{d}H(v)$ is the Laplace–Stieltjes transform of the df H of W. Based on (6.20) we use the notation $\boldsymbol{X} \sim M_d(\boldsymbol{\mu}, \Sigma, \hat{H})$ for normal variance mixtures.

Assuming that Σ is positive definite and that the distribution of W has no point mass at zero, we may derive the joint density of a normal variance mixture distribution. Writing $f_{\boldsymbol{X}|W}$ for the (Gaussian) conditional density of $\boldsymbol{X}$ given W, the density of $\boldsymbol{X}$ is given by

$$\begin{aligned} f(\boldsymbol{x}) &= \int f_{\boldsymbol{X}|W}(\boldsymbol{x} \mid w)\,\mathrm{d}H(w) \\ &= \int \frac{w^{-d/2}}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2w}\right\}\mathrm{d}H(w), \end{aligned} \tag{6.21}$$

in terms of the Lebesgue–Stieltjes integral; when H has density h we simply mean the Riemann integral $\int_0^\infty f_{\boldsymbol{X}|W}(\boldsymbol{x} \mid w)h(w)\,\mathrm{d}w$. All such densities will depend on $\boldsymbol{x}$ only through the quadratic form $(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$, and this means they are the densities of elliptical distributions, as will be discussed in Section 6.3.

Example 6.6 (multivariate two-point normal mixture distribution). Simple examples of normal mixtures are obtained when W is a discrete rv. For example, the two-point normal mixture model is obtained by taking W in (6.18) to be a discrete rv that assumes the distinct positive values k_1 and k_2 with probabilities p and $1-p$, respectively. By setting k_2 large relative to k_1 and choosing p large, this distribution might be used to define two regimes: an *ordinary* regime that holds most of the time and a *stress* regime that occurs with small probability $1-p$. Obviously this idea extends to k-point mixture models.

Example 6.7 (multivariate t distribution). If we take W in (6.18) to be an rv with an inverse gamma distribution $W \sim \mathrm{Ig}(\frac{1}{2}\nu, \frac{1}{2}\nu)$ (which is equivalent to saying that $\nu/W \sim \chi^2_\nu$), then $\boldsymbol{X}$ has a multivariate t distribution with ν degrees of freedom (see Section A.2.6 for more details concerning the inverse gamma distribution). Our notation for the multivariate t is $\boldsymbol{X} \sim t_d(\nu, \boldsymbol{\mu}, \Sigma)$, and we note that Σ is not the covariance matrix of $\boldsymbol{X}$ in this definition of the multivariate t. Since $E(W) = \nu/(\nu-2)$ we have $\mathrm{cov}(\boldsymbol{X}) = (\nu/(\nu-2))\Sigma$, and the covariance matrix (and correlation matrix) of this distribution is only defined if $\nu > 2$.

Using (6.21), the density can be calculated to be

$$f(\boldsymbol{x}) = \frac{\Gamma(\frac{1}{2}(\nu+d))}{\Gamma(\frac{1}{2}\nu)(\pi\nu)^{d/2}|\Sigma|^{1/2}}\left(1 + \frac{(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{\nu}\right)^{-(\nu+d)/2}. \tag{6.22}$$

Clearly, the locus of points with equal density is again an ellipsoid with equation $(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c$ for some $c > 0$. A bivariate example with four degrees of freedom is given in Figure 6.1. In comparison with the multivariate normal, the contours of equal density rise more quickly in the centre of the distribution and decay more gradually on the "lower slopes" of the distribution. In comparison with the multivariate normal, the multivariate t has heavier marginal tails (as discussed in Section 5.1.2) and a more pronounced tendency to generate simultaneous extreme values (see also Section 7.3.1).

Example 6.8 (symmetric generalized hyperbolic distribution). A flexible family of normal variance mixtures is obtained by taking W in (6.18) to have a generalized inverse Gaussian (GIG) distribution, $W \sim N^-(\lambda, \chi, \psi)$ (see Section A.2.5). Using (6.21), it can be shown that a normal variance mixture constructed with this mixing density has the joint density

$$f(\boldsymbol{x}) = \frac{(\sqrt{\chi\psi})^{-\lambda}\psi^{d/2}}{(2\pi)^{d/2}|\Sigma|^{1/2}K_\lambda(\sqrt{\chi\psi})} \frac{K_{\lambda-(d/2)}(\sqrt{(\chi+(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))\psi})}{(\sqrt{(\chi+(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))\psi})^{(d/2)-\lambda}}, \tag{6.23}$$

where K_λ denotes a modified Bessel function of the third kind (see Section A.2.5 for more details). This distribution is a special case of the more general family of multivariate generalized hyperbolic distributions, which we will discuss in greater detail in Section 6.2.2. The more general family can be obtained as *mean–variance* mixtures of normals, which are not necessarily elliptical distributions.

The GIG mixing distribution is very flexible and contains the gamma and inverse gamma distributions as special boundary cases (corresponding, respectively, to $\lambda > 0$, $\chi = 0$ and to $\lambda < 0$, $\psi = 0$). In these cases the density in (6.23) should be interpreted as a limit as $\chi \to 0$ or as $\psi \to 0$. (Information on the limits of Bessel functions is found in Section A.2.5.) The gamma mixing distribution yields Laplace distributions or so-called symmetric variance-gamma (VG) models, and the inverse gamma yields the t as in Example 6.7; to be precise, the t corresponds to the case when $\lambda = -\nu/2$ and $\chi = \nu$. The special cases $\lambda = -0.5$ and $\lambda = 1$ have also attracted attention in financial modelling. The former gives rise to the symmetric normal inverse Gaussian (NIG) distribution; the latter gives rise to a symmetric multivariate distribution whose one-dimensional margins are known simply as hyperbolic distributions.

To calculate the covariance matrix of distributions in the symmetric generalized hyperbolic family, we require the mean of the GIG distribution, which is given in (A.15) for the case $\chi > 0$ and $\psi > 0$. The covariance matrix of the multivariate distribution in (6.23) follows from (6.19).

Normal variance mixture distributions are easy to work with under linear operations, as shown in the following simple proposition.

Proposition 6.9. *If* $\boldsymbol{X} \sim M_d(\boldsymbol{\mu}, \Sigma, \hat{H})$ *and* $\boldsymbol{Y} = B\boldsymbol{X} + \boldsymbol{b}$, *where* $B \in \mathbb{R}^{k\times d}$ *and* $\boldsymbol{b} \in \mathbb{R}^k$, *then* $\boldsymbol{Y} \sim M_k(B\boldsymbol{\mu} + \boldsymbol{b}, B\Sigma B', \hat{H})$.

Proof. The characteristic function in (6.20) may be used to show that

$$\phi_{\boldsymbol{Y}}(\boldsymbol{t}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'(B\boldsymbol{X}+\boldsymbol{b})}) = \mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{b}}\phi_{\boldsymbol{X}}(B'\boldsymbol{t}) = \mathrm{e}^{\mathrm{i}\boldsymbol{t}'(B\boldsymbol{\mu}+\boldsymbol{b})}\hat{H}(\tfrac{1}{2}\boldsymbol{t}'B\Sigma B'\boldsymbol{t}).$$

□

The subclass of mixture distributions specified by $\hat{H}$ is therefore closed under linear transformations. For example, if $\boldsymbol{X}$ has a multivariate t distribution with ν degrees of freedom, then so does any linear transformation of $\boldsymbol{X}$; the linear combination $\boldsymbol{a}'\boldsymbol{X}$ would have a univariate t distribution with ν degrees of freedom (more precisely, the distribution $\boldsymbol{a}'\boldsymbol{X} \sim t_1(\nu, \boldsymbol{a}'\boldsymbol{\mu}, \boldsymbol{a}'\Sigma\boldsymbol{a})$).

Normal variance mixture distributions (and the mean–variance mixtures considered later in Section 6.2.2) are easily simulated, the method being obvious from Definition 6.4. To generate a variate $\boldsymbol{X} \sim M_d(\boldsymbol{\mu}, \Sigma, \hat{H})$ with Σ positive definite, we use the following algorithm.

Algorithm 6.10 (simulation of normal variance mixtures).

(1) Generate $\boldsymbol{Z} \sim N_d(\boldsymbol{0}, \Sigma)$ using Algorithm 6.2.

(2) Generate independently a positive mixing variable W with df H (corresponding to the Laplace–Stieltjes transform $\hat{H}$).

(3) Set $\boldsymbol{X} = \boldsymbol{\mu} + \sqrt{W}\boldsymbol{Z}$.

To generate $\boldsymbol{X} \sim t_d(\nu, \boldsymbol{\mu}, \Sigma)$, the mixing variable W should have an $\mathrm{Ig}(\frac{1}{2}\nu, \frac{1}{2}\nu)$ distribution; it is helpful to note that in this case $\nu/W \sim \chi^2_\nu$, a chi-squared distribution with ν degrees of freedom. Sampling from a generalized hyperbolic distribution with density (6.23) requires us to generate $W \sim N^-(\lambda, \chi, \psi)$. Sampling from the GIG distribution can be accomplished using a rejection algorithm proposed by Atkinson (1982).

6.2.2 Normal Mean–Variance Mixtures

All of the multivariate distributions we have considered so far have elliptical symmetry (see Section 6.3.2 for explanation) and this may well be an oversimplified model for real risk-factor return data. Among other things, elliptical symmetry implies that all one-dimensional marginal distributions are rigidly symmetric, which contradicts the frequent observation for stock returns that negative returns (losses) have heavier tails than positive returns (gains). The models we now introduce attempt to add some asymmetry to the class of normal mixtures by mixing normal distributions with different means as well as different variances; this yields the class of multivariate normal mean–variance mixtures.

Definition 6.11. The random vector $\boldsymbol{X}$ is said to have a (multivariate) normal mean–variance mixture distribution if

$$\boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{m}(W) + \sqrt{W}A\boldsymbol{Z}, \tag{6.24}$$

where

(i) $\boldsymbol{Z} \sim N_k(\mathbf{0}, I_k)$,

(ii) $W \geqslant 0$ is a non-negative, scalar-valued rv which is independent of $\boldsymbol{Z}$,

(iii) $A \in \mathbb{R}^{d \times k}$ is a matrix, and

(iv) $\boldsymbol{m}\colon [0, \infty) \to \mathbb{R}^d$ is a measurable function.

In this case we have that

$$\boldsymbol{X} \mid W = w \sim N_d(\boldsymbol{m}(w), w\Sigma), \tag{6.25}$$

where $\Sigma = AA'$ and it is clear why such distributions are known as mean–variance mixtures of normals. In general, such distributions are not elliptical.

A possible concrete specification for the function $\boldsymbol{m}(W)$ in (6.25) is

$$\boldsymbol{m}(W) = \boldsymbol{\mu} + W\boldsymbol{\gamma}, \tag{6.26}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ are parameter vectors in $\mathbb{R}^d$. Since $E(\boldsymbol{X} \mid W) = \boldsymbol{\mu} + W\boldsymbol{\gamma}$ and $\operatorname{cov}(\boldsymbol{X} \mid W) = W\Sigma$, it follows in this case by simple calculations that

$$E(\boldsymbol{X}) = E(E(\boldsymbol{X} \mid W)) = \boldsymbol{\mu} + E(W)\boldsymbol{\gamma}, \tag{6.27}$$

$$\begin{aligned} \operatorname{cov}(\boldsymbol{X}) &= E(\operatorname{cov}(\boldsymbol{X} \mid W)) + \operatorname{cov}(E(\boldsymbol{X} \mid W)) \\ &= E(W)\Sigma + \operatorname{var}(W)\boldsymbol{\gamma}\boldsymbol{\gamma}' \end{aligned} \tag{6.28}$$

when the mixing variable W has finite variance. We observe from (6.27) and (6.28) that the parameters $\boldsymbol{\mu}$ and Σ are not, in general, the mean vector and covariance matrix of $\boldsymbol{X}$ (or a multiple thereof). This is only the case when $\boldsymbol{\gamma} = \mathbf{0}$, so that the distribution is a normal variance mixture and the simpler moment formulas given in (6.19) apply.

6.2.3 Generalized Hyperbolic Distributions

In Example 6.8 we looked at the special subclass of the generalized hyperbolic (GH) distributions consisting of the elliptically symmetric normal variance mixture distributions. The full GH family is obtained using the mean–variance mixture construction (6.24) and the conditional mean specification (6.26). For the mixing distribution we assume that $W \sim N^-(\lambda, \chi, \psi)$, a GIG distribution with density (A.14).

Remark 6.12. This class of distributions has received a lot of attention in the financial-modelling literature, particularly in the univariate case. An important reason for this attention is their link to Lévy processes, i.e. processes with independent and stationary increments (like Brownian motion or the compound Poisson distribution) that are used to model price processes in continuous time. For every GH distribution it is possible to construct a Lévy process so that the value of the increment of the process over a fixed time interval has that distribution; this is only possible because the GH law is a so-called infinitely divisible distribution, a property that it inherits from the GIG mixing distribution of W.

The joint density in the non-singular case (Σ has rank d) is

$$f(\boldsymbol{x}) = \int_0^\infty \frac{\mathrm{e}^{(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}\boldsymbol{\gamma}}}{(2\pi)^{d/2}|\Sigma|^{1/2}w^{d/2}} \times \exp\left\{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2w} - \frac{\boldsymbol{\gamma}'\Sigma^{-1}\boldsymbol{\gamma}}{2/w}\right\} h(w)\,\mathrm{d}w,$$

where $h(w)$ is the density of W. Evaluation of this integral gives the GH density

$$f(\boldsymbol{x}) = c\frac{K_{\lambda-(d/2)}(\sqrt{(\chi+(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))(\psi+\boldsymbol{\gamma}'\Sigma^{-1}\boldsymbol{\gamma})})\mathrm{e}^{(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}\boldsymbol{\gamma}}}{(\sqrt{(\chi+(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))(\psi+\boldsymbol{\gamma}'\Sigma^{-1}\boldsymbol{\gamma})})^{(d/2)-\lambda}}, \tag{6.29}$$

where the normalizing constant is

$$c = \frac{(\sqrt{\chi\psi})^{-\lambda}\psi^{\lambda}(\psi+\boldsymbol{\gamma}'\Sigma^{-1}\boldsymbol{\gamma})^{(d/2)-\lambda}}{(2\pi)^{d/2}|\Sigma|^{1/2}K_{\lambda}(\sqrt{\chi\psi})}.$$

Clearly, if $\boldsymbol{\gamma} = \mathbf{0}$, the distribution reduces to the symmetric GH special case of Example 6.8. In general, we have a non-elliptical distribution with asymmetric margins. The mean vector and covariance matrix of the distribution are easily calculated from (6.27) and (6.28) using the information on the GIG and its moments given in Section A.2.5. The characteristic function of the GH distribution may be calculated using the same approach as in (6.20) to yield

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}}) = \mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{\mu}}\hat{H}(\tfrac{1}{2}\boldsymbol{t}'\Sigma\boldsymbol{t} - \mathrm{i}\boldsymbol{t}'\boldsymbol{\gamma}), \tag{6.30}$$

where $\hat{H}$ is the Laplace–Stieltjes transform of the GIG distribution.

