Copula-Based Risk Aggregation Modelling

Marie-Pier Côté

Master of Science

Department of Mathematics and Statistics

McGill University

Montréal, Québec

2014-06-10

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

ACKNOWLEDGEMENTS

I am grateful to my supervisor Professor Christian Genest for his valuable advice, his trust and his time. The massive amount of technical knowledge and the writing strategies he shared with me were more than useful in writing this thesis, and will surely be an asset for my future work. I would also like to thank Professor Johanna G. Nešlehová for her insightful comments and thoughts.

I thank my research collaborators Hélène Cossette, Christian Genest, Mélina Mailhot, Etienne Marceau and Khouzeima Moutanabbir for helping me and involving me into interesting projects. I am grateful to the École d'actuariat at Université Laval and the Department of Management Sciences at HEC Montréal for giving me the opportunity to teach courses during my Master's degree. Moral support was provided to me by my dear fiancé Jacques, my parents, my brothers, and my friends Andréane Gaudreault, Daniel Hains-Côté, Andrée-Anne Mailhot, Etienne Marceau, Nathalie Perron, Sarah-Émilie Racine, and Alex Soucy. I thank them for the cheers.

A key aspect of this work is its application to real insurance data. This would not have been possible without the collaboration of a Canadian insurance company, which prefers to not be identified.

Funding in support of this work was provided to me through an Alexander Graham Bell Canada Graduate Scholarships of the Natural Sciences and Engineering Research Council of Canada (NSERC), and a Masters Research Scholarship of the

Fonds de recherche du Québec – Nature et technologies (FRQNT). Additional funding was provided to me through grants to my supervisor from NSERC, FRQNT and the Canada Research Chairs Program.

ABSTRACT

A flexible approach is proposed for risk aggregation. The model consists of a tree structure, bivariate copulas, and marginal distributions. The construction relies on a conditional independence assumption whose implications are studied. A procedure for selecting the tree structure is developed using hierarchical clustering techniques, along with a distance metric based on Kendall's tau. Estimation, simulation, and model validation are also discussed. The approach is illustrated using data from a Canadian property and casualty insurance company.

ABRÉGÉ

On propose une approche flexible pour l'agrégation de risques. Le modèle est constitué d'une arborescence, de copules bivariées et de lois marginales. La construction s'appuie sur un postulat d'indépendance conditionnelle dont les ramifications sont étudiées. On montre comment choisir l'arborescence au moyen de techniques de classification et d'une métrique définie à partir du tau de Kendall. L'estimation, la simulation et l'adéquation du modèle sont aussi abordées. L'approche est illustrée à l'aide de données d'une compagnie canadienne en assurance IARD.

PREFACE AND AUTHOR CONTRIBUTIONS

This project started when I drew Professor Christian Genest's attention to the work of Arbenz, Hummel & Mainik (2012). In their paper, the authors explain a simulation algorithm for a copula-based hierarchical aggregation model used in the life insurance industry (although without much theoretical foundations). I was thrilled by the idea of giving strong arguments in favor of the relevance of that simple and practical model.

The work on this thesis was first theoretical. To understand the implications of the conditional independence assumption, Professor Genest suggested that I prepare some simple examples on the propagation of the dependence in the structure. I performed this task and then derived the joint density of the individual risks. Professor Genest and I investigated a way to build the tree structure, and he pointed out the usefulness of hierarchical clustering theory. After some work, I was able to prove that the distance based on Kendall's tau fulfills the triangle inequality. However, my proof was laborious and was greatly simplified with the help of my supervisor. Once the theoretical framework for the model was ready, I programmed the procedure in R. I tested it on simulated data from the Clayton–Liouville distribution, following a suggestion of Professor Nešlehová. I also tried the procedure on different sets of simulated data from the multivariate Normal distribution.

Professor Genest obtained real insurance data, which was very useful for the completion of a convincing model application. Many statistical techniques came handy to analyze the data: time series analysis, linear regression, maximum (pseudo)

likelihood estimation, goodness of fit tests, etc. I was guided through these techniques by Professor Genest while I performed the data analysis. I showed him my findings and he helped me select the marginal models and copula families.

Chapter 2 of this thesis is a manuscript co-authored by Professor Genest and me. I was responsible for the draft paper, but we worked together on the presentation, content and writing. I finally composed Chapter 3 to provide details on the data analysis performed previously. Professor Genest edited Chapters 1, 3 and 4.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

The first life insurance company, a mutual organization named Equitable Life Assurance Society, was founded in 1762 in London, England. Its success, mainly due to the sound practices established by the mathematician and actuary James Dodson, aroused the interest of investors in the life insurance industry. According to [8], as many as 149 life insurance companies were formed between 1844 and 1853 in Great Britain, but only 59 were still active at the end of this same period. The shocking number of failures outlined the need for legislation: the long-term nature of insurance is incompatible with immediate distribution of profits to investors without proper reserving. Hence, the development of mortality tables and the valuation of life insurance and annuities were the main interest of actuaries in the 18th century, and these problems were studied by Condorcet, De Moivre and Laplace to name a few (see, e.g., [41] for an historical account on their contributions).

Capital requirements are designed to protect policyholders; the intent of the legislation is to ensure that the insurer holds enough assets to support policy liabilities and to withstand unexpected adverse experience in the environment. However, the classical actuarial models are usually based on the assumption that the risks are independent. This assumption is clearly inappropriate to reflect the impact of environmental shocks. For example, many risks in a given geographical area would

typically be affected by a natural (or man-made) catastrophe. At a more granular level, the lifetimes of a husband and his wife are dependent in many ways, e.g., through the broken heart syndrome, or due to an increased likelihood of a common car accident. There is thus a need for actuarial models that can account for dependence.

Bivariate extensions of univariate models, such as the exponential or Pareto distributions, were used early on to consider dependence between two risks. A complete review is available in [32], where the authors also present common shocks models, such as the Marshall–Olkin bivariate exponential distribution. Common shock models are used in life insurance to account for the probability of simultaneous death on joint last-to-die policies. Dependence can also be introduced via a common mixture, used to represent, e.g., environmental or socio-economic conditions (see, e.g., [36] and [37]). In most of these models, however, the margins are of the same family and their parameters define the degree of dependence between the risks. Copula models (see, e.g., [39] or [21]) provide a simple way to remove these constraints and to model the joint behaviour separately from the marginals. In actuarial science, the use of copulas was illustrated in [15] as a way to model the joint survival of lives and the joint distribution of claims and allocated expenses. This seminal paper led to many other applications in the field. As noted in [19], copulas are now used extensively for risk management and risk measurement (see, e.g., [36] and [37]). Stochastic bounds on functions of dependent risks help to identify the worst case scenario for the insurer; a survey of the research on this topic is available in [11].

Risk aggregation is central in insurance: pooling the individual risks allows a more accurate prediction of the claim amounts. If claims are dependent, their aggregation is not as simple as in the independent case; the Fast Fourier Transform cannot be used and the Central Limit Theorem does not hold. Although deterministic aggregation methods for dependent risks were developed recently (see [36] for a survey), Monte Carlo simulations are frequently used to this end. Once the overall exposure to risk is determined, an amount of capital required to support the portfolio of risks can be calculated using risk measures such as VaR or TVaR (see, e.g., [3] for a discussion of coherent risk measures). On the other side, it is useful for risk management and strategic planning to understand the share of the risk associated to each sub-portfolio, i.e., to allocate capital to each individual risk. Capital allocation requires the knowledge of the joint distribution of the risks. Capital allocations for dependent risks were studied in [4], [12], and [17], for example.

This thesis was motivated by a risk aggregation and capital allocation problem for eight portfolios of automobile and home insurance of a large Canadian insurance company. As the relations between the risks are complex, they cannot be modelled with commonly used multivariate dependence structures such as Archimedean, elliptical or extreme-value copula families. Vine copula constructions or nested Archimedean copulas could be considered, but the inference is complex and the models are not easily interpretable. In Chapter 2, a flexible copula-based approach for modelling risk vectors is presented. The idea is to aggregate risks successively, two at a time. The approach is thus particularly well suited to aggregate claim modelling, but it could also be applied in other cases where the overall exposure is of

interest. Under a specific assumption of conditional independence, the joint density of the risks is uniquely determined by the model, which allows capital allocation. Chapter 2 is a manuscript that was submitted for publication; it was written in collaboration with Dr. Christian Genest.

Chapter 3 presents a summary of the data analysis that was performed for the application of the model to the insurance portfolios. A thorough analysis was necessary to fit marginal distributions and to model appropriately the dependence structure. Details are omitted for conciseness.

CHAPTER 2
Copula-Based Risk Aggregation Model

2.1 Introduction

It is now widely recognized that accounting for dependence between individual risks is crucial for overall exposure assessment. For example, an insurance company's total claim payment in a given year arises from different lines of business, regions or subsidiaries. To ensure the payment of all future claims arising from the aggregate portfolio, the company must determine an adequate level of capital. This amount is of critical importance for the company's solvency and profitability. Given that risks often share common environmental and socioeconomic conditions, they are generally dependent. It is thus in a company's interest to build a multivariate risk model tailored to its exposure. In fact, it is not only advisable, but also encouraged by the recent regulatory frameworks such as Basel II, Solvency II or the Guideline E-19 of the Office of the Superintendent of Financial Institutions in Canada.

Copulas provide a flexible tool for modelling the dependence between random variables. They are now used extensively, both in statistics and in substantive fields; see, e.g., [9], [11] and [37] for applications in quantitative finance, actuarial science and risk management, respectively. For problems involving large numbers of variables, Archimedean and elliptical copulas ([21]), vine copula constructions ([33]) and hierarchical copula models ([35]) are currently the most popular options. However,

their use is often limited, either because they provide a restrictive range of dependence structures, or because they involve complex inference procedures.

This chapter considers an alternative modelling strategy, previously investigated in [2], which can accommodate a large spectrum of dependence structures while relying entirely on well established, rank-based inference procedures for bivariate copulas. This iterative approach, which is simple to implement even in high-dimensional contexts, leads to an easily interpretable model when risk aggregation is of interest. To be specific, suppose that $X_1, \ldots, X_d$ are random variables whose partial sums are meaningful. Think for example of future claim amounts from different portfolios in property and casualty insurance. In such an application, the partial sum

$$S_A = \sum_{i \in A} X_i$$

is interpretable and observable for any subset $A \subseteq D = \{1, \ldots, d\}$. A model for the vector $\mathbf{X}_D = (X_1, \ldots, X_d)$ can then be constructed iteratively as follows. First, select the two risks X_i, X_j that are most dependent (in some sense) and combine them through a bivariate copula model. Then, replace the individual risks X_i and X_j by their sum $S_{\{i,j\}}$ and treat the latter as a new, combined risk. This leaves one with $d-1$ risks, on which the procedure can be repeated. Proceed iteratively until all the risks have been aggregated in a single sum.

This procedure, which involves $d \geq 2$ marginal distributions $F_1, \ldots, F_d$ and $d-1$ bivariate copulas, leads to a unique model for $\mathbf{X}_D$ under a conditional independence assumption formulated by [2]. This assumption and its implications are discussed in Section 2.2. As shown in Section 2.3, a notion of distance based on Kendall's tau can

be used in conjunction with classical hierarchical clustering techniques to determine the order in which risks are aggregated. The resulting model is then easy to construct, as illustrated in Section 2.4. Simulation from the model can be performed using an algorithm of the Iman–Conover type detailed in [2]. This procedure is described in Section 2.5, where the good performance of the proposed modelling strategy is illustrated on simulated data. Finally, Section 2.6 presents a real-life modelling exercise aimed at allocating capital between eight portfolios representing car and property coverages for different subsidiaries of a large Canadian insurance company.

2.2 Multivariate hierarchical copula-based model

Let $X_1, \ldots, X_d$ be $d \geq 2$ random variables with cumulative distribution functions (cdf) $F_1, \ldots, F_d$, respectively. For each $A \subseteq D = \{1, \ldots, d\}$, let $\mathbf{F}_A$ denote the cdf of the vector $\mathbf{X}_A = (X_i, i \in A)$. Let also F_A be the cdf of the sum S_A of all components in $\mathbf{X}_A$.

The hierarchical copula-based models considered here are derived from tree structures in which each node links exactly two random variables of the form S_A, S_B with $A, B \subset D$ and $A \cap B = \emptyset$; recall that if $A = \{i\}$ for some $i \in D$, then $S_A = X_i$. A simple example of such a structure involving $d = 3$ variables is depicted in Figure 2–1. In this specific case, a joint model for the pair (X_1, X_2) would first be constructed using marginal distributions F_1 and F_2, and copula $C_{\{1,2\}}$. This determines the distribution $F_{\{1,2\}}$ of $S_{\{1,2\}} = X_1 + X_2$ implicitly. Next, a joint model for the pair $(S_{\{1,2\}}, X_3)$ would be built by combining $F_{\{1,2\}}$ with F_3 using bivariate copula $C_{\{1,2,3\}}$. This sequential aggregation procedure can be abbreviated symbolically as $(X_1 + X_2) + X_3$.

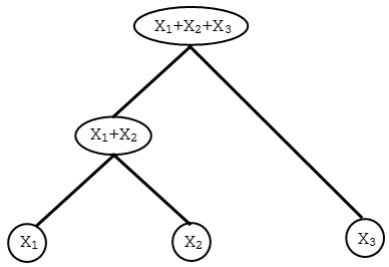

Figure 2–1: Illustration of a tree structure involving three variables.

This procedure has several advantages from a dependence modelling perspective. First, the model involves only bivariate copulas, as opposed to distribution functions in higher dimensions. Second, each of the $d-1$ copulas in the model can be chosen freely, thereby providing great flexibility. Third, classical rank-based procedures ([21]) can be used for inference purposes, as all pairs modelled are directly observable.

As already noted by [2], the joint cdf of $X_1, \ldots, X_d$ cannot be determined uniquely from their marginal distributions and $d-1$ bivariate copulas, unless $d = 2$ or additional assumptions are made. This is illustrated below in a simple case.

Example 1 *Consider the aggregation tree in Figure 2–1, and assume that* $X_i \sim \mathcal{N}(0,1)$*, for* $i = 1, 2, 3$*. Further assume that* $C_{\{1,2\}}$ *and* $C_{\{1,2,3\}}$ *are Gaussian copulas with parameters* ρ_1 *and* ρ_2*, respectively. If* $\operatorname{cov}(X_1, X_3) = a$*, then*

$$\operatorname{cov}(X_2, X_3) = \operatorname{cov}(X_1 + X_2, X_3) - \operatorname{cov}(X_1, X_3) = \rho_2\sqrt{2(1+\rho_1)} - a$$

because $\operatorname{var}(X_1 + X_2) = 2(1+\rho_1)$*. As* $\operatorname{cov}(X_2, X_3) = \operatorname{corr}(X_2, X_3)$*, one finds*

$$\max\{\rho_2\sqrt{2(1+\rho_1)} - 1, -1\} \leq a \leq \min\{\rho_2\sqrt{2(1+\rho_1)} + 1, 1\}.$$

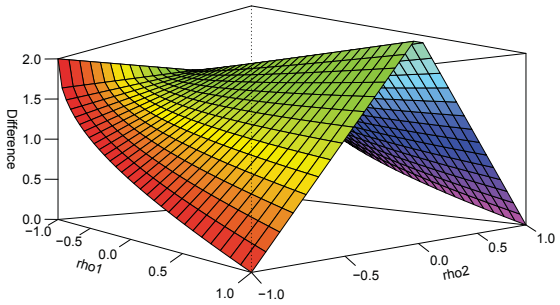

Figure 2–2: Difference between the upper and lower bounds for $\mathrm{corr}(X_1, X_3)$ in Example 1.

Figure 2–2 shows the difference between these bounds in terms of ρ_1 and ρ_2. Thus, even if the copulas of the pairs (X_1, X_3) and (X_2, X_3) are assumed to be Gaussian, they are not entirely determined.

An additional assumption which makes the joint distribution unique is proposed by [2]. In order to state this assumption, let $N_1, \ldots, N_{d-1}$ denote the branching nodes corresponding to the aggregation steps $1, \ldots, d-1$, respectively. (When the order in which the aggregation takes place is not entirely obvious from the tree structure, any convention can be used to make the numbering unique; this has no implication in what follows.) For each $i \in \{1, \ldots, d-1\}$, let S_{A_i} be the sum computed at node N_i and set $\bar{A}_i = D \setminus A_i$.

Conditional Independence Assumption: For each $i \in \{1, \ldots, d-2\}$, the random vectors $\mathbf{X}_{A_i}$ and $\mathbf{X}_{\bar{A}_i}$ are conditionally independent given S_{A_i}.

For the structure displayed in Figure 2–1, for example, one would have $A_1 = \{1, 2\}$ and $A_2 = \{1, 2, 3\}$. The above assumption then states that X_3 is independent of

(X_1, X_2) given $S_{\{1,2\}}$. Thus, once $X_1 + X_2$ is known, no information concerning the components of the sum affects the distribution of X_3.

Under this assumption, the joint distribution of $\mathbf{X}_D$ can be computed from the marginal distributions $F_1, \ldots, F_d$ and the copulas of the branching nodes. The following proposition gives an explicit (and original) expression for its density under the additional condition that all distributions involved are absolutely continuous. In what follows, $A_{i,1}$ and $A_{i,2}$ form the unique partition of A_i such that $S_{A_{i,1}}$ and $S_{A_{i,2}}$ are being modelled at node N_i.

Proposition 1 *Given a tree structure characterized by the sets $A_1, \ldots, A_{d-1}$, the joint density function of the vector $\mathbf{X}_D$ (assuming it exists) is given, for all $x_1, \ldots, x_d \in \mathbb{R}$, by*

$$\mathbf{f}_D(x_1, \ldots, x_d) = \prod_{i=1}^{d-1} c_{A_i} \left\{ F_{A_{i,1}}\left(\sum_{j \in A_{i,1}} x_j \right), F_{A_{i,2}}\left(\sum_{j \in A_{i,2}} x_j \right) \right\} \prod_{i=1}^{d} f_i(x_i)$$

in terms of the marginal densities $f_1, \ldots, f_d$ of $X_1, \ldots, X_d$ and the copula densities c_{A_i} linking $S_{A_{i,1}}$ and $S_{A_{i,2}}$ at node N_i.

A proof of this result is given in the Appendix. Assuming, for example, that $X_1, \ldots, X_d$ are Normally distributed and that the $d-1$ bivariate copulas in the tree structure are Gaussian, the formula in Proposition 1 reduces to a multivariate Normal density. This can be shown either directly through cumbersome calculations or deduced from Proposition 2.8 in [2], where a recursive formula is given for the covariance matrix of $\mathbf{X}_D$.

Example 1 (continued). *Under the conditional independence assumption, the joint density of the vector (X_1, X_2, X_3) is trivariate Normal with mean vector $\mathbf{0}$ and*

covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \rho_1 & a \\ \rho_1 & 1 & a \\ a & a & 1 \end{pmatrix},$$

where $a = \rho_2\sqrt{(1+\rho_1)/2} \in [-1,1]$. *Note that* $a = 0$ *when* $\rho_2 = 0$, *as might be expected under the conditional independence assumption. Further note that* $a = 0$ *when* $\rho_1 = -1$, *because it is not possible for* X_3 *to have the same correlation with two counter-monotonic risks, except if this correlation is* 0.

