# 4

# *Modeling Dependence beyond Correlation*

**Christian Genest and Johanna G. Nešlehová**
*McGill University, Montréal, QC*

As scientists, looking for relationships between variables is an essential part of our efforts to understand the world, identify the causes of illnesses, assess climate change, predict economic cycles or guard against catastrophic events such as tsunamis or financial crises. Statistical models are often used for this purpose. To many scientists and engineers, this is synonymous with correlation and regression because these techniques are typically the first, if not the only ones, to which they were exposed. However, thanks in part to the work of Canadian statisticians, we now know that there is much more to dependence than correlation, regression, and the omnipresent "bell-shaped curve" of basic statistics textbooks. As we will demonstrate in this chapter, those who are oblivious to new tools may miss important messages hidden in their data.

## 4.1 Beyond the Normal Brave Old World

It is often said that a little knowledge is a dangerous thing, and so it is with those who believe that the Normal law, or "bell-shaped curve," is the appropriate model for just about every phenomenon measured on a continuous scale. In reality, many variables of interest just aren't Normally distributed, whatever scale you use. This is typically the case for financial losses, insurance claims, precipitation amounts and the height of storm surges, among others. The problem is even more pervasive when several variables are involved.

For example, McNeil (1997) analyzed losses above one million Danish kroner arising from fire claims made to the Copenhagen Re insurance company between 1980 and 1990. Each of them, adjusted for inflation, can be divided into a claim amount for damage to buildings ($X$), loss of contents ($Y$), and loss of profits ($Z$). For simplicity, let's focus on the 517 cases in which these three amounts were non-zero. Because some losses are very large, these data are best displayed on the logarithmic scale. This is done in Figure 4.1, where the best fitting Normal bell-shaped curve is superposed on the histogram of

losses of each type. A data analyst who is satisfied with the fit would conclude that the distributions of $X$, $Y$, $Z$ are well approximated by the "Log-Normal distribution," i.e., a bell-shaped curve on the logarithmic scale. The use of the logarithmic scale is not just ad hoc in this context; it has been found that the Log-Normal distribution often provides a good first approximation for many monetary insurance losses.

However, the amounts claimed for damage to buildings, loss of contents and loss of profits are naturally related because they are generated by the same fire. More importantly, a severe conflagration is likely to cause substantial damage on all accounts. When this happens, the insurer may incur large losses. The dependence between the claim amounts $X$, $Y$, $Z$ is thus of paramount interest to the company. A statistical model that captures the simultaneous variation of these amounts can be used, among others, to determine sufficient capital reserves in order to avoid insolvency.

To assess this dependence, people typically start by looking at the plots of all pairs of variables. To visualize the relationship between damage to buildings and loss of contents, for example, we can plot the pairs $(\log X_i, \log Y_i)$, where the subscript refers to fire event $\#i$, so that $i = 1, \ldots, 517$. The scatterplots

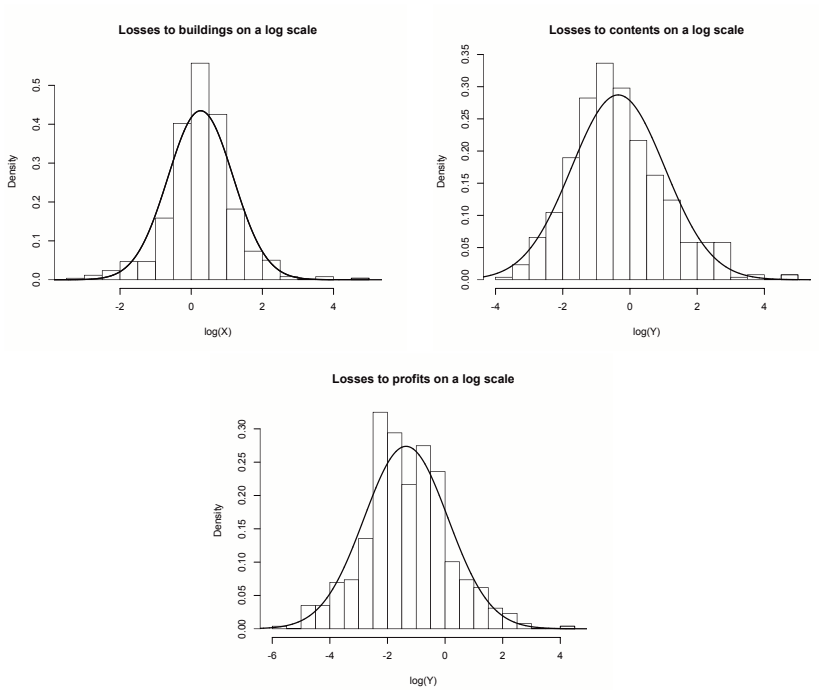


FIGURE 4.1: Histograms for $X$, $Y$, $Z$ after transformation to the logarithmic scale.

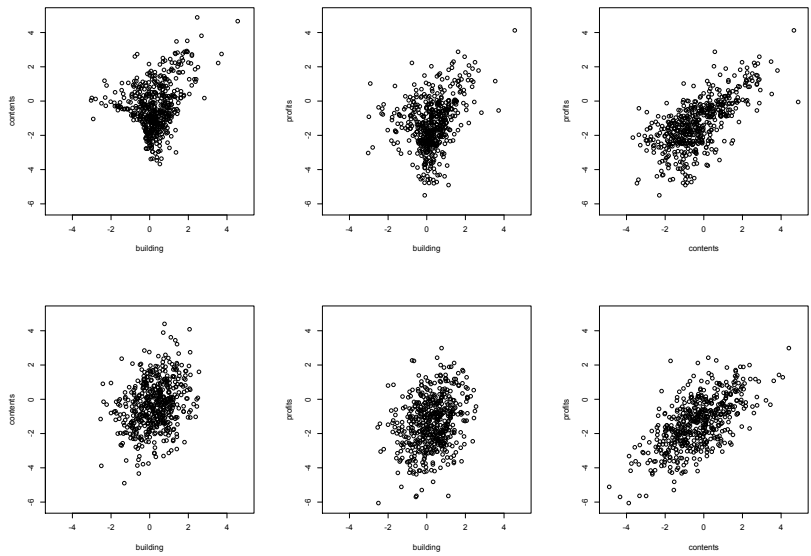


FIGURE 4.2: Top: pairwise scatterplots for $X$, $Y$, $Z$ on the logarithmic scale; bottom: random sample of the same size from a fitted trivariate Normal law.

corresponding to the three possible pairs are displayed in Figure 4.2. The point clouds clearly show that the losses are positively associated. This is plausible because small fires generate smaller losses than big ones do.

