Struct Multidisc Optim (2010) 40:35–51
DOI 10.1007/s00158-009-0385-1

RESEARCH PAPER

Identification of marginal and joint CDFs using Bayesian method for RBDO

Yoojeong Noh · K. K. Choi · Ikjin Lee

Received: 3 September 2008 / Revised: 15 February 2009 / Accepted: 19 March 2009 / Published online: 23 April 2009

Abstract In RBDO, input uncertainty models such as marginal and joint cumulative distribution functions (CDFs) need to be used. However, only limited data exists in industry applications. Thus, identification of the input uncertainty model is challenging especially when input variables are correlated. Since input random variables, such as fatigue material properties, are correlated in many industrial problems, the joint CDF of correlated input variables needs to be correctly identified from given data. In this paper, a Bayesian method is proposed to identify the marginal and joint CDFs from given data where a copula, which only requires marginal CDFs and correlation parameters, is used to model the joint CDF of input variables. Using simulated data sets, performance of the Bayesian method is tested for different numbers of samples and is compared with the goodness-of-fit (GOF) test. Two examples are used to demonstrate how the Bayesian method is used to identify correct marginal CDFs and copula.

Keywords Reliability-based design optimization · Input model uncertainty · Identification of marginal and joint CDFs · Correlated input variables · Copula · Bayesian method · Goodness-of-fit test

Y. Noh · K. K. Choi (✉) · I. Lee
Department of Mechanical & Industrial Engineering, College of Engineering, The University of Iowa, Iowa City, IA 52242, USA
e-mail: kkchoi@engineering.uiowa.edu

Y. Noh
e-mail: noh@engineering.uiowa.edu

I. Lee
e-mail: ilee@engineering.uiowa.edu

Nomenclature

n	number of random variables
$\mathbf{X}$	vector of random variables, $\mathbf{X} = [X_1, \ldots, X_n]^T$
$\mathbf{x}$	realization of vector $\mathbf{X}$, $\mathbf{x} = [x_1, \ldots, x_n]^T$
$F_{X_i}(\cdot)$	marginal CDF of X_i
$F_{X_1 \cdots X_n}(\cdot)$	joint CDF of $X_1, \ldots, X_n$
$f_{X_1, \cdots, X_n}(\cdot)$	joint probability density function (PDF) of $X_1, \ldots, X_n$
θ	matrix of correlation parameters of $X_1, \ldots, X_n$
τ	Kendall's tau
$C(\cdot\|\theta)$	copula with θ
$c(\cdot\|\theta)$	copula density function with θ
$\Phi(\cdot)$	marginal Gaussian CDF
$\Phi^{-1}(\cdot)$	inverse Gaussian CDF
ρ_{ij}	Pearson's correlation coefficient between X_i and X_j
$\mathbf{P}$	covariance matrix consisting of ρ_{ij}
$\Phi_{\mathbf{P}}(\cdot\|\mathbf{P})$	multivariate Gaussian CDF with $\mathbf{P}$
$C_{\Phi}(\cdot\|\mathbf{P})$	Gaussian copula with $\mathbf{P}$

1 Introduction

For RBDO problems, even though input random variables such as material properties and fatigue properties are correlated (Socie 2003; Annis 2004; Efstratios et al. 2004), these variables have often been assumed to be independent because of the difficulty in constructing the joint CDF of correlated input variables. Even when the correlation is considered in the reliability analysis, often the joint Gaussian CDF has been used while the

correct input joint CDF could be non-Gaussian (Nataf 1962; Melchers 1999; Ditlevsen and Madsen 1996; Noh et al. 2008). In addition, certain input variables are known to follow a specific marginal CDF type; for example, some fatigue material properties are known to follow lognormal CDFs. However, when the input marginal CDF type is not known, it is necessary to identify a marginal CDF type from the given data. If an incorrect input joint CDF or incorrect marginal CDFs are used, wrong RBDO results will be obtained. Thus, before carrying out the RBDO, accurate identification of the joint CDF and marginal CDFs is necessary to obtain accurate optimum design results.

To model the joint CDF of input variables, in this paper, it is proposed to use the copula, which is a link between a joint CDF and marginal CDFs. Since the copula requires only marginal CDFs and correlation parameters, which can be obtained from limited data, a joint CDF can be readily modeled by the copula. Furthermore, since the copula decouples marginal CDFs and the joint CDF, it allows the joint CDF type to be different from marginal CDF types. For example, having Gaussian marginal CDFs does not necessarily mean the joint CDF is Gaussian. Thus, it is necessary to use the copula for constructing the input joint CDF with various marginal CDF types.

To correctly identify the marginal CDFs and joint CDF (copula), the Bayesian method, which selects a marginal CDF or copula with the highest normalized weight among candidates based on experimental data (Huard et al. 2006), is used in this paper. By observing the normalized weight, the effectiveness of the Bayesian method is studied for a different number of samples and compared with that of the GOF test. Two mathematical examples and a fatigue problem are used to show how the Bayesian method effectively identifies marginal CDFs and a joint CDF.

2 Modeling of joint CDFs using copulas

As mentioned earlier, when the input variables are correlated, it is difficult to obtain the true joint CDF using only limited experimental data. Thus, in this paper, it is proposed to use the copula to model the input joint CDF using marginal CDFs and correlation measures, which can be obtained from the given experimental data. Section 2.1 provides the definition of the copula, and correlation measures associated with copulas are explained in Section 2.2. Section 2.3 describes commonly used copulas such as the Gaussian and Archimedean copulas.

2.1 Definition of copulas

The word *copula* originated from a Latin word for "link" or "tie" that connects different things. In statistics, the definition of copula is stated by Nelsen (1999): "Copulas are functions that join or couple multivariate distribution functions to their one-dimensional marginal distribution functions. Alternatively, copulas are multivariate distribution functions whose one-dimensional margins are uniform on the interval [0, 1]."

According to Sklar's theorem (Nelsen 1999), if the random variables have marginal CDFs, then there exists an n-dimensional copula C such that

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = C\left(F_{X_1}(x_1),\ldots,F_{X_n}(x_n)\,|\,\theta\right) \tag{1}$$

where θ is the correlation matrix between X_1, …, X_n. If marginal CDFs are all continuous, then C is unique. Conversely, if C is an n-dimensional copula and $F_{X_1}(x_1),\ldots,F_{X_n}(x_n)$ are marginal CDFs, then the joint CDF is an n-dimensional function of marginal CDFs (Nelsen 1999). By taking the derivative of (1), the joint PDF is obtained as

$$f(x_1,\ldots,x_n) = c\left(F_{X_1}(x_1),\ldots,F_{X_n}(x_n)\,|\,\theta\right)\prod_{i=1}^{n} f_{X_i}(x_i) \tag{2}$$

where $c(u_1,\ldots,u_n) = \frac{\partial^n C(u_1,\ldots,u_n)}{\partial u_1\ldots\partial u_n}$ with $u_i = F(x_i)$, and $f_{X_i}(x_i)$ is the marginal PDF for $i = 1,\ldots,n$. To model the joint CDF using the copula, the correlation matrix θ needs to be obtained from experimental data. Since various types of copulas have their own correlation parameters, it is desirable to have a common correlation measure to obtain the correlation parameters from the experimental data.

2.2 Correlation measure

To measure the correlation between two random variables, Pearson's rho and Kendall's tau, can be used. Pearson's rho, which is also called a product moment correlation coefficient, was first discovered by Bravais (1846) and was developed by Pearson (1896). Pearson's rho indicates the degree of linear relationship between two random variables as

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y} \tag{3}$$

where σ_X and σ_Y are standard deviations of X and Y, respectively, and $\mathrm{Cov}(X,Y)$ is the covariance between X and Y. Since Pearson's rho only indicates the linear

relationship between two random variables, it is valid only when the joint CDF is Gaussian. Therefore, it is not a good measure for a nonlinear relationship between two random variables, which often occurs in practical engineering applications.

Unlike Pearson's rho, Kendall's tau does not assume that the relationship between two random variables is linear. Since the Kendall's tau measures the correspondence of rankings between correlated random variables, it is called a rank correlation coefficient. The Kendall's tau was first introduced by Kendall (1938) and is defined as

$$\tau = 4 \int \int_{I^2} C\left(u, v \,|\theta\right) dC\left(u, v\right) - 1 \tag{4}$$

where $I^2 = I \times I\,(I = [0, 1])$, and $u = F_{X_1}(x_1)$ and $v = F_{X_2}(x_2)$ are marginal CDFs of X_1 and X_2. Equation (4) is the population version of Kendall's tau. The sample version of Kendall's tau is

$$t = \frac{c-d}{c+d} = (c-d)/\binom{ns}{2} \tag{5}$$

where c represents the number of concordant pairs, d is the number of discordant pairs, and ns is the number of samples. Using the estimated Kendall's tau, the correlation parameter of the copula, θ, can be calculated because Kendall's tau can be expressed as a function of the correlation parameter as shown in (4). The explicit functions of (4) for some copulas are presented in Table 1. More detailed information on Kendall's tau is presented by Kruskal (1958).

2.3 Commonly used copulas

There are various types of copulas, such as the elliptical copula and the Archimedean copula. In this section, the Gaussian copula, which belongs to an elliptical copula family, and several Archimedean copulas, such as Clayton, Ali-Mikhail-Haq (AMH), Gumbel, Frank, A12, and A14, are introduced as shown in Table 2.