In the above example, one has $\mathrm{cov}(X_1, X_3) = \mathrm{cov}(X_2, X_3)$. This is a consequence of the conditional independence assumption which holds true in greater generality. Given any tree structure involving $d \geq 2$ individual risks, one has

$$\mathrm{cov}(X_i, X_j) = \mathrm{cov}(X_i, X_k)$$

whenever X_j and X_k are identically distributed random variables such that $A_\ell = \{j, k\}$ for some $\ell \in \{1, \ldots, d-1\}$. Thus, this assumption imposes constraints on the correlation structures that can be modelled with the proposed approach. For example, the conditional independence assumption, coupled with the requirement that only bivariate copulas are used, makes it impossible to model identically distributed risks with an autocorrelation matrix of order 1.

The following examples shed additional light into the propagation of dependence in a tree structure. For simplicity, Gaussian copulas and identically distributed risks are considered, although this is not required in the general setting.

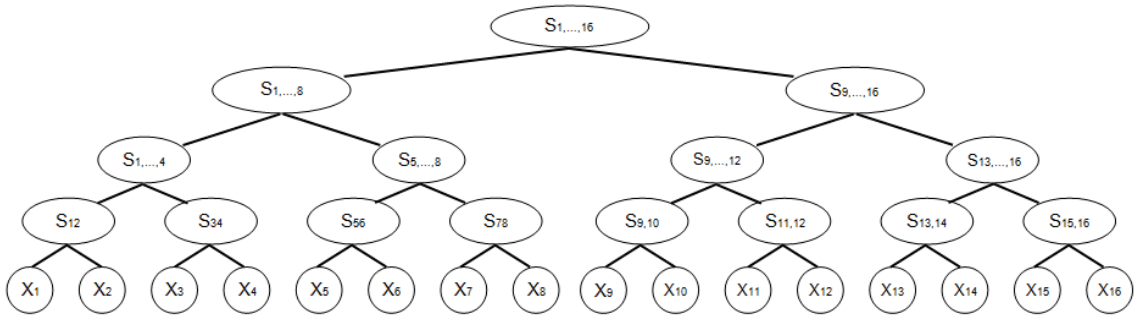

Figure 2–3: Illustration of the tree structure for Example 2.

Example 2 *Consider a perfectly symmetric tree structure involving* 2^ℓ *identically distributed risks, as illustrated in Figure 2–3 in the case* $\ell = 4$*. Suppose further that for each* $i \in \{1, \ldots, d-1\}$*, the copula* C_{A_i} *is such that* $\mathrm{corr}(S_{A_{i,1}}, S_{A_{i,2}}) = \rho \in (-1, 1)$*. In graph theoretical terms, pairs such as* (X_1, X_2)*,* (X_3, X_4) *or* (X_{15}, X_{16}) *are said to be parents of degree* 1*. Similarly, pairs such as* (X_1, X_3) *or* (X_6, X_8) *are called parents of degree* 2*, and so on. Simple calculations show that, under the conditional independence assumption, the correlation between parents of degree* $k \in \{1, \ldots, \ell\}$ *is given by*

$$\rho \left(\frac{1+\rho}{2} \right)^{k-1}.$$

Note that for fixed ρ*, this is a monotone function of* k*, which is decreasing if* $\rho > 0$ *and increasing if* $\rho < 0$*. In the limit, the correlation between parents of degree* k *tends to* 0 *as* $k \to \infty$*. This is shown on Figure 2–4.*

The next example considers a related problem for the same tree structure.

Example 3 *Consider the same symmetric tree structure as in Example 2. Suppose that* $X_1, \ldots, X_{2^\ell}$ *are identically distributed and that it is desired to model them in*

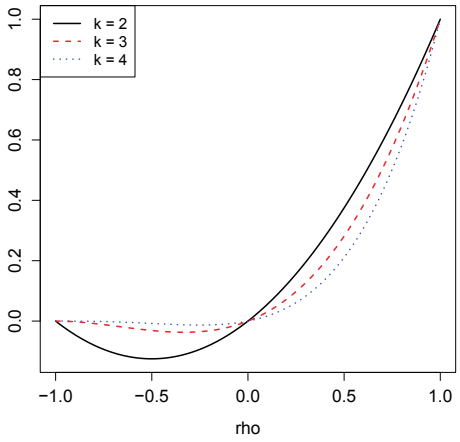

Figure 2–4: Correlation between two parents of degree k as a function of ρ, for symmetric tree structures considered in Example 2.

such a way that $\text{corr}(X_i, X_j) = \rho$ *for all distinct* $i, j \in \{1, \ldots, 2^\ell\}$. *As is well-known, this can only be done if* $\rho \geq 1/(1 - 2^\ell)$. *If the conditional independence property must hold, all the copulas linking individual risks that are parents of degree* 1 *must induce correlation* ρ. *Similarly, all copulas linking two sums of two risks must induce correlation* $2\rho/(1+\rho)$. *More generally, all copulas linking two sums of* 2^{k-1} *risks each is* $k\rho/\{1 + (k-1)\rho\}$, *where* $k \in \{1, \ldots, \ell\}$. *Thus if* $\rho > 0$, *the correlation increases with the number* k *of terms in each sum.*

While this conclusion is consistent with Example 2, it is important to realize that the tree structure plays a key role in defining the model and interpreting its parameters. This is illustrated below.

Example 4 *Consider the same problem as in Example 3, but assume that the structure is now such that variables* $S_{\{1,\ldots,k\}}$ *and* X_{k+1} *are aggregated at step* k. *It can*

then be seen that the copula at aggregation step k must be such that

$$\operatorname{corr}(S_{1,\ldots,k}, X_{k+1}) = \frac{k\rho}{\sqrt{k + 2\binom{k}{2}\rho}},$$

which converges to $\sqrt{\rho}$ as $k \to \infty$.

2.3 Building the hierarchical structure

The main advantage of the proposed model is that, given a tree structure, the selection, estimation, and validation of marginal distributions and copulas can be made using standard inferential techniques. At node N_k, the choice of the bivariate copula family linking variables $S_{A_{k,1}}$ and $S_{A_{k,2}}$ can be guided by a rank plot of the observed pairs of sums $(S_{A_{k,1}}, S_{A_{k,2}})$. Rank-based techniques can be used for parameter estimation and goodness-of-fit techniques; see, e.g., [18] or [21] for details. However, the model depends critically on the manner in which the risks are aggregated.

If $X_1, \ldots, X_d$ are $d \geq 2$ risks, the number M_d of ways in which these risks and their partial sums can be aggregated two at a time may be found recursively by setting $M_1 = 1$ and, for $k \in \{2, \ldots, d\}$,

$$M_k = \frac{1}{2} \sum_{i=1}^{k-1} \binom{d}{i} M_i M_{k-i}.$$

As shown in Table 2–1, M_d grows rapidly. Therefore, a systematic investigation of all possible tree structures is infeasible. In certain applications, a meaningful structure could perhaps be based on subjective arguments. In insurance, for example, one might consider location or line of business as a criterion for risk aggregation. When the number of risks is large, however, this would still leave the modeler with a large number of options. Therefore, a systematic method for building the tree is needed.

Table 2–1: Number of tree structures for aggregating d variables and partial sums thereof, two at a time.

d	3	4	5	6	7	8	9	10
M_d	3	15	105	945	10,395	135,135	2,027,025	34,459,425

In the absence of prior information about an adequate tree structure for dependence modelling, classical hierarchical clustering techniques can be adapted to help a user sieve through the large number of possibilities. These exploratory techniques, which are frequently used in data mining, provide a systematic way of aggregating risks (or sums thereof) two at a time based on an appropriate measure of distance between these elements; see, e.g., [26] and [31] for details on classical clustering analysis.

Consider two arbitrary risks X and Y and a measure of distance $D(X, Y)$ between them. Given a collection $X_1, \ldots, X_d$ of risks, the principle of hierarchical clustering is to identify the two risks for which D is minimal and to combine them into a group. The procedure is then repeated iteratively. In general, this requires defining a notion of distance between two clusters but this is not necessary here, because the risks are summed as one moves up the hierarchy. The elements being compared at each step of the procedure are then all of the same nature. The resulting tree structure, called a dendrogram, may depend on the choice of D.

In the present context, it is natural to measure the proximity between risks by their absolute degree of dependence. The same idea occurs in vine copula model selection, where [10] and [14] proposed to start by modelling pairs exhibiting the largest degree of association. This makes sense because in a bottom-up approach,

decisions taken early on affect the subsequent modelling stages. It is thus crucial to get strong dependencies right from the start.

Clustering works best when D is a (pseudo) metric, i.e., if for all X, Y and Z, $D(X,X) = 0$, $D(X,Y) = D(Y,X)$ and $D(X,Y) \leq D(X,Z) + D(Z,Y)$. From the work of [45], these conditions hold when

$$D(X,Y) = \sqrt{1 - r^2(X,Y)} \quad \text{or} \quad D(X,Y) = \sqrt{1 - \rho^2(X,Y)},$$

where $r(X,Y)$ and $\rho(X,Y)$ denote Pearson's correlation and Spearman's rank correlation between X and Y, respectively. Another option would be to define a metric based on Kendall's tau. Although τ itself has been used as a similarity measure for hierarchical clustering (see, e.g., [24] and [30]), it has apparently never been shown that $\sqrt{1 - \tau^2(X,Y)}$ is indeed a metric. This result is formally stated below and proved in the Appendix.

Proposition 2 *The map $D(X,Y) = \sqrt{1 - \tau^2(X,Y)}$ defines a (pseudo) metric.*

The following example illustrates the application of the hierarchical clustering algorithm in a classical case.

Example 5 *Consider a random sample of size n = 10,000 from the multivariate Normal random vector (X_1, X_2, X_3, X_4) with mean $\mathbf{0}$ and covariance matrix*

$$R_1 = \begin{pmatrix} 1 & 0.5 & 0.2 & 0.2 \\ 0.5 & 1 & 0.2 & 0.2 \\ 0.2 & 0.2 & 1 & 0.3 \\ 0.2 & 0.2 & 0.3 & 1 \end{pmatrix},$$

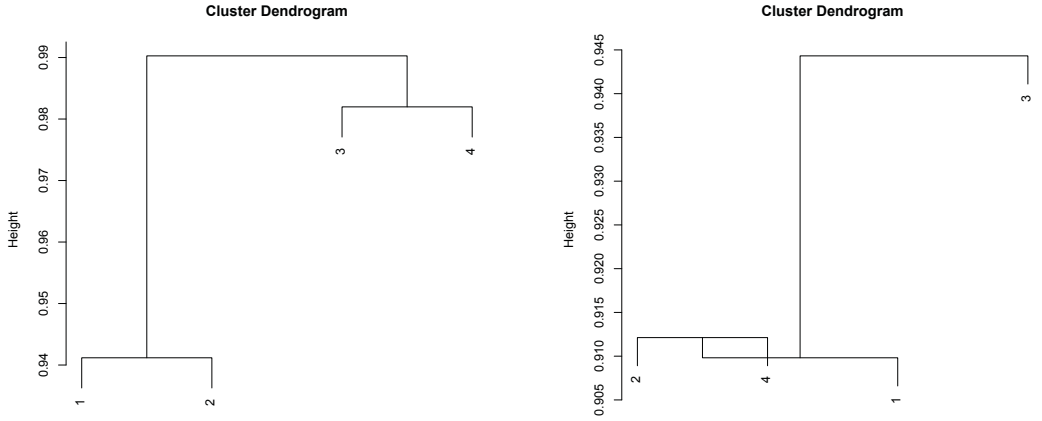

Figure 2–5: Dendrogram for Example 5 with correlation matrix R_1 (left) and R_2 (right).

so that $\mathrm{corr}(X_1+X_2, X_3+X_4) \approx 0.286$. *This multivariate model can be expressed as a hierarchical structure of the form* $(X_1+X_2)+(X_3+X_4)$, *with standard Normal margins and Gaussian copulas at each node. In addition, the conditional independence assumption is satisfied.*

An application of the hierarchical clustering algorithm based on the Kendall tau metric is shown in the left panel of Figure 2–5. As can be seen, the dependence structure of the multivariate distribution is reproduced faithfully. As the correlation between X_1 *and* X_2 *is the largest, they are aggregated first;* X_3 *and* X_4 *are aggregated at step 2, and the two clusters* $S_{\{1,2\}}$ *and* $S_{\{3,4\}}$ *are summed at step 3. The scale displayed on the graph represents the empirical distance* $\sqrt{1-\tau^2(X,Y)}$ *at which* X *and* Y *are aggregated. It matches very closely the theoretical value of* $D(X,Y)$ *which can be computed using the relation* $\tau(X,Y) = 2\arcsin\{r(X,Y)\}/\pi$ *valid for bivariate Normal pairs* (X,Y). *Dendrograms based on the Pearson and Spearman metrics (not shown) are identical except for the scale.*

Now suppose that the correlation matrix is

$$R_2 = \begin{pmatrix} 1 & 0.5 & 0.6 & 0.6 \\ 0.5 & 1 & 0.6 & 0.6 \\ 0.6 & 0.6 & 1 & 0.3 \\ 0.6 & 0.6 & 0.3 & 1 \end{pmatrix},$$

so that $\mathrm{corr}(X_1 + X_2, X_3 + X_4) \approx 0.859$. *The corresponding dendrogram derived from the hierarchical clustering algorithm based on the Kendall tau metric is shown in the right panel of Figure 2–5. In this particular sample, the largest observed value of Kendall's tau is* $\tau_n(X_2, X_4)$, *and hence* X_2 *and* X_4 *are aggregated first. In the second step,* X_1 *is aggregated with* $X_2 + X_4$. *The curious shape of the tree reflects the fact that* $\tau_n(X_2 + X_4, X_1) > \tau_n(X_2, X_4)$. *Finally,* X_3 *is aggregated with* $X_1 + X_2 + X_4$ *in the last step.*

In the second example, therefore, the decision to aggregate the risks in decreasing order of dependence has failed to reproduce the underlying hierarchical structure, even though the latter satisfied the conditional independence assumption. Whether the association is measured using Kendall's tau, Spearman's rho or even Pearson's correlation does not affect this conclusion. The need for model validation techniques is thus obvious; this will be addressed in Section 2.5.

2.4 Other aspects of modelling

Once a tree structure has been chosen, the construction of the hierarchical model involves the selection of a copula at each node, in addition to distributions for the individual risks. Both steps can be carried out in a straightforward manner with

classical inference techniques. In this section, a stylized example is used to show how rank-based methods can help to guide the choice of copula families.

A random vector $\mathbf{X} = (X_1, \ldots, X_d)$ is said to have a Liouville distribution if it can be expressed in the form $\mathbf{X} = R\mathbf{D}$ in terms of a non-negative random variable R which is independent of a d-variate random vector $\mathbf{D}$ having a Dirichlet distribution with parameter $(\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ on the unit simplex

$$\mathcal{S} = \{(s_1, \ldots, s_d) \in [0,1]^d : s_1 + \cdots + s_d = 1\}.$$

As mentioned by [38], if $\mathbf{X}$ has a Liouville distribution, then so does any subset of its components. Furthermore, any vector consisting of sums of distinct components of $\mathbf{X}$ is also of the Liouville type.

In particular, set $\alpha = \alpha_1 + \cdots + \alpha_d \geq 2$ and let F denote the distribution function of R. The underlying copula of the vector $\mathbf{X}$ is then said to be of the Clayton–Liouville type if the Williamson α-transform of F is given, for all $t \geq 0$, by $\psi(t) = \{\max(0, 1 + \theta t)\}^{-1/\theta}$ for some $\theta \geq -1/(\alpha - 1)$. This terminology, introduced by [38], is motivated by the fact that ψ is the generator of a Clayton Archimedean copula with parameter θ.

A sample of size 1000 from the Clayton–Liouville distribution with parameters $\alpha = (4, 3, 2, 1)$ and $\theta = 3$ was generated using the algorithm provided by [38].[1] An

[1] Thanks to Dr. Johanna Nešlehová for providing me with R programs for generating data from the Clayton–Liouville copula

application of the clustering procedure based on Kendall's tau led to the tree structure depicted in Figure 2–6. In order to select copulas $C_{\{1,2\}}$, $C_{\{1,2,3\}}$ and $C_{\{1,2,3,4\}}$, rank plots of the pairs (X_1, X_2), $(X_1 + X_2, X_3)$ and $(X_1 + X_2 + X_3, X_4)$ were drawn. These graphs are displayed in the top row of Figure 2–7. Visual inspection of these plots reveals strong upper tail dependence, as well as a moderate degree of lower tail dependence, at least for the pairs (X_1, X_2) and $(X_1 + X_2, X_3)$. In addition, the rank-based test of [22] reveals significant asymmetry in the copula of the pair $(X_1 + X_2 + X_3, X_4)$ $(p \ll .001)$.

For these reasons, common choices of bivariate Archimedean, elliptical, and extreme-value copulas are ruled out. However, the Clayton–Liouville survival copula family does look like a reasonable option in all cases. In fact, the theoretical survival copula of each of the three pairs is indeed from the Clayton–Liouville family with parameters as specified in Table 2–2. In practice, of course, the user would not know this and might consider other modelling options.

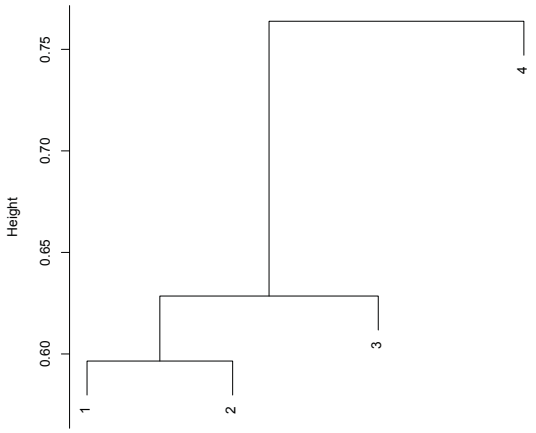

Figure 2–6: Dendrogram for the sample from the Clayton–Liouville distribution.

