Chapter 4
Multivariate Distributions

The preceding chapter showed that by using the two first moments of a multivariate distribution (the mean and the covariance matrix), a lot of information on the relationship between the variables can be made available. Only basic statistical theory was used to derive tests of independence or of linear relationships. In this chapter we give an introduction to the basic probability tools useful in statistical multivariate analysis.

Means and covariances share many interesting and useful properties, but they represent only part of the information on a multivariate distribution. Section 4.1 presents the basic probability tools used to describe a multivariate random variable, including marginal and conditional distributions and the concept of independence. In Sect. 4.2, basic properties on means and covariances (marginal and conditional ones) are derived.

Since many statistical procedures rely on transformations of a multivariate random variable, Sect. 4.3 proposes the basic techniques needed to derive the distribution of transformations with a special emphasis on linear transforms. As an important example of a multivariate random variable, Sect. 4.4 defines the multinormal distribution. It will be analysed in more detail in Chap. 5 along with most of its "companion" distributions that are useful in making multivariate statistical inferences.

The normal distribution plays a central role in statistics because it can be viewed as an approximation and limit of many other distributions. The basic justification relies on the central limit theorem presented in Sect. 4.5. We present this central theorem in the framework of sampling theory. A useful extension of this theorem is also given: it is an approximate distribution to transformations of asymptotically normal variables. The increasing power of computers today makes it possible to consider alternative approximate sampling distributions. These are based on resampling techniques and are suitable for many general situations. Section 4.8 gives an introduction to the ideas behind bootstrap approximations.

4.1   Distribution and Density Function

Let $X = (X_1, X_2, \ldots, X_p)^\top$ be a random vector. The cumulative distribution function (cdf) of X is defined by

$$F(x) = P(X \leq x) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_p \leq x_p).$$

For continuous X, a nonnegative probability density function (pdf) f exists that

$$F(x) = \int_{-\infty}^{x} f(u)du. \tag{4.1}$$

Note that

$$\int_{-\infty}^{\infty} f(u)\, du = 1.$$

Most of the integrals appearing below are multidimensional. For instance, $\int_{-\infty}^{x} f(u)du$ means $\int_{-\infty}^{x_p} \ldots \int_{-\infty}^{x_1} f(u_1, \ldots, u_p)du_1 \ldots du_p$. Note also that the cdf F is differentiable with

$$f(x) = \frac{\partial^p F(x)}{\partial x_1 \cdots \partial x_p}.$$

For discrete X, the values of this random variable are concentrated on a countable or finite set of points $\{c_j\}_{j \in J}$, the probability of events of the form $\{X \in D\}$ can then be computed as

$$P(X \in D) = \sum_{\{j : c_j \in D\}} P(X = c_j).$$

If we partition X as $X = (X_1, X_2)^\top$ with $X_1 \in \mathbb{R}^k$ and $X_2 \in \mathbb{R}^{p-k}$, then the function

$$F_{X_1}(x_1) = P(X_1 \leq x_1) = F(x_{11}, \ldots, x_{1k}, \infty, \ldots, \infty) \tag{4.2}$$

is called the *marginal cdf*. $F = F(x)$ is called the joint cdf. For continuous X the marginal pdf can be computed from the joint density by "integrating out" the variable not of interest.

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_2. \tag{4.3}$$

The conditional pdf of X_2 given $X_1 = x_1$ is given as

$$f(x_2 \mid x_1) = \frac{f(x_1, x_2)}{f_{X_1}(x_1)}. \tag{4.4}$$

Example 4.1 Consider the pdf

$$f(x_1, x_2) = \begin{cases} \frac{1}{2}x_1 + \frac{3}{2}x_2 & 0 \le x_1, x_2 \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

$f(x_1, x_2)$ is a density since

$$\int f(x_1, x_2) dx_1 dx_2 = \frac{1}{2}\left[\frac{x_1^2}{2}\right]_0^1 + \frac{3}{2}\left[\frac{x_2^2}{2}\right]_0^1 = \frac{1}{4} + \frac{3}{4} = 1.$$

The marginal densities are

$$f_{X_1}(x_1) = \int f(x_1, x_2) dx_2 = \int_0^1 \left(\frac{1}{2}x_1 + \frac{3}{2}x_2\right) dx_2 = \frac{1}{2}x_1 + \frac{3}{4};$$

$$f_{X_2}(x_2) = \int f(x_1, x_2) dx_1 = \int_0^1 \left(\frac{1}{2}x_1 + \frac{3}{2}x_2\right) dx_1 = \frac{3}{2}x_2 + \frac{1}{4}.$$

The conditional densities are therefore

$$f(x_2 \mid x_1) = \frac{\frac{1}{2}x_1 + \frac{3}{2}x_2}{\frac{1}{2}x_1 + \frac{3}{4}} \quad \text{and} \quad f(x_1 \mid x_2) = \frac{\frac{1}{2}x_1 + \frac{3}{2}x_2}{\frac{3}{2}x_2 + \frac{1}{4}}.$$

Note that these conditional pdf's are nonlinear in x_1 and x_2 although the joint pdf has a simple (linear) structure.

Independence of two random variables is defined as follows.

Definition 4.1 X_1 and X_2 are independent iff $f(x) = f(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$.

That is, X_1 and X_2 are independent if the conditional pdf's are equal to the marginal densities, i.e. $f(x_1 \mid x_2) = f_{X_1}(x_1)$ and $f(x_2 \mid x_1) = f_{X_2}(x_2)$. Independence can be interpreted as follows: knowing $X_2 = x_2$ does not change the probability assessments on X_1, and conversely.

⚠ Different joint pdf's may have the same marginal pdf's.

Example 4.2 Consider the pdf's

$$f(x_1, x_2) = 1, \quad 0 < x_1, x_2 < 1,$$

and

$$f(x_1, x_2) = 1 + \alpha(2x_1 - 1)(2x_2 - 1), \quad 0 < x_1, \ x_2 < 1, \quad -1 \le \alpha \le 1.$$

We compute in both cases the marginal pdf's as

$$f_{X_1}(x_1) = 1, \quad f_{X_2}(x_2) = 1.$$

Indeed

$$\int_0^1 1 + \alpha(2x_1 - 1)(2x_2 - 1)dx_2 = 1 + \alpha(2x_1 - 1)[x_2^2 - x_2]_0^1 = 1.$$

Hence we obtain identical marginals from different joint distributions.

Let us study the concept of independence using the bank notes example. Consider the variables X_4 (lower inner frame) and X_5 (upper inner frame). From Chap. 3, we already know that they have significant correlation, so they are almost surely not independent. Kernel estimates of the marginal densities, $\hat{f}_{X_4}$ and $\hat{f}_{X_5}$, are given in Fig. 4.1. In Fig. 4.2 (left) we show the product of these two densities. The kernel density technique was presented in Sect. 1.3. If X_4 and X_5 are independent, this product $\hat{f}_{X_4} \cdot \hat{f}_{X_5}$ should be roughly equal to $\hat{f}(x_4, x_5)$, the estimate of the joint density of (X_4, X_5). Comparing the two graphs in Fig. 4.2 reveals that the two densities are different. The two variables X_4 and X_5 are therefore not independent.

An elegant concept of connecting marginals with joint cdfs is given by *copulae*. Copulae are important in Value-at-Risk calculations and are an essential tool in quantitative finance (Härdle, Hautsch, & Overbeck, 2009).

For simplicity of presentation we concentrate on the $p = 2$ dimensional case. A two-dimensional copula is a function $C : [0, 1]^2 \rightarrow [0, 1]$ with the following properties:
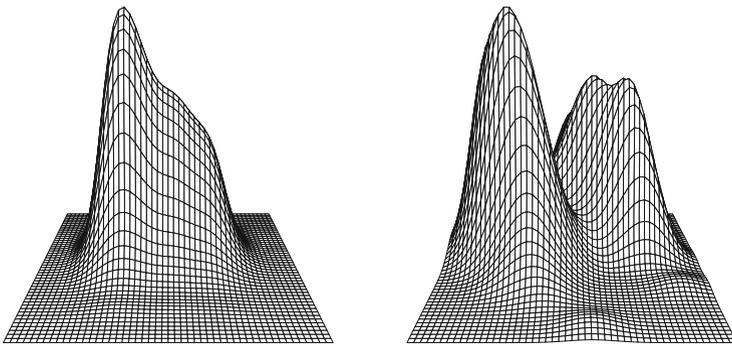
- For every $u \in [0, 1]$: $C(0, u) = C(u, 0) = 0$.
- For every $u \in [0, 1]$: $C(u, 1) = u$ and $C(1, u) = u$.
- For every $(u_1, u_2), (v_1, v_2) \in [0, 1] \times [0, 1]$ with $u_1 \le v_1$ and $u_2 \le v_2$:

$$C(v_1, v_2) - C(v_1, u_2) - C(u_1, v_2) + C(u_1, u_2) \ge 0.$$

The usage of the name "copula" for the function C is explained by the following theorem.
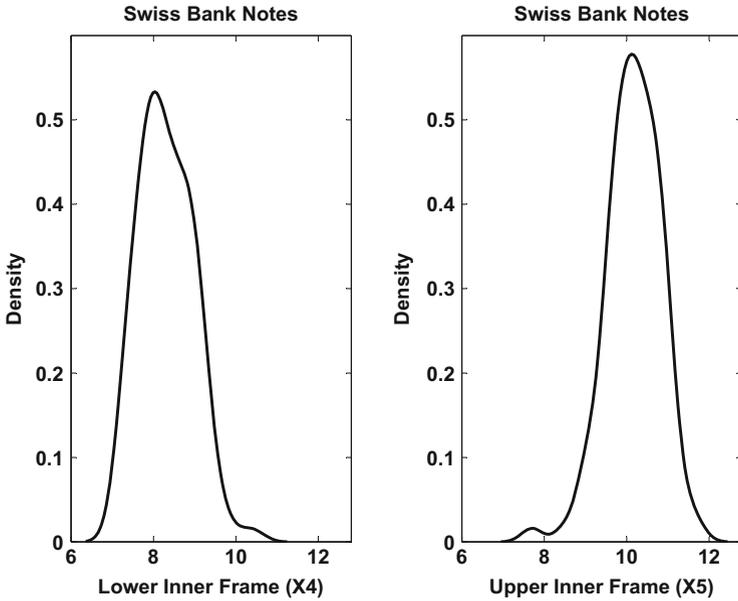
Fig. 4.1 Univariate estimates of the density of X_4 (*left*) and X_5 (*right*) of the bank notes ▣ MVAdenbank2

Fig. 4.2 The product of univariate density estimates (*left*) and the joint density estimate (*right*) for X_4 (*left*) and X_5 of the bank notes ▣ MVAdenbank3

Theorem 4.1 (Sklar's Theorem) *Let F be a joint distribution function with marginal distribution functions F_{X_1} and F_{X_2}. Then a copula C exists with*

$$F(x_1, x_2) = C\{F_{X_1}(x_1), F_{X_2}(x_2)\} \tag{4.5}$$

for every $x_1, x_2 \in \mathbb{R}$. If F_{X_1} and F_{X_2} are continuous, then C is unique. On the other hand, if C is a copula and F_{X_1} and F_{X_2} are distribution functions, then the function F defined by (4.5) is a joint distribution function with marginals F_{X_1} and F_{X_2}.

With Sklar's Theorem, the use of the name "copula" becomes obvious. It was chosen to describe "a function that links a multidimensional distribution to its one-dimensional margins" and appeared in the mathematical literature for the first time in Sklar (1959).

Example 4.3 The structure of independence implies that the product of the distribution functions F_{X_1} and F_{X_2} equals their joint distribution function F,

$$F(x_1, x_2) = F_{X_1}(x_1) \cdot F_{X_2}(x_2). \tag{4.6}$$

Thus, we obtain the *independence copula* $C = \Pi$ from

$$\Pi(u_1, \ldots, u_n) = \prod_{i=1}^{n} u_i \ .$$

Theorem 4.2 *Let X_1 and X_2 be random variables with continuous distribution functions F_{X_1} and F_{X_2} and the joint distribution function F. Then X_1 and X_2 are independent if and only if $C_{X_1, X_2} = \Pi$.*

Proof From Sklar's Theorem we know that there exists an unique copula C with

$$\mathrm{P}(X_1 \leq x_1, X_2 \leq x_2) = F(x_1, x_2) = C\{F_{X_1}(x_1), F_{X_2}(x_2)\}. \tag{4.7}$$

Independence can be seen using (4.5) for the joint distribution function F and the definition of Π,

$$F(x_1, x_2) = C\{F_{X_1}(x_1), F_{X_2}(x_2)\} = F_{X_1}(x_1) F_{X_2}(x_2) \ . \tag{4.8}$$

$\square$

Example 4.4 The *Gumbel–Hougaard* family of copulae (Nelsen, 1999) is given by the function

$$C_\theta(u, v) = \exp\left[-\left\{(-\log u)^\theta + (-\log v)^\theta\right\}^{1/\theta}\right] . \tag{4.9}$$

The parameter θ may take all values in the interval $[1, \infty)$. The Gumbel–Hougaard copulae are suited to describe bivariate extreme value distributions.

For $\theta = 1$, the expression (4.9) reduces to the product copula, i.e. $C_1(u, v) = \Pi(u, v) = u \, v$. For $\theta \to \infty$ one finds for the Gumbel–Hougaard copula:

$$C_\theta(u, v) \longrightarrow \min(u, v) = M(u, v),$$

where the function M is also a copula such that $C(u, v) \leq M(u, v)$ for arbitrary copula C. The copula M is called the *Fréchet–Hoeffding upper bound*.

Similarly, we obtain the *Fréchet–Hoeffding lower bound* $W(u, v) = \max(u + v - 1, 0)$ which satisfies $W(u, v) \leq C(u, v)$ for any other copula C.

1+1=4	Summary
$\hookrightarrow$	The cumulative distribution function (cdf) is defined as $F(x) = P(X < x)$.
$\hookrightarrow$	If a probability density function (pdf) f exists then $F(x) = \int_{-\infty}^{x} f(u)du$.
$\hookrightarrow$	The pdf integrates to one, i.e. $\int_{-\infty}^{\infty} f(x)dx = 1$.
$\hookrightarrow$	Let $X = (X_1, X_2)^\top$ be partitioned into sub-vectors X_1 and X_2 with joint cdf F. Then $F_{X_1}(x_1) = P(X_1 \leq x_1)$ is the marginal cdf of X_1. The marginal pdf of X_1 is obtained by $f_{X_1}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_2$. Different joint pdf's may have the same marginal pdf's.
$\hookrightarrow$	The conditional pdf of X_2 given $X_1 = x_1$ is defined as $f(x_2 \mid x_1) = \dfrac{f(x_1, x_2)}{f_{X_1}(x_1)}$.
$\hookrightarrow$	Two random variables X_1 and X_2 are called independent iff $f(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$. This is equivalent to $f(x_2 \mid x_1) = f_{X_2}(x_2)$.
$\hookrightarrow$	Different joint pdf's may have identical marginal pdf's.
$\hookrightarrow$	Copula is a function which connects marginals to form joint cdfs.

4.2 Moments and Characteristic Functions

Moments: Expectation and Covariance Matrix

If X is a random vector with density $f(x)$ then the expectation of X is

$$\mathsf{E}\, X = \begin{pmatrix} \mathsf{E}\, X_1 \\ \vdots \\ \mathsf{E}\, X_p \end{pmatrix} = \int x f(x)dx = \begin{pmatrix} \int x_1 f(x)dx \\ \vdots \\ \int x_p f(x)dx \end{pmatrix} = \mu. \tag{4.10}$$

Accordingly, the expectation of a matrix of random elements has to be understood component by component. The operation of forming expectations is linear:

$$\mathsf{E}\,(\alpha X + \beta Y) = \alpha\,\mathsf{E}\,X + \beta\,\mathsf{E}\,Y. \tag{4.11}$$

If $\mathcal{A}(q \times p)$ is a matrix of real numbers, we have:

$$\mathsf{E}(\mathcal{A}X) = \mathcal{A}\,\mathsf{E}\,X. \tag{4.12}$$

When X and Y are independent,

$$\mathsf{E}(XY^\top) = \mathsf{E}\,X\,\mathsf{E}\,Y^\top. \tag{4.13}$$

The matrix

$$\mathsf{Var}(X) = \Sigma = \mathsf{E}(X - \mu)(X - \mu)^\top \tag{4.14}$$

is the (theoretical) covariance matrix. We write for a vector X with mean vector μ and covariance matrix Σ,

$$X \sim (\mu, \Sigma). \tag{4.15}$$

The $(p \times q)$ matrix

$$\Sigma_{XY} = \mathsf{Cov}(X, Y) = \mathsf{E}(X - \mu)(Y - \nu)^\top \tag{4.16}$$

is the covariance matrix of $X \sim (\mu, \Sigma_{XX})$ and $Y \sim (\nu, \Sigma_{YY})$. Note that $\Sigma_{XY} = \Sigma_{YX}^\top$ and that $Z = \begin{pmatrix} X \\ Y \end{pmatrix}$ has covariance $\Sigma_{ZZ} = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}$. From

$$\mathsf{Cov}(X, Y) = \mathsf{E}(XY^\top) - \mu\nu^\top = \mathsf{E}(XY^\top) - \mathsf{E}\,X\,\mathsf{E}\,Y^\top \tag{4.17}$$

it follows that $\mathsf{Cov}(X, Y) = 0$ in the case where X and Y are independent. We often say that $\mu = \mathsf{E}(X)$ is the first order moment of X and that $\mathsf{E}(XX^\top)$ provides the second order moments of X:

$$\mathsf{E}(XX^\top) = \{\mathsf{E}(X_i X_j)\}, \text{ for } i = 1, \ldots, p \text{ and } j = 1, \ldots, p. \tag{4.18}$$

Properties of the Covariance Matrix $\Sigma = $ Var(X)

$$\Sigma = (\sigma_{X_i X_j}), \quad \sigma_{X_i X_j} = \mathsf{Cov}(X_i, X_j), \quad \sigma_{X_i X_i} = \mathsf{Var}(X_i) \tag{4.19}$$

$$\Sigma = \mathsf{E}(XX^\top) - \mu\mu^\top \tag{4.20}$$

$$\Sigma \geq 0 \tag{4.21}$$

Properties of Variances and Covariances

$$\mathsf{Var}(a^\top X) = a^\top \mathsf{Var}(X)a = \sum_{i,j} a_i a_j \sigma_{X_i X_j} \qquad (4.22)$$

$$\mathsf{Var}(\mathcal{A}X + b) = \mathcal{A}\,\mathsf{Var}(X)\mathcal{A}^\top \qquad (4.23)$$

$$\mathsf{Cov}(X + Y, Z) = \mathsf{Cov}(X, Z) + \mathsf{Cov}(Y, Z) \qquad (4.24)$$

$$\mathsf{Var}(X + Y) = \mathsf{Var}(X) + \mathsf{Cov}(X, Y) + \mathsf{Cov}(Y, X) + \mathsf{Var}(Y) \qquad (4.25)$$

$$\mathsf{Cov}(\mathcal{A}X, \mathcal{B}Y) = \mathcal{A}\,\mathsf{Cov}(X, Y)\mathcal{B}^\top. \qquad (4.26)$$

Let us compute these quantities for a specific joint density.