We adopt the notation $\boldsymbol{X} \sim \mathrm{GH}_d(\lambda, \chi, \psi, \boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma})$. Note that the distributions $\mathrm{GH}_d(\lambda, \chi/k, k\psi, \boldsymbol{\mu}, k\Sigma, k\boldsymbol{\gamma})$ and $\mathrm{GH}_d(\lambda, \chi, \psi, \boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma})$ are identical for any $k > 0$, which causes an *identifiability problem* when we attempt to estimate the parameters in practice. This can be solved by constraining the determinant $|\Sigma|$ to be a particular value (such as one) when fitting. Note that, while such a constraint will have an effect on the values of χ and ψ that we estimate, it will not have an effect on the value of $\chi\psi$, so this product is a useful summary parameter for the GH distribution.

Linear combinations. The GH class is closed under linear operations.

Proposition 6.13. *If $\boldsymbol{X} \sim \mathrm{GH}_d(\lambda, \chi, \psi, \boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma})$ and $\boldsymbol{Y} = B\boldsymbol{X} + \boldsymbol{b}$, where $B \in \mathbb{R}^{k\times d}$ and $\boldsymbol{b} \in \mathbb{R}^k$, then $\boldsymbol{Y} \sim \mathrm{GH}_k(\lambda, \chi, \psi, B\boldsymbol{\mu}+\boldsymbol{b}, B\Sigma B', B\boldsymbol{\gamma})$.*

Proof. We calculate, using (6.30) and a similar method to Proposition 6.9, that

$$\phi_{\boldsymbol{Y}}(\boldsymbol{t}) = \mathrm{e}^{\mathrm{i}\boldsymbol{t}'(B\boldsymbol{\mu}+\boldsymbol{b})}\hat{H}(\tfrac{1}{2}\boldsymbol{t}'B\Sigma B'\boldsymbol{t} - \mathrm{i}\boldsymbol{t}'B\boldsymbol{\gamma}).$$

□

The parameters inherited from the GIG mixing distribution therefore remain unchanged under linear operations. This means, for example, that margins of $\boldsymbol{X}$ are easy to calculate; we have that $X_i \sim \mathrm{GH}_1(\lambda, \chi, \psi, \mu_i, \Sigma_{ii}, \gamma_i)$.

Parametrizations. There is a bewildering array of alternative parametrizations for the GH distribution in the literature, and it is more common to meet this distribution in a reparametrized form. In one common version the dispersion matrix we call Σ is renamed Δ and the constraint that $|\Delta| = 1$ is imposed; this addresses the identifiability problem mentioned above. The skewness parameters $\boldsymbol{\gamma}$ are replaced by parameters $\boldsymbol{\beta}$, and the non-negative parameters χ and ψ are replaced by the non-negative parameters δ and α according to

$$\boldsymbol{\beta} = \Delta^{-1}\boldsymbol{\gamma}, \qquad \delta = \sqrt{\chi}, \qquad \alpha = \sqrt{\psi + \boldsymbol{\gamma}'\Delta^{-1}\boldsymbol{\gamma}}.$$

These parameters must satisfy the constraints $\delta \geqslant 0$, $\alpha^2 > \boldsymbol{\beta}'\Delta\boldsymbol{\beta}$ if $\lambda > 0$; $\delta > 0$, $\alpha^2 > \boldsymbol{\beta}'\Delta\boldsymbol{\beta}$ if $\lambda = 0$; and $\delta > 0$, $\alpha^2 \geqslant \boldsymbol{\beta}'\Delta\boldsymbol{\beta}$ if $\lambda < 0$. Blæsild (1981) uses this parametrization to show that GH distributions form a closed class of distributions under linear operations and conditioning. However, the parametrization does have the problem that the important parameters α and δ are not generally invariant under either of these operations.

It is useful to be able to move easily between our χ–ψ–Σ–$\boldsymbol{\gamma}$ parametrization, as in (6.29), and the α–δ–Δ–$\boldsymbol{\beta}$ parametrization; λ and $\boldsymbol{\mu}$ are common to both parametrizations. If the χ–ψ–Σ–$\boldsymbol{\gamma}$ parametrization is used, then the formulas for obtaining the other parametrization are

$$\Delta = |\Sigma|^{-1/d}\Sigma, \qquad \boldsymbol{\beta} = \Sigma^{-1}\boldsymbol{\gamma},$$
$$\delta = \sqrt{\chi|\Sigma|^{1/d}}, \qquad \alpha = \sqrt{|\Sigma|^{-1/d}(\psi + \boldsymbol{\gamma}'\Sigma^{-1}\boldsymbol{\gamma})}.$$

If the α–δ–Δ–$\boldsymbol{\beta}$ form is used, then we can obtain our parametrization by setting

$$\Sigma = \Delta, \qquad \gamma = \Delta\boldsymbol{\beta}, \qquad \chi = \delta^2, \qquad \psi = \alpha^2 - \boldsymbol{\beta}'\Delta\boldsymbol{\beta}.$$

Special cases. The multivariate GH family is extremely flexible and, as we have mentioned, contains many special cases known by alternative names.

- If $\lambda = \frac{1}{2}(d+1)$, we drop the word "generalized" and refer to the distribution as a d-dimensional hyperbolic distribution. Note that the univariate margins of this distribution also have $\lambda = \frac{1}{2}(d+1)$ and are not one-dimensional hyperbolic distributions.
- If $\lambda = 1$, we get a multivariate distribution whose univariate margins are one-dimensional hyperbolic distributions. The one-dimensional hyperbolic distribution has been widely used in univariate analyses of financial return data (see Notes and Comments).
- If $\lambda = -\frac{1}{2}$, then the distribution is known as an NIG distribution. In the univariate case this model has also been used in analyses of return data; its functional form is similar to the hyperbolic distribution but with a slightly heavier tail. (Note that the NIG and the GIG are different distributions!)
- If $\lambda > 0$ and $\chi = 0$, we get a limiting case of the distribution known variously as a generalized Laplace, Bessel function or VG distribution.

- If $\lambda = -\frac{1}{2}\nu$, $\chi = \nu$ and $\psi = 0$, we get another limiting case that seems to have been less well studied; it could be called an asymmetric or skewed t distribution. Evaluating the limit of (6.29) as $\psi \to 0$ yields the multivariate density

$$f(\boldsymbol{x}) = c\frac{K_{(\nu+d)/2}(\sqrt{(\nu + Q(\boldsymbol{x}))\boldsymbol{\gamma}'\Sigma^{-1}\boldsymbol{\gamma}})\exp((\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}\boldsymbol{\gamma})}{(\sqrt{(\nu + Q(\boldsymbol{x}))\boldsymbol{\gamma}'\Sigma^{-1}\boldsymbol{\gamma}})^{-(\nu+d)/2}(1 + (Q(\boldsymbol{x})/\nu))^{(\nu+d)/2}}, \quad (6.31)$$

  where $Q(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ and the normalizing constant is

$$c = \frac{2^{1-(\nu+d)/2}}{\Gamma(\frac{1}{2}\nu)(\pi\nu)^{d/2}|\Sigma|^{1/2}}.$$

  This density reduces to the standard multivariate t density in (6.22) as $\boldsymbol{\gamma} \to \mathbf{0}$.

6.2.4 Empirical Examples

In this section we fit the multivariate GH distribution to real data and examine which of the subclasses—such as t, hyperbolic or NIG—are most useful; we also explore whether the general mean–variance mixture models can be replaced by (elliptically symmetric) variance mixtures. Our first example prepares the ground for multivariate examples by looking briefly at univariate models. The univariate distributions are fitted by straightforward numerical maximization of the log-likelihood. The multivariate distributions are fitted by using a variant of the EM algorithm, as described in Section 15.1.1.

Example 6.14 (univariate stock returns). In the literature, the NIG, hyperbolic and t models have been particularly popular special cases. We fit symmetric and asymmetric cases of these distributions to the data used in Example 6.3, restricting attention to daily and weekly returns, where the data are more plentiful ($n = 2020$ and $n = 468$, respectively). Models are fitted using maximum likelihood under the simplifying assumption that returns form iid samples; a simple quasi-Newton method provides a viable alternative to the EM algorithm in the univariate case.

In the upper two panels of Table 6.2 we show results for symmetric models. The t, NIG and hyperbolic models may be compared directly using the log-likelihood at the maximum, since all have the same number of parameters: for daily data we find that eight out of ten stocks prefer the t distribution to the hyperbolic and NIG distributions; for weekly returns the t distribution is favoured in six out of ten cases. Overall, the second best model appears to be the NIG distribution. The mixture models fit much better than the Gaussian model in all cases, and it may be easily verified using the Akaike information criterion (AIC) that they are preferred to the Gaussian model in a formal comparison (see Section A.3.6 for more on the AIC).

For the asymmetric models, we only show cases where at least one of the asymmetric t, NIG or hyperbolic models offered a significant improvement ($p < 0.05$) on the corresponding symmetric model according to a likelihood ratio test. This occurred for weekly returns on Citigroup (C) and Intel (INTC) but for no daily returns. For Citigroup the p-values of the tests were, respectively, 0.06, 0.04 and

Table 6.2. Comparison of univariate models in the GH family, showing estimates of selected parameters and the value of the log-likelihood at the maximum; bold numbers indicate the models that give the largest values of the log-likelihood. See Example 6.14 for commentary.

	Gauss	t model		NIG model		Hyperbolic model	
Stock	$\ln L$	ν	$\ln L$	$\sqrt{\chi\psi}$	$\ln L$	$\sqrt{\chi\psi}$	$\ln L$
Daily returns: symmetric models							
AXP	4945.7	5.8	5001.8	1.6	**5002.4**	1.3	5002.1
EK	5112.9	3.8	**5396.2**	0.8	5382.5	0.6	5366.0
BA	5054.9	3.8	**5233.5**	0.8	5229.1	0.5	5221.2
C	4746.6	6.3	**4809.5**	1.9	4806.8	1.7	4805.0
KO	5319.6	5.1	**5411.0**	1.4	5407.3	1.3	5403.3
MSFT	4724.3	5.8	**4814.6**	1.6	4809.5	1.5	4806.4
HWP	4480.1	4.5	**4588.8**	1.1	4587.2	0.9	4583.4
INTC	4392.3	5.4	**4492.2**	1.5	4486.7	1.4	4482.4
JPM	4898.3	5.1	4967.8	1.3	4969.5	0.9	**4969.7**
DIS	5047.2	4.4	**5188.3**	1	5183.8	0.8	5177.6
Weekly returns: symmetric models							
AXP	719.9	8.8	724.2	3.0	**724.3**	2.8	724.3
EK	718.7	3.6	**765.6**	0.7	764.0	0.5	761.3
BA	732.4	4.4	**759.2**	1.0	758.3	0.8	757.2
C	656.0	5.7	**669.6**	1.6	669.3	1.3	669
KO	757.1	6.0	765.7	1.7	766.2	1.3	**766.3**
MSFT	671.5	6.3	**683.9**	1.9	683.2	1.8	682.9
HWP	627.1	6.0	637.3	1.8	**637.3**	1.5	637.1
INTC	595.8	5.2	**611.0**	1.5	610.6	1.3	610
JPM	681.7	5.9	**693.0**	1.7	692.9	1.5	692.6
DIS	734.1	6.4	742.7	1.9	**742.8**	1.7	742.7
Weekly returns: asymmetric models							
C	NA	6.1	**671.4**	1.7	671.3	1.3	671.2
INTC	NA	6.3	**614.2**	1.8	613.9	1.7	613.3

0.04 for the t, NIG and hyperbolic cases; for Intel the p-values were 0.01 in all cases, indicating quite strong asymmetry.

In the case of Intel we have superimposed the densities of various fitted asymmetric distributions on a histogram of the data in Figure 6.3. A plot of the log densities shown alongside reveals the differences between the distributions in the tail area. The left tail (corresponding to losses) appears to be heavier for these data, and the best-fitting distribution according to the likelihood comparison is the asymmetric t distribution.

Example 6.15 (multivariate stock returns). We fitted multivariate models to the full ten-dimensional data set of log-returns used in the previous example. The resulting values of the maximized log-likelihood are shown in Table 6.3 along with p-values for a likelihood ratio test of all special cases against the (asymmetric) GH model. The number of parameters in each model is also given; note that the general

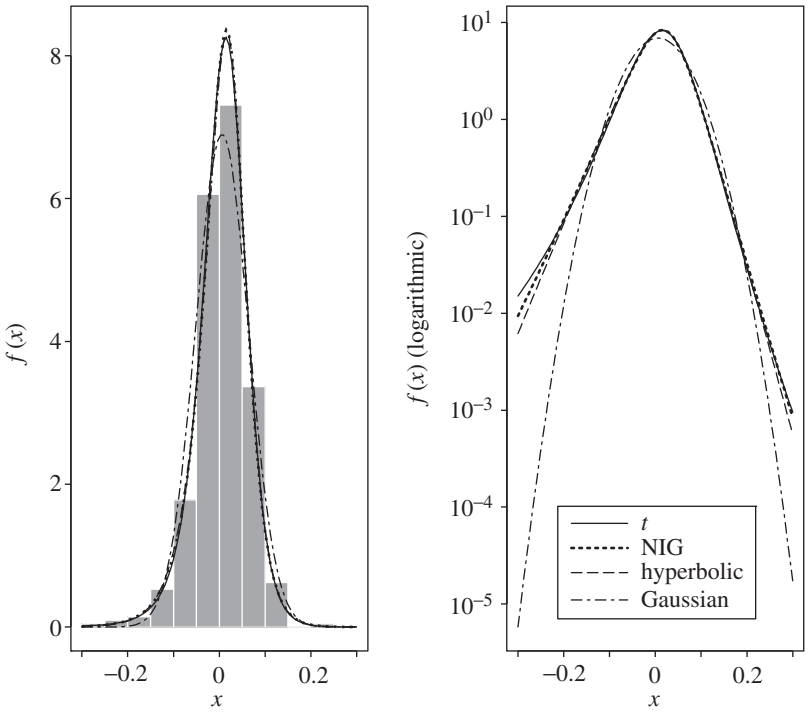

Figure 6.3. Models for weekly returns on Intel (INTC).

d-dimensional GH model has $\frac{1}{2}d(d+1)$ dispersion parameters, d location parameters, d skewness parameters and three parameters coming from the GIG mixing distribution, but is subject to one identifiability constraint; this gives $\frac{1}{2}(d(d+5)+4)$ free parameters.