Table 1 Kendall's tau and its domain

Copula	$\tau = g(\theta)$	$\tau \in \Omega^{\tau}$
Clayton	$1 - \frac{2}{2+\theta}$	$(0, 1]$
AMH	$1 - \frac{2}{3}\frac{(\theta-1)^2 \ln(1-\theta) + \theta}{\theta^2}$	$[-0.181726, 1/3]$
Gumbel	$1\text{-}\theta^{-1}$	$[0, 1]$
Frank	$1 - \frac{4}{\theta}\left(1 - \frac{1}{\theta}\int_0^{\theta} \frac{t}{e^t - 1} dt\right)$	$[-1, 1]\backslash\{0\}$
A12	$1 - \frac{2}{3\theta}$	$[1/3, 1]$
A14	$1 - \frac{2}{1+2\theta}$	$[1/3, 1]$
FGM	$2\theta/9$	$[-2/9, 2/9]$
Gaussian	$\frac{2}{\pi}\arcsin\theta$	$[-1, 1]$

Table 2 Copula functions and their parameter domains

Copula	$C(u, v \,\|\theta)$	$\theta \in \Omega^{\theta}$
Clayton	$\left(u^{-\theta} + v^{-\theta} - 1\right)^{-1/\theta}$	$(0, \infty)$
AMH	$uv/[1 - \theta\,(1-u)\,(1-v)]$	$[-1, 1)$
Gumbel	$\exp\left\{-\left[(-\ln u)^{\theta} + (-\ln v)^{\theta}\right]^{1/\theta}\right\}$	$[1, \infty)$
Frank	$-\frac{1}{\theta}\ln\left[1 + \left(e^{-\theta u} - 1\right)\left(e^{-\theta v} - 1\right)/\left(e^{-\theta} - 1\right)\right]$	$(-\infty, \infty)$
A12	$\left\{1 + \left[\left(u^{-1} - 1\right)^{\theta} + \left(v^{-1} - 1\right)^{\theta}\right]^{1/\theta}\right\}^{-1}$	$[1, \infty)$
A14	$\left\{1 + \left[\left(u^{-1/\theta} - 1\right)^{\theta} + \left(v^{-1/\theta} - 1\right)^{\theta}\right]^{1/\theta}\right\}^{-\theta}$	$[1, \infty)$
FGM	$uv + \theta uv\,(1-u)\,(1-v)$	$[-1, 1]$
Gaussian	$\int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{\exp\left(\frac{2\theta sw - s^2 - w^2}{2(1-\theta^2)}\right)}{2\pi\sqrt{1-\theta^2}} dsdw$	$[-1, 1]$

The Gaussian copula is defined as the joint Gaussian CDF of standard Gaussian variables $\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)$ as

$$C_{\Phi}\left(u_1, \cdots, u_n \,|\mathbf{P}\right) = \Phi_{\mathbf{P}}\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n) \,|\mathbf{P}\right), \ \mathbf{u} \in I^n \tag{6}$$

where $u_i = F_{X_i}(x_i)$ is the marginal CDF of X_i for $i = 1, \ldots, n$, and $\mathbf{P}$ is the covariance matrix consisting of correlation coefficients, Pearson's rho, between correlated input variables. $\Phi(\cdot)$ represents the marginal standard Gaussian CDF, $\Phi(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$, and $\Phi_{\mathbf{P}}(\cdot)$ is the joint Gaussian CDF defined as $\Phi_{\mathbf{P}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{P})^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$ for $\mathbf{x} = [x_1, \ldots, x_n]^T$ with a mean vector $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]^T$. If the marginal CDFs are Gaussian, then the joint CDF modeled by the Gaussian copula is the joint Gaussian CDF.

An Archimedean copula is constructed in a completely different way from the Gaussian copula. An important component of constructing Archimedean copulas is a generator function φ_{θ} with θ. For example, the generator function of the Clayton copula is $\varphi_{\theta}(t) = t^{\theta} - 1$. If φ_{θ} is a continuous and strictly decreasing function from [0,1] to [0,∞) such that $\varphi_{\theta}(0) = \infty$ and $\varphi_{\theta}(1) = 0$ and the inverse φ_{θ}^{-1} is completely monotonic on [0,∞), then the Archimedean copula can be defined as

$$C\left(u_1, \cdots, u_n \,|\theta\right) = \varphi^{-1}\left[\varphi_{\theta}(u_1) + \cdots + \varphi_{\theta}(u_n)\right] \tag{7}$$

Each Archimedean copula has a corresponding unique generator function, and the generator function provides

a multivariate copula, as shown in (7). Once the generator function is provided, Kendall's tau can be obtained as

$$\tau = 1 + 4 \int_0^1 \frac{\varphi_\theta(t)}{\varphi'_\theta(t)} dt \tag{8}$$

The Archimedean copula can have a multivariate CDF, but it is hard to expand to an n-dimensional copula because, as shown in (7), it has one generator function and thus has only one correlation parameter even if n variables are correlated. Therefore, most copula applications consider bivariate data. For multivariate data, the data are analyzed pair by pair using the bivariate copula. This paper also considers only the bivariate copula and paired samples.

3 Methods for identification of marginal CDFs

The two most commonly used methods that determine a marginal CDF for the given data are the GOF test and the Bayesian method. The GOF test has been developed and widely used to identify the marginal CDF and calculate its parameters, such as mean and standard deviation. However, since the GOF test relies on the parameters estimated from samples, if the parameters are incorrectly estimated, then the wrong marginal CDF might be identified. On the other hand, since the Bayesian method calculates weights to identify the marginal CDF by integrating the likelihood function over the parameter, it is less dependent on the estimation of the parameter than the GOF test. Thus, the Bayesian method is preferred over the GOF test.

3.1 Goodness-of-fit test

The most natural way of checking the adequacy of a hypothesized CDF is to compare the empirical CDF and the hypothesized CDF. There are several types of GOF tests: Chi-square, Cramér-von Mises, and Kolmogorov–Smirnov (K-S), etc. The Chi-square test, which compares the difference between the empirical PDF and the hypothesized PDF, requires sufficient data, so that it can be used only for a large number of data. The Cramér–von Mises test is based on the integrated difference between an empirical CDF and a hypothesized CDF, weighted by the hypothesized PDF. The Cramér–von Mises test is known as a more powerful method than the K-S test, but its application is limited to the symmetric and right-skewed CDFs, unlike the K-S test, which is applicable to all types of CDFs (Cirrone et al. 2004). Thus, in this paper, the K-S test is compared with the Bayesian method.

The K-S test compares the empirical CDF and the hypothesized (or theoretical) CDF as

$$D_n = \max_x |F_X(x) - G_n(x)| \tag{9}$$

where $F_X(x)$ and $G_n(x)$ are the hypothesized CDF and the empirical CDF, respectively. Since D_n is a mathematically random variable, the CDF of D_n is related to a significance level α as

$$P\left(D_n \le D_n^\alpha\right) = 1 - \alpha \tag{10}$$

for the confidence level, $1 - \alpha$. D_n^α is a critical value obtained from a standard mathematical table presented by Haldar and Mahadevan (2000). The maximum difference between $F_X(x_i)$ and $G_n(x_i)$ of the ith sample x_i for $i = 1, \ldots, ns$ is calculated as

$$D_n^{\max} = \max_{1 \le i \le ns} |F_X(x_i) - G_n(x_i)| \tag{11}$$

If the maximum difference $D_n^{\max}$ calculated from samples is smaller than D_n^α, the null hypothesis that the given samples come from the hypothesized CDF is accepted; otherwise, it is rejected. Likewise, a p-value can be used as a measure to test the null hypothesis. The p-value is a measure of how much evidence we have against the null hypothesis, and is calculated as a Kolmogorov CDF value of $D_n^{\max}$ in the K-S test. If the p-value is larger than α, then the null hypothesis is accepted; otherwise, it is rejected. The larger the p-value, the more strongly the test accepts the null hypothesis. Much like the Bayesian method uses calculated weights of candidate marginal CDFs to identify a marginal CDF, which will be explained in Section 3.2, the K-S test uses calculated p-values of candidate marginal CDFs to identify a marginal CDF in this paper. Accordingly, using the calculated p-values, a marginal CDF with the highest p-value among candidates is selected as the best fitted marginal CDF in the K-S test.

3.2 Bayesian method

In this paper, the Bayesian method is used to identify the correct marginal CDF from candidate CDFs by calculating the weights of the candidate CDFs to select the one with the highest weight. Consider a finite set $s_q \subset s$ consisting of candidate marginal CDFs M_k, $k = 1, \ldots, q$, where s is a set of all marginal CDFs and q is the number of the candidate marginal CDFs. The Bayesian method consists of defining q hypotheses:

h_k: The data come from the marginal CDF M_k, $k = 1, \ldots, q$.

Table 3 Mean and variance

Distribution	μ and σ^2
Gaussian	μ, σ^2
Weibull	$\mu = a\Gamma(1+1/b),\ \sigma^2 = a^2\Gamma(1+2/b) - \mu^2$
Gamma	$\mu = ab,\ \sigma^2 = ab^2$
Lognormal	$\mu = e^{a+b^2/2},\ \sigma^2 = (e^{b^2}-1)e^{2a+b^2}$
Gumbel	$\mu = a + 0.5772b,\ \sigma^2 = b^2\pi^2/6$
Extreme	$\mu = a - 0.5772b,\ \sigma^2 = b^2\pi^2/6$
Extreme type-II	$\mu = b\Gamma(1-1/a)$, $\sigma^2 = b^2[\Gamma(1-2/b) - \Gamma^2(1-1/b)]$

The probability of each hypothesis h_k given the data D is defined as

$$\Pr(h_k|D, I) = \frac{\Pr(D|h_k, I)\Pr(h_k|I)}{\Pr(D|I)} \tag{12}$$

where $\Pr(D|h_k,I)$ is the likelihood function, $\Pr(h_k|I)$ is the prior on the marginal CDF, and $\Pr(D|I)$ is the normalization constant with any relevant additional knowledge I. The additional knowledge I is explained in detail in Section 3.2.2.

3.2.1 Likelihood function

Under the hypothesis h_k that the data D come from the marginal CDF M_k, the probability of drawing the data D for the hypothesis on M_k is expressed as a likelihood function as

$$\Pr(D|h_k, \mu, \sigma, I) = \prod_{i=1}^{ns} f_k(x_i|a(\mu,\sigma), b(\mu,\sigma)) \tag{13}$$

where x_i is the ith sample value. Since each marginal PDF f_k has its own parameters a and b, common parameters, such as mean or standard deviation, need to be used. For most marginal CDFs, their own parameters (a and b) are expressed as functions of mean and standard deviation as shown in Table 3, and thus the likelihood function can be expressed in terms of mean and standard deviation for the given samples. In Table 3, the domains of mean for Gaussian, Gumbel, and Extreme type-II distributions have $\Omega^\mu = (-\infty, \infty)$ and those for other distributions have $\Omega^\mu = (0, \infty)$. All distributions have the same domain of the standard deviation, $\Omega^\sigma = (0, \infty)$.

