51. Multivariate Modeling with Copulas and Engineering Applications

This chapter reviews multivariate modeling with copulas and provides novel applications in engineering. A copula separates the dependence structure of a multivariate distribution from its marginal distributions. Properties and statistical inferences of copula-based multivariate models are discussed in detail. Applications in engineering are illustrated via examples of bivariate process control and degradation analysis, using existing data in the literature. A software package has been developed to promote the development and application of copula-based methods.

Section 51.1 introduces the concept of copulas and its connection to multivariate distributions. The most important result about copulas is Sklar's theorem, which shows that any continuous multivariate distribution has a canonical representation by a unique copula and all its marginal distributions. A general algorithm to simulate random vectors from a copula is also presented.

Section 51.2 introduces two commonly used classes of copulas: elliptical copulas and Archimedean copulas. Simulation algorithms are also presented.

Section 51.3 presents the maximum-likelihood inference of copula-based multivariate distributions given the data. Three likelihood approaches are introduced. The exact maximum-likelihood approach estimates the marginal and copula parameters simultaneously by maximizing the exact parametric likelihood. The inference functions for margins approach is a two-step approach, which estimates the marginal parameters separately for each margin in a first step, and then estimates the copula parameters given the the marginal parameters.

The canonical maximum-likelihood approach is for copula parameters only, using uniform pseudo-observations obtained from transforming all the margins by their empirical distribution functions.

Section 51.4 presents two novel engineering applications. The first example is a bivariate process-control problem, where the marginal normality seems appropriate but joint normality is suspicious. A Clayton copula provides a better fit to the data than a normal copula. Through simulation, the upper control limit of Hotelling's T^2 chart based on normality is shown to be misleading when the true copula is a Clayton copula. The second example is a degradation analysis, where all the margins are skewed and heavy-tailed. A multivariate gamma distribution with normal copula fits the data much better than a multivariate normal distribution.

Section 51.5 concludes and points to references about other aspects of copula-based multivariate modeling that are not discussed in this chapter.

An open-source software package for the R project has been developed to promote copula-related methodology development and applications. An introduction to the package and illustrations are provided in the Appendix.

Multivariate methods are needed wherever independence cannot be assumed among the variables under investigation. Multivariate data are encountered in real life much more often than univariate data. This is especially true nowadays with the rapid growth of data-acquisition technology. For example, a quality-control engineer may have simultaneous surveillance of several related quality characteristics or process variables; a reliability analyst may measure the amount of degradation for a certain product repeatedly over time. Because of the dependence among the multiple quality characteristics and repeated measurements, univariate methods are invalid or inefficient. Multivariate methods that can account for the multivariate dependence are needed.

Classic multivariate statistical methods are based on the multivariate normal distribution. Under multivariate normality, an elegant set of multivariate techniques, such as principle-component analysis and factor analysis, has become standard tools and been successful in a variety of application fields. These methods have become so popular that they are often applied without a careful check about whether multivariate normality can reasonably be assumed.

In many applications, the multivariate normal assumption may be inappropriate or too strong to be made. Non-normality can occur in different ways. First, the marginal distribution of some variables may not be normal. For instance, in the degradation analysis in Sect. 51.5, the error rates of magnetic-optic disks at all time points are skewed and heavy-tailed, and hence cannot be adequately modeled by normal distributions. Second, even if all the marginal distributions are normal, jointly these variables may not be multivariate normal. For instance, in the bivariate process-control problem in Sect. 51.5, marginal normality seems appropriate but joint normality is suspicious. In both examples, multivariate distributions that are more flexible than the multivariate normal distribution are needed.

Non-normal multivariate distributions constructed from copulas have proved very useful in recent years in many applications. A copula is a multivariate distribution function whose marginals are all uniform over the unit interval. It is well known that any continuous random variable can be transformed to a uniform random variable over the unit interval by its probability integral transformation. Therefore, a copula can be used to *couple* different margins together and construct new multivariate distributions. This method separates a multivariate distribution into two components, all the marginals and a copula, providing a very flexible framework in multivariate modeling. Comprehensive book references on this subject are *Nelsen* [51.1] and *Joe* [51.2]. For widely accessible introductions, see, for example, *Genest* and *MacKay* [51.3] and *Fisher* [51.4].

Copula-based models have gained much attention in various fields. Actuaries have used copulas when modeling dependent mortality and losses [51.5–7]. Financial and risk analysts have used copulas in asset allocation, credit scoring, default risk modeling, derivative pricing, and risk management [51.8–10]. Biostatisticians have used copulas when modeling correlated event times and competing risks [51.11, 12]. The aim of this chapter is to provide a review of multivariate modeling with copulas and to show that it can be extensively used in engineering applications.

The chapter is organized as follows. Section 51.1 presents the formal definition of copulas and the construction of multivariate distribution with copulas. Section 51.2 presents details about two commonly used classes of copulas: elliptical copulas and Archimedean copulas. Section 51.3 presents likelihood-based statistical inferences for copula-based multivariate modeling. Section 51.4 presents two engineering applications: multivariate process control and degradation analysis. Section 51.5 concludes and suggests future research directions. An open-source software package `copula` [51.13] for the `R` project [51.14] has been developed by the author. A brief introduction to the package and illustrations are presented in the Appendix.

51.1 Copulas and Multivariate Distributions

51.1.1 Copulas

Consider a random vector $(U_1, \ldots, U_p)^\top$, where each margin U_i, $i = 1, \ldots, p$, is a uniform random variable over the unit interval. Suppose the joint cumulative distribution function (CDF) of $(U_1, \ldots, U_p)^\top$ is

$$C(u_1, \ldots, u_p) = \Pr(U_1 \le u_1, \ldots, U_p \le u_p) . \quad (51.1)$$

Then, the function C is called a p-dimensional copula. As *Embrechts* et al. [51.9] noted, this definition of a copula masks some of the problems when construct-

ing copulas using other techniques, by not explicitly specifying what properties a function must have to be a multivariate distribution function; for a more rigorous definition, see for example *Nelsen* [51.1]. However, this definition is operational and very intuitive. For example, one immediately obtains with this definition that, for any p-dimensional copula C, $p \geq 3$, each $k \leq p$ margin of C is a k-dimensional copula and that independence leads to a product copula

$$\Pi_p(u_1, \ldots, u_p) = \prod_{i=1}^{p} u_i \,. \tag{51.2}$$

Every continuous multivariate distribution function defines a copula. Consider a continuous random vector $(X_1, \ldots, X_p)^\top$ with joint CDF $F(x_1, \ldots, x_p)$. Let F_i, $i = 1, \ldots, p$, be the marginal CDF of X_i. Then, $U_i = F_i(X_i)$ is a uniform random variable over the unit interval. One can define a copula C as

$$C(u_1, \ldots, u_p) = F\left\{F_1^{-1}(u_1), \ldots, F_p^{-1}(u_p)\right\} . \tag{51.3}$$

The elliptical copulas in Sect. 51.2.1 are constructed this way. Another important class of copulas, Archimedean copulas, is constructed differently (Sect. 51.2.2).

A copula (51.1) can be used to construct multivariate distributions with arbitrary margins. Suppose that it is desired that the i-th margin X_i has marginal CDF G_i. A multivariate distribution function G can be defined via a copula C as

$$G(x_1, \ldots, x_p) = C\{G_1(x_1), \ldots, G_p(x_p)\} . \tag{51.4}$$

This multivariate distribution will have the desired marginal distributions.

Clearly, there is a close connection between copulas and multivariate distributions. It is natural to investigate the converse of (51.4). That is, for a given multivariate distribution function G, does there always exist a copula C such that (51.4) holds? If so, is this C unique? These problems are solved rigorously by *Sklar*'s [51.15] theorem in the next section.

51.1.2 Copulas to Multivariate Distributions

Sklar's theorem is the most important result about copulas. The bivariate version of the theorem was established by *Sklar* [51.15] almost half a century ago in the probability metrics literature. The proof in the general p-dimensional case is more involved and can be found in *Sklar* [51.16]. A formal statement of the theorem is as follows [51.1].

Theorem 51.1

Let F be a p-dimensional distribution function with margins $F_1, \ldots, F_p$. Then there exists a p-dimensional copula C such that, for all x in the domain of F,

$$F(x_1, \ldots, x_p) = C\{F_1(x_1), \ldots, F_p(x_p)\} . \tag{51.5}$$

If $F_1, \ldots, F_p$ are all continuous, the C is unique; otherwise, C is uniquely determined on $\mathrm{Ran}F_1 \times \cdots \times \mathrm{Ran}F_p$, where $\mathrm{Ran}H$ is the range of H. Conversely, if C is a p-dimensional copula and $F_1, \ldots, F_p$ are distribution functions, then the function F defined by (51.5) is a p-dimensional distribution function with marginal distributions $F_1, \ldots, F_p$.

Sklar's theorem ensures that a continuous multivariate distribution can be separated into two components, univariate margins and multivariate dependence, where the dependence structure is represented by a copula. The dependence structure of a multivariate distribution can be analyzed separately from its margins. It is sufficient to study the dependence structure of a multivariate distribution by focusing on its copula.