For the daily data the best of the special cases is the skewed t distribution, which gives a value for the maximized likelihood that cannot be discernibly improved by the more general model with its additional parameter. All other non-elliptically symmetric submodels are rejected in a likelihood ratio test. Note, however, that the elliptically symmetric t distribution cannot be rejected when compared with the most general model, so that this seems to offer a simple parsimonious model for these data (the estimated degree of freedom is 6.0).

For the weekly data the best special case is the NIG distribution, followed closely by the skewed t; the hyperbolic and VG are rejected. The best elliptically symmetric special case seems to be the t distribution (the estimated degree of freedom this time being 6.2).

Example 6.16 (multivariate exchange-rate returns). We fitted the same multivariate models to a four-dimensional data set of exchange-rate log-returns, these being sterling, the euro, Japanese yen and Swiss franc against the US dollar for the period January 2000 to the end of March 2004 (1067 daily returns and 222 weekly returns). The resulting values of the maximized log-likelihood are shown in Table 6.4.

Table 6.3. A comparison of models in the GH family for ten-dimensional stock-return data. For each model, the table shows the value of the log-likelihood at the maximum ($\ln L$), the numbers of parameters ("# par.") and the p-value for a likelihood ratio test against the general GH model. The log-likelihood values for the general model, the best special case and the best elliptically symmetric special case are in bold type. See Example 6.15 for details.

	GH	NIG	Hyperbolic	t	VG	Gauss
Daily returns: asymmetric models						
$\ln L$	**52 174.62**	52 141.45	52 111.65	**52 174.62**	52 063.44	
# par.	77	76	76	76	76	
p-value		0.00	0.00	1.00	0.00	
Daily returns: symmetric models						
$\ln L$	52 170.14	52 136.55	52 106.34	**52 170.14**	52 057.38	50 805.28
# par.	67	66	66	66	66	65
p-value	0.54	0.00	0.00	0.63	0.00	0.00
Weekly returns: asymmetric models						
$\ln L$	**7 639.32**	**7 638.59**	7 636.49	7 638.56	7 631.33	
p-value		0.23	0.02	0.22	0.00	
Weekly returns: symmetric models						
$\ln L$	7 633.65	7 632.68	7 630.44	**7 633.11**	7 625.4	7 433.77
p-value	0.33	0.27	0.09	0.33	0.00	0.00

Table 6.4. A comparison of models in the GH family for four-dimensional exchange-rate return data. For each model, the table shows the value of the log-likelihood at the maximum ($\ln L$), the numbers of parameters ("# par.") and the p-value for a likelihood ratio test against the general GH model. The log-likelihood values for the general model, the best special case and the best elliptically symmetric special case are in bold type. See Example 6.16 for details.

	GH	NIG	Hyperbolic	t	VG	Gauss
Daily returns: asymmetric models						
$\ln L$	**17 306.44**	**17 306.43**	17 305.61	17 304.97	17 302.5	
# par.	20	19	19	19	19	
p-value		0.85	0.20	0.09	0.00	
Daily returns: symmetric models						
$\ln L$	17 303.10	**17 303.06**	17 302.15	17 301.85	17 299.15	17 144.38
# par.	16	15	15	15	15	14
p-value	0.15	0.24	0.13	0.10	0.01	0.00
Weekly returns: asymmetric models						
$\ln L$	**2 890.65**	2 889.90	2 889.65	**2 890.65**	2 888.98	
p-value		0.22	0.16	1.00	0.07	
Weekly returns: symmetric models						
$\ln L$	2 887.52	2 886.74	2 886.48	**2 887.52**	2 885.86	2 872.36
p-value	0.18	0.17	0.14	0.28	0.09	0.00

For the daily data the best of the special cases (both in general and if we restrict ourselves to symmetric models) is the NIG distribution, followed by the hyperbolic, t and VG distributions in that order. In a likelihood ratio test of the special cases against the general GH distribution, only the VG model is rejected at the 5% level; the skewed t model is rejected at the 10% level. When tested against the full model, certain elliptical models could not be rejected, the best of these being the NIG.

For the weekly data the best special case is the t distribution, followed by the NIG, hyperbolic and VG; none of the special cases can be rejected in a test at the 5% level, although the VG model is rejected at the 10% level. Among the elliptically symmetric distributions the Gauss distribution is clearly rejected, and the VG is again rejected at the 10% level, but otherwise the elliptical special cases are accepted; the best of these seems to be the t distribution, which has an estimated degrees-of-freedom parameter of 5.99.

Notes and Comments

Important early papers on multivariate normal mixtures are Kelker (1970) and Cambanis, Huang and Simons (1981). See also Bingham and Kiesel (2002), which contains an overview of the connections between the normal mixture, elliptical and hyperbolic models, and discusses their role in financial modelling. Fang, Kotz and Ng (1990) discuss the symmetric normal mixture models as special cases in their account of the more general family of spherical and elliptical distributions.

The GH distributions (univariate and multivariate) were introduced in Barndorff-Nielsen (1978) and further explored in Barndorff-Nielsen and Blæsild (1981). Useful references on the multivariate distribution are Blæsild (1981) and Blæsild and Jensen (1981). Generalized hyperbolic distributions (particularly in the univariate case) have been popularized as models for financial returns in recent papers by Eberlein and Keller (1995) and Eberlein, Keller and Prause (1998) (see also Bibby and Sørensen 2003). The PhD thesis of Prause (1999) is also a compendium of useful information in this context.

The reasons for their popularity in financial applications are both empirical and theoretical: they appear to provide a good fit to financial return data (again mostly in univariate investigations); they are consistent with continuous-time models, where logarithmic asset prices follow univariate or multivariate Lévy processes (thus generalizing the Black–Scholes model, where logarithmic prices follow Brownian motion); see Eberlein and Keller (1995) and Schoutens (2003).

For the NIG special case see Barndorff-Nielsen (1997), who discusses both univariate and multivariate cases and argues that the NIG is slightly superior to the hyperbolic as a univariate model for return data, a claim that our analyses support for stock-return data. Kotz, Kozubowski and Podgórski (2001) is a useful reference for the VG special case; the distribution appears here under the name generalized Laplace distribution and a (univariate or multivariate) Lévy process with VG-distributed increments is called a Laplace motion. The univariate Laplace motion is essentially the model proposed by Madan and Seneta (1990), who derived it as a Brownian motion under a stochastic time change and referred to it as the VG model

(see also Madan, Carr and Chang 1998). The multivariate t distribution is discussed in Kotz and Nadarajah (2004); the asymmetric or skewed t distribution presented in this chapter is also discussed in Bibby and Sørensen (2003). For alternative skewed extensions of the multivariate t, see Kotz and Nadarajah (2004) and Genton (2004).

6.3 Spherical and Elliptical Distributions

In the previous section we observed that normal variance mixture distributions—particularly the multivariate t and symmetric multivariate NIG—provided models that were far superior to the multivariate normal for daily and weekly US stock-return data. The more general asymmetric mean–variance mixture distributions did not seem to offer much of an improvement on the symmetric variance mixture models. While this was a single example, other investigations suggest that multivariate return data for groups of returns of a similar type often show similar behaviour.

The normal variance mixture distributions are so-called elliptical distributions, and in this section we look more closely at the theory of elliptical distributions. To do this we begin with the special case of spherical distributions.

6.3.1 Spherical Distributions

The spherical family constitutes a large class of distributions for random vectors with *uncorrelated* components and identical, symmetric marginal distributions. It is important to note that within this class, $N_d(\mathbf{0}, I_d)$ is the only model for a vector of mutually independent components. Many of the properties of elliptical distributions can best be understood by beginning with spherical distributions.

Definition 6.17. A random vector $\boldsymbol{X} = (X_1, \ldots, X_d)'$ has a spherical distribution if, for every orthogonal map $U \in \mathbb{R}^{d\times d}$ (i.e. maps satisfying $UU' = U'U = I_d$),

$$U\boldsymbol{X} \stackrel{\mathrm{d}}{=} \boldsymbol{X}.$$

Thus spherical random vectors are distributionally invariant under rotations. There are a number of different ways of defining distributions with this property, as we demonstrate below.

Theorem 6.18. *The following are equivalent.*

(1) $\boldsymbol{X}$ *is spherical.*

(2) *There exists a function* ψ *of a scalar variable such that, for all* $\boldsymbol{t} \in \mathbb{R}^d$,

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}}) = \psi(\boldsymbol{t}'\boldsymbol{t}) = \psi(t_1^2 + \cdots + t_d^2). \tag{6.32}$$

(3) *For every* $\boldsymbol{a} \in \mathbb{R}^d$,

$$\boldsymbol{a}'\boldsymbol{X} \stackrel{\mathrm{d}}{=} \|\boldsymbol{a}\| X_1, \tag{6.33}$$

where $\|\boldsymbol{a}\|^2 = \boldsymbol{a}'\boldsymbol{a} = a_1^2 + \cdots + a_d^2$.

Proof. (1) $\Rightarrow$ (2). If $\boldsymbol{X}$ is spherical, then for any orthogonal matrix U we have

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \phi_{U\boldsymbol{X}}(\boldsymbol{t}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'U\boldsymbol{X}}) = \phi_{\boldsymbol{X}}(U'\boldsymbol{t}).$$

This can only be true if $\phi_{\boldsymbol{X}}(\boldsymbol{t})$ only depends on the length of $\boldsymbol{t}$, i.e. if $\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \psi(\boldsymbol{t}'\boldsymbol{t})$ for some function ψ of a non-negative scalar variable.

(2) $\Rightarrow$ (3). First observe that $\phi_{X_1}(t) = E(\mathrm{e}^{\mathrm{i}tX_1}) = \phi_{\boldsymbol{X}}(t\boldsymbol{e}_1) = \psi(t^2)$, where $\boldsymbol{e}_1$ denotes the first unit vector in $\mathbb{R}^d$. It follows that for any $\boldsymbol{a} \in \mathbb{R}^d$,

$$\phi_{\boldsymbol{a}'\boldsymbol{X}}(t) = \phi_{\boldsymbol{X}}(t\boldsymbol{a}) = \psi(t^2\boldsymbol{a}'\boldsymbol{a}) = \psi(t^2\|\boldsymbol{a}\|^2) = \phi_{X_1}(t\|\boldsymbol{a}\|) = \phi_{\|\boldsymbol{a}\|X_1}(t).$$

(3) $\Rightarrow$ (1). For any orthogonal matrix U we have

$$\phi_{U\boldsymbol{X}}(\boldsymbol{t}) = E(\mathrm{e}^{\mathrm{i}(U'\boldsymbol{t})'\boldsymbol{X}}) = E(\mathrm{e}^{\mathrm{i}\|U'\boldsymbol{t}\|X_1}) = E(\mathrm{e}^{\mathrm{i}\|\boldsymbol{t}\|X_1}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}}) = \phi_{\boldsymbol{X}}(\boldsymbol{t}).$$

□

Part (2) of Theorem 6.18 shows that the characteristic function of a spherically distributed random vector is fully described by a function ψ of a scalar variable. For this reason ψ is known as the *characteristic generator* of the spherical distribution and the notation $\boldsymbol{X} \sim S_d(\psi)$ is used. Part (3) of Theorem 6.18 shows that linear combinations of spherical random vectors always have a distribution of the same *type*, so that they have the same distribution up to changes of location and scale (see Section A.1.1). This important property will be used in Chapter 8 to prove the subadditivity of value-at-risk for linear portfolios of elliptically distributed risk factors. We now give examples of spherical distributions.

Example 6.19 (multivariate normal). A random vector $\boldsymbol{X}$ with the standard uncorrelated normal distribution $N_d(\boldsymbol{0}, I_d)$ is clearly spherical. The characteristic function is

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}}) = \mathrm{e}^{-\boldsymbol{t}'\boldsymbol{t}/2},$$

so that, using part (2) of Theorem 6.18, $\boldsymbol{X} \sim S_d(\psi)$ with characteristic generator $\psi(t) = \mathrm{e}^{-t/2}$.

Example 6.20 (normal variance mixtures). A random vector $\boldsymbol{X}$ with a standardized, uncorrelated normal variance mixture distribution $M_d(\boldsymbol{0}, I_d, \hat{H})$ also has a spherical distribution. Using (6.20), we see that $\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \hat{H}(\frac{1}{2}\boldsymbol{t}'\boldsymbol{t})$, which obviously satisfies (6.32), and the characteristic generator of the spherical distribution is related to the Laplace–Stieltjes transform of the mixture distribution function of W by $\psi(t) = \hat{H}(\frac{1}{2}t)$. Thus $\boldsymbol{X} \sim M_d(\boldsymbol{0}, I_d, \hat{H}(\cdot))$ and $\boldsymbol{X} \sim S_d(\hat{H}(\frac{1}{2}\cdot))$ are two ways of writing the same mixture distribution.

A further, extremely important, way of characterizing spherical distributions is given by the following result.

Theorem 6.21. *$\boldsymbol{X}$ has a spherical distribution if and only if it has the stochastic representation*

$$\boldsymbol{X} \stackrel{\mathrm{d}}{=} R\boldsymbol{S}, \tag{6.34}$$

where $\boldsymbol{S}$ is uniformly distributed on the unit sphere $\mathcal{S}^{d-1} = \{\boldsymbol{s} \in \mathbb{R}^d : \boldsymbol{s}'\boldsymbol{s} = 1\}$ and $R \geqslant 0$ is a radial rv, independent of $\boldsymbol{S}$.