Introducing the mean or standard deviation as the nuisance variable, (12) can be rewritten as

$$\begin{aligned}\Pr(h_k|D, I) &= \int_{-\infty}^{\infty} \Pr(h_k, \mu, \sigma|D, I)\, d\mu \\ &= \int_{-\infty}^{\infty} \frac{\Pr(D|h_k, \mu, \sigma, I)\Pr(h_k|\mu, I)\Pr(\mu|I)}{\Pr(D|I)} d\mu\end{aligned} \tag{14}$$

or

$$\begin{aligned}\Pr(h_k|D, I) &= \int_0^{\infty} \Pr(h_k, \mu, \sigma|D, I)\, d\sigma \\ &= \int_0^{\infty} \frac{\Pr(D|h_k, \mu, \sigma, I)\Pr(h_k|\sigma, I)\Pr(\sigma|I)}{\Pr(D|I)} d\sigma\end{aligned} \tag{15}$$

In (14), $\Pr(h_k, \mu, \sigma|D, I)$ is a function of mean with the standard deviation calculated from samples. Conversely, in (15), $\Pr(h_k, \mu, \sigma|D, I)$ is a function of standard deviation with the mean calculated from samples. To calculate (12), (13) is used as the likelihood function, and the candidate marginal CDFs in the set s_q are Gaussian, Weibull, Gamma, Lognormal, Gumbel, Extreme, and Extreme Type II in this paper. The formulas of candidate marginal PDFs are shown in Table 4 with domains Ω^a and Ω^b of parameters a and b, respectively.

3.2.2 Priors

Let the additional information I be as follows:

I_1: Mean or standard deviation belongs to the set Λ^μ or Λ^σ, respectively, and each estimated $\mu \in \Lambda^\mu$ or $\sigma \in \Lambda^\sigma$ is equally likely;

I_2: For given μ or σ, all marginal CDFs satisfying $\mu \in \Omega_k^\mu$ or $\sigma \in \Omega_k^\sigma$ are equally probable, where Ω_k^μ and Ω_k^σ are domains of μ and σ for the marginal CDF M_k.

The set Λ^μ or Λ^σ provides information on the interval of mean or standard deviation, respectively, that the user might know. For example, if the user knows the specific domain of Λ^μ or Λ^σ, the domain can be used to integrate the likelihood function for calculation of the weights of each candidate marginal CDF. However, if information on the specific domain of Λ^μ or Λ^σ is not known, it can be assumed that $\Lambda^\mu = (-\infty, \infty)$ or $\Lambda^\sigma = (0, \infty)$. In that case, the infinite domain cannot practically be used to integrate the likelihood function, and thus a finite range of Λ^μ or Λ^σ needs to be determined from samples such that Λ^μ or Λ^σ cover the wide range of the parameter. Using the first additional information I_1, the prior on μ or σ can be defined as

$$\Pr(\mu|I_1) = \begin{cases} \dfrac{1}{\lambda(\Lambda^\mu)}, & \mu \in \Lambda^\mu \\ 0, & \mu \notin \Lambda^\mu \end{cases} \tag{16}$$

or

$$\Pr(\sigma|I_1) = \begin{cases} \dfrac{1}{\lambda(\Lambda^\sigma)}, & \sigma \in \Lambda^\sigma \\ 0, & \sigma \notin \Lambda^\sigma \end{cases} \tag{17}$$

Table 4 Marginal PDFs and domains of parameters

Distribution	$f(x \mid a, b)$	$a \in \Omega^a$	$b \in \Omega^b$
Gaussian	$\frac{1}{b\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-a}{b}\right)^2\right]$	$(-\infty, \infty)$	$(0, \infty)$
Weibull	$\frac{b}{a}\left(\frac{x}{a}\right)^{b-1} \exp\left[-\left(\frac{x}{a}\right)^b\right]$	$(0, \infty)$	$(0, \infty)$
Gamma	$x^{a-1}\frac{\exp[-x/b]}{\Gamma(a)\, b^a}$	$(0, \infty)$	$(0, \infty)$
Lognormal	$\frac{1}{bx\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln x - a}{b}\right)^2\right]$	$(-\infty, \infty)$	$(0, \infty)$
Gumbel	$\frac{1}{b} \exp\left[-\frac{x-a}{b} - \exp\left[-\frac{x-a}{b}\right]\right]$	$(-\infty, \infty)$	$(0, \infty)$
Extreme	$\frac{1}{b} \exp\left[\frac{x-a}{b} - \exp\left[\frac{x-a}{b}\right]\right]$	$(-\infty, \infty)$	$(0, \infty)$
Extreme type-II	$\frac{a}{b}\left[\frac{b}{x}\right]^{a+1} \exp\left[-\left(\frac{b}{x}\right)^a\right]$	$(-\infty, \infty)$	$(0, \infty)$

where $\lambda(\cdot)$ is the Lebesgue measure, which is the interval length of Λ^μ or Λ^σ. Likewise, since all marginal CDFs are equally probable for $\mu \in \Omega_k^\mu$ or $\sigma \in \Omega_k^\sigma$, the prior on the marginal CDF, M_k, is defined as

$$\Pr(h_k \mid \mu, I_2) = \begin{cases} 1, & \mu \in \Omega_k^\mu \\ 0, & \mu \notin \Omega_k^\mu \end{cases} \tag{18}$$

or

$$\Pr(h_k \mid \sigma, I_2) = \begin{cases} 1, & \sigma \in \Omega_k^\sigma \\ 0, & \sigma \notin \Omega_k^\sigma \end{cases} \tag{19}$$

In this paper, it is assumed that the prior follows a uniform distribution, which means there is no information on the distribution of μ or σ. If it is known that the prior of μ or σ follows a specific distribution, $\Pr(\mu)$ or $\Pr(\sigma)$ might be expressed as a PDF and can be used as the prior instead of (16) or (17). However, since the prior of μ or σ is usually unknown and the effect of the prior is negligible when the number of samples is enough (i.e., larger than 100 samples), (16) or (17) can be used in most cases.

3.2.3 Normalization of weights

Substituting (13) and (16)–(19) into (14) and (15), (14) and (15) can be rewritten as

$$\Pr(h_k \mid D, I) = \frac{\int_{\Omega_k^\mu \cap \Lambda^\mu} \prod_{i=1}^{ns} f_k(x_i \mid a(\mu, \sigma), b(\mu, \sigma)) d\mu}{\lambda(\Lambda^\mu) \Pr(D \mid I)} \tag{20}$$

or

$$\Pr(h_k \mid D, I) = \frac{\int_{\Omega_k^\sigma \cap \Lambda^\sigma} \prod_{i=1}^{ns} f_k(x_i \mid a(\mu, \sigma), b(\mu, \sigma)) d\sigma}{\lambda(\Lambda^\sigma) \Pr(D \mid I)} \tag{21}$$

In (20) and (21), $\Pr(D|I)$ can be expressed as

$$\Pr(D \mid I) = \sum_{k=1}^{q} \Pr(D \mid h_k, I) \Pr(h_k \mid I) \tag{22}$$

Since $\Pr(D|I)$ is a constant, for convenience, it is not included to calculate weights in this paper. Accordingly, (20) and (21) can be expressed as

$$W_k = \frac{1}{\lambda(\Lambda^\mu)} \int_{\Omega_k^\mu \cap \Lambda^\mu} \prod_{i=1}^{ns} f_k(x_i \mid a(\mu, \sigma), b(\mu, \sigma)) d\mu \tag{23}$$

or

$$W_k = \frac{1}{\lambda(\Lambda^\sigma)} \int_{\Omega_k^\sigma \cap \Lambda^\sigma} \prod_{i=1}^{ns} f_k(x_i \mid a(\mu, \sigma), b(\mu, \sigma)) d\sigma \tag{24}$$

The normalized weight for the marginal CDF M_k is calculated as

$$w_k = \frac{W_k}{\sum_{i=1}^{q} W_i} \tag{25}$$

In the Bayesian method for identifying the marginal CDF, there are two approaches that use mean or standard deviation as variables for calculating normalized weights, and thus it might be necessary to select one

Table 5 Averaged normalized weights over 100 trials using two Bayesian approaches

Original distribution	*ns*	$\mu = 2$		$\mu = 10$	
		Mean	Std.	Mean	Std.
Gaussian	30	0.300	0.273	0.264	0.232
	100	0.462	0.454	0.347	0.308
	300	0.731	0.702	0.427	0.389
Weibull	30	0.328	0.261	0.300	0.310
	100	0.544	0.484	0.495	0.477
	300	0.706	0.705	0.636	0.573
Gamma	30	0.240	0.213	0.246	0.200
	100	0.410	0.322	0.339	0.301
	300	0.750	0.617	0.347	0.337
Lognormal	30	0.240	0.233	0.253	0.200
	100	0.362	0.362	0.332	0.291
	300	0.597	0.558	0.399	0.383
Gumbel	30	0.230	0.224	0.308	0.327
	100	0.395	0.388	0.465	0.440
	300	0.636	0.590	0.651	0.616
Extreme	30	0.552	0.602	0.374	0.403
	100	0.885	0.927	0.497	0.527
	300	0.993	1.000	0.584	0.565
Extreme type-II	30	0.481	0.586	0.338	0.392
	100	0.778	0.787	0.465	0.516
	300	0.875	0.916	0.605	0.669

method. To compare the performance of the two approaches, averaged normalized weights over 100 trials are used.

Let "Mean" and "Std." be the methods using mean and standard deviation, respectively, as variables for the calculation of weights. Table 5 shows the averaged normalized weights for different means and for different samples when Gaussian, Weibull, Gamma, Lognormal, Gumbel, Extreme, and Extreme Type II are the original CDFs. The larger the normalized weights, the better identified each original CDF is. When Gaussian, Weibull, Gamma, Lognormal, and Gumbel are the original CDFs, the normalized weights using "Mean" are slightly better than those using "Std." On the other hand, when Extreme and Extreme Type II are the original CDFs, the normalized weights using "Std." are slightly better than those using "Mean." For all cases, since one method is not always better than the other approach and the normalized weights calculated from two approaches are similar, both approaches can be used. However, "Mean" is better than "Std." in more cases, so "Mean" is used in this paper.