Traditionally, this association is summarized by Pearson's correlation coefficient, which measures how close a point cloud is to a straight line. This coefficient always lies between $-1$ and 1; it equals either of these two values if the scatterplot is a perfect line. Here, the observed correlations are

$$\begin{aligned}\operatorname{corr}(\log X, \log Y) &\approx .282,\\ \operatorname{corr}(\log X, \log Z) &\approx .307,\\ \operatorname{corr}(\log Y, \log Z) &\approx .667.\end{aligned}$$

Thus, on the logarithmic scale, losses of contents ($Y$) and profits ($Z$) are much more highly correlated with each other than they are with damage to buildings ($X$). Because the distributions of $\log(X)$, $\log(Y)$ and $\log(Z)$ appear to be Normal, the simplest joint statistical model for these three variables then consists of assuming that they are jointly Normal. Among others, this implies that the 3D histograms of all pairs of losses are bell-shaped on the logarithmic scale. What makes this model especially seductive is that the relationship among the three losses is then fully explained by the correlation coefficients computed above.

But is such a joint model plausible? To address this question, we simulated an artificial sample of 517 data points from the best fitting trivariate Normal distribution. The resulting pairwise scatterplots are displayed in the bottom row of Figure 4.2. The shapes of the point clouds are strikingly different from those of the original data that appear in the top row of the same figure. This suggests that the data are not jointly Normal, a fact that is confirmed by formal statistical tests (Cox and Small, 1978). Of course, more sophisticated transformations might improve the fit, at the cost of interpretability. However, we shall see that try as we may, no transformation of these data can make them jointly Normal.

## 4.2 Dangers of Ignoring Non-Normality

For many years now, statisticians have been pointing out that modelers and data analysts may miss something important, or worse, reach disastrous conclusions, when they assume that the Normal distribution applies universally. Inadequate protection against insolvency is a case in point. Other examples include the underestimation of risks associated with floods and catastrophic storms. The paper by Embrechts et al. (2002) and the subsequent book by McNeil et al. (2005) have been particularly effective at delivering this message. In spite of this admonition, however, the notions of linear correlation and Normality are so deeply entrenched that many people remain blind to their limitations.

First, Pearson's correlation is merely a number, and as such, it cannot possibly tell the whole story. In Figure 4.2, each of the three datasets displayed in the top row has the same correlation as the Normal dataset appearing just below it, yet their shapes are strikingly different. In particular, there are many more points in the upper right corner of the real data plots than in the plots generated from the trivariate Normal model. In other words, the chances of large losses occurring simultaneously is much higher in reality than the Normal model would predict. An insurance company that fixes its premiums based on the Normal paradigm would thus be underestimating the risk of large claims.

Second, linear correlation is not easy to interpret outside of a specific model. People often view it as a measure of association between two variables, but they do not always realize that it is affected by the choice of scale. In the Danish fire loss data, for instance, the correlations between the original variables are

$$\mathrm{corr}(X, Y) \approx .627, \quad \mathrm{corr}(X, Z) \approx .791, \quad \mathrm{corr}(Y, Z) \approx .617.$$

These figures are very different from those computed from the same data once they have been transformed to the logarithmic scale. While the losses of contents and profits had the largest correlation on the latter scale, it is the

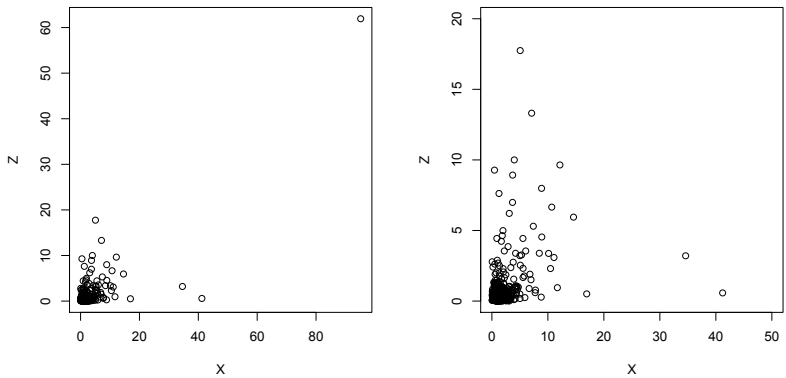


FIGURE 4.3: Scatterplot of $X$ and $Z$ on the original scale (left) and after removing the largest observation (right).

damage to buildings and losses of profits that are the most correlated on the original scale. Given that $\mathrm{corr}(X, Z)$ is large, we might be tempted to think that the relationship between $X$ and $Z$ is nearly linear. However, the left panel of Figure 4.3 reveals that this is not at all so. The high correlation between the two variables is actually driven by a few influential points in the upper right-hand corner of the graph. Yet these points should not be ignored. In fact, they are exactly the type of event that the insurer wants to anticipate and guard against! But even if these points are removed, as in the right panel of Figure 4.3, a linear pattern remains hard to discern.