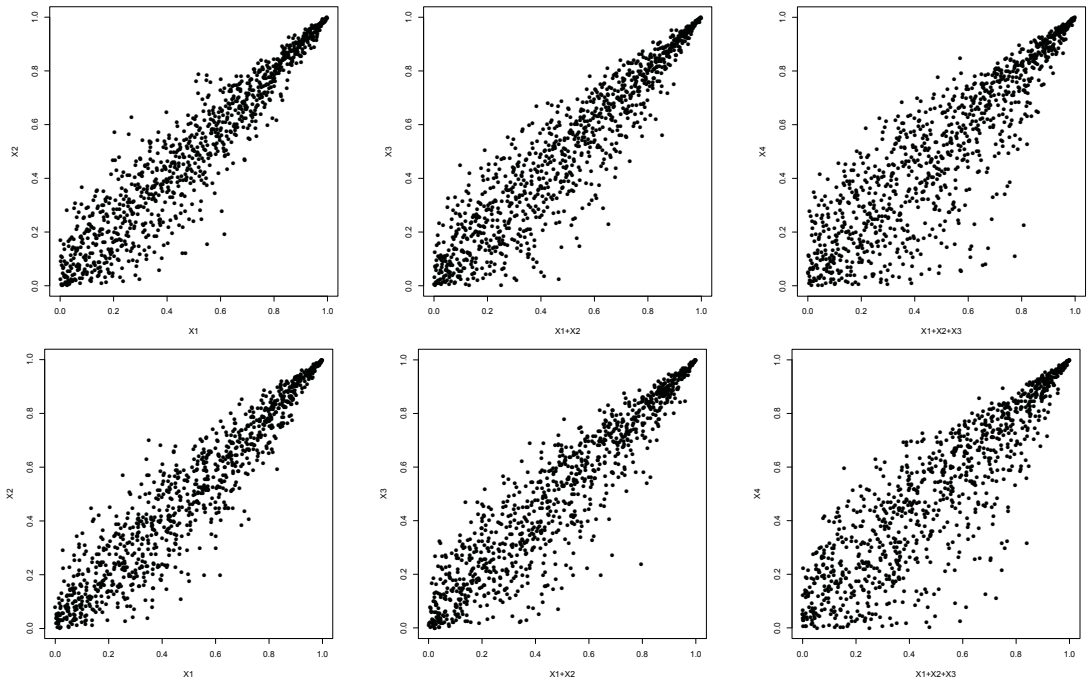

Figure 2–7: Rank plots for the observed pairs of variables (top row) and simulated samples from the selected copula (bottom row) at each aggregation step.

Suppose that a Clayton–Liouville survival copula indexed by parameters $(\alpha_1, \alpha_2) \in \mathbb{N}^2$ and $\theta > 0$ is fitted to each of the three pairs. Expressions for this copula density and Kendall's tau are given in [38]. For each pair, a rank-based estimate $\hat{\theta} = \hat{\theta}(\alpha_1, \alpha_2)$ of θ is obtained by inverting Kendall's tau for fixed values of α_1, α_2 over a suitably large grid. The pseudo-likelihood is then computed at each combinations of $(\alpha_1, \alpha_2, \hat{\theta})$, and the vector of parameters that maximize this pseudo-likelihood is selected.[2]

[2] The idea of quickly estimating θ by inverting Kendall's tau was from Léo Belzile, who kindly shared his findings and his R code with me.

Table 2–2: Theoretical and estimated parameter values for three Clayton–Liouville copulas.

	α_1	α_2	θ	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\theta}$	Bootstrap 95% C.I. on θ
C_{12}	4	3	3	6	5	2.32	$(1.61, 3.36)$
C_{123}	7	2	3	10	3	2.22	$(1.53, 3.21)$
C_{1234}	9	1	3	6	1	3.15	$(2.12, 4.50)$

The resulting estimates are given in Table 2–2 along with a 95% bootstrap confidence interval for $\theta = \theta(\hat{\alpha}_1, \hat{\alpha}_2)$ based on 10,000 replications. The estimates are close to their theoretical values, especially considering that α_1 and α_2 are integer-valued. The discrepancies result in part from the fact that the pseudo-likelihood is quite flat, as illustrated in Figure 2–8 for the first aggregation step. As a corollary, two distributions with slightly different parameters are almost identical. This is illustrated in Figure 2–7, where rank plots of the three pairs of variables (top row) are compared with scatter plots of random samples of the same size from the fitted survival copulas.

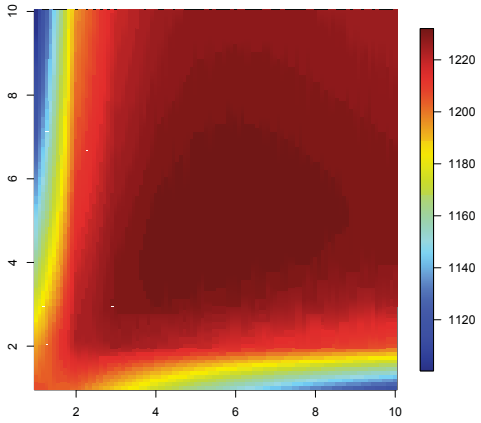

Figure 2–8: Interpolation of the log pseudo-likelihood $\ell(\alpha_1, \alpha_2, \hat{\theta})$ as a function of $\alpha_1, \alpha_2 \in \{1, \ldots, 10\}$ for the survival copula of (X_1, X_2).

At this stage, the components of the hierarchical model appear to be valid. However, the question remains as to whether the overall structure faithfully represents the joint distribution of the risks $X_1, \ldots, X_4$. This validation issue is addressed next.

2.5 Simulation and validation

Although the approach described above leads to appropriate choices of copulas for the pairs that are modelled explicitly, there is no guarantee that the multivariate distribution induced by the construction provides an adequate fit overall. This is because there is no provision in this iterative procedure for checking the conditional independence assumption, which is critical to the validity of the model.

An intuitive way in which this assumption can be validated consists of checking the extent to which the model mimics the distribution of pairs that were not considered explicitly. To this end, one must first know how to simulate observations from the hierarchical model. The following procedure, inspired by the work of [29], can be used to this end.

Algorithm 1 *To generate a sample of size $n \geq 2$ from a hierarchical copula model with tree structure $A_1, \ldots, A_{d-1}$, copulas $C_{A_1}, \ldots, C_{A_{d-1}}$ and marginal distributions $F_1, \ldots, F_d$, choose an integer $m \gg n$ and proceed as follows:*

1. *For each $i \in \{1, \ldots, d\}$, generate a sample of size m from distribution F_i and denote by $\mathbf{X} = (x_{ik})$ the $d \times m$ matrix whose ith row is the sample from F_i.*
2. *For each $j \in \{1, \ldots, d-1\}$,*
   a) *Generate a random sample of size m from copula C_{A_j} and label the resulting pairs $(U_{j1}, V_{j1}), \ldots, (U_{jm}, V_{jm})$ in such a way that $U_{j1} < \cdots < U_{jm}$.*

b) Denote by $\pi_{C_{A_j}}$ *the permutation defined, for each* $k \in \{1, \ldots, m\}$*, by setting* $\pi_{C_{A_j}}(k) = p$ *if and only if* V_{jp} *has rank* k *among* $V_{j1}, \ldots, V_{jm}$*.*

c) For $\ell = 1, 2$*, let* $\pi_{S_{A_{j,\ell}}}$ *be the permutation of the vector* $(1, \ldots, m)$ *induced by the reordering of* $S_{A_{j,\ell}}$ *in ascending order. Formally,* $\pi_{S_{A_{j,\ell}}}$ *is the permutation defined, for each* $p \in \{1, \ldots, m\}$*, by setting* $\pi_{S_{A_{j,\ell}}}(p) = k$ *if and only if* $S_{A_{j,\ell},p}$ *has rank* k *among* $S_{A_{j,\ell},1}, \ldots, S_{A_{j,\ell},m}$*.*

3. For any $j \in \{1, \ldots, d-1\}$ *and* $\ell = 1, 2$*, let* $\pi_{S_{A_{j,\ell}}}(\mathbf{X})$ *and* $\pi_{C_{A_j}}(\mathbf{X})$ *be the* $d \times m$ *matrices whose* (i, k)*th entry is given, respectively, by*

$$[\pi_{S_{A_{j,\ell}}}(\mathbf{X})]_{ik} = \begin{cases} x_{i\pi_{S_{A_{j,\ell}}}(k)} & \text{if } i \in A_{j,\ell}, \\ x_{ik} & \text{if } i \notin A_{j,\ell}, \end{cases}$$

and

$$[\pi_{C_{A_j}}(\mathbf{X})]_{ik} = \begin{cases} x_{i\pi_{C_{A_j}}(k)} & \text{if } i \in A_{j,2}, \\ x_{ik} & \text{if } i \notin A_{j,2}. \end{cases}$$

4. Set $\mathbf{X}_0 = \mathbf{X}$ *and for each* $j \in \{1, \ldots, d-1\}$*, compute iteratively*

$$\mathbf{X}_j = \pi_{C_{A_j}} \circ \pi_{S_{A_{j,2}}} \circ \pi_{S_{A_{j,1}}}(\mathbf{X}_{j-1}).$$

The required sample is any randomly selected collection of n *columns from the* $d \times m$ *matrix* $\mathbf{X}_{d-1}$*.*

To understand heuristically why this procedure works, consider the bivariate case. Let C be the unique copula linking continuous random variables X and Y with marginal distributions F and G, respectively. Let F_m and G_m denote empirical counterparts of F and G based on random samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_m$, respectively. Further assume that $(U_1, V_1), \ldots, (U_m, V_m)$ is a random sample from C

and let $(\hat{U}_1, \hat{V}_1), \ldots, (\hat{U}_m, \hat{V}_m)$ be the corresponding pairs of standardized ranks which form the support of the empirical copula C_m. The sample of size m generated by Algorithm 1 then consists of the pairs defined, for each $i \in \{1, \ldots, m\}$, by

$$(\hat{X}_i, \hat{Y}_i) = \{F_m^{-1}(\hat{U}_i), G_m^{-1}(\hat{V}_i)\}.$$

It is intuitively plausible that for large m, the pairs $(\hat{X}_1, \hat{Y}_1), \ldots, (\hat{X}_m, \hat{Y}_m)$ would resemble a sample from the joint distribution $H = C(F, G)$ because F_m^{-1}, G_m^{-1}, and C_m are consistent estimators of F^{-1}, G^{-1}, and C, respectively; see, e.g., Chapter 21 in [44] and [43]. For a more formal (but partial) argument, see the proof of Theorem 3.4 [2] and the surrounding discussion. It goes without saying that if $(\hat{X}_1, \hat{Y}_1), \ldots, (\hat{X}_m, \hat{Y}_m)$ is a sample from the model, then so is any randomly selected subset of size n thereof.

Algorithm 1 is very quick and easy to implement regardless of the complexity of the tree structure. The following example illustrates this procedure in the simple but unrealistic case where, for pedagogical purposes, $m = n = 4$.

Example 6 *Consider the symmetric structure* $(X_1 + X_2) + (X_3 + X_4)$. *Assume that the ordered observations from the marginals are given by*

$$\mathbf{X}_0 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 10 & 20 & 30 & 40 \\ 100 & 200 & 300 & 400 \\ 1000 & 2000 & 3000 & 4000 \end{pmatrix}.$$

Further assume that the observed pairs of ranks from the copulas are as follows:

$$\begin{aligned} C_{\{1,2\}}: &\quad (1,4), \quad (2,2), \quad (3,1), \quad (4,3), \\ C_{\{3,4\}}: &\quad (1,2), \quad (2,1), \quad (3,4), \quad (4,3), \\ C_{\{1,2,3,4\}}: &\quad (1,3), \quad (2,4), \quad (3,2), \quad (4,1). \end{aligned}$$

At Step 1, $\pi_{S_{\{1\}}}(i) = \pi_{S_{\{2\}}}(i) = i$ *for* $i \in \{1, \ldots, 4\}$ *because the observations are already ordered. Also,* $\pi_{C_{\{1,2\}}}$ *permutes* $(1,2,3,4)$ *into* $(3,2,4,1)$*. Hence*

$$\mathbf{X}_1 = \pi_{C_{\{1,2\}}}(\mathbf{X}_0) = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 40 & 20 & 10 & 30 \\ 100 & 200 & 300 & 400 \\ 1000 & 2000 & 3000 & 4000 \end{pmatrix}$$

and $S_{\{1,2\}} = (41, 22, 13, 34)$*. At Step 2,* $\pi_{S_{\{3\}}}(i) = \pi_{S_{\{4\}}}(i) = i$ *for* $i \in \{1, \ldots, 4\}$*, again because the observations are ordered. Furthermore,* $\pi_{C_{\{3,4\}}}$ *permutes* $(1,2,3,4)$ *into* $(2,1,4,3)$*. Hence*

$$\mathbf{X}_2 = \pi_{C_{\{3,4\}}}(\mathbf{X}_1) = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 40 & 20 & 10 & 30 \\ 100 & 200 & 300 & 400 \\ 2000 & 1000 & 4000 & 3000 \end{pmatrix}$$

and $S_{\{3,4\}} = (2100, 1200, 4300, 3400)$*. At Step 3,* $\pi_{S_{\{1,2\}}}$ *permutes* $(1,2,3,4)$ *into* $(4,2,1,3)$ *and* $\pi_{S_{\{3,4\}}}$ *permutes* $(1,2,3,4)$ *into* $(2,1,4,3)$*. Finally,* $\pi_{C_{\{1,2,3,4\}}}$ *permutes*

$(1,2,3,4)$ *into* $(4,3,1,2)$. *Thus,*

$$\begin{aligned}\mathbf{X}_3 &= \pi_{C_{\{1,2,3,4\}}} \circ \pi_{S_{\{3,4\}}} \circ \pi_{S_{\{1,2\}}}(\mathbf{X}_2) \\ &= \pi_{C_{\{1,2,3,4\}}} \begin{pmatrix} 3 & 2 & 4 & 1 \\ 10 & 20 & 30 & 40 \\ 200 & 100 & 400 & 300 \\ 1000 & 2000 & 3000 & 4000 \end{pmatrix} = \begin{pmatrix} 3 & 2 & 4 & 1 \\ 10 & 20 & 30 & 40 \\ 400 & 300 & 100 & 200 \\ 3000 & 4000 & 2000 & 1000 \end{pmatrix}.\end{aligned}$$

A simple way in which Algorithm 1 can be used to test the validity of a given hierarchical model is as follows:

Algorithm 2 *For some large integer m, carry out the following steps:*

1. *Generate a sample of size m from the proposed model using Algorithm 1.*
2. *Compute the empirical copula C_m^* associated with this sample.*
3. *Determine how close C_m^* is to the empirical copula C_n of the original sample by computing the rank-based Cramér–von Mises statistic*

$$T_{m,n} = \frac{mn}{m+n} \int\!\!\cdots\!\!\int_{[0,1]^d} \{C_m^*(u_1,\ldots,u_d) - C_n(u_1,\ldots,u_d)\}^2 \mathrm{d}u_1 \cdots \mathrm{d}u_d.$$

Assuming that Algorithm 1 is valid in full generality, then for large enough m, C_m^* is a consistent estimator of the copula C^* induced by the model. Furthermore, C_n is known to be a consistent estimator of the true underlying copula C of the risk vector. Therefore, large values of $T_{m,n}$ should lead to the rejection of the null hypothesis $\mathcal{H}_0 : C^* = C$. The limiting null distribution of the statistic $T_{m,n}$ is given by [42], who also show how to find an approximate p-value for the test using the Multiplier Central Limit Theorem. Their procedure is coded in the R package

`TwoCop`. For the artificial data in Section 2.4, for example, one finds $T_{n,n} = 0.006$ and the corresponding p-value is 80%, which comforts the choice of model.

As an additional reality check, rank plots of pairs (X_i, X_j) that were not modelled explicitly can be compared visually to the corresponding graphs for model-generated data. This is illustrated in Figure 2–9, where rank plots of three selected pairs for the data in Section 2.4 (top row) are compared with scatter plots of random samples of the same size generated from the fitted model (bottom row). The similarity between the two samples is striking. Had the statistic $T_{m,n}$ led to the rejection of $\mathcal{H}_0$, Algorithm 2 could have been used iteratively on sub-vectors to identify the problematic spots. In so doing, however, one would need to adjust the p-values to account for multiple testing. This issue will be addressed in future work, along with the validity of Algorithm 1 in its most general setting and the development of a formal test of the conditional independence assumption.

2.6 Application to insurance portfolio modelling

This work was originally motivated by a capital allocation problem for a portfolio of insurance risks held by a large Canadian insurance company. The data available consist of 102 monthly earned premiums and incurred claim amounts for eight different risks from January 2004 to June 2012, inclusively. Monthly claim development is known up to August 2013.

Table 2–3 gives a summary description of the risks. Home insurance in Québec covers fire, theft and other property damage. Québec auto insurance mainly covers loss caused by car damage, as well as property damage to third party. All Ontario portfolios pertain to automobile insurance coverages. Third Party Liability (TPL)

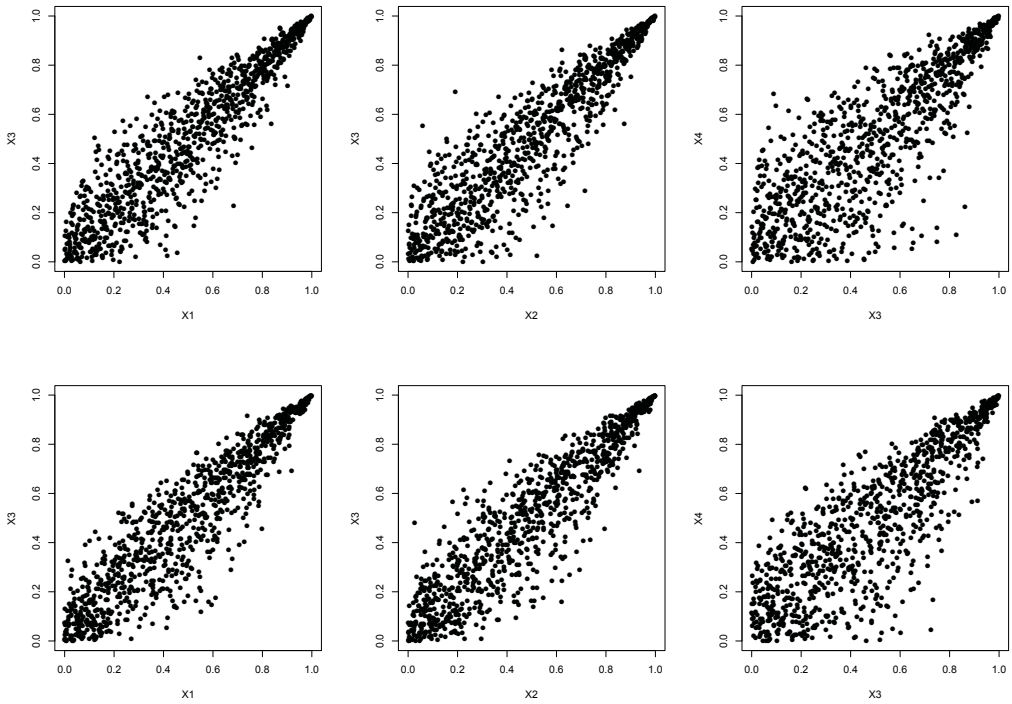

Figure 2–9: Graphs of pairs of random variables that were not explicitly modelled.

covers financial loss due to body injuries caused to others. Accident Benefits (AB) cover the insured's own body injuries up to a specified maximum; the remaining amount falls under the TPL coverage of the responsible party.