Example 4.5   Consider the pdf of Example 4.1. The mean vector $\mu = \binom{\mu_1}{\mu_2}$ is

$$\mu_1 = \int \int x_1 f(x_1, x_2) dx_1 dx_2 = \int_0^1 \int_0^1 x_1 \left(\frac{1}{2}x_1 + \frac{3}{2}x_2 \right) dx_1 dx_2$$

$$= \int_0^1 x_1 \left(\frac{1}{2}x_1 + \frac{3}{4} \right) dx_1 = \frac{1}{2} \left[\frac{x_1^3}{3} \right]_0^1 + \frac{3}{4} \left[\frac{x_1^2}{2} \right]_0^1$$

$$= \frac{1}{6} + \frac{3}{8} = \frac{4 + 9}{24} = \frac{13}{24},$$

$$\mu_2 = \int \int x_2 f(x_1, x_2) dx_1 dx_2 = \int_0^1 \int_0^1 x_2 \left(\frac{1}{2}x_1 + \frac{3}{2}x_2 \right) dx_1 dx_2$$

$$= \int_0^1 x_2 \left(\frac{1}{4} + \frac{3}{2}x_2 \right) dx_2 = \frac{1}{4} \left[\frac{x_2^2}{2} \right]_0^1 + \frac{3}{2} \left[\frac{x_2^3}{3} \right]_0^1$$

$$= \frac{1}{8} + \frac{1}{2} = \frac{1 + 4}{8} = \frac{5}{8}.$$

The elements of the covariance matrix are

$$\sigma_{X_1 X_1} = \mathsf{E}\,X_1^2 - \mu_1^2 \quad \text{with}$$

$$\mathsf{E}\,X_1^2 = \int_0^1 \int_0^1 x_1^2 \left(\frac{1}{2}x_1 + \frac{3}{2}x_2 \right) dx_1 dx_2 = \frac{1}{2} \left[\frac{x_1^4}{4} \right]_0^1 + \frac{3}{4} \left[\frac{x_1^3}{3} \right]_0^1 = \frac{3}{8}$$

$$\sigma_{X_2 X_2} = \mathsf{E}\,X_2^2 - \mu_2^2 \quad \text{with}$$

$$\mathsf{E}\,X_2^2 = \int_0^1 \int_0^1 x_2^2 \left(\frac{1}{2}x_1 + \frac{3}{2}x_2 \right) dx_1 dx_2 = \frac{1}{4} \left[\frac{x_2^3}{3} \right]_0^1 + \frac{3}{2} \left[\frac{x_2^4}{4} \right]_0^1 = \frac{11}{24}$$

$$\sigma_{X_1 X_2} = \mathsf{E}(X_1 X_2) - \mu_1 \mu_2 \quad \text{with}$$

$$\mathsf{E}(X_1 X_2) = \int_0^1 \int_0^1 x_1 x_2 \left(\frac{1}{2} x_1 + \frac{3}{2} x_2 \right) dx_1 dx_2 = \int_0^1 \left(\frac{1}{6} x_2 + \frac{3}{4} x_2^2 \right) dx_2$$

$$= \frac{1}{6} \left[\frac{x_2^2}{2} \right]_0^1 + \frac{3}{4} \left[\frac{x_2^3}{3} \right]_0^1 = \frac{1}{3}.$$

Hence the covariance matrix is

$$\Sigma = \begin{pmatrix} 0.0815 & 0.0052 \\ 0.0052 & 0.0677 \end{pmatrix}.$$

Conditional Expectations

The conditional expectations are

$$\mathsf{E}(X_2 \mid x_1) = \int x_2 f(x_2 \mid x_1) \, dx_2 \quad \text{and} \quad \mathsf{E}(X_1 \mid x_2) = \int x_1 f(x_1 \mid x_2) \, dx_1. \tag{4.27}$$

$\mathsf{E}(X_2|x_1)$ represents the location parameter of the conditional pdf of X_2 given that $X_1 = x_1$. In the same way, we can define $\mathsf{Var}(X_2|X_1 = x_1)$ as a measure of the dispersion of X_2 given that $X_1 = x_1$. We have from (4.20) that

$$\mathsf{Var}(X_2|X_1 = x_1) = \mathsf{E}(X_2 X_2^\top|X_1 = x_1) - \mathsf{E}(X_2|X_1 = x_1) \, \mathsf{E}(X_2^\top|X_1 = x_1).$$

Using the conditional covariance matrix, the conditional correlations may be defined as:

$$\rho_{X_2 X_3 | X_1 = x_1} = \frac{\mathsf{Cov}(X_2, X_3 | X_1 = x_1)}{\sqrt{\mathsf{Var}(X_2 | X_1 = x_1) \, \mathsf{Var}(X_3 | X_1 = x_1)}}.$$

These conditional correlations are known as partial correlations between X_2 and X_3, conditioned on X_1 being equal to x_1.

Example 4.6 Consider the following pdf

$$f(x_1, x_2, x_3) = \frac{2}{3}(x_1 + x_2 + x_3) \text{ where } 0 < x_1, x_2, x_3 < 1.$$

Note that the pdf is symmetric in x_1, x_2 and x_3 which facilitates the computations. For instance,

$$f(x_1, x_2) = \tfrac{2}{3}(x_1 + x_2 + \tfrac{1}{2}) \quad 0 < x_1, x_2 < 1$$

$$f(x_1) \quad = \tfrac{2}{3}(x_1 + 1) \qquad 0 < x_1 < 1$$

and the other marginals are similar. We also have

$$f(x_1, x_2 | x_3) = \frac{x_1 + x_2 + x_3}{x_3 + 1}, \quad 0 < x_1, x_2 < 1$$

$$f(x_1 | x_3) = \frac{x_1 + x_3 + \tfrac{1}{2}}{x_3 + 1}, \quad 0 < x_1 < 1.$$

It is easy to compute the following moments:

$$\mathsf{E}(X_i) = \tfrac{5}{9}; \; \mathsf{E}(X_i^2) = \tfrac{7}{18}; \; \mathsf{E}(X_i X_j) = \tfrac{11}{36} \quad (i \neq j \text{ and } i, j = 1, 2, 3)$$

$$\mathsf{E}(X_1 | X_3 = x_3) = \mathsf{E}(X_2 | X_3 = x_3) = \tfrac{1}{12} \left(\tfrac{6x_3 + 7}{x_3 + 1} \right);$$

$$\mathsf{E}(X_1^2 | X_3 = x_3) = \mathsf{E}(X_2^2 | X_3 = x_3) = \tfrac{1}{12} \left(\tfrac{4x_3 + 5}{x_3 + 1} \right)$$

and

$$\mathsf{E}(X_1 X_2 | X_3 = x_3) = \tfrac{1}{12} \left(\tfrac{3x_3 + 4}{x_3 + 1} \right).$$

Note that the conditional means of X_1 and of X_2, given $X_3 = x_3$, are not linear in x_3. From these moments we obtain:

$$\Sigma = \begin{pmatrix} \frac{13}{162} & -\frac{1}{324} & -\frac{1}{324} \\ -\frac{1}{324} & \frac{13}{162} & -\frac{1}{324} \\ -\frac{1}{324} & -\frac{1}{324} & \frac{13}{162} \end{pmatrix} \text{ in particular } \rho_{X_1 X_2} = -\frac{1}{26} \approx -0.0385.$$

The conditional covariance matrix of X_1 and X_2, given $X_3 = x_3$ is

$$\mathsf{Var}\left(\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \Big| X_3 = x_3 \right) = \begin{pmatrix} \frac{12x_3^2 + 24x_3 + 11}{144(x_3 + 1)^2} & \frac{-1}{144(x_3 + 1)^2} \\ \frac{-1}{144(x_3 + 1)^2} & \frac{12x_3^2 + 24x_3 + 11}{144(x_3 + 1)^2} \end{pmatrix}.$$

In particular, the partial correlation between X_1 and X_2, given that X_3 is fixed at x_3, is given by $\rho_{X_1 X_2 | X_3 = x_3} = -\frac{1}{12x_3^2 + 24x_3 + 11}$ which ranges from -0.0909 to -0.0213 when x_3 goes from 0 to 1. Therefore, in this example, the partial correlation may be larger or smaller than the simple correlation, depending on the value of the condition $X_3 = x_3$.

Example 4.7 Consider the following joint pdf

$$f(x_1, x_2, x_3) = 2x_2(x_1 + x_3); \quad 0 < x_1, x_2, x_3 < 1.$$

Note the symmetry of x_1 and x_3 in the pdf and that X_2 is independent of (X_1, X_3). It immediately follows that

$$f(x_1, x_3) = (x_1 + x_3) \qquad 0 < x_1, x_3 < 1$$

$$f(x_1) = x_1 + \frac{1}{2};$$
$$f(x_2) = 2x_2;$$
$$f(x_3) = x_3 + \frac{1}{2}.$$

Simple computations lead to

$$\mathsf{E}(X) = \begin{pmatrix} \frac{7}{12} \\ \frac{2}{3} \\ \frac{7}{12} \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \frac{11}{144} & 0 & -\frac{1}{144} \\ 0 & \frac{1}{18} & 0 \\ -\frac{1}{144} & 0 & \frac{11}{144} \end{pmatrix}.$$

Let us analyse the conditional distribution of (X_1, X_2) given $X_3 = x_3$. We have

$$f(x_1, x_2 | x_3) = \frac{4(x_1 + x_3)x_2}{2x_3 + 1} \quad 0 < x_1, x_2 < 1$$

$$f(x_1 | x_3) = 2\left(\frac{x_1 + x_3}{2x_3 + 1}\right) \quad 0 < x_1 < 1$$

$$f(x_2 | x_3) = f(x_2) = 2x_2 \quad 0 < x_2 < 1$$

so that again X_1 and X_2 are independent conditional on $X_3 = x_3$. In this case

$$\mathsf{E}\left(\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \middle| X_3 = x_3\right) = \begin{pmatrix} \frac{1}{3}\left(\frac{2+3x_3}{1+2x_3}\right) \\ \frac{2}{3} \end{pmatrix}$$

$$\mathsf{Var}\left(\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \middle| X_3 = x_3\right) = \begin{pmatrix} \frac{1}{18}\left(\frac{6x_3^2 + 6x_3 + 1}{(2x_3 + 1)^2}\right) & 0 \\ 0 & \frac{1}{18} \end{pmatrix}.$$

Properties of Conditional Expectations

Since $E(X_2|X_1 = x_1)$ is a function of x_1, say $h(x_1)$, we can define the random variable $h(X_1) = E(X_2|X_1)$. The same can be done when defining the random variable $Var(X_2|X_1)$. These two random variables share some interesting properties:

$$E(X_2) = E\{E(X_2|X_1)\} \tag{4.28}$$

$$Var(X_2) = E\{Var(X_2|X_1)\} + Var\{E(X_2|X_1)\}. \tag{4.29}$$

Example 4.8 Consider the following pdf

$$f(x_1, x_2) = 2e^{-\frac{x_2}{x_1}}; \ 0 < x_1 < 1, x_2 > 0.$$

It is easy to show that

$$f(x_1) = 2x_1 \ \text{for} \ 0 < x_1 < 1; \quad E(X_1) = \frac{2}{3} \ \text{and} \ Var(X_1) = \frac{1}{18}$$

$$f(x_2|x_1) = \frac{1}{x_1}e^{-\frac{x_2}{x_1}} \ \text{for} \ x_2 > 0; \quad E(X_2|X_1) = X_1 \ \text{and} \ Var(X_2|X_1) = X_1^2.$$

Without explicitly computing $f(x_2)$, we can obtain:

$$E(X_2) = E\{E(X_2|X_1)\} = E(X_1) = \frac{2}{3}$$

$$Var(X_2) = E\{Var(X_2|X_1)\} + Var\{E(X_2|X_1)\}$$

$$= E(X_1^2) + Var(X_1) = \frac{2}{4} + \frac{1}{18} = \frac{10}{18}.$$

The conditional expectation $E(X_2|X_1)$ viewed as a function $h(X_1)$ of X_1 (known as the regression function of X_2 on X_1), can be interpreted as a conditional approximation of X_2 by a function of X_1. The error term of the approximation is then given by:

$$U = X_2 - E(X_2|X_1).$$

Theorem 4.3 *Let $X_1 \in \mathbb{R}^k$ and $X_2 \in \mathbb{R}^{p-k}$ and $U = X_2 - E(X_2|X_1)$. Then we have:*

1. $E(U) = 0$
2. $E(X_2|X_1)$ *is the best approximation of X_2 by a function $h(X_1)$ of X_1 where h :* $\mathbb{R}^k \longrightarrow \mathbb{R}^{p-k}$. *"Best" is the minimum mean squared error (MSE) sense, where*

$$MSE(h) = E[\{X_2 - h(X_1)\}^\top \{X_2 - h(X_1)\}].$$

Characteristic Functions

The characteristic function (cf) of a random vector $X \in \mathbb{R}^p$ (respectively its density $f(x)$) is defined as

$$\varphi_X(t) = \mathsf{E}(e^{\mathbf{i} t^\top X}) = \int e^{\mathbf{i} t^\top x} f(x) \, dx, \quad t \in \mathbb{R}^p,$$

where $\mathbf{i}$ is the complex unit: $\mathbf{i}^2 = -1$. The cf has the following properties:

$$\varphi_X(0) = 1 \quad \text{and} \quad |\varphi_X(t)| \leq 1. \tag{4.30}$$

If φ is absolutely integrable, i.e. the integral $\int_{-\infty}^{\infty} |\varphi(x)| dx$ exists and is finite, then

$$f(x) = \frac{1}{(2\pi)^p} \int_{-\infty}^{\infty} e^{-\mathbf{i} t^\top x} \varphi_X(t) \, dt. \tag{4.31}$$

If $X = (X_1, X_2, \ldots, X_p)^\top$, then for $t = (t_1, t_2, \ldots, t_p)^\top$

$$\varphi_{X_1}(t_1) = \varphi_X(t_1, 0, \ldots, 0), \ldots, \varphi_{X_p}(t_p) = \varphi_X(0, \ldots, 0, t_p). \tag{4.32}$$

If $X_1, \ldots, X_p$ are independent random variables, then for $t = (t_1, t_2, \ldots, t_p)^\top$

$$\varphi_X(t) = \varphi_{X_1}(t_1) \cdot \ldots \cdot \varphi_{X_p}(t_p). \tag{4.33}$$

If $X_1, \ldots, X_p$ are independent random variables, then for $t \in \mathbb{R}$

$$\varphi_{X_1 + \cdots + X_p}(t) = \varphi_{X_1}(t) \cdot \ldots \cdot \varphi_{X_p}(t). \tag{4.34}$$

The characteristic function can recover all the cross-product moments of any order: $\forall j_k \geq 0, k = 1, \ldots, p$ and for $t = (t_1, \ldots, t_p)^\top$ we have

$$\mathsf{E}\left(X_1^{j_1} \cdot \ldots \cdot X_p^{j_p}\right) = \frac{1}{\mathbf{i}^{j_1 + \cdots + j_p}} \left[\frac{\partial \varphi_X(t)}{\partial t_1^{j_1} \ldots \partial t_p^{j_p}} \right]_{t=0}. \tag{4.35}$$

Example 4.9 The cf of the density in Example 4.5 is given by

$$\varphi_X(t) = \int_0^1 \int_0^1 e^{\mathbf{i} t^\top x} f(x) dx$$

$$= \int_0^1 \int_0^1 \{\cos(t_1 x_1 + t_2 x_2) + \mathbf{i} \sin(t_1 x_1 + t_2 x_2)\} \left(\frac{1}{2} x_1 + \frac{3}{2} x_2\right) dx_1 dx_2,$$

$$= \frac{0.5\,e^{\mathbf{i}t_1}\left(3\,\mathbf{i}\,t_1 - 3\,\mathbf{i}\,e^{\mathbf{i}t_2}\,t_1 + \mathbf{i}\,t_2 - \mathbf{i}\,e^{\mathbf{i}t_2}\,t_2 + t_1\,t_2 - 4\,e^{\mathbf{i}t_2}\,t_1\,t_2\right)}{t_1{}^2\,t_2{}^2}$$

$$-\,\frac{0.5\left(3\,\mathbf{i}\,t_1 - 3\,\mathbf{i}\,e^{\mathbf{i}t_2}\,t_1 + \mathbf{i}\,t_2 - \mathbf{i}\,e^{\mathbf{i}t_2}\,t_2 - 3\,e^{\mathbf{i}t_2}\,t_1\,t_2\right)}{t_1{}^2\,t_2{}^2}.$$

Example 4.10 Suppose $X \in \mathbb{R}^1$ follows the density of the standard normal distribution

$$f_X(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

(see Sect. 4.4) then the cf can be computed via

$$\varphi_X(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\mathbf{i}tx}\exp\left(-\frac{x^2}{2}\right)\,dx$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2}(x^2 - 2\mathbf{i}tx + \mathbf{i}^2 t^2)\right\}\exp\left\{\frac{1}{2}\mathbf{i}^2 t^2\right\}\,dx$$

$$= \exp\left(-\frac{t^2}{2}\right)\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(x - \mathbf{i}t)^2}{2}\right\}\,dx$$

$$= \exp\left(-\frac{t^2}{2}\right),$$

since $\mathbf{i}^2 = -1$ and $\int \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(x-\mathbf{i}t)^2}{2}\right\}\,dx = 1$.

A variety of distributional characteristics can be computed from $\varphi_X(t)$. The standard normal distribution has a very simple cf, as was seen in Example 4.10. Deviations from normal covariance structures can be measured by the deviations from the cf (or characteristics of it). In Table 4.1 we give an overview of the cf's for a variety of distributions.

Theorem 4.4 (Cramer–Wold)   *The distribution of $X \in \mathbb{R}^p$ is completely determined by the set of all (one-dimensional) distributions of $t^\top X$ where $t \in \mathbb{R}^p$.*

Table 4.1 Characteristic functions for some common distributions

	pdf	cf
Uniform	$f(x) = I(x \in [a, b])/(b - a)$	$\varphi_X(t) = (e^{\mathbf{i}bt} - e^{\mathbf{i}at})/(b - a)\mathbf{i}t$
$N_1(\mu, \sigma^2)$	$f(x) = (2\pi\sigma^2)^{-1/2}\exp\{-(x - \mu)^2/2\sigma^2\}$	$\varphi_X(t) = e^{\mathbf{i}\mu t - \sigma^2 t^2/2}$
$\chi^2(n)$	$f(x) = I(x > 0)x^{n/2-1}e^{-x/2}/\{\Gamma(n/2)2^{n/2}\}$	$\varphi_X(t) = (1 - 2\mathbf{i}t)^{-n/2}$
$N_p(\mu, \Sigma)$	$f(x) = |2\pi\Sigma|^{-1/2}\exp\{-(x - \mu)^\top\Sigma(x - \mu)/2\}$	$\varphi_X(t) = e^{\mathbf{i}t^\top\mu - t^\top\Sigma t/2}$

This theorem says that we can determine the distribution of X in $\mathbb{R}^p$ by specifying all of the one-dimensional distributions of the linear combinations

$$\sum_{j=1}^{p} t_j X_j = t^\top X, \quad t = (t_1, t_2, \ldots, t_p)^\top.$$

Cumulant Functions

Moments $m_k = \int x^k f(x) dx$ often help in describing distributional characteristics. The normal distribution in $d = 1$ dimension is completely characterised by its standard normal density $f = \varphi$ and the moment parameters are $\mu = m_1$ and $\sigma^2 = m_2 - m_1^2$. Another helpful class of parameters are the cumulants or semi-invariants of a distribution. In order to simplify notation we concentrate here on the one-dimensional ($d = 1$) case.

For a given one-dimensional random variable X with density f and finite moments of order k the characteristic function $\varphi_X(t) = \mathsf{E}(e^{\mathbf{i}tX})$ has the derivative

$$\frac{1}{\mathbf{i}^j} \left[\frac{\partial^j \log \{\varphi_X(t)\}}{\partial t^j} \right]_{t=0} = \kappa_j, \qquad j = 1, \ldots, k.$$

The values κ_j are called cumulants or semi-invariants since κ_j does not change (for $j > 1$) under a shift transformation $X \mapsto X + a$. The cumulants are natural parameters for dimension reduction methods, in particular the Projection Pursuit method (see Sect. 20.2).

The relationship between the first k moments $m_1, \ldots, m_k$ and the cumulants is given by

$$\kappa_k = (-1)^{k-1} \begin{vmatrix} m_1 & 1 & \ldots & 0 \\ m_2 & \begin{pmatrix} 1 \\ 0 \end{pmatrix} m_1 & \ldots & \\ \vdots & \vdots & \ddots & \vdots \\ m_k & \begin{pmatrix} k-1 \\ 0 \end{pmatrix} m_{k-1} & \ldots & \begin{pmatrix} k-1 \\ k-2 \end{pmatrix} m_1 \end{vmatrix}. \tag{4.36}$$

Example 4.11 Suppose that $k = 1$, then formula (4.36) above yields

$$\kappa_1 = m_1.$$

For $k = 2$ we obtain

$$\kappa_2 = - \begin{vmatrix} m_1 & 1 \\ m_2 & \binom{1}{0} m_1 \end{vmatrix} = m_2 - m_1^2.$$

For $k = 3$ we have to calculate

$$\kappa_3 = \begin{vmatrix} m_1 & 1 & 0 \\ m_2 & m_1 & 1 \\ m_3 & m_2 & 2m_1 \end{vmatrix}.$$

Calculating the determinant we have:

$$\kappa_3 = m_1 \begin{vmatrix} m_1 & 1 \\ m_2 & 2m_1 \end{vmatrix} - m_2 \begin{vmatrix} 1 & 0 \\ m_2 & 2m_1 \end{vmatrix} + m_3 \begin{vmatrix} 1 & 0 \\ m_1 & 1 \end{vmatrix}$$

$$= m_1(2m_1^2 - m_2) - m_2(2m_1) + m_3$$

$$= m_3 - 3m_1 m_2 + 2m_1^3. \tag{4.37}$$

Similarly one calculates

$$\kappa_4 = m_4 - 4m_3 m_1 - 3m_2^2 + 12m_2 m_1^2 - 6m_1^4. \tag{4.38}$$

The same type of process is used to find the moments from the cumulants:

$$m_1 = \kappa_1$$

$$m_2 = \kappa_2 + \kappa_1^2$$

$$m_3 = \kappa_3 + 3\kappa_2 \kappa_1 + \kappa_1^3$$

$$m_4 = \kappa_4 + 4\kappa_3 \kappa_1 + 3\kappa_2^2 + 6\kappa_2 \kappa_1^2 + \kappa_1^4. \tag{4.39}$$

A very simple relationship can be observed between the semi-invariants and the central moments $\mu_k = \mathsf{E}(X - \mu)^k$, where $\mu = m_1$ as defined before. In fact, $\kappa_2 = \mu_2, \kappa_3 = \mu_3$ and $\kappa_4 = \mu_4 - 3\mu_2^2$.

