International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:8, No:1, 2014

Estimation of the Upper Tail Dependence Coefficient for Insurance Loss Data Using an Empirical Copula-Based Approach

Adrian O'Hagan, Robert McLoughlin

***Abstract*—**Considerable focus in the world of insurance risk quantification is placed on modeling loss values from lines of business (LOBs) that possess upper tail dependence. Copulas such as the Joe, Gumbel and Student-t copula may be used for this purpose. The copula structure imparts a desired level of tail dependence on the joint distribution of claims from the different LOBs. Alternatively, practitioners may possess historical or simulated data that already exhibit upper tail dependence, through the impact of catastrophe events such as hurricanes or earthquakes. In these circumstances, it is not desirable to induce additional upper tail dependence when modeling the joint distribution of the loss values from the individual LOBs. Instead, it is of interest to accurately assess the degree of tail dependence already present in the data. The empirical copula and its associated upper tail dependence coefficient are presented in this paper as robust, efficient means of achieving this goal.

***Keywords*—**Empirical copula, extreme events, insurance loss reserving, upper tail dependence coefficient.

I. Introduction

THE Gaussian copula [1] has enjoyed great popularity among actuaries and other practitioners in the insurance industry for the purpose of modeling aggregate losses arising from multiple risk sources. Its widespread appeal derives in large part from the necessity to evaluate only one input parameter, the Pearson Correlation Coefficient [2], in order to fit the copula. It retains the inherent appeal of all copulas as an alternative to modeling the multivariate distribution of several variables: one can focus on accurately modeling the univariate marginal loss distributions prior to unifying them under the copula structure.

However, with the advent of the 2008 global financial crash, the Gaussian copula received immense criticism as being implicit in the economic meltdown [3]. The negativity centered on its inability to incorporate tail dependence between marginal distributions, a feature that proved to be markedly unrealistic for the financial products it was being used to model. The Student-t copula has been suggested as a natural successor to the Gaussian copula, since it does incorporate right tail dependence between its marginal distributions [4]. Consequently its use has been promoted in financial and insurance regulation, including Solvency II. Unfortunately the Student-t copula assumes that the input data arises from an elliptical distribution for each marginal, which is clearly unrealistic for many practical applications.

Alternatively, right-tailed Archimedean copulas including the Joe and Gumbel copulas may be used, and have been documented extensively in industry literature for the benefit of actuarial practitioners [5]. These Archimedean copulas show significant promise as a means of progressing beyond the Gaussian approach to assimilating marginal distributions. They do suffer from their own drawbacks in terms of underlying assumptions however, primarily their inability to capture lower tail dependence. Hence some firms may choose instead to "manually" construct a loss model based on the underlying real-world processes that generate losses, such as extreme weather events. Irrespective of the type of model used to simulate losses, it is imperative that the simulations from such models be examined to determine the extent to which they reflect the risk of an extreme total loss to the company within a fixed time period. Such extreme losses occur when most or all lines of business suffer extreme individual losses simultaneously, a phenomenon known as "tail dependence".

A natural approach to gauging a model's ability to effectively simulate extreme total loss events is to use the empirical copula [6] to assess the level of upper tail dependence between lines of business in the model. This provides the standard benefit of all non-parametric approaches in that it does not make any distributional assumptions as to the form of the copula or the nature of the underlying data. Through novel manipulation of the link between the joint empirical distribution of the data and the empirical copula, a robust estimate of upper tail dependence can be extracted.

II. Data

The data comprise of 50,000 simulated general insurance loss claims across each of 9 lines of business (LOBs). For the purposes of confidentiality, the nature of each line of business was not disclosed, nor was the exact nature of the simulation process. However, it is known that each loss value can be decomposed into three components, namely the large loss (L), attritional loss (A) and catastrophe loss (CAT) layers. The sum of these produces the overall loss value for each LOB in each simulation.

Dr. Adrian O'Hagan is the Director of Postgraduate Actuarial Science in the School of Mathematical Sciences at University College Dublin, Belfield, Dublin, Ireland (phone: +353-1716-7377; e-mail: Adrian.ohagan@ucd.ie).

