T E C H N I C A L
R E P O R T

11003

Statistical models and methods for dependence in insurance data

VAN KEILEGOM, I. and N. VERAVERBEKE

I A P S T A T I S T I C S
N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

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

Discussion on: 'Statistical Models and Methods for Dependence in Insurance Data'

(by S. Haug, C. Klüppelberg and L. Peng)

Ingrid Van Keilegom
Université catholique de Louvain*

Noël Veraverbeke
Universiteit Hasselt†

January 10, 2011

We first wish to congratulate the authors for this insightful and inspiring review on copulas, and in particular on extreme value and tail copulas. In this comment we would like to discuss briefly an outlook on two possible extensions of the ideas put forward in this review. The first one is on the influence of covariates, and the second one is on the exploration of insurance data that are subject to right censoring.

1 Covariates

Consider the data on the Danish fire insurance claims: (X_1, X_2), with X_1 the loss to buildings and X_2 the loss to contents. Kendall's tau is significantly greater than zero and there is a positive dependence between the two variables X_1 and X_2. Suppose now that there is also information on some covariate Z and that we observe data on (X_1, X_2, Z). For example, Z could be a continuous covariate representing e.g. the age of the building, or a categorical variable representing the type of building (public building, housing, industry,

*I. Van Keilegom acknowledges financial support from IAP research network P6/03 of the Belgian Government (Belgian Science Policy), and from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650.

†N. Veraverbeke acknowledges financial support from IAP research network P6/03 of the Belgian Government (Belgian Science Policy), and from research grant MTM 2008-03129 of the Spanish Ministerio de Ciencia e Innovacion.

...), or the insurance company. Then the dependence structure between X_1 and X_2 may vary with the value z of the covariate Z and we could think of plotting Kendall's tau as a function of z. Formally we can introduce the joint and marginal distributions of (X_1, X_2), conditionally on $Z = z$. According to Sklar's theorem we then have a conditional copula C_z given by

$$C_z(u_1, u_2) = P(F_{1z}(X_1) \leq u_1, F_{2z}(X_2) \leq u_2),$$

where

$$F_{1z}(x_1) = P(X_1 \leq x_1 | Z = z)$$

and

$$F_{2z}(x_2) = P(X_2 \leq x_2 | Z = z).$$

This conditional copula will be unique provided F_{1z} and F_{2z} are continuous. The first paper mentioning the extension of Sklar's theorem to the conditional case is Patton (2006). He studied financial time series with dependence structure varying in time.

When Z is a continuous covariate, nonparametric estimators of the conditional copula have recently been studied in Veraverbeke, Omelka and Gijbels (2010) and Gijbels, Veraverbeke and Omelka (2010). The estimators are similar to those in formulas (3) and (4) in the paper, but the weights $1/n$ are replaced by a sequence of weights $\{w_{ni}(z, g_n)\}$ $(i = 1, \ldots, n)$ that smooth over the covariate space (where g_n is an appropriate bandwidth sequence). The conditional copula estimators can then be used to obtain conditional association measures that can be expressed as functionals of the former. For example, a conditional version of Kendall's tau can be constructed.

In the context of semiparametric estimation we mention that the pseudo maximum likelihood method of Genest, Ghoudi and Rivest (1995) given in Section 2.1 has recently been considered for conditional copulas. If C_θ is a parametric family of copulas, then the conditional copula is considered to be $C_{\theta(z)}$, where $\theta(\cdot)$ is now a function of the observed value z of Z. This has been studied in the case of known marginals by Acer, Craiu and Yao (2010) and in the case of estimated marginals by Abegaz, Gijbels and Veraverbeke (2010).

An interesting further line of development is the consideration of covariates Z that are in fact curves, or more generally Z is a functional covariate taking values in some infinite dimensional space. We refer to the influential monograph of Ferraty and Vieu (2006) for more details.

The message of this section is that the strength of dependence between two or more variables may vary according to the value of an observed covariate, and that this should be taken into account. It would be interesting to study the nonparametric estimation of a conditional copula, when the copula is an extreme value or tail copula. These two particular types of copulas require special attention, since the nonparametric estimators proposed by Veraverbeke, Omelka and Gijbels (2010) are not necessarily extreme value or tail copulas by construction. To the best of our knowledge, this has not been studied so far in the literature, and we would like to hear the authors' thoughts on this.