Proof. First we prove that if $\boldsymbol{S}$ is uniformly distributed on the unit sphere and $R \geqslant 0$ is an independent scalar variable, then $R\boldsymbol{S}$ has a spherical distribution. This is seen by considering the characteristic function

$$\phi_{R\boldsymbol{S}}(\boldsymbol{t}) = E(\mathrm{e}^{\mathrm{i}R\boldsymbol{t}'\boldsymbol{S}}) = E(E(\mathrm{e}^{\mathrm{i}R\boldsymbol{t}'\boldsymbol{S}} \mid R)).$$

Since $\boldsymbol{S}$ is itself spherically distributed, its characteristic function has a characteristic generator, which is usually given the special notation Ω_d. Thus, by Theorem 6.18 (2) we have that

$$\phi_{R\boldsymbol{S}}(\boldsymbol{t}) = E(\Omega_d(R^2\boldsymbol{t}'\boldsymbol{t})) = \int \Omega_d(r^2\boldsymbol{t}'\boldsymbol{t})\,\mathrm{d}F(r), \tag{6.35}$$

where F is the df of R. Since this is a function of $\boldsymbol{t}'\boldsymbol{t}$, it follows, again from Theorem 6.18 (2), that $R\boldsymbol{S}$ has a spherical distribution.

We now prove that if the random vector $\boldsymbol{X}$ is spherical, then it has the representation (6.34). For any arbitrary $\boldsymbol{s} \in \mathcal{S}^{d-1}$, the characteristic generator ψ of $\boldsymbol{X}$ must satisfy $\psi(\boldsymbol{t}'\boldsymbol{t}) = \phi_{\boldsymbol{X}}(\boldsymbol{t}) = \phi_{\boldsymbol{X}}(\|\boldsymbol{t}\|\boldsymbol{s})$. It follows that, if we introduce a random vector $\boldsymbol{S}$ that is uniformly distributed on the sphere $\mathcal{S}^{d-1}$, we can write

$$\psi(\boldsymbol{t}'\boldsymbol{t}) = \int_{\mathcal{S}^{d-1}} \phi_{\boldsymbol{X}}(\|\boldsymbol{t}\|\boldsymbol{s})\,\mathrm{d}F_{\boldsymbol{S}}(\boldsymbol{s}) = \int_{\mathcal{S}^{d-1}} E(\mathrm{e}^{\mathrm{i}\|\boldsymbol{t}\|\boldsymbol{s}'\boldsymbol{X}})\,\mathrm{d}F_{\boldsymbol{S}}(\boldsymbol{s}).$$

Interchanging the order of integration and using the Ω_d notation for the characteristic generator of $\boldsymbol{S}$ we have

$$\psi(\boldsymbol{t}'\boldsymbol{t}) = E(\Omega_d(\|\boldsymbol{t}\|^2\|\boldsymbol{X}\|^2)) = \int \Omega_d(\boldsymbol{t}'\boldsymbol{t}r^2)\,\mathrm{d}F_{\|\boldsymbol{X}\|}(r), \tag{6.36}$$

where $F_{\|\boldsymbol{X}\|}$ is the df of $\|\boldsymbol{X}\|$. By comparison with (6.35) we see that (6.36) is the characteristic function of $R\boldsymbol{S}$, where R is an rv with df $F_{\|\boldsymbol{X}\|}$ that is independent of $\boldsymbol{S}$. □

We often exclude from consideration distributions that place point mass at the origin; that is, we consider spherical rvs $\boldsymbol{X}$ in the subclass $S_d^+(\psi)$ for which $P(\boldsymbol{X} = \boldsymbol{0}) = 0$. A particularly useful corollary of Theorem 6.21 is then the following result, which is used in Section 15.1.2 to devise tests for spherical and elliptical symmetry.

Corollary 6.22. *Suppose* $\boldsymbol{X} \stackrel{\mathrm{d}}{=} R\boldsymbol{S} \sim S_d^+(\psi)$. *Then*

$$\left(\|\boldsymbol{X}\|, \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}\right) \stackrel{\mathrm{d}}{=} (R, \boldsymbol{S}). \tag{6.37}$$

Proof. Let $f_1(\boldsymbol{x}) = \|\boldsymbol{x}\|$ and $f_2(\boldsymbol{x}) = \boldsymbol{x}/\|\boldsymbol{x}\|$. It follows from (6.34) that

$$\left(\|\boldsymbol{X}\|, \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}\right) = (f_1(\boldsymbol{X}), f_2(\boldsymbol{X})) \stackrel{\mathrm{d}}{=} (f_1(R\boldsymbol{S}), f_2(R\boldsymbol{S})) = (R, \boldsymbol{S}).$$

□

Example 6.23 (working with R and S). Suppose $\boldsymbol{X} \sim N_d(\mathbf{0}, I_d)$. Since $\boldsymbol{X}'\boldsymbol{X} \sim \chi^2_d$, a chi-squared distribution with d degrees of freedom, it follows from (6.37) that $R^2 \sim \chi^2_d$.

We can use this fact to calculate $E(\boldsymbol{S})$ and $\mathrm{cov}(\boldsymbol{S})$, the first two moments of a uniform distribution on the unit sphere. We have that

$$\begin{aligned} \mathbf{0} = E(\boldsymbol{X}) = E(R)E(\boldsymbol{S}) &\Rightarrow E(\boldsymbol{S}) = \mathbf{0}, \\ I_d = \mathrm{cov}(\boldsymbol{X}) = E(R^2)\,\mathrm{cov}(\boldsymbol{S}) &\Rightarrow \mathrm{cov}(\boldsymbol{S}) = I_d/d, \end{aligned} \tag{6.38}$$

since $E(R^2) = d$ when $R^2 \sim \chi^2_d$.

Now suppose that $\boldsymbol{X}$ has a spherical normal variance mixture distribution $\boldsymbol{X} \sim M_d(\mathbf{0}, I_d, \hat{H})$ and we wish to calculate the distribution of $R^2 \stackrel{\mathrm{d}}{=} \boldsymbol{X}'\boldsymbol{X}$ in this case. Since $\boldsymbol{X} \stackrel{\mathrm{d}}{=} \sqrt{W}\boldsymbol{Y}$, where $\boldsymbol{Y} \sim N_d(\mathbf{0}, I_d)$ and W is independent of $\boldsymbol{Y}$, it follows that $R^2 \stackrel{\mathrm{d}}{=} W\tilde{R}^2$, where $\tilde{R}^2 \sim \chi^2_d$ and W and $\tilde{R}$ are independent. If we can calculate the distribution of the product of W and an independent chi-squared variate, then we have the distribution of R^2.

For a concrete example suppose that $\boldsymbol{X} \sim t_d(\nu, \mathbf{0}, I_d)$. For a multivariate t distribution we know from Example 6.7 that $W \sim \mathrm{Ig}(\frac{1}{2}\nu, \frac{1}{2}\nu)$, which means that $\nu/W \sim \chi^2_\nu$. Using the fact that the ratio of independent chi-squared rvs divided by their degrees of freedom is F-distributed, it may be calculated that $R^2/d \sim F(d, \nu)$, the F distribution on d with ν degrees of freedom (see Section A.2.3). Since an $F(d, \nu)$ distribution has mean $\nu/(\nu - 2)$, it follows from (6.38) that

$$\mathrm{cov}(\boldsymbol{X}) = E(\mathrm{cov}(R\boldsymbol{S} \mid R)) = E(R^2 I_d/d) = (\nu/(\nu - 2))I_d.$$

The normal mixtures with $\boldsymbol{\mu} = \mathbf{0}$ and $\Sigma = I_d$ represent an easily understood subgroup of the spherical distributions. There are other spherical distributions that cannot be represented as normal variance mixtures; an example is the distribution of the uniform vector $\boldsymbol{S}$ on $\mathcal{S}^{d-1}$ itself. However, the normal mixtures have a special role in the spherical world, as summarized by the following theorem.

Theorem 6.24. *Denote by Ψ_∞ the set of characteristic generators that generate a d-dimensional spherical distribution for arbitrary $d \geqslant 1$. Then $\boldsymbol{X} \sim S_d(\psi)$ with $\psi \in \Psi_\infty$ if and only if $\boldsymbol{X} \stackrel{\mathrm{d}}{=} \sqrt{W}\boldsymbol{Z}$, where $\boldsymbol{Z} \sim N_d(\mathbf{0}, I_d)$ is independent of $W \geqslant 0$.*

Proof. This is proved in Fang, Kotz and Ng (1990, pp. 48–51). □

Thus, the characteristic generators of normal mixtures generate spherical distributions in arbitrary dimensions, while other spherical generators may only be used in certain dimensions. A concrete example is given by the uniform distribution on the unit sphere. Let Ω_d denote the characteristic generator of the uniform vector $\boldsymbol{S} = (S_1, \ldots, S_d)'$ on $\mathcal{S}_{d-1}$. It can be shown that $\Omega_d((t_1, \ldots, t_{d+1})'(t_1, \ldots, t_{d+1}))$ is not the characteristic function of a spherical distribution in $\mathbb{R}^{d+1}$ (for more details see Fang, Kotz and Ng (1990, pp. 70–72)).

If a spherical distribution has a density f, then, by using the inversion formula

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^d} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\boldsymbol{t}'\boldsymbol{x}} \phi_{\boldsymbol{X}}(\boldsymbol{t})\, \mathrm{d}t_1 \cdots \mathrm{d}t_d,$$

it is easily inferred from Theorem 6.18 that $f(\boldsymbol{x}) = f(U\boldsymbol{x})$ for any orthogonal matrix U, so that the density must be of the form

$$f(\boldsymbol{x}) = g(\boldsymbol{x}'\boldsymbol{x}) = g(x_1^2 + \cdots + x_d^2) \tag{6.39}$$

for some function g of a scalar variable, which is referred to as the *density generator*. Clearly, the joint density is constant on hyperspheres $\{\boldsymbol{x} : x_1^2 + \cdots + x_d^2 = c\}$ in $\mathbb{R}^d$. To give a single example, the density generator of the multivariate t (i.e. the model $\boldsymbol{X} \sim t_d(\nu, \mathbf{0}, I_d)$ of Example 6.7) is

$$g(x) = \frac{\Gamma(\frac{1}{2}(\nu + d))}{\Gamma(\frac{1}{2}\nu)(\pi\nu)^{d/2}} \left(1 + \frac{x}{\nu}\right)^{-(\nu+d)/2}.$$

6.3.2 Elliptical Distributions

Definition 6.25. $\boldsymbol{X}$ has an elliptical distribution if

$$\boldsymbol{X} \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + A\boldsymbol{Y},$$

where $\boldsymbol{Y} \sim S_k(\psi)$ and $A \in \mathbb{R}^{d\times k}$ and $\boldsymbol{\mu} \in \mathbb{R}^d$ are a matrix and vector of constants, respectively.

In other words, elliptical distributions are obtained by multivariate *affine* transformations of spherical distributions. Since the characteristic function is

$$\phi_X(\boldsymbol{t}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'(\boldsymbol{\mu}+A\boldsymbol{Y})}) = \mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{\mu}} E(\mathrm{e}^{\mathrm{i}(A'\boldsymbol{t})'\boldsymbol{Y}}) = \mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{\mu}}\psi(\boldsymbol{t}'\Sigma\boldsymbol{t}),$$

where $\Sigma = AA'$, we denote the elliptical distributions by

$$\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$$

and refer to $\boldsymbol{\mu}$ as the location vector, Σ as the dispersion matrix and ψ as the characteristic generator of the distribution.

Remark 6.26. Knowledge of $\boldsymbol{X}$ does not uniquely determine its elliptical representation $E_d(\boldsymbol{\mu}, \Sigma, \psi)$. Although $\boldsymbol{\mu}$ is uniquely determined, Σ and ψ are only determined up to a positive constant. For example, the multivariate normal distribution $N_d(\boldsymbol{\mu}, \Sigma)$ can be written as $E_d(\boldsymbol{\mu}, \Sigma, \psi(\cdot))$ or $E_d(\boldsymbol{\mu}, c\Sigma, \psi(\cdot/c))$ for $\psi(u) = \mathrm{e}^{-u/2}$ and any $c > 0$. Provided that variances are finite, then an elliptical distribution is fully specified by its mean vector, covariance matrix and characteristic generator, and it is possible to find an elliptical representation $E_d(\boldsymbol{\mu}, \Sigma, \psi)$ such that Σ is the covariance matrix of $\boldsymbol{X}$, although this is not always the standard representation of the distribution.

We now give an alternative stochastic representation for the elliptical distributions that follows directly from Definition 6.25 and Theorem 6.21.

Proposition 6.27. *$\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ if and only if there exist $\boldsymbol{S}$, R and A satisfying*

$$\boldsymbol{X} \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + RA\boldsymbol{S}, \tag{6.40}$$

with

(i) $\boldsymbol{S}$ *uniformly distributed on the unit sphere* $\mathcal{S}^{k-1} = \{\boldsymbol{s} \in \mathbb{R}^k : \boldsymbol{s}'\boldsymbol{s} = 1\}$,

(ii) $R \geqslant 0$, *a radial rv, independent of* $\boldsymbol{S}$, *and*

(iii) $A \in \mathbb{R}^{d\times k}$ *with* $AA' = \Sigma$.

For practical examples we are most interested in the case where Σ is positive definite. The relation between the elliptical and spherical cases is then clearly

$$\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi) \iff \Sigma^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}) \sim S_d(\psi). \tag{6.41}$$

In this case, if the spherical vector $\boldsymbol{Y}$ has density generator g, then $\boldsymbol{X} = \boldsymbol{\mu} + \Sigma^{1/2}\boldsymbol{Y}$ has density

$$f(\boldsymbol{x}) = \frac{1}{|\Sigma|^{1/2}} g((\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})).$$

The joint density is always constant on sets of the form $\{\boldsymbol{x} : (\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c\}$, which are ellipsoids in $\mathbb{R}^d$. Clearly, the full family of multivariate normal variance mixtures with general location and dispersion parameters $\boldsymbol{\mu}$ and Σ are elliptical, since they are obtained by affine transformations of the spherical special cases considered in the previous section.

It follows from (6.37) and (6.41) that for a non-singular elliptical variate $\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ with no point mass at $\boldsymbol{\mu}$, we have

$$\left(\sqrt{(\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu})}, \frac{\Sigma^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})}{\sqrt{(\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu})}}\right) \stackrel{\mathrm{d}}{=} (R, \boldsymbol{S}), \tag{6.42}$$

where $\boldsymbol{S}$ is uniformly distributed on $\mathcal{S}^{d-1}$ and R is an independent scalar rv. This forms the basis of a test of elliptical symmetry described in Section 15.1.2.