In order to exclude the effect of business growth and inflation, loss ratios can be used in modelling purposes provided that pricing is consistent over the study period. The loss ratio for risk i in month t, denoted $LR_{i,t}$, is defined as the ultimate claim amount for accident month t divided by the earned premiums for this month. The development of the claim amounts to ultimate was done using the standard Chain Ladder Method; see, e.g., [34].

Table 2–3: Summary description of eight risks for a Canadian insurance company.

Risk	Subsidiary	Province	Coverage
1	A	Québec	Automobile Insurance
2	B	Québec	Automobile Insurance
3	A	Québec	Home Insurance
4	B	Québec	Home Insurance
5	C	Ontario	Accident Benefits
6	C	Ontario	Third Party Liability
7	D	Ontario	Accident Benefits
8	D	Ontario	Third Party Liability

2.6.1 Choice of marginal distributions for the risks

The model construction first involves fitting univariate distributions to the individual risks. To this end, standard tests of randomness were first applied to all series. Risks 3, 4, 6, and 8 exhibit no trend, seasonality or serial dependence, as confirmed by Ljung–Box tests whose p-values are reported in Table 2–4. Skewed t distributions with negligible mass ($< 10^{-8}$) on $(-\infty, 0)$ were used to model both Risks 3 and 4; Gamma distributions led to a better fit for Risks 6 and 8. Table 2–5 provides estimates based on the parametrization of these distributions given by [37].

Table 2–4: Results of the Ljung–Box test for randomness for Risks 1–8.

Series	Statistic	p-value
Y_1	17.571	6.3%
Y_2	65.921	$< .01\%$
LR_3	13.327	20.6%
LR_4	17.445	6.5%
Y_5	17.267	6.9%
LR_6	11.886	29.3%
Y_7	10.855	36.9%
LR_8	12.456	25.6%

Table 2–5: Selected marginal models for Risks 1–8.

Risk	Trend or Seasonality	Time Series	Distribution
LR_1	Seasonality	White Noise	$t_5(\hat{\mu} = 0.584, \hat{\sigma} = 0.060, \hat{\gamma} = 0.034)$
LR_2	Seasonality	ARMA$(1,1)$ $0.935, 0.780$	$t_5(\hat{\mu} = -0.031, \hat{\sigma} = 0.048, \hat{\gamma} = 0.030)$
LR_3	None	White Noise	$t_8(\hat{\mu} = 0.101, \hat{\sigma} = 0.002, \hat{\gamma} = 0.589)$
LR_4	None	White Noise	$t_{16}(\hat{\mu} = 0.046, \hat{\sigma} = 0.029, \hat{\gamma} = 0.555)$
LR^*_{5b}	Trend	White Noise	$\ln(LR_{5,t}) = -0.936 + 0.012t + Y_{5,t}$, $Y_{5,t} \sim \mathcal{N}(0, 0.389^2)$
LR_{5a}	None	White Noise	$\mathcal{LN}(\hat{\mu} = -0.782, \hat{\sigma} = 0.389)$
LR_6	None	White Noise	$\mathcal{G}(\hat{\alpha} = 3.41, \hat{\beta} = 4.67)$
LR_{7b}	None	White Noise	$\mathcal{LN}(\hat{\mu} = -0.309, \hat{\sigma} = 0.319)$
LR_{7a}	None	White Noise	$\mathcal{LN}(\hat{\mu} = -1.070, \hat{\sigma} = 0.319)$
LR_8	None	White Noise	$\mathcal{G}(\hat{\alpha} = 7.86, \hat{\beta} = 12.96)$

*The indices b and a refer to the series before and after the change point, respectively.

Models for Risks 1, 2, 5, and 7 are more complex. The first two exhibit significant seasonality, as might be expected given the large weather variations that affect Québec winter road conditions. It was thus assumed that, for $\ell = 1, 2$,

$$LR_{\ell,t} = s_{\ell,t} + Y_{\ell,t},$$

where $s_{\ell,t} = s_{\ell,t+12}$ is a seasonal component with $s_{\ell,1} + \cdots + s_{\ell,12} = 0$. The seasonal component estimates shown in Table 2–6 were derived using classical methods ([7]). In computing them, the large loss ratio observed in June 2008 caused by the hail storm of June 10th was treated as an outlier.

Table 2–6: Seasonal components for Québec automobile insurance coverages.

t	1	2	3	4	5	6
$\hat{s}_{1,t}$	0.148	0.068	0.013	−0.100	−0.079	−0.042
$\hat{s}_{2,t}$	0.117	0.049	0.003	−0.089	−0.083	−0.046
t	7	8	9	10	11	12
$\hat{s}_{1,t}$	−0.065	−0.039	−0.084	−0.030	0.031	0.180
$\hat{s}_{2,t}$	−0.039	−0.019	−0.068	−0.009	0.033	0.151

An application of the Ljung–Box test to Y_1 suggests (Table 2–4) that upon deseasonalization, Risk 1 reduces to white noise; a skewed t distribution provided an excellent fit (Table 2–5). However, serial correlation is present in Y_2. This serial dependence can be accounted for by an ARMA$(1, 1)$ model whose residuals can also be modelled using a skewed t distribution. The parameter estimates of the ARMA model given in Table 2–5 are those that maximize the Gaussian likelihood; see [6, pp. 248–249] for a justification.

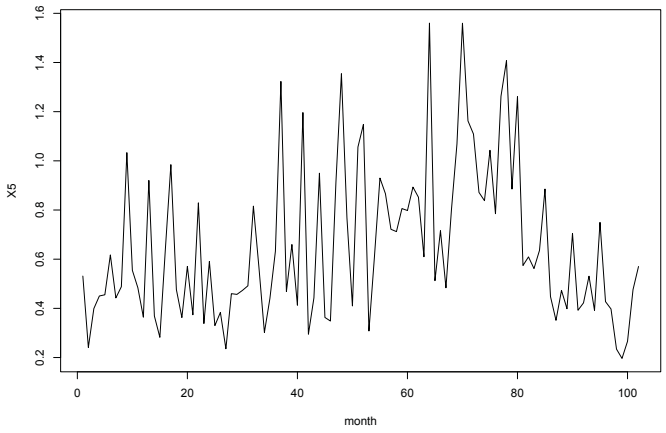

Figure 2–10: Time series data for LR_5.

The time series for Risks 5 and 7 are affected by a structural change point. The AB coverages in Ontario went through a period of high inflation in 2009–10, as reflected by the increasing trend in LR_5 depicted in Figure 2–10. This led to a reform of the Ontario Insurance Act effective September 1, 2010. Detailed information on the changes are available on the website of the Financial Services Commission of Ontario ([16]). This reform did not affect TPL coverages either directly or indirectly, as data inspection confirms. Prior to the reform, a discernible trend in Risk 5 can be modelled adequately by setting

$$\ln(LR_{5,t}) = -0.936 + 0.012\,t + Y_{5,t},$$

where $Y_{5,t} \sim \mathcal{N}(0, 0.394^2)$. The overall fit of the model is good ($R^2 = 0.32$) and the parameter estimates reported in Table 2–5 are highly significant. After the reform, the time series is a white noise to which a lognormal distribution provides an excellent fit; its variance appears to be the same as that of $Y_{5,1}, \ldots, Y_{5,80}$ ($F = 1.66$, $p = 9.44\%$), and hence a pooled variance estimate can be used, viz. $\hat{\sigma}_p^2 = 0.79\,\hat{\sigma}_1^2 + 0.21\,\hat{\sigma}_2^2 = 0.389$. Finally, adequate models for Risk 7 before (b) and after (a) the reform are given by

$$LR_{7b} \sim \mathcal{LN}(-0.309, 0.313), \quad LR_{7a} \sim \mathcal{LN}(-1.070, 0.342),$$

respectively. The centered samples, denoted Y_7, can be assumed to have the same variance ($F = 0.837$, $p = 72.11\%$) and the pooled variance estimate is 0.319.

2.6.2 Model construction

Of greatest interest to the insurance company is a predictive model for the total claim amount S_t associated with the entire portfolio at time t. This quantity, which

Table 2–7: Proportions of premiums, multiplied by 1000.

i	1	2	3	4	5	6	7	8
Q_i	100	250	90	230	70	65	90	105

is typically expressed in dollars per \$1000 premium, is given by

$$S_t = \sum_{i=1}^{8} Q_{i,t} LR_{i,t},$$

where, for $i \in \{1, \ldots, 8\}$, $Q_{i,t}/1000$ is the proportion of premiums for sub-portfolio i at time t. It was assumed that the portfolio distribution is stable over time, i.e., for all $t \in \mathbb{N}$, $Q_{i,t} \equiv Q_i$, with $Q_1, \ldots, Q_8$ as given in Table 2–7.

Given the choice of marginals for the risks, a model for S_t can be written as

$$S_t = Q_1 \hat{s}_{1,t} + Q_2(0.551 + \hat{s}_{2,t} + 0.935 Y_{2,t-1} - 0.779 X_{2,t-1}) + \sum_{i=1}^{8} Q_i X_i,$$

where $X_1 = Y_1$, X_2 is the innovation of the ARMA$(1,1)$ model on Y_2, and $X_i = LR_i$ for $i \in \{3, \ldots, 8\}$. In order to account for the possible dependence between $Q_1 X_1, \ldots, Q_8 X_8$, a copula-based hierarchical approach was used. Considering the effect on the marginals of the legislative reform of September 1, 2010, it was deemed preferable to construct separate dendrograms for the data before and after that date. These dendrograms are displayed in Figure 2–11, where one can clearly see increased dependence within the AB and TPL coverages for the two Ontarian subsidiaries after the reform. The dependence between the Québec risks is unaffected by this reform and generally higher than between Ontario risks.

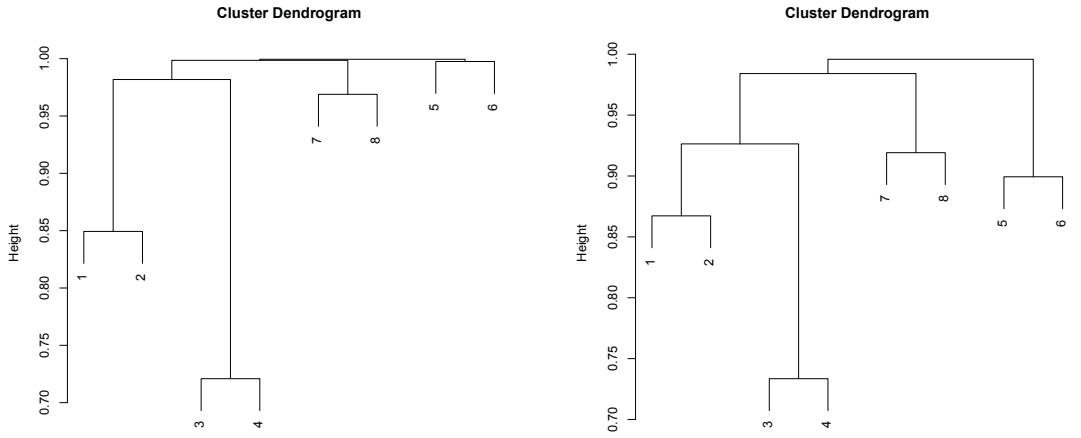

Figure 2–11: Hierarchical structures for $Q_1X_1, \ldots, Q_8X_8$ before (left panel) and after (right panel) the legal reform of September 2010.

2.6.3 Choice of copulas

Based on Figure 2–11, it was resolved to fit nine different copulas, i.e.,

a) copulas $C_{\{3,4\}}$, $C_{\{1,2\}}$ and $C_{\{1,2,3,4\}}$ pertaining to Québec risks, which can be estimated using the entire sample;

b) copulas $C_{\{7,8\}b,a}$ and $C_{\{5,6\}b,a}$ for the Ontario risks, where the indices b and a refer to before and after the legal reform.

c) the product copula for the last two aggregation steps, as independence between Québec and Ontario risks cannot be rejected at any reasonable level.

In order to guide the choice of families for steps a) and b), tests of independence, symmetry, and extremeness were carried out; see [18], [25], and [22] for descriptions of various rank-based procedures that can be used to this end. Based on rank plots and the results of these tests, parametric families of bivariate copulas were selected and fitted by maximum pseudo-likelihood ([20]). The final choices are summarized

in Table 2–8. Details are provided in Chapter 3 of this thesis. It is hard to distinguish between copula families based on the after-reform sample, which has only 22 observations; for this reason, the same copula families were used for $C_{\{7,8\}b}$ and $C_{\{7,8\}a}$, as well as for $C_{\{5,6\}b}$ and $C_{\{5,6\}a}$.

As a final validation step, random samples of sizes 80 and 22 were generated using Algorithm 1 for the overall models before and after the reform. These samples were then compared to the original data with the test statistic of [42]. The null hypothesis that the two samples are coming from the same copula cannot be rejected in either case ($p = 0.51$ and 0.69, respectively).

2.6.4 Risk measures and capital allocation

The model described above is useful for risk measurement and capital allocation. Insurance companies often use the Value-at-Risk (VaR) or the Tail Value-at-Risk (TVaR) for this purpose; both of them are defined in terms of a risk tolerance level $\kappa \in (0, 1)$. Consistent estimates of these quantities can be derived easily from n independent copies $(X_{11}, \ldots, X_{d1}), \ldots, (X_{1n}, \ldots, X_{dn})$ of $(X_1, \ldots, X_d)$ generated

Table 2–8: Copula family and parameter estimates for each aggregation step.

Step	Copula	Parameter	Standard Deviation
$C_{\{3,4\}}$	Galambos	2.429	0.353
$C_{\{1,2\}}$	t_8	0.755	0.058
$C_{\{1,2,3,4\}}$	tev_{10}	0.783	0.072
$C_{\{7,8\}b}$	t_{10}	0.390	0.108
$C_{\{7,8\}a}$	t_{10}	0.637	0.140
$C_{\{5,6\}b}$	Product		
$C_{\{5,6\}a}$	Gaussian	0.700	0.171
$C_{\{1,2,3,4,7,8\}}$	Product		
$C_{\{1,\ldots,8\}}$	Product		

from the model. Denote by $S_1, \ldots, S_n$ the simulated values of $S = X_1 + \cdots + X_d$ and let F_n be their empirical distribution function. Then

$$\widehat{\mathrm{VaR}}_\kappa(S) = \inf\{s \in \mathbb{R} | F_n(s) \geq \kappa\} = s_\kappa$$

and

$$\widehat{\mathrm{TVaR}}_\kappa(S) = \frac{1}{1-\kappa}\left[\frac{1}{n}\sum_{j=1}^{n} S_j \mathbf{1}(S_j > s_\kappa) + s_\kappa\{F_n(s_\kappa) - \kappa\}\right].$$

In order to assess the company's exposure to risk for the period extending from September 2010 to June 2012, it suffices to generate a large random sample from the model using Algorithm 1, assuming that for all $i \in \{1, \ldots, 8\}$, Q_i dollars in monthly premiums are invested in risk i. One can then compute various risk measures using the sum of claims over that period. Such estimates are given in Table 2–9 at level $\kappa = 0.95$ for two different random samples of size 10,000 from the model. In Scenario I, the data were generated from the model as it stood just before the reform of September 2010. In Scenario II, the model fitted to data after the reform was used. In this fashion, it is possible to appreciate the effect of the introduction of limits on AB claims in Ontario in September 2010.

From Table 2–9, one can see that before the reform, the projected average claim amount for Risk 5 would have been \$1881 over the 22-month period. This is in excess of the total premiums earned for that risk over the same period, i.e., $22 \times \$70 = \$1540$. This would have been unsustainable. The need for a reform is even more obvious when the VaR or TVaR for that risk are taken into account.

The impact of the reform on the marginal risks for AB coverages can be assessed by comparing the VaR and TVaR for Risks 5 and 7 in Scenarios I and II. The drop

in the VaR and TVaR of the entire portfolio is approximately \$2000; this represents about 10% of the \$22,000 premium collected over that period.

Under Scenario II, $\widehat{\text{TVaR}}_{0.95}(S) = \$13{,}946$ would represent the company's economic capital for the entire portfolio. This amount is smaller than the sum of the TVaR of the individual components, which totals \$14,857. This illustrates the benefit of diversification. By comparison, the $\text{TVaR}_{0.95}$ can be estimated by simulation at \$13,504 when the risks are assumed to be mutually independent. This unwarranted assumption would lead to an underestimation of the risk exposure.

The proposed model can also be used for capital allocation purposes, i.e., to determine the share of the total risk that relates to each of the components. The TVaR is typically used to this end. For each $i \in \{1, \ldots, d\}$, the TVaR-based capital

Table 2–9: Estimated risk measures with $\kappa = 0.95$ and corresponding capital allocations for the period from September 2010 to June 2012 using the model prior to (I) and after (II) the reform.

		Risk								
		1	2	3	4	5	6	7	8	Total
	Q×LR	1376	2937	1141	2696	749	875	718	1203	11,694
	Mean	1371	2911	1368	3038	1881	1042	1528	1401	14,541
	St. Dev.	35	156	193	246	164	121	107	107	583
I	VaR	1429	3179	1683	3465	2162	1247	1711	1581	15,545
	TVaR	1456	3296	1847	3603	2245	1302	1760	1628	15,903
	ATVaR	1416	3153	1769	3512	1951	1081	1571	1450	
	Mean	1371	2910	1367	3041	760	1046	714	1400	12,609
	St. Dev.	39	163	173	248	66	122	50	106	561
II	VaR	1430	3176	1676	3484	874	1254	799	1577	13,586
	TVaR	1459	3315	1802	3613	908	1313	823	1624	13,946
	ATVaR	1419	3175	1736	3531	792	1113	735	1446	

allocation for risk X_i is defined by

$$\mathrm{TVaR}_\kappa(X_i;S) = \frac{\mathrm{E}[X_i\mathbf{1}\{S > \mathrm{VaR}_\kappa(S)\}] + \beta_\kappa\,\mathrm{E}[X_i\mathbf{1}\{S = \mathrm{VaR}_\kappa(S)\}]}{1-\kappa},$$

where

$$\beta_\kappa = \frac{F_S\{\mathrm{VaR}_\kappa(S)\} - \kappa}{\Pr\{S = \mathrm{VaR}_\kappa(S)\}}$$

if $\Pr\{S = \mathrm{VaR}_\kappa(S)\} > 0$ and 0 otherwise. It can be estimated consistently by

$$\widehat{\mathrm{TVaR}}_\kappa(X_i;S) = \frac{1}{(1-\kappa)n}\left\{\sum_{j=1}^{n} X_{ij}\mathbf{1}(S_j > s_\kappa) + \frac{F_n(s_\kappa)-\kappa}{\sum_{\ell=1}^{n}\mathbf{1}(S_\ell = s_\kappa)/n}\sum_{j=1}^{n} X_{ij}\mathbf{1}(S_j = s_\kappa)\right\}.$$

Table 2–9 reports estimates of this quantity based on the same random samples of size 10,000 from the selected models under Scenarios I and II. Given that the risks are not comonotonic, the TVaR-based allocations are smaller than the TVaR of the individual risks, illustrating diversification benefits. While the reform did not affect capital allocations for Québec, as expected, it had a large impact in Ontario. In addition, note the small increase in the allocation to Risk 6, caused by its stronger dependence with Risk 5 after the reform.