2 Censored data

Incomplete data are often encountered when analyzing insurance data. We focus here on right censored data, although left truncated data are also common (think e.g. of claims that are not reported to the insurance company because the amount of the claim does not exceed the deductible amount). Suppose X is the loss for a single claim. In practice, each claim will have a policy limit Y, i.e. the maximal claim amount insured by the company, which is specific to each contract. When the amount of the claim exceeds the policy limit (i.e. when $X \geq Y$), the loss variable will be right censored. More precisely, for a data set of n insurance claims one observes pairs (T_i, Δ_i) $(i = 1, \ldots .n)$, where $T_i = \min(X_i, Y_i)$, X_i is the ith loss, Y_i the associated policy limit, and

$$\Delta_i = I(X_i \leq Y_i) = \begin{cases} 1 & \text{if } X_i \leq Y_i \quad \text{(uncensored claim)} \\ 0 & \text{if } X_i > Y_i \quad \text{(censored claim).} \end{cases}$$

A lot of research has been done to extend procedures for completely observed data to censored data, but only very little work has been done for censored data in the context of extreme value analysis. The difficulty is that when no parametric assumptions are made on the distribution of X, the right tail of this distribution will be poorly estimated due to the presence of right censoring. This problem is well known in survival analysis, and often complicates the data analysis substantially. In the context of extreme value analysis, this problem is even more problematic, since there are only few data available in the tail. Two important papers that deal with this problem are Beirlant, Guillou, Dierckx and Fils-Villetard (2007) and Einmahl, Fils-Villetard and Guillou (2008). In these papers estimation of the extreme value index and of the extreme values is considered based on the product limit estimator of Kaplan and Meier (1958), which is the nonparametric maximum likelihood estimator under right censoring.

Suppose now that we have a vector (X_1, X_2) of insurance claims, and that X_1 and X_2 are both subject to censoring by variables Y_1 and Y_2, respectively. The dependence structure between X_1 and X_2 could then be expressed by means of a copula. Nonparametric copula estimators could be built from a bivariate Kaplan-Meier estimator for the joint distribution of X_1 and X_2 and univariate Kaplan-Meier estimators for the marginals. To the best of our knowledge this has not been considered yet.

On the other hand, when the copula is assumed to belong to a parametric family of copulas, Wang and Wells (2000) considered the estimation of this copula, extending the work of Genest and Rivest (1993) to bivariate censored data. Denuit, Purcaru and Van Keilegom (2006) revisited their estimator and applied it to insurance data, for which one variable is subject to right censoring and the other one is completely observed. This situation is encountered when e.g. X_1 represents the loss of a claim (subject to censoring), and X_2 represents the allocated loss adjustment expenses (ALAE's, in short) on a single claim (not subject to censoring). Here ALAE's are expenses made by an insurance company, 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. It would be interesting to investigate interval estimation and goodness-of-fit tests based on the proposed estimator.

References

Abegaz, F., Gijbels, I. and Veraverbeke, N. (2010). Semiparametric estimation of conditional copulas (submitted).

Acer, E.F., Craiu, R.V. and Yao, F. (2010). Dependence calibration in conditional copulas: a nonparametric approach. *Biometrics* (to appear).

Beirlant, J., Guillou, A., Dierckx, G. and Fils-Villetard, A. (2007). Estimation of the extreme value index and extreme quantiles under random censoring. *Extremes*, **10**, 151-174.

Denuit, M., Purcaru, O. and Van Keilegom, I. (2006). Bivariate Archimedean copula models for censored data in non-life insurance. *J. Actuar. Pract.*, **13**, 5-32.

Einmahl, J.H.J., Fils-Villetard, A. and Guillou, A. (2008). Statistics of extremes under random censoring. *Bernoulli*, **14**, 207-227.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis. Theory and Practice.* Springer, New York.

Genest, C. and Rivest, L. (1993). Statistical inference procedures for bivariate Archimedean copulas. *J. Amer. Statist. Assoc.*, **88**, 1034-1043.

Gijbels, I., Veraverbeke, N. and Omelka, M. (2010). Conditional copulas, association measures and their applications. *Comput. Statist. Data Anal.* (to appear).

Kaplan, E.L. and Meier, P. (1958). Non-parametric estimator from incomplete observations. *J. Amer. Statist. Assoc.*, **53**, 547-481.

Patton, A.J. (2006). Modelling asymmetric exchange rate dependence. *Intern. Econ. Review*, **47**, 527-556.

Veraverbeke, N., Omelka, M. and Gijbels, I. (2010). Estimation of a conditional copula and association measure (under minor revision).

Wang, W. and Wells, M.T. (2000). Model selection and semiparametric inference for bivariate failure-time data. *J. Amer. Statist. Assoc.*, **95**, 62-72.