Skewness γ_3 and kurtosis γ_4 are defined as:

$$\gamma_3 = \mathsf{E}(X - \mu)^3 / \sigma^3$$

$$\gamma_4 = \mathsf{E}(X - \mu)^4 / \sigma^4. \tag{4.40}$$

The skewness and kurtosis determine the shape of one-dimensional distributions. The skewness of a normal distribution is 0 and the kurtosis equals 3. The relation of these parameters to the cumulants is given by:

$$\gamma_3 = \frac{\kappa_3}{\kappa_2^{3/2}} \tag{4.41}$$

From (4.39) and Example 4.11

$$\gamma_4 = \frac{\kappa_4 + 3\kappa_2^2 + \kappa_1^4 - m_1^4}{\sigma^4} = \frac{\kappa_4 + 3\kappa_2^2}{\kappa_2^2} = \frac{\kappa_4}{\kappa_2^2} + 3. \tag{4.42}$$

These relations will be used later in Sect. 20.2 on Projection Pursuit to determine deviations from normality.

Summary

↪ The expectation of a random vector X is $\mu = \int xf(x)\ dx$, the covariance matrix $\Sigma = \mathsf{Var}(X) = \mathsf{E}(X - \mu)(X - \mu)^\top$. We denote $X \sim (\mu, \Sigma)$.

↪ Expectations are linear, i.e. $\mathsf{E}(\alpha X + \beta Y) = \alpha\,\mathsf{E}\,X + \beta\,\mathsf{E}\,Y$. If X and Y are independent, then $\mathsf{E}(XY^\top) = \mathsf{E}\,X\,\mathsf{E}\,Y^\top$.

↪ The covariance between two random vectors X and Y is $\Sigma_{XY} = \mathsf{Cov}(X, Y) = \mathsf{E}(X - \mathsf{E}\,X)(Y - \mathsf{E}\,Y)^\top = \mathsf{E}(XY^\top) - \mathsf{E}\,X\,\mathsf{E}\,Y^\top$. If X and Y are independent, then $\mathsf{Cov}(X, Y) = 0$.

↪ The characteristic function (cf) of a random vector X is $\varphi_X(t) = \mathsf{E}(e^{\mathbf{i}t^\top X})$.

↪ The distribution of a p-dimensional random variable X is completely determined by all one-dimensional distributions of $t^\top X$ where $t \in \mathbb{R}^p$ (Theorem of Cramer–Wold).

↪ The conditional expectation $\mathsf{E}(X_2|X_1)$ is the MSE best approximation of X_2 by a function of X_1.

4.3   Transformations

Suppose that X has pdf $f_X(x)$. What is the pdf of $Y = 3X$? Or if $X = (X_1, X_2, X_3)^\top$, what is the pdf of

$$Y = \begin{pmatrix} 3X_1 \\ X_1 - 4X_2 \\ X_3 \end{pmatrix}?$$

This is a special case of asking for the pdf of Y when

$$X = u(Y) \tag{4.43}$$

for a one-to-one transformation $u: \mathbb{R}^p \to \mathbb{R}^p$. Define the Jacobian of u as

$$\mathcal{J} = \left(\frac{\partial x_i}{\partial y_j} \right) = \left(\frac{\partial u_i(y)}{\partial y_j} \right)$$

and let abs($|\mathcal{J}|$) be the absolute value of the determinant of this Jacobian. The pdf of Y is given by

$$f_Y(y) = \text{abs}(|\mathcal{J}|) \cdot f_X\{u(y)\}. \tag{4.44}$$

Using this we can answer the introductory questions, namely

$$(x_1, \ldots, x_p)^\top = u(y_1, \ldots, y_p) = \frac{1}{3}(y_1, \ldots, y_p)^\top$$

with

$$\mathcal{J} = \begin{pmatrix} \frac{1}{3} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{3} \end{pmatrix}$$

and hence abs($|\mathcal{J}|$) $= \left(\frac{1}{3}\right)^p$. So the pdf of Y is $\frac{1}{3^p} f_X\left(\frac{y}{3}\right)$.

This introductory example is a special case of

$$Y = \mathcal{A}X + b, \text{ where } \mathcal{A} \text{ is nonsingular.}$$

The inverse transformation is

$$X = \mathcal{A}^{-1}(Y - b).$$

Therefore

$$\mathcal{J} = \mathcal{A}^{-1},$$

and hence

$$f_Y(y) = \text{abs}(|\mathcal{A}|^{-1}) f_X \{\mathcal{A}^{-1}(y - b)\}. \tag{4.45}$$

Example 4.12 Consider $X = (X_1, X_2) \in \mathbb{R}^2$ with density $f_X(x) = f_X(x_1, x_2)$,

$$\mathcal{A} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Then

$$Y = \mathcal{A}X + b = \begin{pmatrix} X_1 + X_2 \\ X_1 - X_2 \end{pmatrix}$$

and

$$|\mathcal{A}| = -2, \quad \text{abs}(|\mathcal{A}|^{-1}) = \frac{1}{2}, \quad \mathcal{A}^{-1} = -\frac{1}{2}\begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Hence

$$f_Y(y) = \text{abs}(|\mathcal{A}|^{-1}) \cdot f_X(\mathcal{A}^{-1}y)$$

$$= \frac{1}{2} f_X \left\{ \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right\}$$

$$= \frac{1}{2} f_X \left\{ \frac{1}{2}(y_1 + y_2), \frac{1}{2}(y_1 - y_2) \right\}. \tag{4.46}$$

Example 4.13 Consider $X \in \mathbb{R}^1$ with density $f_X(x)$ and $Y = \exp(X)$. According to (4.43) $x = u(y) = \log(y)$ and hence the Jacobian is

$$\mathcal{J} = \frac{dx}{dy} = \frac{1}{y}.$$

The pdf of Y is therefore:

$$f_Y(y) = \frac{1}{y} f_X \{\log(y)\}.$$

1+1=⌀ Summary
↪ If X has pdf $f_X(x)$, then a transformed random vector Y, i.e. $X = u(Y)$, has pdf $f_Y(y) = \text{abs}(|\mathcal{J}|) \cdot f_X\{u(y)\}$, where $\mathcal{J}$ denotes the Jacobian $\mathcal{J} = \left(\frac{\partial u(y_i)}{\partial y_j}\right)$.
↪ In the case of a linear relation $Y = \mathcal{A}X + b$ the pdf's of X and Y are related via $f_Y(y) = \text{abs}(|\mathcal{A}|^{-1})f_X\{\mathcal{A}^{-1}(y - b)\}$.

4.4 The Multinormal Distribution

The multinormal distribution with mean μ and covariance $\Sigma > 0$ has the density

$$f(x) = |2\pi\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right\} . \tag{4.47}$$

We write $X \sim N_p(\mu, \Sigma)$.

How is this multinormal distribution with mean μ and covariance Σ related to the multivariate standard normal $N_p(0, \mathcal{I}_p)$? Through a linear transformation using the results of Sect. 4.3, as shown in the next theorem.

Theorem 4.5 *Let $X \sim N_p(\mu, \Sigma)$ and $Y = \Sigma^{-1/2}(X - \mu)$ (Mahalanobis transformation). Then*

$$Y \sim N_p(0, \mathcal{I}_p),$$

i.e. the elements $Y_j \in \mathbb{R}$ are independent, one-dimensional $N(0, 1)$ variables.

Proof Note that $(X - \mu)^\top \Sigma^{-1}(X - \mu) = Y^\top Y$. Application of (4.45) gives $\mathcal{J} = \Sigma^{1/2}$, hence

$$f_Y(y) = (2\pi)^{-p/2} \exp\left(-\frac{1}{2}y^\top y\right) \tag{4.48}$$

which is by (4.47) the pdf of a $N_p(0, \mathcal{I}_p)$. $\qquad\square$

Note that the above Mahalanobis transformation yields in fact a random variable $Y = (Y_1, \ldots, Y_p)^\top$ composed of independent one-dimensional $Y_j \sim N_1(0, 1)$ since

$$f_Y(y) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2} y^\top y\right)$$

$$= \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} y_j^2\right)$$

$$= \prod_{j=1}^{p} f_{Y_j}(y_j).$$

Here each $f_{Y_j}(y)$ is a standard normal density $\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)$. From this it is clear that $\mathsf{E}(Y) = 0$ and $\mathsf{Var}(Y) = \mathcal{I}_p$.

How can we create $N_p(\mu, \Sigma)$ variables on the basis of $N_p(0, \mathcal{I}_p)$ variables? We use the inverse linear transformation

$$X = \Sigma^{1/2} Y + \mu. \tag{4.49}$$

Using (4.11) and (4.23) we can also check that $\mathsf{E}(X) = \mu$ and $\mathsf{Var}(X) = \Sigma$. The following theorem is useful because it presents the distribution of a variable after it has been linearly transformed. The proof is left as an exercise.

Theorem 4.6 *Let $X \sim N_p(\mu, \Sigma)$ and $\mathcal{A}(p \times p)$, $c \in \mathbb{R}^p$, where $\mathcal{A}$ is nonsingular. Then $Y = \mathcal{A}X + c$ is again a p-variate Normal, i.e.*

$$Y \sim N_p(\mathcal{A}\mu + c, \mathcal{A}\Sigma\mathcal{A}^\top). \tag{4.50}$$

Geometry of the $N_p(\mu, \Sigma)$ Distribution

From (4.47) we see that the density of the $N_p(\mu, \Sigma)$ distribution is constant on ellipsoids of the form

$$(x - \mu)^\top \Sigma^{-1} (x - \mu) = d^2. \tag{4.51}$$

Example 4.14 Figure 4.3 shows the contour ellipses of a two-dimensional normal distribution. Note that these contour ellipses are the iso-distance curves (2.34) from the mean of this normal distribution corresponding to the metric Σ^{-1}.

Normal sample
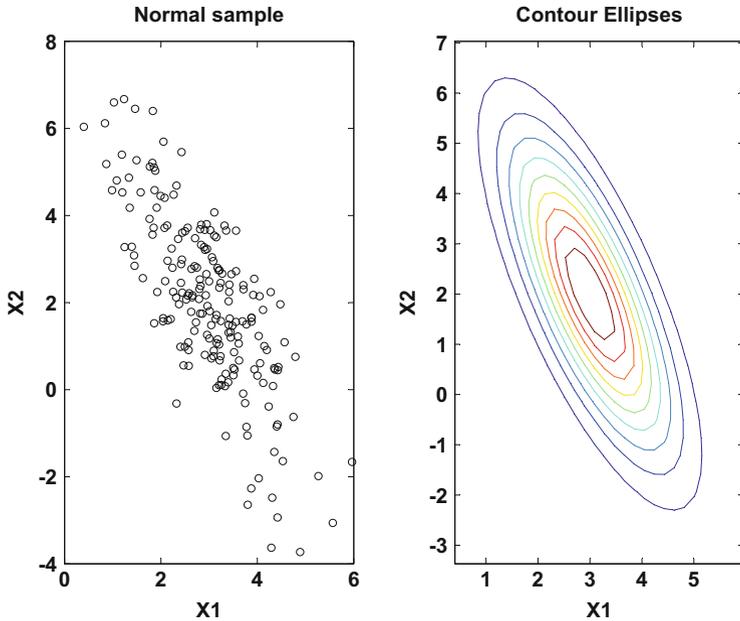
Contour Ellipses

Fig. 4.3 Scatterplot of a normal sample and contour ellipses for $\mu = \binom{3}{2}$ and $\Sigma = \left(\begin{smallmatrix} 1 & -1.5 \\ -1.5 & 4 \end{smallmatrix} \right)$ ⚠
MVAcontnorm

According to Theorem 2.7 in Sect. 2.6 the half-lengths of the axes in the contour ellipsoid are $\sqrt{d^2 \lambda_i}$ where λ_i are the eigenvalues of Σ. If Σ is a diagonal matrix, the rectangle circumscribing the contour ellipse has sides with length $2d\sigma_i$ and is thus naturally proportional to the standard deviations of X_i ($i = 1, 2$).

The distribution of the quadratic form in (4.51) is given in the next theorem.

Theorem 4.7 *If $X \sim N_p(\mu, \Sigma)$, then the variable $U = (X - \mu)^\top \Sigma^{-1}(X - \mu)$ has a χ_p^2 distribution.*

Theorem 4.8 *The characteristic function (cf) of a multinormal $N_p(\mu, \Sigma)$ is given by*

$$\varphi_X(t) = \exp\left(i\, t^\top \mu - \frac{1}{2} t^\top \Sigma t \right). \tag{4.52}$$

We can check Theorem 4.8 by transforming the cf back:

$$f(x) = \frac{1}{(2\pi)^p} \int \exp\left(-\mathbf{i} t^\top x + \mathbf{i} t^\top \mu - \frac{1}{2} t^\top \Sigma t \right) dt$$

$$= \frac{1}{|2\pi \Sigma^{-1}|^{1/2} |2\pi \Sigma|^{1/2}} \int \exp\left[-\frac{1}{2} \{ t^\top \Sigma t + 2\mathbf{i} t^\top (x - \mu) - (x - \mu)^\top \Sigma^{-1}(x - \mu) \} \right]$$

$$\times \exp\left[-\frac{1}{2}\{(x-\mu)^{\top}\Sigma^{-1}(x-\mu)\}\right] dt$$

$$= \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left[-\frac{1}{2}\{(x-\mu)^{\top}\Sigma(x-\mu)\}\right]$$

since

$$\int \frac{1}{|2\pi\Sigma^{-1}|^{1/2}} \exp\left[-\frac{1}{2}\{t^{\top}\Sigma t + 2\mathbf{i}t^{\top}(x-\mu) - (x-\mu)^{\top}\Sigma^{-1}(x-\mu)\}\right] dt$$

$$= \int \frac{1}{|2\pi\Sigma^{-1}|^{1/2}} \exp\left[-\frac{1}{2}\{(t+\mathbf{i}\Sigma^{-1}(x-\mu))^{\top}\Sigma(t+\mathbf{i}\Sigma^{-1}(x-\mu))\}\right] dt$$

$$= 1.$$

Note that if $Y \sim N_p(0, \mathcal{I}_p)$, then

$$\varphi_Y(t) = \exp\left(-\frac{1}{2}t^{\top}\mathcal{I}_p t\right) = \exp\left(-\frac{1}{2}\sum_{i=1}^{p} t_i^2\right)$$

$$= \varphi_{Y_1}(t_1)\cdot\ldots\cdot\varphi_{Y_p}(t_p)$$

which is consistent with (4.33).

Singular Normal Distribution

Suppose that we have $\operatorname{rank}(\Sigma) = k < p$, where p is the dimension of X. We define the (singular) density of X with the aid of the G-Inverse Σ^{-} of Σ,

$$f(x) = \frac{(2\pi)^{-k/2}}{(\lambda_1\cdots\lambda_k)^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^{\top}\Sigma^{-}(x-\mu)\right\} \tag{4.53}$$

where

1. x lies on the hyperplane $\mathcal{N}^{\top}(x-\mu) = 0$ with $\mathcal{N}(p \times (p-k)) : \mathcal{N}^{\top}\Sigma = 0$ and $\mathcal{N}^{\top}\mathcal{N} = \mathcal{I}_k$.
2. Σ^{-} is the G-Inverse of Σ, and $\lambda_1, \ldots, \lambda_k$ are the nonzero eigenvalues of Σ.

What is the connection to a multinormal with k-dimensions? If

$$Y \sim N_k(0, \Lambda_1) \text{ and } \Lambda_1 = \operatorname{diag}(\lambda_1, \ldots, \lambda_k), \tag{4.54}$$

then an orthogonal matrix $\mathcal{B}(p \times k)$ with $\mathcal{B}^{\top}\mathcal{B} = \mathcal{I}_k$ exists that means $X = \mathcal{B}Y + \mu$ where X has a singular pdf of the form (4.53).

Gaussian Copula

In Examples 4.3 and 4.4 we have introduced copulae. Another important copula is the *Gaussian* or *normal copula*,

$$C_\rho(u, v) = \int_{-\infty}^{\Phi_1^{-1}(u)} \int_{-\infty}^{\Phi_2^{-1}(v)} f_\rho(x_1, x_2) dx_2 dx_1 \; , \tag{4.55}$$

see Embrechts, McNeil, and Straumann (1999). In (4.55), f_ρ denotes the bivariate normal density function with correlation ρ for $n = 2$. The functions Φ_1 and Φ_2 in (4.55) refer to the corresponding one-dimensional standard normal cdfs of the margins.

In the case of vanishing correlation, $\rho = 0$, the Gaussian copula becomes

$$\begin{aligned}
C_0(u, v) &= \int_{-\infty}^{\Phi_1^{-1}(u)} f_{X_1}(x_1) dx_1 \int_{-\infty}^{\Phi_2^{-1}(v)} f_{X_2}(x_2) dx_2 \\
&= u\, v \\
&= \Pi(u, v) \; .
\end{aligned}$$

Summary

$\hookrightarrow$ The pdf of a p-dimensional multinormal $X \sim N_p(\mu, \Sigma)$ is

$$f(x) = |2\pi\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right\} \; .$$

The contour curves of a multinormal are ellipsoids with half-lengths proportional to $\sqrt{\lambda_i}$, where λ_i denotes the eigenvalues of Σ ($i = 1, \ldots, p$).

$\hookrightarrow$ The Mahalanobis transformation transforms $X \sim N_p(\mu, \Sigma)$ to $Y = \Sigma^{-1/2}(X - \mu) \sim N_p(0, \mathcal{I}_p)$. Going in the other direction, one can create a $X \sim N_p(\mu, \Sigma)$ from $Y \sim N_p(0, \mathcal{I}_p)$ via $X = \Sigma^{1/2}Y + \mu$.

$\hookrightarrow$ If the covariance matrix Σ is singular (i.e. rank(Σ) $< p$), then it defines a singular normal distribution.

Summary (continued)
$\hookrightarrow$ The Gaussian copula is given by $$C_\rho(u, v) = \int_{-\infty}^{\Phi_1^{-1}(u)} \int_{-\infty}^{\Phi_2^{-1}(v)} f_\rho(x_1, x_2) dx_2 dx_1 \ .$$
$\hookrightarrow$ The density of a singular normal distribution is given by $$\frac{(2\pi)^{-k/2}}{(\lambda_1 \cdots \lambda_k)^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^-(x-\mu)\right\} \ .$$

4.5   Sampling Distributions and Limit Theorems

In multivariate statistics, we observe the values of a multivariate random variable X and obtain a sample $\{x_i\}_{i=1}^n$, as described in Chap. 3. Under random sampling, these observations are considered to be realisations of a sequence of i.i.d. random variables $X_1, \ldots, X_n$, where each X_i is a p-variate random variable which replicates the *parent* or *population* random variable X. Some notational confusion is hard to avoid: X_i is not the ith component of X, but rather the ith replicate of the p-variate random variable X which provides the ith observation x_i of our sample.

For a given random sample $X_1, \ldots, X_n$, the idea of statistical inference is to analyse the properties of the population variable X. This is typically done by analysing some characteristic θ of its distribution, like the mean, covariance matrix, etc. Statistical inference in a multivariate setup is considered in more detail in Chaps. 6 and 7.

Inference can often be performed using some observable function of the sample $X_1, \ldots, X_n$, i.e. a *statistics*. Examples of such statistics were given in Chap. 3: the sample mean $\overline{x}$, the sample covariance matrix $\mathcal{S}$. To get an idea of the relationship between a statistics and the corresponding population characteristic, one has to derive the sampling distribution of the statistic. The next example gives some insight into the relation of $(\overline{x}, S)$ to (μ, Σ).

