Bivariate Archimedean Copula Models for Censored Data in Non-Life Insurance

Michel Denuit
Universite Catholique de Louvain, Belgium, denuit@stat.ucl.ac.be

Oana Purcaru
Universite Catholique de Louvain, Belgium., purcaru@stat.ucl.ac.be

Ingrid Van Keilegom
Universite Catholique de Louvain., vankeilegorn@stat.ucl.ac.be

Follow this and additional works at: http://digitalcommons.unl.edu/joap

Part of the Accounting Commons, Business Administration, Management, and Operations Commons, Corporate Finance Commons, Finance and Financial Management Commons, Insurance Commons, and the Management Sciences and Quantitative Methods Commons

http://digitalcommons.unl.edu/joap/5

This Article is brought to you for free and open access by the Finance Department at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Journal of Actuarial Practice 1993-2006 by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
Bivariate Archimedean Copula Models for Censored Data in Non-Life Insurance

Michel Denuit,∗ Oana Purcaru,† and Ingrid Van Keilegom‡

Abstract§

We describe a methodology based on Archimedean copulas for analyzing nonlife insurance data with censoring present. Specifically, we propose a graphical selection procedure for the nonparametric estimation of the generator. An actual loss-ALAE data set is used for the numerical illustrations and for comparisons of our approach to a few others.

Key words and phrases: dependence, Archimedean copula, model selection, censored data

∗Michel Denuit, Ph.D., is a fellow of the Belgian Actuarial Society, and a professor of actuarial mathematics at the Université Catholique de Louvain, Belgium.

Dr. Denuit’s address is: Institut des Sciences Actuarielles, Université Catholique de Louvain, 6 rue des Wallons, B-1348 Louvain-la-Neuve, BELGIUM. E-mail: denuit@stat.ucl.ac.be

†Oana Purcaru, Ph.D., is a researcher in statistics at the Université Catholique de Louvain, Belgium.

Dr. Purcaru’s address is: Institut de Statistique, Université Catholique de Louvain, 20 Voie du Roman Pays, B-1348 Louvain-la-Neuve, BELGIUM. E-mail: purcaru@stat.ucl.ac.be

‡Ingrid Van Keilegom, Ph.D., is professor of statistics at the Université Catholique de Louvain.

Dr. Van Keilegom’s address is: Institut de Statistique, Université Catholique de Louvain, 20 Voie du Roman Pays, B-1348 Louvain-la-Neuve, BELGIUM. E-mail: vankeilegom@stat.ucl.ac.be

§We thank Philippe Lambert, Jean-Marie Rolin, Ronald Pruitt, and Mark van der Laan. Michel Denuit received financial support from Communauté française de Belgique under contract “Projet d’Actions de Recherche Concertées” ARC 04/09-320 and from Banque Nationale de Belgique under grant “Risk Measures and Economic Capital.” Ingrid Van Keilegom was supported from the IAP research network nr. P5/24 of the Belgian Government (Belgian Science Policy).
1 Introduction

Various processes in casualty insurance involve correlated pairs of variables. A prominent example is the loss and allocated loss adjustment expense (ALAE)\(^1\) associated with a single claim. As expensive claims generally take longer to be settled (thus inducing considerable costs for the insurance company), one may expect a positive dependence between losses and their associated ALAEs, i.e., large values for losses tend to be associated with large values for ALAEs.

This positive association has some practical implications in the pricing of certain reinsurance treaties such as excess-of-loss treaties.\(^2\) This positive association also contributes to the reinsurer's expenses associated with settlement costs on a prorata basis. Neglecting the dependence exhibited by reinsurance data may lead to underestimation of the expected reinsurer's payment. It is therefore crucial for the reinsurer to have an appropriate model for the random pair (loss, ALAE).

Typically, a given amount of loss is divided between the insurer and the reinsurer as follows. The insurer pays the loss from ground up to a specified amount \(r\) called the insurer's retention. The reinsurer covers the claim from \(r\) up to a maximum limit of \(w\). The excess over \(w\) remains with the direct insurer (but a policy limit, i.e., an upper bound to the amount paid by the insurer to the policyholder, may be specified in the contract). Let \(X\) denote the loss and \(Y\) denote the associated ALAE. Assuming a prorata sharing of expenses, the reinsurer's payment for a given realization of loss and associated ALAE pair, \((X, Y)\), is given by

\[
g(X, Y) = \begin{cases} 
0 & \text{if } X < r, \\
X - r + \left(\frac{X-r}{X}\right) Y & \text{if } r \leq X < w, \\
w - r + \left(\frac{w-r}{w}\right) Y & \text{if } X \geq w.
\end{cases}
\]

The net premium of this treaty involves the computation of \(\mathbb{E}[g(X, Y)]\), which in turn requires the knowledge of the joint distribution for the pair \((X, Y)\).

The copula construction is very useful for the analysis of dependence in actuarial science. Applications of copulas to insurance data

---

\(^1\)The allocated loss adjustment expense is the insurance company's expense (e.g., lawyers' fees and claims investigation expenses) that is specifically attributable to the settlement of individual claims.

\(^2\)In an excess-of-loss treaty the reinsurer covers the largest losses, i.e., those losses exceeding some high threshold called the retention limit of the direct insurer, and pays that part of the loss exceeding this threshold.
modeling have been proposed by several authors, e.g., Carrière (2000), Frees, Carrière and Valdez (1996), Frees and Valdez (1998), Klugman and Parsa (1999), Valdez (2001) and Embrechts et al. (2002). Recently, research has focused on a subclass of copulas called the Archimedean copula class, which indexes the copula by a univariate function (called the generator) and therefore yields more tractable analytical properties. Many well-known systems of bivariate distributions belong to the Archimedean class. Frailty models also fall under that general description. As shown by Genest and McKay (1986a and b), this class of copulas is wide and analytically tractable. Its elements have stochastic properties that make them attractive for the statistical treatment of data. The joint modeling in parametric settings of loss-ALAE data has been examined by Frees and Valdez (1998) (Pareto marginals and Gumbel copula) and Klugman and Parsa (1999) (inverse paralogistic for loss, inverse Burr for ALAE and Frank copula).

