I N S T I T U T D E
S T A T I S T I Q U E

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N
P A P E R

0423

BIVARIATE ARCHIMEDEAN COPULA MODELLING FOR LOSS-ALAE DATA IN NON-LIFE INSURANCE

M. DENUIT, O. PURCARU and I. VAN KEILEGOM

http://www.stat.ucl.ac.be

BIVARIATE ARCHIMEDEAN COPULA MODELLING FOR LOSS-ALAE DATA IN NON-LIFE INSURANCE

Michel Denuit[⋆,†], Oana Purcaru[⋆] & Ingrid Van Keilegom[⋆]
Institut de Statistique[⋆] & Institut des Sciences Actuarielles[†]
Université Catholique de Louvain
B-1348 Louvain-la-Neuve, Belgium

September 10, 2004

Abstract

This paper considers the bivariate loss-ALAE modelling problem in actuarial science, taking into account the particular form of the censorship present in the data. Specifically, a graphical selection procedure for the generator of the underlying archimedean copula is proposed, based on a nonparametric estimation of the generator. The approach is in line with GENEST & RIVEST (1993) who considered complete data and WANG & WELLS (2000b) who treated doubly censored data. A loss-ALAE data set provided by the US Insurance Services Office is used for the numerical illustrations, and comparisons with previous results appeared in the actuarial literature are performed.

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

1 Introduction and Motivation

1.1 Losses and their associated ALAE's

Various processes in casualty insurance involve correlated pairs of variables. A prominent example is the loss and allocated loss adjustment expenses (ALAE's, in short) on a single claim. Here ALAE's are type of insurance company expenses that are specifically attributable to the settlement of individual claims such as lawyers' fees and claims investigation expenses.

Expensive claims generally need some time to be settled and induce considerable costs for the insurance company. Actuaries therefore expect some positive dependence between losses and their associated ALAE's, i.e. large values for losses tend to be associated with large values for ALAE's.

This positive association has some practical implications in the pricing of reinsurance treaties. The reinsurer covers the largest losses (i.e. those exceeding some high threshold called the retention of the direct insurer, and pays that part exceeding this threshold). It also contributes to pay the associated settlement costs on a prorata basis. In many cases, neglecting the dependence exhibited by the data leads to serious underestimation of the expected reinsurer's payment. It is therefore crucial for the reinsurer to have an appropriate model for the random couple loss-ALAE at its diposal.

Typically, a given amount of loss is divided between the insurer and the reinsurer as follows. The insurance company pays the loss from ground up to a specified amount r called the insurer's retention. Then, the reinsurer covers the claim from r, but with a maximum limit equal to ω. The excess over the limit ω 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). If we denote the loss as X and the associated ALAE as Y, assuming a prorata sharing of expenses, the reinsurer's payment for a given realization (X, Y) of loss and associated ALAE is described by the function:

$$g(X,Y) = \begin{cases} 0, & \text{if } X < r, \\ \\ X - r + \frac{X-r}{X}Y, & \text{if } r \leq X < \omega, \\ \\ \omega - r + \frac{\omega - r}{\omega}Y, & \text{if } X \geq \omega. \end{cases}$$

The net premium of such a 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 possible dependence between losses and ALAE's has therefore to be accounted for when reinsurance treaties are priced.

1.2 Presentation of the ISO data set

The data used in the present paper were collected by the US Insurance Services Office, and comprise general liability claims randomly chosen from late settlement lags. The data consist in $n = 1,500$ observations, each accompanied by a policy limit ℓ (that is, the maximal claim amount) specific to each contract. Therefore the loss variable will be censored when the amount of the claim exceeds the policy limit. More precisely, one observes a triple (T_i, Y_i, Δ_i), where $T_i = \min(X_i, \ell_i)$, X_i is the ith loss and Y_i the associated ALAE, $i = 1, \ldots, n$, and

$$\Delta_i = \mathbb{I}[T_i = \ell_i] = \begin{cases} 1, & \text{if} \quad X_i \leq \ell_i \quad (\text{ uncensored claim}) \\ 0, & \text{if} \quad X_i > \ell_i \quad (\text{ censored claim}). \end{cases} \tag{1.1}$$

Some summary statistics of the data are gathered in Table 1.1. It appears clearly that, even if the vast majority of losses are uncensored (1,466 among the 1,500 observations), the 34 censored data points have a much higher mean than the 1,466 complete data (217,941 US$ versus 37,110 US$). A scatterplot of (loss, ALAE) on the log scales is depicted in Figure 1.1. Its shape suggests some positive relationship between loss and ALAE: large losses tend to be associated with large ALAE's, as expected. Moreover, censored data points (represented by triangles in Figure 1.1) clearly cluster to the right.

	loss	ALAE	loss (uncensored)	loss (censored)
Total N	1,500	1,500	1,466	34
Min	10 US$	15 US$	10 US$	5,000 US$
1st Qu.	4,000 US$	2,333 US$	3,750 US$	50,000 US$
Mean	41,208 US$	12,588 US$	37,110 US$	217,941 US$
Median	12,000 US$	5,471 US$	11,049 US$	100,000 US$
3rd Qu.	35,000 US$	12,577 US$	32,000 US$	300,000 US$
Max	2,173,595 US$	501,863 US$	2,173,595 US$	1,000,000 US$
Std Dev.	102,748	28,146	92,513	258,205

Table 1.1: *Summary statistics for losses and ALAE's.*

1.3 Modelling loss-ALAE data with archimedean copulas

The copula construction turns out to be very useful for the analysis of dependence in actuarial science. Applications of copulas to insurance data modelling have been proposed e.g. by Carrière (2000), Frees, Carrière & Valdez (1996), Frees & Valdez (1998), Klugman & Parsa (1999), Valdez (2001) and Embrechts et al. (2002).

