Insurance: Mathematics and Economics 61 (2015) 135–145

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

Tail negative dependence and its applications for aggregate loss modeling

Lei Hua

Division of Statistics, Northern Illinois University, DeKalb, IL, 60115, United States

ARTICLE INFO

Article history:
Received September 2014
Received in revised form
January 2015
Accepted 4 January 2015
Available online 14 January 2015

Keywords:
Tail order
Scale mixture
Loss frequency
Loss severity
MEPS data
Archimedean copula
GGS copula

ABSTRACT

Tail order of copulas can be used to describe the strength of dependence in the tails of a joint distribution. When the value of tail order is larger than the dimension, it may lead to tail negative dependence. First, we prove results on conditions that lead to tail negative dependence for Archimedean copulas. Using the conditions, we construct new parametric copula families that possess upper tail negative dependence. Among them, a copula based on a scale mixture with a generalized gamma random variable (GGS copula) is useful for modeling asymmetric tail negative dependence. We propose mixed copula regression based on the GGS copula for aggregate loss modeling of a medical expenditure panel survey dataset. For this dataset, we find that there exists upper tail negative dependence between loss frequency and loss severity, and the introduction of tail negative dependence structures significantly improves the aggregate loss modeling.

1. Introduction

As more data become available for statistical inference, one can now observe more new dependence patterns that have not previously attracted enough attention, probably due to technical obstacles or the non-availability of relevant dataset. The existing statistical models may not be capable of well capturing those new dependence structures. To this end, copula becomes a very flexible tool. The challenge, however, is to create a new copula family that is suitable for describing new dependence patterns, as well being computable and easy to implement. Observing an asymmetric tail negative dependence structure in a medical expenditure dataset, we find the necessity of developing new copula families to account for such dependence structures.

In actuarial science, aggregate loss modeling has been a very important task. How to model them appropriately is extremely important for insurers or governments to assess and predict the associated costs. Aggregate loss modeling usually consists of modeling of the loss frequency and severity. The former corresponds to the modeling of the average number of loss events associated with an insurance policy during a policy term, say one year, and the latter corresponds to the modeling of the average amount of the losses.

There are different ways to do aggregate loss modeling by considering loss frequency and loss severity separately. In what follows, we list two methods, and we refer the reader to Frees (2010) for more details about regression analysis for insurance applications. Let a random variable Y be loss severity, and a random variable N be loss frequency. One way is to model the loss frequency N and the conditional loss severity $Y|N > 0$ separately, by regression models such as generalized linear models, and then treat their product as the aggregate loss. If the data for aggregate losses are available, then an alternative way is to model the frequency and severity simultaneously by a mixture model, say the Tweedie model. In the first method N and $Y|N > 0$ are assumed to be independent, while in the second method N and Y are assumed to be independent.

Recently, Gschlößl and Czado (2007) find that the independence assumption commonly assumed in the literature between loss severity and frequency may not hold, and copula models are employed to account for the dependence structure between loss frequency and severity in Czado et al. (2012) and Krämer et al. (2013). The latter two papers incorporate some commonly used parametric copula families into respective regression models for loss frequency and severity, and the authors claim that there is a moderate positive dependence between loss severity and frequency in a German auto claim dataset.

The method aforementioned can also be applied to other datasets as long as there is a suitable parametric copula family that

E-mail address: hua@math.niu.edu.

http://dx.doi.org/10.1016/j.insmatheco.2015.01.001
0167-6687/

can capture the dependence structure. Based on a Medical Expenditure Panel Survey (MEPS) dataset[1] from the Agency for Healthcare Research and Quality (AHRQ), we observe an interesting dependence pattern between loss severity and frequency. In general, the average expense per visit is independent of the number of visits during each year. However, for patients who use medical services more frequently, the average cost per visit and the number of visits tend to be more negatively dependent. To the best of our knowledge, none of the commonly used copulas can capture this special dependence pattern. Therefore, we need to develop new copulas that can capture different degrees of upper tail negative dependence while keeping the rest parts approximately independent. If a suitable copula can be constructed, we will be able to incorporate the new copula into the mixed copula regression model proposed in Czado et al. (2012) and Krämer et al. (2013) for aggregate loss modeling. We refer the interested readers to the first panel of Fig. 4 for an example of the upper tail negative dependence pattern in the MEPS dataset.

After considering the families of extreme value copula, elliptical copula and Archimedean copula, we find that only Archimedean copula is suitable for our purpose. Furthermore, in order to obtain tail negative dependence for an Archimedean copula, we shall use a scale mixture approach studied in McNeil and Nešlehová (2009) for constructing the new copula. Tail behavior of Archimedean copulas has been studied in Charpentier and Segers (2009), Hua and Joe (2011, 2013) and Larsson and Nešlehová (2011), but none of them give the conditions for tail negative dependence.

Our main contributions in this paper are the following: first of all, we prove general conditions that lead to upper tail negative dependence for an Archimedean copula, which also generalize some results in Hua and Joe (2011, 2013); second, we construct some new Archimedean copulas and study their properties, and one of these copulas is very useful in modeling the asymmetric tail negative dependence pattern observed in the MEPS dataset; finally, we implement our new copula into mixed copula regression and conduct a data analysis for the medical expenditure dataset, and find that the new copula can significantly improve the aggregate loss modeling.

In what follows, we will briefly introduce some basic concepts and notations in Section 2. We will discuss in Section 3 how to construct a desirable asymmetric tail negative dependence structure based on the notion of tail order. Some parametric copulas will be constructed. Among the new copulas, a two-parameter Archimedean copula family based on generalized gamma simplex mixture is useful for modeling such an asymmetric upper tail negative dependence structure. In Section 4, we implement the new copula into the mixed copula regression model, and conduct aggregate loss modeling for a medical expenditure dataset. The introduction of the new tail negative dependence copula significantly improves the aggregate loss modeling. We conclude the paper in Section 5. Some technical details are presented in the Appendix.

2. Preliminaries

Due to its growing popularity in the last decade and its flexibility in modeling non-Gaussian dependence structures, the notion of copula has been used widely in the actuarial literature. A copula $C : [0, 1]^d \to [0, 1]$ for a d-dimensional random vector can be defined as $C(u_1, \ldots, u_d) = F(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d))$, where F is the joint cumulative distribution function (cdf), F_i is the univariate cdf for the ith marginal, and F_i^{-1} is the generalized inverse function defined as $F_i^{-1}(u) = \inf\{x : F_i(x) \geq u\}$. We refer the reader to Joe (1997) and Nelsen (2006) for references of copulas.

[1] http://meps.ahrq.gov/mepsweb/index.jsp.

For a copula C, the lower tail order of C is defined as a constant κ_L such that $C(u, \ldots, u) \sim u^{\kappa_L}\ell(u)$ as $u \to 0^+$, where the notation $g(x) \sim h(x)$ as $x \to x_0$ means that $\lim_{x\to x_0} g(x)/h(x) = 1$, and $\ell(x)$ is a slowly varying function as $x \to 0^+$. For a measurable function $g : \Re_+ \to \Re_+$, if for any constant $r > 0$, $\lim_{x\to 0^+} g(rx)/g(x) = 1$, then g is said to be slowly varying at 0^+, denoted as $g \in \mathrm{RV}_0(0^+)$; if for any constant $r > 0$, $\lim_{x\to\infty} g(rx)/g(x) = r^\alpha$, $\alpha \in \Re$, then g is said to be regularly varying at ∞ with variation exponent α, and is denoted as $g \in \mathrm{RV}_\alpha$. For a random variable X, we usually use F_X to represent the cdf of X. When we say that X is regularly varying at ∞, it actually means that the survival function $\overline{F}_X \in \mathrm{RV}_\alpha$ with some variation exponent $\alpha < 0$. Similarly, the upper tail order of C is defined as a constant κ_U such that $\overline{C}(1-u, \ldots, 1-u) \sim u^{\kappa_U}\ell(u)$ as $u \to 0^+$, where $\overline{C}$ is the survival function of C. Tail order is a flexible quantity for capturing the degree of dependence in the tails. It can be used for upper and lower tails separately, and quantify the strength of dependence ranging from tail negative dependence to tail positive dependence. The range of tail order κ is $1 \leq \kappa \leq \infty$, and generally speaking, a smaller κ implies a stronger dependence in the tail. For the bivariate case, if the tail order $\kappa > 2$, then there is negative dependence in the tail. We refer the reader to Hua and Joe (2011) for more details about the notion of tail order.