Archimedean copulas are appealing in life insurance, where they naturally arise from frailty models: assuming that a group of individuals share a common frailty yields an Archimedean copula for the remaining lifetimes (with the inverse of the frailty Laplace transform as generator). This construction loses its appeal in nonlife insurance. The Archimedean construction remains nevertheless attractive because it allows for flexibility and keeps the model mathematically tractable.

Of course, (Archimedean) copula modeling is not the only approach to take dependence into account in nonlife insurance problems. When the data are heavy tailed, multivariate extreme value theory can also be helpful. We will come back to the modeling issue in the conclusion to this paper.

Because copulas characterize the dependence structure of random vectors once the effect of the marginals has been factored out, identifying and fitting a copula to data is not an easy task. In practice, it is often preferable to restrict the search of an appropriate copula to some reasonable family, such as the Archimedean one. Then, it is useful to have simple graphical procedures to select the best fitting model among some competing alternatives for the data at hand.

Starting from the assumption that the Archimedean dependence structure is appropriate (an assumption that we will retain throughout this paper), Genest and Rivest (1993) proposed a procedure for selecting a parametric generator. Their method relies on the estimation of the univariate distribution function associated with the probability integral transformation and requires complete data. Specifically, the best fitting Archimedean model is the one where its probability integral transformation distribution is the closest to its empirical estimate.
Wang and Wells (2000b) extended the idea of Genest and Rivest (1993) to right-censored bivariate failure-time data. This type of censorship is not the one typically encountered in actuarial work. Because the censoring issue is handled in the stage of estimating the bivariate distribution function, however, the approach proposed by Wang and Wells (2000b) is flexible enough to deal with other censoring mechanisms. This is precisely the route we follow in this paper to deal with the modeling of losses and ALAE.

Frees and Valdez (1998) have applied techniques developed by Genest and Rivest (1993) for complete data to loss-ALAE data in order to select the appropriate generator. As pointed out by Frees and Valdez (1998, Section 4.2.1), censoring in the loss variable is ignored in the identification process. We will develop in this paper an appropriate nonparametric estimator of the joint distribution of loss-ALAE taking into account the particular censorship present in the data. Specifically, we follow the general approach described in Wang and Wells (2000b), but instead of using Dabrowska (1988) estimator for the bivariate distribution, we use the estimator proposed in Akritas (1994), because only the loss variable is subject to censoring.

This paper is organized as follows: Section 2 reviews the notion of copulas and gives some examples from the Archimedean family. In Section 3, we propose a new nonparametric estimator for the generator, that takes into account the fact that losses may be censored whereas ALAEs are completely observable. This nonparametric estimation then serves as a benchmark to select an appropriate parametric Archimedean copula. Numerical illustrations are given in Section 4 using actual data. Section 5 concludes.

2 Archimedean Copulas

The word “copula” was first employed in a statistical sense by Sklar (1959) in a theorem that now bears his name. His idea was to separate a joint distribution function into two parts: one that describes the dependence structure (the copula) and parts that describe the marginal behavior only. Broadly speaking, a copula is (the restriction to the unit square \([0, 1]^2\) of) a joint distribution function for a bivariate random vector with unit uniform marginals.

Sklar’s theorem elucidates the role that copulas play in the relationship between multivariate distribution functions and their univariate margins. Specifically, given a bivariate distribution function \(F\) with univariate marginal distribution functions \(F_X\) and \(F_Y\), there exists a copula
C such that for all \((x, y) \in \mathbb{R}^2\) the joint distribution function \(F\) can be represented as:

\[
F(x, y) = C(F_X(x), F_Y(y)), \quad (x, y) \in \mathbb{R}^2.
\] (1)

When the marginals \(F_X\) and \(F_Y\) are continuous, then the copula \(C\) in (1) is unique. Otherwise \(C\) is uniquely determined on \(\text{Range}(F_X) \times \text{Range}(F_Y)\). Conversely, if \(C\) is a copula and \(F_X\) and \(F_Y\) are distribution functions, then the function \(F\) defined by equation (1) is a bivariate distribution function with margins \(F_X\) and \(F_Y\). Formal proofs can be found, e.g., in Nelsen (1999). Next we define the Archimedean family of copulas.

Consider a twice-differentiable strictly decreasing and convex function \(> : [0,1] \to [0,\infty]\) satisfying \(>(1) = 0\). These requirements are enough to guarantee that \(>\) has an inverse \(<^{-1}\) having also two derivatives. Every such function \(>\) generates a bivariate distribution function \(C_{<}\) whose marginals are uniform on the unit interval (i.e., a copula) given by

\[
C_{<}(u, v) = \begin{cases} 
>^{-1} \{> (u) + > (v)\} & \text{if } > (u) + > (v) \leq > (0), \\
0 & \text{otherwise}, 
\end{cases}
\] (2)

for \(0 \leq u, v \leq 1\). Copulas \(C_{<}\) of the form equation (2) are referred to as Archimedean copulas. The function \(>\) is called the generator of the copula. Only \(>\) functions satisfying \(\lim_{t \to 0^+} > (t) = \infty\) are used in this paper. This ensures that \(C_{<}\) is absolutely continuous. Now, a bivariate distribution function \(F\) with marginals \(F_X\) and \(F_Y\) is said to be generated by an Archimedean copula if, and only if, equation (1) holds with an Archimedean copula \(C_{<}\).

A useful tool for studying Archimedean copulas is the bivariate probability integral transformation, which is the bivariate analog of the probability integral transform (PIT).\(^3\) In particular, the copula \(C\) for \((X, Y)\) is just the joint distribution function for the random couple \((F_X(X), F_Y(Y))\) provided \(F_X\) and \(F_Y\) are continuous. Let us define the bivariate PIT of \((X, Y)\) with joint distribution function \(F\) as \(Z = F(X, Y)\). It is not generally true that the distribution function \(K\) of \(Z\) is uniform on \([0,1]\), even when \(F\) is continuous. Moreover, \(K\) does not characterize \(F\) because \(K\) does not contain any information about the marginals \(F_X\) and \(F_Y\). Indeed, we have that \(Z = F(X, Y) = C(U, V)\) where \((U, V)\) admits \(C\) as joint distribution function.