Example 4.15 Consider an iid sample of n random vectors $X_i \in \mathbb{R}^p$ where $\mathsf{E}(X_i) = \mu$ and $\mathsf{Var}(X_i) = \Sigma$. The sample mean $\overline{x}$ and the covariance matrix $\mathcal{S}$ have already been defined in Sect. 3.3. It is easy to prove the following results:

$$\mathsf{E}(\overline{x}) \ = n^{-1} \sum_{i=1}^n \mathsf{E}(X_i) = \mu$$

$$\mathsf{Var}(\overline{x}) = n^{-2} \sum_{i=1}^n \mathsf{Var}(X_i) = n^{-1}\Sigma = \mathsf{E}(\overline{x}\,\overline{x}^\top) - \mu\mu^\top$$

$$\mathsf{E}(\mathcal{S}) = n^{-1}\,\mathsf{E}\left\{\sum_{i=1}^{n}(X_i - \overline{x})(X_i - \overline{x})^{\top}\right\}$$

$$= n^{-1}\,\mathsf{E}\left\{\sum_{i=1}^{n} X_i X_i^{\top} - n\,\overline{x}\,\overline{x}^{\top}\right\}$$

$$= n^{-1}\left\{n\left(\Sigma + \mu\mu^{\top}\right) - n\left(n^{-1}\Sigma + \mu\mu^{\top}\right)\right\}$$

$$= \tfrac{n-1}{n}\Sigma.$$

This shows in particular that $\mathcal{S}$ is a biased estimator of Σ. By contrast, $\mathcal{S}_u = \frac{n}{n-1}\mathcal{S}$ is an unbiased estimator of Σ.

Statistical inference often requires more than just the mean and/or the variance of a statistic. We need the sampling distribution of the statistics to derive confidence intervals or to define rejection regions in hypothesis testing for a given significance level. Theorem 4.9 gives the distribution of the sample mean for a multinormal population.

Theorem 4.9 *Let $X_1, \ldots, X_n$ be i.i.d. with $X_i \sim N_p(\mu, \Sigma)$. Then $\overline{x} \sim N_p(\mu, n^{-1}\Sigma)$.*

Proof $\overline{x} = n^{-1}\sum_{i=1}^{n} X_i$ is a linear combination of independent normal variables, so it has a normal distribution (see Chap. 5). The mean and the covariance matrix were given in the preceding example. □

With multivariate statistics, the sampling distributions of the statistics are often more difficult to derive than in the preceding Theorem. In addition they might be so complicated that approximations have to be used. These approximations are provided by limit theorems. Since they are based on asymptotic limits, the approximations are only valid when the sample size is large enough. In spite of this restriction, they make complicated situations rather simple. The following central limit theorem shows that even if the parent distribution is not normal, when the sample size n is large, the sample mean $\overline{x}$ has an approximate normal distribution.

Theorem 4.10 (Central Limit Theorem (CLT)) *Let $X_1, X_2, \ldots, X_n$ be i.i.d. with $X_i \sim (\mu, \Sigma)$. Then the distribution of $\sqrt{n}(\overline{x} - \mu)$ is asymptotically $N_p(0, \Sigma)$, i.e.*

$$\sqrt{n}(\overline{x} - \mu) \xrightarrow{\mathcal{L}} N_p(0, \Sigma) \qquad as \quad n \longrightarrow \infty.$$

The symbol "$\xrightarrow{\mathcal{L}}$" denotes *convergence in distribution* which means that the distribution function of the random vector $\sqrt{n}(\overline{x} - \mu)$ converges to the distribution function of $N_p(0, \Sigma)$.

Example 4.16 Assume that $X_1, \ldots, X_n$ are i.i.d. and that they have Bernoulli distributions where $p = \frac{1}{2}$ (this means that $P(X_i = 1) = \frac{1}{2}, \ P(X_i = 0) = \frac{1}{2}$). Then $\mu = p = \frac{1}{2}$ and $\Sigma = p(1-p) = \frac{1}{4}$. Hence,

$$\sqrt{n}\left(\overline{x} - \frac{1}{2}\right) \xrightarrow{\mathcal{L}} N_1\left(0, \frac{1}{4}\right) \qquad \text{as} \quad n \longrightarrow \infty.$$

The results are shown in Fig. 4.4 for varying sample sizes.

Example 4.17 Now consider a two-dimensional random sample $X_1, \ldots, X_n$ that is i.i.d. and created from two independent Bernoulli distributions with $p = 0.5$. The joint distribution is given by $P(X_i = (0, 0)^{\top}) = \frac{1}{4}, P(X_i = (0, 1)^{\top}) = \frac{1}{4}, P(X_i = (1, 0)^{\top}) = \frac{1}{4}, P(X_i = (1, 1)^{\top}) = \frac{1}{4}$. Here we have

$$\sqrt{n}\left\{\overline{x} - \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}\right\} = N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{4} \end{pmatrix}\right) \qquad \text{as} \quad n \longrightarrow \infty.$$

Figure 4.5 displays the estimated two-dimensional density for different sample sizes.

The asymptotic normal distribution is often used to construct confidence intervals for the unknown parameters. A confidence interval at the level $1 - \alpha$, $\alpha \in (0, 1)$, is an interval that covers the true parameter with probability $1 - \alpha$:

$$P(\theta \in [\widehat{\theta}_l, \widehat{\theta}_u]) = 1 - \alpha,$$

where θ denotes the (unknown) parameter and $\widehat{\theta}_l$ and $\widehat{\theta}_u$ are the lower and upper confidence bounds, respectively.

Example 4.18 Consider the i.i.d. random variables $X_1, \ldots, X_n$ with $X_i \sim (\mu, \sigma^2)$ and σ^2 known. Since we have $\sqrt{n}(\overline{x} - \mu) \xrightarrow{\mathcal{L}} N(0, \sigma^2)$ from the CLT, it follows that

$$P\left(-u_{1-\alpha/2} \leq \sqrt{n}\frac{(\overline{x} - \mu)}{\sigma} \leq u_{1-\alpha/2}\right) \longrightarrow 1 - \alpha, \qquad \text{as} \quad n \longrightarrow \infty$$

where $u_{1-\alpha/2}$ denotes the $(1 - \alpha/2)$-quantile of the standard normal distribution. Hence the interval

$$\left[\overline{x} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, \overline{x} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}\right]$$

is an approximate $(1 - \alpha)$-confidence interval for μ.

But what can we do if we do not know the variance σ^2? The following corollary gives the answer.

Fig. 4.4 The CLT for
Bernoulli distributed random
variables. Sample size $n = 5$
(*up*) and $n = 35$ (*down*) Q
`MVAcltbern`

Asymptotic Distribution, n=5

Asymptotic Distribution, n=35
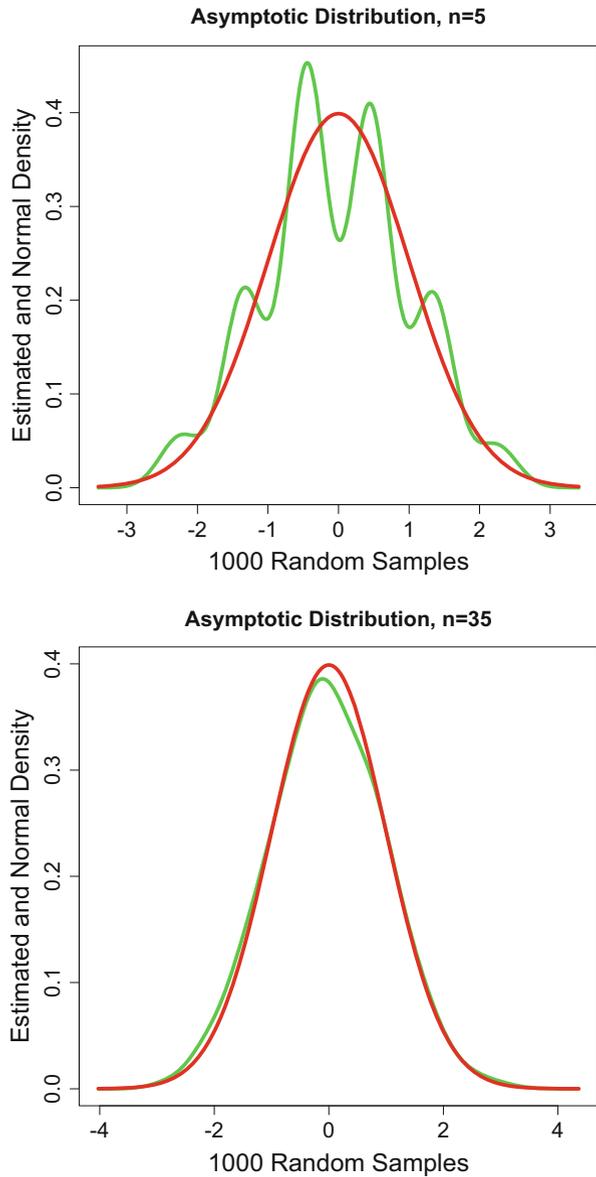
Corollary 4.1 *If $\hat{\Sigma}$ is a consistent estimate for Σ, then the CLT still holds, namely*

$$\sqrt{n}\; \hat{\Sigma}^{-1/2}(\overline{x} - \mu) \xrightarrow{\mathcal{L}} N_p(0, \mathcal{I}) \qquad as \quad n \longrightarrow \infty.$$
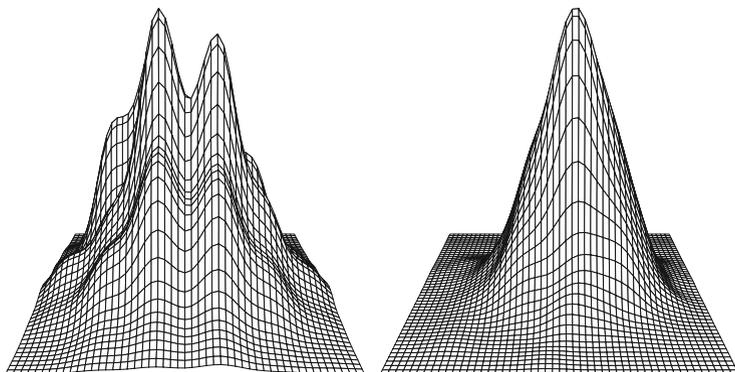
Fig. 4.5 The CLT in the two-dimensional case. Sample size $n = 5$ (*left*) and $n = 85$ (*right*) Q
MVAcltbern2

Example 4.19 Consider the i.i.d. random variables $X_1, \ldots, X_n$ with $X_i \sim (\mu, \sigma^2)$, and now with an unknown variance σ^2. From Corollary 4.1 using $\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$ we obtain

$$\sqrt{n} \left(\frac{\overline{x} - \mu}{\widehat{\sigma}} \right) \xrightarrow{\mathcal{L}} N(0, 1) \qquad \text{as} \quad n \longrightarrow \infty.$$

Hence we can construct an approximate $(1 - \alpha)$-confidence interval for μ using the variance estimate $\widehat{\sigma}^2$:

$$C_{1-\alpha} = \left[\overline{x} - \frac{\widehat{\sigma}}{\sqrt{n}} u_{1-\alpha/2}, \ \overline{x} + \frac{\widehat{\sigma}}{\sqrt{n}} u_{1-\alpha/2} \right].$$

Note that by the CLT

$$P(\mu \in C_{1-\alpha}) \longrightarrow 1 - \alpha \qquad \text{as} \quad n \longrightarrow \infty.$$

Remark 4.1 One may wonder how large should n be in practice to provide reasonable approximations. There is no definite answer to this question: it mainly depends on the problem at hand (the shape of the distribution of the X_i and the dimension of X_i). If the X_i are normally distributed, the normality of $\overline{x}$ is achieved from $n = 1$. In most situations, however, the approximation is valid in one-dimensional problems for n larger than, say, 50.

Transformation of Statistics

Often in practical problems, one is interested in a function of parameters for which one has an asymptotically normal statistic. Suppose for instance that we are interested in a cost function depending on the mean μ of the process: $f(\mu) = \mu^\top \mathcal{A} \mu$ where $\mathcal{A} > 0$ is given. To estimate μ we use the asymptotically normal statistic $\overline{x}$. The question is: how does $f(\overline{x})$ behave? More generally, what happens to a statistic t that is asymptotically normal when we transform it by a function $f(t)$? The answer is given by the following theorem.

Theorem 4.11 *If $\sqrt{n}(t - \mu) \overset{\mathcal{L}}{\longrightarrow} N_p(0, \Sigma)$ and if $f = (f_1, \ldots, f_q)^\top : \mathbb{R}^p \to \mathbb{R}^q$ are real valued functions which are differentiable at $\mu \in \mathbb{R}^p$, then $f(t)$ is asymptotically normal with mean $f(\mu)$ and covariance $\mathcal{D}^\top \Sigma \mathcal{D}$, i.e.*

$$\sqrt{n}\{f(t) - f(\mu)\} \overset{\mathcal{L}}{\longrightarrow} N_q(0, \mathcal{D}^\top \Sigma \mathcal{D}) \qquad for \quad n \longrightarrow \infty, \qquad (4.56)$$

where

$$\mathcal{D} = \left(\frac{\partial f_j}{\partial t_i} \right)(t) \Big|_{t=\mu}$$

is the $(p \times q)$ matrix of all partial derivatives.

Example 4.20 We are interested in seeing how $f(\overline{x}) = \overline{x}^\top \mathcal{A} \overline{x}$ behaves asymptotically with respect to the quadratic cost function of μ, $f(\mu) = \mu^\top \mathcal{A} \mu$, where $\mathcal{A} > 0$.

$$D = \frac{\partial f(\overline{x})}{\partial \overline{x}} \Big|_{\overline{x}=\mu} = 2\mathcal{A}\mu.$$

By Theorem 4.11 we have

$$\sqrt{n}(\overline{x}^\top \mathcal{A} \overline{x} - \mu^\top \mathcal{A} \mu) \overset{\mathcal{L}}{\longrightarrow} N_1 \left(0, 4\mu^\top \mathcal{A} \Sigma \mathcal{A} \mu\right).$$

Example 4.21 Suppose

$$X_i \sim (\mu, \Sigma); \quad \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \quad p = 2.$$

We have by the CLT (Theorem 4.10) for $n \to \infty$ that

$$\sqrt{n}(\overline{x} - \mu) \overset{\mathcal{L}}{\longrightarrow} N(0, \Sigma).$$

Suppose that we would like to compute the distribution of $\left(\begin{array}{c} \overline{x}_1^2 - \overline{x}_2 \\ \overline{x}_1 + 3\overline{x}_2 \end{array} \right)$. According

to Theorem 4.11 we have to consider $f = (f_1, f_2)^\top$ with

$$f_1(x_1, x_2) = x_1^2 - x_2, \quad f_2(x_1, x_2) = x_1 + 3x_2, \quad q = 2.$$

Given this $f(\mu) = \binom{0}{0}$ and

$$\mathcal{D} = (d_{ij}), \quad d_{ij} = \left(\frac{\partial f_j}{\partial x_i} \right)\Bigg|_{x=\mu} = \left(\begin{array}{cc} 2x_1 & 1 \\ -1 & 3 \end{array} \right)\Bigg|_{x=0}.$$

Thus

$$\mathcal{D} = \left(\begin{array}{cc} 0 & 1 \\ -1 & 3 \end{array} \right).$$

The covariance is

$$\underbrace{\left(\begin{array}{cc} 0 & -1 \\ 1 & 3 \end{array} \right)}_{\mathcal{D}^\top} \underbrace{\left(\begin{array}{cc} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{array} \right)}_{\Sigma} \underbrace{\left(\begin{array}{cc} 0 & 1 \\ -1 & 3 \end{array} \right)}_{\mathcal{D}} = \underbrace{\left(\begin{array}{cc} 0 & -1 \\ 1 & 3 \end{array} \right)}_{\mathcal{D}^\top} \underbrace{\left(\begin{array}{cc} -\frac{1}{2} & \frac{5}{2} \\ -1 & \frac{7}{2} \end{array} \right)}_{\Sigma \mathcal{D}} = \underbrace{\left(\begin{array}{cc} 1 & -\frac{7}{2} \\ -\frac{7}{2} & 13 \end{array} \right)}_{\mathcal{D}^\top \Sigma \mathcal{D}},$$

which yields

$$\sqrt{n} \left(\begin{array}{c} \overline{x}_1^2 - \overline{x}_2 \\ \overline{x}_1 + 3\overline{x}_2 \end{array} \right) \xrightarrow{\mathcal{L}} N_2 \left(\binom{0}{0}, \left(\begin{array}{cc} 1 & -\frac{7}{2} \\ -\frac{7}{2} & 13 \end{array} \right) \right).$$

Example 4.22 Let us continue the previous example by adding one more component to the function f. Since $q = 3 > p = 2$, we might expect a singular normal distribution. Consider $f = (f_1, f_2, f_3)^\top$ with

$$f_1(x_1, x_2) = x_1^2 - x_2, \quad f_2(x_1, x_2) = x_1 + 3x_2, \quad f_3 = x_2^3, \quad q = 3.$$

From this we have that

$$\mathcal{D} = \left(\begin{array}{ccc} 0 & 1 & 0 \\ -1 & 3 & 0 \end{array} \right) \quad \text{and thus} \quad \mathcal{D}^\top \Sigma \mathcal{D} = \left(\begin{array}{ccc} 1 & -\frac{7}{2} & 0 \\ -\frac{7}{2} & 13 & 0 \\ 0 & 0 & 0 \end{array} \right).$$

The limit is in fact a singular normal distribution!

	## Summary

| |
|---|
| ↪   If $X_1, \ldots, X_n$ are i.i.d. random vectors with $X_i \sim N_p(\mu, \Sigma)$, then $\bar{x} \sim N_p(\mu, \frac{1}{n}\Sigma)$. |
| ↪   If $X_1, \ldots, X_n$ are i.i.d. random vectors with $X_i \sim (\mu, \Sigma)$, then the distribution of $\sqrt{n}(\overline{x}-\mu)$ is asymptotically $N(0, \Sigma)$ (Central Limit Theorem). |
| ↪   If $X_1, \ldots, X_n$ are i.i.d. random variables with $X_i \sim (\mu, \sigma)$, then an asymptotic confidence interval can be constructed by the CLT: $\overline{x} \pm \frac{\hat{\sigma}}{\sqrt{n}} u_{1-\alpha/2}$. |
| ↪   If t is a statistic that is asymptotically normal, i.e. $\sqrt{n}(t - \mu) \xrightarrow{\mathcal{L}} N_p(0, \Sigma)$, then this holds also for a function $f(t)$, i.e. $\sqrt{n}\{f(t) - f(\mu)\}$ is asymptotically normal. |

4.6   Heavy-Tailed Distributions

Heavy-tailed distributions were first introduced by the Italian-born Swiss economist Pareto and extensively studied by Paul Lévy. Although in the beginning these distributions were mainly studied theoretically, nowadays they have found many applications in areas as diverse as finance, medicine, seismology, structural engineering. More concretely, they have been used to model returns of assets in financial markets, stream flow in hydrology, precipitation and hurricane damage in meteorology, earthquake prediction in seismology, pollution, material strength, teletraffic and many others.

A distribution is called heavy-tailed if it has higher probability density in its tail area compared with a normal distribution with same mean μ and variance σ^2. Figure 4.6 demonstrates the differences of the pdf curves of a standard Gaussian distribution and a Cauchy distribution with location parameter $\mu = 0$ and scale parameter $\sigma = 1$. The graphic shows that the probability density of the Cauchy distribution is much higher than that of the Gaussian in the tail part, while in the area around the centre, the probability density of the Cauchy distribution is much lower.

In terms of kurtosis, a heavy-tailed distribution has kurtosis greater than 3 (see Chap. 4, formula (4.40)), which is called leptokurtic, in contrast to mesokurtic distribution (kurtosis = 3) and platykurtic distribution (kurtosis < 3). Since univariate heavy-tailed distributions serve as basics for their multivariate counterparts and their density properties have been proved useful even in multivariate cases, we will start from introducing some univariate heavy-tailed distributions. Then we will move on to analyse their multivariate counterparts and their tail behaviour.