In order to capture the reflectively asymmetric tail dependence pattern (i.e., the upper and lower tails are different) in the left panel of Fig. 4, we need to construct a bivariate copula of which the upper tail order is greater than 2, and the lower tail order is allowed to be close to 2. For elliptical copulas, the upper and lower tails are symmetric, so they are not suitable. For a bivariate extreme value copula, the upper tail order is either 1 or 2, so it cannot be an upper tail negative dependence copula. For an Archimedean copula, it can be written in the following form:

$$C(u_1, \ldots, u_d) = \psi(\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_d)), \tag{1}$$

where ψ^{-1} is the inverse of the generator ψ. There are mainly two ways of constructing an Archimedean copula. One way is based on the Laplace transform (LT) of a positive random variable. Namely, let the generator ψ in (1) be the LT of a positive random variable. That is,

$$\psi(s) = \int_0^\infty \exp\{-sy\}F_Y(dy), \quad s \geq 0,$$

where Y is a positive random variable. It is well known that such a generator ψ is completely monotone and can be used to construct an Archimedean copula for any dimension. The other way is based on the survival copula for a scale mixture with a uniform distribution on a simplex (see McNeil and Nešlehová, 2009). More specifically, if a random vector $\boldsymbol{X} := (X_1, \ldots, X_d) \stackrel{d}{=} R \times (S_1, \ldots, S_d)$ satisfies some regularity conditions, where $\stackrel{d}{=}$ means "equality in distribution", then the survival copula of $\boldsymbol{X}$ is an Archimedean copula of dimension d. We will use this representation throughout the paper, and give a more formal introduction at the beginning of Section 3.

In Section 3, we will prove that the tail behavior of the above random variable R will affect the strength of dependence in the tails for the corresponding Archimedean copula. In order to characterize the tail behavior of a univariate random variable, a well-developed and mathematically tractable tool is the maximum domain of attraction (MDA), which the univariate random variable belongs to. For example, a gamma random variable belongs to the MDA of Gumbel, and a Pareto random variable belongs to the MDA of Fréchet. More mathematically, a random variable X is said to belong to the MDA of an extreme value distribution H, if there exist normalizing constants $\sigma_n > 0$ and $\mu_n \in \Re$ such that

$$(M_n - \mu_n)/\sigma_n \xrightarrow{d} H, \quad n \to \infty,$$

where M_n is the first order statistic (i.e., maximum) of a random sample of X with sample size n, and $\xrightarrow{d}$ means "convergence in distribution". This is written as $X \in \mathrm{MDA}(H)$. It is well known that there are only three non-degenerate univariate extreme value distributions: Fréchet, Gumbel and Weibull. Roughly speaking, MDA of Fréchet (denoted as $\mathrm{MDA}(\Phi_\alpha)$, where α is the shape parameter of the Fréchet distribution) includes univariate distributions that have heavier right distributional tails, MDA of Gumbel (denoted as $\mathrm{MDA}(\Lambda)$) consists of univariate distributions that have lighter right distributional tails, while MDA of Weibull corresponds to bounded random variables that are often irrelevant to actuarial applications. We refer the reader to Embrechts et al. (1997) for a classical reference on the extreme value theory and relevant applications in insurance and finance.

3. Tail negative dependence

It is clear from the brief discussion in Section 2 that, extreme value copula and elliptical copula are not suitable for constructing an asymmetric tail negative dependence structure. So, we will focus on Archimedean copula in this section.

To provide a parametric Archimedean copula that has a simple form, one often considers ψ to be an LT of a positive random variable. We refer the reader to Joe and Hu (1996) and Joe (1997) for many implementable parametric Archimedean copulas. However, Archimedean copulas generated by such LTs do not provide tail negative dependence. For a bivariate random vector (X_1, X_2), if $\mathbb{P}[X_1 \le x_1, X_2 \le x_2] \ge \mathbb{P}[X_1 \le x_1]\mathbb{P}[X_2 \le x_2]$ for any $x_1, x_2 \in \Re$, then (X_1, X_2) is said to be positive quadrant dependent (PQD). If a bivariate Archimedean copula is constructed by the LT of a positive random variable, then the copula is PQD, and thus positive upper quadrant dependent (PUQD) and positive lower quadrant dependent (PLQD) (see Section 2.1.1 of Joe, 1997), and the tail orders $\kappa \le 2$ for both upper and lower tails (Proposition 2 of Hua and Joe, 2011). So Archimedean copulas based on the LT of a positive random variable cannot be used to construct such an asymmetric upper tail negative dependence structure. However, if an Archimedean copula is derived from the survival copula of a scale mixture with a uniform distribution on the simplex, then we will show that conditions on the mixing random variable can lead to a very flexible tail for the corresponding copula. The tail can be tail dependent, intermediate tail dependent (Hua and Joe, 2011), or tail negative dependent.

Instead of using an LT of a positive random variable, one can construct an Archimedean copula using any generator ψ that satisfies certain regularity conditions (see Malov, 2001 or McNeil and Nešlehová, 2009). In McNeil and Nešlehová (2009), an Archimedean copula can be the survival copula of a random vector

$$\boldsymbol{X} := (X_1, \dots, X_d) \stackrel{d}{=} R \times (S_1, \dots, S_d), \tag{2}$$

where R and $(S_1, \dots, S_d)$ are independent, R is a positive random variable and $(S_1, \dots, S_d)$ is uniformly distributed on the simplex $\{\boldsymbol{x} \in \Re^d_+ : \sum_i x_i = 1\}$. In this case, ψ can be the Williamson d-transform of the cdf F_R with $F_R(0) = 0$. That is,

$$\psi(s) = \int_s^\infty (1 - s/r)^{d-1} F_R(dr), \quad s \in [0, \infty). \tag{3}$$

In Section 3.1, we will prove that the tail behavior of R will affect the strength of dependence in the tails of the Archimedean copula, and upper tail negative dependence can be derived from the representation (2).

3.1. Conditions

In Hua and Joe (2013), we find that when the right tail of $1/R$ follows a power law, a lighter right tail of $1/R$ tends to increase the upper tail order of the associated Archimedean copula, thus decreasing the degree of positive dependence in the upper tail. In what follows, unless otherwise specified, the tail of a univariate random variable or distribution is always referred to the right distributional tail. From Hua and Joe (2011), we know that if tail order $\kappa > d$, where d is the dimension, then the copula may have tail negative dependence. So, in order to obtain tail negative dependence, we shall decrease the tail heaviness of the random variable $1/R$. However, by observing Example 4 of Hua and Joe (2013), even if $1/R$ has a very light tail, it cannot provide tail negative dependence. The reason is not that the tail of $1/R$ is not sufficiently light, but that the Archimedean copula is constructed by the LT of a positive random variable. So, in the following Proposition 1, we will instead use the scale mixture method to construct Archimedean copulas that can have tail negative dependence.

In this section, all distribution functions and density functions are assumed to be ultimately monotone to the left and right endpoints; this condition is very mild and is satisfied by all the commonly used distributions. Since the theoretical results developed in this section are for distributional tails, without loss of generality, we further assume that the cdfs of the marginal distributions are all continuous so that the copula is uniquely determined to avoid cumbersome arguments.