\(^3\)The probability integral transform theorem states that given any random variable \(X\) with continuous distribution function \(F_X, F_X(X)\) is uniformly distributed on the interval \([0,1]\). This fundamental result underlies many statistical procedures.
Genest and Rivest (1993) studied the bivariate PIT for Archimedean copulas and obtained the following result: Let \((U, V)\) be a random couple with unit uniform marginals and joint distribution function \(C_\phi\). The distribution function \(K\) of \(Z = C_\phi(U, V)\) is given by

\[
K(z) = z - \lambda(z) \quad \text{where} \quad \lambda(\xi) = \frac{\phi(\xi)}{\phi'(1)(\xi)}, \quad \text{for} \quad 0 < \xi \leq 1. \tag{3}
\]

Once the copula is known, it is important to measure the extent to which \(X\) and \(Y\) are dependent. Loosely speaking, the objective of dependence measures is to capture the fact that the probability of having large (or small) values for both components is high, while the probability of having large values for the first component together with small values for the second component, or vice versa, is small. In general, the covariance will not reveal the whole information on the dependence structure of a random couple. Hence, practitioners should also be aware of other dependence concepts such as rank correlations. Kendall's rank correlation coefficient (often called Kendall's tau) is a nonparametric measure of association based on the number of concordances and discordances in a sample of paired observations. Concordance occurs when pairs of observations vary together, and discordance occurs when pairs of observations vary differently.

More specifically, a pair of observations is concordant if the observation with the larger value of \(X\) has also the larger value for \(Y\). The pair is discordant if the observation with the larger value of \(X\) has the smaller value of \(Y\). If \((X, Y)\) and \((X', Y')\) are independent and identically distributed, then they are said to be concordant if \((X - X')(Y - Y') > 0\) holds true. They are said to be discordant when the reverse inequality is valid. Henceforth, we denote

\[
\mathbb{P}\{\text{concordance}\} = \mathbb{P}\{(X - X')(Y - Y') > 0\} \quad \text{and} \quad \mathbb{P}\{\text{discordance}\} = \mathbb{P}\{(X - X')(Y - Y') < 0\}.
\]

The idea of using the concordance and discordance probabilities comes from the fact that probabilities of events involving only inequality relationships between two random variables are invariant with respect to increasing transformations of these variables. Hence, defining dependence measures from these probabilities ensures that they will only depend on the underlying copula.
Having defined the notion of concordance and discordance, we are now ready to introduce Kendall's rank correlation coefficient: Kendall’s rank correlation coefficient for a random couple \((X, Y)\) is defined as

\[
\tau(X, Y) = \mathbb{P}\text{[concordance]} - \mathbb{P}\text{[discordance]}.
\]

If the marginals of \(X\) and \(Y\) are continuous with copula \(C\), then \(\tau\) can be rewritten as

\[
\tau(X, Y) = 4 \int_0^1 \int_0^1 C(u_1, u_2) \, dC(u_1, u_2) - 1 \tag{5}
\]

so that the value of Kendall’s rank correlation coefficient only depends on the copula for \((X, Y)\). In general, evaluating \(\tau\) requires the evaluation of a double integral. For an Archimedean copula, the situation is simpler in that \(\tau\) can be evaluated directly from the generator \(\phi\), as explained in equation (9) below.

We will now briefly state the definition of several Archimedean copulas used in this paper.

- **Clayton's copula** is given by

\[
C_\alpha(u, v) = (u^{-\alpha} + v^{-\alpha} - 1)^{-1/\alpha}, \quad \alpha > 0.
\]

It is the Archimedean copula associated with the generator

\[
\phi_\alpha(t) = t^{-\alpha} - 1, \quad \alpha > 0,
\]

with Kendall’s tau given by \(\tau = \alpha / (\alpha + 2)\).

- **Frank's copula** is given by

\[
C_\alpha(u, v) = -\frac{1}{\alpha} \ln \left(1 + \frac{(\exp(-\alpha u) - 1)(\exp(-\alpha v) - 1)}{\exp(-\alpha) - 1}\right), \quad \alpha \neq 0.
\]

It is the Archimedean copula generated by

\[
\phi_\alpha(t) = -\ln \left(\frac{\exp(-\alpha t) - 1}{\exp(-\alpha) - 1}\right), \quad \alpha \neq 0,
\]

with Kendall’s tau given by

\[
\tau = 1 + \frac{4}{\alpha} \left(\int_0^\alpha \frac{\xi}{\alpha(\exp(\xi) - 1)} \, d\xi - 1\right).
\]
• Gumbel-Hougaard's copula has the form

\[ C_\alpha(u, v) = \exp \left( \{- \ln u\}^{\alpha} + \{- \ln v\}^{\alpha} \right)^{1/\alpha}, \quad \alpha \geq 1. \]

It is the Archimedean copula associated with

\[ \phi_\alpha(t) = (-\ln(t))^{\alpha}, \quad \alpha \geq 1, \]

with Kendall's tau given by \( \tau = 1 - 1/\alpha. \)

• Joe's copula is given by

\[ C_\alpha(u, v) = 1 - \left( \bar{u}^\alpha + \bar{v}^\alpha - \bar{u}^\alpha \bar{v}^\alpha \right)^{1/\alpha}, \quad \alpha \geq 1, \]

where \( \bar{u} = 1 - u \) and \( \bar{v} = 1 - u. \) It is the Archimedean copula associated with

\[ \phi_\alpha(t) = -\ln \left( 1 - (1 - t)^\alpha \right), \quad \alpha \geq 1. \]

For this copula, there is no simple form to compute Kendall's tau.