In effect, linear correlation does not make all that much sense beyond the Normal paradigm. As additional evidence, consider the widespread misconception that any degree of correlation between $-1$ and 1 can be achieved. For instance when two losses are Normally distributed on the logarithmic scale, it can be shown that the theoretical value of the correlation coefficient actually lies between $\varrho_{\min}$ and $\varrho_{\max}$, where

$$\varrho_{\min} = \frac{e^{-\sigma_1\sigma_2} - 1}{\sqrt{(e^{\sigma_1^2} - 1)(e^{\sigma_2^2} - 1)}}, \quad \varrho_{\max} = \frac{e^{\sigma_1\sigma_2} - 1}{\sqrt{(e^{\sigma_1^2} - 1)(e^{\sigma_2^2} - 1)}}.$$

Here, $\sigma_1 > 0$ and $\sigma_2 > 0$ quantify the variability of the Normal distributions about their respective means. These so-called standard deviations are also related to the height of the bell-shaped curves. For the Danish fire loss data, with $\sigma_1$ and $\sigma_2$ estimated from the sample, the correlations should satisfy

$$
\begin{aligned}
-.12 \leq \text{corr}(X, Y) &\leq .63, \\
-.09 \leq \text{corr}(X, Z) &\leq .52, \\
-.02 \leq \text{corr}(Y, Z) &\leq .98.
\end{aligned}
$$

The fact that the observed correlations did not all fall within these bounds could simply be due to the "luck of the draw" or sampling variation but in view of the large sample size, it is more likely a sign that the losses are not Log-Normal after all. Indeed, the presence of very large claims, visible in the left panel of Figure 4.3, suggests that the individual loss distributions are heavy-tailed. This suspicion can be confirmed with a Q–Q plot or the test of Normality due to American statistician Samuel Shapiro and Canadian Martin Wilk (Shapiro and Wilk, 1965). If the loss distributions are really heavy-tailed, the sample linear correlation coefficient is trying to estimate a quantity that may not even be defined theoretically. The observed correlations would then be meaningless!

## 4.3 Copulas to the Rescue

When dealing with non-Normal data, it is crucial to think beyond correlation, linear regression, and joint Normality. An interesting concept that makes it possible to study dependence in broader terms was proposed by the American mathematician Abe Sklar in response to a question posed by his French colleague Maurice Fréchet.

### 4.3.1 Fréchet's Problem

Cast in the simplest possible terms, Fréchet's problem is the following. Suppose that $X$ and $Y$ are two fire claim amounts, say, for which we know how to compute the probabilities $\Pr(X \leq x)$ and $\Pr(Y \leq y)$ for any values $x$ and $y$. Viewed as functions of $x$ and $y$, the probabilities $F(x) = \Pr(X \leq x)$ and $G(y) = \Pr(Y \leq y)$ are called the marginal distribution functions, or margins, of $X$ and $Y$. The question is then how to construct a model for the probability of the events $X \leq x$ and $Y \leq y$ occurring simultaneously, denoted $\Pr(X \leq x \text{ and } Y \leq y)$, while ensuring that $X$ has distribution $F$ and $Y$ has distribution $G$.

Sklar's solution can in fact be traced back to work done in the 1940s by Wassily Hoeffding, a statistician of Danish origin. The idea is to write

$$\Pr(X \leq x \text{ and } Y \leq y) = C\{\Pr(X \leq x), \Pr(Y \leq y)\}, \tag{4.1}$$

where $C$ is a specific function of two variables called a copula. The purpose

of a copula is to "glue together the margins" or "couple the individual probabilities" (hence the Latin term "copula") in order to generate dependence between the variables. Any copula can be expressed as

$$C(u, v) = \Pr(U \leq u \text{ and } V \leq v) \tag{4.2}$$

in terms of variables $U$ and $V$ that are uniformly distributed on the interval $[0, 1]$, meaning that they are equally likely to lie anywhere between 0 and 1.

While regression imposes a linear relationship between the variables (after transformations), Sklar's formula provides complete flexibility in capturing the relation between $X$ and $Y$ because any copula $C$ can be combined with any margins $F$ and $G$ to construct a valid joint model for the pair $(X, Y)$. Conversely, any joint model for a pair $(X, Y)$ can be expressed in terms of its margins and some copula $C$. This is "the joy of copulas" (Genest and MacKay, 1986a,b).

### 4.3.2 Measuring Association

Sklar (1959) showed that the copula $C$ of a pair $(X, Y)$ is unique when the variables are measured on a continuous scale. It then contains *all* the relevant information about the dependence between $X$ and $Y$. Thus it makes sense to measure their association using $C$ or, equivalently, in terms of the uniform variables $U$ and $V$ appearing in formula (4.2). These variables, which satisfy the relations $U = F(X)$ and $V = G(Y)$, govern the dependence between $X$ and $Y$.

Surprisingly, the fact that association should be measured in terms of $U$ and $V$, rather than $X$ and $Y$, was recognized as early as 1904 by the British psychologist Charles Spearman. His measure of association, now termed Spearman's rank correlation coefficient (Spearman, 1904), turns out to be the sample analog of Pearson's linear correlation between $U$ and $V$. It can be computed from $C$ alone from a formula involving a double integral, viz.

$$\varrho(X, Y) = \mathrm{corr}(U, V) = -3 + 12 \int_0^1 \int_0^1 C(u, v) \mathrm{d}v \mathrm{d}u.$$

Similarly, the difference between the proportions of "concordant" and "discordant" pairs of observations popularized by the British statistician Maurice Kendall is a sample analog of another copula-based quantity called Kendall's tau (Kendall, 1938). [Two observations $(X_1, Y_1)$ and $(X_2, Y_2)$ are called concordant if the ranking of $X_1$ and $X_2$ is the same as that for $Y_1$ and $Y_2$; otherwise, the observations are said to be discordant.] As mentioned earlier, working on the uniform scale was also favored by Hoeffding in his seminal dissertation (Hoeffding, 1940).

The fundamental role of copulas in measuring association was crystallized in the work of Schweizer and Wolff (1981), who highlighted the fact that copulas are invariant with respect to increasing transformations of the variables,

and hence scale-free. In the context of the Danish fire loss data, for instance, this means that the pairs $(X, Y)$ and $(\log(X), \log(Y))$ share the same copula. As a consequence of this fact, Spearman's rho, Kendall's tau and other measures of dependence based on copulas yield the same values whether they are computed from the original data or after transforming them while preserving their order. In all cases, the sample values of Spearman's rho and Kendall's tau are

$$\begin{aligned} \varrho(X,Y) \approx .19, \quad \varrho(X,Z) \approx .29, \quad \varrho(Y,Z) \approx .64, \\ \tau(X,Y) \approx .12, \quad \tau(X,Z) \approx .2, \quad \tau(Y,Z) \approx .46. \end{aligned}$$

By focusing on the copula, what we have achieved is a complete separation of the dependence pattern of the variables from their individual behavior. This has several advantages. For example, both Spearman's rho and Kendall's tau can now reach all possible values between $-1$ and 1; the bounds are attained if one variable is a monotone (but not necessarily linear) function of the other.