Proposition 1. *Suppose a random vector $\boldsymbol{X} := (X_1, \dots, X_d)$ is defined as in (2). If $1/R \in \mathrm{MDA}(\Phi_\alpha)$ and $\mathbb{E}[1/R] < \infty$, then the lower tail order of $\boldsymbol{X}$ is $\kappa = \alpha$. That is, the upper tail order of the corresponding Archimedean copula is α.*

Proof. Let F be the identical univariate cdf for X_i's, and C be the copula for $\boldsymbol{X}$. Since the survival copula for $\boldsymbol{X}$ is an Archimedean copula, in order to study the upper tail of the Archimedean copula, it suffices to study the lower tail of $\boldsymbol{X}$. Due to Eq. (1) of Hua and Joe (2013), the upper tail order κ of the Archimedean copula can be derived as

$$\kappa = \lim_{u \to 0^+} \frac{\log C(u, \dots, u)}{\log(u)} = \lim_{x \to 0^+} \frac{\log(C(F(x), \dots, F(x)))}{\log(F(x))}$$
$$= \lim_{x \to 0^+} \frac{\log(\mathbb{P}[X_1 \le x, \dots, X_d \le x])}{\log(\mathbb{P}[X_1 \le x])}.$$

Letting $T := 1/R$, $y = 1/x$ and $s_* = \max\{s_1, \dots, s_d\}$, we have

$$\mathbb{P}[X_1 \le x, \dots, X_d \le x] = \mathbb{P}[RS_1 \le x, \dots, RS_d \le x]$$
$$= \int_{\mathbf{s} \ge \mathbf{0}, \|\mathbf{s}\|_1 = 1} \mathbb{P}[T \ge s_*/x] F_{\mathbf{S}}(ds_1, \dots, ds_d).$$

Since $T \in \mathrm{MDA}(\Phi_\alpha)$, $\mathbb{P}[T \ge \cdot] \in \mathrm{RV}_{-\alpha}$ and there exists a slowly varying function $\ell(\cdot)$ such that $\mathbb{P}[T \ge t] = t^{-\alpha}\ell(t)$, we have equations (4) and (5) given in Box I where $B(\cdot, \cdot)$ is a Beta function. Eq. (4) is implied by the fact that univariate marginals of $\mathbf{S}$ are distributed as Beta$(1, d-1)$ (Ferguson, 1973). Eq. (5) holds due to the following:

(a) $0 < \lim_{y \to \infty} \int_0^y \mathbb{P}[T \ge x](1 - x/y)^{d-2} dx \le \lim_{y \to \infty} \int_0^y \mathbb{P}[T \ge x] dx = \mathbb{E}[T] < \infty$ from the condition;
(b) by Proposition 1.3.6 (i) of Bingham et al. (1987), $\lim_{y \to \infty} \log(\ell(y))/\log(y) = 0$;
(c) since $1/d \le s_* \le 1$, and as $y \to \infty$, $\mathbb{P}[T \ge s_* y]/\mathbb{P}[T \ge y] \to s_*^{-\alpha}$ uniformly in $s_* \in [1/d, 1]$,

$$\lim_{y \to \infty} \log\left(\int_{\mathbf{s} \ge \mathbf{0}, \|\mathbf{s}\|_1 = 1} \frac{\mathbb{P}[T \ge s_* y]}{\mathbb{P}[T \ge y]} F_{\mathbf{S}}(ds_1, \dots, ds_d)\right)$$
$$= \log\left(\int_{\mathbf{s} \ge \mathbf{0}, \|\mathbf{s}\|_1 = 1} s_*^{-\alpha} F_{\mathbf{S}}(ds_1, \dots, ds_d)\right)$$
$$\le \alpha \log(d) < \infty.$$

That is, $\kappa = \alpha$, which completes the proof. $\square$

$$
\begin{aligned}
\kappa &= \lim_{x\to 0^+} \frac{\log(\mathbb{P}[X_1 \le x, \ldots, X_d \le x])}{\log(\mathbb{P}[X_1 \le x])} \\
&= \lim_{y\to\infty} \frac{\log\left(\mathbb{P}[T \ge y] \times \int_{\mathbf{s}\ge \mathbf{0}, \|\mathbf{s}\|_1=1} \mathbb{P}[T \ge s_* y]/\mathbb{P}[T \ge y] F_{\mathbf{S}}(ds_1, \ldots, ds_d)\right)}{\log\left(\int_0^1 \mathbb{P}[T \ge s_1 y] F_{S_1}(ds_1)\right)} \\
&= \lim_{y\to\infty} \frac{\log(\mathbb{P}[T \ge y]) + \log\left(\int_{\mathbf{s}\ge \mathbf{0}, \|\mathbf{s}\|_1=1} \mathbb{P}[T \ge s_* y]/\mathbb{P}[T \ge y] F_{\mathbf{S}}(ds_1, \ldots, ds_d)\right)}{-\log(y) - \log(\mathrm{B}(1, d-1)) + \log\left(\int_0^y \mathbb{P}[T \ge x](1 - x/y)^{d-2} dx\right)} \qquad (4) \\
&= \lim_{y\to\infty} \frac{-\alpha \log y + \log(\ell(y)) + \log\left(\int_{\mathbf{s}\ge \mathbf{0}, \|\mathbf{s}\|_1=1} \mathbb{P}[T \ge s_* y]/\mathbb{P}[T \ge y] F_{\mathbf{S}}(ds_1, \ldots, ds_d)\right)}{-\log(y) - \log(\mathrm{B}(1, d-1)) + \log\left(\int_0^y \mathbb{P}[T \ge x](1 - x/y)^{d-2} dx\right)} \\
&= \alpha, \qquad (5)
\end{aligned}
$$

Box I.

Remark 1. In Proposition 1, the condition on the random variable R has the following equivalent relationships: $1/R \in \mathrm{MDA}(\Phi_\alpha) \iff \overline{F}_{1/R} \in \mathrm{RV}_{-\alpha} \iff F_R \in \mathrm{RV}_\alpha(0^+)$.

Remark 2. Proposition 1 generalizes Proposition 6 in Hua and Joe (2013), where only intermediate tail dependence has been studied. Moreover, we use a different method to prove Proposition 1 in this paper and the proof is shorter than that in Hua and Joe (2013).

From Example 4 in Hua and Joe (2013), we notice that, for a d-dimensional Archimedean copula constructed by the LT of an inverse gamma random variable with the shape parameter α, the corresponding R in (2) cannot satisfy that $F_R \in \mathrm{RV}_\alpha(0^+)$ for $\alpha > d$. This is a reason why this copula does not have tail negative dependence.

More generally, we want to know whether multivariate Archimedean copula constructed by the LT of a positive random variable can have upper tail order that is larger than the dimension of the copula. As discussed in Section 1, this is not true for a bivariate Archimedean copula constructed by the LT of a positive random variable, as such a copula is PQD and thus PUQD and PLQD. So, for the bivariate case, neither the upper nor the lower tail of such an Archimedean copula can have tail negative dependence. For the multivariate case with dimension $d \ge 2$, we know that an Archimedean copula constructed by the LT of a positive random variable is positive lower orthant dependent (PLOD) (see Corollary 4.6.3 of Nelsen, 2006). Therefore, by Proposition 2 of Hua and Joe (2011), the lower tail order must be less than or equal to the dimension d for an Archimedean copula constructed by the LT of a positive random variable. However, for the upper tail, we do not know whether such a multivariate Archimedean copula is positive upper orthant dependent (PUOD). The following result implies that the upper tail order of a d-dimensional Archimedean copula constructed by the LT of a positive random variable must be less than or equal to d.

Corollary 2. *Let C be a d-dimensional Archimedean copula constructed as (1) with ψ the LT of a positive random variable Y.*

1. *If $\overline{F}_Y = o(\overline{F}_W)$, where $W \sim$ inverse-gamma$(d, 1)$, then the upper tail order κ_U of the copula C exists, and $\kappa_U = d$.*
2. *If $\overline{F}_Y \in \mathrm{RV}_{-\alpha}$ for an α such that $d \ge \alpha > 1$, then the upper tail order κ_U of C exists, and $\kappa_U = \alpha$.*
3. *If $\overline{F}_Y \in \mathrm{RV}_{-1}$ and $\mathbb{E}[Y] < \infty$, then the upper tail order κ_U of C exists, and $\kappa_U = 1$.*

Proof. From Proposition 1 of McNeil and Nešlehová (2010), we know that for an Archimedean copula constructed by the LT of a positive random variable, one can also write it as the survival copula for the random vector in (2) with a scaling random variable R, and the relationship between R and Y is that

$$\frac{1}{R} \stackrel{d}{=} Y \times \frac{1}{\mathrm{gamma}(d, 1)} =: YW,$$

where W is distributed as inverse-gamma$(d, 1)$ and independent of Y. Since

$$
\begin{aligned}
\overline{F}_W(w) &= \int_w^\infty \frac{1}{\Gamma(d)} x^{-d-1} \exp\{-1/x\} dx \\
&= \frac{1}{\Gamma(d)} \int_0^{1/w} t^{d-1} \exp\{-t\} dt = \frac{1}{\Gamma(d)} \gamma(d, 1/w) \\
&\sim \frac{1}{d\Gamma(d)} w^{-d}, \quad w \to \infty, \qquad (6)
\end{aligned}
$$

where $\Gamma(\cdot)$ is the gamma function, $\gamma(\cdot,\cdot)$ is the lower incomplete gamma function, and the asymptotic equivalence is referred to Abramowitz and Stegun (1964), $\overline{F}_W \in \mathrm{RV}_{-d}$.