The probability density function (PDF) of the CDF F in (51.5) can be found from the PDF of C and $F_1, \ldots, F_p$. The PDF c of the copula C in (51.1) is

$$c(u_1, \ldots, u_p) = \frac{\partial^p C(u_1, \ldots, u_p)}{\partial u_1 \ldots \partial u_p} . \tag{51.6}$$

When the density c is known, the density f of the multivariate distribution F in (51.5) is

$$f(x_1, \ldots, x_p) = c\left\{F_1(x_1), \ldots, F_p(x_p)\right\} \prod_{i=1}^{p} f_i(x_i) \,, \tag{51.7}$$

where f_i is the density function of the distribution F_i. Expression (51.7) is called the canonical representation of a multivariate PDF. It will be used to construct likelihood for observed data.

51.1.3 Concordance Measures

The copula of two random variables completely determines any dependence measures that are scale-invariant, that is, measures that remain unchanged under monotonically increasing transformations of the random variables. The construction of the multivariate distribution (51.5) implies that the copula function C is invariant

under monotonically increasing transformations of its margins. Therefore, scale-invariant dependence measures can be expressed in terms of the copulas of the random variables.

Concordance measures of dependence are based on a form of dependence known as concordance. The most widely used concordance measures are Kendall's tau and Spearman's rho. Both of them can be defined by introducing a concordance function between two continuous random vectors (X_1, X_2) and (X_1', X_2') with possibly different joint distributions G and H, but with common margins F_1 and F_2. This concordance function Q is defined as

$$Q = \Pr\left\{(X_1 - X_1')(X_2 - X_2') > 0\right\} - \Pr\left\{(X_1 - X_1')(X_2 - X_2') < 0\right\}, \quad (51.8)$$

which is the difference between the probability of concordance and dis-concordance of (X_1, X_2) and (X_1', X_2'). It can be shown that

$$Q = Q(C_G, C_H) = 4\int_0^1\int_0^1 C_G(u, v)\,\mathrm{d}C_H(u, v) - 1, \quad (51.9)$$

where C_G and C_H are the copulas of G and H, respectively.

For a bivariate random vector (X_1, X_2) with copula C, Kendall's tau is defined as $Q(C, C)$, interpreted as the difference between the probability of concordance and dis-concordance of two independent and identically distributed observations. Therefore, we have

$$\tau = 4\int_0^1\int_0^1 C(u_1, u_2)\,\mathrm{d}C(u_1, u_2) - 1, \quad (51.10)$$

where the range of τ can be shown to be $[-1, 1]$. Spearman's rho, on the other hand, is defined as $3Q(C, \Pi)$, where Π is the product copula obtained under independence. That is,

$$\rho = 12\int_0^1\int_0^1 u_1 u_2\,\mathrm{d}C(u_1, u_2) - 3. \quad (51.11)$$

The constant 3 scales this measure into the range of $[-1, 1]$ (see for example *Nelson* [51.1] p.129). Spearman's rho is proportional to the difference between the probability of concordance and dis-concordance of two vectors: both have the same margins, but one has copula C while the other has the product copula Π. It is straightforward to show that Spearman's rho equals Pearson's product-moment correlation coefficient for the probability-integral-transformed variables $U_1 = F_1(X)$ and $U_2 = F_2(Y)$:

$$\rho = 12E(U_1U_2) - 3 = \frac{E(U_1U_2) - 1/4}{1/12} = \frac{E(U_1U_2) - E(U_1)E(U_2)}{\sqrt{\mathrm{Var}(U_1)\mathrm{Var}(U_2)}}. \quad (51.12)$$

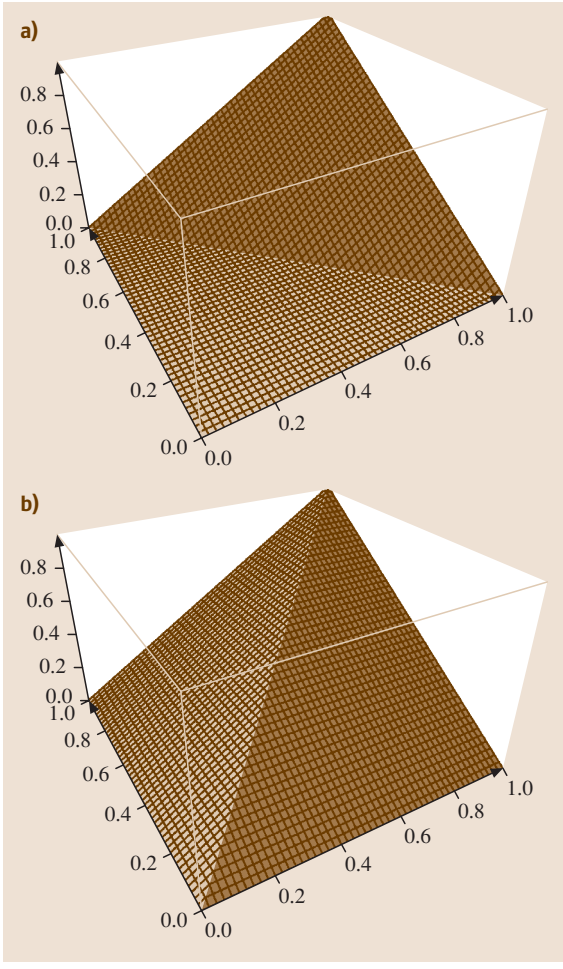

Fig. 51.1a,b Perspective plots of the Fréchet–Hoeffding bounds. **(a)** lower bound; **(b)** upper bound

There are other dependence measures based on copulas. For example, tail dependence is a very important measure when studying the dependence between extreme events. Details can be found in *Joe* [51.2].

51.1.4 Fréchet–Hoeffding Bounds

Important bounds are defined for copulas and multivariate distributions. These bounds are called the Fréchet–Hoeffding bounds, named after the pioneering work of Fréchet and Hoeffding, who independently published their work on this in 1935 and 1940, respectively [51.17]. Define the functions M_p and W_p on $[0,1]^p$ as follows:

$$M_p(u_1,\dots,u_p) = \min(u_1,\dots,u_p),$$
$$W_p(u_1,\dots,u_p) = \max(u_1+\cdots+u_p-n+1,0)\,.$$

Then for every copula C,

$$W_p(u_1,\dots,u_p) \le C(u_1,\dots,u_p) \le M_p(u_1,\dots,u_p)\,. \tag{51.13}$$

These bounds are general bounds, regardless of whether the margins are continuous or not. The function M_p is always a p-dimensional copula for $p \ge 2$. The function W_p fails to be a copula for $p \ge 2$, but it is the best possible lower bound since, for any $u = (u_1,\dots,u_p) \in [0,1]^p$, there exists a copula C (which depends on u) such that $C(u) = W_p(u)$. In the bivariate case, these bounds correspond to perfect negative dependence and perfect positive dependence, respectively. Within a given family of copulas, they may or may not be attained (see for example [51.1] Table 4.1). Figure 51.1 shows the perspective plots of the Fréchet–Hoeffding bounds copulas and the product copula.

Intuitively, perfect dependence should lead to extremes of concordance measures. It can be shown that, for continuous random vector (X_1, X_2) with copula C, $\tau = -1$ (or $\rho = -1$) is equivalent to $C = W_2$ and $\tau = 1$ (or $\rho = 1$) is equivalent to $C = M_2$; see *Embrechts* et al. [51.18] for a proof.

51.1.5 Simulation

Random-number generation from a copula is very important in statistical practice. Consider the p-dimensional copula in (51.1). Let $C_k(u_1,\dots,u_k) = C(u_1,\dots,u_k,1,\dots,1)$ for $k = 2,\dots,p-1$. The conditional CDF of U_k given $U_1 = u_1,\dots,U_{k-1} = u_{k-1}$ is

$$C_k(u_k|u_1,\dots,u_{k-1}) = \frac{\dfrac{\partial^{k-1}C_k(u_1,\dots,u_k)}{\partial u_1\dots\partial u_{k-1}}}{\dfrac{\partial^{k-1}C_{k-1}(u_1,\dots,u_{k-1})}{\partial u_1\dots\partial u_{k-1}}}\,. \tag{51.14}$$

Algorithm (51.1) is a general algorithm to generate a realization $(u_1,\dots,u_p)$ from C via a sequence of conditioning. When the expression for $C_k(\cdot|u_1,\dots,u_{k-1})$ is available, a root-finding routine is generally needed in generating u_k using the inverse CDF method. With realizations from C, one can easily generate realizations from the multivariate distribution (51.4) by applying the inverse CDF method at each margin.

Algorithm 51.1
Generating a random vector from a copula

1. Generate u_1 from a uniform over $[0,1]$.
2. For $k = 2,\dots,p$, generate u_k from $C_k(\cdot|u_1,\dots,u_{k-1})$.

51.2 Some Commonly Used Copulas

We introduce two commonly used copula classes in this section: elliptical copulas and Archimedean copulas. A third class of copulas, extreme-value copulas, is very useful in multivariate extreme-value theory but is omitted here to limit the scope of this chapter; more details about extreme-value copulas can be found in *Joe* [51.2].