3.3 Comparison of two methods

It is stated that the Bayesian method performs better in identifying the correct marginal CDF than the GOF test, but it is still valuable to compare the two methods numerically.

Consider a set of random data with a different number of samples: $ns = 30$, 100, and 300, given that the original distribution is Gamma with GM (400, 0.025). The Gaussian, Weibull, Gumbel, Lognormal, Extreme, and Extreme Type II are selected as candidate distributions. All parameters of each CDF are calculated from $\mu = 10.0$ and $\sigma = 0.5$ using Table 3. The PDF shapes of the candidate distributions are shown in Fig. 1. In the GOF test, p-values are calculated and used to test the null hypothesis of each candidate CDF. In the Bayesian method, normalized weights are calculated to identify the correct CDF. To compare the two methods, the p-value and normalized weights should be comparable, but they are not. The p-value is the probability of obtaining a *value* of the test statistic (D_n) at least as extreme as the actually observed value, given that the null hypothesis is true. On the other hand, it is not the probability of the hypothesis (Sterne and Smith 2000) from which the normalized weights originate. Thus, instead of directly using two values, it might be better to observe how many p-values and normalized weights are assigned to correct marginal CDF or CDFs with shapes similar to the correct one.

Table 6 shows the sum of p-values and normalized weights over 100 trials when both methods are used. For a small number of samples, such as 30, the Bayesian method assigns most normalized weights to Gaussian, Gamma, and Lognormal distributions because the PDF shapes of Gaussian and Lognormal are very close to Gamma for the given mean $\mu = 10$, as shown in Fig. 1. On the other hand, the GOF test assigns p-values to all candidate distributions almost evenly, which means the

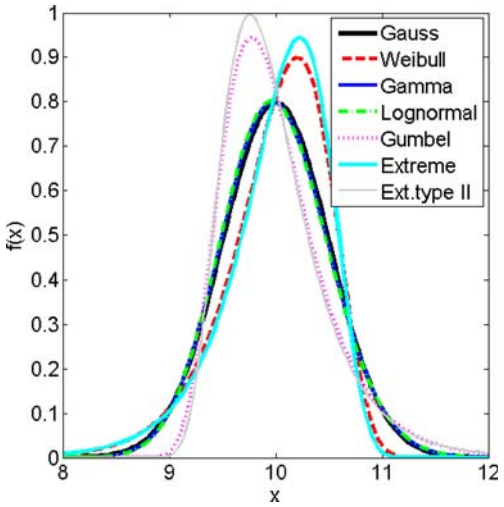

Fig. 1 Marginal PDFs for $\mu = 10.0$ and $\sigma = 0.5$

Table 6 Comparison of GOF with Bayesian method for identification of gamma marginal CDF

Distribution	$ns = 30$		$ns = 100$		$ns = 300$	
	GOF	Bay.	GOF	Bay.	GOF	Bay.
Gaussian	16.3	24.1	21.9	32.9	29.3	31.0
Weibull	13.1	8.59	10.4	1.12	3.29	0.00
Gamma	16.4	24.6	22.0	33.0	30.8	34.2
Lognormal	16.5	24.7	21.9	32.4	30.7	34.8
Gumbel	13.4	7.46	9.78	0.42	3.01	0.00
Extreme	12.2	6.08	8.01	0.20	1.56	0.00
Extreme type II	12.0	4.45	5.97	0.00	1.39	0.00

GOF test does not identify correct CDFs for a small number of samples. Further, since the p-values of all distributions are larger than 5 for a given significance level, 5%, the GOF test accepts all distributions as correct even though the PDF shapes of Weibull, Gumbel, Extreme, and Extreme Type II are quite different from Gamma (original), as shown in Fig. 1. For a large number of samples, such as 300, the GOF test correctly identifies Gaussian, Gamma, and Lognormal CDFs. Still, the Bayesian method shows better performance than the GOF tests in this example. It might be some concern that the normalized weight of the original distribution is still not high, even for a large number of samples. However, the PDF shapes of those distributions are almost identical even at the tail end, which lead to similar RBDO results. Thus, it does not matter which of the three distributions is selected for this example.

4 Methods for identification of copulas

Just as two methods—the GOF test and the Bayesian method—can be used to identify marginal CDFs, two methods can be used to identify the correct copula.

4.1 Goodness-of-fit test

The GOF test compares the empirical copula C_{ns} calculated from the given ns samples and a theoretical copula $C_{\theta_{ns}}$ with some parameter θ_{ns} calculated from the data. The GOF test for identification of copula can be carried out by using the parametric bootstrap (Genest and Rémillard 2005). Using the Cramér–von Mises statistic S_{ns}, the null hypothesis H_0 (that is, the data follow a specific copula type C_k) can be tested by

$$S_{ns} = \sum_{i=1}^{ns} \left\{ C_{ns}\left(\hat{U}_i\right) - C_{\theta_{ns}}\left(\hat{U}_i\right) \right\}^2 \tag{26}$$

where $\hat{U}_i = R_i/(ns+1)$ for $i = 1, \ldots, ns$ is the empirical marginal CDF value, and $R_i = [R_{i1}, \ldots, R_{in}]^T$ is the rank of the samples for $X_1, \ldots, X_n$, and the empirical copula is defined as

$$C_{ns}(u) = \frac{1}{ns} \sum_{i=1}^{ns} 1\left(\hat{U}_i \le u\right), \; u \in [0,1]^n \tag{27}$$

where $1\left(\hat{U}_i \le u\right)$ indicates "1" if $\hat{U}_i \le u$; otherwise, it is "0". After obtaining the Cramer–von Mises statistic S_{ns} from the samples, the empirical copula $C^*_{ns,k}$ and the Cramér–von Mises statistic $S^*_{ns,k}$ need to be repeatedly estimated for every $k = 1, \ldots, N$ where N is a large number (e.g., 1,000–10,000).

As a result, an approximate p-value for the test based on the Cramer–von Mises statistic S_{ns} is given by

$$p = \frac{1}{N} \sum_{k=1}^{ns} 1\left(S^*_{ns,k} > S_{ns}\right) \tag{28}$$

where $1\left(S^*_{ns,k} > S_{ns}\right)$ indicates "1" if $S^*_{ns,k} > S_{ns}$; otherwise, it is "0". For some significance levels, such as 5%, according to the p-value, the null hypothesis H_0 might be accepted or rejected. When the number of random variables is larger than three, a two-level parametric boot strap similar to the previously described one-level boot strap needs to be used. The detailed algorithm is implemented by Genest and Rémillard (2005).

In recent years, various other GOF tests have been developed. For Archimedean copulas, a GOF test statistic that compares the one-dimensional empirical function K_{ns} and the theoretical function $K_{\theta_{ns}}$ is proposed as (Genest and Rivest 1993; Genest and Favre 2007)

$$\mathrm{K}(t) = \sqrt{ns}\left\{K_{ns}(t) - K_{\theta_{ns}}(t)\right\} \tag{29}$$

where $K_\theta(t) = \Pr(C(u, v\,|\theta) < t)$ and $K_{ns}(t) = \frac{1}{ns}\sum_{k=1}^{ns} 1\left(V_{k,ns} \le t\right)$ with $V_{k,ns} = \frac{1}{ns}\sum_{j=1}^{ns} 1\left(R_{1j} \le R_{1k}, R_{2j} \le R_{2k}\right)$. As in the previous parametric bootstrap, using another type of the statistic K, the p-value can be calculated.

The GOF tests suggest a way to identify the correct copula, but it depends on the estimation of the correlation parameter θ_{ns}. Further, for the second method, since the function K is explicitly expressed only for Archimedean copulas, it cannot be used for non-Archimedean copulas because the function K cannot be algebraically formulated for non-Archimedean copulas (Genest and Rémillard 2005).

4.2 Bayesian method

Just as the Bayesian method is used to identify the correct marginal CDFs, it can also be used to identify the correct copula by calculating the weights of the candidate copulas to select the one with the highest weight. To calculate the normalized weights, Kendall's tau is used as the variable for integrating the likelihood function.

Let a finite set $S_Q \subset S$ consist of candidate copulas C_k, $k = 1, \ldots, Q$ where S is a set of all copulas and Q is the number of the candidate copulas. The Bayesian method consists of defining Q hypotheses (Huard et al. 2006):

H_k: The data come from copula C_k, $k = 1, \ldots, Q$.

The probability of each hypothesis H_k given the data D is defined as

$$\Pr(H_k \,|D, I) = \frac{\Pr(D \,|H_k, I)\Pr(H_k \,|I)}{\Pr(D \,|I)} \tag{30}$$

where $\Pr(D|H_k, I)$ is the likelihood function, $\Pr(H_k|I)$ is the prior on the copula family, and $\Pr(D|I)$ is the normalization constant with any relevant additional knowledge I, which will be explained in Section 4.2.2.

4.2.1 Likelihood function

Under the hypothesis H_k that the data D come from the copula C_k, the probability of drawing the data D for the hypothesis on C_k is expressed as a likelihood function as

$$\Pr(D \,|H_k, \tau, I) = \prod_{i=1}^{ns} c_k(u_i, v_i \,|\theta) \tag{31}$$

where (u_i, v_i) are ns mutually independent pairs of the data and calculated as $u_i = F_X(x_i)$ and $v_i = F_Y(y_i)$, where $F_X(x_i)$ and $F_Y(y_i)$ are the marginal CDF values obtained from the given paired data (x_i, y_i). Since it is assumed that the data D come from the copula C_k, the probability of drawing D from the copula C_k (likelihood function) is expressed as a copula density function c_k. The paired data are independent of each other, so that the likelihood function is expressed as multiplications of the copula density function values evaluated at all the data. Since each copula C_k has its own correlation parameter θ, a common correlation measure, Kendall's tau, needs to be used. Using the relationship between the parameter and Kendall's tau $\tau = g_k(\theta)$ for $k = 1, \ldots, Q$, as shown in Table 1, the correlation parameter can be expressed as $\theta = g_k^{-1}(\tau)$.