To prove 1, since $\overline{F}_Y = o(\overline{F}_W)$ and $\overline{F}_W \in \mathrm{RV}_{-d}$, by the corollary of Theorem 3 in Embrechts and Goldie (1980), $\overline{F}_{1/R} \in \mathrm{RV}_{-d}$ and thus $\mathbb{E}[1/R] < \infty$ as $d = 2, 3, \ldots$. Then, by Proposition 1, $\kappa_U = d$.

To prove 2, if $\overline{F}_Y \in \mathrm{RV}_{-d}$, then by the corollary of Theorem 3 in Embrechts and Goldie (1980), $\overline{F}_{1/R} \in \mathrm{RV}_{-d}$, so the claim is proved. If $\overline{F}_Y \in \mathrm{RV}_{-\alpha}$ with $d > \alpha > 1$, then clearly $\overline{F}_W = o(\overline{F}_Y)$, and thus $\overline{F}_{1/R} \in \mathrm{RV}_{-\alpha}$ and $\mathbb{E}[1/R] < \infty$. Proposition 1 leads to the claim.

To prove 3, it is similar to the second case, but we need the extra condition $\mathbb{E}[Y] < \infty$ so that $\mathbb{E}[1/R] = \mathbb{E}[Y]\mathbb{E}[W] < \infty$, which completes the proof. □

Remark 3. Corollary 2 supplements Proposition 4 in Hua and Joe (2011) where the conditions are proposed on the LT of Y instead of the survival function $\overline{F}_Y$ of Y. However, we shall note that the condition in case 3 of Corollary 2 is a sufficient but not a necessary condition for the upper tail order being equal to 1. Based on Proposition 4 of Hua and Joe (2011) and Proposition 3 of Hua and Joe (2012), even when the right tail of $\overline{F}_Y$ is heavier so that $\mathbb{E}[Y]$ does not exist, one may still get $\kappa_U = 1$.

From Proposition 1, we find that when the right tail of $1/R$ becomes lighter, the upper tail order of the corresponding Archimedean copula becomes larger, and thus the dependence in the upper tail becomes weaker and can even negative. When $1/R \in \mathrm{MDA}(\Lambda)$, the right tail of $1/R$ becomes even lighter. In this case, we may naturally expect that the upper tail order of the corresponding Archimedean copula may be even larger or infinite. The following is the result in this sense.

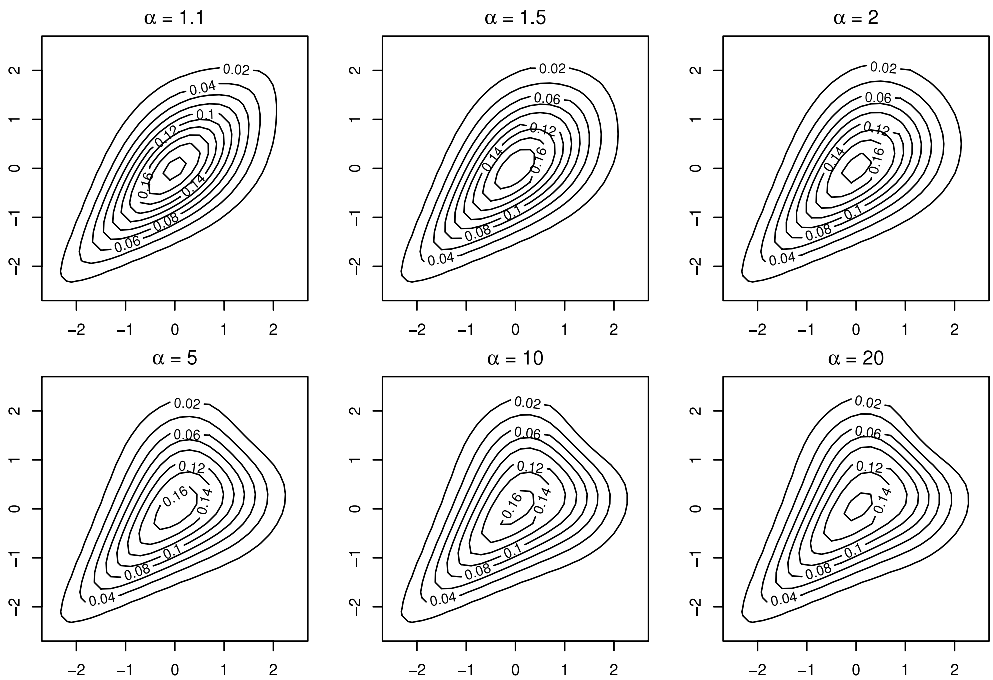

Fig. 1. Normalized contour plots of the IPS copula.

Proposition 3. *Suppose a random vector* $\boldsymbol{X} := (X_1, \ldots, X_d)$ *is defined as in (2). If* $1/R \in \mathrm{MDA}(\Lambda)$*, then the lower tail order of* $\boldsymbol{X}$ *is* $\kappa = \infty$*; that is, the upper tail order of the corresponding Archimedean copula is* $\kappa_U = \infty$.

Proof. Let $T = 1/R$, and thus $\mathbb{E}[T] < \infty$. From the proof of Proposition 1, we can write equation given in Box II which completes the proof. □

3.2. Examples

After proving the above propositions, we will have many choices of parametric distributions for the random variable R, because MDA(Fréchet) and MDA(Gumbel) are very large classes of distributions (Embrechts et al., 1997). In this section, we give some examples of parametric copulas that have upper tail negative dependence.

Example 1 (*Inverse-Pareto — Simplex Copula, aka, IPS Copula*). Let $X_i \stackrel{d}{=} RS_i$, $i = 1, 2$, (S_1, S_2) be uniformly distributed on $\{\boldsymbol{x} \geq \boldsymbol{0} : x_1 + x_2 = 1\}$, and $T := 1/R$ follow a Pareto distribution with cdf $F(x) = 1 - (1 + x)^{-\alpha}$, $x \geq 0$, $\alpha > 1$. Then the generator defined in (3) for the Archimedean copula is

$$\psi(s) = \frac{s}{1-\alpha}\left[1 - (1 + 1/s)^{-\alpha+1}\right] + 1, \quad s \geq 0, \alpha > 1.$$

Clearly, $\overline{F}_{1/R} \in \mathrm{RV}_{-\alpha}$, $\alpha > 1$. Therefore, the upper tail order of the survival copula for (X_1, X_2) is $\kappa_U = \alpha$. Depending on the value of $\alpha > 1$, this upper tail changes from intermediate tail dependence to tail negative dependence as the dependence parameter α becomes larger. Fig. 1 shows some contour plots for the IPS copula. It is clear that: (1) when $1 < \alpha < 2$, the IPS copula has intermediate tail dependence in the upper tail; (2) when $\alpha = 2$, the upper tail looks like independence; (3) when $\alpha > 2$, the upper tail appears to be negatively dependent, and a larger α indicates stronger tail negative dependence. For the lower tail, since $\psi \in \mathrm{RV}_{-1}$, by Proposition 6 of Hua and Joe (2011), the IPS copula always has lower tail order $\kappa_L = 1$, which is also consistent to the contour plots in Fig. 1.

However, in order to be useful for analyzing the expenditure dataset, the candidate copulas should possess not only upper tail negative dependence but also a lower tail that is close to independence. Further investigation will be conducted to seek simple forms of the corresponding Williamson's d-transform ψ that can lead to such an Archimedean copula. The effect of R on the lower tail of an Archimedean copula is referred to Larsson and Nešlehová (2011). After considering the upper and lower tails together, we construct a copula in Example 2.