3 Estimation of the Generator

Given \( K \) from equation (3), it is possible to recover \( \phi \) by solving the differential equation

\[ \frac{\phi(v)}{\phi'(v)} = v - K(v), \]

which yields

\[ \phi(v) = \exp \left\{ \int_{v_0}^{v} \frac{1}{\xi - K(\xi)} d\xi \right\} \]

where \( 0 < v_0 < 1 \) is an arbitrary constant. From equation (2), \( \phi \) is defined up to a positive factor. The function \( \phi \) defined in equation (6) generates an Archimedean copula whenever \( v - K(v) \) is negative and remains bounded away from 0 on the unit interval. Specifically, Genest and Rivest (1993) proved that the function \( \phi \) given in equation (6) is decreasing and convex and satisfies \( \phi(1) = 0 \) if, and only if,

\[ K(v-) = \lim_{t\to v^-} K(t) > v, \quad \text{for all } 0 < v < 1. \]
The condition in equation (7) has to be fulfilled by the estimator of $K$ in order to recover a proper generator from equation (6). More specifically, under the assumption that the dependence function associated with $K$ is Archimedean, a natural estimator $\hat{A}$ of $A$ can be derived from an estimator $\hat{K}$ of $K$ through the relation $\hat{A}(v) = v - \hat{K}(v)$, $0 < v < 1$. Provided $\hat{K}(v-) > v$ for all $0 < v < 1$, formula (6) then provides an estimator of $C_\phi$ within the class of Archimedean copulas.

Kendall's tau is given by

$$\tau = \tau(X, Y) = 4E[F(X, Y)] - 1,$$

which in the Archimedean case reduces to

$$\tau = 4 \int_0^1 \lambda(\xi)d\xi + 1 = 3 - 4 \int_0^1 K(\xi)d\xi. \quad (9)$$

As the estimation of $K$ takes into account the censoring mechanism, the estimated $\tau$ obtained from equation (9) is suitable for censored data.

The problem of estimating the generator was studied by Genest and Rivest (1993), who were the first to propose a procedure for identifying a generator in empirical applications with complete data. Given observations from a random pair $(X, Y)$ with joint distribution function $F$, this procedure relies on the estimation of the distribution function associated with the probability integral transformation $Z = F(X, Y)$. As pointed out by Genest and Rivest (1993), because the empirical estimate of the bivariate distribution function is always larger than $1/n$ and as the estimator takes values only on a $(0, 1)$ range, $K$ can be estimated as

$$\hat{K}_n(z) = \frac{1}{n} \# \{i \mid Z_i \leq z\} \quad \text{where}$$

$$Z_i = \frac{1}{n-1} \# \{(x(j), y(j)) \mid x(j) < x(i), y(j) < y(i)\}, \quad (11)$$

the symbol $\#$ stands for the cardinality of a set and $\{(x_i, y_i), i = 1, \ldots, n\}$ are the observed data.

The Genest-Rivest technique, however, is not appropriate for censored data. In the case of censored data, Wang and Wells (2000b) proposed a modified estimator of $K$. As $K$ can be written as

$$K(v) = P[F(X,Y) \leq v] = E[I\{F(X,Y) \leq v\}] \quad (12)$$
the suggested estimator is given by

$$\hat{K}_n(v) = \int_0^\infty \int_0^\infty I[\hat{F}(x, y) \leq v]d\hat{F}(x, y)$$  \hspace{1cm} (13)$$

where $\hat{F}$ stands for a nonparametric estimator of the joint distribution function $F$ taking censoring into account. As mentioned by Wang and Wells (2000b), this approach is sufficiently flexible to deal with various censorship mechanisms, as long as $\hat{F}$ is an appropriate estimator for $F$.

Several authors have proposed nonparametric estimators of a bivariate distribution, e.g., Dabrowska (1988), Prentice and Cai (1992), van der Laan (1996), and Prentice, Moodie, and Wu (2004). A widely used estimator of the bivariate survival function is the one developed by Dabrowska (1988). This estimator is a generalization of the univariate Kaplan-Meier estimator and is based on the product-integral of a suitably defined bivariate cumulative hazard function. The marginals used are univariate Kaplan-Meier estimates. However, as mentioned in Dabrowska (1988, Section 3), this estimator is not monotonic. The weak convergence of the estimator of the bivariate survival function is given in Dabrowska (1989).

When only one variable is subject to censoring, Akritas (1994) proposed a nonparametric estimator for the bivariate distribution. This estimator is an average (over the uncensored variable) of estimates of the conditional distribution function of the censored variable given the uncensored variable. The estimates of the conditional distribution function used are nearest neighbor estimators. Properties of the proposed estimator for the bivariate distribution, such as asymptotic optimality and weak convergence, are proved in Akritas (1994).

In order to use Akritas' (1994) estimator for random right censoring, we need first to justify the applicability of the techniques to the data at hand. Loss-ALAE data are subject to a generalized type I-censoring in the terminology of Klein and Moeschberger (1997, page 57). The censoring variable in this case is the policy limit, which is constant and varies from policy to policy. We now prove that this type of censoring leads to the same likelihood function as random right censoring up to a factor not depending on the unknown survival distribution. This will show that Akritas' (1994) estimator (defined in equation (14)) remains consistent when type I-censoring is present.

Let $(t_i, y_i, \delta_i), i = 1, \ldots, n$, denote the observed data set, $g$ be a known probability density function (a kernel function), and $(h_n)$ denote a sequence of positive constants such that $h_n \rightarrow 0$ as $n \rightarrow \infty$ (a bandwidth sequence). The conditional local likelihood of $X$ given $Y$ at the point $y$ is then given by
\[ L(y) = \prod_{i=1}^{n} g\left(\frac{y - y_i}{h_n}\right) \mathbb{P}[T_i = t_i, \Delta_i = \delta_i | Y_i = y_i] \]
\[ = \prod_{i=1}^{n} g\left(\frac{y - y_i}{h_n}\right) \mathbb{P}[X_i = t_i | Y_i = y_i]^{\delta_i} \mathbb{P}[X_i > t_i | Y_i = y_i]^{1-\delta_i} \]
\[ = \prod_{i|\delta_i=1}^{n} g\left(\frac{y - y_i}{h_n}\right) (F_{X|Y}(t_i | y_i) - F_{X|Y}(t_i - | y_i)) \]
\[ \times \prod_{i|\delta_i=0}^{n} g\left(\frac{y - y_i}{h_n}\right) (1 - F_{X|Y}(t_i | y_i)), \]

whereas the likelihood for randomly censored data contains an extra factor depending solely on the conditional censoring distribution. As this factor has no influence on the maximization problem, both likelihoods reach their maximum at the same distribution function, i.e., when \( F_{X|Y} \) equals the Beran estimator (defined in equation (15)).