Figure 2–12 displays TVaR-based allocations for September 2010, expressed in percentage of the TVaR, before and after the reform under the proposed models (top row) and under independence (bottom row). The difference in the top charts shows the combined effect of the change in marginals and dependence before and after the reform. The difference in the bottom charts isolates the effect of the margins. After the reform, the Ontario risks represent a smaller portion of the overall capital, but

the share of Risk 6 is increased. The smaller share of capital allocated to Risk 3 under independence reflects its strong dependence with Risk 4 and the large exposure to the latter.

2.7 Discussion

In this chapter, a strategy for modelling high-dimensional data is proposed; it is based on successive aggregation of pairs of dependent random variables. The model can be compared in essence to vine copula constructions, as it is defined with a tree structure, marginal distributions, and bivariate copulas. However, inference is simpler in the proposed approach, as all partial sums of variables are observable. This modelling technique is particularly well-suited for actuarial applications, where risk aggregation, and thus partial sums, is a primary concern. The proposed model provided a good fit to real data from a Canadian insurance company, and simulation was used to gain insight on risk exposure and capital allocations.

As for any model, the validity of the construction relies on assumptions. Vine copula models assume a "simplifying assumption," which has been the object of some debate; see, e.g., [27] and [1]. Here, the main caveat is a conditional independence assumption described in Section 2.2. This constraint on the dependence structure makes sense intuitively and can be checked heuristically. However, a formal validation procedure remains to be found. Also, Algorithm 1 for simulating from the resulting model is partially, but not entirely, motivated by [2]. These issues will be addressed in future work.

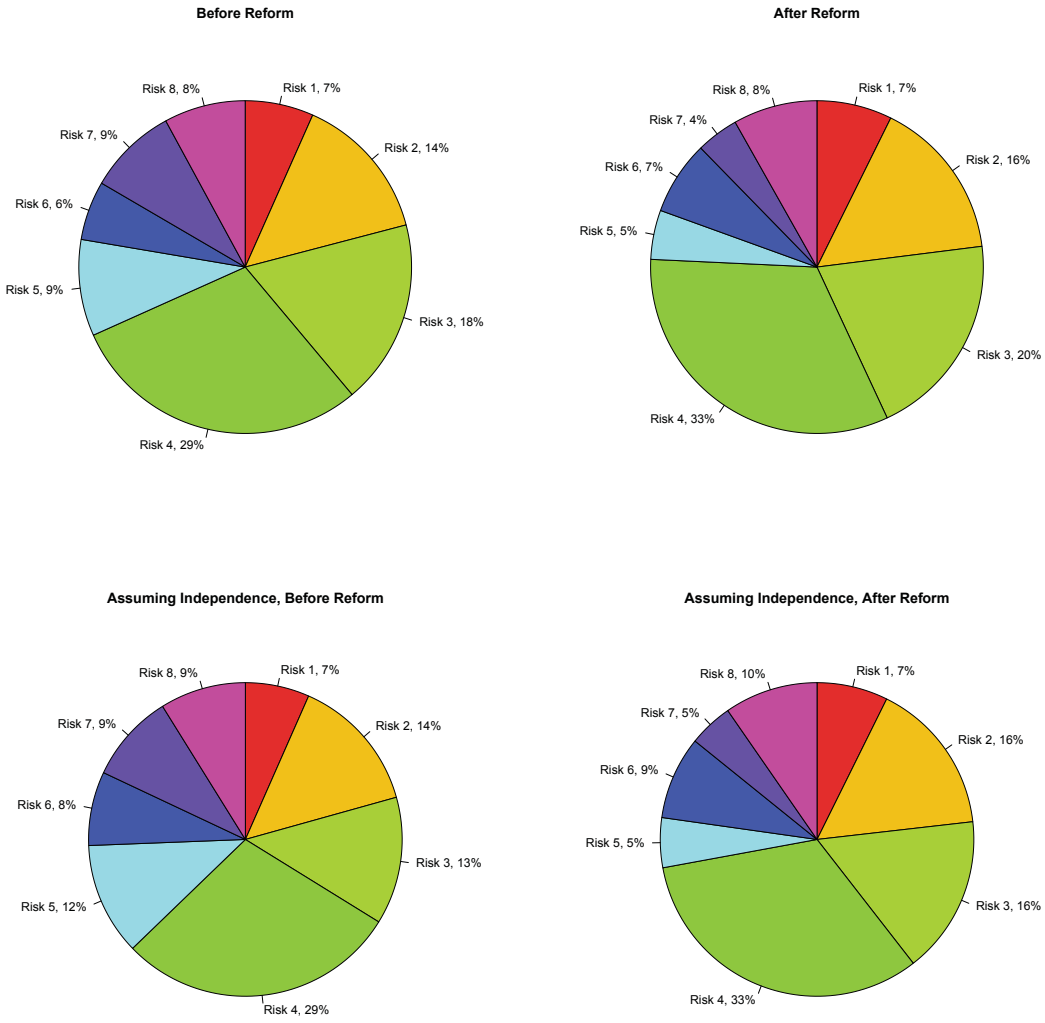

Figure 2–12: TVaR-based allocations, expressed in percentage of $\text{TVaR}_{0.95}$, before (left) and after (right) the reform under the proposed models (top row) and independence between the risks (bottom row).

Appendix

Proof of Proposition 1. The formula is established by induction on the number of risks. In the case $d = 2$, it simply states that the joint density of (X_i, X_j) is

$$\mathbf{f}_{A_1}(x_i, x_j) = c_{A_1}\{F_i(x_i), F_j(x_j)\} f_i(x_i) f_j(x_j),$$

with $A_1 = \{i, j\}$, which is trivially true.

Next, assume that the formula holds for any tree structure linking at least 2 and at most $d \geq 2$ individual variables in such a way that the conditional independence assumption holds. Consider a structure with $d + 1$ variables which satisfies the conditional independence assumption, and let $S_{A_{d,1}}$ and $S_{A_{d,2}}$ be the two sums that are aggregated at the last node, N_d. Letting $s = \sum_{i \in A_{d,1}} x_i$, one can then express the joint density of $(X_1, \ldots, X_{d+1})$ as

$$\begin{aligned}&\mathbf{f}_{\{1,\ldots,d+1\}}(x_1, \ldots, x_{d+1}) \\ &\qquad = \mathbf{f}_{A_{d,1}}(\mathbf{x}_{A_{d,1}}|S_{A_{d,1}} = s)\mathbf{f}_{A_{d,2}}(\mathbf{x}_{A_{d,2}}|S_{A_{d,1}} = s) f_{A_{d,1}}(s), \quad (2.1)\end{aligned}$$

where the conditional independence assumption has been used. Let K be the set of indices representing the aggregation steps leading to $S_{A_{d,1}}$. By the induction hypothesis,

$$\begin{aligned}&\mathbf{f}_{A_{d,1}}(\mathbf{x}_{A_{d,1}}|S_{A_{d,1}} = s) f_{A_{d,1}}(s) \\ &\qquad = \prod_{i \in K} c_{A_i} \left\{ F_{A_{i,1}}\left(\sum_{j \in A_{i,1}} x_j\right), F_{A_{i,2}}\left(\sum_{j \in A_{i,2}} x_j\right) \right\} \prod_{i \in A_{d,1}} f_i(x_i). \quad (2.2)\end{aligned}$$

Two cases must be distinguished according as $A_{d,2}$ is a singleton or not. If $A_{d,2} = \{k\}$, then, by construction,

$$\mathbf{f}_{A_{d,2}}(x_k|S_{A_{d,1}} = s) = c_{A_d}\{F_{A_{d,1}}(s), F_k(x_k)\}f_k(x_k). \tag{2.3}$$

Upon substitution of (2.2) and (2.3) into (2.1), one gets the desired result, in view of the fact that $A_{d,1} = \{1, \ldots, d+1\} \setminus \{k\}$ and $K = \{1, \ldots, d-1\}$.

If $A_{d,2}$ has more than one element and $t = \sum_{i \in A_{d,2}} x_i$, then

$$\begin{aligned}\mathbf{f}_{A_{d,2}}(\mathbf{x}_{A_{d,2}}|S_{A_{d,1}} = s) &= \mathbf{f}_{A_{d,2}}(\mathbf{x}_{A_{d,2}}|S_{A_{d,1}} = s, S_{A_{d,2}} = t)f_{A_{d,2}}(t|S_{A_{d,1}} = s) \\ &= \mathbf{f}_{A_{d,2}}(\mathbf{x}_{A_{d,2}}|S_{A_{d,2}} = t)f_{A_{d,2}}(t|S_{A_{d,1}} = s),\end{aligned}$$

where the conditional independence assumption was used once more. By construction, one has

$$\mathbf{f}_{A_{d,2}}(\mathbf{x}_{A_{d,2}}|S_{A_{d,1}} = s) = \mathbf{f}_{A_{d,2}}(\mathbf{x}_{A_{d,2}})c_{A_d}\{F_{A_{d,1}}(s), F_{A_{d,2}}(t)\}.$$

Now if $L = \{1, \ldots, d-1\} \setminus K$, the induction hypothesis yields

$$\mathbf{f}_{A_{d,2}}(\mathbf{x}_{A_{d,2}}) = \prod_{i \in L} c_{A_i}\left\{F_{A_{i,1}}\left(\sum_{j \in A_{i,1}} x_j\right), F_{A_{i,2}}\left(\sum_{j \in A_{i,2}} x_j\right)\right\} \prod_{i \in A_{d,2}} f_i\left(x_i\right).$$

To conclude, it suffices to replace this expression and identity (2.2) into (2.1). □

Proof of Proposition 2. Clearly, $D(X, X) = 0$ and $D(X, Y) = D(Y, X)$ for all choices of X and Y. To establish the triangle inequality, first write

$$D(X, Y) = f\{T(X, Y)\} = 2\sqrt{T(X, Y)\{1 - T(X, Y)\}},$$

in terms of the probability $T(X,Y)$ that X and Y are discordant. Given variables X, Y, Z with $a = T(X,Y)$, $b = T(X,Z)$ and $c = T(Z,Y)$, one must show that

$$f(a) \leq f(b) + f(c). \tag{2.4}$$

As T is a pseudo metric ([13]), one has $a \leq b + c$. Furthermore, it can be assumed without loss of generality that $b, c \leq 1/2$. This is because if $b \leq 1/2$ and $c > 1/2$, say, then (2.4) holds if and only if $f(a') \leq f(b') + f(c')$, where $a' = T(X,-Y)$, $b' = T(X,Z)$ and $c' = T(Z,-Y)$, with $a' \leq b' + c'$, $b' = b \leq 1/2$ and $c' = 1 - c \leq 1/2$. Similarly if $b > 1/2$ and $c \leq 1/2$, simply replace X by $-X$, while if $b, c > 1/2$, use $-Z$ instead of Z.

Now suppose $b \leq c$ without loss of generality. If $a \leq c$, then $f(a) \leq f(c) \leq f(b)+f(c)$ because f is increasing on $[0, 1/2]$. The same argument applies if $1-a \leq c$, because $f(a) = f(1-a)$. Finally, suppose $a > c$ and $1 - a > c$. Two cases must be considered. If $b + c \leq 1/2$, then $f(a) \leq f(b+c) \leq f(b) + f(c)$ because $f(0) = 0$ and f is concave increasing on $[0, 1/2]$, which implies that it is subadditive on this interval. If, however, $b + c > 1/2$, then there exists $\delta > 0$ such that $s + t - \delta = 1/2$, and hence, from the properties of f,

$$f(a) \leq f(1/2) \leq f(s) + f(t - \delta) \leq f(s) + f(t).$$

This completes the argument. □

CHAPTER 3
Data Analysis and Model Fitting

This chapter contains additional details concerning the data analysis that led to the choice of a copula-based risk aggregation model for the insurance portfolio data presented in Section 2.6. The data available consist of claim amounts and associated earned premiums for the eight portfolios described in Table 2–3.

Claims data were provided in the form of monthly cumulative claim payment development triangles. Data were available for accident months from January 2004 to June 2012, and claim development data were given as of August 2013. As some of the latest months were recent, the classical Chain Ladder factors (see, e.g., [34]) were used to project the claims to their ultimate values. This is quite straightforward and sensitive to confidentiality, so the development factors for the eight risks are not reported here.

The earned premium for month t represents the portion of the premium paid to the insurer that is allocated to month t, regardless of the actual payment frequency.[1]

[1] For example, if a policy has an annual premium of \$120 paid in January, the earned premium in January is only \$10 even if \$120 is entered in the company's book. In February, the policyholder doesn't pay a new premium, but the earned premium for that month is again 10\$, and so on.

Assuming that pricing is consistent over the entire period, the earned premium is a measure of the volume of the portfolio, which is directly related to the claim amount.

The purpose of this application is to build an appropriate model for the total claim payment arising from the eight portfolios, in order to determine risk measures and capital allocations. However, the claim payments for different months are not directly comparable, because of the upward trend induced by inflation and business growth. In fact, material increase in volume is visible in the data; for example, the ratio of earned premiums for Subsidiary C in January 2012 to those in January 2004 is 263%. Thus, the fluctuation in volume has to be taken into account. This can be accomplished by modelling the loss ratio, defined as the ultimate claim paid over the earned premium, rather than by modelling the claim amounts. In this fashion, the effect of inflation is also eliminated.

The loss ratios are analyzed for trend, seasonality and autocorrelation as detailed in Section 3.1. It is necessary to extract these marginal effects before fitting the multivariate copula model. Once the behaviour of the individual risks is determined, the tree structure for the hierarchical dependence model is selected in Section 3.2 and the choices of copulas are explained. The proposed risk aggregation model seems adequate for this portfolio, as can be seen in Section 3.3. Finally, conclusions obtained with the fitted model are outlined, along with further comments in Section 3.4.

3.1 Univariate analysis

For each risk, the loss-ratio time series is reviewed for trend, seasonality or structural change point. Autocorrelation and mean reversion are taken into account if significant. Then, a parametric distribution is chosen, and the parameters are

fitted using maximum likelihood estimation. The distributions considered for each risk included the Gamma, Pareto, Lognormal and truncated Normal distributions. In most cases, the fat tail of the loss ratio distribution is hard to capture accurately, so generalized hyperbolic distributions, described in [37], were also considered as an option. Using the parametrization in [5], Y follows a generalized hyperbolic distribution if there exist $\mu, \gamma \in \mathbb{R}$ and $\sigma > 0$ such that

$$Y \stackrel{d}{=} \mu + W\gamma + \sqrt{W}\sigma Z,$$

where $Z \sim \mathcal{N}(0,1)$, $W \sim$ Generalized Inverse Gaussian(λ, χ, ψ), and $\stackrel{d}{=}$ denotes equality in distribution. The parameters may be estimated with the R package `ghyp`, where for simplicity and identifiability, one sets $\mathrm{E}(W) = 1$ and $\alpha = \sqrt{\chi\psi}$. A special case of this distribution is the skewed $t(\nu)$ distribution, obtained when $W \sim$ Inverse Gamma$\{\nu/2, (\nu-2)/2\}$. These distributions are defined on $\mathbb{R}$, so care was taken to verify that there was essentially no weight on $(-\infty, 0)$. In the following subsections, the univariate analysis is performed on each type of coverage.

3.1.1 Québec automobile insurance coverages

Loss ratios for Risks 1 and 2, Québec automobile insurance coverages, exhibit large seasonality. As displayed in Figure 3–1, larger loss ratios are incurred in the winter, and there is a small increase in the incidence during summer vacation time. The loss ratios are thus decomposed into a random component and a deterministic monthly seasonality, denoted $s_{\ell,t}$ for $\ell = 1, 2$ and such that $s_{\ell,t} = s_{\ell,t+12}$. The seasonal components, shown in Table 2–6, are estimated by setting $\hat{s}_{\ell,k} = w_{\ell,k} - \sum_{j=1}^{12} w_{\ell,j}$ for $k = 1, \ldots, 12$ and $\ell = 1, 2$, where $w_{\ell,k}$ is the average of the centered loss ratios for

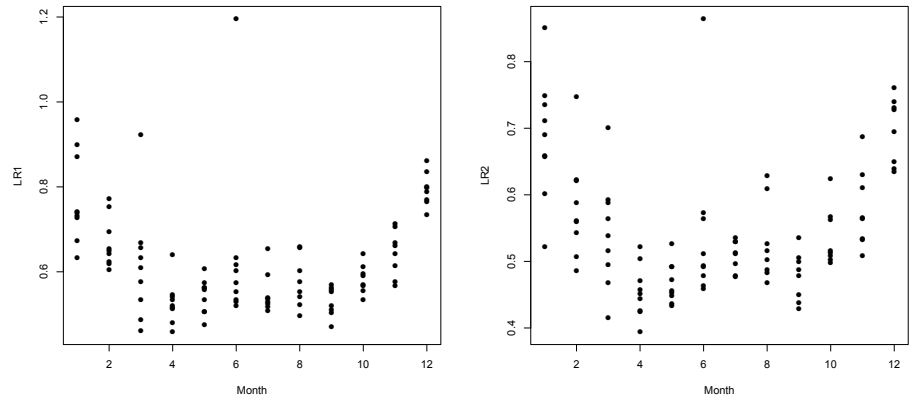

Figure 3–1: Scatterplots of the loss ratios for Risks 1 (left) and 2 (right) in terms of the month.

month k over all the years in the study period (see, e.g., [7] for details). In computing $w_{\ell,6}$, for $\ell = 1, 2$, the outlier data points for June 2008 are removed. These large loss ratios are due to the hail storm that affected many regions of Québec, and this storm is listed as a catastrophic event by the Insurance Bureau of Canada in [28]. No trend was discernible in the time series for Risks 1 and 2, and one can see in Figure 3–2 that the deterministic component, representing a constant mean plus the seasonality, explains most of the seasonal variations in the data.