A lot of recent 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 prescription. As illustrated by Genest & Mc Kay (1986a,b), this class of copulas is wide and analytically tractable and its elements have stochastic properties that make them attractive for the statistical treatment of data. The joint modelling in parametric settings of loss-ALAE data has been examined by Frees & Valdez (1998) (Pareto marginals and Gumbel copula) and Klugman & Parsa (1999) (inverse paralogistic for loss, inverse Burr for ALAE and Frank copula).

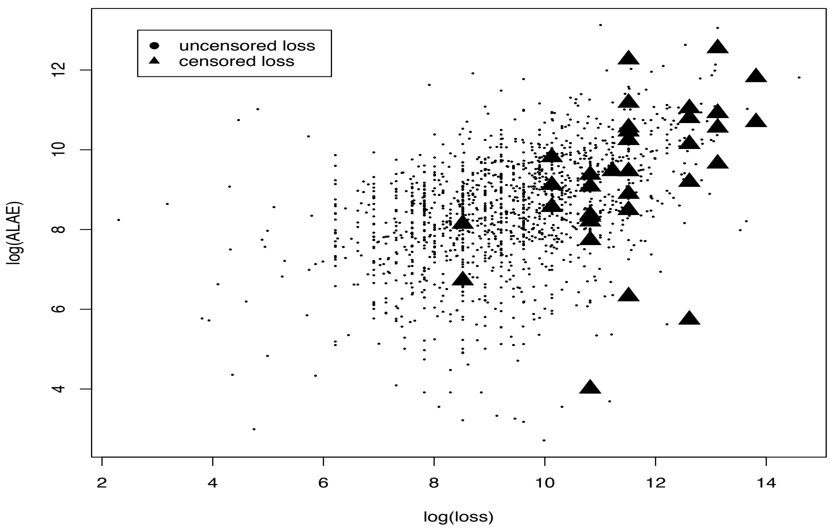

Figure 1.1: *Scatterplot for loss and ALAE (on the log-scale).*

1.4 Aim of the 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, like the archimedean one. Then, it is extremely 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 work), GENEST & 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 whose probability integral transformation distribution is the closest to its empirical estimate.

WANG & WELLS (2000b) extended the idea of GENEST & RIVEST (1993) to right-censored bivariate failure-time data. This kind of censorship is not the one encountered in actuarial problems but, as pointed out by WANG & WELLS (2000b), because the censoring issue is handled in the stage of estimating the bivariate distribution function, the approach they propose is flexible enough to deal with other censoring mechanisms. This is precisely the route we follow in this paper to deal with the modelling of loss-ALAE.

In FREES & VALDEZ (1998), techniques developed by GENEST & RIVEST (1993) for complete data have been applied to loss-ALAE data in order to select the appropriate generator. As pointed out by FREES & VALDEZ (1998) in their 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 & WELLS (2000b), but instead of using DABROWSKA (1988) estimator for the bivariate distribution, we use the estimator proposed in AKRITAS (1994), since only the loss variable is subject to censoring.

1.5 Agenda

In Section 2, we briefly present the notion of copula and give 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 ALAE's are completely observable. This nonparametric estimation then serves as a benchmark to select an appropriate parametric archimedean copula. The paper ends with numerical illustrations, given in Section 4.

2 Archimedean copulas

2.1 Sklar's theorem

The word "copula" was first employed in a statistical sense by SKLAR (1959) in the theorem which now bears his name. His idea was to separate a joint distribution function into a part which describes the dependence structure (the copula) and parts which 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:

$$F(x,y) = C\left(F_X(x), F_Y(y)\right), \quad (x,y) \in \mathbb{R}^2. \tag{2.1}$$

When the marginals F_X and F_Y are continuous, then the copula C in (2.1) is unique. Otherwise C is uniquely determined on Range$(F_X)\times$Range(F_Y). Conversely, if C is a copula and F_X and F_Y are distribution functions then the function F defined by (2.1) is a bivariate distribution function with margins F_X and F_Y. Formal proofs can be found e.g. in NELSEN (1999).

2.2 Archimedean family

2.2.1 Definition

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

$$C_\phi(u,v) = \begin{cases} \phi^{-1}\left\{\phi(u)+\phi(v)\right\} & \text{if } \phi(u)+\phi(v) \le \phi(0), \\ 0 \text{ otherwise,} \end{cases} \tag{2.2}$$

for $0 \le u, v \le 1$. Copulas C_ϕ of the form (2.2) are referred to as archimedean copulas. The function ϕ is called the generator of the copula. Only ϕ functions satisfying $\lim_{t\to 0+} \phi(t) = +\infty$ are used in this work. This ensures that C_ϕ is absolutely continuous.

Now, a bivariate distribution function F with marginals F_1 and F_2 is said to be generated by an archimedean copula if, and only if (2.1) holds with an archimedean copula C_ϕ.