The bivariate distribution function \( F \) can be written as:
\[
F(x, y) = \mathbb{P}[X \leq x, Y \leq y] = \int_{0}^{y} F_{X|Y}(x | z) dF_{Y}(z).
\]

The proposed estimator of \( F \) will be based on the estimate of the conditional distribution \( F_{X|Y}(x | y) = \mathbb{P}[X \leq x | Y = y] \), i.e.,
\[
\hat{F}(x, y) = \int_{0}^{y} \hat{F}_{X|Y}(x | z) d\hat{F}_{Y}(z)
\]
\[
= \frac{1}{n} \sum_{k=1}^{n} I[0 \leq y_k \leq y] \hat{F}_{X|Y}(x | y_k), \tag{14}
\]

where
\[
\hat{F}_{X|Y}(x | z) = 1 - \prod_{i|T_i \leq z \cap \delta_i=1} \left(1 - \frac{W_{ni}(z; h_n)}{\sum_{j=1}^{n} W_{nj}(z; h_n) I[t_j \geq t_i]} \right) \tag{15}
\]
is the Beran (1981) estimator and
\[
W_{ni}(z; h_n) = \frac{g\left(\frac{z-y_i}{h_n}\right)}{\sum_{j=1}^{n} g\left(\frac{z-y_j}{h_n}\right)} \tag{16}
\]
4 Application to Loss-ALAE Modeling

We use data collected by the Insurance Services Office. The data comprise 1,500 (sample size of $n = 1,500$) general liability claims randomly chosen from late settlement lags. Each claim is accompanied by a policy limit $\ell$ (the maximum claim amount covered) that is specific to each contract. Therefore, the loss variable is censored when the claim amount exceeds the policy limit. More precisely, one observes a triple $(T_i, Y_i, \Delta_i)$, where $T_i = \min(X_i, -\ell_i)$, $X_i$ is the $i^{th}$ loss, and $Y_i$ is the associated ALAE, $i = 1, \ldots, n$, and

$$
\Delta_i = I[T_i = -\ell_i] = \begin{cases} 1, & \text{if } X_i \leq -\ell_i \text{ (uncensored claim)} \\ 0, & \text{if } X_i > -\ell_i \text{ (censored claim)} \end{cases}
$$

(17)

where $I[A]$ denotes the indicator of the occurrence of the event $A$. Some summary statistics of the data are gathered in Table 1. There are 34 censored data points, and they have a much higher mean than the 1,466 complete data ($\$217,941$ versus $\$37,110$). A scatterplot of (loss, ALAE) on a log scale is depicted in Figure 1. Its shape suggests some positive relationship between loss and ALAE: large losses tend to be associated with large ALAEs, as expected. Moreover, censored data points (represented by triangles in Figure 1) clearly cluster to the right.

We will now derive a nonparametric estimate of the generator, $\hat{K}_n$, then compare it to several parametric analogs $K_\alpha$ corresponding, for instance to the Clayton, Gumbel, Frank, or Joe copulas, in order to select the best parametric model.

4.1 Nonparametric Estimation of the Generator

The kernel density function used in equation (16) is the biweight kernel, i.e.,

$$
g(u) = \frac{15}{16} (1 - u^2)^2 I\{|u| \leq 1\}.
$$

Other kernel functions, such as the Epanechnikov kernel, the uniform kernel or the Gaussian kernel, can be used as well and yield very similar

---

4 We thank Professors Edward Frees and Emiliano Valdez for providing access to the loss-ALAE data that were originally collected by the Insurance Services Office (ISO), New Jersey, USA.

5 The library "bivsurv" of Statlib (available from lib.stat.cmu.edu) has been used for the numerical illustrations. This library contains functions for nonparametric survival curve analysis (the Unix version has been contributed by Ronald Pruitt).
Table 1
Summary Statistics for Loss and ALAE Data Set

<table>
<thead>
<tr>
<th></th>
<th>All Losses</th>
<th>Uncensored Losses</th>
<th>Censored Losses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>1,500</td>
<td>1,500</td>
<td>1,466</td>
</tr>
<tr>
<td>Minimum</td>
<td>10</td>
<td>15</td>
<td>10</td>
</tr>
<tr>
<td>1st Quartile</td>
<td>4,000</td>
<td>2,333</td>
<td>3,750</td>
</tr>
<tr>
<td>Mean</td>
<td>41,208</td>
<td>12,588</td>
<td>37,110</td>
</tr>
<tr>
<td>Median</td>
<td>12,000</td>
<td>5,471</td>
<td>11,049</td>
</tr>
<tr>
<td>3rd Quartile</td>
<td>35,000</td>
<td>12,577</td>
<td>32,000</td>
</tr>
<tr>
<td>Maximum</td>
<td>2,173,595</td>
<td>501,863</td>
<td>2,173,595</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>102,748</td>
<td>28,146</td>
<td>92,513</td>
</tr>
</tbody>
</table>

results. One important step in estimating the joint distribution function of loss and ALAE is the selection of the bandwidth appearing in the estimation of the conditional distribution of $X$ given $Y$. We choose the bandwidth such that it minimizes the average mean squared error (AMSE in short) of the empirical estimate $\hat{K}_n$ of the distribution $K$. As the AMSE has a complicated structure and depends on a number of unknown quantities, it will be used by means of a bootstrap procedure. The procedure is based on Van Keilegom and Veraverbeke (1997). Let us describe this procedure performed for a fixed value of $h_n$.