Distribution Comparison
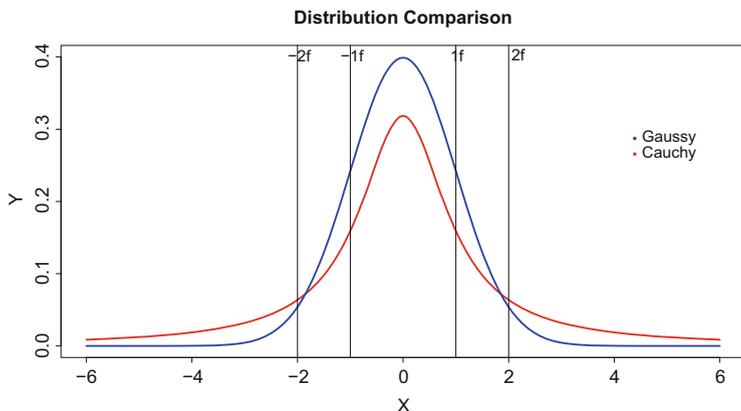
Fig. 4.6 Comparison of the pdf of a standard Gaussian (*blue*) and a Cauchy distribution (*red*) with location parameter 0 and scale parameter 1 ☖ MVAgausscauchy
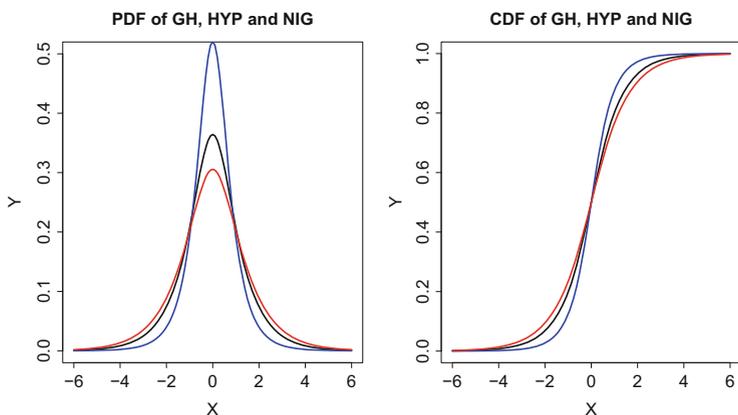
Fig. 4.7 pdf (*left*) and cdf (*right*) of GH ($\lambda = 0.5$, *black*), HYP (*red*), and NIG (*blue*) with $\alpha = 1, \beta = 0, \delta = 1, \mu = 0$ ☖ MVAghdis

Generalised Hyperbolic Distribution

The generalised hyperbolic distribution was introduced by Barndorff-Nielsen and at first applied to model grain size distributions of wind blown sands. Today one of its most important uses is in stock price modelling and market risk measurement. The name of the distribution is derived from the fact that its log-density forms a hyperbola, while the log-density of the normal distribution is a parabola (Fig. 4.7).

The density of a one-dimensional generalised hyperbolic (GH) distribution for $x \in \mathbb{R}$ is

$$
\begin{aligned}
&f_{\text{GH}}(x; \lambda, \alpha, \beta, \delta, \mu) \\
&= \frac{\left(\sqrt{\alpha^2 - \beta^2}/\delta\right)^{\lambda}}{\sqrt{2\pi} K_{\lambda}(\delta\sqrt{\alpha^2 - \beta^2})} \frac{K_{\lambda-1/2}\{\alpha\sqrt{\delta^2 + (x-\mu)^2}\}}{\sqrt{\delta^2 + (x-\mu)^2/\alpha})^{1/2-\lambda}} e^{\beta(x-\mu)}
\end{aligned} \quad (4.57)
$$

where K_{λ} is a modified Bessel function of the third kind with index λ

$$
K_{\lambda}(x) = \frac{1}{2} \int_0^{\infty} y^{\lambda-1} e^{-\frac{x}{2}(y+y^{-1})} dy \quad (4.58)
$$

The domain of variation of the parameters is $\mu \in \mathbb{R}$ and

$$
\begin{aligned}
\delta \geq 0, |\beta| < \alpha, &\quad \text{if} \quad \lambda > 0 \\
\delta > 0, |\beta| < \alpha, &\quad \text{if} \quad \lambda = 0 \\
\delta > 0, |\beta| \leq \alpha, &\quad \text{if} \quad \lambda < 0
\end{aligned}
$$

The generalised hyperbolic distribution has the following mean and variance

$$
\mathsf{E}[X] = \mu + \frac{\delta\beta}{\sqrt{\alpha^2 - \beta^2}} \frac{K_{\lambda+1}(\delta\sqrt{\alpha^2 - \beta^2})}{K_{\lambda}(\delta\sqrt{\alpha^2 - \beta^2})} \quad (4.59)
$$

$$
\begin{aligned}
\mathsf{Var}[X] = \delta^2 \Bigg[&\frac{K_{\lambda+1}(\delta\sqrt{\alpha^2 - \beta^2})}{\delta\sqrt{\alpha^2 - \beta^2} K_{\lambda}(\delta\sqrt{\alpha^2 - \beta^2})} + \frac{\beta^2}{\alpha^2 - \beta^2} \Bigg[\frac{K_{\lambda+2}(\delta\sqrt{\alpha^2 - \beta^2})}{K_{\lambda}(\delta\sqrt{\alpha^2 - \beta^2})} \\
&- \left\{ \frac{K_{\lambda+1}(\delta\sqrt{\alpha^2 - \beta^2})}{K_{\lambda}(\delta\sqrt{\alpha^2 - \beta^2})} \right\}^2 \Bigg] \Bigg]
\end{aligned} \quad (4.60)
$$

Where μ and δ play important roles in the density's location and scale respectively. With specific values of λ, we obtain different sub-classes of GH such as hyperbolic (HYP) or normal-inverse Gaussian (NIG) distribution.

For $\lambda = 1$ we obtain the hyperbolic distributions (HYP)

$$
f_{\text{HYP}}(x; \alpha, \beta, \delta, \mu) = \frac{\sqrt{\alpha^2 - \beta^2}}{2\alpha\delta K_1(\delta\sqrt{\alpha^2 - \beta^2})} e^{\{-\alpha\sqrt{\delta^2 + (x-\mu)^2} + \beta(x-\mu)\}} \quad (4.61)
$$

where $x, \mu \in \mathbb{R}, \delta \geq 0$ and $|\beta| < \alpha$. For $\lambda = -1/2$ we obtain the NIG distribution

$$
f_{\text{NIG}}(x; \alpha, \beta, \delta, \mu) = \frac{\alpha\delta}{\pi} \frac{K_1\left(\alpha\sqrt{(\delta^2 + (x-\mu)^2)}\right)}{\sqrt{\delta^2 + (x-\mu)^2}} e^{\{\delta\sqrt{\alpha^2 - \beta^2} + \beta(x-\mu)\}} \quad (4.62)
$$

Student's t-Distribution

The t-distribution was first analysed by Gosset (1908) who published it under pseudonym "Student" by request of his employer. Let X be a normally distributed random variable with mean μ and variance σ^2, and Y be the random variable such that Y^2/σ^2 has a chi-square distribution with n degrees of freedom. Assume that X and Y are independent, then

$$t \stackrel{\text{def}}{=} \frac{X\sqrt{n}}{Y} \tag{4.63}$$

is distributed as Student's t with n degrees of freedom. The t-distribution has the following density function

$$f_t(x;n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \tag{4.64}$$

where n is the number of degrees of freedom, $-\infty < x < \infty$, and Γ is the gamma function:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx. \tag{4.65}$$

The mean, variance, skewness and kurtosis of Student's t-distribution ($n > 4$) are:

$$\mu = 0$$
$$\sigma^2 = \frac{n}{n-2}$$
$$\text{Skewness} = 0$$
$$\text{Kurtosis} = 3 + \frac{6}{n-4}.$$

The t-distribution is symmetric around 0, which is consistent with the fact that its mean is 0 and skewness is also 0 (Fig. 4.8).

Student's t-distribution approaches the normal distribution as n increases, since

$$\lim_{n\to\infty} f_t(x;n) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}. \tag{4.66}$$

In practice the t-distribution is widely used, but its flexibility of modelling is restricted because of the integer-valued tail index.

In the tail area of the t-distribution, x is proportional to $|x|^{-(n+1)}$. In Fig. 4.13 we compared the tail-behaviour of t-distribution with different degrees of freedom. With higher degree of freedom, the t-distribution decays faster.
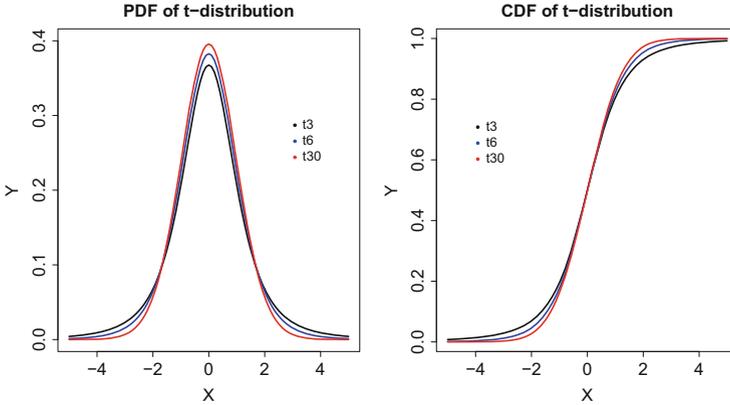
Fig. 4.8 pdf (*left*) and cdf (*right*) of *t*-distribution with different degrees of freedom (t3 stands for *t*-distribution with degree of freedom 3) ⓠ MVAtdis

Laplace Distribution

The univariate Laplace distribution with mean zero was introduced by Laplace (1774). The Laplace distribution can be defined as the distribution of differences between two independent variates with identical exponential distributions. Therefore it is also called the double exponential distribution (Fig. 4.9).

The Laplace distribution with mean μ and scale parameter θ has the pdf

$$f_{\text{Laplace}}(x; \mu, \theta) = \frac{1}{2\theta} e^{-\frac{|x-\mu|}{\theta}} \tag{4.67}$$

and the cdf

$$F_{\text{Laplace}}(x; \mu, \theta) = \frac{1}{2}\left\{1 + \text{sign}(x - \mu)(1 - e^{-\frac{|x-\mu|}{\theta}})\right\}, \tag{4.68}$$

where sign is sign function. The mean, variance, skewness and kurtosis of the Laplace distribution are

$$\mu = \mu$$
$$\sigma^2 = 2\theta^2$$
$$\text{Skewness} = 0$$
$$\text{Kurtosis} = 6$$

With mean 0 and $\theta = 1$, we obtain the standard Laplace distribution

PDF of Laplace distribution   **CDF of Laplace distribution**
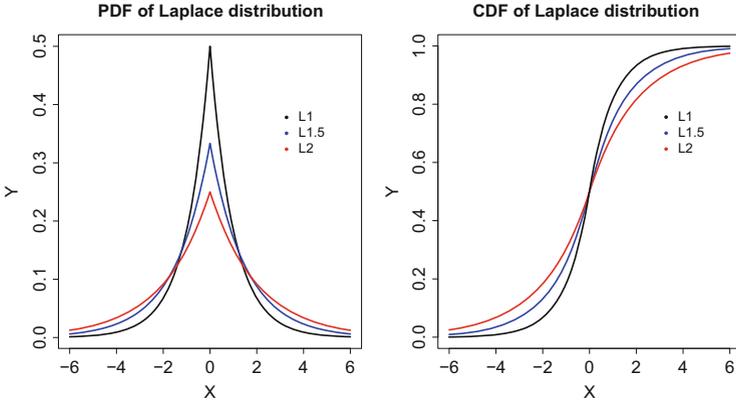
Fig. 4.9 pdf (*left*) and cdf (*right*) of Laplace distribution with zero mean and different scale parameters (L1 stands for Laplace distribution with $\theta = 1$) 🔍 MVAlaplacedis

$$f(x) = \frac{e^{-|x|}}{2} \tag{4.69}$$

$$F(x) = \begin{cases} \frac{e^x}{2} & \text{for} \quad x < 0 \\ 1 - \frac{e^{-x}}{2} & \text{for} \quad x \geq 0 \end{cases} \tag{4.70}$$

Cauchy Distribution

The Cauchy distribution is motivated by the following example.

Example 4.23 A gangster has just robbed a bank. As he runs to a point s metres away from the wall of the bank, a policeman reaches the crime scene behind the wall of the bank. The robber turns back and starts to shoot but he is such a poor shooter that the angle of his fire (marked in Fig. 4.10 as α) is uniformly distributed. The bullets hit the wall at distance x (from the centre). Obviously the distribution of x, the random variable where the bullet hits the wall, is of vital knowledge to the policeman in order to identify the location of the gangster. (Should the policeman calculate the mean or the median of the observed bullet hits $\{x_i\}_{i=1}^{n}$ in order to identify the location of the robber?)

Since α is uniformly distributed:

$$f(\alpha) = \frac{1}{\pi} I(\alpha \in [-\pi/2, \pi/2])$$

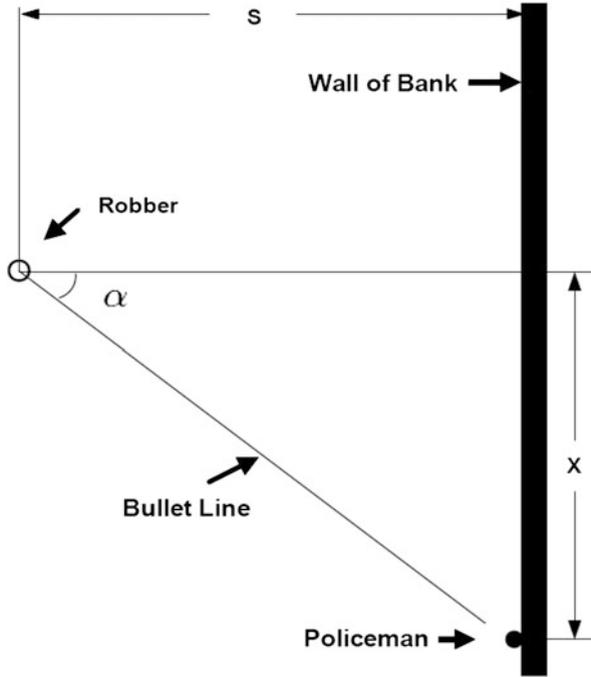**Fig. 4.10** Introduction to Cauchy distribution—robber vs. policeman

and

$$\tan \alpha = \frac{x}{s}$$

$$\alpha = \arctan\left(\frac{x}{s}\right)$$

$$d\alpha = \frac{1}{s} \frac{1}{1 + \left(\frac{x}{s}\right)^2} dx$$

For a small interval $d\alpha$, the probability is given by

$$f(\alpha)d\alpha = \frac{1}{\pi}d\alpha$$

$$= \frac{1}{s\pi} \frac{1}{1 + \left(\frac{x}{s}\right)^2} dx$$

with

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1}{\pi} d\alpha = 1$$

$$\int_{-\infty}^{\infty} \frac{1}{s\pi} \frac{1}{1+\left(\frac{x}{s}\right)^2} dx = \frac{1}{\pi} \left\{ \arctan\left(\frac{x}{s}\right) \right\}_{-\infty}^{\infty}$$

$$= \frac{1}{\pi} \left\{ \frac{\pi}{2} - \left(-\frac{\pi}{2}\right) \right\}$$

$$= 1$$

So the pdf of x can be written as:

$$f(x) = \frac{1}{s\pi} \frac{1}{1+\left(\frac{x}{s}\right)^2}$$

The general formula for the pdf and cdf of the Cauchy distribution is

$$f_{\text{Cauchy}}(x; m, s) = \frac{1}{s\pi} \frac{1}{1+\left(\frac{x-m}{s}\right)^2} \tag{4.71}$$

$$F_{\text{Cauchy}}(x; m, s) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{x-m}{s}\right) \tag{4.72}$$

where m and s are location and scale parameter respectively. The case in the above example where $m = 0$ and $s = 1$ is called the standard Cauchy distribution with pdf and cdf as following,

$$f_{\text{Cauchy}}(x) = \frac{1}{\pi(1+x^2)} \tag{4.73}$$

$$F_{\text{Cauchy}}(x; m, s) = \frac{1}{2} + \frac{\arctan(x)}{\pi} \tag{4.74}$$

The mean, variance, skewness and kurtosis of Cauchy distribution are all undefined, since its moment generating function diverges. But it has mode and median, both equal to the location parameter m (Fig. 4.11).

Mixture Model

Mixture modelling concerns modelling a statistical distribution by a mixture (or weighted sum) of different distributions. For many choices of component density functions, the mixture model can approximate any continuous density to arbitrary accuracy, provided that the number of component density functions is sufficiently large and the parameters of the model are chosen correctly. The pdf of a mixture distribution consists of n distributions and can be written as:
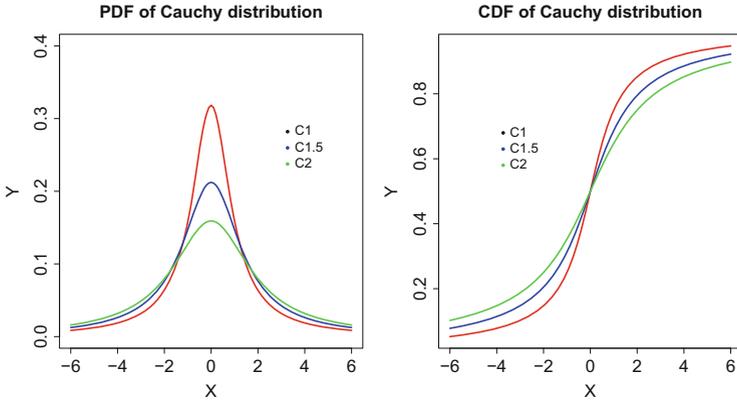
Fig. 4.11 pdf (*left*) and cdf (*right*) of Cauchy distribution with $m = 0$ and different scale parameters (C1 stands for Cauchy distribution with $s = 1$) 🔲 MVAcauchy

$$f(x) = \sum_{l=1}^{L} w_l \, p_l(x) \tag{4.75}$$

under the constraints:

$$0 \le w_l \le 1$$

$$\sum_{l=1}^{L} w_l = 1$$

$$\int p_l(x)dx = 1$$

where $p_l(x)$ is the pdf of the l'th component density and w_l is a weight. The mean, variance, skewness and kurtosis of a mixture are

$$\mu = \sum_{l=1}^{L} w_l \mu_l \tag{4.76}$$

$$\sigma^2 = \sum_{l=1}^{L} w_l \{\sigma_l^2 + (\mu_l - \mu)^2\} \tag{4.77}$$

$$\text{Skewness} = \sum_{l=1}^{L} w_l \left\{ \left(\frac{\sigma_l}{\sigma}\right)^3 SK_l + \frac{3\sigma_l^2(\mu_l - \mu)}{\sigma^3} + \left(\frac{\mu_l - \mu}{\sigma}\right)^3 \right\} \tag{4.78}$$

$$\text{Kurtosis} = \sum_{l=1}^{L} w_l \left\{ \left(\frac{\sigma_l}{\sigma} \right)^4 K_l + \frac{6(\mu_l - \mu)^2 \sigma_l^2}{\sigma^4} + \frac{4(\mu_l - \mu)\sigma_l^3}{\sigma^4} SK_l \right.$$

$$\left. + \left(\frac{\mu_l - \mu}{\sigma} \right)^4 \right\}, \tag{4.79}$$

where μ_l, σ_l, SK_l and K_l are respectively mean, variance, skewness and kurtosis of l'th distribution.

Mixture models are ubiquitous in virtually every facet of statistical analysis, machine learning and data mining. For data sets comprising continuous variables, the most common approach involves mixture distributions having Gaussian components.