51.2.1 Elliptical Copulas

Elliptical copulas are copulas of elliptical distributions. A multivariate elliptical distribution of random vector $(X_1,\dots,X_p)$ centered at zero has density of the form $\phi(t) = \psi(t^\top \Sigma t)$, where $t \in R^p$ and Σ is a $p \times p$ dispersion matrix, which can be parameterized such that $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$ [51.19]. Let R_{ij} and τ_{ij} be Pearson's linear correlation coefficient and Kendall's tau between X_i and X_j, respectively. For an elliptical distribution, they are connected through

$$\tau_{ij} = \frac{2}{\pi}\arcsin(R_{ij})\,. \tag{51.15}$$

This relationship makes elliptical copulas very attractive in applications since the similarity between Kendall's tau matrix and the correlation matrix can offer a wide range of dependence structures. Tractable properties similar to those of multivariate normal distributions are another

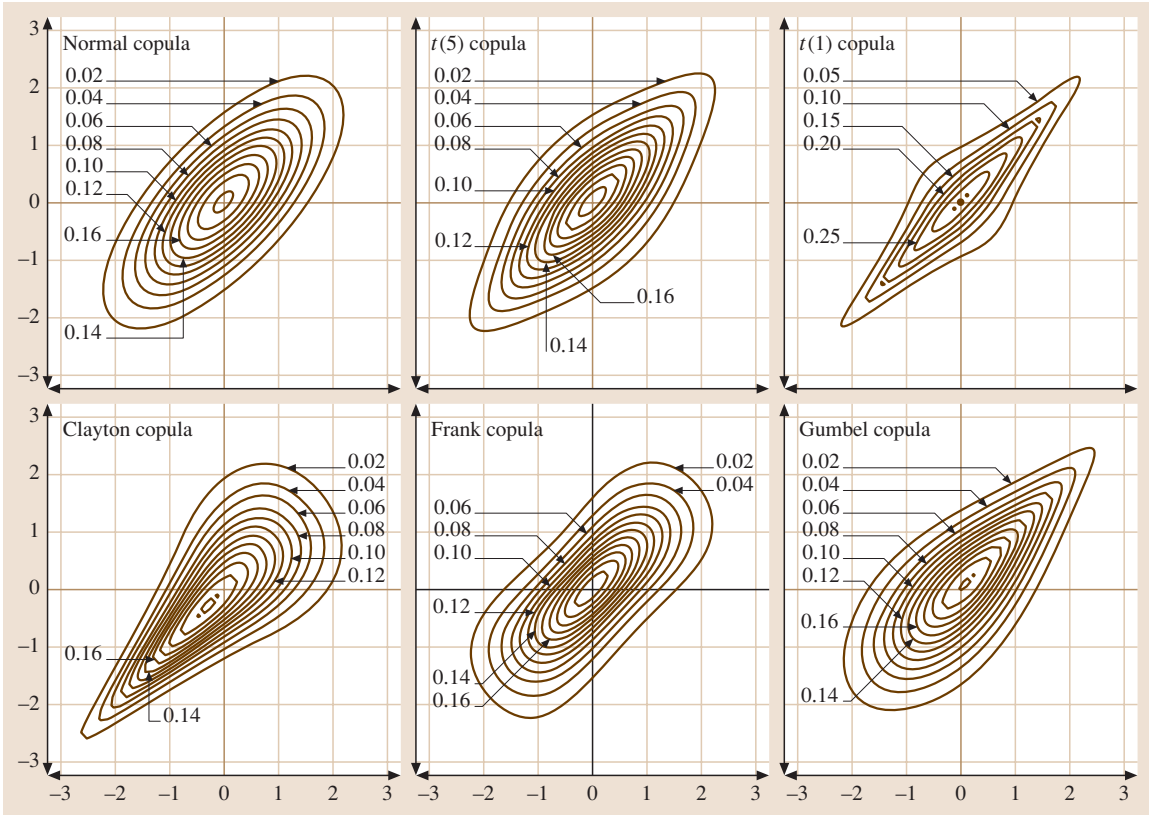

Fig. 51.2 Contours of bivariate distributions with the same marginals but different copulas. Both marginal distributions are standard normal

attractive feature of elliptical copulas. The most popular elliptical distributions are multivariate normal and multivariate t, providing two popular copulas: normal copulas and t copulas.

The normal copula has been widely used in financial applications for its tractable calculus [51.8,20]. Consider the joint CDF Φ_Σ of a multivariate normal distribution with correlation matrix Σ. Let Φ be the CDF of a standard normal variable. A normal copula with dispersion matrix Σ is defined as

$$C(u_1, \ldots, u_p; \Sigma) = \Phi_\Sigma\left[\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p)\right] . \quad (51.16)$$

The functions Φ, Φ^{-1} and Φ_Σ are available in any reasonably good statistical softwares, which makes their application widely accessible.

The t copula can be constructed similarly [51.21]. Consider the joint CDF $T_{\Sigma,\nu}$ of the standardized multivariate Student's t distribution with correlation matrix Σ and ν degrees of freedom. Let F_{t_ν} be the CDF of the univariate t distribution with ν degrees of freedom. A t copula with dispersion matrix Σ and degrees-of-freedom parameter ν is defined as

$$C(u_1, \ldots, u_p; \Sigma, \nu) = T_{\Sigma,\nu}\left[F_{t_\nu}^{-1}(u_1), \ldots, F_{t_\nu}^{-1}(u_p)\right] . \quad (51.17)$$

These copulas can be used to construct multivariate distributions using (51.5). Note that a normal copula with normal margins is the same as a multivariate normal distribution. However, a t copula with t margins is not necessarily a multivariate t distribution. A multivariate t distribution must have the same degrees of freedom at all the margins. In contrast, a t copula with t margins can have different degrees of freedom at different margins. It offers a lot more flexibility in modeling multivariate heavy-tailed data.

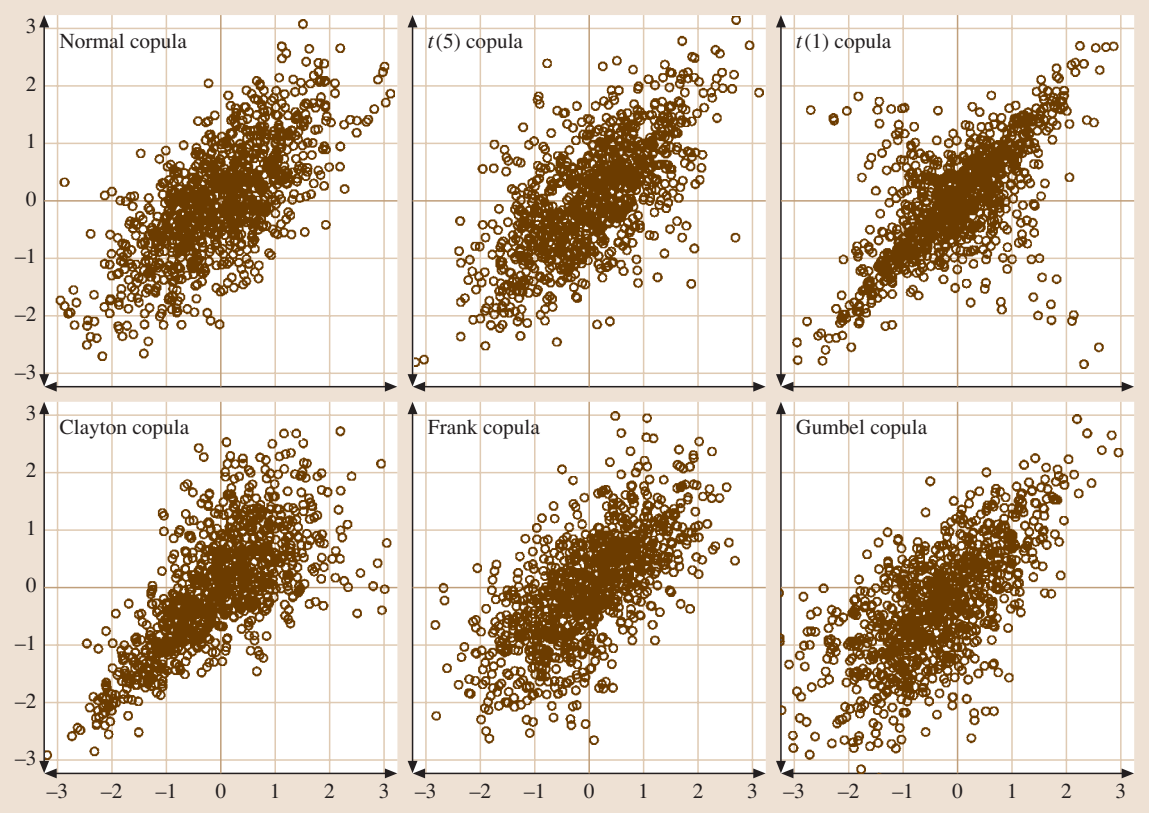

Fig. 51.3 1000 random points from bivariate distributions with the same marginals but different copulas. Both marginal distributions are standard normal

Figure 51.2 shows the density contours of bivariate distributions with the same margins but different copulas. These distributions all have standard normal as both margins, and their values of Kendall's tau are all 0.5. The three plots in the first row of Fig. 51.2 are for a normal copula, t copula with five degrees of freedom, and t copula with one degree of freedom (or Cauchy copula). These densities are computed with (51.7). Note that a normal copula can be viewed as a t copula with infinite degrees of freedom. Figure 51.2 illustrates that the dependence in the tails gets stronger as the number of degrees of freedom decreases.

Simulation from normal copulas and t copulas are straightforward if random-number generators for multivariate normal and t distributions are available. In `R`, the package `mvtnorm` [51.22] provides CDF, PDF and random-number generation for multivariate normal and multivariate t distributions. These facilities are used in the implementation of the package `copula` [51.13].

Figure 51.3 shows 1000 points from the corresponding bivariate distributions in Fig. 51.2.