First generate two uniform random numbers on $[0, 1]$, $u^*$ and $v^*$. Then we construct the uncensored bootstrap data $Y_t^*$ from the empirical distribution of ALAE, i.e., $Y_t^* = \hat{F}_Y^{-1}(v^*)$, and the censored bootstrap data $X_t^*$ from the conditional distribution, i.e., $X_t^* = \hat{F}_{X|Y}^{-1}(u^*|Y_t^*)$. As $Y_t^*$ equals the value of a certain $Y_j$ from the original data, we will take as the censoring bootstrap data the policy limit associated with $Y_j$, i.e., $\ell_t^* = \ell_j$. With the bootstrapped data $(T_t^*, Y_t^*) = (\min(X_t^*, \ell_t^*), Y_t^*)$ and the indicator $\Delta_t^* = I[T_t^* = X_t^*]$, $i = 1, \ldots, n$, we then estimate the distribution $K$ and compute the AMSE. Specifically, consider a grid $\nu = (\nu_1, \ldots, \nu_m)$ on the unit interval $[0, 1]$ and let $\hat{K}_n^*(\nu) = \{\hat{K}_{n,b}^*(\nu)\}_{b=1}^B$ denote the $B \times m$ matrix of the empirical estimates of $K$ given by equation (13) for the $B$ resamples computed on this grid.

For each $\nu_l$ of the grid, $l = 1, \ldots, m$, we estimate the bias, the variance, and the MSE as follows:
Figure 1: Scatterplot for log-Loss and log-ALAE

\[ \text{Bias}_h[\hat{K}_n^*(v_l)] = \frac{1}{B} \sum_{b=1}^{B} \hat{K}_{n,b}^*(v_l) - \hat{K}_n(v_l) \]

\[ \text{Var}_h[\hat{K}_n^*(v_l)] = \frac{1}{B} \sum_{b=1}^{B} [\hat{K}_{n,b}^*(v_l)]^2 - \left( \frac{1}{B} \sum_{b=1}^{B} \hat{K}_{n,b}^*(v_l) \right)^2 \]

\[ \text{MSE}_h[\hat{K}_n^*(v_l)] = \text{Var}_h[\hat{K}_n^*(v_l)] + \left( \text{Bias}_h[\hat{K}_n^*(v_l)] \right)^2, \]

for \( l = 1, \ldots, m \). The optimal bandwidth will be then the one that minimizes
\[ \text{AMSE}_n = \frac{1}{m} \sum_{t=1}^{m} \text{MSE}_h[\hat{K}_n^*(\nu_t)]. \]

The validity of this bootstrap procedure has been established in Van Keilegom and Veraverbeke (1997) for the Beran estimator. Starting from this result, the validity of the bootstrap for the estimator of \( K \) can be derived. The results based on \( B = 500 \) resamples are plotted in Figure 2, which shows the optimal value is \( h_n \approx 0.4 \).

![Figure 2: AMSE of \( \hat{K}_n \) vs. Bandwidth \( h_n \)](image)

The estimation of \( K \) then follows from equation (13), and the resulting \( \hat{K}_n \) is depicted in Figure 3. The generator of the Archimedean...
copula is then obtained by plugging $\hat{K}_n$ into equation (6). The estimate of the generator for the loss-ALAE data is depicted in Figure 3.

It is interesting to compare the estimates of the distribution $K$ when the bivariate distribution function used is either Dabrowska's (1988) estimator or Genest and Rivest's (1993) estimator in the uncensored case (i.e., we ignore the censored loss variables and work only with 1,466 observations over the initial 1,500 data points). These estimates are depicted in Figure 4, together with $\hat{K}_n$ of Figure 3. We see that the three curves are close to each other. This may be explained by the limited amount of censored points present in the data set.

Figure 4 demonstrates that the estimated functions $K$ never intersect each other: the uncensored $\hat{K}_n$ dominates the Dabrowska $\hat{K}_n$, which in turn dominates the Akritas $\hat{K}_n$. From Caperaa, Fougeres, and Genest (1997b), the domination of the uncensored estimator suggests that neglecting censorship when it is present in the data or failing to taking into account the particular form of censorship tends to underestimate the strength of the positive quadrant dependence in the data. Thus, neglecting censorship may be a dangerous strategy for actuaries.

4.2 Graphical Model Selection for Generators

We can compare the empirical estimator with several parametric analogs $K_\alpha$ corresponding, for instance, to Clayton, Gumbel, Frank, or Joe copulas, in order to select the best parametric model. The selection criterion is the minimization of the $L^2$-norm distance:

$$S(\alpha) = \int (K_\alpha(\xi) - \hat{K}_n(\xi))^2 d\xi$$

where $\alpha$ is the dependence parameter. Specifically, $\alpha$ is estimated for different parametric models using the omnibus procedure described below, and $S(\alpha)$ is then computed with the estimated $\alpha$ taking for $\hat{K}_n$ the Akritas estimator. The optimal parametric model is the one minimizing $S(\alpha)$ over the alternatives considered.

Note that the chosen statistic $S(\alpha)$ is of the integral type, and thus it considers the whole range of the data—it does not specifically focus on the tails to test the goodness of the fit. Other statistics of Kolmogorov-Smirnov type are considered in Genest, Quessy, and Remillard (2006). In addition to computing $S(\alpha)$, a comparison of the graph of $K$ and $\lambda$ for the parametric models and the nonparametric benchmarks is often helpful. Two procedures for estimating $\alpha$ are provided: the Wang and Wells estimation procedure and the omnibus estimation procedure.
Figure 3: \( \hat{K}_n \) (top) and \( \hat{\phi} \) (bottom) for Loss-ALAE Data
The Wang and Wells estimation procedure requires an initial value for $\alpha$, which can easily be obtained using the method of moments and the one-to-one relationship between the population version of Kendall’s tau and $\alpha$ given in equation (9). Nonparametric estimation of Kendall’s tau under censoring is a complex problem. One estimation of Kendall’s tau can be obtained from (9), with $K$ replaced by its empirical estimator $\hat{K}_n$, given in equation (13). For loss-ALAE data, we get $\hat{\tau} = 0.3669$.

