Multivariate Distribution Models

Model Description

While the probability distribution for an individual random variable is called "marginal," the probability distribution for multiple random variables is called a "multivariate" or "joint" distribution. The joint PDF of two continuous random variables X and Y is defined as

$$f(x,y)\cdot dx\cdot dy = \mathrm{P}\left(x < X \le x+dx \bigcap y < Y \le y+dy\right) \tag{1}$$

The joint PDF has the following properties:

$$\begin{aligned} & f(x,y) \ge 0 \\ & f(y) = \int_{-\infty}^{\infty} f(x,y)\,dx \\ & f(x) = \int_{-\infty}^{\infty} f(x,y)\,dy \\ & \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1 \end{aligned} \tag{2}$$

The relationship between the joint PDF and the joint CDF is

$$F(x,y) = P(X \le x \bigcap Y \le y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x,y)\,dx\,dy \tag{3}$$

which implies that

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y} \tag{4}$$

Having the joint distribution, conditional distributions are defined in accordance with the conditional probability rule for events:

$$f(x\,|\,y)\cdot dx = \frac{f(x,y)\cdot dx\cdot dy}{f(y)\cdot dy} \Rightarrow f(x\,|\,y) = \frac{f(x,y)}{f(y)} \tag{5}$$

As a result, the joint distribution can be expressed in terms of conditional distributions as

$$f(x,y) = f(x\,|\,y)\cdot f(y) = f(y\,|\,x)\cdot f(x) \tag{6}$$

Two random variables, X and Y, are said to be statistically independent if

$$f(x\,|\,y) = f(x) \quad \text{or} \quad f(y\,|\,x) = f(y) \tag{7}$$

Statistical independence implies that the joint distribution for two statistically independent random variables, X and Y, is the product of the marginals:

$$f(x,y) = f(y)\cdot f(x) \tag{8}$$

Model Parameters

In the context of joint distributions, partial descriptors include the mean product:

$$\mathrm{E}[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \cdot y \cdot f(x,y)\,dx\,dy \tag{9}$$

and the covariance:

$$\mathrm{Cov}[X,Y] = \mathrm{E}[(X-\mu_X)(Y-\mu_Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)\cdot(y-\mu_Y)\cdot f(x,y)\,dx\,dy \tag{10}$$

Expansion of the integrand in Eq. (10) reveals that the covariance is equal to the mean product minus the product of the means:

$$\mathrm{Cov}[X,Y] = \mathrm{E}[XY] - \mu_X\mu_Y \tag{11}$$

This echoes the fact that the variance of a marginal distribution equals the mean square minus the square of the means. As discussed later in this document, the covariance between two random variables is a measure of linear dependence between them. A normalized, i.e., dimensionless measure of this linear dependence is the correlation coefficient, which is defined as:

$$\rho_{XY} = \frac{\mathrm{Cov}[X,Y]}{\sigma_X\sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y} = \mathrm{E}\left[\frac{X-\mu_X}{\sigma_X}\cdot\frac{Y-\mu_Y}{\sigma_Y}\right] \tag{12}$$

Matrix Notation

When dealing with second-moment information, i.e., mean, variance, and correlation of multiple random variables it is convenient to use matrix notation. As an illustration, let **X** be a vector of random variables, while **x** is the vector of realizations of **X**. Then, the vector of means is

$$\mathbf{M_X} = \begin{Bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{Bmatrix} \tag{13}$$

where m_i is the mean of random variable number i. The covariance matrix is a symmetric matrix that contains the variances of the random variables, and the covariance between them:

$$\mathbf{\Sigma_{XX}} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\cdot\sigma_1\cdot\sigma_2 & \rho_{13}\cdot\sigma_1\cdot\sigma_3 \\ \rho_{12}\cdot\sigma_1\cdot\sigma_2 & \sigma_2^2 & \rho_{23}\cdot\sigma_2\cdot\sigma_3 \\ \rho_{13}\cdot\sigma_1\cdot\sigma_3 & \rho_{23}\cdot\sigma_2\cdot\sigma_3 & \sigma_3^2 \end{bmatrix} \tag{14}$$

By defining the matrix $\mathbf{D_{XX}}$ to be a square matrix with the standard deviations on the diagonal the covariance matrix is written as the decomposition

$$\mathbf{\Sigma_{XX}} = \mathbf{D_{XX}}\mathbf{R_{XX}}\mathbf{D_{XX}} \tag{15}$$

where $\mathbf{R_{XX}}$ is the correlation matrix, which is also symmetric:

$$\mathbf{R_{XX}} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix} \tag{16}$$

Common Distribution Types

Unlike the situation for univariate distributions, only a few standard multivariate distribution types are encountered. By far the most common is the joint Normal distribution.