Moreover, there are many other aspects of dependence that can be studied by looking at the copula. For example, UBC professor Harry Joe (Joe, 1997) used copulas to construct a coefficient $\lambda(X, Y)$ which quantifies the tendency of variables $X$ and $Y$ to take large values simultaneously. His upper-tail coefficient, which varies between 0 (no tail dependence) and 1 (comonotonic dependence), is especially relevant for measuring the riskiness of a financial position. In the fire loss data, for example, tail dependence turns out be quite strong, viz.

$$\lambda(X,Y) \approx .56, \quad \lambda(X,Z) \approx .5, \quad \lambda(Y,Z) \approx .68.$$

This nicely echoes the simultaneous presence of large values of $X$, $Y$, $Z$ which is clearly visible in the plots in the top row of Figure 4.2, and in Figure 4.3.

Given the complexity of relationships between variables, dependence can be quantified in more sophisticated ways than merely through one-number summaries. Originating in the work of Erich Lehmann, a professor of statistics at the University of California at Berkeley, many dependence concepts and orderings have been proposed (Lehmann, 1966; Müller and Stoyan, 2002) and much of this methodology is still under development. Many Canadian statisticians have contributed to this line of research, including Philippe Capéraà and Louis-Paul Rivest at Université Laval and Mhamed Mesfioui in Trois-Rivières. Many actuarial implications of these results have also been considered, notably by Hélène Cossette and Étienne Marceau, also at Laval (Denuit et al., 2005).

### 4.3.3 Stress Testing

When the individual behavior of risks is well understood but the way in which they interact is uncertain, Sklar's formula is the ideal tool to explore the impact of various dependence scenarios on quantities of interest.

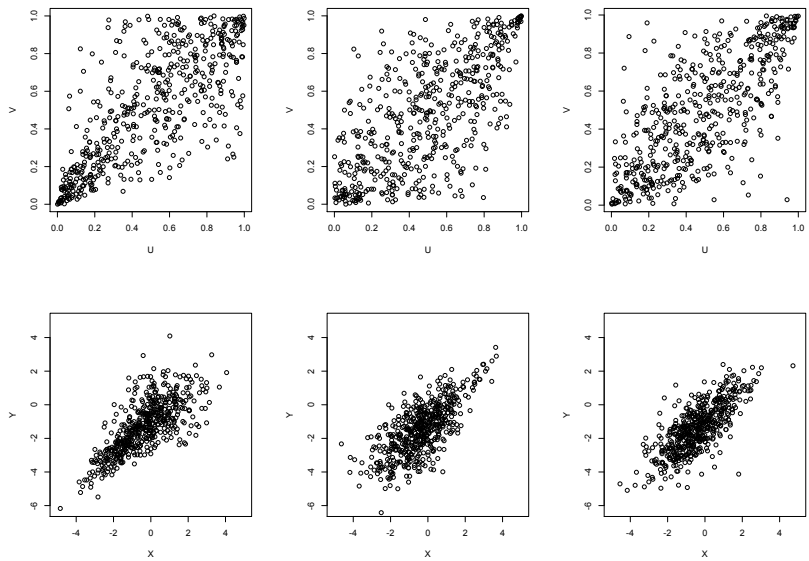


FIGURE 4.4: Top: scatterplots of random samples of size 517 from the Clayton (left), Gumbel (middle) and Student $t_4$ copulas (right); bottom: corresponding samples with fitted Normal margins.

For example, suppose that the losses of contents ($Y$) and profits ($Z$) are indeed Log-Normal. By calibrating these distributions to the data at hand, we can compute $\Pr(Y \leq y)$ and $\Pr(Z \leq z)$ for any real numbers $y$ and $z$. By varying $C$ in expression (4.1), we can then generate a host of models that can be used to investigate, say, the distribution of the total loss of contents *and* profits, $Y + Z$.

To illustrate, consider the Clayton, Gumbel, and Student $t_4$ copulas. These are merely three common choices among the wealth of copula families listed in the popular 1999 book *An Introduction to Copulas* by the American mathematician Roger Nelsen. The word "family" is used here because copulas typically involve a constant $\theta$, called a parameter, that can be used to adjust the degree of dependence. For example, the copula defined, for all $u, v \in (0, 1)$, by

$$C(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta} \tag{4.3}$$

is the Clayton copula with parameter $\theta > 0$. The theoretical value of Kendall's tau for this model is $\tau = \theta/(\theta + 2)$, which increases from 0 to 1 as $\theta \to \infty$.

Displayed in Figure 4.4 are simulated samples of size 517 from the models corresponding to a Clayton (left), Gumbel (middle) or $t_4$ (right) copula. The top row displays samples from the copulas themselves; their parameters were calibrated so that Kendall's tau is the same as in the data. For example, to

adjust the Clayton copula to the pair $(Y, Z)$, we solved the equation $\theta/(\theta+2) = .46$, leading to $\theta = 1.7$. The very same samples were then transformed so as to reproduce the marginal behavior of $Y$ and $Z$. To generate a pair from $(X, Y)$, one generates a pair $(U, V) = (u, v)$ from the chosen copula and then one finds the values of $x$ and $y$ for which $F(x) = u$ and $G(y) = v$; this is done by numerical inversion. One then sets $X = x$ and $Y = y$.

The result of this transformation is shown on the logarithmic scale in the bottom row of Figure 4.4. The shape of the point clouds makes it abundantly clear that different copulas can induce very different dependence patterns. The same conclusion applies when the models are used to compute probabilities or make predictions. This is illustrated in two different ways in Tables 4.1–4.2.

Table 4.1 reports estimates of the probability that the total loss of contents and profits, $Y + Z$, exceeds either 50 or 100 million (M) Danish kroner. All the estimates are based on the same marginal distributions; only the copula varies. In addition to the Clayton, Gumbel and Student $t_4$ copulas, we included the Normal copula which corresponds to the joint Normal model. The estimates were obtained by simulating one million pairs from each of these models. As one can see, the probability that $Y + Z > 50$M varies by a factor of about 1.75 between the most (Clayton) and the least (Gumbel) optimistic scenario. This factor grows to 2.18 for the probability that the total loss exceeds 100M.