The following proposition shows that a particular conditional distribution of an elliptically distributed random vector $\boldsymbol{X}$ has the same correlation matrix as $\boldsymbol{X}$ and can also be used to test for elliptical symmetry.

Proposition 6.28. *Let* $\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ *and assume that* Σ *is positive definite and* $\mathrm{cov}(\boldsymbol{X})$ *is finite. For any* $c \geqslant 0$ *such that* $P((\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \geqslant c) > 0$, *we have*

$$\rho(\boldsymbol{X} \mid (\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \geqslant c) = \rho(\boldsymbol{X}). \tag{6.43}$$

Proof. It follows easily from (6.42) that

$$\boldsymbol{X} \mid (\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \geqslant c \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + R\Sigma^{1/2}\boldsymbol{S} \mid R^2 \geqslant c,$$

where $R \stackrel{\mathrm{d}}{=} \sqrt{(\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu})}$ and $\boldsymbol{S}$ is independent of R and uniformly distributed on $\mathcal{S}^{d-1}$. Thus we have

$$\boldsymbol{X} \mid (\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \geqslant c \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + \tilde{R}\Sigma^{1/2}\boldsymbol{S},$$

where $\tilde{R} \stackrel{\mathrm{d}}{=} R \mid R^2 \geqslant c$. It follows from Proposition 6.27 that the conditional distribution remains elliptical with dispersion matrix Σ and that (6.43) holds. □

6.3.3 Properties of Elliptical Distributions

We now summarize some of the properties of elliptical distributions in a format that allows their comparison with the properties of multivariate normal distributions in Section 6.1.3. Many properties carry over directly and others only need to be modified slightly. These parallels emphasize that it would be fairly easy to base many standard procedures in risk management on an assumption that risk-factor changes have an approximately elliptical distribution, rather than the patently false assumption that they are multivariate normal.

Linear combinations. If we take linear combinations of elliptical random vectors, then these remain elliptical with the same characteristic generator ψ. Let $\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ and take any $B \in \mathbb{R}^{k\times d}$ and $\boldsymbol{b} \in \mathbb{R}^k$. Using a similar argument to that in Proposition 6.9 it is then easily shown that

$$B\boldsymbol{X} + \boldsymbol{b} \sim E_k(B\boldsymbol{\mu} + \boldsymbol{b}, B\Sigma B', \psi). \tag{6.44}$$

As a special case, if $\boldsymbol{a} \in \mathbb{R}^d$, then

$$\boldsymbol{a}'\boldsymbol{X} \sim E_1(\boldsymbol{a}'\boldsymbol{\mu}, \boldsymbol{a}'\Sigma\boldsymbol{a}, \psi). \tag{6.45}$$

Marginal distributions. It follows from (6.45) that marginal distributions of $\boldsymbol{X}$ must be elliptical distributions with the same characteristic generator. Using the $\boldsymbol{X} = (\boldsymbol{X}_1', \boldsymbol{X}_2')'$ notation from Section 6.1.3 and again extending this notation naturally to $\boldsymbol{\mu}$ and Σ,

$$\boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}, \qquad \Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix},$$

we have that $\boldsymbol{X}_1 \sim E_k(\boldsymbol{\mu}_1, \Sigma_{11}, \psi)$ and $\boldsymbol{X}_2 \sim E_{d-k}(\boldsymbol{\mu}_2, \Sigma_{22}, \psi)$.

Conditional distributions. The conditional distribution of $\boldsymbol{X}_2$ given $\boldsymbol{X}_1$ may also be shown to be elliptical, although in general it will have a *different* characteristic generator $\tilde{\psi}$. For details of how the generator changes see Fang, Kotz and Ng (1990, pp. 45, 46). In the special case of multivariate normality the generator remains the same.

Quadratic forms. If $\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ with Σ non-singular, then we observed in (6.42) that

$$Q := (\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \overset{\mathrm{d}}{=} R^2, \tag{6.46}$$

where R is the radial rv in the stochastic representation (6.40). As we have seen in Example 6.23, for some particular cases the distribution of R^2 is well known: if $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$, then $R^2 \sim \chi^2_d$; if $\boldsymbol{X} \sim t_d(\nu, \boldsymbol{\mu}, \Sigma)$, then $R^2/d \sim F(d, \nu)$. For all elliptical distributions, Q must be independent of $\Sigma^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})/\sqrt{Q}$.

Convolutions. The convolution of two independent elliptical vectors with the *same dispersion matrix* Σ is also elliptical. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent d-dimensional random vectors satisfying $\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ and $\boldsymbol{Y} \sim E_d(\tilde{\boldsymbol{\mu}}, \Sigma, \tilde{\psi})$, then we may take the product of characteristic functions to show that

$$\boldsymbol{X} + \boldsymbol{Y} \sim E_d(\boldsymbol{\mu} + \tilde{\boldsymbol{\mu}}, \Sigma, \bar{\psi}), \tag{6.47}$$

where $\bar{\psi}(u) = \psi(u)\tilde{\psi}(u)$.

If the dispersion matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$ differ by more than a constant factor, then the convolution will not necessarily remain elliptical, even when the two generators ψ and $\tilde{\psi}$ are identical.

6.3.4 Estimating Dispersion and Correlation

Suppose we have risk-factor return data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ that we believe come from some elliptical distribution $E_d(\boldsymbol{\mu}, \Sigma, \psi)$ with heavier tails than the multivariate normal. We recall from Remark 6.26 that the dispersion matrix Σ is not uniquely determined, but rather is only fixed up to a constant of proportionality; when covariances are finite, the covariance matrix is proportional to Σ.

In this section we briefly consider the problem of estimating the location parameter $\boldsymbol{\mu}$, a dispersion matrix Σ and the correlation matrix P, assuming finiteness of second moments. We could use the standard estimators of Section 6.1.2. Under an assumption of iid or uncorrelated vector observations we observed that $\bar{\boldsymbol{X}}$ and S in (6.9) are unbiased estimators of the mean vector and the covariance matrix, respectively. They will also be consistent under quite weak assumptions. However, this does not necessarily mean they are the best estimators of location and dispersion for any given finite sample of elliptical data. There are many alternative estimators that may be more efficient for heavy-tailed data and may enjoy better robustness properties for contaminated data.

One strategy would be to fit a number of normal variance mixture models, such as the t and NIG, using the approach of Section 6.2.4. From the best-fitting model we would obtain an estimate of the mean vector and could easily calculate the implied estimates of the covariance and correlation matrices. In this section we give simpler, alternative methods that do not require a full fitting of a multivariate distribution; consult Notes and Comments for further references to robust dispersion estimation.

M-estimators. Maronna's M-estimators (Maronna 1976) of location and dispersion are a relatively old idea in robust statistics, but they have the virtue of being particularly simple to implement. Let $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ denote estimates of the mean vector and the dispersion matrix. Suppose for every observation $\boldsymbol{X}_i$ we calculate $D_i^2 = (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})' \hat{\Sigma}^{-1} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})$. If we wanted to calculate improved estimates of location and dispersion, particularly for heavy-tailed data, it might be expected that this could be achieved by reducing the influence of observations for which D_i is large, since these are the observations that might tend to distort the parameter estimates most. M-estimation uses decreasing weight functions $w_j : \mathbb{R}^+ \to \mathbb{R}^+$, $j = 1, 2$, to reduce the weight of observations with large D_i values. This can be turned into an iterative procedure that converges to so-called M-estimates of location and dispersion; the dispersion matrix estimate is generally a biased estimate of the true covariance matrix.

Algorithm 6.29 (M-estimators of location and dispersion).

(1) As starting estimates take $\hat{\boldsymbol{\mu}}^{[1]} = \bar{\boldsymbol{X}}$ and $\hat{\Sigma}^{[1]} = S$, the standard estimators in (6.9). Set iteration count $k = 1$.

(2) For $i = 1, \ldots, n$ set $D_i^2 = (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^{[k]})' \hat{\Sigma}^{[k]\,-1} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^{[k]})$.

(3) Update the location estimate using

$$\hat{\boldsymbol{\mu}}^{[k+1]} = \frac{\sum_{i=1}^{n} w_1(D_i)\boldsymbol{X}_i}{\sum_{i=1}^{n} w_1(D_i)},$$

where w_1 is a weight function, as discussed below.

(4) Update the dispersion matrix estimate using

$$\hat{\Sigma}^{[k+1]} = \frac{1}{n}\sum_{i=1}^{n} w_2(D_i^2)(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^{[k]})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^{[k]})',$$

where w_2 is a weight function.

(5) Set $k = k + 1$ and repeat steps (2)–(4) until estimates converge.

Popular choices for the weight functions w_1 and w_2 are the decreasing functions $w_1(x) = (d + \nu)/(x^2 + \nu) = w_2(x^2)$ for some positive constant ν. Interestingly, use of these weight functions in Algorithm 6.29 exactly corresponds to fitting a multivariate $t_d(\nu, \boldsymbol{\mu}, \Sigma)$ distribution with known degrees of freedom ν using the EM algorithm (see, for example, Meng and van Dyk 1997).

There are many other possibilities for the weight functions. For example, the observations in the central part of the distribution could be given full weight and only the more outlying observations downweighted. This can be achieved by setting $w_1(x) = 1$ for $x \leqslant a$, $w_1(x) = a/x$ for $x > a$, for some value a, and $w_2(x^2) = (w_1(x))^2$.

Correlation estimates via Kendall's tau. A method for estimating correlation that is particularly easy to carry out is based on Kendall's rank correlation coefficient; this method will turn out to be related to a method in Chapter 7 that is used for estimating the parameters of certain copulas. The theoretical version of Kendall's rank correlation (also known as Kendall's tau) for two rvs X_1 and X_2 is denoted by $\rho_\tau(X_1, X_2)$ and is defined formally in Section 7.2.3; it is shown in Proposition 7.43 that if $(X_1, X_2) \sim E_2(\boldsymbol{\mu}, \Sigma, \psi)$, then

$$\rho_\tau(X_1, X_2) = \frac{2}{\pi}\arcsin(\rho), \tag{6.48}$$

where $\rho = \sigma_{12}/(\sigma_{11}\sigma_{22})^{1/2}$ is the *pseudo-correlation coefficient* of the elliptical distribution, which is always defined (even when correlation coefficients are undefined because variances are infinite). This relationship can be inverted to provide a method for estimating ρ from data; we simply replace the left-hand side of (6.48) by the standard textbook estimator of Kendall's tau, which is given in (7.52), to get an estimating equation that is solved for $\hat{\rho}$. This method estimates correlation by exploiting the geometry of an elliptical distribution and does not require us to estimate variances and covariances.

The method can be used to estimate a correlation matrix of a higher-dimensional elliptical distribution by applying the technique to each bivariate margin. This does, however, result in a matrix of pairwise correlation estimates that is not necessarily positive definite; this problem does not always arise, and if it does, a matrix

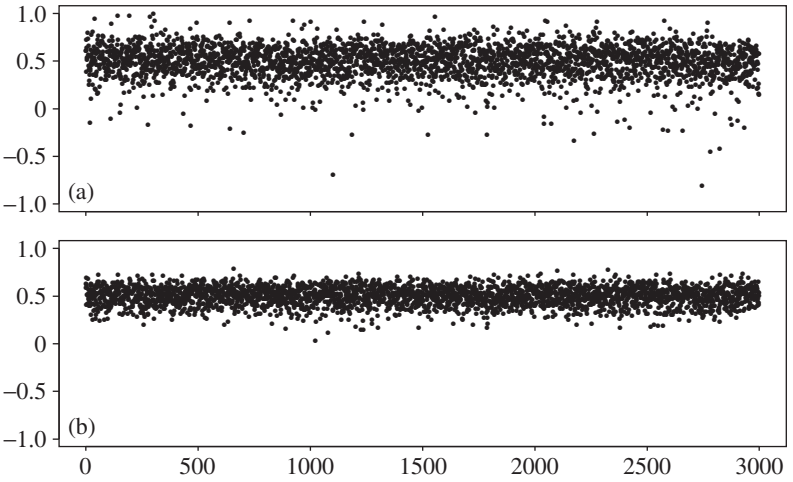

Figure 6.4. For 3000 independent samples of size 90 from a bivariate t distribution with three degrees of freedom and linear correlation 0.5: (a) the standard (Pearson) estimator of correlation; (b) the Kendall's tau transform estimator. See Example 6.30 for commentary.

adjustment method can be used, such as the eigenvalue method of Rousseeuw and Molenberghs (1993), which is given in Algorithm 7.57.

Note that to turn an estimate of a bivariate correlation matrix into a robust estimate of a dispersion matrix we could estimate the ratio of standard deviations $\lambda = (\sigma_{22}/\sigma_{11})^{1/2}$, e.g. by using a ratio of *trimmed* sample standard deviations; in other words, we leave out an equal number of outliers from each of the univariate data sets $X_{1,i}, \ldots, X_{n,i}$ for $i = 1, 2$ and calculate the sample standard deviations with the remaining observations. This would give us the estimate

$$\hat{\Sigma} = \begin{pmatrix} 1 & \hat{\lambda}\hat{\rho} \\ \hat{\lambda}\hat{\rho} & \hat{\lambda}^2 \end{pmatrix}. \tag{6.49}$$

Example 6.30 (efficient correlation estimation for heavy-tailed data). Suppose we calculate correlations of asset or risk-factor returns based on 90 days (somewhat more than three trading months) of data; it would seem that this ought to be enough data to allow us to accurately estimate the "true" underlying correlation under an assumption that we have identically distributed data for that period.

Figure 6.4 displays the results of a simulation experiment where we have generated 3000 bivariate samples of iid data from a t distribution with three degrees of freedom and correlation $\rho = 0.5$; this is a heavy-tailed elliptical distribution. The distribution of the values of the standard correlation coefficient (also known as the Pearson correlation coefficient) is not particularly closely concentrated around the true value and produces some very poor estimates for a number of samples. On the other hand, the Kendall's tau transform method produces estimates that are generally much closer to the true value, and thus provides a more efficient way of estimating ρ.

Notes and Comments

A comprehensive reference for spherical and elliptical distributions is Fang, Kotz and Ng (1990); we have based our brief presentation of the theory on this account. Other references for the theory are Kelker (1970), Cambanis, Huang and Simons (1981) and Bingham and Kiesel (2002), the latter in the context of financial modelling. The original reference for Theorem 6.21 is Schoenberg (1938). Frahm (2004) suggests a generalization of the elliptical class to allow asymmetric models while preserving many of the attractive properties of elliptical distributions. For a more historical discussion (going back to Archimedes) and some surprising properties of the uniform distribution on the unit d-sphere, see Letac (2004).