The pdf for a Gaussian mixture is:

$$f_{\text{GM}}(x) = \sum_{l=1}^{L} \frac{w_l}{\sqrt{2\pi}\sigma_l} e^{-\frac{(x-\mu_l)^2}{2\sigma_l^2}}. \tag{4.80}$$

For a Gaussian mixture consisting of Gaussian distributions with mean 0, this can be simplified to:

$$f_{\text{GM}}(x) = \sum_{l=1}^{L} \frac{w_l}{\sqrt{2\pi}\sigma_l} e^{-\frac{x^2}{2\sigma_l^2}}, \tag{4.81}$$

with variance, skewness and kurtosis
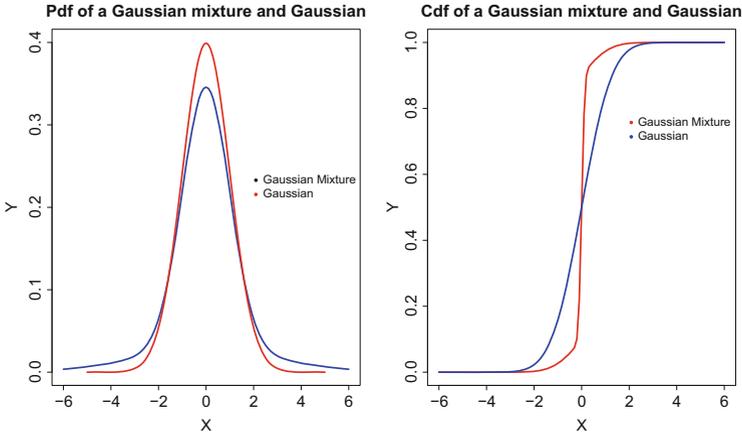
$$\sigma^2 = \sum_{l=1}^{L} w_l \sigma_l^2 \tag{4.82}$$

$$\text{Skewness} = 0 \tag{4.83}$$

$$\text{Kurtosis} = \sum_{l=1}^{L} w_l \left(\frac{\sigma_l}{\sigma} \right)^4 3 \tag{4.84}$$

Example 4.24 Consider a Gaussian Mixture which is 80 % $N(0,1)$ and 20 % $N(0,9)$. The pdf of $N(0,1)$ and $N(0,9)$ are (Fig. 4.12):

$$f_{N(0,1)}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$$f_{N(0,9)}(x) = \frac{1}{3\sqrt{2\pi}} e^{-\frac{x^2}{18}}$$

so the pdf of the Gaussian Mixture is

Fig. 4.12 pdf (*left*) and cdf (*right*) of a Gaussian mixture (Example 4.24) Q MVAmixture

Table 4.2 Basic statistics of t, Laplace and Cauchy distribution

	t	Laplace	Cauchy
Mean	0	μ	Not defined
Variance	$\frac{n}{n-2}$	$2\theta^2$	Not defined
Skewness	0	0	Not defined
Kurtosis	$3 + \frac{6}{n-4}$	6	Not defined

$$f_{\mathrm{GM}}(x) = \frac{1}{5\sqrt{2\pi}}\left(4e^{-\frac{x^2}{2}} + \frac{1}{3}e^{-\frac{x^2}{18}}\right)$$

Notice that the Gaussian Mixture is not a Gaussian distribution:

$$\mu = 0$$
$$\sigma^2 = 0.8 \times 1 + 0.2 \times 9 = 2.6$$
$$\text{Skewness} = 0$$
$$\text{Kurtosis} = 0.8 \times \left(\frac{1}{\sqrt{2.6}}\right)^4 \times 3 + 0.2 \times \left(\frac{\sqrt{9}}{\sqrt{2.6}}\right)^4 \times 3 = 7.54$$

The kurtosis of this Gaussian mixture is higher than 3.

A summary of the basic statistics is given in Tables 4.2 and 4.3.

Table 4.3 Basic statistics of GH distribution and mixture model

GH	
Mean	$\mu + \dfrac{\delta\beta}{\sqrt{\alpha^2+\beta^2}}\dfrac{K_{\lambda+1}(\delta\sqrt{\alpha^2+\beta^2})}{K_\lambda(\delta\sqrt{\alpha^2+\beta^2})}$
Variance	$\delta^2\left[\dfrac{K_{\lambda+1}(\delta\sqrt{\alpha^2+\beta^2})}{\delta\sqrt{\alpha^2+\beta^2}K_\lambda(\delta\sqrt{\alpha^2+\beta^2})} + \dfrac{\beta^2}{\alpha^2+\beta^2}\left[\dfrac{K_{\lambda+2}(\delta\sqrt{\alpha^2+\beta^2})}{K_\lambda(\delta\sqrt{\alpha^2+\beta^2})} - \left\{\dfrac{K_{\lambda+1}(\delta\sqrt{\alpha^2+\beta^2})}{K_\lambda(\delta\sqrt{\alpha^2+\beta^2})}\right\}^2\right]\right]$

Mixture	
Mean	$\sum_{l=1}^{L} w_l \mu_l$
Variance	$\sum_{l=1}^{L} w_l\{\sigma_l^2 + (\mu_l - \mu)^2\}$
Skewness	$\sum_{l=1}^{L} w_l\left\{\left(\dfrac{\sigma_l}{\sigma}\right)^3 SK_l + \dfrac{3\sigma_l^2(\mu_l-\mu)}{\sigma^3} + \left(\dfrac{\mu_l-\mu}{\sigma}\right)^3\right\}$
Kurtosis	$\sum_{l=1}^{L} w_l\left\{\left(\dfrac{\sigma_l}{\sigma}\right)^4 K_l + \dfrac{6(\mu_l-\mu)^2\sigma_l^2}{\sigma^4} + \dfrac{4(\mu_l-\mu)\sigma_l^3}{\sigma^4} SK_l + \left(\dfrac{\mu_l-\mu}{\sigma}\right)^4\right\}$

Multivariate Generalised Hyperbolic Distribution

The multivariate Generalised Hyperbolic Distribution (GH_d) has the following pdf

$$f_{GH_d}(x;\lambda,\alpha,\beta,\delta,\Delta,\mu) = a_d \frac{K_{\lambda-\frac{d}{2}}\left\{\alpha\sqrt{\delta^2 + (x-\mu)^\top \Delta^{-1}(x-\mu)}\right\}}{\left\{\alpha^{-1}\sqrt{\delta^2 + (x-\mu)^\top \Delta^{-1}(x-\mu)}\right\}^{\frac{d}{2}-\lambda}} e^{\beta^\top(x-\mu)}$$

(4.85)

$$a_d = a_d(\lambda,\alpha,\beta,\delta,\Delta) = \frac{\left(\sqrt{\alpha^2 - \beta^\top \Delta\beta}/\delta\right)^\lambda}{(2\pi)^{\frac{d}{2}} K_\lambda(\delta\sqrt{\alpha^2 - \beta^\top \Delta\beta})},$$

(4.86)

and characteristic function

$$\phi(t) = \left(\frac{\alpha^2 - \beta^\top \Delta\beta}{\alpha^2 - \beta^\top \Delta\beta + \frac{1}{2}t^\top \Delta t - i\beta^\top \Delta t}\right)^{\frac{\lambda}{2}}$$

$$\times \frac{K_\lambda\left(\delta\sqrt{\alpha^2 - \beta^\top \Delta\beta^\top + \frac{1}{2}t^\top \Delta t - i\beta^\top \Delta t}\right)}{K_\lambda\left(\delta\sqrt{\alpha^2 - \beta^\top \Delta\beta^\top}\right)}$$

(4.87)

These parameters have the following domain of variation:

$$
\begin{aligned}
&\lambda \in \mathbb{R}, \qquad \beta, \mu \in \mathbb{R}^d \\
&\delta > 0, \qquad \alpha > \beta^\top \Delta \beta \\
&\Delta \in \mathbb{R}^{d \times d} \text{ positive definite matrix} \\
&|\Delta| = 1
\end{aligned}
$$

For $\lambda = \frac{d+1}{2}$ we obtain the multivariate hyperbolic (HYP) distribution; for $\lambda = -\frac{1}{2}$ we get the multivariate normal inverse Gaussian (NIG) distribution.

Blæsild and Jensen (1981) introduced a second parameterisation (ζ, Π, Σ), where

$$
\zeta = \delta \sqrt{\alpha^2 - \beta^\top \Delta \beta} \tag{4.88}
$$

$$
\Pi = \beta \sqrt{\frac{\Delta}{\alpha^2 - \beta^\top \Delta \beta}} \tag{4.89}
$$

$$
\Sigma = \delta^2 \Delta \tag{4.90}
$$

The mean and variance of $X \sim \mathrm{GH}_d$

$$
\mathsf{E}[X] = \mu + \delta R_\lambda(\zeta) \Pi \Delta^{\frac{1}{2}} \tag{4.91}
$$

$$
\mathsf{Var}[X] = \delta^2 \{ \zeta^{-1} R_\lambda(\zeta) \Delta + S_\lambda(\zeta)(\Pi \Delta^{\frac{1}{2}})^\top (\Pi \Delta^{\frac{1}{2}}) \} \tag{4.92}
$$

where

$$
R_\lambda(x) = \frac{K_{\lambda+1}(x)}{K_\lambda(x)} \tag{4.93}
$$

$$
S_\lambda(x) = \frac{K_{\lambda+2}(x) K_\lambda(x) - K_{\lambda+1}^2(x)}{K_\lambda^2(x)} \tag{4.94}
$$

Theorem 4.12 *Suppose that X is a d-dimensional variate distributed according to the generalised hyperbolic distribution GH_d. Let (X_1, X_2) be a partitioning of X, let r and k denote the dimensions of X_1 and X_2, respectively, and let (β_1, β_2) and (μ_1, μ_2) be similar partitions of β and μ, let*

$$
\Delta = \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{pmatrix} \tag{4.95}
$$

be a partition of Δ such that Δ_{11} is a $r \times r$ matrix. Then one has the following

1. *The distribution of X_1 is the r-dimensional generalised hyperbolic distribution,*
   $\mathrm{GH}_r(\lambda^*, \alpha^*, \beta^*, \delta^*, \mu^*, \Delta^*)$, *where*

$$\lambda^* = \lambda$$

$$\alpha^* = |\Delta_{11}|^{-\frac{1}{2r}}\{\alpha^2 - \beta_2(\Delta_{22} - \Delta_{21}\Delta_{11}^{-1}\Delta_{12})\beta_2^\top\}^{\frac{1}{2}}$$

$$\beta^* = \beta_1 + \beta_2\Delta_{21}\Delta_{11}^{-1}$$

$$\delta^* = \delta|\Delta_{11}|^{\frac{1}{2\rho}}$$

$$\mu^* = \mu_1$$

$$\Delta^* = |\Delta|^{-\frac{1}{r}}\Delta_{11}$$

2. *The conditional distribution of X_2 given $X_1 = x_1$ is the k-dimensional generalised hyperbolic distribution $\mathrm{GH}_k(\tilde{\lambda}, \tilde{\alpha}, \tilde{\beta}, \tilde{\delta}, \tilde{\mu}, \tilde{\Delta})$, where*

$$\tilde{\lambda} = \lambda - \frac{r}{2}$$

$$\tilde{\alpha} = \alpha|\Delta_{11}|^{\frac{1}{2k}}$$

$$\tilde{\beta} = \beta_2$$

$$\tilde{\delta} = |\Delta_{11}|^{-\frac{1}{2k}}\{\delta^2 + (x_1 - \mu_1)\Delta_{11}^{-1}(x_1 - \mu_1)^\top\}^{\frac{1}{2}}$$

$$\tilde{\mu} = \mu_2 + (x_1 - \mu_1)\Delta_{11}^{-1}\Delta_{12}$$

$$\tilde{\Delta} = |\Delta_{11}|^{\frac{1}{k}}(\Delta_{22} - \Delta_{21}\Delta_{11}^{-1}\Delta_{12})$$

3. *Let $Y = XA + B$ be a regular affine transformation of X and let $||A||$ denote the absolute value of the determinant of A. The distribution of Y is the d-dimensional generalised hyperbolic distribution $\mathrm{GH}_d(\lambda^+, \alpha^+, \beta^+, \delta^+, \mu^+, \Delta^+)$, where*

$$\lambda^+ = \lambda$$

$$\alpha^+ = \alpha||A||^{-\frac{1}{d}}$$

$$\beta^+ = \beta(A^{-1})^\top$$

$$\delta^+ = ||A||^{\frac{1}{d}}$$

$$\mu^+ = \mu A + B$$

$$\Delta^+ = ||A||^{-\frac{2}{d}}A^\top\Delta A$$

Multivariate t-Distribution

If X and Y are independent and distributed as $N_p(\mu, \Sigma)$ and $\mathcal{X}_n^2$ respectively, and $X\sqrt{n/Y} = t - \mu$, then the pdf of t is given by

$$f_t(t; n, \Sigma, \mu) = \frac{\Gamma\{(n+p)/2\}}{\Gamma(n/2)n^{p/2}\pi^{p/2}|\Sigma|^{1/2}\{1 + \frac{1}{n}(t-\mu)^\top\Sigma^{-1}(t-\mu)\}^{(n+p)/2}}$$

(4.96)

The distribution of t is the noncentral t-distribution with n degrees of freedom and the noncentrality parameter μ, Giri (1996).

Multivariate Laplace Distribution

Let g and G be the pdf and cdf of a d-dimensional Gaussian distribution $N_d(0, \Sigma)$, the pdf and cdf of a multivariate Laplace distribution can be written as

$$f_{M\text{Laplace}_d}(x; m, \Sigma) = \int_0^\infty g(z^{-\frac{1}{2}}x - z^{\frac{1}{2}}m)z^{-\frac{d}{2}}e^{-z}dz$$

(4.97)

$$F_{M\text{Laplace}_d}(x, m, \Sigma) = \int_0^\infty G(z^{-\frac{1}{2}}x - z^{\frac{1}{2}}m)e^{-z}dz$$

(4.98)

the pdf can also be described as

$$f_{M\text{Laplace}_d}(x; m, \Sigma) = \frac{2e^{x^\top\Sigma^{-1}m}}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\left(\frac{x^\top\Sigma^{-1}x}{2 + m^\top\Sigma^{-1}m}\right)^{\frac{\lambda}{2}}$$
$$\times K_\lambda\left(\sqrt{(2 + m^\top\Sigma^{-1}m)(x^\top\Sigma^{-1}x)}\right)$$

(4.99)

where $\lambda = \frac{2-d}{2}$ and $K_\lambda(x)$ is the modified Bessel function of the third kind

$$K_\lambda(x) = \frac{1}{2}\left(\frac{x}{2}\right)^\lambda \int_0^\infty t^{-\lambda-1}e^{-t-\frac{x^2}{4t}}dt, \qquad x > 0$$

(4.100)

Multivariate Laplace distribution has mean and variance

$$\mathsf{E}[X] = m$$

(4.101)

$$\mathsf{Cov}[X] = \Sigma + mm^\top$$

(4.102)

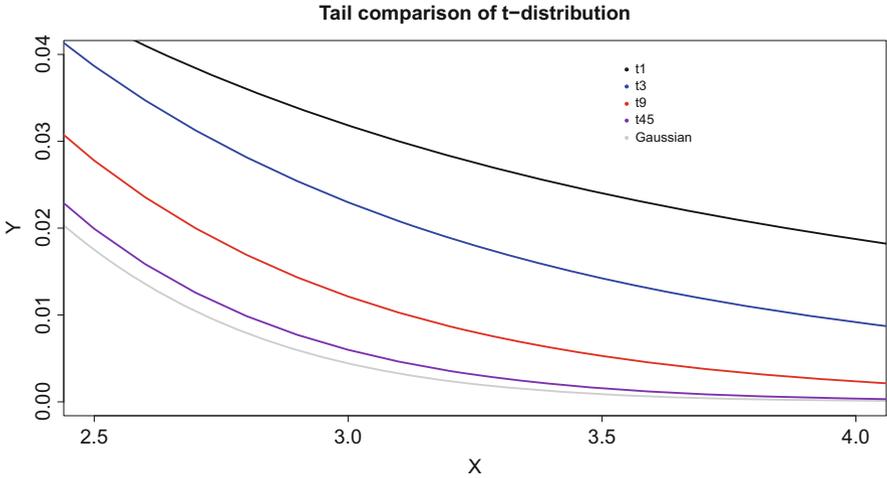**Tail comparison of t−distribution**

Fig. 4.13 Tail comparison of t-distribution ⊙ MVAtdistail

Multivariate Mixture Model

A multivariate mixture model comprises multivariate distributions, e.g. the pdf of a multivariate Gaussian distribution can be written as

$$f(x) = \sum_{l=1}^{L} \frac{w_l}{|2\pi \Sigma_l|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_l)^\top \Sigma^{-1}(x-\mu_l)} \tag{4.103}$$

Generalised Hyperbolic Distribution

The GH distribution has an exponential decaying speed

$$f_{\text{GH}}(x; \lambda, \alpha, \beta, \delta, \mu = 0) \sim x^{\lambda-1} e^{-(\alpha-\beta)x} \qquad \text{as} \quad x \to \infty, \tag{4.104}$$

As a comparison to tail behaviour of t-distribution depicted in Fig. 4.13, the Fig. 4.14 illustrates the tail behaviour of GH distributions with different value of λ with $\alpha = 1, \beta = 0, \delta = 1, \mu = 0$. It is clear that among the four distributions, GH with $\lambda = 1.5$ has the lowest decaying speed, while NIG decays faster.

In Fig. 4.15, Chen, Härdle, and Jeong (2008), four distributions and especially their tail-behaviour are compared. In order to keep the comparability of these distributions, we specified the means to 0 and standardised the variances to 1. Furthermore we used one important subclass of the GH distribution: the NIG
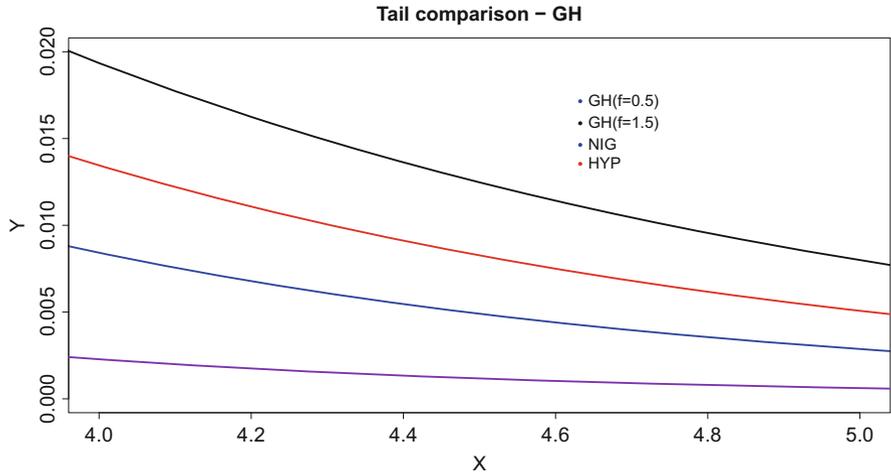
Tail comparison − GH

- GH(f=0.5)
- GH(f=1.5)
- NIG
- HYP

Fig. 4.14 Tail comparison of GH distribution (pdf) ⬛ MVAghdistail

Distribution comparison

- Laplace
- NIG
- Cauchy
- Gaussian

Tail comparison

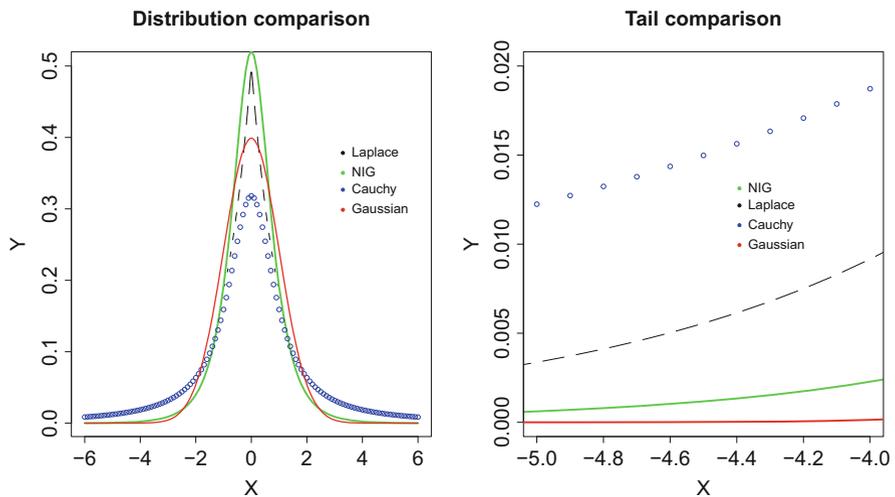- NIG
- Laplace
- Cauchy
- Gaussian

Fig. 4.15 Graphical comparison of the NIG distribution (*line*), standard normal distribution ⬛
MVAghadatail

distribution with $\lambda = -\frac{1}{2}$ introduced above. On the left panel, the complete forms
of these distributions are revealed. The Cauchy (dots) distribution has the lowest
peak and the fattest tails. In other words, it has the flattest distribution. The NIG
distribution decays second fast in the tails although it has the highest peak, which is
more clearly displayed on the right panel.