51.2.2 Archimedean Copulas

Archimedean copulas are constructed via a completely different route without referring to distribution functions or random variables. A key component in this way of construction is a complete monotonic function. A function $g(t)$ is completely monotonic on an interval J if it is continuous there and has derivatives of all orders which alternate in sign, that is,

$$(-1)^k \frac{\mathrm{d}}{\mathrm{d}t^k}\varphi(t) \geq 0, \quad k = 1, 2, \cdots, \tag{51.18}$$

for all t in the interior of J. Let φ be a continuous strictly decreasing function form $[0, 1]$ to $[0, \infty]$ such that $\varphi(0) = \infty$ and $\varphi(1) = 0$, and let φ^{-1} be the inverse of φ. A function defined by

$$C(u_1, \ldots, u_p) = \varphi^{-1}\left[\varphi(u_1) + \cdots + \varphi(u_p)\right] \tag{51.19}$$

Table 51.1 Some one-parameter (α) Archimedean copulas

Family	Generator $\varphi(t)$	Frailty distribution	Laplace transformation of frailty $\mathcal{L}(s) = \varphi^{-1}(s)$
Clayton	$t^{-\alpha} - 1$	Gamma	$(1+s)^{-1/\alpha}$
Frank	$\ln \frac{e^{\alpha t}-1}{e^{\alpha}-1}$	Log series	$\alpha^{-1} \ln [1 + e^{s}(e^{\alpha} - 1)]$
Gumbel	$(-\ln t)^{\alpha}$	Positive stable	$\exp(-s^{1/\alpha})$

is a p-dimensional copula for all $p \geq 2$ if and only if φ^{-1} is completely monotonic over $[0, \infty)$; (see for example [51.1]). The copula C in (51.19) is called an Archimedean copula. The name Archimedean for these copulas comes from a property of the unit cube and copula C which is an analog of the Archimedean axiom for positive real numbers (see [51.1] p. 98 for more details). The function φ is called the generator of the copula. A generator uniquely (up to a scalar multiple) determines an Archimedean copula.

In the bivariate case, an Archimedean copula may be obtained with weaker conditions on the generator φ and its pseudo-inverse $\varphi^{[-1]}$:

$$C(u_1, u_2) = \max \left\{ \varphi^{[-1]} \left[\varphi(u_1) + \varphi(u_2) \right], 0 \right\} , \tag{51.20}$$

where the generator φ is a function with two continuous derivatives such that $\varphi(1) = 0$, $\varphi'(u) < 0$, and $\varphi''(u) > 0$ for all $u \in [0, 1]$, and $\varphi^{[-1]}$ is the pseudo-inverse of φ defined as

$$\varphi^{[-1]}(v) = \begin{cases} \varphi^{-1}(v) & 0 \leq v \leq \varphi(0) , \\ 0 & \varphi(0) \leq v \leq \infty . \end{cases}$$

The generator φ is called a strict generator if $\varphi(0) = \infty$, in which case $\varphi^{[-1]} = \varphi$. *Genest* and *McKay* [51.3] give proofs for some basic properties of bivariate copulas.

The generator φ plays an important role in the properties of an Archimedean copulas. It can be shown that Kendall's tau for an Archimedean copula with generator φ is

$$\tau = 4 \int_0^1 \int_0^1 \frac{\varphi(v)}{\varphi'(v)} \mathrm{d}v + 1 . \tag{51.21}$$

This relationship can be used to construct estimating equations that equate the sample Kendall's tau to the theoretical value from the assumed parametric copula family.

Due to the exchangeable structure in (51.19), the associations among all the variables are exchangeable too. As a consequence, an Archimedean copula cannot accommodate negative association unless $p = 2$. For Archimedean copulas with positive associations, there is a mixture representation due to *Marshall* and *Olkin* [51.23]. Suppose that, conditional on a positive latent random variable called the frailty, γ, the distribution of U_i is $F_i(U_i|\gamma) = U_i^{\gamma}$, $i = 1, \ldots, p$, and $U_1, \ldots, U_p$ are independent. Then the copula C of $U_1, \ldots, U_p$ is

$$C(u_1, \ldots, u_p) = E \left(\prod_{i=1}^{p} u_i^{\gamma} \right) , \tag{51.22}$$

where the expectation is taken with respect to the distribution of γ, F_{γ}. Recall that the Laplace transform of γ is

$$\mathcal{L}(s) = E\gamma(\mathrm{e}^{-s\gamma}) = \int_0^{\infty} \mathrm{e}^{-sx} \, \mathrm{d}F_{\gamma}(x) .$$

The Laplace transform has a well-defined inverse $\mathcal{L}^{-1}$. *Marshall* and *Olkin* [51.23] show that the copula in (51.22) is

$$C(u_1, \ldots, u_p) = \mathcal{L} \left[\mathcal{L}^{-1}(u_1) + \cdots + \mathcal{L}^{-1}(u_p) \right] . \tag{51.23}$$

This result suggests that an Archimedean copula can be constructed using the inverse of a Laplace transform as the generator.

Table 51.1 summarizes three commonly used one-parameter Archimedean copulas. A comprehensive list of one-parameter bivariate Archimedean copulas and their properties can be found in Table 4.1 of *Nelson* [51.1]. The three copulas in Table 51.1 all have inverse transforms of some positive random variables as their generators. The Clayton copula was introduced by *Clayton* [51.24] when modeling correlated survival times with a gamma frailty. The Frank copula first appeared in *Frank* [51.25]. It can be shown that the inverse of its generator is the Laplace transform of a log series random variables defined on positive integers. The Gumbel copula can be traced back to *Gumbel* [51.26]. *Hougaard* [51.27] uses a positive stable random variable to derive the multivariate distribution based on a Gumbel copula.

Density contours of bivariate distributions constructed from these three Archimedean copulas are presented in the second row of Fig. 51.2. Both margins of these distributions are still standard normals. The parameters of these copulas are chosen such that the value of Kendall's tau is 0.5. The density of an Archimedean copula can be found by differentiating the copula as in (51.6). When the dimension p is high, the differentiation procedure can be tedious. Symbolic calculus softwares can be used for this purpose. The package `copula` uses the simple symbolic derivative facility in `R` combined with some programming to construct PDF expressions for copulas given the generator function and its inverse function. From Fig. 51.2, one observes that the Frank copula has symmetric dependence. The dependence of the distribution based on the Clayton copula is stronger in the lower-left region than in the upper-right region. In contrast, the dependence of the distribution based on the Gumbel copula is stronger in the upper-right region than in the lower-left region.

Simulation from a general Archimedean can be done using the general Algorithm (51.1) in Sect. 51.2. When the inverse of the generator is known to be the Laplace transform of some positive random variable, an algorithm based on (51.23) is summarized in Algorithm (51.2) [51.6]. This algorithm is very easy to implement, given that a random-number generator of the frailty is available. Gamma-variable generator is available in most softwares. Algorithms for generating positive stable and log series variables can be found in *Chambers* et al. [51.28] and *Kemp* [51.29], respectively. For the bivariate case, the general algorithm (51.1) can be simplified, avoiding numerical root-finding. These algorithms have been implemented in the package `copula` [51.13]. The lower panel of Fig. 51.3 shows 1000 random points generated from the corresponding bivariate distributions with Archimedean copulas in Fig. 51.2.

Algorithm 51.2

tbp Generating a random vector from an Archimedean copula with a known frailty distribution

1. Generate a latent variable γ whose Laplace transformation $\mathcal{L}$ is the inverse generator function φ^{-1}.
2. Generate independent uniform observations $v_1, \ldots, v_p, i = 1, \ldots, p$.
3. Output $u_i = \mathcal{L}(-\gamma^{-1} \log v_i), i = 1, \ldots, p$.

51.3 Statistical Inference

This section presents the maximum-likelihood (ML) estimation for multivariate distributions constructed from copulas. Other methods, such as moment methods and nonparametric methods, are less developed for copula-based models and hence omitted.

Suppose that we observe a random sample of size n from a multivariate distribution (51.5):

$$(X_{i1}, \ldots, X_{ip})^\top, \qquad i = 1, \ldots, n\,.$$

The parameter of interest is $\theta = (\beta^\top, \alpha^\top)^\top$, where β is the marginal parameter vector for the marginal distributions $F_i, i = 1, \ldots, p$, and α is the association parameter vector for the copula C. Regression models for the marginal variables can be incorporated easily by assuming that the residuals follow a multivariate distribution (51.5).

51.3.1 Exact Maximum Likelihood

The exact log-likelihood $l(\theta)$ of the parameter vector θ can be expressed from (51.7):

$$l(\theta) = \sum_{i=1}^{n} \log c\left[F_1(X_{i1}; \beta), \ldots, F_p(X_{ip}; \beta); \alpha\right] + \sum_{i=1}^{n}\sum_{j=1}^{p} \log f_i(X_{ij}; \beta)\,. \tag{51.24}$$

The ML estimator of θ is

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta \in \Theta} l(\theta)\,,$$

where Θ is the parameter space.

Under the usual regularity conditions for the asymptotic ML theory, the ML estimator $\hat{\theta}_{\mathrm{ML}}$ is consistent and asymptotically efficient, with limiting distribution

$$\sqrt{n}(\hat{\theta}_{\mathrm{ML}} - \theta_0) \to N\left[0, I^{-1}(\theta_0)\right]\,,$$

where θ_0 is the true parameter value and I is the Fisher information matrix. The asymptotic variance matrix $I^{-1}(\theta_0)$ can be estimated consistently by an empirical variance matrix of the influence functions evaluated at $\hat{\theta}_{\mathrm{ML}}$.