There is a vast literature on alternative estimators of dispersion and correlation matrices, particularly with regard to better robustness properties. Textbooks with relevant sections include Hampel et al. (1986), Marazzi (1993), Wilcox (1997) and Huber and Ronchetti (2009); the last of those books is recommended more generally for applications of robust statistics in econometrics and finance.

We have concentrated on M-estimation of dispersion matrices, since this is related to the maximum likelihood estimation of alternative elliptical models. M-estimators have a relatively long history and are known to have good local robustness properties (insensitivity to small data perturbations); they do, however, have relatively low breakdown points in high dimensions, so their performance can be poor when data are more contaminated. A small selection of papers on M-estimation is Maronna (1976), Devlin, Gnanadesikan and Kettenring (1975, 1981) and Tyler (1983, 1987); see also Frahm (2004), in which an interesting alternative derivation of a Tyler estimator is given. The method based on Kendall's tau was suggested in Lindskog, McNeil and Schmock (2003).

6.4 Dimension-Reduction Techniques

The techniques of dimension reduction, such as factor modelling and principal components, are central to multivariate statistical analysis and are widely used in econometric model building. In the high-dimensional world of financial risk management they are essential tools.

6.4.1 Factor Models

By using a factor model we attempt to explain the randomness in the components of a d-dimensional vector $\boldsymbol{X}$ in terms of a smaller set of *common factors*. If the components of $\boldsymbol{X}$ represent, for example, equity returns, it is clear that a large part of their variation can be explained in terms of the variation of a smaller set of market index returns. Formally, we define a factor model as follows.

Definition 6.31 (linear factor model). The random vector $\boldsymbol{X}$ is said to follow a p-factor model if it can be decomposed as

$$\boldsymbol{X} = \boldsymbol{a} + B\boldsymbol{F} + \boldsymbol{\varepsilon}, \tag{6.50}$$

where

(i) $\boldsymbol{F} = (F_1, \ldots, F_p)'$ is a random vector of *common factors* with $p < d$ and a covariance matrix that is positive definite,

(ii) $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_d)'$ is a random vector of *idiosyncratic error terms*, which are uncorrelated and have mean 0,

(iii) $B \in \mathbb{R}^{d\times p}$ is a matrix of constant *factor loadings* and $\boldsymbol{a} \in \mathbb{R}^d$ is a vector of constants, and

(iv) $\operatorname{cov}(\boldsymbol{F}, \boldsymbol{\varepsilon}) = E((\boldsymbol{F} - E(\boldsymbol{F}))\boldsymbol{\varepsilon}') = 0$.

The assumptions that the errors are uncorrelated with each other (ii) and also with the common factors (iv) are important parts of this definition. We do not in general require independence, only uncorrelatedness. However, if the vector $\boldsymbol{X}$ is multivariate normally distributed and follows the factor model in (6.50), then it is possible to find a version of the factor model where $\boldsymbol{F}$ and $\boldsymbol{\varepsilon}$ are Gaussian and the errors can be assumed to be mutually independent and independent of the common factors. We elaborate on this assertion in Example 6.32 below.

It follows from the basic assumptions that factor models imply a special structure for the covariance matrix Σ of $\boldsymbol{X}$. If we denote the covariance matrix of $\boldsymbol{F}$ by Ω and that of $\boldsymbol{\varepsilon}$ by the diagonal matrix Υ, it follows that

$$\Sigma = \operatorname{cov}(\boldsymbol{X}) = B\Omega B' + \Upsilon. \tag{6.51}$$

If the factor model holds, the common factors can always be transformed so that they have mean 0 and are orthogonal. By setting $\boldsymbol{F}^* = \Omega^{-1/2}(\boldsymbol{F} - E(\boldsymbol{F}))$ and $B^* = B\Omega^{1/2}$, we have a representation of the factor model of the form $\boldsymbol{X} = \boldsymbol{\mu} + B^*\boldsymbol{F}^* + \boldsymbol{\varepsilon}$, where $\boldsymbol{\mu} = E(\boldsymbol{X})$, as usual, and $\Sigma = B^*(B^*)' + \Upsilon$.

Conversely, it can be shown that whenever a random vector $\boldsymbol{X}$ has a covariance matrix that satisfies

$$\Sigma = BB' + \Upsilon \tag{6.52}$$

for some $B \in \mathbb{R}^{d\times p}$ with $\operatorname{rank}(B) = p < d$ and diagonal matrix Υ, then $\boldsymbol{X}$ has a factor-model representation for some p-dimensional factor vector $\boldsymbol{F}$ and d-dimensional error vector $\boldsymbol{\varepsilon}$.

Example 6.32 (equicorrelation model). Suppose $\boldsymbol{X}$ is a random vector with standardized margins (zero mean and unit variance) and an *equicorrelation matrix*; in other words, the correlation between each pair of components is equal to $\rho > 0$. This means that the covariance matrix Σ can be written as $\Sigma = \rho J_d + (1-\rho)I_d$, where J_d is the d-dimensional square matrix of ones and I_d is the identity matrix, so that Σ is obviously of the form (6.52) for the d-vector $B = \sqrt{\rho}\mathbf{1}$.

To find a factor decomposition of $\boldsymbol{X}$, take *any* zero-mean, unit-variance rv Y that is *independent* of $\boldsymbol{X}$ and define a single common factor F and errors $\boldsymbol{\varepsilon}$ by

$$F = \frac{\sqrt{\rho}}{1+\rho(d-1)}\sum_{j=1}^{d} X_j + \sqrt{\frac{1-\rho}{1+\rho(d-1)}}Y, \qquad \varepsilon_j = X_j - \sqrt{\rho}F,$$

where we note that in this construction F also has mean 0 and variance 1. We therefore have the factor decomposition $\boldsymbol{X} = B\boldsymbol{F} + \boldsymbol{\varepsilon}$, and it may be verified by calculation that $\text{cov}(F, \varepsilon_j) = 0$ for all j and $\text{cov}(\varepsilon_j, \varepsilon_k) = 0$ when $j \neq k$, so that the requirements of Definition 6.31 are satisfied. A random vector with an equicorrelation matrix can be thought of as following a factor model with a single common factor.

Since we can take any Y, the factors and errors in this decomposition are non-unique. Consider the case where the vector $\boldsymbol{X}$ is Gaussian; it is most convenient to take Y to also be Gaussian, since in that case the common factor is normally distributed, the error vector is multivariate normally distributed, Y is independent of ε_j, for all j, and ε_j and ε_k are independent for $j \neq k$. Since $\text{var}(\varepsilon_j) = 1 - \rho$, it is most convenient to write the factor model implied by the equicorrelation model as

$$X_j = \sqrt{\rho}F + \sqrt{1-\rho}Z_j, \quad j = 1, \ldots, d, \tag{6.53}$$

where $F, Z_1, \ldots, Z_d$ are mutually independent standard Gaussian rvs. This model will be used in Section 11.1.5 in the context of modelling homogeneous credit portfolios. For the more general construction on which this example is based, see Mardia, Kent and Bibby (1979, Exercise 9.2.2).

6.4.2 Statistical Estimation Strategies

Now assume that we have data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ representing risk-factor changes at times $t = 1, \ldots, n$. Each vector observation $\boldsymbol{X}_t$ is assumed to be a realization from a factor model of the form (6.50) so that we have

$$\boldsymbol{X}_t = \boldsymbol{a} + B\boldsymbol{F}_t + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, n, \tag{6.54}$$

for common-factor vectors $\boldsymbol{F}_t = (F_{t,1}, \ldots, F_{t,p})'$, error vectors $\boldsymbol{\varepsilon}_t$, a vector of constants $\boldsymbol{a} \in \mathbb{R}^d$, and loading matrix $B \in \mathbb{R}^{d \times p}$. There are occasionally situations where we might wish to model $\boldsymbol{a}$ and B as time dependent, but mostly they are assumed to be fixed over time.

The model (6.54) is clearly an idealization. Data will seldom be perfectly explained by a factor model; the aim is to find an approximating factor model that captures the main sources of variability in the data. Three general types of factor model are commonly used in financial risk applications; they are known as *macroeconomic*, *fundamental* and *statistical* factor models.

Macroeconomic factor models. In these models we assume that appropriate factors $\boldsymbol{F}_t$ are also observable and we collect time-series data $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_n \in \mathbb{R}^p$. The name comes from the fact that, in many applications of these models in economics and finance, the observed factors are macroeconomic variables, such as changes in GDP, inflation and interest rates.

A simple example of a macroeconomic model in finance is Sharpe's single-index model, where $F_1, \ldots, F_n$ are observations of the return on a market index and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are individual equity returns that are explained in terms of the market return. Fitting of the model (estimation of B and $\boldsymbol{a}$) is accomplished by time-series regression techniques; it is described in Section 6.4.3.

Fundamental factor models. In contrast to the macroeconomic factor models, here we assume that the loading matrix B is known but that the underlying factors $\boldsymbol{F}_t$ are unobserved. Factor values $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_n$ have to be estimated from the data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ using cross-sectional regression at each time point.

The name comes from applications in modelling equity returns where the stocks are classified according to their "fundamentals", such as country, industry sector and size (small cap, large cap, etc.). These are generally categorical variables and it is assumed that there are underlying, unobserved factors associated with each level of the categorical variable, e.g. a factor for each country or each industry sector.

If each risk-factor change $X_{t,i}$ can be identified with a unique set of values for the fundamentals, e.g. a unique country or industry, then the matrix B is a matrix consisting of zeros and ones. If $X_{t,i}$ is attributed to different values of the fundamental variable, then B might contain factor weights summing to 1; for example, 60% of a stock return for a multinational company might be attributed to an unobserved US factor and 40% to an unobserved UK factor. There may also be situations in fundamental factor modelling where time-dependent loading matrices B_t are used.

Statistical factor models. In these models we observe neither the factors $\boldsymbol{F}_t$ nor the loadings B. Instead, we use statistical techniques to estimate both from the data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. This can be a very powerful approach to explaining the variability in data, but we note that the factors we obtain, while being explanatory in a statistical sense, may not have any obvious interpretation.

There are two general methods for finding factors. The first method, which is quite common in finance, is to use *principal component analysis* to construct factors; we discuss this technique in detail in Section 6.4.5. The second method, *classical statistical factor analysis*, is less commonly used in finance (see Notes and Comments).

Factor models and systematic risk. In the context of risk management, the goal of all approaches to factor modelling is either to identify or to estimate appropriate factor data $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_n$. If this is achieved, we can then concentrate on modelling the distribution or dynamics of the factors, which is a lower-dimensional problem than modelling $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

The factors describe the *systematic risk* and are of primary importance. The unobserved errors $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ describe the *idiosyncratic risk* and are of secondary importance. In situations where we have many risk factors, the risk embodied in the errors is partly mitigated by a diversification effect, whereas the risk embodied in the common factors remains. The following simple example gives an idea why this is the case.

Example 6.33. We continue our analysis of the one-factor model in Example 6.32. Suppose that the random vector $\boldsymbol{X}$ in that example represents the return on d different companies so that the rv $Z_{(d)} = (1/d)\sum_{j=1}^{d} X_j$ can be thought of as the portfolio return for an equal investment in each of the companies. We calculate that

$$Z_{(d)} = \frac{1}{d}\mathbf{1}'BF + \frac{1}{d}\mathbf{1}'\boldsymbol{\varepsilon} = \sqrt{\rho}F + \frac{1}{d}\sum_{j=1}^{d}\varepsilon_j.$$

The risk in the first term is not affected by increasing the size of the portfolio d, whereas the risk in the second term can be reduced. Suppose we measure risk by simply calculating variances; we get

$$\operatorname{var}(Z_{(d)}) = \rho + \frac{1-\rho}{d} \to \rho, \quad d \to \infty,$$

showing that the systematic factor is the main contributor to the risk in a large-portfolio situation.

6.4.3 Estimating Macroeconomic Factor Models

Two equivalent approaches may be used to estimate the model parameters in a macroeconomic factor model of the form (6.54). In the first approach we perform d univariate regression analyses, one for each component of the individual return series. In the second approach we estimate all parameters in a single multivariate regression.

Univariate regression. Writing $X_{t,j}$ for the observation at time t of instrument j, we consider the univariate regression model

$$X_{t,j} = a_j + \boldsymbol{b}_j' \boldsymbol{F}_t + \varepsilon_{t,j}, \quad t = 1, \ldots, n.$$

This is known as a time-series regression, since the responses $X_{1,j}, \ldots, X_{n,j}$ form a univariate time series and the factors $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_n$ form a possibly multivariate time series. Without going into technical details we simply remark that the parameters a_j and $\boldsymbol{b}_j$ are estimated using the standard ordinary least-squares (OLS) method found in all textbooks on linear regression. To justify the use of the method and to derive statistical properties of the method it is usually assumed that, conditional on the factors, the errors $\varepsilon_{1,j}, \ldots, \varepsilon_{n,j}$ are identically distributed and serially uncorrelated. In other words, they form a white noise process as defined in Chapter 4.

The estimate $\hat{a}_j$ obviously estimates the jth component of $\boldsymbol{a}$, while $\hat{\boldsymbol{b}}_j$ is an estimate of the jth row of the matrix B. By performing a regression for each of the univariate time series $X_{1,j}, \ldots, X_{n,j}$ for $j = 1, \ldots, d$, we complete the estimation of the parameters $\boldsymbol{a}$ and B.

Multivariate regression. To set the problem up as a multivariate linear-regression problem, we construct a number of large matrices:

$$X = \underbrace{\begin{pmatrix} \boldsymbol{X}_1' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix}}_{n \times d}, \qquad F = \underbrace{\begin{pmatrix} 1 & \boldsymbol{F}_1' \\ \vdots & \vdots \\ 1 & \boldsymbol{F}_n' \end{pmatrix}}_{n \times (p+1)}, \qquad B_2 = \underbrace{\begin{pmatrix} \boldsymbol{a}' \\ B' \end{pmatrix}}_{(p+1) \times d}, \qquad E = \underbrace{\begin{pmatrix} \boldsymbol{\varepsilon}_1' \\ \vdots \\ \boldsymbol{\varepsilon}_n' \end{pmatrix}}_{n \times d}.$$

Each row of the data X corresponds to a vector observation at a fixed time point t, and each column corresponds to a univariate time series for one of the individual returns. The model (6.54) can then be expressed by the matrix equation

$$X = FB_2 + E, \tag{6.55}$$

where B_2 is the matrix of regression parameters to be estimated.