At a first glance, the probabilities reported in Table 4.1 may seem very small and the differences between them appear to be inconsequential. Nevertheless, these differences can have a substantial impact on the financial reserves that regulators require the insurance companies to keep in order to cover potentially large claims. A quantity that often enters the calculations of such reserves is the Value-at-Risk, commonly abbreviated VaR. In statistical terms, the VaR at level $\alpha$ is the upper $\alpha$th percentile of the loss distribution or, in other words, the "maximum loss which is not exceeded with probability $\alpha$"; see Section 2.2.2 of McNeil et al. (2005) for further details and discussion.

Typically, the VaR is computed at $\alpha = .99$ or even .999. Table 4.2 reports the VaR (in millions of Danish kroner) at both levels when Log-Normal marginal loss distributions are linked by different copulas. While the VaR estimates are quite close at the level .99, they vary much more when $\alpha = .999$. These differences show how important is the choice of dependence structure.

TABLE 4.1: Estimates of the probability of a total loss of contents and profits $(Y + Z)$ based on four different copulas and Log-Normal margins.

| | Copula | | | |
|---|---|---|---|---|
| | Clayton | Normal | Student $t_4$ | Gumbel |
| $10^3 \times \Pr(Y + Z > 50\text{M})$ | 1.3 | 1.8 | 2.0 | 2.3 |
| $10^3 \times \Pr(Y + Z > 100\text{M})$ | .22 | .3 | .37 | .48 |

TABLE 4.2: Estimates of the Value-at-Risk for the total loss of contents and profits ($Y + Z$) based on four different copulas and Log-Normal margins.

| | Copula | | | |
|---|---|---|---|---|
| | Clayton | Normal | Student $t_4$ | Gumbel |
| VaR at $\alpha = .99$ | 21.44 | 23.07 | 23.61 | 24.21 |
| VaR at $\alpha = .999$ | 55.05 | 63.42 | 68.75 | 71.69 |

Imagine, for instance, that a financial reserve is determined from the VaR at level .999 using a Clayton copula. If the true dependence structure happens to be Gumbel, the reserve would then be underestimated by over 16M Danish kroner, or nearly 3M Canadian dollars!

Needless to say, errors of judgment in the choice of margins can also lead to serious underestimation of the VaR or other measures of risk at such high levels. In the Danish fire loss example, there is evidence that the upper tails of the marginal distributions are much heavier than those of the Log-Normal distribution. If this were taken into account, the resulting VaR estimates would be quite a bit higher. More details and additional illustrations can be found in McNeil et al. (2005); see also Chapter 15 by Bruno Rémillard.

### 4.3.4 Validate, Validate, Validate!

The financial crisis of 2008 is arguably the most famous case of risk underestimation due to an inappropriate choice of copula and marginal distributions. It was described by the English journalist Felix Salmon in an award winning article entitled "Recipe for disaster: The formula that killed Wall Street." Originally published online (*Wired Magazine*: 17.03) and later reproduced in print (Salmon, 2012), the article recounts how quants (i.e., quantitative analysts from the financial sector) underestimated a joint probability of default by assuming that loans in different pools, say $A$ and $B$, were linked by a Normal copula. In his article, Salmon cites

$$\Pr(T_A < 1, T_B < 1) = \Phi_2[\Phi^{-1}\{F_A(1)\}, \Phi^{-1}\{F_B(1)\}, \gamma].$$

as the formula that was "instrumental in causing the unfathomable losses that brought the world financial system to its knees." This formula simply states that the probability of observing loan defaults in pools $A$ and $B$ within one year is given by a distribution of the form (4.1) in which the margins are denoted $F_A$ and $F_B$ rather than $F$ and $G$, and

$$C(u, v) = \Phi_2\{\Phi^{-1}(u), \Phi^{-1}(v), \gamma\}$$

is the Normal (or Gaussian) copula of the bivariate Normal distribution $\Phi_2$ with standard Normal margins $\Phi$ and Pearson's correlation $\gamma$.

David X. Li, a quant of Chinese origin, proposed this model (Li, 2000). While he and his formula were repeatedly blamed for bringing on the financial crisis, this is clearly an oversimplification, though it makes a good story with a Canadian twist, given that Li did his graduate work at Laval and Waterloo. In effect, Li himself was aware of the limitations of his model (Meissner, 2008, p. 71), but this did not prevent the bankers from using it extensively. They failed to realize that it led to an underestimation of the risk of joint defaults because it has no tail dependence (i.e., Joe's coefficient is 0).

Compounded with the fact that in financial markets, dependence often changes over time, the static Normal copula model led to prices and capital reserves that were too low, and therefore unduly attractive for traders and bankers in quest of financial gain. Does this mean that the Normal copula should be locked up, never to be used again? Hardly. It may not be appropriate for modeling default times or fire insurance claims, but there are situations in which it makes perfect sense. Copula models are versatile tools that must be used wisely. What we need is to understand their limitations and, even more important, to make sure that they are fit for use in the situation at hand. This requires good data and proper model validation.

## 4.4 Proof of the Pudding Is in the Ranking

The strategy for building a statistical model is always more or less the same. First look at the data and choose a class of models that captures its main characteristics. Next, use the data to estimate the model parameters efficiently and check whether the fit is satisfactory. If it isn't, try over again and make sure to validate your choice before you use the model to compute probabilities or make predictions that inform a decision process. What is meant by "efficient" and "satisfactory" depends on the context and involves a fair dose of pragmatism.

When it comes to copula modeling, the main stumbling block is the fact that except in the rare instance where the individual behavior of the variables is known with absolute certainty, we never actually observe data from the copula. Remember from (4.2) that $C$ is the distribution function of the uniform variables $U = F(X)$ and $V = G(Y)$. What we collect and observe, however, are data from $X$ and $Y$, say $(X_1, Y_1), \ldots, (X_n, Y_n)$. If the formulas for $F$ and $G$ were available (that is, if we knew the exact distributions of $X$ and $Y$), we could simply transform these data into a sample from $(U, V)$ by setting $U_1 = F(X_1)$ and $V_1 = G(Y_1)$, etc. We could then plot the pairs $(U_1, V_1), \ldots, (U_n, V_n)$ to start the model building process. However, we simply don't have this luxury.