To carry out the ML estimation, one feeds the log-likelihood function $l(\theta)$ to an optimization routine. The asymptotic variance matrix can be obtained from the inverse of an estimated Fisher information matrix, which is the negative Hessian matrix of $l(\theta)$. In `R`, one constructs the likelihood function using copula densities supplied in the `copula` package, and uses `optim` to maximize it.

The maximization of $l(\theta)$ in (51.24) may be a difficult task, especially when the dimension is high and/or the number of parameters is large. The separation of the margins and copula in (51.24) suggests that one may estimate the marginal parameters and association parameters in two steps, leading to the method in the next subsection.

51.3.2 Inference Functions for Margins (IFM)

The IFM estimation method was proposed by *Joe* and *Xu* [51.30]. This method estimates the marginal parameters β in a first step by

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \sum_{j=1}^{p} \log f_i(X_{ij}; \beta) \,, \tag{51.25}$$

and then estimates the association parameters α given $\hat{\beta}$ by

$$\hat{\alpha} = \arg\max_{\alpha} \sum_{i=1}^{n} \log c \times \Big[F_1(X_{i1}; \hat{\beta}), \ldots, F_p(X_{ip}; \hat{\beta}); \alpha \Big] \,. \tag{51.26}$$

When each marginal distribution F_i has its own parameters β_i so that $\beta = (\beta_1^\top, \ldots, \beta_p^\top)^\top$, the first step consists of an ML estimation for each margin $j = 1, \ldots, p$:

$$\hat{\beta}_j = \arg\max_{\beta_j} \sum_{i=1}^{n} \log f(X_{ij}; \beta_j) \,. \tag{51.27}$$

In this case, each maximization task has a very small number of parameters, greatly reducing the computational difficulty. This approach is called the two-stage parametric ML method by *Shih* and *Louis* [51.31] in a censored data setting.

The IFM estimator from (51.25) and (51.26), $\hat{\theta}_{\mathrm{IFM}}$, is in general different from the ML estimate $\hat{\theta}_{\mathrm{ML}}$. The limiting distribution of $\hat{\theta}_{\mathrm{IFM}}$ is

$$\sqrt{n}(\hat{\theta}_{\mathrm{IFM}} - \theta_0) \to N\Big[0, G^{-1}(\theta_0)\Big] \,,$$

where G is the Godambe information matrix [51.32]. This matrix has a sandwich form like the usual robust estimation with estimating functions. Detailed expressions can be found in *Joe* [51.2].

Compared to the ML estimator, the IFM estimator has advantages in numerical computations and is asymptotically efficient. Even in finite samples, it is highly efficient relative to the exact ML estimator [51.2]. The IFM estimate can be used as a starting value in an exact ML estimation.

51.3.3 Canonical Maximum Likelihood (CML)

When the association is of explicit interest, the parameter α can be estimated with the CML method without specifying the marginal distribution. This approach uses the empirical CDF of each marginal distribution to transform the observations $(X_{i1}, \ldots, X_{ip})^\top$ into pseudo-observations with uniform margins $(U_{i1}, \ldots, U_{ip})^\top$ and then estimates α as

$$\hat{\alpha}_{\mathrm{CML}} = \arg\max_{\alpha} \sum_{i=1}^{n} \log c(U_{i1}, \ldots, U_{ip}; \alpha) \,. \tag{51.28}$$

The CML estimator $\hat{\alpha}_{\mathrm{CML}}$ is consistent, asymptotically normal, and fully efficient at independence [51.31, 33].

51.4 Engineering Applications

Two engineering applications of copulas are considered in this section: multivariate process control and degradation analysis. An important third application is the modeling of multivariate failure times, which may be censored. We focus on complete-data applications in this chapter. In the example of multivariate process control, marginal normality seems appropriate but joint normality is suspicious. In the example of degradation analysis, the margins are right-skewed and have long tails. We use a gamma distribution for each margin and a normal copula for the association.

51.4.1 Multivariate Process Control

In quality management, multiple process characteristics necessitate a multivariate method for process

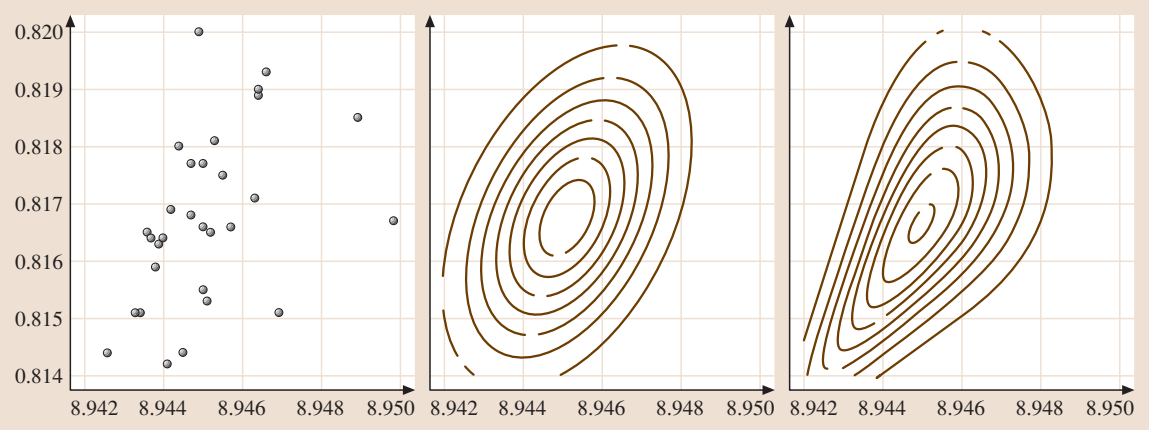

Fig. 51.4 Bivariate process characteristics and parametric fits. *Left*: scatter plot of the data; *center*: contours of bivariate normal fit; *right*: contours of bivariate fit with normal margins and Clayton copula

control. There are three major control charts used in practice: Hotelling's T^2, multivariate cumulative sum (MCUSUM), and multivariate exponentially weighted moving average (MEWMA); see *Lowry* and *Montgomery* [51.34] for a review. The most popular multivariate control chart is the T^2 chart, which has a long history since *Hotelling* [51.35]. *Mason* and *Young* [51.36] give details on how to use it in industrial applications. This method assumes that the multiple characteristics under surveillance are jointly normally distributed. The control limit of the chart is based on the sampling distribution of the statistic T^2, which can be shown to have an F distribution. When the multivariate normal assumption does not hold, due to either univariate or multivariate non-normality, the T^2 control chart based on multivariate normality can be inaccurate and misleading.

Copula-based multivariate distributions open a new avenue for the statistical methods of multivariate process control. The parametric form of the multivariate distribution can be determined from a large amount of historical in-control data. Given a sample of observations when the process is in-control, one can estimate the parameters and propose a statistic that measures the deviation from the target. The exact distribution of this statistic is generally unknown, and the control limit needs to be obtained from bootstrap; see for example *Liu* and *Tang* [51.37].

As an illustration, consider the example of bivariate process control in *Lu* and *Rudy* [51.38]. The data consists of 30 pairs of bivariate measurements from an exhaust manifold used on a Chrysler 5.21 engine in a given model year. They were collected from a machine capability study performed on the machine builder's floor. The sample correlation coefficient is 0.44. The left panel of Fig. 51.4 shows the scatter plot of the 30 observations. The assumption of normality for each margin seems fine from the normal Q–Q plots (not shown). However, the joint distribution may not be a bivariate normal. The scatter plot suggests that the association may be stronger in the lower end than in the higher end of the data. This nonsymmetric association cannot be captured by a symmetric copula, such as those elliptical copula and Frank copula in Fig. 51.2. A better fit of the data may be obtained from a Clayton copula, which allows the bivariate dependence to be stronger in the left tail than in the right tail. The center panel of Fig. 51.4 shows the contours of the ML bivariate normal fit. The right panel of Fig. 51.4 shows the contours of the ML bivariate fit with normal margins and the Clayton copula. The maximized log-likelihood of the two models are 307.64 and 309.87, respectively. A formal test of the difference, which is beyond the scope of this chapter, can be done by comparing non-nested models without knowing the true model based on Kullback–Leibler information [51.39].

The T^2 control chart of *Lu* and *Rudy* [51.38] is a phase II chart for single observations to detect any departure of the underlying process from the standard values. Suppose that we observe a random sample of p-dimensional multivariate observations with sample size m. Let $\bar{X}_m$ and S_m be the sample mean vector and sample covariance matrix, respectively. For a future p-dimensional multivariate observation X, the T^2 is defined as

$$T^2 = (X - \bar{X}_m)^\top S_m^{-1} (X - \bar{X}) \,. \tag{51.29}$$

Table 51.2 Comparison of T^2 percentiles when the true copula is normal and when the true copula is Clayton with various Kendall's τ. The percentiles under Clayton copulas are obtained from 100 000 simulations

Percentiles	Normal copula	Clayton copula			
		$\tau = 0.2$	$\tau = 0.4$	$\tau = 0.6$	$\tau = 0.8$
90%	5.357	5.373	5.416	5.590	5.868
95%	7.150	7.253	7.468	8.061	9.396
99%	11.672	12.220	13.080	15.764	23.526
99.73%	15.754	16.821	18.611	24.173	41.123

Under joint normality, it can be shown that the exact distribution of

$$\frac{m^2 - mp}{p(m+1)(m-1)} T^2$$

is F with degrees of freedom p and $m-p$. The exact upper control limit (UCL) for T^2 with level α is then

$$\mathrm{UCL}_\alpha = \frac{p(m+1)(m-1)}{m^2 - mp} F_{1-\alpha;\,p,m-p}\,, \qquad (51.30)$$

where $F_{1-\alpha;\,p,m-p}$ is the $100(1-\alpha)$ percentile of an F distribution with p and $m-p$ degrees of freedom. In this example, $m = 30$, $p = 2$. The exact upper control limit for T^2 with level α is then

$$\begin{aligned}\mathrm{UCL}_\alpha &= 2(30+1)(30-1)/[30^2 - 2(30)]F_{1-\alpha;2,28} \\ &= 2.14 F_{1-\alpha;2,28}\,.\end{aligned}$$

With $\alpha = 0.9973$, the control limit UCL = 15.75.