If we assume that the unobserved error vectors $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ comprising the rows of E are identically distributed and serially uncorrelated, conditional on $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_n$, then the equation (6.55) defines a standard multivariate linear regression (see, for example, Mardia, Kent and Bibby (1979) for the standard assumptions). An estimate of B_2 is obtained by multivariate OLS according to the formula

$$\hat{B}_2 = (F'F)^{-1}F'X. \tag{6.56}$$

The factor model is now essentially calibrated, since we have estimates for $\boldsymbol{a}$ and B. The model can now be critically examined with respect to the original conditions of Definition 6.31. Do the error vectors $\boldsymbol{\varepsilon}_t$ come from a distribution with diagonal covariance matrix, and are they uncorrelated with the factors?

To learn something about the errors we can form the model residual matrix $\hat{E} = X - F\hat{B}_2$. Each row of this matrix contains an inferred value of an error vector $\hat{\boldsymbol{\varepsilon}}_t$ at a fixed point in time. Examination of the sample correlation matrix of these inferred error vectors will hopefully show that there is little remaining correlation in the errors (or at least much less than in the original data vectors $\boldsymbol{X}_t$). If this is the case, then the diagonal elements of the sample covariance matrix of the $\hat{\boldsymbol{\varepsilon}}_t$ could be taken as an estimator $\hat{\Upsilon}$ for Υ. It is sometimes of interest to form the covariance matrix implied by the factor model and compare this with the original sample covariance matrix S of the data. The implied covariance matrix is

$$\hat{\Sigma}^{(\mathrm{F})} = \hat{B}\hat{\Omega}\hat{B}' + \hat{\Upsilon}, \quad \text{where } \hat{\Omega} = \frac{1}{n-1}\sum_{t=1}^{n}(\boldsymbol{F}_t - \bar{\boldsymbol{F}})(\boldsymbol{F}_t - \bar{\boldsymbol{F}})'.$$

We would hope that $\hat{\Sigma}^{(\mathrm{F})}$ captures much of the structure of S and that the correlation matrix $R^{(\mathrm{F})} := \wp(\hat{\Sigma}^{(\mathrm{F})})$ captures much of the structure of the sample correlation matrix $R = \wp(S)$.

Example 6.34 (single-index model for Dow Jones 30 returns). As a simple example of the regression approach to fitting factor models we have fitted a single factor model to a set of ten Dow Jones 30 daily stock-return series from 1992 to 1998. Note that these are different returns to those analysed in previous sections of this chapter. They have been chosen to be of two types: technology-related titles such as Hewlett-Packard, Intel, Microsoft and IBM; and food- and consumer-related titles such as Philip Morris, Coca-Cola, Eastman Kodak, McDonald's, Wal-Mart and Disney. The factor chosen is the corresponding return on the Dow Jones 30 index itself.

The estimate of B implied by formula (6.56) is shown in the first line of Table 6.5. The highest values of B correspond to so-called *high-beta* stocks; since a one-factor model implies the relationship $E(X_j) = a_j + B_j E(F)$, these stocks potentially offer high expected returns relative to the market (but are often riskier titles); in this case, the four technology-related stocks have the highest beta values. In the second row, values of r^2, the so-called coefficient of determination, are given for each of the univariate regression models. This number measures the strength of the regression relationship between X_j and F and can be interpreted as the proportion of the variation of the stock return that is explained by variation in the market return;

Table 6.5. The first line gives estimates of B for a multivariate regression model fitted to ten Dow Jones 30 stocks where the observed common factor is the return on the Dow Jones 30 index itself. The second row gives r^2 values for a univariate regression model for each individual time series. The next ten lines of the table give the sample correlation matrix of the data R, while the middle ten lines give the correlation matrix implied by the factor model. The final ten lines show the estimated correlation matrix of the residuals from the regression model, with entries less than 0.1 in absolute value being omitted. See Example 6.34 for full details.

	MO	KO	EK	HWP	INTC	MSFT	IBM	MCD	WMT	DIS
$\hat{B}$	0.87	1.01	0.77	1.12	1.12	1.11	1.07	0.86	1.02	1.03
r^2	0.17	0.33	0.14	0.18	0.17	0.21	0.22	0.23	0.24	0.26
MO	1.00	0.27	0.14	0.17	0.16	0.25	0.18	0.22	0.16	0.22
KO	0.27	1.00	0.17	0.22	0.21	0.25	0.18	0.36	0.33	0.32
EK	0.14	0.17	1.00	0.17	0.17	0.18	0.15	0.14	0.17	0.16
HWP	0.17	0.22	0.17	1.00	0.42	0.38	0.36	0.20	0.22	0.23
INTC	0.16	0.21	0.17	0.42	1.00	0.53	0.36	0.19	0.22	0.21
MSFT	0.25	0.25	0.18	0.38	0.53	1.00	0.33	0.22	0.28	0.26
IBM	0.18	0.18	0.15	0.36	0.36	0.33	1.00	0.20	0.20	0.20
MCD	0.22	0.36	0.14	0.20	0.19	0.22	0.20	1.00	0.26	0.26
WMT	0.16	0.33	0.17	0.22	0.22	0.28	0.20	0.26	1.00	0.28
DIS	0.22	0.32	0.16	0.23	0.21	0.26	0.20	0.26	0.28	1.00
MO	1.00	0.24	0.16	0.18	0.17	0.19	0.20	0.20	0.20	0.21
KO	0.24	1.00	0.22	0.24	0.23	0.26	0.27	0.28	0.28	0.29
EK	0.16	0.22	1.00	0.16	0.15	0.17	0.18	0.18	0.18	0.19
HWP	0.18	0.24	0.16	1.00	0.17	0.19	0.20	0.20	0.21	0.22
INTC	0.17	0.23	0.15	0.17	1.00	0.19	0.19	0.19	0.20	0.21
MSFT	0.19	0.26	0.17	0.19	0.19	1.00	0.22	0.22	0.22	0.23
IBM	0.20	0.27	0.18	0.20	0.19	0.22	1.00	0.23	0.23	0.24
MCD	0.20	0.28	0.18	0.20	0.19	0.22	0.23	1.00	0.23	0.24
WMT	0.20	0.28	0.18	0.21	0.20	0.22	0.23	0.23	1.00	0.25
DIS	0.21	0.29	0.19	0.22	0.21	0.23	0.24	0.24	0.25	1.00
MO	1.00									
KO		1.00					−0.12	0.12		
EK			1.00							
HWP				1.00	0.30	0.24	0.20			
INTC				0.30	1.00	0.43	0.20			
MSFT				0.24	0.43	1.00	0.14			
IBM		−0.12		0.20	0.20	0.14	1.00			
MCD		0.12						1.00		
WMT										
DIS										1.00

the highest r^2 corresponds to Coca-Cola (33%), and in general it seems that about 20% of individual stock-return variation is explained by market-return variation.

The next ten lines of the table give the sample correlation matrix of the data R, while the middle ten lines give the correlation matrix implied by the factor model

(corresponding to $\hat{\Sigma}^{(\mathrm{F})}$). The latter matrix picks up much, but not all, of the structure of the former matrix. The final ten lines show the estimated correlation matrix of the residuals from the regression model, but only those elements that exceed 0.1 in absolute value. The residuals are indeed much less correlated than the original data, but a few larger entries indicate imperfections in the factor-model representation of the data, particularly for the technology stocks. The index return for the broader market is clearly an important common factor, but further systematic effects that are not captured by the index appear to be present in these data.

6.4.4 Estimating Fundamental Factor Models

To estimate a fundamental factor model we consider, at each time point t, a cross-sectional regression model of the form

$$\boldsymbol{X}_t = B\boldsymbol{F}_t + \boldsymbol{\varepsilon}_t, \tag{6.57}$$

where $\boldsymbol{X}_t \in \mathbb{R}^d$ are the risk-factor change data, $B \in \mathbb{R}^{d\times p}$ is a known matrix of factor loadings (which may be time dependent in some applications), $\boldsymbol{F}_t \in \mathbb{R}^p$ are the factors to be estimated, and $\boldsymbol{\varepsilon}_t$ are errors with diagonal covariance matrix Υ. There is no need for an intercept $\boldsymbol{a}$ in the estimation of a fundamental factor model, as this can be absorbed into the factor estimates.

To obtain precision in the estimation of $\boldsymbol{F}_t$, the dimension d of the risk-factor vector needs to be large with respect to the number of factors p to be estimated. Note also that the components of the error vector $\boldsymbol{\varepsilon}_t$ cannot generally be assumed to have equal variance, so (6.57) is a regression problem with so-called heteroscedastic errors.

We recall that, in typical applications in equity return modelling, the factors are frequently identified with country, industry-sector and company-size effects. The rows of the matrix B can consist of zeros and ones, if $X_{t,i}$ is associated with a single country or industry sector, or weights, if $X_{t,i}$ is attributed to more than one country or industry sector. This kind of interpretation for the factors is also quite common in the factor models used for modelling portfolio credit risk, as we discuss in Section 11.5.1.

Unbiased estimators of the factors $\boldsymbol{F}_t$ may be obtained by forming the OLS estimates

$$\hat{\boldsymbol{F}}_t^{\mathrm{OLS}} = (B'B)^{-1}B'\boldsymbol{X}_t,$$

and these are the best linear unbiased estimates in the case where the errors are homoscedastic, so that $\Upsilon = \upsilon^2 I_d$ for some scalar υ. However, in general, the OLS estimates are not efficient and it is possible to obtain linear unbiased estimates with a smaller covariance matrix using the method of generalized least squares (GLS). If Υ were a known matrix, the GLS estimates would be given by

$$\hat{\boldsymbol{F}}_t^{\mathrm{GLS}} = (B'\Upsilon^{-1}B)^{-1}B'\Upsilon^{-1}\boldsymbol{X}_t. \tag{6.58}$$

In practice, we replace Υ in (6.58) with an estimate $\hat{\Upsilon}$ obtained as follows. Under an assumption that the model (6.57) holds at every time point $t = 1, \ldots, n$, we first

carry out OLS estimation at each time t and form the model residual vectors

$$\hat{\boldsymbol{\varepsilon}}_t = \boldsymbol{X}_t - B\hat{\boldsymbol{F}}_t^{\mathrm{OLS}}, \quad t = 1, \ldots, n.$$

We then form the sample covariance matrix of the residuals $\hat{\boldsymbol{\varepsilon}}_1, \ldots, \hat{\boldsymbol{\varepsilon}}_n$. This matrix should be approximately diagonal, if the factor model assumption holds. We can set off-diagonal elements equal to zero to form an estimate of Υ.

We give an example of the estimation of a fundamental factor model in the context of modelling the yield curve in Section 9.1.4.

6.4.5 Principal Component Analysis

The aim of principal component analysis (PCA) is to reduce the dimensionality of highly correlated data by finding a small number of uncorrelated linear combinations that account for most of the variance of the original data. PCdimensional reductionA is not itself a model, but rather a data-rotation technique. However, it can be used as a way of constructing factors for use in factor modelling, and this is the main application we consider in this section.

The key mathematical result behind the technique is the *spectral decomposition theorem* of linear algebra, which says that any symmetric matrix $A \in \mathbb{R}^{d\times d}$ can be written as

$$A = \Gamma \Lambda \Gamma', \tag{6.59}$$

where

(i) $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ is the diagonal matrix of *eigenvalues of A* that, without loss of generality, are ordered so that $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_d$, and

(ii) Γ is an orthogonal matrix satisfying $\Gamma\Gamma' = \Gamma'\Gamma = I_d$ whose columns are standardized *eigenvectors of A* (i.e. eigenvectors with length 1).

Theoretical principal components. Obviously we can apply this decomposition to any covariance matrix Σ, and in this case the positive semidefiniteness of Σ ensures that $\lambda_j \geqslant 0$ for all j. Suppose the random vector $\boldsymbol{X}$ has mean vector $\boldsymbol{\mu}$ and covariance matrix Σ and we make the decomposition $\Sigma = \Gamma\Lambda\Gamma'$ as in (6.59). The principal components transform of $\boldsymbol{X}$ is then defined to be

$$\boldsymbol{Y} = \Gamma'(\boldsymbol{X} - \boldsymbol{\mu}), \tag{6.60}$$

and it can be thought of as a rotation and a recentring of $\boldsymbol{X}$. The jth component of the rotated vector $\boldsymbol{Y}$ is known as the *jth principal component* of $\boldsymbol{X}$ and is given by

$$Y_j = \boldsymbol{\gamma}_j'(\boldsymbol{X} - \boldsymbol{\mu}), \tag{6.61}$$

where $\boldsymbol{\gamma}_j$ is the eigenvector of Σ corresponding to the jth ordered eigenvalue; this vector is also known as the jth vector of *loadings*.

Simple calculations show that

$$E(\boldsymbol{Y}) = \boldsymbol{0} \quad \text{and} \quad \mathrm{cov}(\boldsymbol{Y}) = \Gamma'\Sigma\Gamma = \Gamma'\Gamma\Lambda\Gamma'\Gamma = \Lambda,$$

so that the principal components of $\boldsymbol{Y}$ are uncorrelated and have variances $\mathrm{var}(Y_j) = \lambda_j$, $\forall j$. The components are thus ordered by variance, from largest to

smallest. Moreover, the first principal component can be shown to be the standardized linear combination of $\boldsymbol{X}$ that has maximal variance among all such combinations; in other words,

$$\operatorname{var}(\boldsymbol{\gamma}_1'\boldsymbol{X}) = \max\{\operatorname{var}(\boldsymbol{a}'\boldsymbol{X}) : \boldsymbol{a}'\boldsymbol{a} = 1\}.$$

For $j = 2, \ldots, d$, the jth principal component can be shown to be the standardized linear combination of $\boldsymbol{X}$ with maximal variance among all such linear combinations that are *orthogonal* to (and hence uncorrelated with) the first $j - 1$ linear combinations. The final dth principal component has minimum variance among standardized linear combinations of $\boldsymbol{X}$.