4.7   Copulae

The cumulative distribution function (cdf) of a two-dimensional vector (X_1, X_2) is given by

$$F(x_1, x_2) = \mathrm{P}(X_1 \le x_1, X_2 \le x_2). \tag{4.105}$$

For the case that X_1 and X_2 are independent, their joint cumulative distribution function $F(x_1, x_2)$ can be written as a product of their one-dimensional marginals:

$$F(x_1, x_2) = F_{X_1}(x_1)\, F_{X_2}(x_2) = \mathrm{P}(X_1 \le x_1)\,\mathrm{P}(X_2 \le x_2). \tag{4.106}$$

But how can we model dependence of X_1 and X_2? Most people would suggest linear correlation. Correlation is though an appropriate measure of dependence only when the random variables have an elliptical or spherical distribution, which include the normal multivariate distribution. Although the terms "correlation" and "dependency" are often used interchangeably, correlation is actually a rather imperfect measure of dependency, and there are many circumstances where correlation should not be used.

Copulae represent an elegant concept of connecting marginals with joint cumulative distribution functions. Copulae are functions that join or "couple" multivariate distribution functions to their one-dimensional marginal distribution functions. Let us consider a d-dimensional vector $X = (X_1, \ldots, X_d)^\top$. Using copulae, the marginal distribution functions $F_{X_i}(i = 1, \ldots, d)$ can be separately modelled from their dependence structure and then coupled together to form the multivariate distribution F_X. Copula functions have a long history in probability theory and statistics. Their application in finance is very recent. Copulae are important in Value-at-Risk calculations and constitute an essential tool in quantitative finance (Härdle et al., 2009).

First let us concentrate on the two-dimensional case, then we will extend this concept to the d-dimensional case, for a random variable in $\mathbb{R}^d$ with $d \ge 1$. To be able to define a copula function, first we need to represent a concept of the *volume of a rectangle, a 2-increasing function* and *a grounded function*.

Let U_1 and U_2 be two sets in $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$ and consider the function $F : U_1 \times U_2 \longrightarrow \overline{\mathbb{R}}$.

Definition 4.2 The F-volume of a rectangle $B = [x_1, x_2] \times [y_1, y_2] \subset U_1 \times U_2$ is defined as:

$$V_F(B) = F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1) \tag{4.107}$$

Definition 4.3 F is said to be a 2-increasing function if for every $B = [x_1, x_2] \times [y_1, y_2] \subset U_1 \times U_2$,

$$V_F(B) \geq 0 \tag{4.108}$$

Remark 4.2 Note that "to be 2-increasing function" neither implies nor is implied by "to be increasing in each argument".

The following lemmas (Nelsen, 1999) will be very useful later for establishing the continuity of copulae.

Lemma 4.1 *Let U_1 and U_2 be non-empty sets in $\mathbb{R}$ and let $F : U_1 \times U_2 \longrightarrow \mathbb{R}$ be a two-increasing function. Let x_1, x_2 be in U_1 with $x_1 \leq x_2$, and y_1, y_2 be in U_2 with $y_1 \leq y_2$. Then the function $t \mapsto F(t, y_2) - F(t, y_1)$ is non-decreasing on U_1 and the function $t \mapsto F(x_2, t) - F(x_1, t)$ is non-decreasing on U_2.*

Definition 4.4 If U_1 and U_2 have a smallest element $\min U_1$ and $\min U_2$ respectively, then we say that a function $F : U_1 \times U_2 \longrightarrow \mathbb{R}$ is grounded if :

$$\text{for all } x \in U_1 \ : \ F(x, \min U_2) = 0 \ \text{ and} \tag{4.109}$$

$$\text{for all } y \in U_2 \ : \ F(\min U_1, y) = 0 \tag{4.110}$$

In the following, we will refer to this definition of a cdf.

Definition 4.5 A cdf is a function from $\overline{\mathbb{R}}^2 \mapsto [0, 1]$ which

(i) is grounded
(ii) is 2-increasing
(iii) satisfies $F(\infty, \infty) = 1$

Lemma 4.2 *Let U_1 and U_2 be non-empty sets in $\mathbb{R}$ and let $F : U_1 \times U_2 \longrightarrow \mathbb{R}$ be a grounded two-increasing function. Then F is non-decreasing in each argument.*

Definition 4.6 If U_1 and U_2 have a greatest element $\max U_1$ and $\max U_2$ respectively, then we say that a function $F : U_1 \times U_2 \longrightarrow \mathbb{R}$ has margins and that the margins of F are given by:

$$F(x) = F(x, \max U_2) \text{ for all } x \in U_1 \tag{4.111}$$

$$F(y) = F(\max U_1, y) \text{ for all } y \in U_2 \tag{4.112}$$

Lemma 4.3 *Let U_1 and U_2 be non-empty sets in $\mathbb{R}$ and let $F : U_1 \times U_2 \longrightarrow \mathbb{R}$ be a grounded two-increasing function which has margins. Let $(x_1, y_1), (x_2, y_2) \in S_1 \times S_2$. Then*

$$|F(x_2, y_2) - F(x_1, y_1)| \leq |F(x_2) - F(x_1)| + |F(y_2) - F(y_1)| \tag{4.113}$$

Definition 4.7 A two-dimensional copula is a function C defined on the unit square $I^2 = I \times I$ with $I = [0, 1]$ such that

(i) for every $u \in I$ holds: $C(u, 0) = C(0, v) = 0$, i.e. C is grounded.
(ii) for every $u_1, u_2, v_1, v_2 \in I$ with $u_1 \leq u_2$ and $v_1 \leq v_2$ holds:

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0, \qquad (4.114)$$

i.e. C is 2-increasing.
(iii) for every $u \in I$ holds $C(u, 1) = u$ and $C(1, v) = v$.

Informally, a copula is a joint distribution function defined on the unit square $[0, 1]^2$ which has uniform marginals. That means that if $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$ are univariate distribution functions, then $C \{F_{X_1}(x_1), F_{X_2}(x_2)\}$ is a two-dimensional distribution function with marginals $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$.

Example 4.25 The functions $\max(u + v - 1, 0)$, uv, $\min(u, v)$ can be easily checked to be copula functions. They are called respectively the minimum, product and maximum copula.

Example 4.26 Consider the function

$$C_\rho^{\text{Gauss}}(u, v) = \Phi_\rho \{\Phi^{-1}(u), \Phi^{-1}(v)\} \qquad (4.115)$$

$$= \int_{-\infty}^{\Phi_1^{-1}(u)} \int_{-\infty}^{\Phi_2^{-1}(v)} f_\rho(x_1, x_2) dx_2 dx_1$$

where Φ_ρ is the joint two-dimensional standard normal distribution function with correlation coefficient ρ, while Φ_1 and Φ_2 refer to standard normal cdfs and
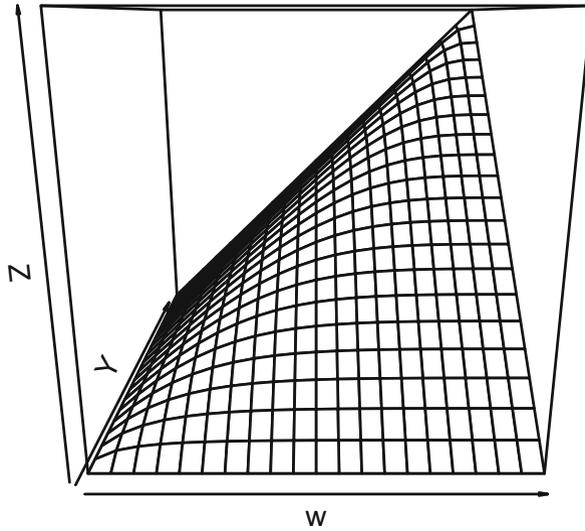
$$f_\rho(x_1, x_2) = \frac{1}{2\pi \sqrt{1 - \rho^2}} \exp \left\{ -\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{2(1 - \rho^2)} \right\} \qquad (4.116)$$

denotes the bivariate normal pdf.

It is easy to see that C^{Gauss} is a copula, the so-called Gaussian or normal copula, since it is 2-increasing and

$$\Phi_\rho \{\Phi^{-1}(u), \Phi^{-1}(0)\} = \Phi_\rho \{\Phi^{-1}(0), \Phi^{-1}(v)\} = 0 \qquad (4.117)$$

$$\Phi_\rho \{\Phi^{-1}(u), \Phi^{-1}(1)\} = u \text{ and } \Phi_\rho \{\Phi^{-1}(1), \Phi^{-1}(v)\} = v \qquad (4.118)$$

Fig. 4.16 Surface plot of the Gumbel–Hougaard copula, $\theta = 3$ Q MVAghsurface

A simple and useful way to represent the graph of a copula is the contour diagram that is, graphs of its level sets—the sets in I^2 given by $C(u, v) = $ a constant. In Figs. 4.16 and 4.17 we present the countour diagrams of the Gumbel–Hougard copula (Example 4.4) for different values of the copula parameter θ.

For $\theta = 1$ the Gumbel–Hougaard copula reduces to the product copula, i.e.

$$C_1(u, v) = \Pi(u, v) = uv \qquad (4.119)$$

For $\theta \to \infty$, one finds for the Gumbel–Hougaard copula:

$$C_\theta(u, v) \longrightarrow \min(u, v) = M(u, v) \qquad (4.120)$$

where M is also a copula such that $C(u, v) \leq M(u, v)$ for an arbitrary copula C. The copula M is called the Fréchet–Hoeffding upper bound.

The two-dimensional function $W(u, v) = \max(u + v - 1, 0)$ defines a copula with $W(u, v) \leq C(u, v)$ for any other copula C. W is called the Fréchet–Hoeffding lower bound.

In Fig. 4.18 we show an example of Gumbel–Hougaard copula sampling for fixed parameters $\sigma_1 = 1$, $\sigma_2 = 1$ and $\theta = 3$.

One can demonstrate the so-called Fréchet–Hoeffding inequality, which we have already used in Example 1.3, and which states that each copula function is bounded by the minimum and maximum one:

$$W(u, v) = \max(u + v - 1, 0) \leq C(u, v) \leq \min(u, v) = M(u, v) \quad (4.121)$$
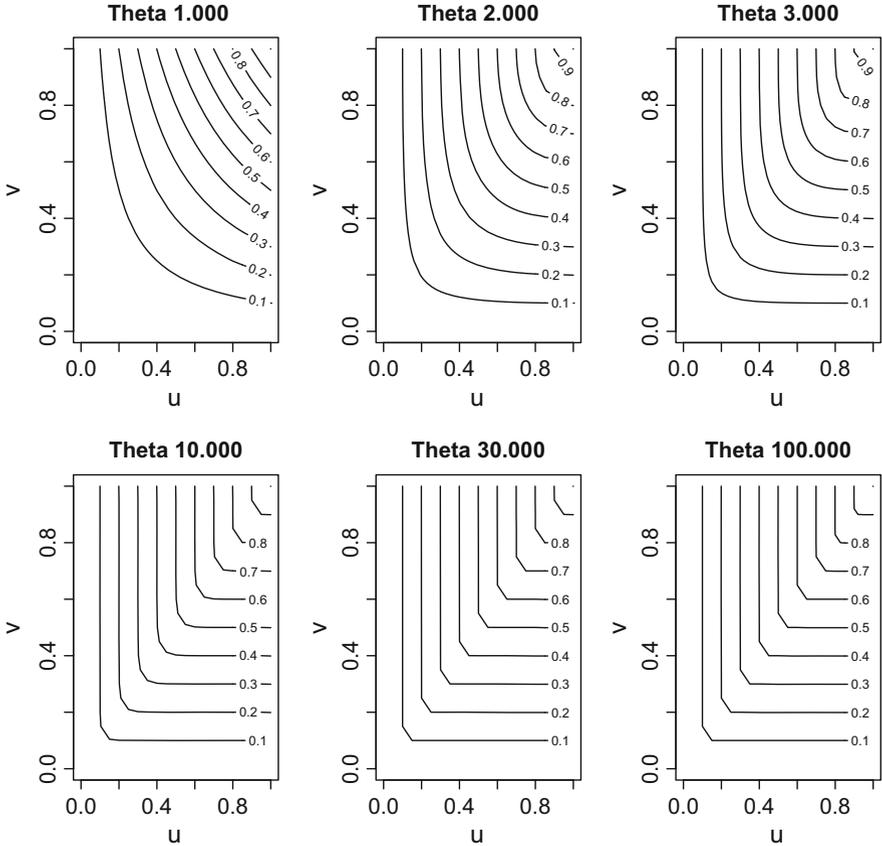
**Fig. 4.17** Contour plots of the Gumbel–Hougard copula 🅀 MVAghcontour

The full relationship between copula and joint cdf depends on Sklar theorem.

Example 4.27 Let us verify that the Gaussian copula satisfies Sklar's theorem in both directions. On the one side, let

$$F(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \frac{1}{2\pi \sqrt{1 - \rho^2}} \exp\left\{ -\frac{u_1^2 - 2\rho u_1 u_2 + u_2^2}{2(1 - \rho^2)} \right\} \, du_2 du_1.$$

(4.122)

be a two-dimensional normal distribution function with standard normal cdf's $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$. Since $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$ are continuous, a unique copula

Fig. 4.18 10,000-sample
output for $\sigma_1 = 1$, $\sigma_2 = 1$,
$\theta = 3$ 🔍 MVAsample1000
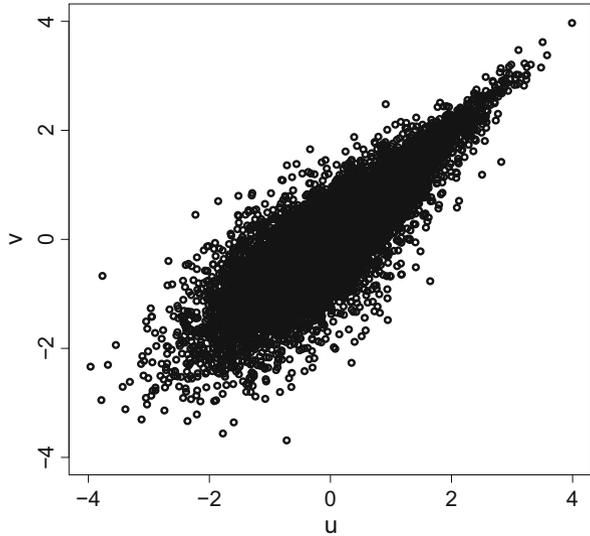
C exists such that for all x_1, $x_2 \in \overline{\mathbb{R}}^2$ a two-dimensional distribution function can be written as a copula in $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$:

$$F(x_1, x_2) = C\{\Phi_{X_1}(x_1), \Phi_{X_2}(x_2)\} \tag{4.123}$$

The Gaussian copula satisfies the above equality, therefore it is the unique copula mentioned in Sklar's theorem. This proves that the Gaussian copula, together with Gaussian marginals, gives the two-dimensional normal distribution.

Conversely, if C is a copula and F_{X_1} and F_{X_2} are standard normal distribution functions, then

$$C\{F_{X_1}(x_1), F_{X_2}(x_2)\} = \int_{-\infty}^{\phi_1^{-1}\{F_{X_1}(x_1)\}} \int_{-\infty}^{\phi_2^{-1}\{F_{X_2}(x_2)\}} \frac{1}{2\pi\sqrt{1-\rho^2}}$$

$$\times \exp\left\{-\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{2(1-\rho^2)}\right\} dx_2 dx_1 \tag{4.124}$$

is evidently a joint (two-dimensional) distribution function. Its margins are

$$C\{F_{X_1}(x_1), F_{X_2}(+\infty)\} = \Phi_\rho\left[\Phi^{-1}\{F_{X_1}(x_1)\}, +\infty\right] = F_{X_1}(x_1) \tag{4.125}$$

$$C\{F_{X_1}(+\infty), F_{X_2}(x_2)\} = \Phi_\rho\left[+\infty, \Phi^{-1}\{F_{X_2}(x_2)\}\right] = F_{X_2}(x_2) \tag{4.126}$$

The following proposition shows one attractive feature of the copula representation of dependence, i.e. that the dependence structure described by a copula

is invariant under increasing and continuous transformations of the marginal distributions.

Theorem 4.13 *If (X_1, X_2) have copula C and set g_1, g_2 two continuously increasing functions, then $\{g_1(X_1), g_2(X_2)\}$ have the copula C, too.*

Example 4.28 Independence implies that the product of the cdf's F_{X_1} and F_{X_2} equals the joint distribution function F, i.e.:

$$F(x_1, x_2) = F_{X_1}(x_1) F_{X_2}(x_2). \tag{4.127}$$

Thus, we obtain the independence or product copula $C = \Pi(u, v) = uv$.

While it is easily understood how a product copula describes an independence relationship, the converse is also true. Namely, the joint distribution function of two independent random variables can be interpreted as a product copula. This concept is formalised in the following theorem:

Theorem 4.14 *Let X_1 and X_2 be random variables with continuous distribution functions F_{X_1} and F_{X_2} and the joint distribution function F. Then X_1 and X_2 are independent if and only if $C_{X_1,X_2} = \Pi$.*

Example 4.29 Let us consider the Gaussian copula for the case $\rho = 0$, i.e. vanishing correlation. In this case the Gaussian copula becomes

$$
\begin{aligned}
C_0^{\text{Gauss}}(u, v) &= \int_{-\infty}^{\Phi_1^{-1}(u)} \varphi(x_1)dx_1 \int_{-\infty}^{\Phi_2^{-1}(v)} \varphi(x_2)dx_2 \\
&= uv \\
&= \Pi(u, v).
\end{aligned}
\tag{4.128}
$$

The following theorem, which follows directly from Lemma 4.3, establishes the continuity of copulae .

Theorem 4.15 *Let C be a copula. Then for any $u_1, v_1, u_2, v_2 \in I$ holds*

$$|C(u_2, v_2) - C(u_1, v_1)| \leq |u_2 - u_1| + |v_2 - v_1| \tag{4.129}$$

From (4.129) it follows that every copula C is uniformly continuous on its domain.

A further important property of copulae concerns the partial derivatives of a copula with respect to its variables:

Theorem 4.16 *Let $C(u, v)$ be a copula. For any $u \in I$, the partial derivative $\frac{\partial C(u,v)}{\partial v}$ exists for almost all $u \in I$. For such u and v one has:*

$$\frac{\partial C(u, v)}{\partial v} \in I \tag{4.130}$$

The analogous statement is true for the partial derivative $\frac{\partial C(u,v)}{\partial u}$:

$$\frac{\partial C(u, v)}{\partial u} \in I \tag{4.131}$$

Moreover, the functions

$$u \mapsto C_v(u) \stackrel{\text{def}}{=} \partial C(u, v)/\partial v \text{ and}$$

$$v \mapsto C_u(v) \stackrel{\text{def}}{=} \partial C(u, v)/\partial u$$

are defined and non-increasing almost everywhere on I.

Until now, we have considered copulae only in a two-dimensional setting. Let us now extend this concept to the d-dimensional case, for a random variable in $\mathbb{R}^d$ with $d \geq 1$.

Let $U_1, U_2, \ldots, U_d$ be non-empty sets in $\overline{\mathbb{R}}$ and consider the function $F : U_1 \times U_2 \times \cdots \times U_d \longrightarrow \overline{\mathbb{R}}$. For $a = (a_1, a_2, \ldots, a_d)$ and $b = (b_1, b_2, \ldots, b_d)$ with $a \leq b$ (i.e. $a_k \leq b_k$ for all k) let $B = [a, b] = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]$ be the d-box with vertices $c = (c_1, c_2, \ldots, c_d)$. It is obvious that each c_k is either equal to a_k or to b_k.

Definition 4.8 The F-volume of a d-box $B = [a, b] = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_d, b_d] \subset U_1 \times U_2 \times \cdots \times U_d$ is defined as follows:

$$V_F(B) = \sum_{k=1}^{d} \text{sign}(c_k) F(c_k) \tag{4.132}$$

where $\text{sign}(c_k) = 1$, if $c_k = a_k$ for even k and $\text{sign}(c_k) = -1$, if $c_k = a_k$ for odd k.

Example 4.30 For the case $d = 3$, the F-volume of a 3-box $B = [a, b] = [x_1, x_2] \times [y_1, y_2] \times [z_1, z_2]$ is defined as:

$$V_F(B) = F(x_2, y_2, z_2) - F(x_2, y_2, z_1) - F(x_2, y_1, z_2) - F(x_1, y_2, z_2)$$
$$+ F(x_2, y_1, z_1) + F(x_1, y_2, z_1) + F(x_1, y_1, z_2) - F(x_1, y_1, z_1)$$

Definition 4.9 F is said to be a d-increasing function if for all d-boxes B with vertices in $U_1 \times U_2 \times \cdots \times U_d$ holds:

$$V_F(B) \geq 0 \tag{4.133}$$

Definition 4.10 If $U_1, U_2, \ldots, U_d$ have a smallest element $\min U_1, \min U_2, \ldots, \min U_d$ respectively, then we say that a function $F : U_1 \times U_2 \times \cdots \times U_d \longrightarrow \mathbb{R}$ is grounded if :

$$F(x) = 0 \text{ for all } x \in U_1 \times U_2 \times \cdots \times U_d \tag{4.134}$$

such that $x_k = \min U_k$ for at least one k.