Another way to estimate Kendall’s tau is to compute it directly from the data, by ignoring the Archimedean assumption. Wang and Wells (2000a) showed that if the largest observations of each of the two variables are uncensored, then a consistent estimate of $\tau$ is given by:

$$\hat{\tau} = 4 \sum_{i=1}^{n} \sum_{j=1}^{n} \hat{F}(x_{(i)}, y_{(j)})\hat{F}(\Delta x_{(i)}, \Delta y_{(j)}) - 1$$  \hspace{1cm} (18)
where \( x_{(i)} \) and \( y_{(j)} \) are the \( i^{th} \) and \( j^{th} \) ordered observation of the \( X_i \)'s and the \( Y_i \)'s, respectively, \( \hat{F} \) is the Dabrowska estimator of the bivariate distribution function, and

\[
\hat{F}(\Delta x_{(i)}, \Delta y_{(j)}) = \hat{F}(x_{(i)}, y_{(j)}) - \hat{F}(x_{(i-1)}, y_{(j)}) - \hat{F}(x_{(i)}, y_{(j-1)}) + \hat{F}(x_{(i-1)}, y_{(j-1)}).
\]

The approach of Wang and Wells (2000a) applied to the loss-ALAE data yields \( \hat{\tau} = 0.3567 \), a value that is close to \( \hat{\tau} \).

The omnibus estimation procedure is a omnibus semiparametric procedure, which is known also as the maximum pseudo-likelihood procedure. It treats marginal distributions as (infinite dimensional) nuisance parameters. This procedure substitutes empirical analogs for the marginal distribution functions in the likelihood for the dependence parameters and then in maximizing the resulting pseudo-likelihood. As shown by Genest, Ghoudi, and Rivest (1995) the resulting estimator is consistent and asymptotically normal, even in the presence of censorship.

The first step consists of estimating the two marginals nonparametrically, by rescaled versions of the Kaplan-Meier estimator (for loss variable) and the empirical estimator (for ALAE variable). As explained in Genest, Ghoudi, and Rivest (1995), these rescaled versions are \( n/(n+1) \) times the empirical distributions and are taken to avoid difficulties due to the potential unboundedness of \( \log(c_\alpha(u, v)) \) as \( u \) or \( v \) tends to one. These two marginal estimators, \( \hat{F}_X \) and \( \hat{F}_Y \), are used in the second step to estimate the dependence parameter.

As only the loss variable is censored, the likelihood function can be written as follows:

\[
L(\alpha) = \prod_{i=1}^{n} c_\alpha(u_i, v_i)^{\delta_i} \left( 1 - \frac{\partial c_\alpha(u_i, v_i)}{\partial v} \right)^{1-\delta_i}
\]

where \( (u_i, v_i) = (\hat{F}_X(t_i), \hat{F}_Y(y_i)) \), \( C_\alpha \) is the Archimedean copula under consideration, and \( c_\alpha \) is its density. The log-likelihood will therefore be given by:

\[
\ln L(\alpha) = \sum_{i=1}^{n} \left( \delta_i \ln(c_\alpha(u_i, v_i)) + (1 - \delta_i) \ln \left( 1 - \frac{\partial c_\alpha(u_i, v_i)}{\partial v} \right) \right).
\]

The derivatives appearing in the expression of the likelihood for the four parametric models considered are given in Table 2. The omnibus estimator for \( \alpha \) maximizes \( \ln L(\alpha) \).
### Table 2

Partial Derivatives of \( C_\alpha \) with \( \hat{u} = -\ln u \) and \( \hat{v} = 1 - u \)

<table>
<thead>
<tr>
<th>Copula</th>
<th>( \frac{\partial C_\alpha(u,v)}{\partial v} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clayton:</td>
<td>( [1 + v^\alpha(u^{-\alpha} - 1)]^{-1 - 1/\alpha} )</td>
</tr>
<tr>
<td>Frank:</td>
<td>( \left[ e^{-\alpha v} - e^{-\alpha(u+v)} \right] \times \left[ (1 - e^{-\alpha}) - (1 - e^{-\alpha u})(1 - e^{-\alpha v}) \right]^{-1} )</td>
</tr>
<tr>
<td>Gumbel-Hougaard:</td>
<td>( v^{-1} \exp\left{-(\hat{u}^\alpha + \hat{v}^\alpha) \right} \left[1 + \left(\frac{\hat{u}}{v}\right)^\alpha \right]^{-1 + 1/\alpha} )</td>
</tr>
<tr>
<td>Joe:</td>
<td>( (1 - \hat{u}^\alpha)(1 - \hat{u}^\alpha + \hat{v}^\alpha \hat{v}^{-\alpha})^{-1 + 1/\alpha} )</td>
</tr>
</tbody>
</table>

#### 4.3 Graphical Representations

We will now identify the appropriate Archimedean copula. Note that all four parametric models considered allow an upper bound approaching 1 for Kendall's tau, which is not the case for other Archimedean copulas (for instance for Ali-Maikhail-Haq family, \( \tau < \frac{1}{3} \)). Table 3 shows the method of moment estimations (associated with the two estimates of Kendall's tau, \( \hat{\tau} \) and \( \hat{\tau} \)) and the omnibus estimations of the dependence parameters. Except for the Frank and Gumbel-Hougaard copulas, for which the three values are quite close, the estimates are different for Joe and Clayton copulas, indicating that these two models might not be appropriate for the data.

<table>
<thead>
<tr>
<th>Method of Moments and Omnibus Estimates of ( \alpha )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copula</td>
</tr>
<tr>
<td>Clayton</td>
</tr>
<tr>
<td>Frank</td>
</tr>
<tr>
<td>Gumbel</td>
</tr>
<tr>
<td>Joe</td>
</tr>
</tbody>
</table>

Figure 5 shows the nonparametric and the four parametric estimates \( \lambda_{\hat{\alpha}} \) (where \( \hat{\alpha} \)'s are the omnibus estimates) of the function \( \lambda \), as well as the nonparametric estimate \( \hat{\lambda}_n \), suggests that the closest parametric models are the Frank and Gumbel-Hougaard models.