2.2.2 Bivariate probability integral transformation theorem

It is well-known 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 is known as the Probability Integral Tranformation (PIT) theorem and underlies many statistical procedures. In particular, the copula C for (X, Y) is just the joint distribution function for the random couple $\big(F_X(X), F_Y(Y)\big)$ provided F_X and F_Y are continuous.

Now, 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 since 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.

Genest & Rivest (1993) studied the bivariate PIT for archimedean copulas. They obtained the following result. Let (U, V) be a random couple with unit uniform marginals and joint distribution function C_ϕ. The distribution function K of $Z = C_\phi(U, V)$ is given by $K(z) = z - \lambda(z)$ where

$$\lambda(\xi) = \frac{\phi(\xi)}{\phi^{(1)}(\xi)}, \;\; 0 < \xi \le 1.$$

2.2.3 Some members of the archimedean family

In this section, we briefly recall the definition of several archimedean copulas that will be used in this paper.

Clayton copula Clayton copula is given by

$$C_\alpha(u, v) = \left(u^{-\alpha} + v^{-\alpha} - 1\right)^{-1/\alpha}, \;\; \alpha > 0.$$

It is the archimedean copula associated to the generator

$$\phi_\alpha(t) = t^{-\alpha} - 1, \qquad \alpha > 0,$$

and possesses Kendall's tau $\tau = \frac{\alpha}{\alpha+2}$.

Frank copula Frank's copula is given by

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

It is the archimedean copula generated by

$$\phi_\alpha(t) = -\ln\left(\frac{\exp(-\alpha t) - 1}{\exp(-\alpha) - 1}\right), \qquad \alpha \neq 0,$$

and possesses Kendall's tau

$$\tau = 1 + \frac{4}{\alpha} \cdot \left(\int_0^\alpha \frac{\xi}{\alpha(\exp(\xi) - 1)} d\xi - 1 \right).$$

Gumbel-Hougaard copula Gumbel copula has the form

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

It is the archimedean copula associated with

$$\phi_\alpha(t) = \Big(- \ln(t) \Big)^\alpha, \ \alpha \geq 1.$$

and possesses Kendall's tau $\tau = 1 - \frac{1}{\alpha}$.

Joe copula Joe copula is given by

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

where $\bar{u} = 1 - u$ and $\bar{v} = 1 - u$. It is the archimedean copula associated with

$$\phi_\alpha(t) = - \ln\Big(1 - (1 - t)^\alpha \Big), \qquad \alpha \geq 1.$$

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

3 Nonparametric estimation of the generator in presence of censored data

3.1 General principle

Given K, it is possible to recover ϕ by solving the differential equation $\frac{\phi(v)}{\phi^{(1)}(v)} = v - K(v)$ which yields

$$\phi(v) = \exp\left\{ \int_{\xi = v_0}^{v} \frac{1}{\xi - K(\xi)} d\xi \right\} \tag{3.1}$$

where $0 < v_0 < 1$ is an arbitrary choosen constant (coming back to (2.2), it is easily seen that ϕ is defined up to a positive factor). The function ϕ defined in (3.1) generates an archimedean copula whenever $v - K(v)$ is negative and remains bounded away from 0 on the unit interval. Specifically, Genest & Rivest (1993) proved that the function ϕ given in (3.1) is decreasing and convex and satisfies $\phi(1) = 0$ if, and only if,

$$K(v-) = \lim_{t \to v-} K(v) > v, \qquad \text{for all } 0 < v < 1. \tag{3.2}$$

The condition (3.2) has to be fulfilled by the estimator of K in order to recover a proper generator from (3.1). More specifically, under the assumption that the dependence function associated with K is archimedean, a natural estimator $\widehat{\lambda}$ of λ can be derived from an estimator $\widehat{K}$ of K through the relation $\widehat{\lambda}(v) = v - \widehat{K}(v)$, $0 < v < 1$. Provided $\widehat{K}(v-) > v$ for all $0 < v < 1$, formula (3.1) then provides an estimator of C_ϕ within the class of archimedean copulas.

3.2 Estimating Kendall's tau

Kendall's tau is given by

$$\tau = \tau(X,Y) = 4\mathbb{E}[F(X,Y)] - 1 \tag{3.3}$$

which in the archimedean case reduces to

$$\tau = 4\int_0^1 \lambda(\xi)d\xi + 1 = 3 - 4\int_0^1 K(\xi)d\xi. \tag{3.4}$$

Since the estimation of K takes into account the censoring mechanism, the estimated τ obtained from (3.4) is suitable for censored data.

3.3 Genest-Rivest estimation procedure for the generator with complete data

Genest & Rivest (1993) were the first to propose a procedure for identifying a generator in empirical applications when data are completely observable. 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 & Rivest (1993), since the empirical estimate of the bivariate distribution function is always larger than 1/n and since the estimator has to take values on (0,1) range, K can be estimated as

$$\widehat{K}_n(z) = \frac{1}{n}\#\{i|z_i \le z\} \text{ where } z_i = \frac{1}{n-1}\#\Big\{(x_{(j)},y_{(j)})\Big|x_{(j)} < x_{(i)}, y_{(j)} < y_{(i)}\Big\}, \tag{3.5}$$