Mr. Robert McLoughlin is a graduate of the UCD M.Sc. Actuarial Science programme, class of 2013.

R J Kiln & Co Limited kindly provided the simulated loss claims data analyzed in this paper; as well as financial sponsorship of Mr. McLoughlin to facilitate his undertaking this research as part of his M.Sc. dissertation.

The large and attritional loss values are simulated from standard statistical distributions such as the exponential or gamma, parameterized using Kiln's historical claims experience. The addition of the CAT loss layer imparts upper tail dependence on the modeled loss process reflecting underlying events or losses to which LOBs are exposed. In each simulation a random number of CAT events are simulated, each with their own loss size. These losses are then added to some, but not all, lines of business, with appropriate scaling to reflect the degree of exposure of that LOB to the type of CAT loss simulated.

For example, general insurance claims from LOBs covering car and home insurance for a coastal city would both be significantly impacted by a simulated hurricane CAT event; whereas similar LOBs for a landlocked city would not. Hence the first pairing of LOBs would be expected to display upper tail dependence in this instance whereas the second pairing would not. Fig. 1 presents the pairs plots of the simulated loss claims for all 9 lines of business. Fig. 2 presents the pairs plot for lines of business 5 and 6 only. This pairing is used for illustrative purposes later in the paper due to the high level of upper tail dependence between the variables it encapsulates.

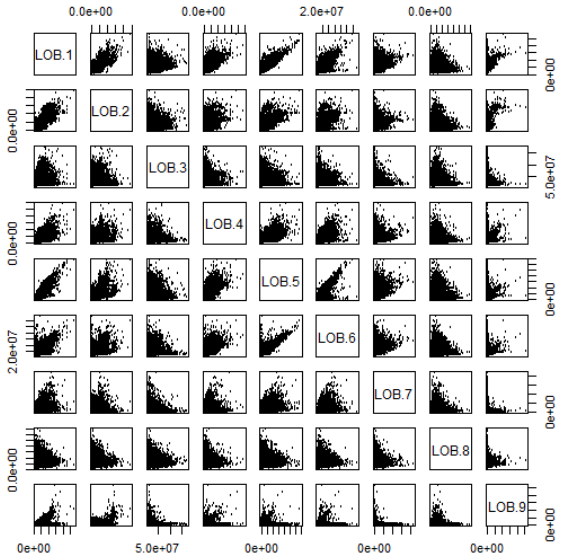

Fig. 1 Pairs plot of the simulated loss data for 9 lines of business

The ultimate goal when producing simulated loss data of this nature is to identify extreme total loss events, which occur when a large number of CAT losses each impact multiple LOBs within a single time period (i.e. multiple LOBs that have upper tail dependence). The insurance company must be able to demonstrate solvency to the regulator under such extreme scenarios, achieved by setting aside sufficiently large capital reserves. Insurance companies increasingly attempt to incorporate upper tail dependence in their loss models, through copula-based or other procedures [7]. The empirical copula provides a natural method of verifying whether or not their attempts have proven successful.

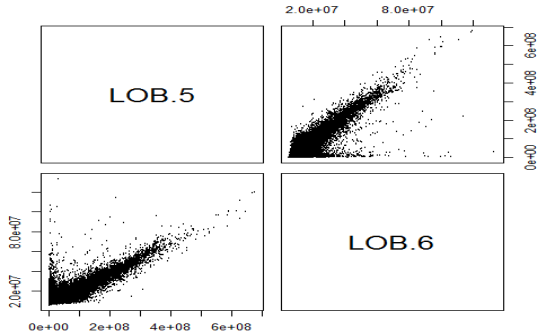

Fig. 2 Pairs plot of the simulated loss data for LOBs 5 and 6.