Multivariate Normal Distribution

The joint normal PDF is

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n}\cdot\sqrt{\det(\mathbf{\Sigma}_{XX})}}\cdot\exp\left(\frac{1}{2}\left(\mathbf{x}-\mathbf{M}_X\right)^T\mathbf{\Sigma}_{XX}^{-1}\left(\mathbf{x}-\mathbf{M}_X\right)\right) \tag{17}$$

where n is the number of random variables. In the bi-variate case it reads

$$f(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\cdot\exp\left(-\frac{z}{2(1-\rho)^2}\right) \tag{18}$$

where

$$z = \left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} \tag{19}$$

A special case is the standard normal distribution, which is characterized by zero means, unit variances, and zero covariances. This PDF is denoted by the symbol φ and takes the form

$$\varphi(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^n}}\cdot\exp\left(-\frac{1}{2}\mathbf{y}^T\mathbf{y}\right) \tag{20}$$

This multivariate distribution has several properties that are important in reliability analysis:

1. The multivariate standard normal PDF is rotationally symmetric and it decays exponentially in the radial and tangential directions
2. The probability content outside a hyper-plane distanced β from the point $\mathbf{y}=\mathbf{0}$ is:

$$p = \Phi(-\beta) \tag{21}$$

which is employed in the document on FORM.

3. The probability content outside a hyper- paraboloid with apex distanced β from the point **y=0** is also available, as described in the document on SORM.

Multivariate t-Distribution

(Yet to be written.)

Classical Inference

Diagrams

In the same way as relative frequency diagrams are instructive visualizations of the realizations of a single random variable, a scatter diagram is valuable when two random variables are observed simultaneously. The scatter diagram visualizes the outcomes of one variable along one axis versus the outcomes of the other variable along the other axis. The plot gives a sense of the dependence between the two variables.

Classical Inference

The formulas for sample mean and sample variance of individual random variables are valid for the inference on joint random variables. In addition, the sample correlation coefficient is:

$$\rho = \frac{1}{n-1} \cdot \left(\frac{\sum_{i=1}^{n} x_i y_i - n \cdot \overline{x} \cdot \overline{y}}{s_x \cdot s_y} \right) \tag{22}$$

Statistical Dependence

Correlation

Statistical dependence between random variables may take different forms. For example, one form of dependence is that one variable varies exponentially with the other. Yet another example is linear dependence, in which the realizations of one random variable tend to be proportional to the outcomes of another random variable. Correlation, defined in Eq. (12), measures linear dependence. In other words, two random variables can be uncorrelated but statistically dependent. It is also emphasized that when statistical dependence is specified by means of correlation then the possibility of a non-positive definite correlation matrix is present. In reliability analysis, this prevents the transformation into standard normal random variables. As a result, some correlation structures are impractical and/or unphysical. Importantly, the range of possible correlation depends upon the marginal probability distributions of the random variables. Hence, in reliability analysis applications, the specification of correlation must be made with care and with knowledge about the marginal probability distributions. This is different with copulas.

Copulas

Copulas represent an alternative technique for specifying statistical dependence between random variables. Currently, its use is more widespread in economics than engineering, but that may change. Copulas extend the options for prescribing statistical dependence beyond the use of the correlation coefficient, which only provides linear statistical dependence. The correlation coefficient is convenient and popular for a few reasons. First, it appears prominently in second-moment theory, together with means and standard deviations. Second, the correlation coefficient appears as a parameter in the powerful joint normal probability distribution, as described earlier in this document. However, the convenience of the correlation coefficient diminishes in the general of circumstances. Consider the example when the joint distribution is sought for a set of random variables with mixed marginal distributions and perhaps nonlinear dependence tendencies. This problem is important in reliability analysis where the Nataf or Rosenblatt transformations are usually applied. Under such circumstances the copulas represent an alternative, although it has yet to become popular in reliability analysis. The key feature of the copula technique is that a variety of dependence structures are possible. One example is stronger dependence in the distribution tails. An interesting class of copulas is the generalized elliptical distributions that are generalizations of the joint normal distribution. The joint normal distribution is also elliptical, but it is a special case of the "infinite" possibilities provided by copulas.