Using the Kendall's tau as the nuisance variable, (30) can be rewritten as (Huard et al. 2006)

$$\begin{aligned}\Pr(H_k \,|D, I) &= \int_{-1}^{1} \Pr(H_k, \tau \,|D, I)\, d\tau \\ &= \int_{-1}^{1} \frac{\Pr(D \,|H_k, \tau, I)\Pr(H_k \,|\tau, I)\Pr(\tau \,|I)}{\Pr(D \,|I)} d\tau \end{aligned} \tag{32}$$

where (31) is used as the likelihood function and the candidate copulas in the set S_Q are Clayton, AMH, Gumbel, Frank, A12, A14, Farlie–Gumbel–Morgenstern (FGM), Gaussian, and independent.

4.2.2 Priors

Let the additional information I on the copula be as follows:

I_1: Kendall's tau belongs to the set Λ^τ, and each estimated $\tau \in \Lambda^\tau$ is equally likely;
I_2: For a given τ, all copula families satisfying $\tau \in \Omega_k^\tau$ are equally probable, where Ω_k^τ is the domain of τ for C_k.

The set Λ^τ provides information on the interval of Kendall's tau that the user might know. For example, based on the user's experience, it might be known that the range of Kendall's tau estimated between two interesting variables can have only a positive range $\Lambda^\tau = [0, 1]$. However, if information on the correlation parameter between variables is not known, it can be assumed as $\Lambda^\tau = [-1, 1]$. Using the first additional information I_1, the prior on τ can be defined as (Huard et al. 2006)

$$\Pr(\tau|\, I_1) = \begin{cases} \frac{1}{\lambda(\Lambda^\tau)}, & \tau \in \Lambda^\tau \\ 0, & \tau \notin \Lambda^\tau \end{cases} \tag{33}$$

where $\lambda(\cdot)$ is the Lebesgue measure, which is the interval length of Λ^τ.

Likewise, since all copula families are equally probable for $\tau \in \Omega_k^r$, the prior on the copula family is defined as

$$\Pr(H_k \,|\tau, I_2) = \begin{cases} 1, & \tau \in \Omega_k^\tau \\ 0, & \tau \notin \Omega_k^\tau \end{cases} \tag{34}$$

If it is known that the prior distribution of τ follows a certain distribution, $\Pr(\tau)$ might be expressed as a specific distribution and can be used as the prior instead of (33). However, the prior of τ is usually unknown, so that (33) is commonly used in practical applications.

Table 7 Comparison of GOF test with Bayesian method for identification of copula

Copula	$ns = 30$		$ns = 100$		$ns = 300$	
	GOF	Bay.	OF	Bay.	GOF	Bay.
Clayton	17.5	39.7	34.8	53.4	41.7	68.0
Gumbel	13.4	5.55	0.39	0.00	0.00	0.00
Gaussian	16.3	9.95	0.07	2.13	0.00	0.00
Frank	16.2	9.81	2.58	2.45	0.00	0.00
A12	18.2	20.1	31.0	25.9	34.9	21.6
A14	18.3	14.9	30.5	16.0	23.3	10.4

4.2.3 Normalization of weights

Substituting (31), (33), and (34) into (32), (32) can be rewritten as

$$\Pr(H_k \,|D, I) = \frac{\int_{\Omega_k^\tau \cap \Lambda^\tau} \prod_{i=1}^{ns} c_k\left(u_i, v_i \,|g_k^{-1}(\tau)\right) d\tau}{\lambda(\Lambda^\tau) \Pr(D|\, I)} \tag{35}$$

where $\Pr(D|I)$ is expressed as (Huard et al. 2006)

$$\Pr(D \,|I) = \sum_{l=1}^{Q} \Pr(D \,|H_k, I) \Pr(H_k \,|I) \tag{36}$$

In this paper, since $\Pr(D|I)$ is a constant, it is not included for convenience. Accordingly, the computation of (35) can be expressed as the computation of the weights as

$$W_k = \frac{1}{\lambda(\Lambda^\tau)} \int_{\Omega_k^\tau \cap \Lambda^\tau} \prod_{i=1}^{ns} c_k\left(u_i, v_i \,|g_k^{-1}(\tau)\right) d\tau \tag{37}$$

The normalized weight of C_k is calculated as

$$w_k = \frac{W_k}{\sum_{i=1}^{Q} W_k} \tag{38}$$

Since the Bayesian method selects one marginal CDF or copula that best describes the given data among candidates, the identified marginal CDF, or copula might not be the correct one. On the other hand, since some commonly used marginal CDFs are known and the total number of marginal CDF types is not large, it is easy to determine the candidate marginal CDFs and identify a correct CDF among them. However, since there exist plenty of copula types, it might be possible that the data come from an unknown copula that is not among the given candidate copulas. A way of solving this problem is presented by Bretthorst (1996).

4.3 Comparison of two methods

Just as the GOF test and the Bayesian method are compared for identification of marginal CDFs, two methods for identification of copulas are compared in this section. Consider a set of random data with a different number of samples: $ns = 30$, 100, and 300 for the original copula as Clayton with $\tau = 0.4$ where Clayton, Gumbel, Gaussian, Frank, A12, and A14 are selected as candidate copulas.

Table 7 shows the sum of p-values and normalized weights over 100 trials. As shown in the table, for a small number of samples, the Bayesian method assigns 39.7 to the Clayton copula, which is the original copula, whereas the GOF test assigns only 17.5 to the correct copula. Further, the Bayesian assigns the normalized weights according to the similarity of the copula shape, i.e., A12 is the most similar with Clayton, A14 is the second most similar, and so on.

On the other hand, the GOF test accepts all candidate copulas as correct copulas on the average (p-values are larger than 5), even though some copulas such as Gumbel have very distinct shapes with the Clayton copula. As the number of samples is increased to 300, the performance of the GOF test is improved, but the Bayesian method is still better than the GOF test at identifying the copula in this example.

5 Effectiveness of Bayesian method

In this section, the effectiveness of the Bayesian method for identification of marginal CDFs and copulas is studied for different numbers of samples.

5.1 Identification of marginal CDF

Given Gaussian, Weibull, Gamma, Lognormal, Gumbel, Extreme, and Extreme type II CDFs, the Bayesian method is tested to identify original CDFs using normalized weights over 100 trials where the number of samples is 30, 100, and 300. Figure 2 shows the sum of normalized weights over 100 trials for different numbers of samples when original CDFs, indicated by boxes on the name of the marginal CDFs, are given with $\mu = 2.0$ and $\sigma = 0.5$. For example, when the Gaussian CDF is original, samples are randomly generated from the Gaussian CDF 100 times. Using the Bayesian method, the normalized weights can be calculated from the 100 data set. Adding up the normalized weights over 100 trials, the sum of normalized weights is approximately 30 for Gaussian, 25 for Weibull, and 45 for the rest

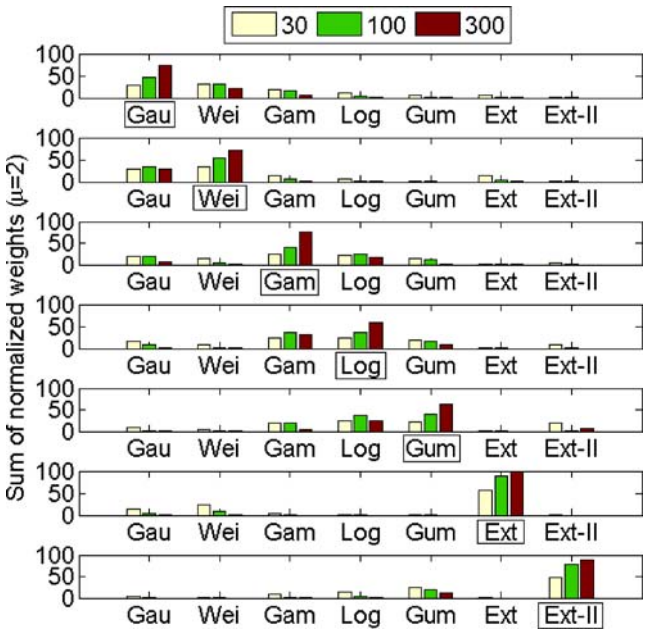

Fig. 2 Sum of normalized weights over 100 trials for $\mu = 2.0$ and $\sigma = 0.5$

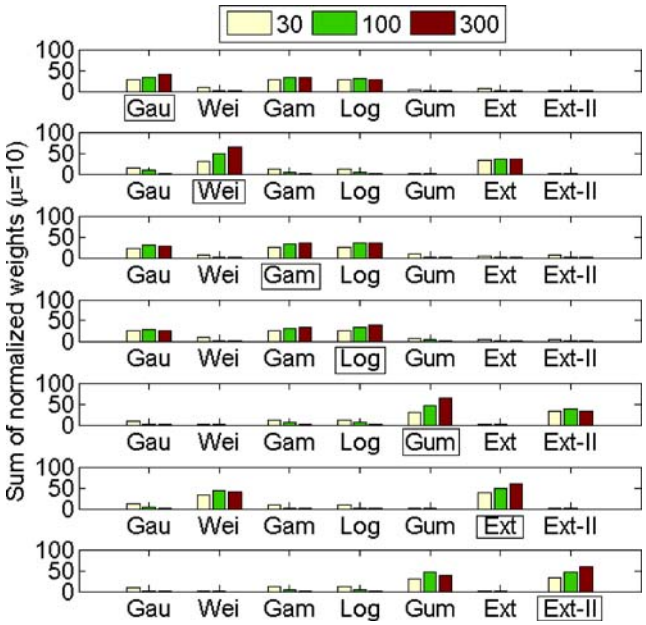

Fig. 4 Sum of normalized weights over 100 trials for $\mu = 10.0$ and $\sigma = 0.5$

of the CDFs when $ns = 30$. Since the PDF shapes of Gaussian and Weibull CDFs are similar, as shown in Fig. 3, the Weibull distribution has the second highest sum of normalized weight among candidate marginal CDFs. Likewise, when Weibull is the original CDF, the Gaussian distribution has the second highest sum of normalized weight. On the other hand, when the CDFs with distinct PDF shapes such as Extreme and Extreme Type II are original, it is much easier to identify the original CDF.