Example 2 (*Generalized-gamma — Simplex Mixture, aka, GGS Copula*). Let $X_i \stackrel{d}{=} RS_i$, $i = 1, 2$, (S_1, S_2) be uniformly distributed on $\{\boldsymbol{x} \geq 0 : x_1 + x_2 = 1\}$, and $R^{1/\beta}$ follow a gamma distribution with shape parameter α so that

$$F_R(x) = \frac{1}{\beta\Gamma(\alpha)}\int_0^x s^{\alpha/\beta-1}\exp\{-s^{1/\beta}\}ds \quad \alpha > 0, \beta > 0, \tag{7}$$

and the Archimedean generator is

$$\begin{aligned}\psi(s) &= \int_s^\infty (1 - s/r)F_R(dr)\\ &= \frac{1}{\Gamma(\alpha)}\left(\Gamma(\alpha, s^{1/\beta}) - s\Gamma(\alpha-\beta, s^{1/\beta})\right),\end{aligned} \tag{8}$$

where $\Gamma(\cdot, \cdot)$ is an upper incomplete gamma function. Note that, although $\Gamma(0) = \infty$, the case with $\alpha = \beta$ is also implementable, and details about the implementation is in the Appendix. Some contour plots of the GGS copula are illustrated in Fig. 2.

Clearly, as $x \to 0^+$, $F_R(x) \sim [\alpha\Gamma(\alpha)]^{-1}x^{\alpha/\beta}$. Then by Proposition 1, the upper tail order of the corresponding Archimedean copula is $\kappa_U = \max\{\alpha/\beta, 1\}$; note that, Corollary 2 of Larsson and Nešlehová (2011) shows that there is upper tail dependence if $\alpha < \beta$. If $R \in \mathrm{MDA}(\Lambda)$ with an auxiliary function $a(\cdot)$, then by Proposition 7 of Larsson and Nešlehová (2011), the lower tail order $\kappa_L = 2^{1-\gamma}$, where γ is the index such that the auxiliary function $a \in RV_\gamma$. A Weibull distribution with cdf $1 - \exp\{-x^{1/\beta}\}$, $\beta > 0$, belongs to MDA of Gumbel, with an auxiliary function $a_*(x) = \beta x^{1-1/\beta}$ (Embrechts et al., 1997); $a_* \in \mathrm{RV}_{1-1/\beta}$. Then by Lemma 1 of Larsson and Nešlehová (2011), a_* can also

$$\begin{aligned}\kappa &= \lim_{y\to\infty} \frac{\log(\mathbb{P}[T\ge y]) + \log\left(\int_{\mathbf{s}\ge\mathbf{0},\|\mathbf{s}\|_1=1} \mathbb{P}[T\ge s_* y]/\mathbb{P}[T\ge y] F_{\mathbf{S}}(ds_1,\dots,ds_d)\right)}{-\log(y) - \log(\mathrm{B}(1,d-1)) + \log\left(\int_0^y \mathbb{P}[T\ge x](1-x/y)^{d-2}dx\right)} \\ &\ge \lim_{y\to\infty} \frac{\log(\mathbb{P}[T\ge y]) + \log\left(\int_{\mathbf{s}\ge\mathbf{0},\|\mathbf{s}\|_1=1} 1 F_{\mathbf{S}}(ds_1,\dots,ds_d)\right)}{-\log(y) - \log(\mathrm{B}(1,d-1)) + \log\left(\int_0^y \mathbb{P}[T\ge x](1-x/y)^{d-2}dx\right)} \\ &= \lim_{y\to\infty} \frac{\log(\mathbb{P}[T\ge y])}{-\log(y)} = \infty,\end{aligned}$$

Box II.

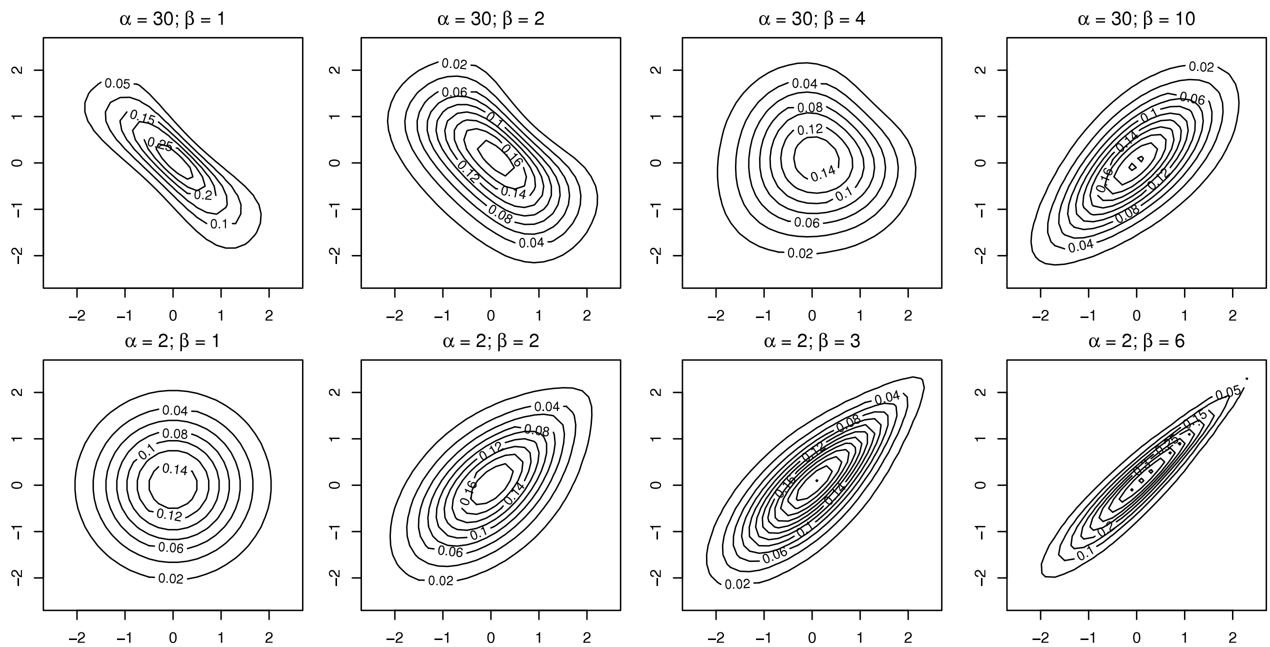

Fig. 2. Normalized contour plots of the GGS copula.

be an auxiliary function for the survival function of R. Therefore, $\gamma = 1 - 1/\beta$, and $\kappa_L = 2^{1/\beta}$. This copula can provide very flexible upper and lower tails, ranging from tail positive to tail negative dependence. Note that, when $\alpha = 2$ and $\beta = 1$, $\kappa_L = \kappa_U = 2$, and moreover, $\psi(s) = \Gamma(2,s) - s\Gamma(1,s) = \exp(-s)$, which is the generator of the independence copula. So the independence copula is a special case of the GGS copula.

We can also re-parameterize the GGS copula by the upper and lower tail orders; that is, $\alpha = \kappa_U \ln(2)/\ln(\kappa_L)$, and $\beta = \ln(2)/\ln(\kappa_L)$.

Note that, the generalized gamma distribution is a special case of the generalized beta distribution of the second kind (GB2). Therefore, a more flexible dependence structure (with more parameters) can be constructed by using GB2 as the scaling random variable. The MGB2 copula studied in Yang et al. (2011) is induced by conditionally independent generalized gamma distributions where the conditioning random variable is used to distort one shape parameter of the generalized gamma distribution, and therefore, the mechanism is completely different than the scale mixture model studied in this paper.

Example 3 (*Inverse-gamma — Simplex Mixture, Example 2 of McNeil and Nešlehová, 2010*). Let $X_i \overset{d}{=} RS_i$, $i = 1, 2$, (S_1, S_2) be uniformly distributed on $\{\boldsymbol{x} \ge 0 : x_1 + x_2 = 1\}$, and $1/R$ follow a gamma distribution with shape parameter α and scale parameter 1 so that the generator of the Archimedean copula is

$$\psi(s) = \frac{\gamma(\theta, 1/s)}{\Gamma(\theta)} - \frac{s\gamma(\theta+1, 1/s)}{\Gamma(\theta)}.$$

By Proposition 3, since $1/R \in$ MDA(Gumbel), the upper tail order is $\kappa_U = \infty$, which implies that the upper tail is always negatively dependent. For the lower tail, since R follows an inverse gamma distribution with shape parameter α and scale parameter 1, due to (6), $\overline{F}_R \in \mathrm{RV}_{-\alpha}$. Therefore, by Theorem 1(a) of Larsson and Nešlehová (2011), the corresponding generator $\psi \in \mathrm{RV}_{-\alpha}$, which implies lower tail dependence (see Hua and Joe, 2011; Larsson and Nešlehová, 2011). The limitation of this copula is that there are no parameters to control the upper tail order that always has tail negative dependence with $\kappa_U = \infty$.