From a philosophical viewpoint, the need to specify statistical dependence between random variables is, in some sense, a symptom of imperfect models. The source of correlation is due to hidden phenomena behind the random variables. If the underlying phenomena were modelled then the need to prescribe statistical dependence might vanish. Consider the example of prescribing correlation between the earthquake intensity at two nearby sites. The need to estimate this correlation disappears if the modelling is expanded to include the hypocentre location, the earthquake magnitude, and the attenuation of the intensity to each site. It is those underlying phenomena that cause correlation in intensity between sites. This philosophical discussion is somewhat akin to the discussion on whether aleatory uncertainty exists. It does, unless all models are perfect, which they are not. However, this paragraph is intended to foster a strong focus on modelling and careful examination of the need to prescribe statistical dependence.

Sklar's Theorem

Sklar's theorem is the foundation for the use of copulas. It states that the joint CDF of some random variables, $\mathbf{X}$, can be written in terms of a copula, C, which is a function of the marginal CDFs of the random variables:

$$F(x_1,x_2,\ldots,x_n) = C\left(F_1(x_1),F_2(x_2),\ldots,F_n(x_n)\right) \tag{23}$$

That is, the joint distribution is composed of the marginal distributions and the copula function. In other words, the copula is a function that couples the marginal distribution functions. This is the means by which dependence is introduced. It is also observed that copulas express dependence on a "quantile scale," namely along the random variable axes. In this manner, the dependence at 10% probability of exceedance can be different from the dependence at 90% probability of exceedance. Several other interpretations of Eq. (23) are possible. First, it is observed that a copula is what remains of a joint

cumulative distribution once the action of the marginal cumulative distribution functions has been removed. In other words, the marginals provide the probability distributions, while the sole purpose of the copula is to provide statistical dependence. Furthermore, Sklar's theorem can be written

$$C\left(p_1,p_2,\ldots,p_n\right)=F\left(F_1^{-1}(p_1),F_2^{-1}(p_2),\ldots,F_n^{-1}(p_n)\right) \tag{24}$$

where p_i are probabilities. This form of Eq. (23) is used to "extract" copulas from existing joint distributions, as described shortly. It is noted that a copula is invariant with respect to strictly increasing transformations of the random variables, such as that of transforming random variables from normal to standard normal.

Explicit and Implicit Copulas

The simplest example of a copula is the one that yields no dependence at all. That is, the copula for independent random variables is:

$$F(x_1,x_2,\ldots,x_n)=F_1(x_1)\cdot F_2(x_2)\cdots F_n(x_n) \tag{25}$$

This expression, which corresponds to the definition of statistical independence between random variables, is an example of an explicit copula. Copulas are either implicit or explicit. Implicit copulas are extracted from known joint distributions. For example, the Gauss copula is extracted from the joint normal probability distribution. Specifically, from Sklar's theorem in Eq. (23) it is understood that when the random variables have the joint CDF F then the copula C is the CDF of the marginal distributions. This is what is emphasized in Eq. (24). Consider two correlated normal random variables, here standard normal for simplicity:

$$F(x_1,x_2)=\Phi(x_1,x_2)=\int_{-\infty}^{x_2}\int_{-\infty}^{x_1}\frac{1}{2\pi\sqrt{1-\rho^2}}\cdot\exp\left(-\frac{s_1^2+s_2^2-2\rho s_1 s_2}{2(1-\rho^2)}\right)ds_1\,ds_2 \tag{26}$$

The Gauss copula is extracted from this joint CDF by substituting the random variables in the original distribution with the marginal CDFs:

$$C(p_1,p_2)=\Phi(x_1,x_2)=\int_{-\infty}^{\Phi^{-1}(p_2)}\int_{-\infty}^{\Phi^{-1}(p_1)}\frac{1}{2\pi\sqrt{1-\rho^2}}\cdot\exp\left(-\frac{s_1^2+s_2^2-2\rho s_1 s_2}{2(1-\rho^2)}\right)ds_1\,ds_2 \tag{27}$$

Thus, the normal copula is extracted.