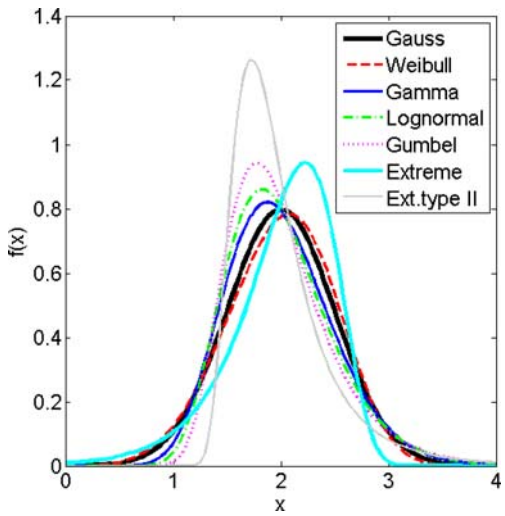

Fig. 3 Marginal PDFs for $\mu = 2.0$ and $\sigma = 0.5$

When $\mu = 2.0$ is changed to 10.0 with the same $\sigma = 0.5$, the PDF shapes become different than those shown in Fig. 3. Figure 4 shows the sum of normalized weights over 100 trials when $\mu = 10.0$ and $\sigma = 0.5$. For the CDFs with distinct PDF shapes, such as Weibull, Gumbel, Extreme, and Extreme Type II, it is easy to identify the correct CDF. However, when Gaussian, Gamma, and Lognormal Distributions are original, since the Gaussian, Gamma, and Lognormal CDFs are almost identical even at the tail ends, as shown in Fig. 1, it is difficult to identify the correct CDF among three CDFs, and their normalized weights are rather similar to each other even if the number of samples is increased to 300. The similar PDF shapes provide similar RBDO results, so that it does not matter which of the three CDFs is selected in this case.

On the other hand, since the identified distribution generally fits to the given data, its tail behavior might be different from the one of the original CDF particularly when the number of samples is small. If the tail behavior of the identified distribution is quite different from the one of the original distribution, different tail behaviors yield different contours for a target reliability index, which lead to different RBDO results. In practical applications, since experimental data is very limited, it could be hard to select a distribution that has the same tail as the original distribution. For this, RBDO with a confidence level of the input model uncertainty is currently being investigated so that reliable optimum designs can be obtained.

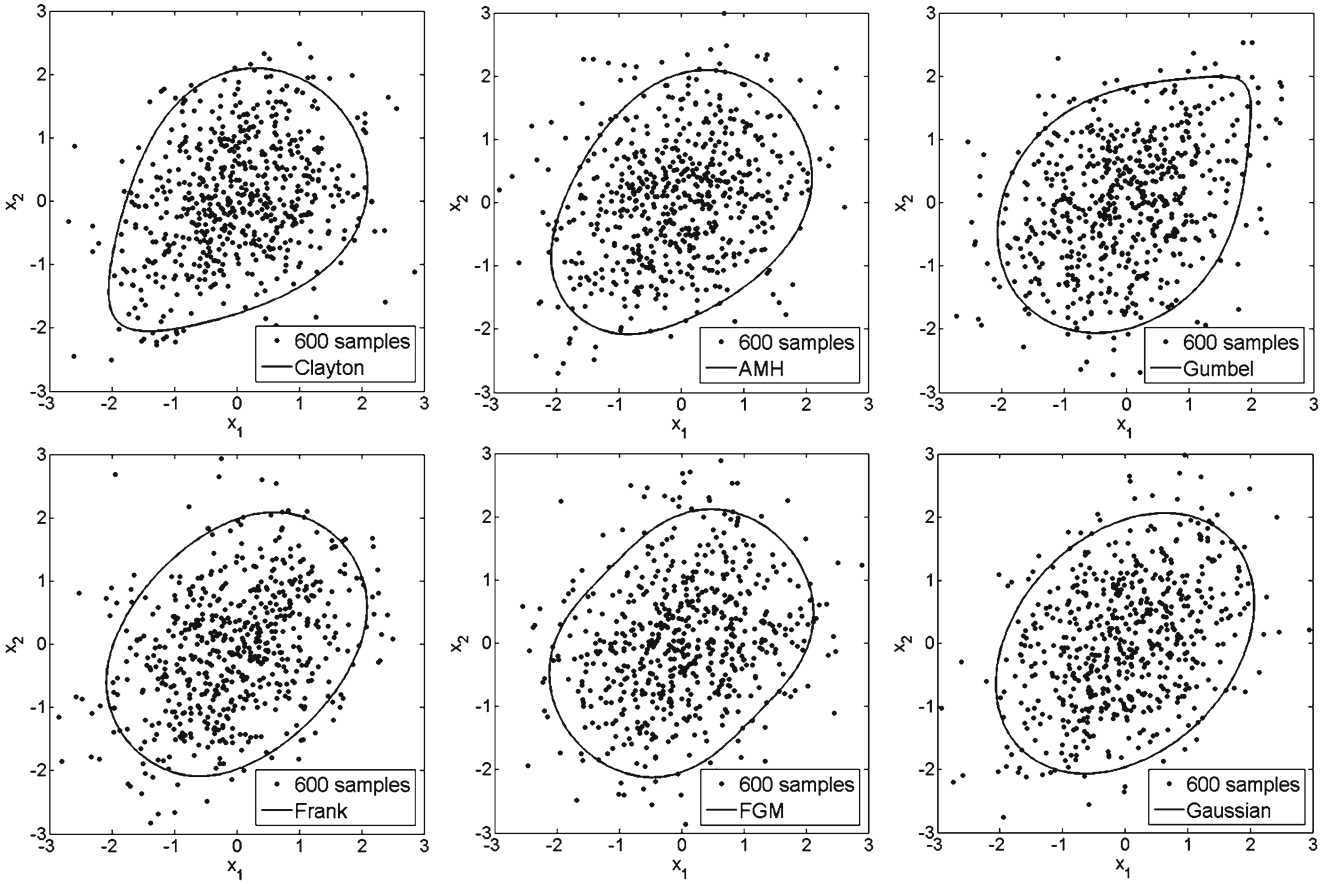

Fig. 5 PDF contours and scatter plots of samples ($ns = 600$) for Clayton, AMH, Gumbel, Frank, FGM, and Gaussian with $\tau = 0.2$

5.2 Identification of copula

In this study, different values of Kendall's tau, i.e., $\tau = 0.2$, and 0.7, are used to study the effect of Kendall's tau on identification of the correct copulas. The candidate copulas are selected as Clayton, AMH, Gumbel, Frank, A12, A14, FGM, Gaussian, and an independent copula, which can be expressed as the multiplication of marginal CDFs, $C(u, v) = uv$. Figures 5 and 6 show the PDF contour and sum of normalized weights over 100 trials when the original copula, which is indicated by a box on the name of each copula for $\tau = 0.2$. Likewise, Figs. 7 and 8 show the PDF contour and sum of normalized weights for $\tau = 0.7$.

For small correlation coefficients such as $\tau = 0.2$, since the PDF contours of most copulas are similar to each other, except Clayton and Gumbel, as shown in Fig. 5, it is not simple to identify the correct one. For instance, even though the original copula is AMH, the normalized weights of incorrect copulas such as the independent copula are high, especially for a small number of samples, $ns = 30$ as shown in Fig. 6. Therefore, a large number of samples is generally required

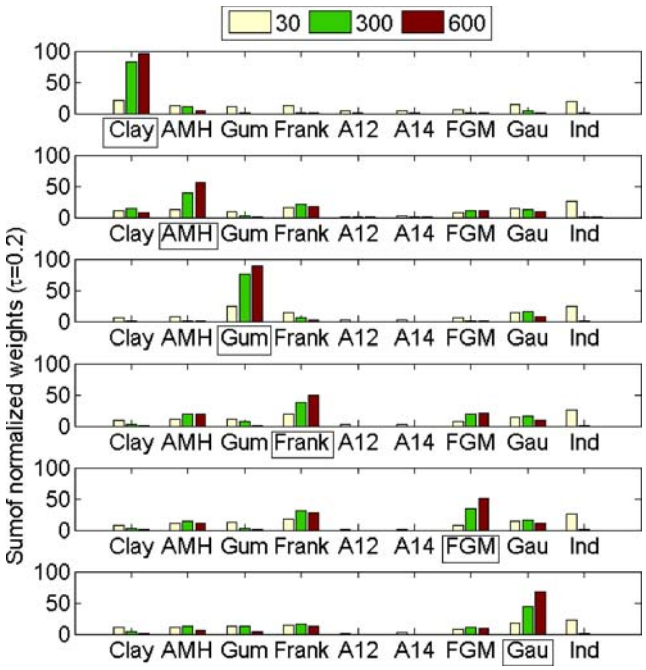

Fig. 6 Sum of normalized weights over 100 trials for $\tau = 0.2$ and $ns = 30$, 300, and 600