4. Aggregate loss — data analysis

4.1. Introduction

The dataset we are analyzing is based on Panel 14 and Panel 15 for the calendar year of 2010 from the 2010 Full Year Consolidated Data File. The dataset was collected on a nationally representative sample of the civilian noninstitutionalized population of the United States.

Based on the description of the data file that was provided by the Agency for Healthcare Research and Quality, it consists of MEPS survey data obtained in Rounds 3–5 of Panel 14 and Rounds 1–3 of Panel 15 (i.e., the rounds for the MEPS panels covering calendar year 2010), and consolidates all of the final 2010 person-level variables. The person ID number (`DUPERSID`) can be used to identify the same person across panels. We have checked that, in the consolidated data file, there are no repeated measures on

Table 1
Summary of the variables.

	Min	1st quantile		Median		Mean	3rd quantile			Max
OPVEXP10	6		183		717	2415		2353		68,370
Average expense	6		132		394	1510		1617		36,680
Age	0		22		42	38		54		64
Number of visits	1	2	3	4	5	6	7	8	9	10
OPDRV10 (#obs)	1127	298	106	54	46	22	18	6	8	5
	11	12	13	14	15	16	17	19	19	20
	4	3	2	2	1	1	0	1	2	1
	21	22	23	25	29	31	32	35	40	42
	1	1	1	1	1	1	1	1	1	2
	46	48	65	98	-	-	-	-	-	-
	1	1	1	1	-	-	-	-	-	-
Insurance coverage		Any private (1)			Public only (2)			Uninsured (3)		
INSCOV10 (#obs)		1077			497			148		
Race		Hispanic (1)		Black (2)		Asian (3)			Other (4)	
RACETHNX (#obs)		391		356		69			906	

the person-level. To illustrate the empirical observation of upper tail negative dependence, we now consider the variables of the number of outpatient department visits to physicians in 2010 (OPDRV10), and the associated facility expenses (OPVEXP10). The average expense per visit used in the data analysis is calculated by the ratio between OPVEXP10 and OPDRV10. We use the average expense as loss severity and the number of visits as loss frequency. There are 32,846 individuals totally, among whom 1722 are below age 65 and have positive numbers of outpatient visits to physicians and positive facility expenses. The records with ages equal to or greater than 65 are excluded, as the Medicare program becomes available for those people, and the utilization patterns of outpatient expenditures would be very different than that of those younger than 65. As discussed in Frees et al. (2011), patients usually make the decision of whether to use medical services, but once they decide to seek treatment, physicians mainly decide on the intensity of expenditures. So it is reasonable to consider the "zeros" and "non-zeros" separately. The dependence structure is better captured by the copula for non-zero data, so the records with response variables being zero are excluded in the analysis. Binary regression, such as logistic regression, can be applied separately if one wants to do inference for an overall population.

Descriptive statistics of the variables are in Table 1. The scatter plot on the original scale for average expense and number of visits is in Fig. 3, from which one may think of an independence structure between the two variables. However, the dependence pattern cannot be observed clearly using the original scale, due to the discrete nature of the frequency data.

In order to visualize the dependence pattern more intuitively, we add tiny random noises (Normal$(0, 1)/1000$) on the numbers of visits to make them look continuous, and then transform the expenses and continualized number of visits respectively into normal scores that are distributed as a standard Normal distribution. One can also use some other techniques such as jitters or the technique used in getting the normalized QQ plot for discrete variables in Fig. 5. The dependence pattern is illustrated in the left panel of Fig. 4. Although the plot is not based on the original data, the pattern of upper tail negative dependence is inherited from the original data. We only use the technique to visualize the dependence structure, and we still use the original number of visits as a discrete variable when we fit the model in what follows. Note that, copula is generally not unique for discrete variables. However, for a joint cdf $F(y_1, \ldots, y_d) = C(F_1(y_1), \ldots, F_d(y_d))$ with non-continuous or discrete marginal F_i's, the copula C is unique on $\mathrm{Range}(F_1) \times \cdots \times \mathrm{Range}(F_d)$. Note that the number of categories of

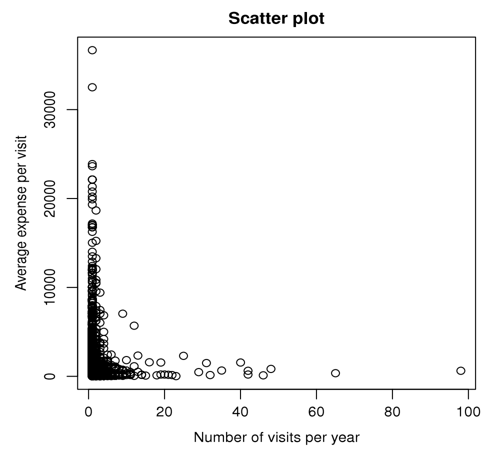

Fig. 3. Scatter plot on the original scale.

the variable "number of visits" is fairly large, and the use of copulas does not impose any restriction.

It seems that there is tail negative dependence in the upper tail, and independence in the other parts. The reason could be flat fees or overhead charges, so that a lower average cost tends to be associated with a larger number of visits. In practice, one needs to be aware of the differences between commonly-used models that are able to account for negative associations (e.g., elliptical copulas) and the models that can have asymmetric tail negative dependence (e.g., the GGS copula). The former is symmetric between upper and lower tails, while the latter has flexible local dependence structures in the upper and lower tails, respectively.

Following the approach proposed in Czado et al. (2012) and Krämer et al. (2013), we now use a mixed copula model to conduct a regression analysis for the aggregate losses. We use the Zipf distribution (Zipf, 1932) to model the loss frequency, a lognormal model for the loss severity, and the GGS copula to model the dependence structure between loss severity and loss frequency. Since our proposed copula model is able to account for the upper tail negative dependence pattern, where larger losses occur, the regression analysis based on the GGS copula is expected to outperform the analysis based on the existing copulas.

For both loss frequency (N) and loss severity (Y), we have considered ages, incomes, sex, education, insurance coverage and races for the covariates. We have tried several marginal regression models such as gamma and lognormal distributions for Y, and zero-truncated Poisson, zero-truncated negative binomial, and

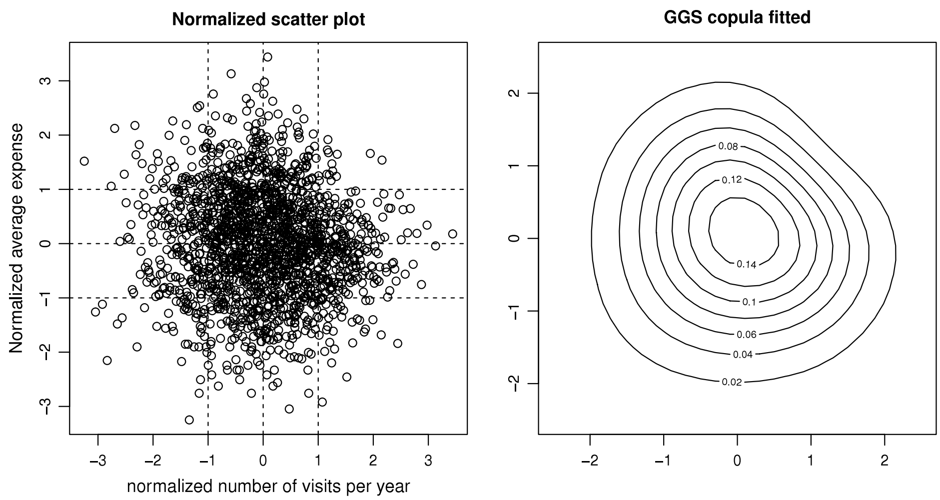

Fig. 4. Asymmetric upper tail negative dependence between loss frequency and severity. In the left plot, the marginals are transformed to standard normals; in the right plot, the pseudo data is fitted by the GGS copula. It is clear that when the number of visits is larger, the relation between the number of visits and average expenses becomes more negatively dependent, while the rest parts seem to be independent. That is, there is upper tail negative dependence between the loss frequency and severity data. Spearman's ρ calculated for the four quadrants $(++)$, $(-+)$, $(--)$ and $(+-)$ are respectively -0.046, 0.031, 0.041 and -0.010, which also suggests a relatively stronger degree of negative dependence in the upper tail.