The lemmas, which we presented for the two-dimensional case, have analogous multivariate versions, see Nelsen (1999).

Definition 4.11 A d-dimensional copula (or d-copula) is a function C defined on the unit d-cube $I^d = I \times I \times \cdots \times I$ such that

(i) for every $u \in I^d$ holds: $C(u) = 0$, if at least one coordinate of u is equal to 0;
   i.e. C is grounded.
(ii) for every $a, b \in I^d$ with $a \leq b$ holds:

$$V_C([a, b]) \geq 0; \tag{4.135}$$

   i.e. C is 2-increasing.
(iii) for every $u \in I^d$ holds: $C(u) = u_k$, if all coordinates of u are 1 except u_k.

Analogously to the two-dimensional setting, let us state the Sklar's theorem for the d-dimensional case.

Theorem 4.17 (Sklar's Theorem in d-Dimensional Case) *Let F be a d-dimensional distribution function with marginal distribution functions $F_{X_1}, F_{X_2}, \ldots, F_{X_d}$. Then a d-copula C exists such that for all $x_1, \ldots, x_d \in \overline{\mathbb{R}}^d$:*

$$F(x_1, x_2, \ldots, x_d) = C\{F_{X_1}(x_1), F_{X_2}(x_2), \ldots, F_{X_d}(x_d)\} \tag{4.136}$$

Moreover, if $F_{X_1}, F_{X_2}, \ldots, F_{X_d}$ are continuous then C is unique. Otherwise C is uniquely determined on the Cartesian product $Im(F_{X_1}) \times Im(F_{X_2}) \times \cdots \times Im(F_{X_d})$.
   Conversely, if C is a copula and $F_{X_1}, F_{X_2}, \ldots, F_{X_d}$ are distribution functions then F defined by (4.136) is a d-dimensional distribution function with marginals $F_{X_1}, F_{X_2}, \ldots, F_{X_d}$.

In order to illustrate the d-copulae we present the following examples:

Example 4.31 Let Φ denote the univariate standard normal distribution function and $\Phi_{\Sigma,d}$ the d-dimensional standard normal distribution function with correlation matrix Σ. Then the function

$$C_\rho^{\text{Gauss}}(u, \Sigma) = \Phi_{\Sigma,d}\{\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\}$$
$$= \int_{-\infty}^{\phi_1^{-1}(u_d)} \cdots \int_{-\infty}^{\phi_2^{-1}(u_1)} f_\Sigma(x_1, \ldots, x_n) dx_1 \ldots dx_d \tag{4.137}$$

is the d-dimensional Gaussian or normal copula with correlation matrix Σ. The function

$$f_\rho(x_1, \ldots, x_d) = \frac{1}{\sqrt{\det(\Sigma)}}$$
$$\times \exp\left\{ -\frac{(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_d))^\top (\Sigma^{-1}-\mathcal{I}_d)(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_d))}{2} \right\} \tag{4.138}$$

is a copula density function. The copula dependence parameter α is the collection of all unknown correlation coefficients in Σ. If $\alpha \neq 0$, then the corresponding normal copula allows to generate joint symmetric dependence. However, it is not possible to model a tail dependence, i.e. joint extreme events have a zero probability.

Example 4.32 Let us consider the following function

$$C_\theta^{\mathrm{GH}}(u_1, \ldots, u_d) = \exp\left[-\left\{ \sum_{j=1}^d \left(-\log u_j\right)^\theta \right\}^{1/\theta} \right] \tag{4.139}$$

One recognise this function is as the d-dimensional Gumbel–Hougaard copula function. Unlike the Gaussian copula, the copula (4.139) can generate an upper tail dependence.

Example 4.33 As in the two-dimensional setting, let us consider the d-dimensional Gumbel–Hougaard copula for the case $\theta = 1$. In this case the Gumbel–Hougaard copula reduces to the d-dimensional product copula, i.e.

$$C_1(u_1, \ldots, u_d) = \prod_{j=1}^d u_j = \Pi^d(u) \tag{4.140}$$

The extension of the two-dimensional copula M, which one gets from the d-dimensional Gumbel–Hougaard copula for $\theta \to \infty$ is denoted $M^d(u)$:

$$C_\theta(u_1, \ldots u_d) \longrightarrow \min(u_1, \ldots, u_d) = M^d(u) \tag{4.141}$$

The d-dimensional function

$$W^d(u) = \max(u_1 + u_2 + \cdots + u_d - d + 1, 0) \tag{4.142}$$

defines a copula with $W(u) \leq C(u)$ for any other d-dimensional copula function $C(u)$. $W^d(u)$ is the Fréchet–Hoeffding lower bound in the d-dimensional case.

The functions M^d and Π^d are d-copulae for all $d \geq 2$, whereas the function W^d fails to be a d-copula for any $d > 2$ (Nelsen, 1999). However, the d-dimensional version of the Fréchet–Hoeffding inequality can be written as follows:

$$W^d(u) \leq C(u) \leq M^d(u) \tag{4.143}$$

As we have already mentioned, copula functions have been widely applied in empirical finance.

Summary

↪ The cumulative distribution function (cdf) is defined as $F(x) = P(X < x)$.

↪ If a probability density function (pdf) f exists then $F(x) = \int_{-\infty}^{x} f(u)du$.

↪ The pdf integrates to one, i.e. $\int_{-\infty}^{\infty} f(x)dx = 1$.

4.8   Bootstrap

Recall that we need large sample sizes in order to sufficiently approximate the critical values computable by the CLT. Here large means $n > 50$ for one-dimensional data. How can we construct confidence intervals in the case of smaller sample sizes? One way is to use a method called the *Bootstrap*. The Bootstrap algorithm uses the data twice:

1. estimate the parameter of interest,
2. simulate from an estimated distribution to approximate the asymptotic distribution of the statistics of interest.

In detail, bootstrap works as follows. Consider the observations $x_1, \ldots, x_n$ of the sample $X_1, \ldots, X_n$ and estimate the empirical distribution function (EDF) F_n. In the case of one-dimensional data

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x). \tag{4.144}$$

This is a step function which is constant between neighbouring data points.
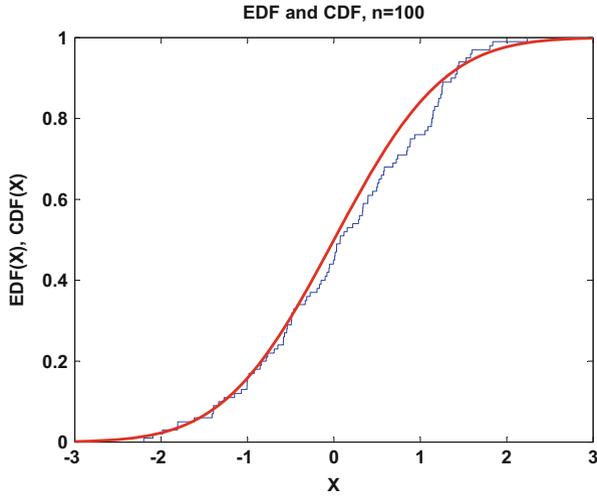
Fig. 4.19 The standard normal cdf (*thick line*) and the empirical distribution function (*thin line*) for $n = 100$ 🔍 MVAedfnormal
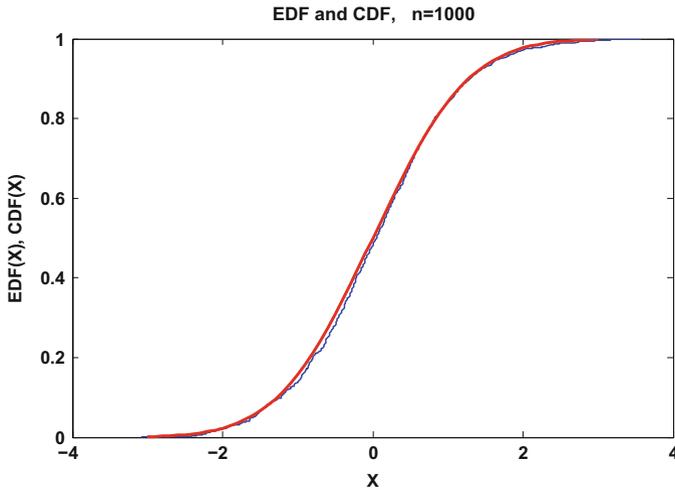
Fig. 4.20 The standard normal cdf (*thick line*) and the empirical distribution function (*thin line*) for $n = 1,000$ 🔍 MVAedfnormal

Example 4.34 Suppose that we have $n = 100$ standard normal $N(0, 1)$ data points $X_i, i = 1, \ldots, n$. The cdf of X is $\Phi(x) = \int_{-\infty}^{x} \varphi(u)du$ and is shown in Fig. 4.19 as the thin, solid line. The EDF is displayed as a thick step function line. Figure 4.20 shows the same setup for $n = 1,000$ observations.

Now draw with replacement a new sample from this empirical distribution. That is we sample with replacement n^* observations $X_1^*, \ldots, X_{n^*}^*$ from the original sample. This is called a Bootstrap sample. Usually one takes $n^* = n$.

Since we sample with replacement, a single observation from the original sample may appear several times in the Bootstrap sample. For instance, if the original sample consists of the three observations x_1, x_2, x_3, then a Bootstrap sample might look like $X_1^* = x_3, X_2^* = x_2, X_3^* = x_3$. Computationally, we find the Bootstrap sample by using a uniform random number generator to draw from the indices $1, 2, \ldots, n$ of the original samples.

The Bootstrap observations are drawn randomly from the empirical distribution, i.e. the probability for each original observation to be selected into the Bootstrap sample is $1/n$ for each draw. It is easy to compute that

$$\mathrm{E}_{F_n}(X_i^*) = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}.$$

This is the expected value given that the cdf is the original mean of the sample $x_1, \ldots, x_n$. The same holds for the variance, i.e.

$$V_{F_n}(X_i^*) = \widehat{\sigma}^2,$$

where $\widehat{\sigma}^2 = n^{-1} \sum (x_i - \overline{x})^2$. The cdf of the bootstrap observations is defined as in (4.144). Figure 4.21 shows the cdf of the $n = 100$ original observations as a solid line and two bootstrap cdf's as thin lines.

The CLT holds for the bootstrap sample. Analogously to Corollary 4.1 we have the following corollary.

EDF and 2 bootstrap EDF's, n=100
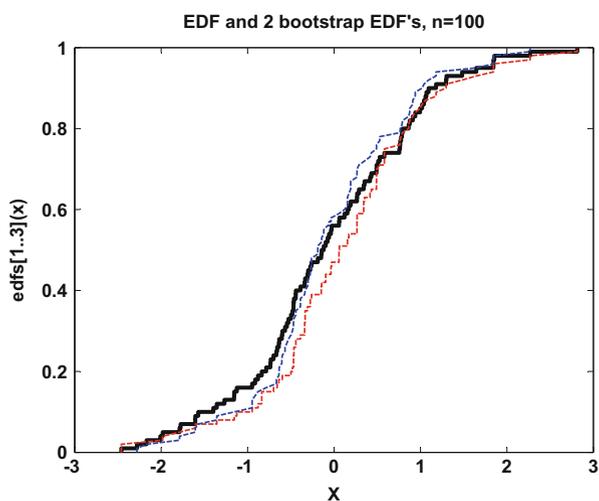
Fig. 4.21 The cdf F_n (*thick line*) and two bootstrap cdf's F_n^* (*thin lines*) 🔍
MVAedfbootstrap

Corollary 4.2 *If* $X_1^*, \ldots, X_n^*$ *is a bootstrap sample from* $X_1, \ldots, X_n$, *then the distribution of*

$$\sqrt{n} \left(\frac{\overline{x}^* - \overline{x}}{\widehat{\sigma}^*} \right)$$

also becomes $N(0, 1)$ *asymptotically, where* $\overline{x}^* = n^{-1} \sum_{i=1}^n X_i^*$ *and* $(\widehat{\sigma}^*)^2 = n^{-1} \sum_{i=1}^n (X_i^* - \overline{x}^*)^2$.

How do we find a confidence interval for μ using the Bootstrap method? Recall that the quantile $u_{1-\alpha/2}$ might be bad for small sample sizes because the true distribution of $\sqrt{n} \left(\frac{\overline{x} - \mu}{\widehat{\sigma}} \right)$ might be far away from the limit distribution $N(0, 1)$. The Bootstrap idea enables us to "simulate" this distribution by computing $\sqrt{n} \left(\frac{\overline{x}^* - \overline{x}}{\widehat{\sigma}^*} \right)$ for **many** Bootstrap samples . In this way we can estimate an empirical $(1 - \alpha/2)$-quantile $u_{1-\alpha/2}^*$. The bootstrap improved confidence interval is then

$$C_{1-\alpha}^* = \left[\overline{x} - \frac{\widehat{\sigma}}{\sqrt{n}} u_{1-\alpha/2}^*, \ \overline{x} + \frac{\widehat{\sigma}}{\sqrt{n}} u_{1-\alpha/2}^* \right].$$

By Corollary 4.2 we have

$$P(\mu \in C_{1-\alpha}^*) \longrightarrow 1 - \alpha \quad \text{as } n \to \infty,$$

but with an improved speed of convergence, see Hall (1992).

Summary

↪ For small sample sizes the bootstrap improves the precision of the confidence interval.

↪ The bootstrap distribution $\mathcal{L}\left\{ \sqrt{n}(\overline{x}^* - \overline{x})/\widehat{\sigma}^* \right\}$ converges to the same asymptotic limit as the distribution $\mathcal{L}\left\{ \sqrt{n}(\overline{x}^* - \overline{x})/\widehat{\sigma} \right\}$.

4.9 Exercises

Exercise 4.1 *Assume that the random vector* Y *has the following normal distribution:* $Y \sim N_p(0, \mathcal{I})$. *Transform it according to* (4.49) *to create* $X \sim N(\mu, \Sigma)$ *with mean* $\mu = (3, 2)^\top$ *and* $\Sigma = \begin{pmatrix} 1 & -1.5 \\ -1.5 & 4 \end{pmatrix}$. *How would you implement the resulting formula on a computer?*

Exercise 4.2 *Prove Theorem 4.7 using Theorem 4.5.*

Exercise 4.3 *Suppose that X has mean zero and covariance $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$. Let $Y = X_1 + X_2$. Write Y as a linear transformation, i.e. find the transformation matrix $\mathcal{A}$. Then compute $\mathsf{Var}(Y)$ via (4.26). Can you obtain the result in another fashion?*

Exercise 4.4 *Calculate the mean and the variance of the estimate $\hat{\beta}$ in (3.50).*

Exercise 4.5 *Compute the conditional moments $E(X_2|x_1)$ and $E(X_1|x_2)$ for the pdf of Example 4.5.*

Exercise 4.6 *Prove the relation (4.28).*

Exercise 4.7 *Prove the relation (4.29).*
Hint: Note that $\mathsf{Var}(E(X_2|X_1)) = E(E(X_2|X_1) E(X_2^\top|X_1)) - E(X_2) E(X_2^\top))$ and that
$E(\mathsf{Var}(X_2|X_1)) = E[E(X_2 X_2^\top|X_1) - E(X_2|X_1) E(X_2^\top|X_1)].$

Exercise 4.8 *Compute (4.46) for the pdf of Example 4.5.*

Exercise 4.9

$$\text{Show that} \quad f_Y(y) = \begin{cases} \frac{1}{2}y_1 - \frac{1}{4}y_2 & 0 \le y_1 \le 2, \ |y_2| \le 1 - |1 - y_1| \\ 0 & \text{otherwise} \end{cases} \quad \text{is a pdf.}$$

Exercise 4.10 *Compute (4.46) for a two-dimensional standard normal distribution. Show that the transformed random variables Y_1 and Y_2 are independent. Give a geometrical interpretation of this result based on iso-distance curves.*

Exercise 4.11 *Consider the Cauchy distribution which has no moment, so that the CLT cannot be applied. Simulate the distribution of $\overline{x}$ (for different n's). What can you expect for $n \to \infty$?*
Hint: The Cauchy distribution can be simulated by the quotient of two independent standard normally distributed random variables.

Exercise 4.12 *A European car company has tested a new model and reports the consumption of petrol (X_1) and oil (X_2). The expected consumption of petrol is 8 l per 100 km (μ_1) and the expected consumption of oil is 1 l per 10,000 km (μ_2). The measured consumption of petrol is 8.1 l per 100 km ($\overline{x}_1$) and the measured consumption of oil is 1.1 l per 10,000 km ($\overline{x}_2$). The asymptotic distribution of $\sqrt{n} \left\{ \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \right\}$ is $N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{pmatrix} \right)$.*
For the American market the basic measuring units are miles (1 mile $\approx$ 1.6 km) and gallons (1 gallon $\approx$ 3.8 l). The consumptions of petrol (Y_1) and oil (Y_2) are usually reported in miles per gallon. Can you express $\overline{y}_1$ and $\overline{y}_2$ in terms of $\overline{x}_1$ and $\overline{x}_2$? Recompute the asymptotic distribution for the American market.

Exercise 4.13 *Consider the pdf $f(x_1, x_2) = e^{-(x_1+x_2)}, x_1, x_2 > 0$ and let $U_1 = X_1 + X_2$ and $U_2 = X_1 - X_2$. Compute $f(u_1, u_2)$.*

Exercise 4.14 *Consider the pdf's*

$$f(x_1, x_2) = 4x_1 x_2 e^{-x_1^2} \quad x_1, x_2 > 0,$$
$$f(x_1, x_2) = 1 \qquad\qquad 0 < x_1, x_2 < 1 \text{ and } x_1 + x_2 < 1$$
$$f(x_1, x_2) = \tfrac{1}{2} e^{-x_1} \qquad x_1 > |x_2|.$$

For each of these pdf's compute $\mathrm{E}(X)$, $\mathsf{Var}(X)$, $\mathrm{E}(X_1|X_2)$, $\mathrm{E}(X_2|X_1)$, $\mathsf{Var}(X_1|X_2)$ *and* $\mathsf{Var}(X_2|X_1)$.

Exercise 4.15 *Consider the pdf* $f(x_1, x_2) = \frac{3}{4} x_1^{-\frac{1}{2}}$, $0 < x_1 < x_2 < 1$. *Compute* $P(X_1 < 0.25)$, $P(X_2 < 0.25)$ *and* $P(X_2 < 0.25 | X_1 < 0.25)$.

Exercise 4.16 *Consider the pdf* $f(x_1, x_2) = \frac{1}{2\pi}$, $0 < x_1 < 2\pi, 0 < x_2 < 1$. *Let* $U_1 = \sin X_1 \sqrt{-2 \log X_2}$ *and* $U_2 = \cos X_1 \sqrt{-2 \log X_2}$. *Compute* $f(u_1, u_2)$.

Exercise 4.17 *Consider* $f(x_1, x_2, x_3) = k(x_1 + x_2 x_3)$; $0 < x_1, x_2, x_3 < 1$.

(a) *Determine* k *so that* f *is a valid pdf of* $(X_1, X_2, X_3) = X$.
(b) *Compute the* (3×3) *matrix* Σ_X.
(c) *Compute the* (2×2) *matrix of the conditional variance of* (X_2, X_3) *given* $X_1 = x_1$.

Exercise 4.18 *Let* $X \sim N_2 \left(\binom{1}{2}, \begin{pmatrix} 2 & a \\ a & 2 \end{pmatrix} \right)$.

(a) *Represent the contour ellipses for* $a = 0; -\frac{1}{2}; +\frac{1}{2}; 1$.
(b) *For* $a = \frac{1}{2}$ *find the regions of* X *centred on* μ *which cover the area of the true parameter with probability* 0.90 *and* 0.95.

Exercise 4.19 *Consider the pdf*

$$f(x_1, x_2) = \frac{1}{8x_2} e^{-\left(\frac{x_1}{2x_2} + \frac{x_2}{4}\right)} \qquad x_1, x_2 > 0.$$

Compute $f(x_2)$ *and* $f(x_1|x_2)$. *Also give the best approximation of* X_1 *by a function of* X_2. *Compute the variance of the error of the approximation.*

Exercise 4.20 *Prove Theorem .*