Common Copulas

The following is a set of common copulas, where the notation, $p=F(x)$ is employed for convenience:

Independent: $$C(p_1,p_2)=p_1\cdot p_2 \tag{28}$$

Normal: $$C(p_1,p_2)=\int_{-\infty}^{\Phi^{-1}(p_2)}\int_{-\infty}^{\Phi^{-1}(p_1)}\frac{1}{2\pi\sqrt{1-\rho^2}}\cdot\exp\left(-\frac{s_1^2+s_2^2-2\rho s_1 s_2}{2(1-\rho^2)}\right)ds_1\,ds_2 \tag{29}$$

$$\text{Student: } C(p_1,p_2)=\int_{-\infty}^{T_\nu^{-1}(p_2)}\int_{-\infty}^{T_\nu^{-1}(p_1)}\frac{1}{2\pi\sqrt{1-\rho^2}}\cdot\exp\left(-\frac{s_1^2+s_2^2-2\rho s_1 s_2}{2(1-\rho^2)}\right)^{-\left(\frac{\nu+2}{2}\right)}ds_1\,ds_2 \tag{30}$$

$$\text{Frank: } C(p_1,p_2)=-\frac{1}{\theta}\cdot\ln\left(1+\frac{\left(e^{-\theta\cdot p_1}-1\right)\cdot\left(e^{-\theta\cdot p_2}-1\right)}{e^{-\theta}-1}\right) \tag{31}$$

$$\text{Clayton: } C(p_1,p_2)=\left(p_1^{-\theta}+p_2^{-\theta}-1\right)^{-\frac{1}{\theta}} \tag{32}$$

$$\text{Gumbel: } C(p_1,p_2)=\exp\left(-\left(\left(-\ln(u_1)\right)^\theta+\left(-\ln(u_2)\right)^\theta\right)^{\frac{1}{\theta}}\right) \tag{33}$$

Sampling of Realizations

(Yet to be written.)

Meta Distributions

A potentially interesting aspect of the use of copulas is the possibility of creating entirely new "meta" distributions. This is achieved by first employing Eq. (24) to extract an implicit copula, followed by utilization of Eq. (23) to substitute "arbitrary" CDFs into the copula. Clearly, a great number of possible joint probability distributions—perhaps more or less useful—then become available. To generate realization of random variables from meta distributions the following sampling procedure may be helpful:

1. Generate outcomes of some random variables **x** from the fundamental distribution, say the normal
2. Obtain the marginal CDF value for each random variable, i.e., $p=F(x)$
3. Transform according to some marginal distribution: $x=F^{-1}(p)$

Copula Densities

$$f(x_1,x_2,\ldots,x_n)=c\left(F_1(x_1),F_2(x_2),\ldots,F_n(x_n)\right)\cdot\prod_{i=1}^{n}f_i(x_i) \tag{34}$$

where c is the derivative of C:

$$c(p_1,p_2,\ldots,p_n)=\frac{\partial^n C(p_1,p_2,\ldots,p_n)}{\partial p_1\partial p_2\cdots\partial p_n} \tag{35}$$

Measures of Tail Dependence

Because copulas facilitate the specification of different statistical dependence at different quantiles, it is of interest to introduce some generic measure of dependence, particularly in the tail of the distributions. This is particularly useful because, in certain applications, it is the extreme outcomes of dependent random variables that are of interest. The coefficient of upper tail dependence is defined as the probability that the random variable X_i exceeds the value associated with its inverse CDF of q, i.e., "the quantile of order q", given that the other random variable X_j exceeds the value associated with its inverse CDF of q, when q tends towards unity. For continuous random variables, the coefficient is:

$$\lambda_{upper} = \lim_{q \to 1} P\left(X_i > F_i^{-1}(q) \;\middle|\; X_j > F_j^{-1}(q) \right) = \lim_{q \to 1} \frac{1 - 2q + C(q,q)}{1 - q} \tag{36}$$

Similarly,

$$\lambda_{lower} = \lim_{q \to 1} P\left(X_i < F_i^{-1}(q) \;\middle|\; X_j < F_j^{-1}(q) \right) = \lim_{q \to 1} \frac{C(q,q)}{q} \tag{37}$$

It is noted that for the normal copula, $\lambda_{upper}=\lambda_{lower}=0$. Hence, it is not possible to take into account tail dependence with this copula, contrary to, say, the Student copula. When using copulas, there are also other measures of dependence other than the measures of tail dependence in Eqs. (36) and (37). These include the "rank-dependent correlation coefficient," such as Kendall's tau and Spearman's rho. For Archimedean copulas, there is a strong connection between Kendall's tau and the parameter of the copula function. (Yet to be written.)