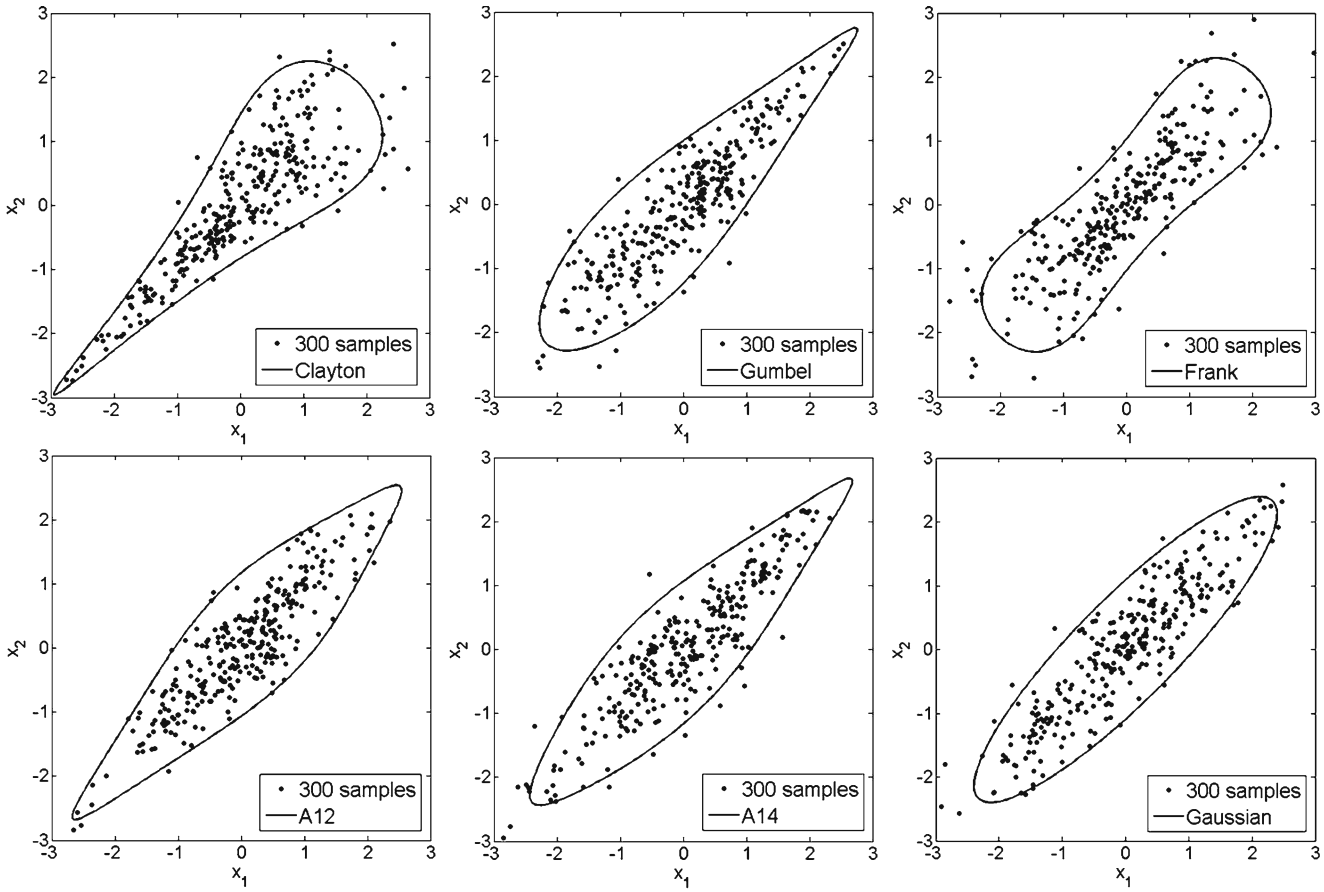

Fig. 7 PDF contours and scatter plots of samples ($ns = 300$) for Clayton, Gumbel, Frank, A12, A14, and Gaussian with $\tau = 0.7$

to identify a correct copula when the Kendall's tau is small.

When the correlation between two variables is more significant, e.g., $\tau = 0.7$, it is easier to identify the correct copula because the copula shapes are quite distinct from each other, as shown in Fig. 7. Accordingly, the correct copula can be easily identified with the highest normalized weight seen in Fig. 8.

In RBDO problems, if the shape of the selected copula is significantly different from the true copula—especially for high correlation—a wrong RBDO result will be obtained. For instance, if the Clayton copula is the original copula, and other copulas such the Frank or Gaussian are wrongly selected, the RBDO result will be very different from the true RBDO result (Noh et al. 2007). However, if the correlation between input variables is not high, copulas such as the AMH, Frank, FGM, and Gaussian, which have similar copula shapes, might provide quite similar RBDO results. Thus, 300 samples could be acceptable to identify the right copula even for small correlation coefficients. For large correlation coefficients, 100, or even 30, samples could be enough to identify correctly for copulas such as Clayton, Gumbel, and Frank.

6 Examples

To show how the Bayesian method identifies marginal CDFs and joint CDF, two problems are tested.

6.1 Mathematical example

Suppose that $X_1 \sim LN(1.62, 0.08)$ and $X_2 \sim N(5, 0.5)$ with a joint CDF modeled by the Frank copula. The Kendall's tau is given as $\tau = 0.5$. From the given marginal and joint CDFs, which are unknown, 100 samples are generated to test the Bayesian method. A strategy of generating the paired data from the given marginal and joint CDF proceeds as follows:

First, U is randomly generated from a uniform distribution with an interval [0, 1]. Second, given U, V

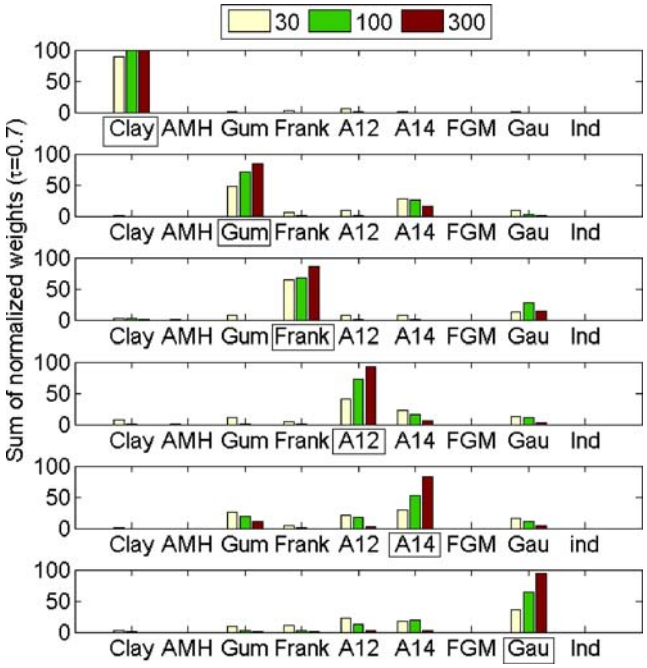

Fig. 8 Sum of normalized weights over 100 trials for $\tau = 0.7$ and $ns = 30$, 100, and 300

is calculated from a derivative of the copula function as

$$Q_U(V) = \frac{\partial C(U, V)}{\partial U} = W \tag{39}$$

by setting $V = Q_U^{-1}(W)$ where W is randomly generated from a uniform distribution with an interval [0, 1], which is independently generated from U. Employing inverse CDFs of X_1 and X_2, $X_1 = F_{X_1}^{-1}(U)$ and $X_2 = F_{X_2}^{-1}(V)$, paired samples are generated.

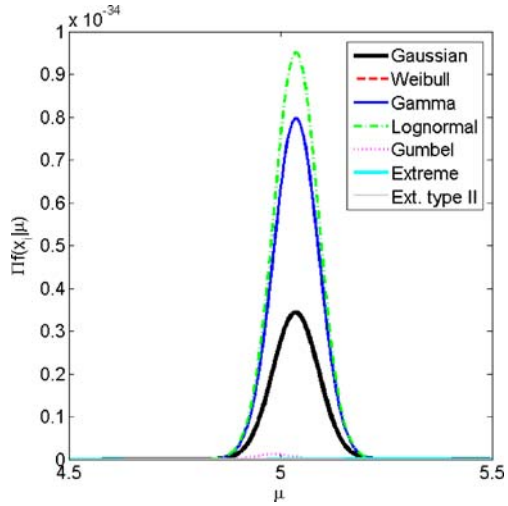

Fig. 9 Likelihood functions of μ for X_1

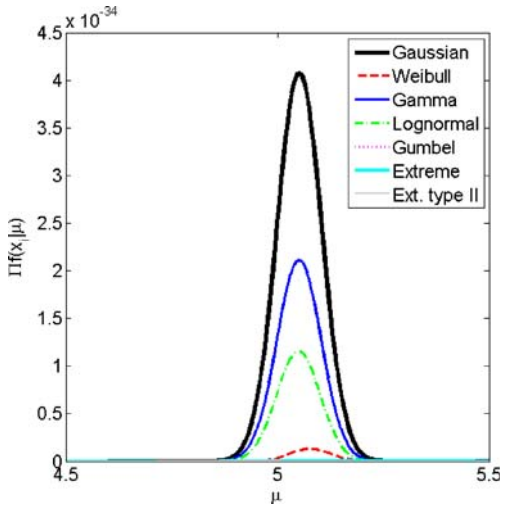

Fig. 10 Likelihood functions of μ for X_2

Using the generated 100 paired samples, the likelihood functions for candidate marginal distributions for the first variable X_1 are obtained. As shown in Fig. 9, the lognormal CDF has the highest peak and is widely spread, which means the normalized weight of the lognormal CDF is the highest among candidate marginal CDFs. Thus, the Bayesian method identifies the lognormal CDF, which is the original CDF of X_1, as the correct marginal CDF. For the second variable X_2, as shown in Fig. 10, the likelihood function of μ for the Gaussian CDF has the highest peak, which means the Gaussian CDF has the highest normalized weights, 0.550 in Table 8. Thus, the Bayesian method correctly identifies the original marginal CDF (Gaussian) of X_2.

Using the identified marginal CDFs, a copula can be identified using the Bayesian method. As shown in Fig. 11, the Frank copula has the highest weight among candidate copulas. Since the Frank copula has a distinct PDF shape among candidate copulas, the normalized weight is very high, 0.996 (Table 9), and correctly identified. According to the PDF contour of the Frank copula shown in Fig. 12, the Frank copula well describes the given data.