III. Methods

A. The Empirical Copula

Sklar's Theorem [8] describes the dependence between two or more random variables $X_1, X_2, \ldots, X_d$. It states that the joint cumulative distribution function (CDF) of the random variables, $H(x_1, \ldots, x_d)$, can be expressed as a function C of the marginal CDFs, $F_1(x_1), \ldots, F_d(x_d)$:

$$H(x_1, \ldots, x_d) = \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d) \quad (1)$$

$$H(x_1, \ldots, x_d) = C\big(F_1(x_1), \ldots, F_d(x_d)\big) \quad (2)$$

C is known as a copula and is unique if all marginal CDFs, $(F_1, \ldots, F_d)$, are continuous (which is true for the simulated loss data).

In the case of the empirical copula, the equivalence of (1) and (2) means that it is straightforward to fit the copula structure by simply calculating the joint empirical CDF of the observed values of the variables. However, some novel adjustments are required if the ultimate goal is to derive the upper tail dependence coefficient from the empirical copula. Primarily, estimated joint CDF values must be sourced for pairings of X and Y values not present in the observed data, or the empirical copula fitted will not be smooth across the range of the variables [9]. Further details are provided in Section III.

B. The Upper Tail Dependence Coefficient

Upper tail dependence between two random variables is the phenomenon whereby knowledge of the realisation of a large (tail) value for one random variable increases the probability of a large value being realised for the other random variable. The concept extends to the general case of multiple random variables. The upper tail dependence coefficient [10], λ_U, is a numerical value that captures this information, calculated as:

$$\lambda_U = \lim_{v \to 1^-} P\big(Y > F_Y^{-1}(v) \big| X > F_X^{-1}(v)\big) \quad (3)$$

In a copula-based setting this can also be expressed as:

$$\lambda_U = \lim_{v \to 1^-} \frac{1 - 2v + C(v,v)}{1 - v} \quad (4)$$

For the family of parametric copulas, the form of C employed in (4) will produce its own specific formulation for

λ_U for each copula. In the case of the non-parametric empirical copula, [11] provides a useful derivation of the form taken by λ_U for two random variables:

$$\lambda_U = 2 - \frac{ln\ C(1-\frac{t}{T},\ 1-\frac{t}{T})}{ln\ (1-\frac{t}{T})} t \approx \sqrt{T} \quad (5)$$

where *ln* denotes the natural logarithm and *T* denotes the number of observations present for each of the marginal distributions.

However, as is the case for other sources presenting such results for tail dependence, there is scant illustration as to how the value might be calculated in a practical setting, and the means by which the process may be automated for speed and efficiency. There is also minimal guidance given as to the sensitivity of the estimate of upper tail dependence to the input parameter *t*. This paper and its accompanying package *EmpCop*, currently under development for the statistical programming language R [12], seek to remedy these deficiencies.

The key intuition required when applying (5) is that, although values *(X=x*)* and *(Y=y*)* yielding CDF values close to $(1-\frac{t}{T})$ may exist for each of the marginal distributions separately; the paired value (*x*, y*)* is unlikely to exist in the observed data and possess a corresponding joint empirical CDF value. Hence the joint empirical CDF must be extrapolated and estimated at regular intervals not present in the observed data in order to find an approximation to the intersection $C(1-\frac{t}{T},\ 1-\frac{t}{T})$.The steps required to calculate the estimate of λ_U for two random variables are provided in algorithmic form in Section III C.

C. Using the Empirical Copula to Estimate the Upper Tail Dependence Coefficient

Assuming that the random variables for losses arising from two lines of business are denoted by *X* and *Y* and that there are *T* observations in total for each of *X* and *Y*:

1. Sort the marginal values of *X* and *Y* in ascending order.
2. Compute the marginal empirical CDFs for *X* and *Y*.
3. Scale each set of marginal empirical CDF values by the multiplicative constant *T/(T+1)*. This prevents the maximum loss in each marginal having a CDF value of 1 (this is desirable since losses above the maximum value simulated are possible).
4. Calculate the value $(1-\frac{t}{T})$ using t = $\sqrt{T}$, the joint CDF intersection point of interest.
5. Find the largest value *x'* such that *P(X<x')* $<(1-\frac{t}{T})$. In other words find the value of X that has marginal CDF as close as possible to the desired value $(1-\frac{t}{T})$.
6. Repeat step 5 for *Y* to identify y'.
7. Calculate the joint CDF for *X* and *Y*, extrapolated for paired values of *X* and *Y* not present in the underlying data, *H*(x,y)*. Many approaches to this problem are available. We employ the nonparametric kernel smoothing method present in the *np* package in R [13].
8. Use the extrapolated joint CDF formed in step 7 to find the value *H*(x',y')*. Using (1) and (2) from Sklar's Theorem this result can be expressed as:

$$H^*(x',y') \approx \mathbb{P}(X<x',\ Y<y') \approx C\left(1-\frac{t}{T},\ 1-\frac{t}{T}\right) \quad (6)$$

9. Calculate the value of the upper tail dependence coefficient λ_U' using (5) and substituting *H*(x',y')* for $C(1-\frac{t}{T},\ 1-\frac{t}{T})$, as detailed in (6).

For small data sets it is desirable to calculate a second estimate of the upper TDC, λ_U''. This is done by returning to step 5 and finding *x''* and *y''* such that *x''* and *y''* are the smallest values of *X* and *Y* resulting in *P(X<x'')*$>(1-\frac{t}{T})$ and *P(Y< y'')* $>(1-\frac{t}{T})$. This can be averaged with λ_U' to give a more robust estimate of the upper TDC. As the number of observations *T* increases, λ_U' and λ_U'' converge to the same value.

IV. Results

Using the methodology described in Section III, tail dependence coefficients were calculated for all pairings of lines of business in the simulated loss data. A fixed subset of size 2,500 claims was used for this purpose. Beyond this value of *T* the computation time becomes cumbersome for a standard laptop processor. However, subdividing the data into samples of size 2,500 permitted the volatility of the results for tail dependence to be checked across separate samples, a marked advantage to not using the data in its entirety at a single pass.

The main source of computational burden associated with this method is the formation of the extrapolated grid of joint cumulative distribution values across the variables, outlined in Section III *C*. However, for industrial purposes, the use of a computer cluster or mainframe would likely render the process feasible for larger values of *T* and in higher dimensional settings than the bivariate case considered in this work. Hence it is expected that the *EmpCop* package for R that is currently under preparation will prove very appealing to end users interested in assessing tail dependence in insurance loss settings and beyond.

The results for upper TDCs proved to be unanimously consistent with expectations as to the nature of the underlying lines of business and how they were affected by simulated catastrophe loss events that induce tail dependence. Table I presents the mean upper TDC values recorded for each LOB pairing across the 20 samples of simulated loss values, each of size 2,500. The results for the upper TDCs also displayed minimal volatility across the 20 samples. For illustrative purposes, consider the case of lines of business 5 and 6, as presented in Fig. 2. This pairing of LOBs is known to be subject to the same CAT events and to similar degrees. Hence high upper tail dependence in the simulated losses is desirable if the underlying loss model is efficiently recreating real world risk sources. The mean upper tail dependence coefficient recorded for this LOB pairing was 0.866 with a standard

deviation of 0.051. Fig. 3 provides a histogram of the values of upper TDC λ_U recorded across the 20 sub-samples of loss claims for LOBs 5 and 6.

It was deemed essential to check the sensitivity of the results for upper TDCs to the assumed value for the input parameter $t = \sqrt{T}$. Hence, for each of the LOB pairings, a fixed sample of 2,500 simulated loss values was used to evaluate λ_U, with t varied from the value 45 to the value 55 in increments of 1 unit, i.e. 5 units either side of its base value of $t = \sqrt{2500} = 50$. Reassuringly the results for upper TDC across LOB pairings displayed minimal volatility for variations in the value of t. As an illustration, Fig. 4 presents a line plot of upper TDC values for lines of business 5 and 6 for the prescribed range of values of t.