Once the seasonality is removed from Risk 1, $X_{1,t} = LR_{1,t} - \hat{s}_{1,t}$ is a White Noise, as supported by the Ljung–Box test with 10 degrees of freedom (p-value of 6.3%). However, in order to account for mean reversion and significant autocorrelation of first order in Risk 2, LR_2 is modelled as $LR_{2,t} = \hat{m}_2 + \hat{s}_{2,t} + Y_{2,t}$, where $\hat{m}_2 = 0.5508$, $\hat{s}_{2,t} = \hat{s}_{2,t+12}$, and

$$Y_{2,t} - 0.9346 Y_{2,t-1} = X_{2,t} - 0.7793 X_{2,t-1},$$

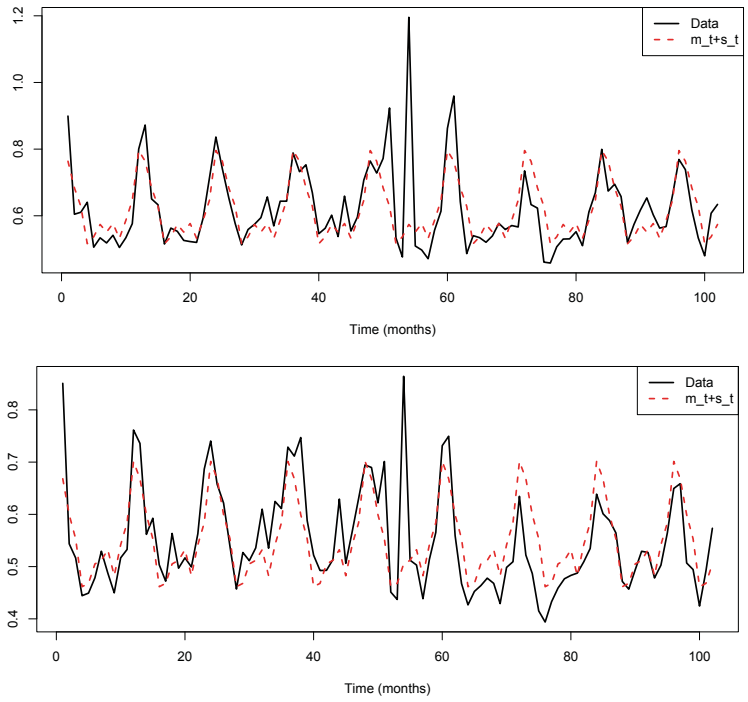

Figure 3–2: Time series and deterministic component for Risks 1 (top) and 2 (bottom).

where $X_{2,t}$ is a White Noise. This ARMA$(1,1)$ model was selected using the Akaike Information Criterion (AIC); it has the lowest AIC at -280.2, compared with -276.2 for ARMA$(2,1)$ and -275.7 for AR(3). Note that the Gaussian likelihood is used even though the innovations are not Gaussian, but this is adequate for model selection purposes as explained in [6].

It remains to choose marginal distributions for X_1 and X_2. The best fit for Risk 2 is provided by the skewed t distribution, with 5 degrees of freedom, $\hat{\mu} = -0.0310, \hat{\sigma} = 0.0483$ and $\hat{\gamma} = 0.0303$. The AIC for that model is the lowest, at -315.7, and is followed by the more complex generalized hyperbolic distribution (AIC of -311.7).

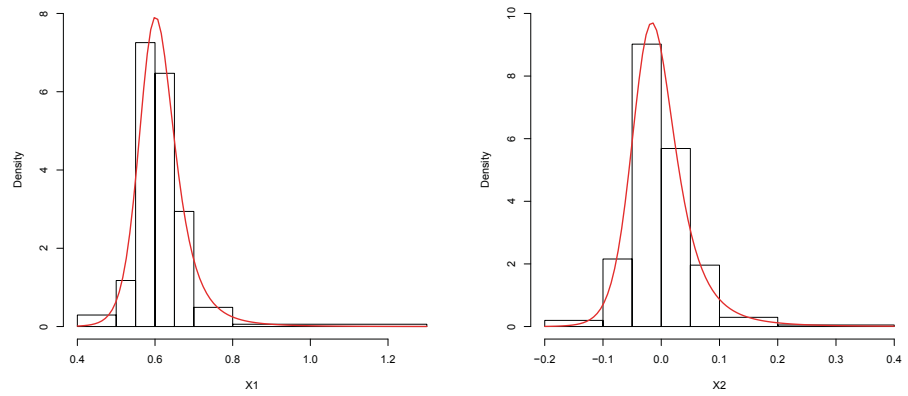

Figure 3–3: Histograms of Risks 1 (left) and 2 (right) and fitted skewed $t(5)$ densities.

The AIC obtained for the other distributions were above -297. The histogram of X_2 is similar to the fitted skewed t density, as shown in the right panel of Figure 3–3. The probability of generating a negative loss ratio (once the seasonal and mean components are added) is equal to 7.2×10^{-12}, which is negligible. Summary statistics are shown in Table 3–1 and are comparable with their empirical counterpart. The quantile-quantile plot shown in the right panel of Figure 3–4 outlines the presence of the catastrophic loss ratio in June 2008. This data point is not so influential in the parameter estimation; the main impact of the removal of this observation is to decrease the estimate of the skewness parameter γ by 0.01.

Table 3–1: Summary statistics for Risks 1 and 2.

Risk	Distribution	Mean	St. Dev.	Median	95th Quantile	99th Quantile
X_1	Empirical	0.6214	0.0890	0.6105	0.7403	1.2374
	Skewed t	0.6187	0.0768	0.6098	0.7337	0.8533
X_2	Empirical	-0.0008	0.0597	-0.0070	0.0940	0.3425
	Skewed t	-0.0006	0.0689	-0.0087	0.0952	0.2019

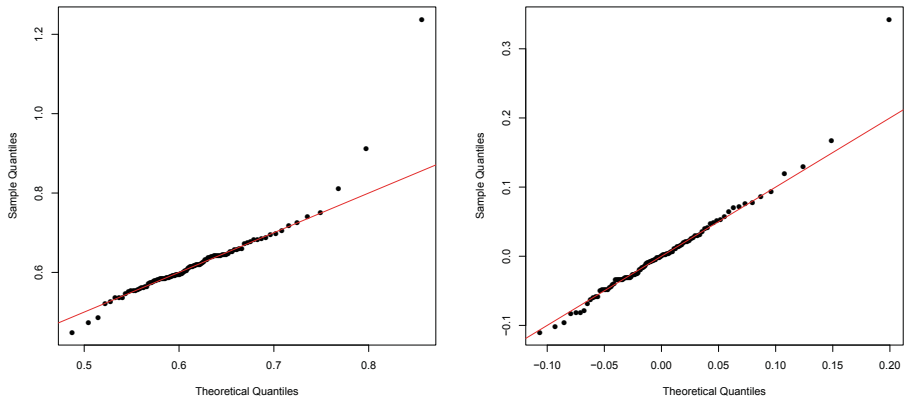

Figure 3–4: Quantile-quantile plots for Risks 1 (left) and 2 (right) and fitted skewed $t(5)$ densities.

Extreme observations for Risk 1 restrict the options available to the heavy-tailed distributions. Once again, the skewed $t(5)$ distribution provided the best fit, although the heaviness of the right tail is not entirely captured by this model, which is illustrated in the left panel of Figure 3–4. The probability of generating a negative loss ratio is immaterial (2.5×10^{-9}). The histogram of Risk 1 is compared with the fitted density in Figure 3–3; the result is satisfying. The AIC for Skewed $t(5)$ is -267.6, compared with -265.1 for the generalized hyperbolic distribution. The other distributions led to very poor fit and an AIC higher than -230. It is interesting to note that even though the quantile-quantile plot is not perfectly aligned, there is a 0.08% probability of generating a loss ratio as high as the one observed in June 2008. In the options considered, no other model with finite variance could generate such a large observation while providing a reasonable fit to the bulk of the distribution. Finally, the outlier observation is not so influential on the overall estimation.

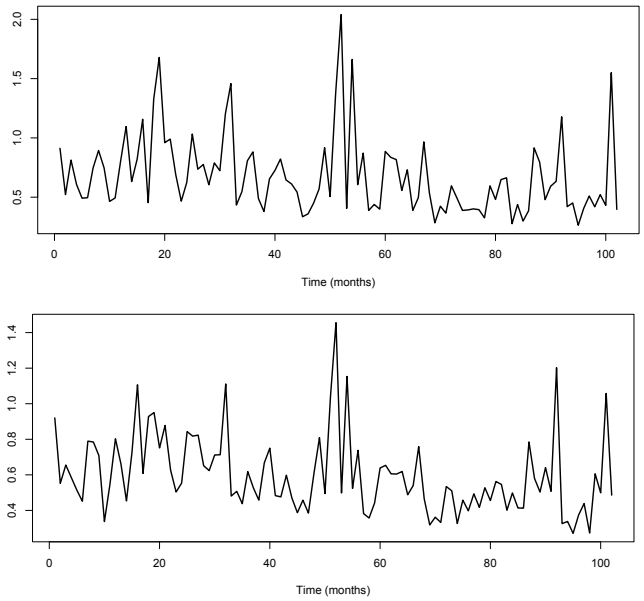

Figure 3–5: Time series for Risks 3 (top) and 4 (bottom).

3.1.2 Québec home insurance coverages

Risks 3 and 4 are home insurance coverages in Québec. These risks are simpler to model than the automobile insurance in the same province. No seasonality or trend is detected, even when the outlier data points for April and June 2008 are removed.[2] Both of these time series are White Noise according to the Ljung–Box test (see Table 2–4), which can be visually assessed by looking at Figure 3–5.

For Risk 3, the lowest AIC, 18.6, is obtained for the skewed t distribution with 9 degrees of freedom. However, in that case, the skewness is too large so it is impossible

[2] Large amounts of snow in the winter of 2008 caused damage to pools and structures, which were reported mainly in April. Many claims were incurred due to the June 2008 hail storm.

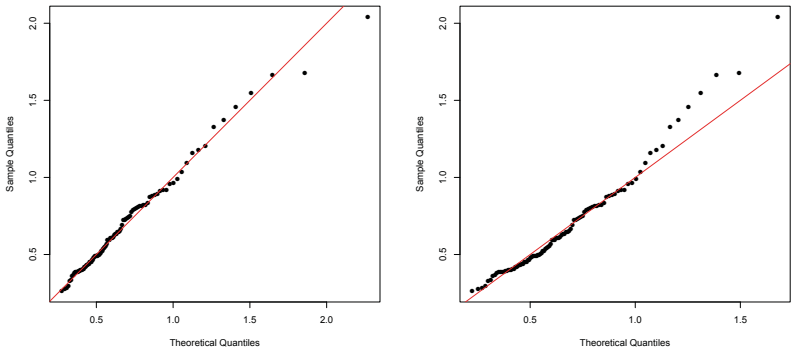

Figure 3–6: Quantile-quantile plots for Risk 3 using skewed $t(8)$ (left) and lognormal (right).

to evaluate the quantiles reliably by simulation, and there is no closed form expression for the cumulative distribution function. Thus, it is more useful to select a skewed $t(8)$, with AIC $= 18.8$, because the quantiles can be evaluated and the difference in the AIC and the fit is immaterial. The lognormal distribution yielded an AIC of 22.0, but the probability in the right tail is clearly underestimated, as shown in Figure 3–6. The skewed $t(8)$ is preferred, even if the high quantiles are slightly overestimated (as can be visually assessed in the left panel of Figure 3–6).

In Table 3–2, some summary statistics of the distribution of X_3 are compared with their empirical versions. The standard deviation of the fitted model is larger than its empirical counterpart due to the heavy tail. It is preferable to overestimate

Table 3–2: Summary statistics for Risks 3 and 4.

Risk	Distribution	Mean	St. Dev.	Median	95th Quantile	99th Quantile
X_3	Empirical	0.6774	0.3328	0.5958	1.46	2.04
	Skewed $t(8)$	0.6904	0.4168	0.5825	1.40	2.25
X_4	Empirical	0.5993	0.2188	0.5429	1.11	1.46
	Skewed $t(16)$	0.6008	0.2286	0.5526	1.03	1.39

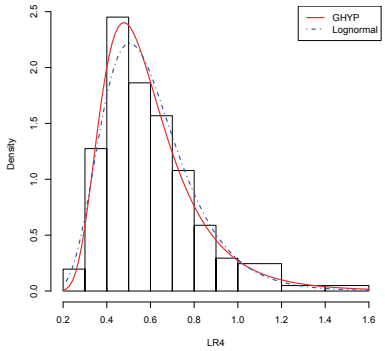

Figure 3–7: Histogram of Risk 4 and fitted densities.

the high quantile than to underestimate them (by using the lognormal distribution, for example), as it leads to a more conservative assessment of the risk related to the insurance portfolio. Finally, there is virtually no weight below zero for this skewed $t(8)$ distribution, and the parameters estimated are given in Table 2–5.

For Risk 4, the skewed $t(16)$, with AIC $= -44.8$ and the lognormal distribution, with AIC $= -44.9$, are considered. Both densities seem to provide a good fit to the histogram of LR_4 in Figure 3–7. However, the skewed t distribution provides a better fit in the right tail, as illustrated in the quantile-quantile plot in Figure 3–8.

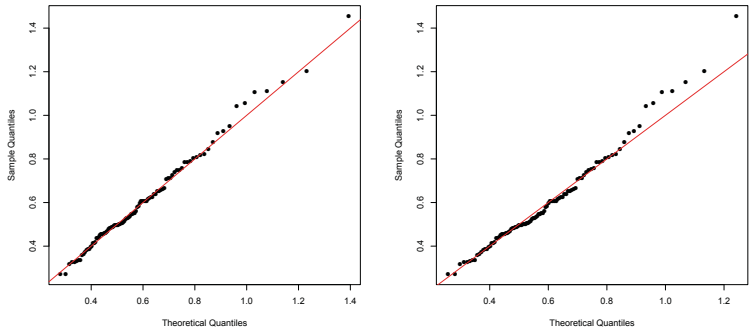

Figure 3–8: Quantile-quantile plots for Risk 4 using skewed $t(16)$ (left) and lognormal (right).

Thus, the latter option is selected, in line with the treatment of Risk 3. Summary statistics in Table 3–2 show that the median, first, and second moments are close to their empirical values.

3.1.3 Ontario Accident Benefits coverages

Risks 5 and 7, Accident Benefits (AB) coverages in Ontario, were affected by a reform of the Ontario Insurance Act effective on September 1, 2010 (month 81 in the time series data). The mitigating effect of the reform is clearly visible in the time series for both risks in Figure 3–9, so it is adequate to assume a structural change point at that date. However, this leaves only 22 data points to estimate the distribution after the reform. In order to guide the choice of distribution, the same family of distribution was fitted before and after the reform.

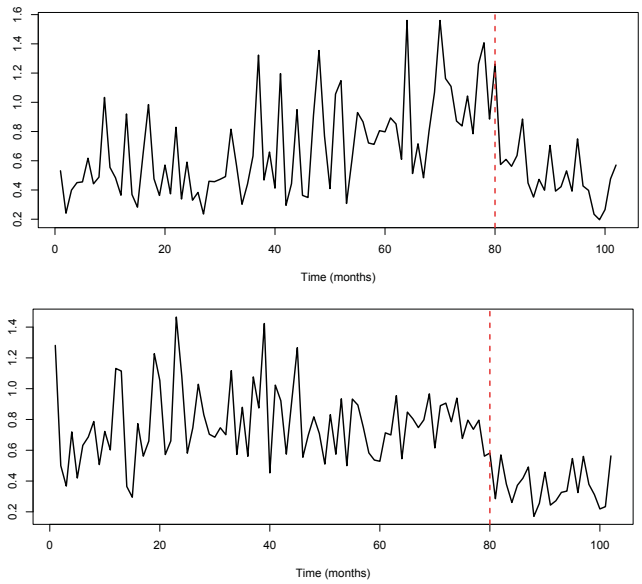

Figure 3–9: Time series for Risks 5 (top) and 7 (bottom).

Risk 5 shows an upward trend before the reform, due to a period of high inflation in the claim amounts. Many options for modelling the trend deterministically were tested. In the end, it was decided that the simple option of fitting a linear regression to the logarithm of the loss ratio provided a satisfactory fit. Thus, for $t = 1, \ldots, 80$,

$$\log(LR_{5,t}) = -0.9357 + 0.0116t + \varepsilon_{5,t},$$

where $\varepsilon_{5,t} \sim \mathcal{N}(0, \hat{\sigma}_b = 0.3939)$. The test of overall significance for this regression model yielded a p-value of 4.5×10^{-8}, and the coefficient of determination is 32.04%. The studentized residuals appear to be Normal and exhibit no clear pattern, as shown in Figure 3–10.

Therefore, the lognormal distribution was also selected for Risk 5 after the reform. It provides an adequate fit to the few data points available. A constant mean is reasonable in this case, and the maximum likelihood estimates of the parameters are $\hat{\mu}_a = -0.7821$ and $\hat{\sigma}_a = 0.3684$. As the loss ratios on the log scale are Normally

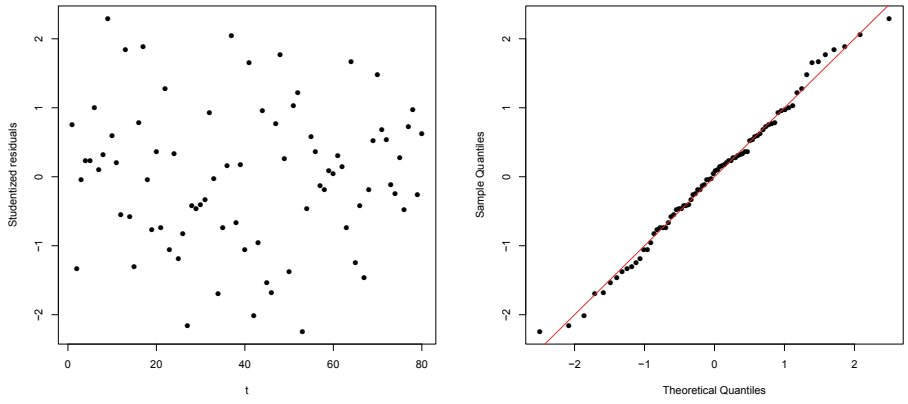

Figure 3–10: Studentized residuals (left) and quantile-quantile plot (right) for the regression on $\log(LR_{5,t})$.

distributed, an F-test of equality of variance can be used and leads to a statistic of 1.66 with 79 and 21 degrees of freedom. Thus, the p-value is 9.4%, and the test does not allow the rejection of the null hypothesis that both variances are equal. The pooled variance estimate $\hat{\sigma} = 0.3887$ is, therefore, the estimated parameter for both before and after the reform.

Risk 7 is a White Noise and no trend is detected in the time series. Before the reform, the lognormal (AIC $= -5.3$) and the gamma (AIC $= -5.8$) distributions are the best options. The quantile-quantile plots are compared in Figure 3–11, and one can see that even though the AIC is lower for the gamma distribution, the right tail is better represented by the lognormal distribution (shown in the left panel). Hence, the lognormal distribution is selected to capture the high quantiles, and the same family is fitted to the data after the reform. Once again, the F-test of equality of variance is performed and the hypothesis that the variance before and after the reform are equal cannot be rejected.

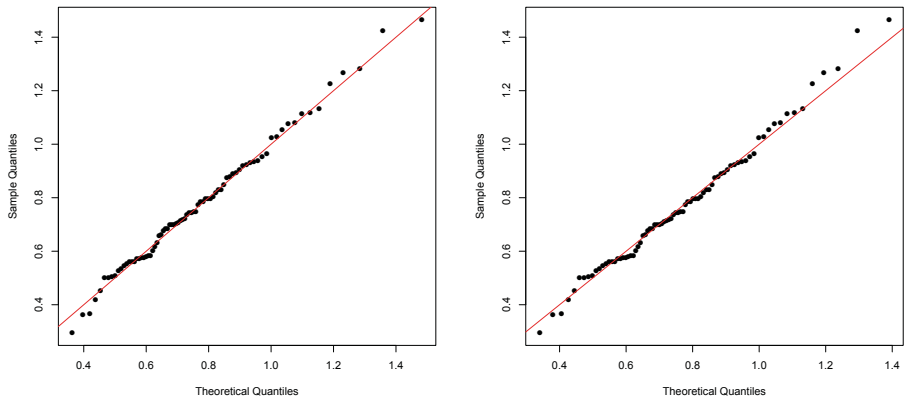

Figure 3–11: Quantile-quantile plots for Risk 7 before the reform using lognormal (left) and gamma (right).