Table 8 Normalized weights of candidate marginal CDFs for X_1 and X_2

Distribution	X_1	X_2
Gaussian	0.166	0.550
Weibull	0.000	0.008
Gamma	0.378	0.285
Lognormal	0.449	0.158
Gumbel	0.007	0.000
Extreme	0.000	0.000
Extreme type-II	0.000	0.000

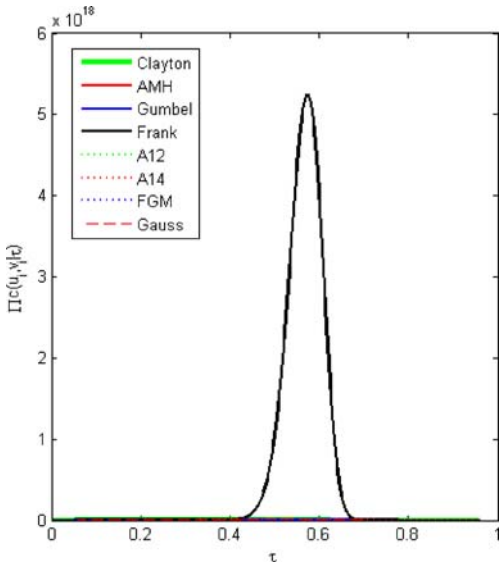

Fig. 11 Likelihood functions of τ

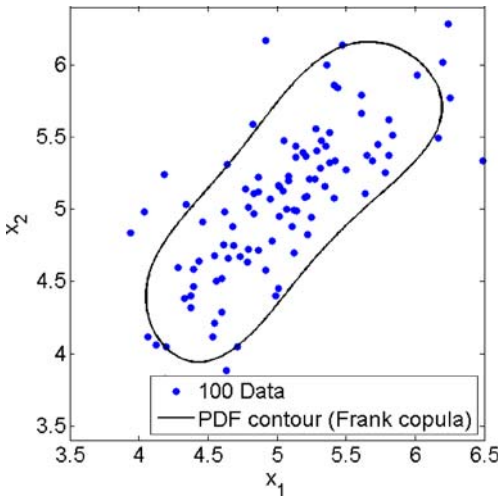

Fig. 12 100 Data and PDF contour

In practical applications, since the true copula is unknown, one of copulas that best describe the given experimental data will be identified.

6.2 Strain fatigue life example

In fatigue problems, strain–life relationship is expressed as

$$\frac{\Delta\varepsilon_f'}{2} = \frac{\sigma_f'}{E}(2N_f)^b + \varepsilon_f'(2N_f)^c \tag{40}$$

where $\Delta\varepsilon_f'$ is the strain amplitude, E is the Young's modulus, N_f is the fatigue life, σ_f' and b are the fatigue strength coefficient and exponent, and ε_f' and c are fatigue ductility coefficient and exponent, respectively. Figure 13 shows 29 data pairs of the fatigue strength coefficient and exponent, and the fatigue ductility coefficient and exponent in 950X steel (Socie 2003). As shown in Fig. 13a and b, σ_f' and b, and ε_f' and c are highly negatively correlated.

Based on the given data, it is necessary to identify marginal CDFs and a joint CDF. Since it is known that σ_f' and ε_f' follow lognormal CDF and b and c follow Gaussian CDF, identification of only the joint CDF is necessary in this example. Since two pairs of variables are highly negatively correlated, only the Frank and Gaussian are selected as candidate copulas because other copulas such as Clayton and Gumbel cannot have negative correlation and some copulas such as AMH and FGM can only have large negative correlation. Using the Bayesian method, the Gaussian copula for σ_f' and b, and the Frank copula for ε_f' and c are identified based on the given marginal CDFs and data as seen in Table 10 where the correlation coefficient between σ_f' and b is calculated as -0.828 (Pearson's rho) and -0.906 (Kendall's tau), respectively. As shown in Fig. 13, the PDF contours of two identified copulas well describe the given experimental data.

Since the number of given experimental data is small, it is possible that the identified copulas may not be correct. Further, even though the mean and standard deviations of two variables are given in this example, those parameters are usually obtained from the given data. Accordingly, even if the copulas are correctly identified, the estimated parameters could be significantly erroneous. Therefore, the confidence level needs to be implemented in the input model for RBDO. Future research will address this issue.

For probabilistic life prediction, when variables are so closely related, another option is to let one variable be a function of the other variables, such as a linear fit. The problem with this approach is that the data in Fig. 13 cannot be properly fitted by linear functions. In addition, the importance of correct modeling of a joint CDF of correlated variables is stated by Annis (2004) using the Paris equation relating crack growth rate. Annis (2004) pointed out that the result, standard deviation of the number of cycles, will be over corrected and will thus underestimate the overall variability, and

Table 9 Normalized weights of candidate copulas

Clayton	AMH	Gumbel	Frank	A12	A14	FGM	Gauss
0.000	0.000	0.000	0.996	0.000	0.001	0.000	0.003

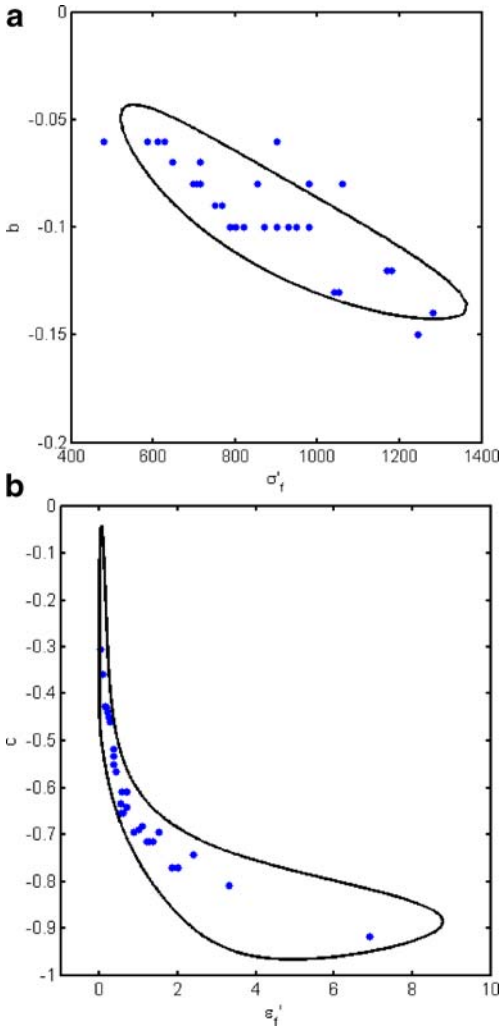

Fig. 13 Paired data obtained from 950X steel (Courtesy of Professor Darrell Socie) and PDF Contours of Gaussian and Frank Copula. **a** Fatigue Strength Coefficient $\left(\sigma'_f\right)$ and Exponent (b). **b** Fatigue Ductility Coefficient $\left(\varepsilon'_f\right)$ and Exponent (c)

the right way is "... to correctly modeling the joint behavior to reduce greater than 700% error in the estimated of the standard deviation to about 1%" in his example. Moreover, in RBDO, it is utmost important to obtain reliable optimum designs. Thus, the linear or nonlinear fit may not be a desirable approach since it will lead to unreliable optimum designs.

Table 10 Normalized weights of candidate copulas

	Frank	Gauss
σ'_f, b	0.355	0.645
ε'_f, c	0.607	0.393

7 Conclusions and future works

To carry out the RBDO, information about the input variables, such as marginal and joint CDF types, needs to be correctly identified, which is especially challenging when input variables are correlated and only limited data are available. In this paper, a copula is utilized to model the joint CDF of the correlated input variables using limited information such as the marginal CDF types (if they are known) and the given test data. Since the correct identification of the copula is necessary to model the correct joint CDF, the Bayesian method is used to identify a copula that best describes the given experimental data. The identification of marginal CDFs is as important as the identification of the copula. Thus, the Bayesian method is proposed for identification of a correct marginal CDF. Since the Bayesian method assigns normalized weights mostly to correct marginal CDFs or copulas and to similar marginal CDFs or copulas among candidates, it is more effective than the GOF test.

However, the identified copulas could be wrong if a very small number of samples is used. Even though marginal CDFs and copulas are correctly identified, estimated parameters such as mean or standard deviation from the limited data could yield inaccurate contours of the target reliability index, which would lead to wrong RBDO results. Thus, for future research, RBDO with a confidence level, which alleviates the effect of the wrong identification and quantification of marginal CDFs and copulas, is being investigated.

Acknowledgements Research is jointly supported by the U.S. Army TARDEC Project # W911NF-07-D-0001 and the Automotive Research Center, which is sponsored by the U.S. Army TARDEC.

References

Annis C (2004) Probabilistic life prediction isn't as easy as it looks. JAI 1(2):3–14

Bravais A (1846) Analyse mathematique sur les probabilites des erreurs de situation d'un point. Memoires par divers Savan 9:255–332

Bretthorst GL (1996) An introduction to model selection using probability theory as logic. In: Heidbreder G (ed) Maximum entropy and Bayesian methods. Kluwer, Dordrecht, pp 1–42

Cirrone GAP, Donadio S, Guatelli S et al (2004) A godness-of-fit statistical toolkit. IEEE Trans Nucl Sci 51(5):2056–2063

Ditlevsen O, Madsen HO (1996) Structural reliability methods. Wiley, New York

Efstratios N, Ghiocel DM, Singhal S (2004) Engineering design reliability handbook. CRC, New York

Genest C, Favre AC (2007) Everything you always wanted to know about copula modeling but were afraid to ask. J Hydrol Eng 12(4):347–368

Genest C, Rémillard B (2005) Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. Technical Rep. G-2005-51, Group d'Études et de Recherche en Analyse des Décision

Genest C, Rivest LP (1993) Statistical inference procedures for bivariate Archimedean copulas. J Am Stat Assoc 88:1034–1043

Haldar A, Mahadevan S (2000) Probability, reliability, and statistical methods in engineering design. Wiley, New York

Huard D, Évin G, Favre AC (2006) Bayesian copula selection. Comput Stat Data Anal 51:809–822

Kendall M (1938) A new measure of rank correlation. Biometrika 30:81–89

Kruskal WH (1958) Ordinal measures of associations. J Am Stat Assoc 53(284):814–861

Melchers RE (1999) Structural reliability analysis and prediction, 2nd edn. Wiley, New York

Nataf A (1962) Détermination des CDFs de probabilités dont les marges sont données. C R Hebd Séances Acad Sci 255: 42–43

Nelsen RB (1999) An introduction to copulas. Springer, New York

Noh Y, Choi KK, Du L (2007) New transformation of dependent input variables using copula for RBDO. In: 7th world congress of structural and multidisciplinary optimization, Seoul, Korea, 21–25 May

Noh Y, Choi KK, Du L (2008) Reliability based design optimization of problems with correlated input variables using copulas. Struct Multidisc Optim 38(1):1–16

Pearson K (1896) Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. Philos Trans Royal Soc London Ser A 187:253–318

Socie DF (2003) Seminar notes: probabilistic aspects of fatigue. http://www.fatiguecaculator.com. Cited 8 May 2008

Sterne JA, Smith GD (2000) Sifting the evidence—what's wrong with significance tests? Br Med J 322:226–231