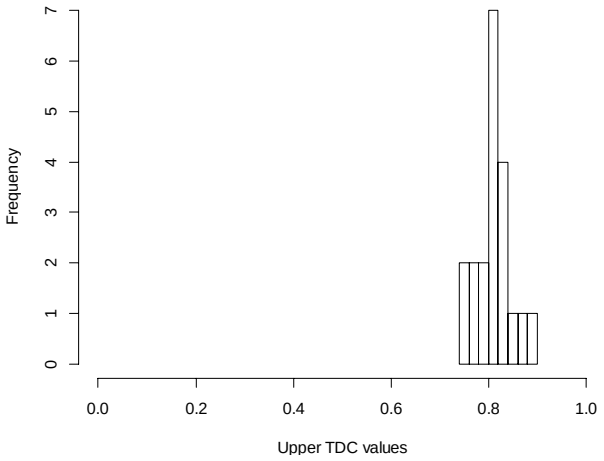

Fig. 3 Histogram of upper tail dependence coefficients using the empirical copula for 20 samples of 2,500 simulated loss claims across lines of business 5 and 6

TABLE I
MEAN UPPER TDC VALUES FOR ALL LOB PAIRINGS USING 20 SAMPLES OF 2,500 SIMULATED LOSS CLAIMS

	LOB1	LOB2	LOB3	LOB4	LOB5	LOB6	LOB7	LOB8	LOB9
LOB1		0.590	0.127	0.445	0.817	0.793	0.258	0.087	0.142
LOB2			0.171	0.479	0.566	0.566	0.238	0.104	0.143
LOB3				0.054	0.090	0.047	0.058	0.089	0.013
LOB4					0.494	0.472	0.254	0.053	0.075
LOB5						0.866	0.424	0.110	0.087
LOB6							0.313	0.101	0.055
LOB7								0.183	0.020
LOB8									0.087
LOB9									

V. CONCLUSION

The empirical copula provides a robust method of estimating the upper tail dependence coefficient between lines of business in an insurance loss setting. It is demonstrated to exhibit low sensitivity to its input parameter and provides stable results upon repeated resampling of the data. The empirical copula based approach offers users the benefit of requiring no assumptions as to the distributional nature of the underlying data or the copula used to assimilate the marginal random variables. In the practical setting of general insurance claims, this gives firms the flexibility to model the loss processes governing their lines of business according to any prescription, copula-based or otherwise, that reflects the tail dependence present between the sources of loss. Simulated values from the fitted model can then be tested using the empirical copula approach to verify that upper tail dependence has been successfully incorporated. This affords the user greater confidence in identifying extreme total loss events and setting aside appropriate reserves to ensure their solvency.

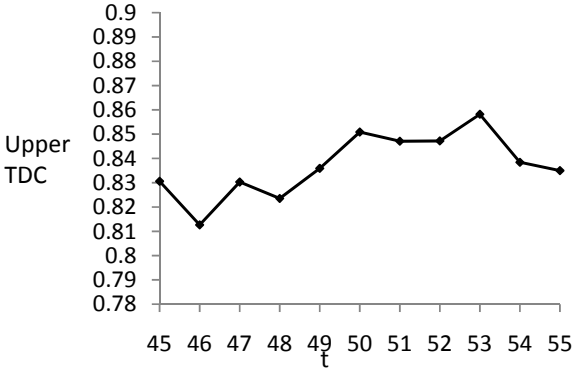

Fig. 4 Line plot showing variation in upper TDC for a sample of 2,500 simulated loss claims for different values of input parameter t

ACKNOWLEDGMENT

Dr. Adrian O'Hagan and Mr. Robert McLoughlin thank Mr. Brian Heffernan (FIA), Chief Actuary at R J Kiln & Co Limited, for hisexpertise in appraising the relevance and reliability of the methods and results presented in this paper.