3.1.4 Ontario Third Party Liability coverages

The last type of coverage to model marginally is the Third Party Liability (TPL) insurance, represented by Risks 6 and 8. The reform of the Ontario Insurance Act did not affect the TPL coverages directly, and the time series data for these risks displayed in Figure 3–12 do not reflect any major change after September 2010. It is therefore simple and justifiable to model these risks without assuming a change point. Both risks are White Noise according to the Ljung–Box test.

The exceptionally large loss ratio in April 2007 for Risk 6 is explained by two litigation cases that were incurred in that month. The presence of this extreme observation reflects the volatile nature of the TPL claims, and it is thus important to take this observation into account. However, it is not easy to find a model that

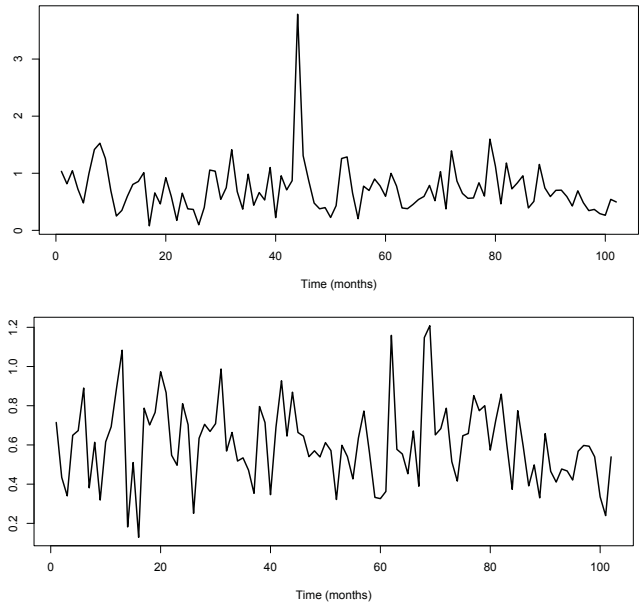

Figure 3–12: Time series for Risks 6 (top) and 8 (bottom).

Table 3–3: Summary statistics for Risks 6 and 8.

Risk	Distribution	Mean	St. Dev.	Median	95th Quantile	99th Quantile
X_6	Empirical	0.7300	0.4494	0.6592	1.41	3.78
	Gamma	0.7299	0.3954	0.6599	1.48	1.95
X_8	Empirical	0.6061	0.2080	0.5978	0.99	1.21
	Gamma	0.6061	0.2162	0.5806	1.00	1.22

can produce such large observations. The skewed t distribution has the smallest AIC (80.5), followed by the gamma (82.7). The skewed t distribution is not skewed enough to remove all the weight on $(-\infty, 0)$, so the gamma distribution was preferred, considering that the model fits were very similar. In Table 3–3, one can see that the standard deviation of the fitted model is smaller than the empirical one. This is due to the large influence of the outlier observation on the second empirical moment. The mean, the median, and the 95th quantile of the fitted distribution match reasonably their empirical counterparts.

For Risk 8, the gamma distribution has the minimum AIC, -27.9, followed by the skewed t distribution (AIC $= -26.9$). However, with the latter, the probability of having a negative loss ratio is 0.04%, which is inadequate. Thus, the gamma distribution with $\hat{\alpha} = 7.86$ and $\hat{\beta} = 12.96$ is selected. In Table 3–3, it is clear that the parametric distribution fits very closely the empirical one, as the moments and the quantiles compared are almost identical. The quantile-quantile plot for Risks 6 and 8 are shown in Figure 3–13 and support the above analysis.

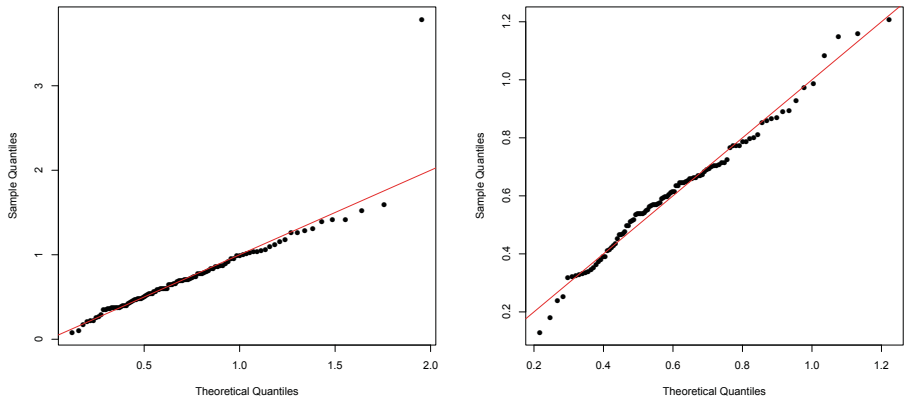

Figure 3–13: Quantile-quantile plots for Risks 6 (left) and 8 (right) using gamma distribution.

3.2 Determination of the tree structure and selection of copulas

The second component of the risk aggregation model is the tree structure defining the order in which the risks are combined. For eight risks, as many as 135,135 different trees could be considered. Although it would be more intuitive to group the risks related to the same province together, many different possibilities are reasonable, e.g., grouping first by coverage, or by subsidiary. This illustrates the usefulness of the clustering method from Section 2.3 along with a dependence-based distance.

A little more notation is needed to describe how the marginal impacts were removed before modelling the dependence structure. Recall that X_1 is the de-seasonnalized series for Risk 1, i.e., $X_{1,t} = LR_{1,t} - \hat{s}_{1,t}$, and $X_{2,t}$ represents the innovation from the ARMA model on the de-seasonalised loss ratio $Y_{2,t}$ for Risk 2. Finally, let $X_{j,t}$ denote the loss ratios $LR_{j,t}$ for $j \in \{3, \ldots, 8\}$ and $t \in \{1, \ldots, 102\}$. The purpose of the approach is to model the total claim amount, so one must translate the loss ratios into claims, by multiplying them by a premium. For an insurance

Table 3–4: Proportions of premiums, multiplied by 1000.

i	1	2	3	4	5	6	7	8
Q_i	100	250	90	230	70	65	90	105

company, future earned premiums are not known exactly due to new business and lapses, but they are fairly stable over consecutive months, and thus are easy to predict for the next month. For purposes of determining the structure and projecting the model, it is assumed that \$1,000 of premiums is invested monthly in the portfolio, in the proportions presented in Table 3–4 (repeated here for convenience). Therefore, the aggregate claim amounts S_t can be modelled as

$$S_t = Q_1\hat{s}_{1,t} + Q_2(0.551 + \hat{s}_{2,t} + 0.935Y_{2,t-1} - 0.779X_{2,t-1}) + \sum_{i=1}^{8} Q_i X_{i,t},$$

where $X_1, \ldots, X_8$ are dependent.

As the legal reform affected the marginal behaviours of Risks 5 and 7, the observations of $X_1, \ldots, X_8$ are not all identically distributed, due to the different marginal distributions for Risks 5 and 7 for $t \in \{81, \ldots, 102\}$. In fact, it is inappropriate to simply use the ranks of these two variables in the estimation, without accounting for the different marginal distributions before and after the change point. Nor was it necessary to standardize the two risks, because the reform seems to have affected the dependence structure between the Ontario coverages as well.

The method presented in Section 2.3, along with the distance based on Kendall's tau, can be applied in two stages to the data before the reform (80 data points) and to the data after the reform (22 data points). The aggregation structures obtained are shown in Figure 3–14. While the tree structure itself is the same, the dependence

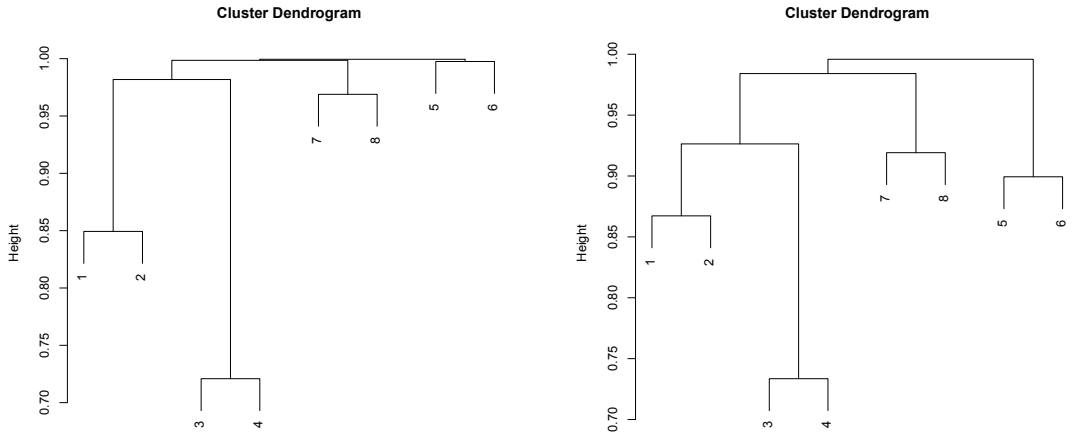

Figure 3–14: Hierarchical structures for $Q_1X_1, \ldots, Q_8X_8$ before (left panel) and after (right panel) the legal reform of September 2010.

between the Ontario risks is increased after the reform. This is supported by the values of Kendall's tau and Spearman's rho for the two samples, shown in Table 3–5. Therefore, it is simple and convenient to estimate different copulas for risks (X_5, X_6) and (X_7, X_8) before and after the reform. The Québec risks are not affected by the change point, so the entire sample is used to determine the appropriate copulas for (X_1, X_2), (X_3, X_4) and $(Q_1X_1 + Q_2X_2, Q_3X_3 + Q_4X_4)$.

The last two aggregation steps involve the Ontario risks, so the dependence should also change on September 1, 2010. However, the null hypothesis of independence is rejected in neither of the four cases on the basis of the tests of independence

Table 3–5: Association measures for Ontario risks before and after the reform.

	Before reform		After reform	
Risks	τ_{80}	ρ_{80}	τ_{22}	ρ_{22}
(X_5, X_6)	0.070	0.100	0.437	0.602
(X_7, X_8)	0.247	0.361	0.394	0.545

Table 3–6: p-values for Kendall and Spearman tests of independence on the last two aggregation steps.

	Before reform		After reform	
Step	Kendall	Spearman	Kendall	Spearman
$C_{\{1,\ldots,4,7,8\}}$	0.480	0.410	0.263	0.223
$C_{\{1,\ldots,8\}}$	0.666	0.637	0.577	0.528

carried out (see the p-values reported in Table 3–6). Hence, the product copula is appropriate for $C_{\{1,\ldots,4,7,8\}}$ and $C_{\{1,\ldots,8\}}$, reflecting risk diversification between the two provinces.

The product copula may not be used for $C_{\{3,4\}}$, following the results of the tests of independence whose p-values are reported in Table 3–7. The choice of copula family for $C_{\{3,4\}}$ can be guided by the result of the test of extremeness as described in [25]. For this test, the null hypothesis is $\mathcal{H}_0 : C_{\{3,4\}} \in \mathcal{C}$, where $\mathcal{C}$ is the class of extreme-value copulas. Thus, large p-values are desired to support the selection of an extreme-value copula. The test statistic and p-value computed with the finite-sample variance were obtained with the function `evTestK` of the `R` package `copula`. The results are shown in Table 3–8, and in this case, the p-value is 63%, which indicates that an extreme-value copula is a viable option.

Goodness-of-fit tests based on the Cramér–von Mises statistic S_n, as detailed in [23], can be used to verify that the chosen copula is reasonable. The null hypothesis

Table 3–7: p-values for Kendall and Spearman tests of independence on the Québec risks.

Step	Kendall	Spearman
$C_{\{3,4\}}$	$< 0.01\%$	$< 0.01\%$
$C_{\{1,2\}}$	$< 0.01\%$	$< 0.01\%$
$C_{\{1,\ldots,4\}}$	0.33%	0.33%

Table 3–8: Results of test of extremeness with finite sample variance.

Step	Statistic	p-value
$C_{\{3,4\}}$	-0.49	62.7%
$C_{\{1,2\}}$	1.79	7.3%
$C_{\{1,\ldots,4\}}$	0.05	95.9%
$C_{\{5,6\}a}$	4.61	$< 0.1\%$
$C_{\{7,8\}b}$	2.05	4.0%
$C_{\{7,8\}a}$	1.27	20.4%

of this test is again of the form $\mathcal{H}_0 : C_{\{3,4\}} \in \mathcal{C}$, for specific choices of parametric copula families. Goodness-of-fit tests on different copulas were conducted and the results are shown in the Appendix. The largest p-value (18%) was obtained for the Galambos copula. Although this is a shaky basis for selecting a copula family, this specific choice also seems to provide the best fit, based on comparisons of the scatterplot of the pairs of ranks and simulated observations from different copulas (shown in Figure 3–15 only for the Galambos copula), and comparisons of pairs

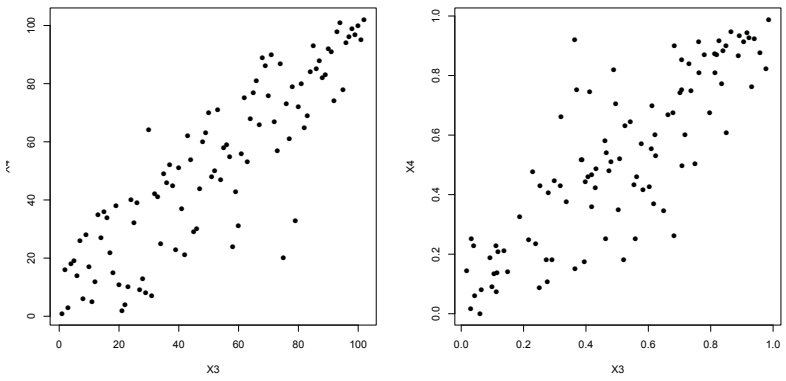

Figure 3–15: Pairs of ranks for X_3 and X_4 (left), and 102 simulated observations from a Galambos copula with $\tau = 0.68$ (right).

of observed and simulated rankits. Hence, the Galambos copula with $\tau = 0.68$ is selected for $C_{\{3,4\}}$.

The same process is used to select the copula family for $C_{\{1,2\}}$. The product copula is not appropriate (as shown in Table 3–7) and there is no strong evidence in favour of an extreme-value copula; the p-value of the test of extremeness, reported in Table 3–8, is only 7.3%. Pairs of ranks are shown in the left panel of Figure 3–16 and are compared with simulated observations for the t_8 copula with parameter 0.755, estimated by maximum pseudo-likelihood. This copula seems to describe appropriately the dependence between X_1 and X_2, and also has the maximum p-value for the goodness-of-fit test performed (refer to Table A–2 in the Appendix).

The last copula for Québec risks links $Q_1X_1 + Q_2X_2$ and $Q_3X_3 + Q_4X_4$. The hypothesis of independence is rejected at the 1% level (p-value of 0.33% for the tests of independence based on Kendall's tau and Spearman's rho). In fact, there is evidence of dependence in the extremes, as the p-value for test of extremeness is

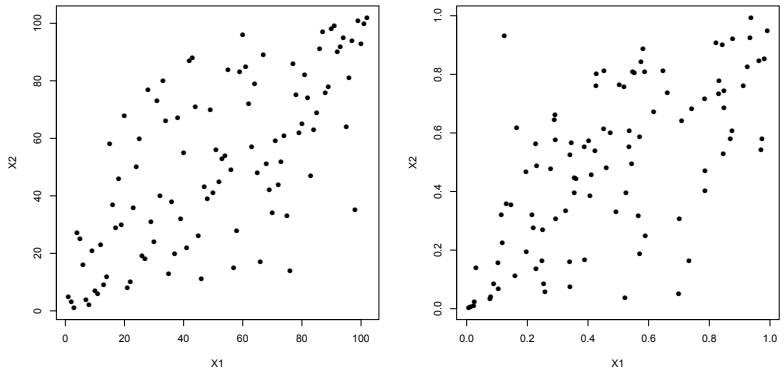

Figure 3–16: Pairs of ranks for X_1 and X_2 (left), and 102 simulated observations from a t_8 with $\tau = 0.54$ (right).

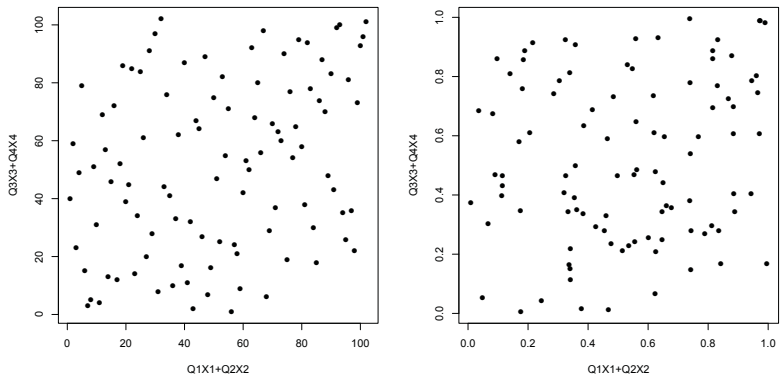

Figure 3–17: Pairs of ranks for $Q_1X_1 + Q_2X_2$ and $Q_3X_3 + Q_4X_4$ (left), and 102 simulated observations from a tev_{10} with $\tau = 0.39$ (right).

95.9%. This is also visible in the upper right corner of the scatterplot of the pairs of ranks in the left panel of Figure 3–17. The t extreme-value copula with 10 degrees of freedom and $\tau = 0.39$ is selected for $C_{\{1,\ldots,4\}}$, and the goodness-of-fit test for this copula yielded a p-value of 94.26%.

For the purpose of determining the tree structure and estimating the copulas, the loss ratios for Risk 5 before the reform are centered to remove the effect of the upward trend. The copula for Risks 5 and 6 before the reform is then the product copula; the null hypothesis of independence cannot be rejected, as can be seen from Table 3–9. However, after the reform, independence is rejected at the 1% level, and it is necessary to fit a copula to the 22 data points available after the reform. As shown in Table 3–8, extreme-value copulas are not adequate. In the absence of more observations, the Gaussian copula was selected, because it led to the highest p-value (98.45%) for the goodness-of-fit test and it had as a special case the product copula. This is thus consistent with the structure observed before the reform.

Table 3–9: p-values for Kendall and Spearman tests of independence on the Ontario risks.