When the true copula is a Clayton copula but is misspecified as a normal copula, the control limit in (51.30) can be inaccurate and hence misleading. By comparing the contours of a normal copula model with those of a Clayton copula model in Fig. 51.2, one can conjecture that, if the true copula is a Clayton copula, then $\Pr(T^2 > \mathrm{UCL}_\alpha)$ will be greater than its nominal level α, because the bivariate density with the Clayton copula is more concentrated on the lower-left part of the plot than the bivariate normal density. In other words, in order to maintain the control level α, one needs to increase the UCL of the T^2 chart. This difference obviously depends on the sample size m and the association parameter of the true Clayton copula. For a given sample size m and a Kendall's τ value, which determines the association strength of a Clayton copula, the control limit of T^2 can be obtained by simulation. Table 51.2 compares the 90%, 95%, 99%, and 99.73% percentiles of T^2 when the true copula is normal and when the true copula is Clayton. The percentiles under Clayton copulas are obtained from 100 000 simulations. The true Clayton copulas are parameterized to give Kendall's τ values 0.2, 0.4, 0.6, and 0.8. From Table 51.2, one observes that the simulated percentiles of T^2 are greater than those based on the F distribution under the normal assumption. The control region based on the normal assumption is smaller than expected, which will result in investigating the process more often than necessary when the process is actually in control. The difference increases with the strength of the association.

This example illustrates that a non-normal joint distribution may have an important influence on the control limit of the widely used T^2 chart, even when both the margins are normals. The T^2 statistic still measures the deviance from the target, but its distribution is unknown under the non-normal model. A comprehensive investigation of multivariate process control using copula is a future research direction.

51.4.2 Degradation Analysis

Performance degradation data has repeated measures over time for each test unit (see for example *Meeker* and *Escobar* [51.40] Chap. 13). These repeated measures on the same unit are correlated. There is a voluminous statistical literature on the analysis of repeated mea-

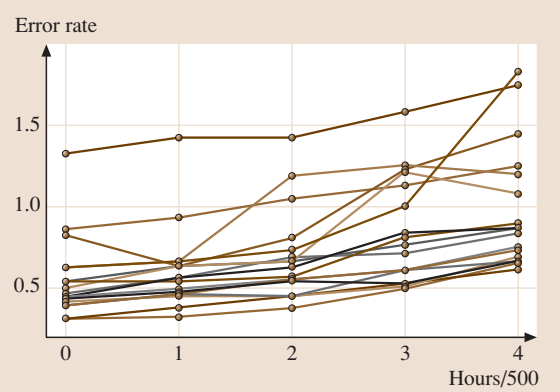

Fig. 51.5 Error rates ($\times 10^5$) of 16 magneto-optic data-storage disks measured every 500 h

Table 51.3 IFM fit for all the margins using normal and gamma distributions, both parameterized by mean and standard deviation. Presented results are log-likelihood (Loglik), estimated mean, and estimated standard deviation (StdDev) for each margin under each model

Time in units of 500 h	Normal margins			Gamma margins		
	Loglik	Mean	StdDev	Loglik	Mean	StdDev
0	−0.484	0.565	0.062	2.568	0.565	0.054
1	−0.526	0.617	0.063	2.538	0.617	0.054
2	−2.271	0.709	0.070	−0.125	0.709	0.064
3	−4.441	0.870	0.080	−3.269	0.870	0.078
4	−6.996	1.012	0.094	−5.205	1.012	0.087

surements (see, for example, *Davis* [51.41]. Analysis of such data has been implemented in popular statistical softwares, for example, `PROC MIXED` of the SAS system [51.42] and the `nlme` package [51.43] for `R` and `Splus`. Continuous response variables are generally assumed to be normally distributed and a multivariate normal distribution is used in likelihood-based approaches. The following example shows that a multivariate gamma distribution with normal copula can provide a much better fit to the data than a multivariate normal distribution.

Degradation data on block error rates of 16 magneto-optic data storage disks are collected every 500 h for 2000 h at 80 °C and 85% relative humidity [51.44]. Figure 51.5 shows these error rates at all five time points. A degradation analysis often needs to fit a curve for the degradation trend in order to allow predictions at unobserved time points. Before choosing a curve to fit, we first carry out exploratory data analysis using the two-step IFM method to look into parametric modeling for each margin and copula separately.

Separate parametric fits for each margin is the first step of the IFM approach in Sect. 51.4. Two parametric models for each margin are used: normal and gamma. To make the parameters comparable across models, the gamma distribution is parameterized by its mean μ and standard deviation σ, giving a density function of

$$f(x;\mu,\sigma)=\frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}e^{-\frac{x}{\beta}}\,, \tag{51.31}$$

where $\alpha=\mu^2/\sigma^2$ and $\beta=\sigma^2/\mu$. Table 51.3 summarizes the separate parametric fits for each margins using normal and gamma distributions. For all the margins, the gamma distribution fit yields higher log-likelihood than the normal distribution fit. The estimated mean from both models are the same for the first three digits after the decimal point. The estimated standard deviation is noticeably lower in the gamma model, especially at earlier time points where the data are more skewed and heavier-tailed. These estimates are consistent with the descriptive statistics of each time point, suggesting that the mean error rate is increasing over time, and their standard errors is increasing with the mean level.

Given the parametric fit for each margins, we can explore copula fitting in the second step of IFM. Due to the small number of observations, we choose single-parameter normal copulas with three dispersion structures: AR(1), exchangeable, and Toeplitz. In particular, with $p=5$, the dispersion matrices with parameter ρ under these structures are, respectively,

$$\begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho^2 & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{pmatrix}, \text{ and}$$

$$\begin{pmatrix} 1 & \rho & & & \\ \rho & 1 & \rho & & \\ & \rho & 1 & \rho & \\ & & \rho & 1 & \rho \\ & & & \rho & 1 \end{pmatrix}. \tag{51.32}$$

Table 51.4 summarizes the log-likelihood and the estimated association parameter ρ for the given estimated margins in Table 51.3. Note that the log-likelihood values are not comparable across models with different margins because the data being used in the estimation are different. They are comparable when the modeled margins are the same. For both normal margins and gamma margins, the AR(1) structure gives the highest log-likelihood value. The estimated parameter is about 0.9, indicating high dependence among repeated measurements.

Table 51.4 also presents the normal copulas estimation using the CML method. No parametric distribution is assumed for each margin. The empirical distribution is used to transform the observations of each margin

Table 51.4 IFM and CML fit for single-parameter normal copulas with dispersion structures: AR(1), exchangeable, and Toeplitz

Dispersion structure	IFM fit				CML fit	
	Normal margins		Gamma margins		Empirical margins	
	Loglik	$\hat{\rho}$	Loglik	$\hat{\rho}$	Loglik	$\hat{\rho}$
AR(1)	39.954	0.917	66.350	0.892	10.380	0.964
Exchangeable	38.618	0.868	62.627	0.791	9.791	0.942
Toeplitz	23.335	0.544	39.975	0.540	5.957	0.568

into uniform variables in [0, 1], which are then used in (51.28). The CML fit also shows that the AR(1) structure gives the highest log-likelihood and that the within-disk dependence is high. Based on these exploratory analysis, the AR(1) structure is used for the dispersion matrix of normal copula in an exact ML analysis.

We now present the exact ML estimation of a degradation model. For the sake of simplicity, we use a linear function of time to model the mean $\mu(t)$ and a linear function of $\mu(t)$ to model the logarithm of the standard deviation $\sigma(t)$. That is,

$$\mu(t) = \phi_0 + \phi_1 t, \tag{51.33}$$

$$\log \sigma(t) = \psi_0 + \psi_1 [\mu(t) - 1.0] , \tag{51.34}$$

where ϕ_0, ϕ_1, ψ_0, and ψ_1 are parameters, and the function of $\log \sigma(t)$ is centered at 1.0 for easier prediction of the variance at higher error rates. Two parametric models are considered for the repeated error rates: (1) multivariate normal and (2) multivariate gamma via a normal copula. Note that the two models both use the normal copula. The marginal distributions of the two models at time t are both parameterized by mean $\mu(t)$ and standard deviation $\sigma(t)$ for comparison purpose. A similar parameterization was used in *Lambert* and *Vandenhende* [51.45] and *Frees* and *Wang* [51.7].

Table 51.5 summarizes the maximum-likelihood estimate of the parameters and their standard errors for both models. These estimates for both marginal parameters and the copula parameter are virtually the same or very close to each other. However, the standard errors of these estimates are noticeably smaller in the multivariate gamma model. The maximized log-likelihood from the gamma model is much higher than that from the normal model. Given that both models have the same number of parameters, the multivariate gamma distribution fits the data much better.