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

3.4 Wang-Wells general estimation procedure for the generator in the presence of censored data

However Genest-Rivest technique is no longer appropriate when the data are subject to censoring. For such cases, Wang & Wells (2000b) proposed a modified estimator of K. Since K can be written as

$$K(v) \;=\; \Pr[F(X,Y) \le v] = \mathbb{E}[\mathbb{I}\{F(X,Y) \le v\}]$$

the suggested estimator is given by

$$\widehat{K}_n(v) = \int_0^\infty \int_0^\infty \mathbb{I}[\widehat{F}(x,y) \le v]d\widehat{F}(x,y) \tag{3.6}$$

where $\widehat{F}$ stands for a nonparametric estimator of the joint distribution function F taking censoring into account. As mentioned by the authors, this approach is sufficiently flexible to deal with various censorship mechanisms, as long as $\widehat{F}$ is an appropriate estimator for F. The estimator we use for modelling the joint distribution of loss-ALAE data is described in details in Section 3.5.

3.5 Akritas estimation procedure for a bivariate distribution function under censoring

Different nonparametric estimators of a bivariate distribution were e.g. proposed by Dabrowska (1988), Prentice & Cai (1992), van der Laan (1996) and Prentice, Moodie & Wu (2004). A widely used estimator of the bivariate survival function is the one developed by Dabrowska (1988). It is a generalisation 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 Section 3 of Dabrowska (1988), the proposed estimator fails to be monotone. The weak convergence of the estimator of the bivariate survival function has been obtained in Dabrowska (1989).

When only one variable is subject to censoring, a nonparametric estimator for the bivariate distribution was proposed by Akritas (1994). 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, were proven in Akritas (1994).

As already pointed out in the introduction, the loss-ALAE data set is made of non identically distributed data. In order to use the estimator developed by Akritas (1994) 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 (i.e. the censoring variable, the policy limit ℓ in this case, is constant and differs from each individual to another), in the terminology of Klein & Moeschberger (1997, page 57). Let us now prove that this type of censoring leads to the same likelihood function than random right censoring (up to a factor not depending on the unknown survival distribution). This will show that the estimator of Akritas (1994) remains consistent when type I-censoring is present.

Let the observed data be denoted by (t_i, y_i, δ_i), $i = 1, \ldots, n$, and let g be a known probability density function (a kernel function) and $\{h_n\}$ a sequence of positive constants tending to zero as n tends to infinity (a bandwidth sequence). The conditional local likelihood of X given Y at the point y is then given by

$$\begin{aligned} L(y) &= \prod_{i=1}^{n} g\left(\frac{y-y_i}{h_n}\right) \Pr[T_i = t_i, \Delta_i = \delta_i | Y_i = y_i] \\ &= \prod_{i=1}^{n} g\left(\frac{y-y_i}{h_n}\right) \Pr[X_i = t_i | Y_i = y_i]^{\delta_i} \cdot \Pr[X_i > t_i | Y_i = y_i]^{1-\delta_i} \\ &= \prod_{i|\delta_i=1} g\left(\frac{y-y_i}{h_n}\right) \left[F_{X|Y}(t_i|y_i) - F_{X|Y}(t_i - |y_i)\right] \\ &\quad \cdot \prod_{i|\delta_i=0} g\left(\frac{y-y_i}{h_n}\right) [1 - F_{X|Y}(t_i|y_i)], \end{aligned}$$

whereas the likelihood for randomly censored data contains in addition a factor depending solely on the conditional censoring distribution. Since 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 below).

The bivariate distribution function F can be written as:

$$F(x,y) = \Pr[X \le x, Y \le y] = \int_0^y F_{X|Y}(x|z)dF_Y(z).$$

The proposed estimator of F will be based on the estimate of the conditional distribution $F_{X|Y}(x|y) = \Pr[X \le x|Y = y]$, i.e.

$$\begin{aligned}\widehat{F}(x,y) &= \int_0^y \widehat{F}_{X|Y}(x|z)d\widehat{F}_Y(z)\\ &= \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}[0 \le y_k \le y] \cdot \widehat{F}_{X|Y}(x|y_k),\end{aligned}$$

where

$$\widehat{F}_{X|Y}(x|z) = 1 - \prod_{i|T_i \le x \text{ and } \delta_i=1} \left[1 - \frac{W_{ni}(z;h_n)}{\sum_{j=1}^{n} W_{nj}(z;h_n)\ \mathbb{I}[t_j \ge t_i]}\right] \tag{3.7}$$

is the Beran (1981) estimator and

$$W_{ni}(z;h_n) = \frac{g\left(\frac{z-Y_i}{h_n}\right)}{\sum_{j=1}^{n} g\left(\frac{z-Y_j}{h_n}\right)}. \tag{3.8}$$

4 Application to loss-ALAE modelling

4.1 Nonparametric estimation of the generator

The kernel density function we use in (3.8) is the biweight kernel, i.e.

$$g(u) = \frac{15}{16}\left(1-u^2\right)^2 \cdot \mathbb{I}\{\ |u| \le 1\}.$$

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 $\widehat{K}_n$ of the distribution K. Since the $AMSE$ has a complicated structure and depends on a number of unknown quantities, it will be used by means of a bootstrap procedure.