It was only in the late 1970s and 1980s that statisticians gradually realized that one could uncover the concealed copula by ranking the observations in

ascending order and plotting the resulting pairs of ranks divided by the sample size $n$. As a simple illustration, suppose that only $n = 3$ fire claims are observed and that fictitious amounts for losses of contents and profits are as follows:

$$(Y_1, Z_1) = (2900, 50), \quad (Y_2, Z_2) = (80, 45), \quad (Y_3, Z_3) = (150, 120).$$

The ranks of the $Y$ values are then $R_1 = 3$, $R_2 = 1$, $R_3 = 2$ while those of the $Z$ values are $S_1 = 2$, $S_2 = 1$, $S_3 = 3$. The rank plot would then show the points

$$\left(\frac{R_1}{n}, \frac{S_1}{n}\right) = \left(\frac{3}{3}, \frac{2}{3}\right), \quad \left(\frac{R_2}{n}, \frac{S_2}{n}\right) = \left(\frac{1}{3}, \frac{1}{3}\right), \quad \left(\frac{R_3}{n}, \frac{S_3}{n}\right) = \left(\frac{2}{3}, \frac{3}{3}\right).$$

This does not amount to much of a plot because the sample is impractically small. However, look at Figure 4.5 and see what happens when this procedure is used on the Danish fire loss data. It transpires from the graphs that the dependence structure differs from one pair to the next.

For simplicity, focus on the dependence between losses of contents and profits. Comparing the point cloud in the lower panel of Figure 4.5 to the three scatterplots in the top row of Figure 4.4, would you rather pick the Clayton, Gumbel or Student $t_4$ copula as a plausible dependence structure between $Y$ and $Z$? Though this strategy may sound simplistic, this is in essence what model builders need to do; and it isn't just black magic. Expressed in mathematical terms, the points on the rank plot constitute the support of what statisticians call the empirical copula, a "random function" defined by

$$\begin{aligned} C_n(u, v) &= \text{proportion of fire events for which } R_i \le nu \text{ and } S_i \le nv, \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left(\frac{R_i}{n} \le u, \frac{S_i}{n} \le v\right). \end{aligned}$$

As it turns out, $C_n$ gets closer and closer to the true $C$ as the sample size $n$ becomes infinitely large. In the limit, $\sqrt{n}\,(C_n - C)$ is related to a multidimensional version of the well-known Brownian motion. This interesting phenomenon was first studied by the German probabilist Ludger Rüschendorf (Rüschendorf, 1976) and the French academician Paul Deheuvels (Deheuvels, 1979). Many refinements and extensions have since been developed, notably by Canadian statisticians Bruno Rémillard and Kilani Ghoudi (Ghoudi and Rémillard, 1998).

At first sight, it seems that in the Danish fire loss data, the Gumbel copula is preferable to the Clayton or the Student $t_4$ as a descriptor of the dependence between losses of contents ($Y$) and profits ($Z$). The formula for the Gumbel copula involves a parameter $\theta$, viz.

$$C(u, v) = e^{-\{|\log(u)|^\theta + |\log(v)|^\theta\}^{1/\theta}}$$

and the value of $\theta \ge 1$ is linked to Kendall's tau via the relation $\tau = 1 - 1/\theta$.

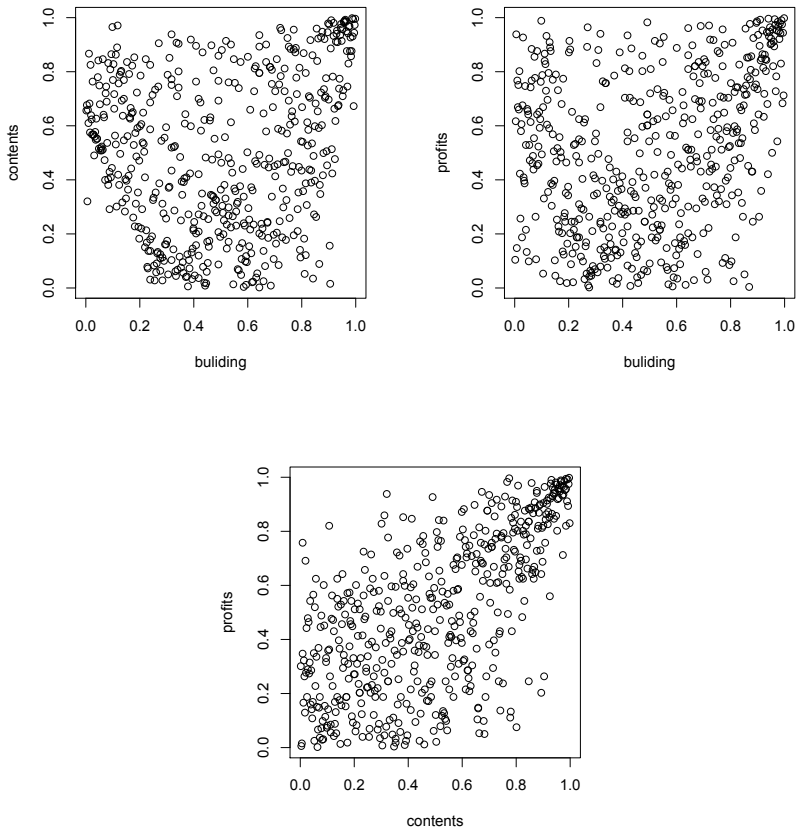


FIGURE 4.5: Rank plots for $(X, Y)$ (top left), $(X, Z)$ (top right) and $(Y, Z)$ (bottom).