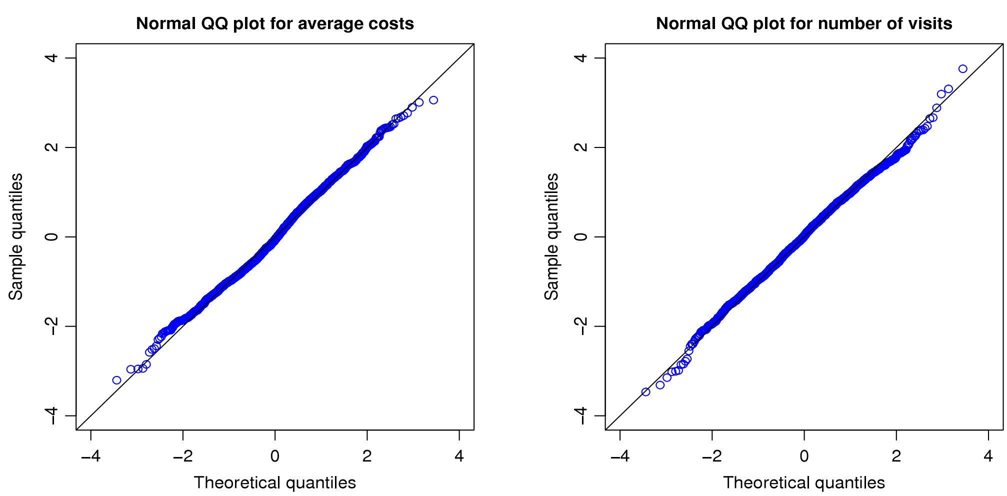

Fig. 5. Normalized QQ plots of quantile residuals for regression on marginals: the left panel is the plot for lognormal regression on average expenses, and the right panel is the plot for Zipf regression on number of visits per year.

Zipf distributions for N. The right tail of the gamma distribution is too light for the average loss Y, and the right tails of the zero-truncated Poisson and zero-truncated negative binomial are also too light for capturing the heavy tails of the loss frequencies. Based on preliminary data analysis, we finally choose age, insurance coverages, and races as the covariates for both N and Y. The other covariates are either non-significant or leading to relatively worse AICs. Note that, for the mixed copula method, the covariates for frequency and severity are not necessarily the same, although we choose the same set of covariates for our data analysis. We have not compared all the distributions for the marginal regression models, but the lognormal and Zipf's models are good enough for capturing the main features in the dataset. In the following, we discuss the two marginal regression analyses, respectively.

4.2. Marginal regression

In the literature, such as in Frees et al. (2011), negative binomial or Poisson distributions are often used to model the number of outpatient visits when zeros are included. The proposed GGS copula focuses on positive medical expenditures, so zero-truncated count distributions would be useful. Here we remark that, for the MEPS data, the percentage of zeros in loss frequency is often not well covered by the probability of "zeros" associated with the Poisson or negative binomial distributions, and zero-inflated count distributions are often needed when the zeros are included for inference; see Xia and Gustafson (2014) and its web-based materials for detailed discussion about this issue.

We have conducted regression analysis for the number of outpatient visits using zero-truncated Poisson, zero-truncated negative binomial, and Zipf distributions, respectively. We used the normalized QQ plots to compare the performance of these three models. Both zero-truncated Poisson and zero-truncated negative binomial models underestimated the upper tail heaviness; see Fig. 6 for the performance of zero-truncated Poisson and zero-truncated negative binomial models. The Zipf model was the best among the three models, and the main reason was that the Zipf distribution is a discretized Pareto distribution so that it is able to account for the heavy tail of the data.

Now, denote a covariate vector as $\boldsymbol{x}_i, i = 1, \ldots, 1722$. The marginal regression model for the loss frequency using the Zipf distribution can be written as

$$f_N(n|s,m) = \mathbb{P}[N=n|s,m] = \frac{n^{-s}}{\sum_{i=1}^{m} i^{-s}}, \quad n=1,2,\ldots,m,\ s>0, \tag{9}$$

where $s > 0$ is the parameter of the Zipf distribution, and m is the maximum value of N; for this dataset, we chose $m = 98$, the maximum number of visits in the dataset. The Zipf distribution has a power law, which means that the right tail of the distribution is heavier than the commonly-used Poisson distribution. The Zipf distribution can be looked at as a discretized Pareto distribution, and the value of s determines the degree of tail heaviness: a larger s corresponds to a lighter right tail of the distribution, and vice versa. Moreover, $m = 98$ is the maximum number of visits in the dataset, which conforms to the definition of the Zipf distribution. Although in practice one can specify a larger m to allow conservative out-of-sample prediction for the number of visits, there seems to be no particular reason to specify a number other than the maximum value observed in the data, as now the main task is to model the current data. The covariates are introduced as follows:

$$\ln(s_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\eta}, \quad i = 1,\ldots,1722,$$

where $\boldsymbol{\eta}$ is the regression coefficients for the loss frequency (including the intercept term). We use a linear form in $\boldsymbol{x}_i$ to demonstrate how to apply the regression model. However, a nonlinear relation between $\ln(s_i)$ and $\boldsymbol{x}_i$ could lead to a better fitting, after a transformation being applied on $\boldsymbol{x}_i$ (e.g., using polynomials or splines).

The lognormal model for the loss severity can be written as

$$f_Y(y|\mu,\sigma) = \frac{1}{\sigma y\sqrt{2\pi}} \exp\left\{-\frac{(\ln y-\mu)^2}{2\sigma^2}\right\},$$

where μ is the location parameter and σ is the scale parameter, and the covariates are introduced through the following equation:

$$\mu_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}, \quad i = 1,\ldots,1722,$$

where $\boldsymbol{\gamma}$ is the corresponding regression coefficients (including the intercept term). We assume a homogeneous scale parameter regardless of the values of covariates.

The maximum likelihood estimates (MLEs) of the regression coefficients and the location parameters are reported in Table 2. In order to diagnose how well the two regression models fit the dataset, we use normalized QQ plots of quantile residuals (Dunn and Smyth, 1996) that are in Fig. 5. From Fig. 5, we find that the lognormal regression and Zipf regression models fit the dataset quite well. So, in the copula regression model, we will keep using these two distributions for fitting the marginals, while incorporating the GGS copula for capturing the dependence structure.

Here we have both continuous and discrete response variables. The procedures of getting normalized QQ plots are different for continuous and for discrete variables. Here we only briefly introduce the main steps used for the discrete case. We refer the reader to Dunn and Smyth (1996) for more details. The procedure for the normalized QQ plot of the number of visits is that: (1) derive the corresponding cumulative probabilities for each response variable y and for $y-1$, respectively; (2) randomly sample a probability from a uniform distribution with the endpoints of the uniform distribution being the corresponding cumulative probabilities for y and $y-1$; (3) transform the probability obtained from step (2) by Φ^{-1} to get the sample quantiles; (4) plot the sample quantiles against the theoretical quantiles that can be obtained as for continuous variables.

Table 2
Maximum Likelihood Estimates of marginal and copula regression models, where AIC $= -2 \times$ log likelihood $+ 2 \times$ number of parameters.

		Marginal	s.e.	GGS	s.e.
Frequency	Intercept	0.885	0.046	0.881	0.045
	age	−0.003	0.001	−0.003	0.001
	ins(2)	−0.120	0.036	−0.109	0.035
	ins(3)	−0.085	0.054	−0.112	0.054
	race(2)	0.042	0.045	0.034	0.044
	race(3)	0.076	0.083	0.081	0.080
	race(4)	0.146	0.039	0.130	0.038
Severity	Intercept	5.737	0.111	5.744	0.108
	age	0.011	0.002	0.010	0.002
	ins(2)	−0.747	0.086	−0.732	0.085
	ins(3)	−0.631	0.133	−0.546	0.133
	race(2)	0.062	0.111	0.099	0.109
	race(3)	−0.019	0.197	0.009	0.193
	race(4)	0.422	0.094	0.437	0.093
	$\ln(\sigma)$	0.403	0.017	0.405	0.017
Dependence	$\ln(\alpha)$	–	–	3.951	1.733
	$\ln(\beta)$	–	–	1.643	0.896
AIC		31,743.68		31,705.27	

4.3. Copula regression

The dependence structure between loss severity and loss frequency can be accounted for by copulas. In Krämer et al. (2013), the joint density functions of continuous loss severity Y and discrete loss frequency N can be written as

$$f_{Y,N}(y,n|\theta) = f_Y\left[D_1(F_Y(y),F_N(n)|\theta) - D_1(F_Y(y),F_N(n-1)|\theta)\right],$$

where $D_1(u,v|\theta) := \partial C(u,v|\theta)/\partial u$, C is the copula for loss severity and loss frequency, and θ is the set of the dependence parameter(s).