Let us describe this procedure performed for a fixed value of h_n. We first generate two uniform random numbers on $[0,1]$, u^* and v^*, say. Then we construct the uncensored bootstrap data Y_i^* from the empirical distribution of ALAE, i.e $Y_i^* = \widehat{F}_Y^{-1}(v^*)$, and the censored bootstrap data X_i^* from the conditional ditribution, i.e. $X_i^* = \widehat{F}_{X|Y}^{-1}(u^*|Y_i^*)$. Since Y_i^* 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. $\ell_i^* = \ell_j$. With the bootstrapped data $(T_i^*, Y_i^*) = (\min(X_i^*, \ell_i^*), Y_i^*)$ and the indicator $\Delta_i^* = \mathbb{I}[T_i^* = X_i^*]$, $i = 1, \ldots, n$, we then estimate the distribution K and compute the $AMSE$.

Specifically, consider a grid $\boldsymbol{v} = (v_1, \ldots, v_m)$ (in the numerical illustration, we take $m = 100$) on the unit interval $[0, 1]$ and denote by $\widehat{K}_n^*(v) = (\widehat{K}_{n,b}^*(v))_b, \quad b = 1, \ldots, B,$ the $B \times m$ -matrix of the empirical estimates of K given by (3.6) for the B resamples, computed on this grid. In each value v_l of the grid we estimate the bias, the variance and the MSE as follows:

$$\begin{aligned}
\widehat{Bias_h}[\widehat{K}_n^*(v_l)] &= \frac{1}{B} \sum_{j=1}^{B} \widehat{K}_{n,b}^*(v_l) - \widehat{K}_n(v_l) \\
\widehat{Var_h}[\widehat{K}_n^*(v_l)] &= \frac{1}{B} \sum_{j=1}^{B} \left[\widehat{K}_{n,b}^*(v_l)\right]^2 - \left(\frac{1}{B} \sum_{j=1}^{B} \widehat{K}_{n,b}^*(v_l)\right)^2 \\
\widehat{MSE_h}[\widehat{K}_n^*(v_l)] &= \widehat{Var_h}[\widehat{K}_n^*(v_l)] + \left(\widehat{Bias_h}[\widehat{K}_n^*(v_l)]\right)^2,
\end{aligned}$$

for $l = 1, \ldots, m$. The optimal bandwidth will be then the one which minimizes

$$\widehat{AMSE_h} = \frac{1}{m} \sum_{l=1}^{m} \widehat{MSE_h}[\widehat{K}_n^*(v_l)].$$

The results based on $B = 500$ resamples are plotted in Figure 4.1. It follows that the optimal value is $h_n = 0.4$.

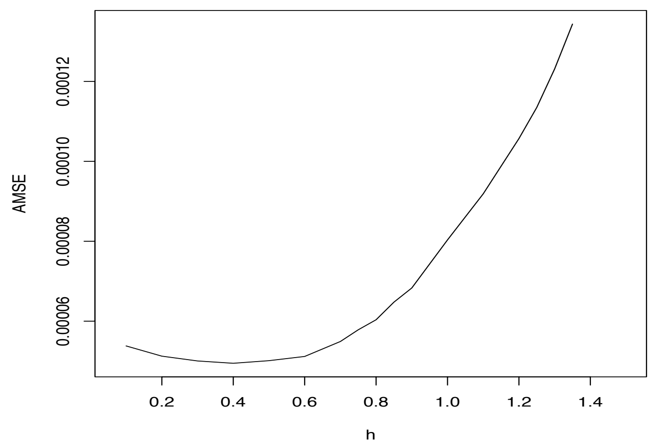

Figure 4.1: *AMSE of $\widehat{K}_n$ for different values of the bandwidth h_n.*

The estimation of K then follows from (3.6) and the resulting $\widehat{K}_n$ is depicted in Figure 4.2. The generator of the archimedean copula is then obtained by plugging $\widehat{K}_n$ into (3.1). The estimate of the generator for the loss-ALAE data is depicted in Figure 4.2.

4.2 Comparison with Dabroswka and Genest-Rivest estimations

It would be interesting to compare the estimates of the distribution K, when for estimating the bivariate distribution function we use either the estimator of Dabrowska (1988) or the estimate in the uncensored case suggested by Genest & Rivest (1993) (i.e. ignore the censored loss variables and work only with 1,466 observations over the initial 1,500 data points). The resulting estimations are depicted in Figure 4.3, together with $\widehat{K}_n$ of Figure 4.2. 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.

4.3 Graphical model selection procedure for the generator

4.3.1 General principle

The idea is now to compare the empirical estimator $\widehat{K}_n$ with several parametric analogues K_α corresponding for instance to Clayton, Gumbel, Frank or Joe copulas, in order to select the best parametric model. It will be selected as to minimize the L^2-norm distance:

$$S(\alpha) = \int \left[K_\alpha(\xi) - \widehat{K}_n(\xi)\right]^2 d\xi.$$

4.3.2 Estimation of the dependence parameter

Wang & Wells estimation procedure An initial value for the dependence parameter α can easily be obtained with the method-of-moments, from the one-to-one relationship (3.4) between the population version of Kendall's tau and α.

Nonparametric estimation of Kendall's tau under censoring is a complex problem. One estimation of Kendall's tau can be obtained from (3.4), with K replaced by its empirical estimator $\widehat{K}_n$, given in (3.6). For loss-ALAE data, we get $\widehat{\tau} = 0.3669$.