A look at the QQ-plot of the nonparametric and parametric quantiles of the distribution \( K \) depicted in Figure 6 confirms our previous conclusion (although for the Frank copula there is a great disparity for the
higher quantiles, corresponding to high losses and expenses). In order to choose the best model between the parametric models considered, we compute the distance $S(\hat{\alpha})$ for each model, where the estimated values of the dependence parameters are the omnibus estimates given in Table 3. The results are summarized in Table 4. It follows that the Gumbel-Hougaard model provides the best fit to the data, even though it is quite close to Frank model (looking at the QQ-plot in Figure 6, the Frank copula seems to perform as well as the Gumbel-Hougaard one, except possibly in the tails of the distribution of the copula).

It may also be interesting to have a closer look at the significant departures of the parametric model from the nonparametric benchmark. Here, we follow the approach suggested by Vandenhende and Lambert (2005), who computed nonparametric confidence bands for the function $-\lambda_n(t) = \hat{K}_n(t) - t$, using selected quantiles obtained from a bootstrap procedure. We compute confidence intervals on the $K$ and $\lambda$ functions at each point. The bootstrap procedure described in Sec-
tion 4.1 can be used to get 95% confidence intervals, by selecting the 2.5% and 97.5% quantiles in the series $\hat{K}_{n,1}^*, \ldots, \hat{K}_{n,B}^*$ and $\hat{\lambda}_{n,1}^*, \ldots, \hat{\lambda}_{n,B}^*$, respectively. We refer the reader to Genest and Remillard (2005) for more details about the use of this bootstrap procedure, as well as for theoretical justifications.

Results obtained with $B = 1000$ resamples are depicted in Figure 7, together with the parametric alternatives. The Clayton and Joe copulas significantly depart from the nonparametric estimation. The Frank and Gumbel models show good adherence to the nonparametric benchmark. This confirms the conclusions drawn from the inspection of the QQ-plots displayed in Figure 6.
Figure 7: Nonparametric and Parametric Estimates of $K$ (left) and $\lambda$ (right) Using Gumbel (top), Clayton, Frank, and Joe (bottom) Copulas, Respectively
Table 4

<table>
<thead>
<tr>
<th>Copula</th>
<th>( S(\hat{\alpha}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clayton</td>
<td>0.00245393</td>
</tr>
<tr>
<td>Frank</td>
<td>0.00028323</td>
</tr>
<tr>
<td>Gumbel-Hougaard</td>
<td>0.00025499</td>
</tr>
<tr>
<td>Joe</td>
<td>0.00099576</td>
</tr>
</tbody>
</table>

5 Summary and Conclusion

In this paper, we proposed a semiparametric modeling strategy for bivariate outcomes commonly encountered in nonlife actuarial practice. We develop an appropriate nonparametric estimator of the joint distribution of loss-ALAE taking into account the particular censorship present in the data. This estimate is then used to identify the appropriate Archimedean copula that fits the data. A selection procedure for the generator of the underlying Archimedean copula was also described.

Even if the choice of the Archimedean copula for the particular data set that we analyzed has not been modified [Gumbel copula ranked first on the basis of the integrated square difference, with and without taking censorship into account as in Frees and Valdez (1998)], we believe that the procedure we proposed should be applied in practice. The proportion of censored data was indeed rather low (34 out of 1,500 data points) and it can be expected that neglecting the censorship may lead to an incorrect choice of the Archimedean copula.

As mentioned in the introduction, there are other approaches to modeling multivariate data in nonlife insurance. Multivariate extreme value theory can also be considered. As nonlife insurance data are often heavy tailed, this approach has some intuitive appeal (even if the componentwise maxima are not in line with loss-ALAE data, where large losses are of interest in reinsurance, whatever the size of ALAEs). Mikosch (2005) contrasted the approach based on multivariate extreme value theory with copulas. In the authors' opinion, no single approach systematically outperforms the others, so actuaries are urged to con-

---

6Specifically, if the actuary agrees to restrict the modeling to Archimedean copulas, then the generator should first be estimated in a nonparametric way, taking a possible censorship into account, to serve as a benchmark when selecting the optimal parametric family.
sider other models, such as extreme value theory, for the data to be analyzed.

Extreme value copulas have been used by Cebrian, Denuit, and Lambert (2003) in a similar context. Such copulas are of the form

\[ C(u_1, u_2) = \exp \left( \left( \log(u_1) + \log(u_2) \right) A \left( \frac{\log(u_2)}{\log(u_1) + \log(u_2)} \right) \right) \]

for \( 0 \leq u_1, u_2 \leq 1 \), where

\[ A(w) = \int_0^1 \max \{ (1 - w)q, w(1 - q) \} dL(q) \]

for some positive finite measure \( L \) on \([0, 1]\). The function \( A \) is called the dependence function. The dependence function \( A \) must satisfy the following properties: \( A(0) = A(1) = 1 \), \( \max(w, 1 - w) \leq A(w) \leq 1 \) for \( 0 \leq w \leq 1 \) and \( A(w) \) is a convex function in the region \( 0 \leq w \leq 1 \). Moreover, if \( A(w) = 1 \), then \((X, Y)\) are independent. If \( A(w) = \max(w, 1 - w) \), then \((X, Y)\) are perfectly dependent (or comonotonic). The family of extreme value copulas includes, e.g., Gumbel copula (the only one that belongs to both Archimedean and extreme value families), the logistic copula, the asymmetric logistic copula, and the mixed copula. The function \( A \) could be estimated nonparametrically, exactly as we did in this paper for the generator of Archimedean copulas. A reference for complete data is Caperaa, Fougeres, and Genest (1997b).

In this paper, we found that the best-fitting Archimedean copula is identified by estimating the generator in a nonparametric way. The Archimedean assumption can then be tested on the basis of some distance between \( C_\phi \) and a fully nonparametric estimation of the underlying copula \( C \) (a Kolmogorov-Smirnov or an integrated distance, for example). In this case also, a bootstrap procedure can be used to compute the \( p \)-values.

Finally, we note that computing the pure premium relating to reinsurance treaty described in Section 1 requires only knowledge of the conditional expectation of loss given ALAE. Therefore, regressions can be conducted to obtain the pure premiums. There are, however, many applications where the knowledge of the joint distribution is needed (such as for computing safety loading).
References


Denuit, Purcaru, and Van Keilegom: Archimedean Copulas