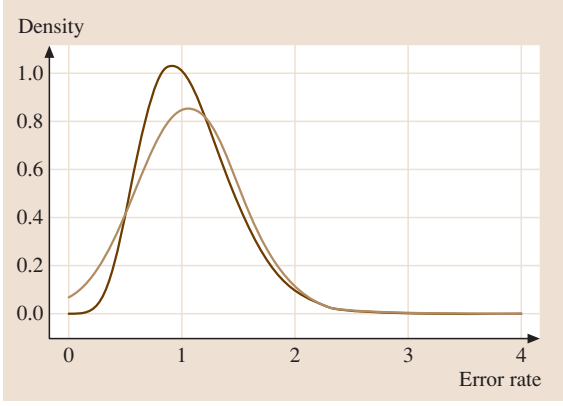

Fig. 51.6 Predictive density of disk error rate at 2500 h. The *dark line* is from the gamma model; the *gray line* is from the normal model

The difference between the two models can also be illustrated by their predictive density of the error rate at 2500 h. Figure 51.6 presents the densities of the error rate at 2500 h using the estimated mean $\mu(2500)$ and

Table 51.5 Maximum-likelihood results for the disk error-rate data. Parameter estimates, standard errors and log-likelihood are provided for both the multivariate normal model and the multivariate gamma model with a normal copula. The second entry of each cell is the corresponding standard error

Model	Marginal parameters				Copula parameter	Loglik
	Mean		StdDev.			
	ϕ_0	ϕ_1	ψ_0	ψ_1	ρ	
Normal	0.564	0.099	−0.849	1.439	0.899	34.719
	0.057	0.019	0.262	0.557	0.034	
Gamma	0.564	0.101	−0.986	1.383	0.900	48.863
	0.051	0.015	0.185	0.442	0.033	

$\sigma(2500)$ obtained with $\hat{\phi}_0$, $\hat{\phi}_1$, $\hat{\psi}_0$, and $\hat{\psi}_1$. The normal model gives mean 1.058 and standard deviation 0.465, while the gamma model gives mean 1.070 and standard deviation 0.411. Although the mean values are close, the gamma model gives a small standard deviation and captures the skewness and long tail of the data.

51.5 Conclusion

This chapter reviews multivariate modeling with copulas and provides novel applications in engineering. Multivariate distribution construction using copulas and their statistical inferences are discussed in detail. Engineering applications are illustrated via examples of bivariate process control and degradation analysis, using existing data in the literature. Copulas offer a flexible modeling strategy that separates the dependence structure from the marginal distributions. Multivariate distributions via copula apply to a much wider range of multivariate scenarios than the traditionally assumed multivariate normal distribution. A publicly available R package has been developed to promote the research on copulas and their applications.

Some important topics about copulas are not discussed in this chapter. The survival function is of great concern in failure-time data analysis. Similarly to (51.5), a multivariate survival function can be constructed via a copula with

$$S(x_1, \ldots, x_p) = C[S_1(x_1), \ldots, S_p(x_p)] ,$$

where S is the joint survival function and $S_i(t) = 1 - F(t)$ is the i-th marginal survival function, $i = 1, \ldots, p$. In this setting C is called a survival copula. Censoring presents an extra difficulty for multivariate failure-time data analysis. *Georges* et al. [51.46] gives an excellent review on multivariate survival modeling with copulas. This chapter has focused on parametric copula models. Standard inferences of the maximum-likelihood method can be applied under the usual regularity conditions. However, which copula to choose and how well it fits the data are important practical problems. Diagnostic tools, particularly graphical tools, can be very useful. There are not many works in this direction; some recent ones are *Wang* and *Wells* [51.11] and *Fermanian* [51.47].

Copulas have had a long history in the probability literature [51.17]. Recent development and application in insurance, finance and biomedical research have been successful. With this chapter, it is hoped to encourage engineering researchers and practitioners to stimulate more advancement on copulas and seek more applications.

51.A Appendix

51.A.1 The R Package Copula

Overview

Software implementation is very important in promoting the development and application of copula-based approaches. Unfortunately, there are few software packages available for copula-based modeling. One exception is the `finmetrics` module [51.48] of `Splus` [51.49]. For an array of commonly used copulas, the `finmetrics` module provides functions to evaluate their CDF and PDF, generate random numbers from them, and fit them for given data. However, these functionalities are limited because only bivariate copulas are implemented. Furthermore, the software is commercial. It is desirable to have an open-source platform for the development of copula methods and applications.

R is a free software environment for statistical computing and graphics [51.14]. It runs on all platforms, including Unix/Linux, Windows, and MacOS. Cutting-edge statistical developments are easily incorporated into R by the mechanism of contributed packages with quality assurance [51.50]. It provides excellent graphics and interfaces easily with lower-level compiled code such as `C/C++` or `FORTRAN`. An active developer–user interaction is available through the R-help mailing list. Therefore, it is a natural choice to write an R package for copulas.

The package `copula` [51.13] is designed using the object-oriented feature of the `S` language [51.51]. It is publicly available at the Comprehensive R Archive Network [CRAN, `http://www.r-project.org`]. S4 classes are created for elliptical copulas and Archimedean copulas with arbitrary dimension; the extreme-value copula class is still to be implemented at the time of writing. For each copula family, methods of density, distribution, and random-number generator are implemented. For visualization, methods of contour and perspective plots are provided for bivariate copulas.

More facilities, such as extreme-value copulas, association measures and tail dependence measures, will be included in future releases of the package.

Illustration

The package `copula` depends on the contributed packages `mvtnorm`, `scatterplot3d`, and `sn`, taking advantages of the existing facilities in these packages that are relevant. The package needs to be loaded before using:

```
> library(copula)
```

The package is well documented following the requirement of the `R` project [51.50]. A list of help topics can be obtained from:

```
> library(help = copula)
```

We illustrate the features of the package from the following aspects by examples.

Constructing copula objects. An object of class `normalCopula` can be created by

```
> mycop1 <- ellipCopula(family =
"normal", param = c(0.707, 0.5,
0.2), dim = 3, dispstr = "un").
```

The created object `mycop1` is of class `normal Copula`, which inherits `ellipCopula` and `copula`. It has dimension three, with an `unstructured` dispersion matrix

$$\begin{pmatrix} 1.000 & 0.707 & 0.500 \\ 0.707 & 1.000 & 0.200 \\ 0.500 & 0.200 & 1.000 \end{pmatrix}.$$

An object of class `tCopula` can be created similarly, with an extra argument for the degrees of freedom, `df`:

```
> mycop2 <- ellipCopula(family =
"t", param = c(0.9, 0.5, 0.2),
df = 5, dim = 3, dispstr = "un")
```

Examples of objects of Archimedean copulas can be created by:

```
> mycop3 <- archmCopula(family =
"clayton", param = 2, dim = 3)
> mycop4 <- archmCopula(family =
"frank", param = 5.736, dim = 3)
> mycop5 <- archmCopula(family =
"gumbel", param = 2, dim = 3)
```

Constructing Multivariate Distribution via Copulas. An object of multivariate distributions via copulas can be constructed by specifying the copula and its marginal distributions. For example:

```
> mymvd1 <- mvdc(copula =
normalCopula(0.5, dim = 2),
margins = c("norm", "gamma"),
paramMargins = list(list(mean = 0,
sd = 2), list(shape = 2, rate = 2)))
```

The created object `mymvd1` is of class `mvdc`. It is a bivariate distribution constructed via a normal copula. One of the marginal distributions is normal with mean 0 and standard deviation 2. The other marginal distribution is gamma with shape 2 and rate 2.

Density, Distribution, and Simulation. The density and distribution of an object of `copula` class are obtained through the generic method functions `dcopula` and `pcopula`. These functions for an object of the `mvdc` class are obtained through the method functions `dmvdc` and `pmvdc`. The density method `dmvdc` for an `mvdc` object can be used to construct the likelihood for a given dataset.

For Archimedean copulas, obtaining the density function by differentiating the copula can be tedious. The `copula` package provides expressions for the PDF from symbolic calculations. The following code returns the CDF and PDF expressions of a Clayton copula with parameter α:

```
> mycop3@exprdist$cdf
(1 + (u1^(-alpha) - 1 + u2^(-alpha) -
1 + u3^(-alpha) - 1))^(-1/alpha)

> mycop3@exprdist$pdf
(1 + (u1^(-alpha) - 1 + u2^(-alpha) -
1 + u3^(-alpha) - 1))^((((-1/alpha)
- 1) - 1) - 1) * ((((-1/alpha) - 1)
- 1) * (u3^((-alpha) - 1) * (-alpha)))
* (((-1/alpha) - 1) * (u2^((-alpha)
- 1) * (-alpha))) * ((-1/alpha)
* (u1^((-alpha) - 1) * (-alpha)))
```

These can be exported into other programming languages with little or minor modification.