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

$$\tilde{\tau} = 4 \sum_{i=1}^{n} \sum_{j=1}^{n} \widehat{F}(t_{(i)}, y_{(j)}) \cdot \widehat{F}(\triangle t_{(i)}, \triangle y_{(j)}) - 1 \tag{4.1}$$

where $x_{(i)}$ ($y_{(j)}$) is the $i-th$ ($j-th$) ordered observation of X_i's (Y_i's respectively) variable, $\widehat{F}$ is the Dabrowska estimator of the bivariate distribution function

$$\begin{aligned} \widehat{F}(\triangle t_{(i)}, \triangle y_{(j)}) &= \widehat{F}(t_{(i)}, y_{(j)}) - \widehat{F}(t_{(i-1)}, y_{(j)}) \\ &\quad - \widehat{F}(t_{(i)}, y_{(j-1)}) + \widehat{F}(t_{(i-1)}, y_{(j-1)}). \end{aligned}$$

The approach of Wang & Wells (2000a) applied to Loss-ALAE data yields $\tilde{\tau} = 0.3567$, a value which is close to $\widehat{\tau}$.

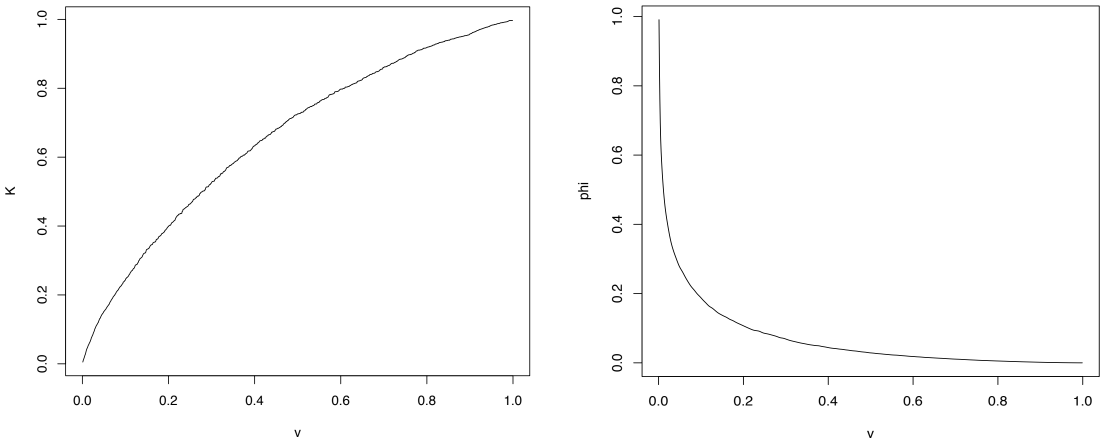

Figure 4.2: *Graph of $\widehat{K}_n$ (left panel) and $\widehat{\phi}$ (right panel) for the loss-ALAE data set.*

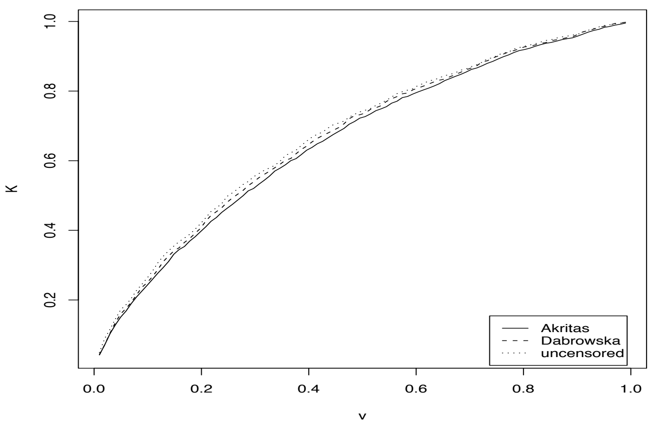

Figure 4.3: *Graph of $\widehat{K}_n$ obtained using Dabrowska and Akritas estimators, and by neglecting censored data points.*

Omnibus estimation procedure The omnibus semiparametric procedure, known also as the maximum pseudo-likelihood procedure, treats marginal distributions as (infinite dimensional) nuisance parameters. It consists in substituting empirical analogues for the marginal distribution functions in the likelihood for the dependence parameters and then in maximizing the resulting peudo-likelihood. As shown by Genest et al. (1995) the resulting estimator is consistent and asymptotically normal, even in the presence of censorship. The first step consists in estimating the two marginals nonparametrically, by rescaled versions of Kaplan-Meier estimator (for loss variable) and empirical estimator (for ALAE variable). As explained by Genest et al. (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, $\widehat{F}_X$ and $\widehat{F}_Y$ respectively, are, in a second step, used to estimate the dependence parameter, as $\widehat{\alpha} = \arg\max_\alpha L(\alpha)$.

Since 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 $(u_i, v_i) = (\widehat{F}_X(t_i), \widehat{F}_Y(y_i))$, C_α is the archimedean copula under consideration and c_α 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 4.1.