REFERENCES

[1] T. Schmidt, "Coping with copulas," in *Copulas: From Theory to Application in Finance* 1st ed. J. Rank, Ed. London: Risk Books, 2006.
[2] S. D. Bolboacă and L. J äntschi, "Pearson versus Spearman, Kendal's Tau Correlation Analysis on the Structure Activity Relationships of Biologic Active Compounds," *Leonardo Journal of Sciences,* no. 9, pp. 179-200, July-December 2006.
[3] S. Lee, "Formula from Hell," *Forbes,* May 2009.
[4] S. Demarta and A. McNeil, "The t Copula and Related Copulas," *International Statistical Review,* vol. 73, no. 1, pp. 111-129, April 2005.
[5] G. G. Venter, "Tails of Copulas," *Proceedings of Casualty Actuarial Society* vol. 89, no. 171, pp.68-113, Mar. 2002.
[6] R. A. Shaw, A. D. Smith, and G. S. Spivak, "Measurement and modelling of dependencies in economic capital. A discussion paper," *The Institute and Faculty of Actuaries Sessional Meeting*, paper 10,pp. 54-71, May 2010.
[7] P. Embrechts, F. Lindskog, and A. McNeil, "Modelling Dependence with Copulas and Applications to Risk Management," in *Handbook of Heavy Tail Distributions in Finance: Handbooks in Finance.* Revised ed. vol. 1, S. T. Rachev, Ed. Amsterdam: North-Holland/Elsevier, 2003, pp. 329-384.
[8] A. Sklar, "Fonctions de répartition à n dimensions et leurs marges", *Publ. Inst. Statist. Univ. Paris,* vol.8, pp. 229–231, 1959.
[9] G. Frahm, M. Junker, and R. Schmidt, "Estimating the tail-dependence coefficient: Properties and pitfalls," *Insurance: Mathematics and Economics,* vol. 37, no. 1, pp. 80-100, 2005.
[10] M. Fischer and I. Klein, "Some results on weak and strong tail dependence coefficients for means of copulas, Discussion paper," Friedrich-Alexander-University Erlangen-Nuremburg, Chair of Statistics and Economics, No. 78, 2007.
[11] P. J. Sweeting and F. Fotiou, "Calculating and communicating tail association and the risk of extreme loss. A discussion paper," *The Institute and Faculty of Actuaries Sessional Research Discussion Paper*, pp.3-20, Sept. 2011.
[12] R Core Team. "R: A language and environment for statistical computing". R Foundation for Statistical Computing, Vienna, Austria, 2012.
[13] T. Hayfield and J. S. Racine, "Nonparametric econometrics: the np package", *Journal of Statistical Software*vol. 27, no.5., 2008.

Dr. Adrian O'Hagan graduated at the top of his degree class as a Bachelor of Actuarial and Financial Studies in 2005 at University College Dublin (UCD); and was awarded an MSc and PhD in Statistics by UCD in 2006 and 2012 respectively.

He completed an industrial placement at Imagine Reinsurance Ltd. From January-June 2005.He worked as an Assistant Lecturer in UCD, teaching subjects including Actuarial Mathematics, Survival Models and Actuarial Statistics while studying for his Ph.D. He currently works as a Lecturer and Director of Postgraduate Actuarial Science in the UCD School of Mathematical Sciences. He served as a proof reader for Statistical and Probabilistic Methods in Actuarial Science (New York, New York, Chapman & Hall, 2007). He published the article "Computational Aspects of Fitting Mixture Models via the EM Algorithm" in the journal Computational Statistics and Data Analysis in 2012. He has the articles "Jackknife and Bootstrap Methods of Variance Estimation in Model-Based Clustering" and "MNIGClust: Clustering Using the Multivariate Normal Inverse Gaussian Distribution" under review at the *Journal of Computational and Graphical Statistics* and *Statistics and Computing*, respectively, at the time of writing. His research interests include actuarial science, modeling extreme loss events, model-based clustering and political science.

Dr. O'Hagan is an Associate Member of the Institute and Faculty of Actuaries (IFA).

Mr. Robert McLoughlin is a 2013 graduate of UCD's MSc Actuarial Science programme. He completed research on the empirical copula as a means of assessing upper tail dependence as part of his MSc dissertation in association with Kiln and under the supervision of Dr. O'Hagan.