The methods `rcopula` and `rmvdc` generate random numbers from a `copula` or `mvdc` object. The following code generates five observations from `mymvd1` and evaluates the density and distribution at these points:

```
> n <- 5
> x <- rmvdc(mymvd1, n)
> x
          [,1]       [,2]
[1,] -2.7465647 0.6404319
[2,] -1.2674922 0.2707347
[3,] -1.8268522 0.4869647
[4,] 0.2742349 1.1763891
[5,] 2.5947601 1.6410892
```

```
> cbind(dmvdc(mymvd1, x), pmvdc
(mymvd1, x))
           [,1]       [,2]
[1,] 0.06250414 0.06282100
[2,] 0.14514281 0.06221677
[3,] 0.13501126 0.09548434
[4,] 0.10241486 0.45210057
[5,] 0.03698266 0.78582431
```

Bivariate Contour and Perspective Plot.
The `contour` and `persp` methods are implemented for the `copula` and `mvdc` classes. The following code examples draw the contours and perspective plot of the CDF for a bivariate t copula with correlation $\rho = 0.707$:

```
> contour(tCopula(0.707), pcopula)
> persp(tCopula(0.707), pcopula)
```

To draw these plots for an `mvdc` object, the ranges of the margins need to be specified:

```
> persp(mymvd1, dmvdc, xlim =
c(-4, 4), ylim = c(0, 3))
> contour(mymvd1, dmvdc, xlim =
c(-4, 4), ylim = c(0, 3))
```


References

51.1 R. B. Nelsen: *An Introduction to Copulas* (Springer, Berlin Heidelberg New York 1999)
51.2 H. Joe: *Multivariate Models and Dependence Concepts* (Chapman Hall, Norwell 1997)
51.3 C. Genest, J. MacKay: The joy of copulas: Bivariate distributions with uniform marginals (Com: 87V41 P248), Am. Statist. **40**, 280–283 (1986)
51.4 N. I. Fisher: Copulas. In: *Encyclopedia of Statistical Sciences*, ed. by S. Kotz, C. B. Read, D. L. Banks (Wiley, New York 1997) pp. 159–163
51.5 E. W. Frees, J. Carriere, E. A. Valdez: Annuity valuation with dependent mortality, J. Risk Insur. **63**, 229–261 (1996)
51.6 E. W. Frees, E. A. Valdez: Understanding relationships using copulas, North Am. Actuar. J. **2**, 1–25 (1998)
51.7 E. W. Frees, P. Wang: Credibility using copulas, North Am. Actuar. J. **9**, 31–48 (2005)
51.8 E. Bouyè, V. Durrleman, A. Bikeghbali, G. Riboulet, T. Roncalli: *Copulas for Finance – A Reading Guide and Some Applications*, *Working Paper* (Groupe de Recherche Opérationnelle, Crédit Lyonnais, Lyon 2000)
51.9 P. Embrechts, F. Lindskog, A. McNeil: Modelling dependence with copulas and applications to risk management. In: *Handbook of Heavy Tailed Distribution in Finance*, ed. by S. Rachev (Elsevier, Amsterdam 2003) pp. 329–384
51.10 U. Cherubini, E. Luciano, W. Vecchiato: *Copula Methods in Finance* (Wiley, New York 2004)
51.11 W. Wang, M. T. Wells: Model selection and semiparametric inference for bivariate failure-time data (C/R: p73–76), J. Am. Statist. Assoc. **95**, 62–72 (2000)
51.12 G. Escarela, J. F. Carrière: Fitting competing risks with an assumed copula, Statist. Methods Med. Res. **12**, 333–349 (2003)
51.13 J. Yan: *Copula: Multivariate Dependence with Copula, R package version 0.3-3* 2005) CRAN, http://cran.r-project.org
51.14 R Development Core Team: *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna 2005)
51.15 A. W. Sklar: Fonctions de répartition à *n* dimension et leurs marges, Publ. Inst. Statist. Univ. Paris **8**, 229–231 (1959)
51.16 A. Sklar: Random variables, distribution functions, and copulas – A personal look backward and forward. In: *Distributions with Fixed Marginals and Related Topics*, *IMS Lecture Notes Monogr. Ser.*, Vol. 28, ed. by L. Rüschendorf, B. Schweizer, M. D. Taylor (Institute of Mathematical Statistics, Bethesda 1996) pp. 1–14
51.17 B. Schweizer: Thirty years of copulas. In: *Advances in Probability Distributions with Given Margins: Beyond the Copulas*, ed. by G. Dall'Aglio, S. Kotz, G. Salinetti (Kluwer Academic, Dordrecht 1991) pp. 13–50
51.18 P. Embrechts, A. McNeil, D. Straumann: Correlation and dependence in risk management: Properties and pitfalls. In: *Risk Management: Value at Risk and Beyond*, ed. by M. Dempster (Cambridge Univ. Press, Cambridge 2002) pp. 176–223
51.19 K.-T. Fang, S. Kotz, K. W. Ng: *Symmetric Multivariate and Related Distributions* (Chapman Hall, Norwell 1990)
51.20 P. X.-K. Song: Multivariate dispersion models generated from Gaussian copula, Scandin. J. Statist. **27**, 305–320 (2000)
51.21 S. Demarta, A. J. McNeil: The *t* copula and related copulas, Int. Statist. Rev. **73**, 111–129 (2005)
51.22 A. Genz, F. Bretz, T. Hothorn: *Mvtnorm: Multivariate Normal and T Distribution, R package version 0.7-2* 2005) CRAN, http://cran.r-project.org
51.23 A. W. Marshall, I. Olkin: Families of multivariate distributions, J. Am. Statist. Assoc. **83**, 834–841 (1988)
51.24 D. G. Clayton: A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in

chronic disease incidence, Biometrika **65**, 141–152 (1978)
51.25 M.J. Frank: On the simultaneous associativity of $F(x,y)$ and $x+y-F(x,y)$, Aequ. Math. **19**, 194–226 (1979)
51.26 E.J. Gumbel: Bivariate exponential distributions, J. Am. Statist. Assoc. **55**, 698–707 (1960)
51.27 P. Hougaard: A class of multivariate failure time distributions (Corr: V75 p395), Biometrika **73**, 671–678 (1986)
51.28 J.M. Chambers, C.L. Mallows, B.W. Stuck: A method for simulating stable random variables (Corr: V82 P704; V83 P581), J. Am. Statist. Assoc. **71**, 340–344 (1976)
51.29 A.W. Kemp: Efficient generation of logarithmically distributed pseudo-random variables, Appl. Statist. **30**, 249–253 (1981)
51.30 H. Joe, J. Xu: *The Estimation Method of Inference Functions for Margins for Multivariate Models*, *Tech. Rep. 166* (Department of Statistics, University of British Columbia, Vancouver 1996)
51.31 J.H. Shih, T.A. Louis: Inferences on the association parameter in copula models for bivariate survival data, Biometrics **51**, 1384–1399 (1995)
51.32 V.P. Godambe: An optimum property of regular maximum likelihood estimation (Ack: V32 p1343), Annal. Math. Statist. **31**, 1208–1212 (1960)
51.33 C. Genest, K. Ghoudi, L.-P. Rivest: A semiparametric estimation procedure of dependence parameters in multivariate families of distributions, Biometrika **82**, 543–552 (1995)
51.34 C.A. Lowry, D.C. Montgomery: A review of multivariate control charts, IIE Trans. **27**, 800–810 (1995)
51.35 H. Hotelling: Multivariate quality control – Illustrated by the air testing of sample bombsights. In: *Techniques of Statistical Analysis*, ed. by C. Eisenhart, M.W. Hastay, W.A. Wallis (McGraw–Hill, New York 1947) pp. 111–184
51.36 R.L. Mason, J.C. Young: *Multivariate Statistical Process Control with Industrial Applications*, ed. by ASA-SIAM Ser. Statist. Appl. Probab. (SIAM, Philadelphia 2001) p. 263
51.37 R.Y. Liu, J. Tang: Control charts for dependent and independent measurements based on bootstrap methods, J. Am. Statist. Assoc. **91**, 1694–1700 (1996)
51.38 M.-W. Lu, R.J. Rudy: Multivariate control chart. In: *Recent Advances in Reliability and Quality Engineering*, ed. by H. Pham (World Scientific, Singapore 2001) pp. 61–74
51.39 Q.H. Vuong: Likelihood ratio tests for model selection and non-nested hypotheses (STMA V31 0456), Econometrica **57**, 307–333 (1989)
51.40 W.Q. Meeker, L.A. Escobar: *Statistical Methods for Reliability Data* (Wiley, New York 1998)
51.41 C.S. Davis: *Statistical Methods for the Analysis of Repeated Measurements* (Springer, Berlin Heidelberg New York 2002)
51.42 R.C. Littell, G.A. Milliken, W.W. Stroup, R.D. Wolfinger: *SAS System for Mixed Models* (SAS Institute, Cary 1996)
51.43 J.C. Pinheiro, D.M. Bates: *Mixed-Effects Models in S and S-PLUS* (Springer, Berlin, New York 2000)
51.44 W.P. Murray: Archival life expectancy of 3M magneto-optic media, J. Magn. Soc. Jpn. **17**, 309–314 (1993)
51.45 P. Lambert, F. Vandenhende: A copula-based model for multivariate non-normal longitudinal data: Analysis of a dose titration safety study on a new antidepressant, Statist. Med. **21**, 3197–3217 (2002)
51.46 P. Georges, A.-G. Lamy, E. Nicolas, G. Quibel, T. Roncalli: *Multivariate survival modelling: A Unified Approach with Copulas*, *Working Paper* (Groupe de Recherche Opérationnelle, Crédit Lyonnais, Lyon 2001)
51.47 J.-D. Fermanian: Goodness-of-fit tests for copulas, J. Multivariate Anal. **95**, 119–152 (2005)
51.48 Insightful Corp.: *S + Finmetrics Reference Manual* (Insightful, Seattle 2002)
51.49 Insightful Corp.: *S-PLUS (Version 7.0)* (Insightful, Seattle 2005)
51.50 R Development Core Team: *Writing R Extensions* (R Foundation for Statistical Computing, Vienna 2005)
51.51 J.M. Chambers: *Programming with Data: A Guide to the S Language* (Springer, Berlin, New York 1998)