Copula	Partial derivative $\frac{\partial C_\alpha(u,v)}{\partial v}$
Clayton	$[1 + v^\alpha(u^{-\alpha} - 1)]^{-1-1/\alpha}$
Frank	$\left[e^{-\alpha v} - e^{-\alpha(u+v)}\right]\left[(1-e^{-\alpha}) - (1-e^{-\alpha u})(1-e^{-\alpha v})\right]^{-1}$
Gumbel-Hougaard	$v^{-1} \cdot \exp\{-(\widetilde{u}^\alpha + \widetilde{v}^\alpha)^{1/\alpha}\} \cdot [1 + (\frac{\widetilde{u}}{\widetilde{v}})^\alpha]^{-1+1/\alpha}$
Joe	$(1-\bar{u}^\alpha)(1-\bar{u}^\alpha + \bar{u}^\alpha \bar{v}^{-\alpha})^{-1+1/\alpha}$

Table 4.1: *Partial derivatives of C_α (with the notations: $\widetilde{u} = -\ln u$ and $\bar{u} = 1-u$).*

4.4 Graphical representations

Let us now identify the appropriate archimedean copula. It is important to mention that all four parametric models considered here 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, $\tau < \frac{1}{3}$). The method-of-moment estimations (associated to the two estimates of Kendall's tau, $\tilde{\tau}$ and $\hat{\tau}$) and the omnibus estimations of the dependence parameters are gathered in Table 4.2. Excepting Frank and Gumbel-Hougaard copulas, for which the three values are quite close, for Joe and Clayton copulas, the estimates are different, indicating that these two models might not be appropriate for the data.

Copula	*Method-of-moments* for $\tilde{\tau} = 0.3567$	for $\hat{\tau} = 0.3669$	Omnibus
Clayton	1.1088	1.1594	0.5174
Frank	3.5919	3.7225	3.0861
Gumbel	1.5544	1.5797	1.4454
Joe	2.0074	2.0549	1.6504

Table 4.2: *Method-of-moment and omnibus estimates for the dependence parameter, α.*

Plotting in Figure 4.4 the nonparametric and the four parametric estimates $\lambda_{\widehat{\alpha}}$ (where $\widehat{\alpha}$'s are the omnibus estimates) of the function λ, as well as the nonparametric estimate $\widehat{\lambda}_n$, suggests that the closest parametric models are Frank and Gumbel-Hougaard models.

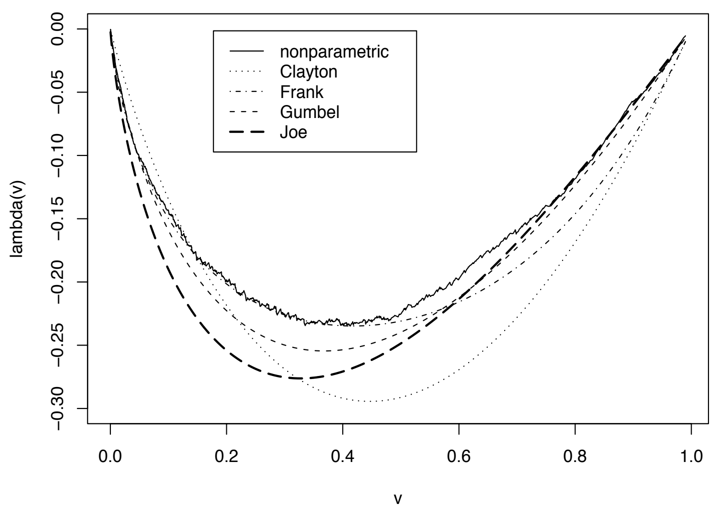

Figure 4.4: *Nonparametric and parametric estimates of λ.*

A look at the QQ-plot of the nonparametric and parametric quantiles of the distribution K depicted in Figure 4.5 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 $S(\widehat{\alpha})$ for these four models, where the estimated values of the dependence parameters are the omnibus estimates given in Table 4.2. The results are summarized in Table 4.3.