Proceeding as we did earlier with the Clayton copula, we could calibrate $\theta$ from the data by solving the equation $1 - 1/\theta = .46$, which yields $\theta = 1/.54 \approx 1.85$. This estimate, denoted by $\hat{\theta}_n$ to remind us that it is based on a data sample of size $n$, gives us a good idea about the value of the unknown parameter $\theta$. To make it even more precise, we can exploit the so-called rank-based likelihood. If

$$c_\theta(u, v) = \frac{\partial^2}{\partial u \partial v} C_\theta(u, v)$$

is the density of the copula $C_\theta$, obtained by differentiating $C_\theta$ once with respect to $u$ and once with respect to $v$, the rank-based likelihood is a function $\ell$ of

the parameter $\theta$ given by

$$\ell(\theta) = \sum_{i=1}^{n} \log \left\{ c_\theta \left( \frac{R_i}{n}, \frac{S_i}{n} \right) \right\}.$$

A much more efficient estimate is then given by the value of $\theta$ that maximizes $\ell(\theta)$. This estimation technique, originally studied by Genest et al. (1995) and Shih and Louis (1995), is now deemed the "industry standard." We could also maximize the full likelihood involving the margins — either globally or by steps as in Joe (2005) — but proceeding from ranks protects us against errors in the choice of the marginal distributions (Kim et al., 2007).

Having settled on an estimate $\hat{\theta}_n$ of the parameter, we need to check whether the selected copula $C_\theta$ with $\theta = \hat{\theta}_n$ does indeed reproduce the pattern observed in the data, represented by the empirical copula $C_n$. The closer $C_{\hat{\theta}_n}$ is to $C_n$, the better the model. This proximity can be conveniently measured by the statistic

$$S_n = \sum_{i=1}^{n} \left\{ C_n \left( \frac{R_i}{n}, \frac{S_i}{n} \right) - C_{\hat{\theta}_n} \left( \frac{R_i}{n}, \frac{S_i}{n} \right) \right\}^2$$

advocated and investigated by Genest et al. (2009b).

For the losses of contents ($Y$) and profits ($Z$), one gets $S_n = .0799$, which seems small, but beware of appearances! What do we know of the sampling variation of $S_n$ in datasets of size 517 when the copula is Gumbel with parameter $\theta = 1.85$? One way to find out is to generate thousands of datasets from that copula and compute the value of the statistic for each one of them. This technique, called the "parametric bootstrap," was shown to be valid for copula models by Genest and Rémillard (2008). Another popular computer-intensive method is based on the Multiplier Central Limit Theorem first applied in this context by Rémillard and Scaillet (2009); see also Kojadinovic et al. (2011).

At the end of the day, it turns out that a value of .0799 or bigger is a pretty rare occurrence. The probability that $S_n$ is at least .0799 when the true copula is Gumbel is then about .0005 or about 5 in 10,000. Because this value is so small, the Gumbel copula no longer seems satisfactory and the search for a more adequate dependence structure must go on. One viable option is the "survival Clayton," defined, for all $u, v \in (0, 1)$, by

$$\bar{C}_\theta(u, v) = 1 - u - v + C_\theta(1 - u, 1 - v)$$

with $C_\theta$ as in Equation (4.3). In that case, $S_n = .0345$ and $\Pr(S_n \geq .0345) \approx .0892$ or 8.92%. This probability is larger than the conventional 5% significance level, meaning that there is no strong statistical evidence against the survival Clayton model.

But just as with the choice of marginal distributions, the selection of an appropriate copula requires much care and all sorts of checks and balances. Many other questions can be asked. For instance, is the copula exchangeable,

that is, does $C(u, v) = C(v, u)$ always hold? Does the copula exhibit tail dependence? Does it belong to one of the general classes of copulas proposed in the literature, such as Archimedean, elliptical or extreme-value? Procedures are actively being developed to answer these and other questions, and here again many Canadians are contributing, including Jean-François Quessy in Trois-Rivières; see, e.g., Genest and Nešlehová (2013).

## 4.5 Applications and Future Challenges

Since the year 2000, the popularity of copula-based modeling has expanded considerably. For a brief history of this development, see Genest et al. (2009a). Today, copulas are commonly used in actuarial science, biostatistics and health sciences, environmental sciences, hydrology, finance, risk management, and financial economics, among others. Major books and review articles that discuss copula modeling at length include Joe (1997), Frees and Valdez (1998), Nelsen (2006), Cherubini et al. (2004), Denuit et al. (2005), McNeil et al. (2005), Genest and Favre (2007), Salvadori et al. (2007), and Patton (2012).

In many fields, researchers encounter data that pose challenges far beyond what we sketched here. For example, financial and environmental data typically exhibit time trends and the dependence patterns may also change over time. In medical studies, variables are often measured on a discrete scale and may be only partially observed. Furthermore, it is often necessary to include explanatory variables in the construction of the marginal distributions. Some of these issues are considered in Genest and Nešlehová (2007), Kolev and Paiva (2009) and researchers at the University of Waterloo (Cook et al., 2010), among others. New copula-based tools are constantly being developed to meet these challenges and many researchers across Canada are at the forefront of this work.

With the proliferation of large datasets from a variety of sources, perhaps the most pressing and ubiquitous challenge is posed by the need to "leverage big data." This calls for ways to build dependence models involving hundreds, and even thousands of variables. Most of what we described here can be extended to any dimension; we limited ourselves to two or three variables for simplicity only. However, the use of these techniques in high-dimensional contexts poses serious statistical and numerical challenges. Moreover, several of the copula families used for two or three variables do not extend easily to these situations or produce unduly restrictive dependence structures.

New model building strategies for high-dimensional data are being actively developed. Among them, the so-called vine copula construction, rooted in the work of Harry Joe (1997), is quite popular at present (Kurowicka and Joe, 2010). For recent books that look at vine copula constructions and their simulation in a financial context, see Ma (2010) and Mai and Scherer (2012). More

broadly, specialized conferences are regularly being held to stimulate research on copula modeling, including recent workshops at the Centre de recherches mathématiques in Montréal and at the Banff International Research Station.

## Acknowledgments

This research was supported by the Canada Research Chairs Program and grants from the Natural Sciences and Engineering Research Council of Canada and the Fonds de recherche du Québec – Nature et technologies.

## About the Authors

**Christian Genest** is the director of the Institut des sciences mathématiques du Québec and a professor of statistics at McGill University, where he holds a Canada Research Chair in Stochastic Dependence Modeling. He received a BSpSc from the Université du Québec à Chicoutimi, an MSc from the Université de Montréal, and a PhD from the University of British Columbia. He has served as editor of *The Canadian Journal of Statistics* and as president of the Statistical Society of Canada. He was the first winner of the CRM–SSC Prize and was awarded the 2011 SSC Gold Medal for research. He is a fellow of the American Statistical Association and the Institute of Mathematical Statistics.