	Before reform		After reform	
Step	Kendall	Spearman	Kendall	Spearman
$C_{\{5,6\}}$	35.64%	37.80%	0.39%	0.36%
$C_{\{7,8\}}$	0.12%	0.11%	0.99%	0.97%

Pairs of ranks of X_7 and X_8 before the legal reform in Ontario are plotted in the left panel of Figure 3–18. Copula $C_{\{7,8\}b}$ is probably not an extreme-value copula, as the test of extremeness mildly rejects the null hypothesis (p-value of 4%). Simulated observations from the t_{10} copula shown in the right panel of Figure 3–18 compares reasonably with the pseudo-observations of $C_{\{7,8\}b}$. This copula also yielded the highest p-value for the goodness-of-fit test performed, as outlined in Table A–5 in the Appendix. In fact, this copula family also provides a reasonable fit to the data after the reform, and it is comforting to use again the t_{10} copula for $C_{\{7,8\}a}$, but the Kendall tau induced by the copula is increased from 0.26 (before reform) to 0.44 (after reform).

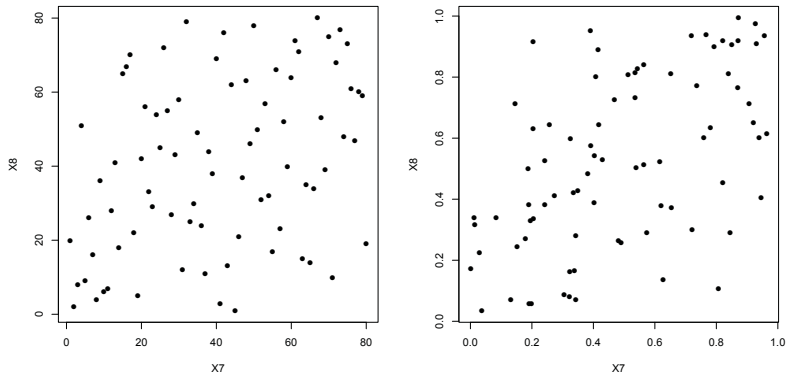

Figure 3–18: Pairs of ranks for X_7 and X_8 before the reform (left), and 80 simulated observations from a t_{10} with $\tau = 0.26$ (right).

3.3 Model validation

Algorithm 2 can be used to broadly validate the models before and after the reform. This is performed using the `TwoCop` library in R, which only supports tests when the two sample sizes are equal. p-values before the reform is 51% and after the reform is 69%. These large p-values indicate the non-rejection of the null hypothesis that the true sample and the simulated sample are coming from the same distribution. This is an argument in favour of the adequacy of the model.

Another way to verify indirectly whether the conditional independence hypothesis is fulfilled is to compare plots of observed and simulated pairs that are not explicitly modelled. Two examples are displayed in Figures 3–19 and 3–20. It is possible to visually compare the observed ranks (left panel) with simulated data from the copula-based aggregation model (right). In both cases, there is no reason to think that the model is inappropriate based on these plots. This technique was used on other implicitly modelled pairs (not shown) and similar conclusions were drawn.

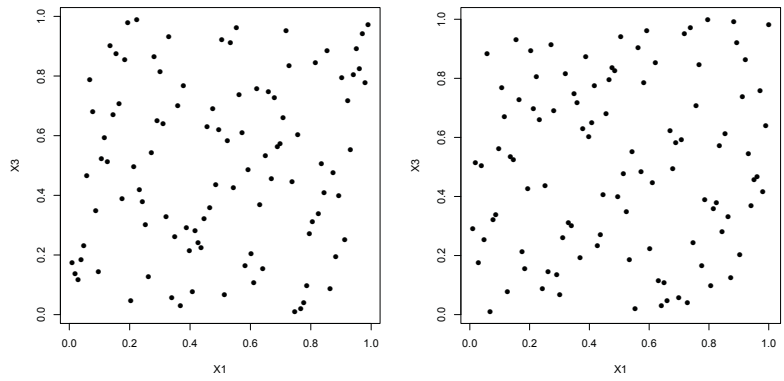

Figure 3–19: Pairs of ranks for observed (X_1, X_3) (left), and 102 simulated observations of (X_1, X_3) (right).

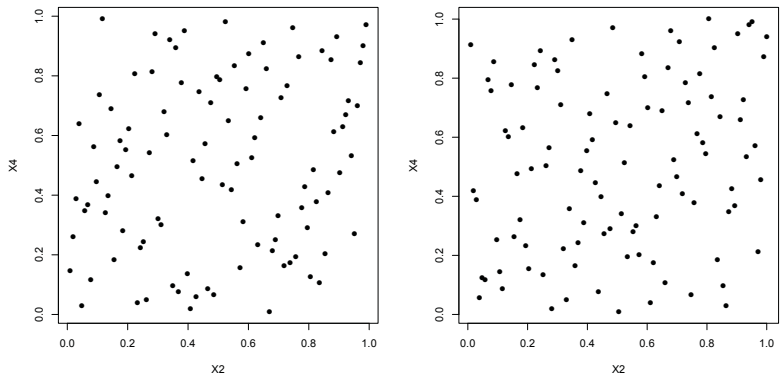

Figure 3–20: Pairs of ranks for observed (X_2, X_4) (left), and 102 simulated observations of (X_2, X_4) (right).

3.4 Further results and conclusions

The model estimated after the reform can be used to project the results for month of July 2012, assuming \$1000 of premiums in the entire portfolio, divided between the risks using proportions Q_i. Table 3–10 presents the estimated risk measures, computed with the formulas in Section 10.3.5 of [36], based on 10,000 simulations using Algorithm 1. In that table, the Q×LR rows represent the real data for the months of July and August 2012, in order to assess the predictive power of the model. However, the claim payments are known only up to August 2013 and are developed to ultimate using the Chain Ladder factors. This means that the ultimate loss ratios are uncertain, especially for Risks 6 and 8 because the claim runoff is longer for the TPL coverage. Despite the uncertainty, the expectations under the model seems similar to the observed values, except maybe for Risk 6, where the results might be overestimated. This could be explained by the fact that no change

Table 3–10: Estimated risk measures with $\kappa = 0.99$ and corresponding capital allocations for the month of July 2012.

Risk	1	2	3	4	5	6	7	8	Total
Q	100	250	90	230	70	65	90	105	1000
Q×LR July 2012	67	143	62	125	26	26	33	60	542
Q×LR August 2012	59	131	83	156	31	18	27	57	562
Mean	55	127	62	138	34	47	33	64	561
St. Dev.	7	16	38	53	14	25	11	23	110
VaR	79	179	198	317	79	125	64	129	912
TVaR	95	214	282	382	94	141	74	143	1093
ATVaR	77	176	270	374	38	52	35	71	

point was assumed after the reform for TPL for simplicity, but there was slight visual evidence of a lower mean in the loss ratios in the last months in Figure 3–12.

Summing the individual TVaRs leads to \$1425, which represents the TVaR of the sum of the risks if they were comonotonic. Using the model, a \$332 diversification benefit is gained, which is material as it represents one third of the earned premiums.

The major portion of the risk for this portfolio is allocated to Home Insurance coverage (Risks 3 and 4). This is partly due to the large premium collected for Risk 4, hence a bigger exposure, but it is also related to the strong dependence between these two risks. Interestingly, the TVaR-based capital allocations for both Home Insurance risks are higher than the earned premiums for these risks, which is not the case for the other types of coverages. For the insurance company, this means that growing the Home insurance business line in Québec requires more capital than focusing on Automobile insurance, for example. Such a multivariate model is thus a useful tool for strategic planning and capital management. This illustrates why developing a model tailored to the company's risks is encouraged under the new

Own Risk and Solvency Assessment guideline of the Office of the Superintendent of Financial Institutions in Canada [40].

Capital requirement calculated with this model could be used for insurance risks. However, other risks have to be taken into account and included in the overall assessment of an insurance company's exposure. For example, the so-called parameter risk (risk of mis-estimation of model parameters) was not considered here; the maximum likelihood parameter estimates were used without margin for adverse deviation.

CHAPTER 4
Conclusion

The copula-based risk aggregation model offers a simple and practical framework to model risk vectors. Such a model for d risks is defined with a tree structure, $d-1$ bivariate copulas and d marginal distributions. Under the conditional independence assumption, the joint distribution of the risks is unique. This assumption introduces constraints on the dependence between risks that are not aggregated together directly, but it is intuitive and it seems to be reasonable in many cases. In this thesis, a procedure to determine the tree structure was presented and an algorithm for generating data from the model was adapted from [2]. The entire process from estimation to model validation was illustrated with simulated data and with an application to insurance portfolio modelling.

The insurance data in the analysis of Chapter 3 are complex. Nevertheless, the copula-based aggregation model is flexible enough to lead to valuable conclusions. The copulas selected are various extreme-value and elliptical copulas inducing different degrees of association, meaning that using nested Archimedian copulas would be inadequate. Parameter estimation is straightforward and does not require a simplifying assumption as in the vine copula approach. The proposed aggregation model is thus an interesting option, and it is in fact already used in practice.

In the illustration, it was observed that a legislative reform had an impact on the degree of dependence between the risks. This structural change point was

obvious, but it highlights the fact that the dependence between risks may change over time, or due to external factors. This trend risk should be monitored by insurance companies, and it would be interesting to develop tools that could help model changes in dependence structure over time.

There are multiple future areas of research on the topic. It would be interesting to either validate Algorithm 1 fully, or to explore other simulation methods exploiting the joint density obtained in Proposition 2, e.g., the Metropolis–Hastings algorithm. It would also be useful to implement the test of [42] for unequal sample sizes to improve the power of the test, until a formal model validation technique is developed. This could be in the form of a goodness-of-fit test, in which the null hypothesis is that the model is adequate, or another validation technique to verify that the conditional independence assumption is fulfilled. Finally, it might also be worthwhile to see how this model can be applied to micro-level data: individual coverages within a policy could be modelled using the copula-based risk aggregation approach.

Appendix

Table A–1: Results of goodness-of-fit tests for $C_{\{3,4\}}$.

Copula	S_n	Parameter	p-value
Gumbel	0.0224	3.13	17.83%
Galambos	0.0221	2.43	18.13%
Hüsler–Reiß	0.0224	3.11	16.93%
Gaussian	0.0319	0.88	2.05%
t_4	0.0401	0.86	0.95%
Joe	0.0385	4.26	2.85%
Frank	0.0285	10.61	5.25%
Plackett	0.0447	26.79	0.65%
Clayton survival	0.0449	3.47	2.05%
tev_4	0.0239	0.96	11.04%
tev_8	0.0222	0.98	13.34%
tev_{10}	0.0220	0.98	17.03%
tev_{15}	0.0219	0.99	16.50%

Table A–2: Results of goodness-of-fit tests for $C_{\{1,2\}}$.

Copula	S_n	Parameter	p-value
Gumbel	0.0230	2.117	22.63%
Galambos	0.0227	1.414	26.82%
Hüsler–Reiß	0.0230	1.946	27.52%
Gaussian	0.0188	0.762	42.41%
t_4	0.0188	0.733	46.20%
t_7	0.0177	0.752	50.00%
t_8	0.0177	0.755	51.10%
t_9	0.0178	0.756	47.70%
t_{10}	0.0178	0.758	48.00%
t_{15}	0.0181	0.761	44.61%
Joe	0.0665	2.514	0.25%
Frank	0.0366	5.970	3.85%
Clayton	0.0727	1.596	0.55%

Table A–3: Results of goodness-of-fit tests for $C_{\{1,\ldots,4\}}$.

Copula	S_n	Parameter	p-value
Gaussian	0.0185	0.338	65.08%
t_4	0.0222	0.298	43.41%
t_8	0.0194	0.320	56.09%
t_{15}	0.0188	0.329	60.09%
Frank	0.0226	1.828	40.21%
Plackett	0.0229	2.411	45.80%
tev_8	0.0129	0.740	92.96%
tev_9	0.0128	0.766	94.26%
tev_{10}	0.0128	0.787	94.26%
tev_{11}	0.0128	0.804	93.26%

Table A–4: Results of goodness-of-fit tests for $C_{\{5,6\}a}$.

Copula	S_n	Parameter	p-value
Gaussian	0.0180	0.700	98.45%
t_4	0.0286	0.616	68.38%
t_{10}	0.0214	0.669	92.26%
t_{15}	0.0202	0.680	96.15%
Frank	0.0303	4.520	73.08%
Plackett	0.0384	5.945	49.50%
Clayton	0.0287	1.566	60.99%

Table A–5: Results of goodness-of-fit tests for $C_{\{7,8\}}$.

	$C_{\{7,8\}b}$			$C_{\{7,8\}a}$		
Copula	S_n	Parameter	p-value	S_n	Parameter	p-value
Gaussian	0.0124	0.398	96.45%	0.0239	0.640	79.87%
t_4	0.0166	0.364	82.47%	0.0252	0.625	81.87%
t_{10}	0.0129	0.390	96.95%	0.0240	0.637	85.26%
t_{15}	0.0125	0.394	96.55%	0.0239	0.638	81.97%
Frank	0.0160	2.359	91.36%	0.0293	4.322	78.47%
Plackett	0.0171	2.990	86.16%	0.0263	7.677	88.26%

References

[1] E.F. Acar, C. Genest, and J. Nešlehová. Beyond simplified pair-copula constructions. *Journal of Multivariate Analysis*, 110:74–90, 2012.

[2] P. Arbenz, C. Hummel, and G. Mainik. Copula based hierarchical risk aggregation through sample reordering. *Insurance: Mathematics and Economics*, 51:122–133, 2012.

[3] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9:203–228, 1999.

[4] M. Bargès, H. Cossette, and E. Marceau. TVaR-based capital allocation with copulas. *Insurance: Mathematics and Economics*, 45:348–361, 2009.

[5] W. Breymann and D. Lüthi. ghyp: A package on generalized hyperbolic distributions. `http://cran.r-project.org/web/packages/ghyp/vignettes/Generalized_Hyperbolic_Distribution.pdf`, 2013. Accessed on February 22, 2014.

[6] P.J. Brockwell and R.A. Davis. *Times Series: Theory and Methods.* Springer, New York, 1987.

[7] P.J. Brockwell and R.A. Davis. *Introduction to Times Series and Forecasting.* Springer, New York, second edition, 2002.

[8] H. Bühlmann. The actuary: The role and limitations of the profession since the mid-19th century. *ASTIN Bulletin*, 27:165–172, 1997.

[9] U. Cherubini, E. Luciano, and W. Vecchiato. *Copula Methods in Finance*. Wiley, New York, 2004.

[10] C. Czado, U. Schepsmeier, and A. Min. Maximum likelihood estimation of mixed C-vines with application to exchange rates. *Statistical Modelling*, 12:229–255, 2012.

[11] M. Denuit, J. Dhaene, M.J. Goovaerts, and R. Kaas. *Actuarial Theory for Dependent Risks: Measures, Orders and Models*. Wiley, New York, 2005.

[12] J. Dhaene, L. Henrard, Z. Landsman, A. Vandendorpe, and S. Vanduffel. Some results on the CTE-based capital allocation rule. *Insurance: Mathematics and Economics*, 42:855–863, 2008.

[13] P. Diaconis and R.L. Graham. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society, Series B*, 39:262–268, 1977.

[14] J. Dißmann, E.C. Brechmann, C. Czado, and D. Kurowicka. Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis*, 59:52–69, 2013.

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

[16] Financial Services Commission of Ontario. Changes to automobile insurance regulations, 2010. `www.fsco.gov.on.ca/en/auto/autobulletins/2010/Pages/a-01_10.aspx`. Accessed on February 12, 2014.

[17] E. Furman and Z. Landsman. Economic capital allocations for non-negative portfolios of dependent risks. *ASTIN Bulletin*, 38:601–619, 2007.

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

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

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

[21] C. Genest and J. Nešlehová. Copulas and copula models. *Encyclopedia of Environmetrics*, Second Edition (El-Shaarawi, A.H. and Piegorsch, W.W., eds), Wiley, Chichester, vol. 2, pp. 541–553, 2012.

[22] C. Genest, J. Nešlehová, and J.-F. Quessy. Tests of symmetry for bivariate copulas. *Annals of the Institute of Statistical Mathematics*, 64:811–834, 2012.

[23] C. Genest, B. Rémillard, and D. Beaudoin. Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44:199–214, 2009.

[24] A.W. Ghent. Kendall's "Tau" coefficient as an index of similarity in comparisons of plant or animal communities. *The Canadian Entomologist*, 95:568–575, 1963.

[25] N. Ben Ghorbal, C. Genest, and J. Nešlehová. On the Ghoudi, Khoudraji, and Rivest test for extreme-value dependence. *The Canadian Journal of Statistics*, 37:534–552, 2009.

[26] J.A. Hartigan. *Clustering Algorithms.* Wiley, New York, 1975.

[27] I. Hobæk Haff, K. Aas, and A. Frigessi. On the simplified pair-copula construction — simply useful or too simplistic? *Journal of Multivariate Analysis*, 101:1296–1310, 2010.

[28] Insurance Bureau of Canada. FACTS of the Property & Casualty Insurance Industry in Canada, 2013. `http://www.ibc.ca/en/need_more_info/facts_book/documents/ibc-facts-2013.pdf`. Accessed on February 21, 2014.

[29] R. Iman and W. Conover. A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics — Simulation and Computation*, 11:311–334, 1982.

[30] L. Jiang and Y. Lu. Effective hierarchical clustering methods for establishing theoretical p38 mapk signaling pathway in head and neck squamous cell carcinoma gene expression profiles of head and neck squamous cell carcinoma. 2007. `http://www.paper.edu.cn`.

[31] L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley, New York, 1990.

[32] S. Kotz, N. Balakrishnan, and N. L. Johnson. *Continuous Multivariate Distributions, Models and Applications*, volume 1. Wiley, New York, 2004.

[33] D. Kurowicka and H. Joe, editors. *Dependence Modeling. Vine Copula Handbook.* World Scientific, Singapore, 2011.

[34] T. Mack. Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin*, 23:213–225, 1993.

[35] J.-F. Mai and M. Scherer. h-extendible copulas. *ASTIN Bulletin*, 110:151–160, 2012.

[36] E. Marceau. *Modélisation et évaluation quantitative des risques en actuariat: Modèles sur une période.* Statistique et probabilités appliquées. Springer, Paris, 2013.

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

[38] A.J. McNeil and J. Nešlehová. From Archimedean to Liouville copulas. *Journal of Multivariate Analysis*, 101:1772–1790, 2010.

[39] R.B. Nelsen. *An Introduction to Copulas (Second edition).* Springer Series in Statistics. Springer, Berlin, 2006.

[40] Office of the Superintendent of Financial Institutions Canada. Guideline E-19: Own Risk and Solvency Assessment, 2014. `http://www.osfi-bsif.gc.ca/Eng/Docs/e19.pdf`. Accessed on March 3, 2014.

[41] P.-C. Pradier. L'actuariat au siècle des lumières: Risque et décisions économiques et statistiques. *Revue économique*, 54:139–156, 2003.

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

[43] J. Segers. Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. *Bernoulli*, 18:764–782, 2012.

[44] A.W. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, Cambridge, UK, 1998.

[45] S. van Dongen and A. J. Enright. Metric distances derived from cosine similarity and Pearson and Spearman correlations. *arXiv preprint arXiv:1208.3145*, 2012.