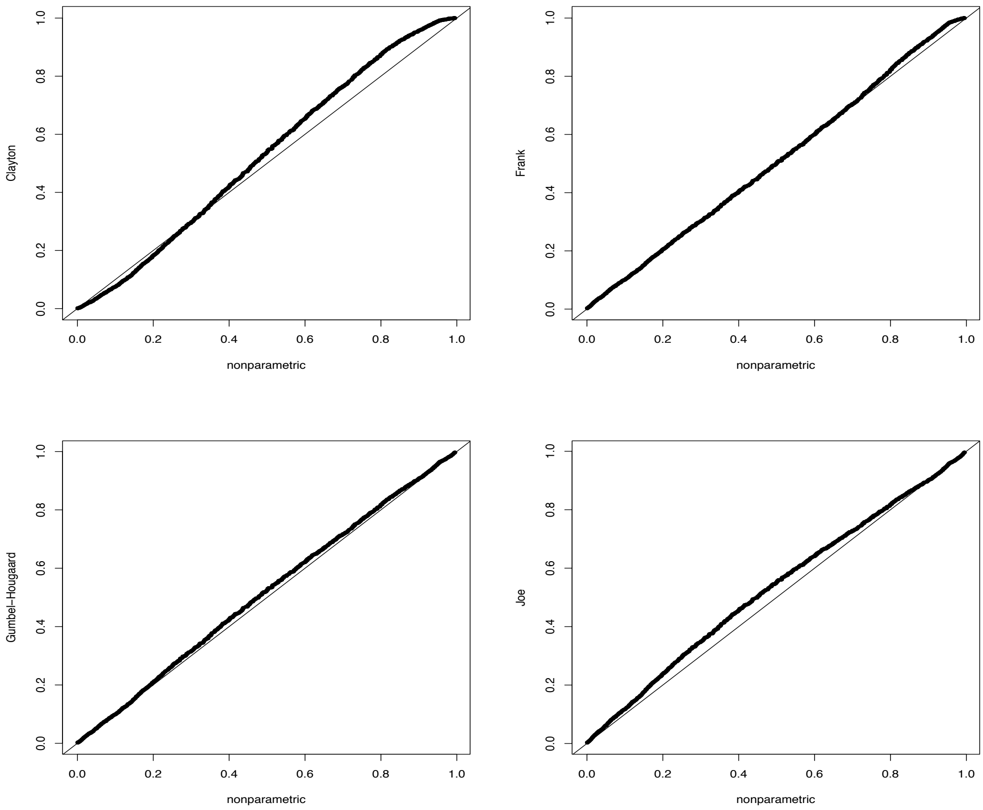

Figure 4.5: *QQ-plots corresponding to the nonparametric and parametric estimates of K.*

Copula	$S(\widehat{\alpha})$
Clayton	0.00245393
Frank	0.00028323
Gumbel-Hougaard	0.00025499
Joe	0.00099576

Table 4.3: *Distances $S(\hat{\alpha})$ for the parametric models, with $\hat{\alpha}$ the omnibus estimates.*

It follows that Gumbel-Hougaard model provides the best fit to the data, even though it is quite close to Frank model.

Conclusion

In this paper, we proposed a semiparametric modelling strategy for bivariate outcomes commonly encountered in 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 which 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 & Valdez (1998)), we believe that the procedure we proposed should be applied in practice. The proportion of censored data was indeed rather low (34 on 1,500 data points) and it can be expected that neglecting the censorship may lead to an incorrect choice of the archimedean copula.

Acknowledgements

We would like to thank Professors Frees and Valdez for kindly providing the loss-ALAE data, which were collected by the US Insurance Services Office (ISO). The authors are also grateful to Ronald Pruitt, Mark van der Laan and Statlib library (lib.stat.cmu.edu). Ingrid Van Keilegom gratefully acknowledges financial support from the IAP research network nr. P5/24 of the Belgian Government (Belgian Science Policy).

References

[1] Akritas, M.G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *The Annals of Statistics* **22**, 1299-1327

[2] Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical Report, University of California, Berkeley

[3] CARRIÈRE, J.F. (2000). Bivariate survival models for coupled lives. *Scandinavian Actuarial Journal* **100**, 17-32.

[4] DABROWSKA, D. M. (1988). Kaplan-Meier estimate on the plane. *The Annals of Statistics* **16**, 1475-1489.

[5] DABROWSKA, D. M. (1989). Kaplan-Meier estimate on the plane: weak convergence, LIL and the bootstrap. *Journal of Multivariate Analysis* **29**, 308-325.

[6] EMBRECHTS, P., MCNEIL, A., & STRAUMANN, D. (2002), Correlation and Dependency in Risk Management: Properties and Pitfalls. In *Risk Management: Value at Risk and Beyond*, eds Dempster M. and Moffatt H., Cambridge University Press, Cambridge.

[7] FREES, E.W., CARRIÈRE, J.F. and VALDEZ, E.A. (1996). Annuity Valuation with Dependent Mortality. *Journal of Risk and Insurance* **63**, 2, 229-261.

[8] FREES, E.W., and VALDEZ, E.A. (1998). Understanding relationships using copulas. *North American Actuarial Journal* **2**, 1-25.

[9] 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.

[10] GENEST, C., and MACKAY, R. J. (1986a). Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données. *Canadian Journal of Statistics* **14**, 145-159.

[11] GENEST, C., and MACKAY, R. J. (1986b). The joy of copulas: Bivariate distributions with uniform marginals. *The American Statistician* **40**, 280-283.

[12] GENEST, C., and RIVEST, L. (1993). Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association* **88**, 1034-1043

[13] KLEIN, J.P., and MOESCHBERGER, M.L. (1997). *Survival Analysis, Techniques for Censored and Truncated Data*. Springer Verlag, New York.

[14] KLUGMAN, S.A., and PARSA, R. (1999) Fitting bivariate loss distributions with copulas. *Insurance: Mathematics and Economics* **24**, 139-148.

[15] NELSEN, R.B.(1999). *An Introduction to Copulas*. Springer-Verlag, New York.

[16] PRENTICE, R. L., and CAI, J. (1992). Covariance and survivor function estimation using censored failure time data. *Biometrika* **79**, 495-512.

[17] PRENTICE, R. L., MOODIE, F. Z., and WU, J. (2004). Hazard-based nonparametric survivor function estimation. *Journal of the Royal Statistical Society* **66**, 305-319.

[18] VALDEZ, E. (2001). Bivariate analysis of survivorship and persistency. *Insurance: Mathematics & Economics* **29**, 357-373.

[19] van der Laan, M. J. (1996). Efficient estimation of the bivariate censoring model and repairing NPMLE. *The Annals of Statistics* **24**, 596-627.

[20] Wang, W., and Wells, M.T. (2000a). Estimation of Kendall's tau under censoring. *Statistica Sinica* **10**, 1199-1215.

[21] Wang, W., and Wells, M.T. (2000b). Model selection and semiparametric inference for bivariate failure-time data. *Journal of the American Statistical Association* **95**, 62-72.