To measure the ability of the first few principal components to explain the variance of $\boldsymbol{X}$, we observe that

$$\sum_{j=1}^{d} \operatorname{var}(Y_j) = \sum_{j=1}^{d} \lambda_j = \operatorname{trace}(\Sigma) = \sum_{j=1}^{d} \operatorname{var}(X_j).$$

If we interpret $\operatorname{trace}(\Sigma) = \sum_{j=1}^{d} \operatorname{var}(X_j)$ as a measure of the total variance of $\boldsymbol{X}$, then, for $k \leqslant d$, the ratio $\sum_{j=1}^{k} \lambda_j / \sum_{j=1}^{d} \lambda_j$ represents the amount of this variance that is explained by the first k principal components.

Principal components as factors. We note that, by inverting the principal components transform (6.60), we obtain

$$\boldsymbol{X} = \boldsymbol{\mu} + \Gamma\boldsymbol{Y} = \boldsymbol{\mu} + \Gamma_1\boldsymbol{Y}_1 + \Gamma_2\boldsymbol{Y}_2,$$

where we have partitioned $\boldsymbol{Y}$ into vectors $\boldsymbol{Y}_1 \in \mathbb{R}^k$ and $\boldsymbol{Y}_2 \in \mathbb{R}^{d-k}$, such that $\boldsymbol{Y}_1$ contains the first k principal components, and we have partitioned Γ into matrices $\Gamma_1 \in \mathbb{R}^{d\times k}$ and $\Gamma_2 \in \mathbb{R}^{d\times(d-k)}$ correspondingly. Let us assume that the first k principal components explain a large part of the total variance and we decide to focus our attention on them and ignore the further principal components in $\boldsymbol{Y}_2$. If we set $\boldsymbol{\varepsilon} = \Gamma_2\boldsymbol{Y}_2$, we obtain

$$\boldsymbol{X} = \boldsymbol{\mu} + \Gamma_1\boldsymbol{Y}_1 + \boldsymbol{\varepsilon}, \tag{6.62}$$

which is reminiscent of the basic factor model (6.50) with the vector $\boldsymbol{Y}_1$ playing the role of the factors and the matrix Γ_1 playing the role of the factor loading matrix. Although the components of the error vector $\boldsymbol{\varepsilon}$ will tend to have small variances, the assumptions of the factor model are generally violated in (6.62) since $\boldsymbol{\varepsilon}$ need not have a diagonal covariance matrix and need not be uncorrelated with $\boldsymbol{Y}_1$. Nevertheless, principal components are often interpreted as factors and used to develop approximate factor models. We now describe the estimation process that is followed when data are available.

Sample principal components. Assume that we have a time series of multivariate data observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ with identical distribution, unknown mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, with the spectral decomposition $\Sigma = \Gamma\Lambda\Gamma'$ as before.

To construct sample principal components we need to estimate the unknown parameters. We estimate $\boldsymbol{\mu}$ by $\bar{\boldsymbol{X}}$, the sample mean vector, and we estimate Σ by the sample covariance matrix

$$S_x = \frac{1}{n}\sum_{t=1}^{n}(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(\boldsymbol{X}_t - \bar{\boldsymbol{X}})'.$$

We apply the spectral decomposition (6.59) to the symmetric, positive-semidefinite matrix S_x to get

$$S_x = GLG', \tag{6.63}$$

where G is the eigenvector matrix, $L = \text{diag}(l_1, \ldots, l_d)$ is the diagonal matrix consisting of ordered eigenvalues, and we switch to roman letters to emphasize that these are now calculated from an empirical covariance matrix. The matrix G provides an estimate of Γ and L provides an estimate of Λ.

By analogy with (6.60) we define vectors of sample principal components

$$\boldsymbol{Y}_t = G'(\boldsymbol{X}_t - \bar{\boldsymbol{X}}), \quad t = 1, \ldots, n. \tag{6.64}$$

The jth component of $\boldsymbol{Y}_t$ is known as the jth sample principal component at time t and is given by

$$Y_{t,j} = \boldsymbol{g}_j'(\boldsymbol{X}_t - \bar{\boldsymbol{X}}),$$

where $\boldsymbol{g}_j$ is the jth column of G, that is, the eigenvector of S_x corresponding to the jth largest eigenvalue.

The rotated vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ have the property that their sample covariance matrix is L, as is easily verified:

$$\begin{aligned} S_y &= \frac{1}{n}\sum_{t=1}^{n}(\boldsymbol{Y}_t - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_t - \bar{\boldsymbol{Y}})' = \frac{1}{n}\sum_{t=1}^{n}\boldsymbol{Y}_t\boldsymbol{Y}_t' \\ &= \frac{1}{n}\sum_{t=1}^{n}G'(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(\boldsymbol{X}_t - \bar{\boldsymbol{X}})'G = G'S_xG = L. \end{aligned}$$

The rotated vectors therefore show no correlation between components, and the components are ordered by their sample variances, from largest to smallest.

Remark 6.35. In a situation where the different components of the data vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ have very different sample variances (particularly if they are measured on very different scales), it is to be expected that the component (or components) with largest variance will dominate the first loading vector $\boldsymbol{g}_1$ and dominate the first principal component. In these situations the data are often transformed to have identical variances, which effectively means that principal component analysis is applied to the sample correlation matrix R_x. Note also that we could derive sample principal components from a robust estimate of the correlation matrix or a multivariate dispersion matrix.

We can now use the sample eigenvector matrix G and the sample principal components $\boldsymbol{Y}_t$ to calibrate an approximate factor model of the form (6.62). We assume that our data are realizations from the model

$$\boldsymbol{X}_t = \bar{\boldsymbol{X}} + G_1 \boldsymbol{F}_t + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, n, \tag{6.65}$$

where G_1 consists of the first k columns of G and $\boldsymbol{F}_t = (Y_{t,1}, \ldots, Y_{t,k})'$, $t = 1, \ldots, n$. The choice of k is based on a subjective choice of the number of sample principal components that are required to explain a substantial part of the total sample variance (see Example 6.36).

Equation (6.65) bears a resemblance to the factor model (6.54) except that, in practice, the errors $\boldsymbol{\varepsilon}_t$ do not generally have a diagonal covariance matrix and are not generally uncorrelated with $\boldsymbol{F}_t$. Nevertheless, the method is a popular approach to constructing time series of statistically explanatory factors from multivariate time series of risk-factor changes.

Example 6.36 (PCA-based factor model for Dow Jones 30 returns). We consider the data in Example 6.34 again. Principal component analysis is applied to the sample covariance matrix of the return data and the results are summarized in Figures 6.5 and 6.6. In the former we see a bar plot of the sample variances of the first eight principal components l_j; the cumulative proportion of the total variance explained by the components is given above each bar; the first two components explain almost 50% of the variation. In the latter figure the first two loading vectors $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ are summarized.

The first vector of loadings is positively weighted for all stocks and can be thought of as describing a kind of index portfolio; of course, the weights in the loading vector do not sum to 1, but they can be scaled to do so and this gives a so-called principal-component-mimicking portfolio. The second vector has positive weights for the consumer titles and negative weights for the technology titles; as a portfolio it can be thought of as prescribing a programme of short selling of technology to buy consumer titles. These first two sample principal components loading vectors are used to define factors.

In Table 6.6 the transpose of the matrix $\hat{B}$ (containing the loadings estimates in the factor model) is shown; the rows are merely the first two loading vectors from the principal component analysis. In the third row, values of r^2, the so-called coefficient of determination, are given for each of the univariate regression models, and these indicate that more of the variation in the data is explained by the two PCA-based factors than was explained by the observed factor in Example 6.34; it seems that the model is best able to explain Intel returns.

The next ten lines give the correlation matrix implied by the factor model (corresponding to $\hat{\Sigma}^{(\mathrm{F})}$). Compared with the true sample correlation matrix in Example 6.34 this seems to pick up more of the structure than did the correlation matrix implied by the observed factor model.

The final ten lines show the estimated correlation matrix of the residuals from the regression model, but only those elements that exceed 0.1 in absolute value. The residuals are again less correlated than the original data, but there are quite a number

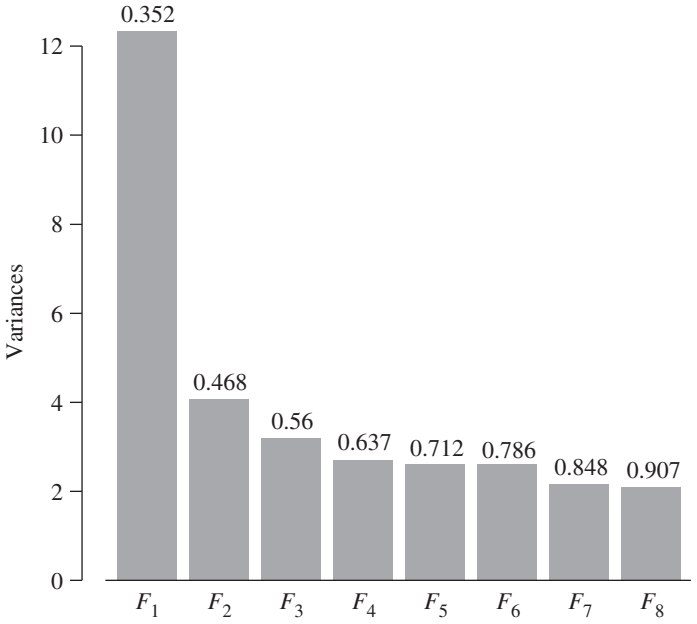

Figure 6.5. Bar plot of the sample variances l_j of the first eight principal components; the cumulative proportion of the total variance explained by the components is given above each bar ($\sum_{j=1}^{k} l_j / \sum_{j=1}^{10} l_j, k = 1, \ldots, 8$).

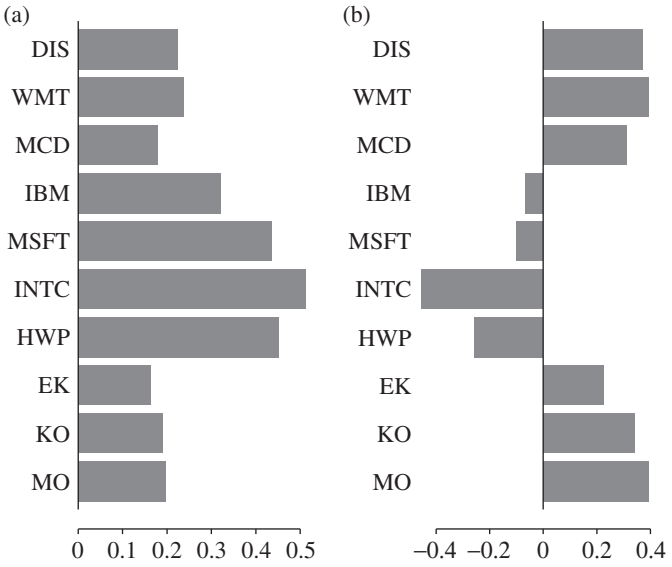

Figure 6.6. Bar plot summarizing the loading vectors $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ defining the first two principal components: (a) factor 1 loadings; (b) factor 2 loadings.

Table 6.6. The first two lines give estimates of the transpose of B for a factor model fitted to ten Dow Jones 30 stocks, where the factors are constructed from the first two sample principal components. The third row gives r^2 values for the univariate regression model for each individual time series. The next ten lines give the correlation matrix implied by the factor model. The final ten lines show the estimated correlation matrix of the residuals from the regression model, with entries less than 0.1 in absolute value omitted. See Example 6.36 for full details.

	MO	KO	EK	HWP	INTC	MSFT	IBM	MCD	WMT	DIS
$\hat{B}'$	0.20	0.19	0.16	0.45	0.51	0.44	0.32	0.18	0.24	0.22
	0.39	0.34	0.23	−0.26	−0.45	−0.10	−0.07	0.31	0.39	0.37
r^2	0.35	0.42	0.18	0.55	0.75	0.56	0.35	0.34	0.42	0.41
MO	1.00	0.39	0.25	0.17	0.13	0.25	0.20	0.35	0.38	0.38
KO	0.39	1.00	0.28	0.21	0.17	0.29	0.23	0.38	0.42	0.42
EK	0.25	0.28	1.00	0.18	0.15	0.22	0.18	0.25	0.28	0.27
HWP	0.17	0.21	0.18	1.00	0.64	0.55	0.43	0.20	0.23	0.23
INTC	0.13	0.17	0.15	0.64	1.00	0.61	0.48	0.16	0.19	0.18
MSFT	0.25	0.29	0.22	0.55	0.61	1.00	0.44	0.27	0.31	0.30
IBM	0.20	0.23	0.18	0.43	0.48	0.44	1.00	0.21	0.25	0.24
MCD	0.35	0.38	0.25	0.20	0.16	0.27	0.21	1.00	0.38	0.37
WMT	0.38	0.42	0.28	0.23	0.19	0.31	0.25	0.38	1.00	0.41
DIS	0.38	0.42	0.27	0.23	0.18	0.30	0.24	0.37	0.41	1.00
MO	1.00	−0.19	−0.15					−0.19	−0.37	−0.26
KO	−0.19	1.00	−0.15		0.11				−0.16	−0.17
EK	−0.15	−0.15	1.00					−0.15	−0.16	−0.16
HWP				1.00	−0.63	−0.37	−0.14			
INTC		0.11		−0.63	1.00	−0.24	−0.31			
MSFT				−0.37	−0.24	1.00	−0.22			
IBM				−0.14	−0.31	−0.22	1.00			
MCD	−0.19		−0.15					1.00	−0.19	−0.19
WMT	−0.37	−0.16	−0.16					−0.19	1.00	−0.23
DIS	−0.26	−0.17	−0.16					−0.19	−0.23	1.00

of larger entries, indicating imperfections in the factor-model representation of the data. In particular, we have introduced a number of larger negative correlations into the residuals; in practice, we seldom expect to find a factor model in which the residuals have a covariance matrix that appears perfectly diagonal.

Notes and Comments

For a more detailed discussion of factor models see the paper by Connor (1995), which provides a comparison of the three types of model, and the book by Campbell, Lo and MacKinlay (1997). An excellent practical introduction to these models with examples in S-Plus is Zivot and Wang (2003). Other accounts of factor models and PCA in finance are found in Alexander (2001) and Tsay (2002).

Much of our discussion of factor models, multivariate regression and principal components is based on Mardia, Kent and Bibby (1979). Statistical approaches to factor models are also treated in Seber (1984) and Johnson and Wichern (2002); these include classical statistical factor analysis, which we have omitted from our account.