**Johanna G. Nešlehová** is an associate professor of statistics at McGill University. She studied mathematics and statistics at Univerzita Karlova v Praze, Universität Hamburg and Carl von Ossietzky Universität Oldenburg. Before moving to Canada, she was at ETH Zürich for five years, as a postdoctoral fellow at RiskLab Switzerland and subsequently as Heinz Hopf Lecturer. Her primary research interests are multivariate extreme-value theory, dependence modeling, and applications of statistics to risk management and health. She was elected a member of the International Statistical Institute in 2011.

## Bibliography

Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula Methods in Finance*. Wiley, New York.

Cook, R. J., Lawless, J. F., and Lee, K.-A. (2010). A copula-based mixed Poisson model for bivariate recurrent events under event-dependent censoring. *Statistics in Medicine*, 29:694–707.

Cox, D. R. and Small, N. J. H. (1978). Testing multivariate normality. *Biometrika*, 65:263–272.

Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. *Académie Royale de Belgique, Bulletin de la Classe des Sciences, 5*e *Série*, 65:274–292.

Denuit, M., Dhaene, J., Goovaerts, M. J., and Kaas, R. (2005). *Actuarial Theory for Dependent Risk: Measures, Orders and Models.* Wiley, New York.

Embrechts, P., McNeil, A. J., and Straumann, D. (2002). Correlation and dependence in risk management: Properties and pitfalls. In *Risk Management: Value at Risk and Beyond (Cambridge, 1998)*, pp. 176–223. Cambridge University Press, Cambridge.

Frees, E. W. and Valdez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, 2:1–25.

Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12:347–368.

Genest, C., Gendron, M., and Bourdeau-Brien, M. (2009a). The advent of copulas in finance. *The European Journal of Finance*, 15:609–618.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82:543–552.

Genest, C. and MacKay, R. J. (1986a). Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données. *The Canadian Journal of Statistics*, 14:145–159.

Genest, C. and MacKay, R. J. (1986b). The joy of copulas: Bivariate distributions with uniform marginals. *The American Statistician*, 40:280–283.

Genest, C. and Nešlehová, J. (2007). A primer on copulas for count data. *ASTIN Bulletin*, 37:475–515.

Genest, C. and Nešlehová, J. G. (2013). Assessing and modeling asymmetry in bivariate continuous data. In *Copulae in Mathematical and Quantitative Finance, Proceedings of the Workshop Held in Cracow, 10–11 July 2012*, pp. 91–114. Springer, Berlin.

Genest, C. and Rémillard, B. (2008). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Annales de l'Institut Henri Poincaré: Probabilités et Statistiques*, 44:1096–1127.

Genest, C., Rémillard, B., and Beaudoin, D. (2009b). Omnibus goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics & Economics*, 44:199–213.

Ghoudi, K. and Rémillard, B. (1998). Empirical processes based on pseudo-observations. In *Asymptotic Methods in Probability and Statistics (Ottawa, ON, 1997)*, pp. 171–197. North-Holland, Amsterdam.

Hoeffding, W. (1940). Maßstabinvariante Korrelationstheorie für diskontinuierliche Verteilungen. *Archiv für mathematische Wirtschafts- und Sozialforschung*, 7:4–70.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall, London.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94:401–419.

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

Kim, G., Silvapulle, M. J., and Silvapulle, P. (2007). Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics & Data Analysis*, 51:2836–2850.

Kojadinovic, I., Yan, J., and Holmes, M. (2011). Fast large-sample goodness-of-fit tests for copulas. *Statistica Sinica*, 21:841–871.

Kolev, N. and Paiva, D. (2009). Copula-based regression models: A survey. *Journal of Statistical Planning and Inference*, 139:3847–3856.

Kurowicka, D. and Joe, H. (2010). *Dependence Modeling: Handbook on Vine Copulae*. World Scientific Publishing, Singapore/SG.

Lehmann, E. L. (1966). Some concepts of dependence. *The Annals of Mathematical Statistics*, 37:1137–1153.

Li, D. X. (2000). On default correlation: A copula function approach. *Journal of Fixed Income*, 9(4):43–54.

Ma, J. (2010). *Higher-dimensional Copula Models and Their Application: Bayesian Inference for D-vine Pair-copula Constructions Based on Different Bivariate Families*. VDM Publishing.

Mai, J.-F. and Scherer, M. (2012). *Simulating Copulas: Stochastic Models, Sampling Algorithms and Applications*. Imperial College Press, London.

McNeil, A. J. (1997). Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin*, 27:117–137.

McNeil, A. J., Frey, R., and Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, Princeton, NJ.

Meissner, G. (2008). *The Definitive Guide to CDOs*. Risk Books, London.

Müller, A. and Stoyan, D. (2002). *Comparison Methods for Stochastic Models and Risks*. Wiley, Chichester.

Nelsen, R. B. (2006). *An Introduction to Copulas,* Second Edition. Springer, New York.

Patton, A. J. (2012). A review of copula models for economic time series. *Journal of Multivariate Analysis*, 110:4–18.

Rémillard, B. and Scaillet, O. (2009). Testing for equality between two copulas. *Journal of Multivariate Analysis*, 100:377–386.

Rüschendorf, L. (1976). Asymptotic distributions of multivariate rank order statistics. *The Annals of Statistics*, 4:912–923.

Salmon, F. (2012). The formula that killed Wall Street. *Significance*, 9(1):16–20.

Salvadori, G., De Michele, C., Kottegoda, N. T., and Rosso, R. (2007). *Extremes in Nature: An Approach Using Copulas.* Springer, Dordrecht.

Schweizer, B. and Wolff, E. F. (1981). On nonparametric measures of dependence for random variables. *The Annals of Statistics*, 9:879–885.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality: Complete samples. *Biometrika*, 52:591–611.

Shih, J. H. and Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 51:1384–1399.

Sklar, A. (1959). Fonctions de répartition à $n$ dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris*, 8:229–231.

Spearman, C. E. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101.