Due to the limitation of space, we refer the reader to Krämer et al. (2013) for a reference on how to incorporate a copula into the two marginal regression models so that we can obtain the MLEs for those regression coefficients $(\boldsymbol{\eta},\boldsymbol{\gamma})$ and for the dependence parameters (α,β). The MLEs are reported in Table 2, where the likelihood for the GGS regression model is based on the overall model with marginals and dependence being fitted simultaneously.

Based on the MLEs of the dependence parameters, we can get the estimated upper and lower tail orders respectively as $\hat{\kappa}_U = 10.05$ and $\hat{\kappa}_L = 1.14$. It is clear that the upper tail order is greater than 2, the dimension. So, it has upper tail negative dependence, and this is consistent with the preliminary plot in the left panel of Fig. 4. While the estimated lower tail order suggests that the dependence structure in the lower tail is positive dependence. The negative regression coefficient associated with the variable `COUNT_OP` in Table 4 of Frees et al. (2011) also partially supports our findings, and the proposed GGS copula further detects that the negative dependence is only in the upper tail.

The likelihood used to calculate the AIC for the mixed copula approach is the likelihood that was obtained when fitting the regression coefficients and the dependence parameters simultaneously. The likelihood used to get the AIC for the independence assumption is the product of the likelihoods associated with the two marginal regression models. Since the GGS copula includes the independence copula as a special case, we can also let the two dependence parameters of the GGS copula be $\alpha = 2$ and $\beta = 1$ to calculate the likelihood of the model for the independence assumption, which actually lead to the same results as reported in Table 2. It is clear that, among these two nested models, the mixed copula model based on the GGS copula fits the overall model relatively better with a smaller AIC.

In order to estimate the aggregate loss, we have two ways based on two different assumptions. First, we can use the mixed

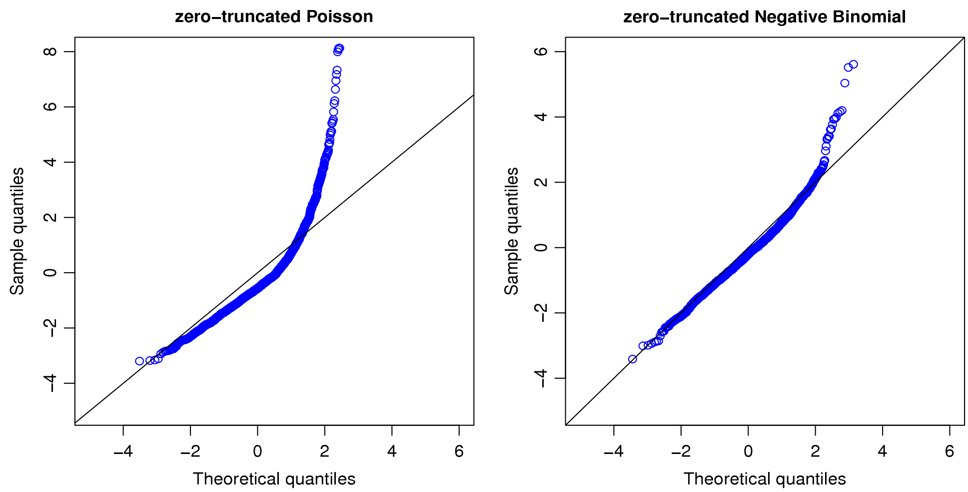

Fig. 6. Normalized QQ plots of quantile residuals for regression on numbers of visits per year: zero-truncated Poisson and zero-truncated negative Binomial models.

copula approach assuming that loss frequency and loss severity are dependent (Krämer et al., 2013) according to the dependence parameters we have obtained from the GGS copula. Second, we may assume that loss frequency and loss severity are independent, conduct the marginal regressions separately, and use the sum of the products between expected number of visits and expected cost per visit as the aggregate loss. The estimated aggregate loss based on the GGS copula was 4,362,626, the estimated aggregate loss based on the independence model was 6,147,354, and the actual loss is 4,159,322. The in-sample estimated aggregate loss based on the GGS copula is far closer to the actual one, which also suggests that the proposed model outperforms its alternative. The empirical analysis suggests that, when the upper tail appears to be negatively dependent, a misspecified independence model that is often used in aggregate loss modeling may overestimate the aggregate loss. Our results also supplement a finding from an auto insurance claim data analyzed in Krämer et al. (2013), where a mild positive dependence between loss frequency and loss severity results in an underestimation of the aggregate loss by the independence model. For aggregate loss modeling accounting for both loss severity and frequency, the dependence structure between them should be taken good care of. Otherwise, modeling the total loss directly as the response variable may be even better.

The MEPS dataset used for empirical analysis actually contains the total expense per year for each individual. For this particular dataset, one can choose to model the aggregate loss and the number of visits jointly as the bivariate response variable, then a copula covering positive upper tail dependence can be used. As suggested by the above empirical analysis, the dependence between yearly total expenses and the number of visits per year is not linear, and conditioning on different values of covariates there may be different degrees of positive dependence. To this end, one can also consider a full-range tail dependence copula and allow the dependence parameter change along different values of covariates. We refer the reader to Hua and Xia (2014) for a more detailed discussion on regression on dependence parameters with full-range tail dependence copulas.

5. Concluding remark

From insurance data, one often observes two unique non-Gaussian phenomenons. First, the univariate marginals are often skewed and the right distributional tails could be light or heavy. Second, the dependence structure between univariate marginals or between their transformed forms often cannot be well captured by a covariance matrix. Copula has been proven to be a very useful tool in dealing with these situations. Moreover, statistical inference on high-risk scenarios often plays a critical role for risk management. For example, the dependence structure in the upper tail for the case we studied in Section 4 influences the overall assessment dramatically. Therefore, when choosing a copula for modeling dependence structures, we have to be particularly careful about the tail behavior of the candidate copula models. An ideal candidate copula shall be the one that has less number of parameters but wider range of dependence in both upper and lower tails, and the range of dependence of the copula shall be able to cover the actual range of dependence suggested by the observed data and beyond. Moreover, implementation of the copula should be achievable.

Keeping these criteria in mind, we first narrow down the families of copulas and consider only Archimedean copulas, because it is suitable for constructing asymmetric dependence structures between upper and lower tails, and for incorporating tail negative dependence as well. Some sufficient conditions for upper tail negative dependence have been derived from a scale mixture representation of Archimedean copulas. Through theoretical study on copulas, we have constructed new parametric copulas that have simple forms and desirable properties. We implement the GGS copula into a mixed copula regression model, and apply it on a medical expenditure dataset that displays an upper tail negative dependence pattern between loss severity and loss frequency. The mixed copula regression with the GGS copula provides a significantly better assessment for total losses. However, assuming that the loss severity and frequency are independent overestimates the total losses.

Since the yearly total expenditure is actually available for the dataset, one may consider using a Tweedie model (Tweedie, 1984) for modeling the aggregate loss directly. A comparison between the Tweedie regression and the mixed copula regression based on the GGS copula would be interesting. Moreover, a more challenging and interesting question is that, can one or two dependence parameters be added in the Tweedie model so that loss frequency, loss severity and their dependence structures can be taken into consideration simultaneously?

Acknowledgments

The research conducted was partially supported by a grant from the Casualty Actuarial Society through an Individual Grants Competition, and partially supported by the Research and Artistry Grant from Northern Illinois University. The author is thankful to the precious time of the reviewers for the above grants, to the helpful comments from Professor Edward W. Frees on a preliminary manuscript, and to the excellent reviewer's report provided by an anonymous reviewer. All remaining errors are the author's own.

Appendix

For a bivariate Archimedean copula $C(u, v) = \psi(\psi^{-1}(u) + \psi^{-1}(v))$, the density function can be written as

$$c(u, v) = \frac{\psi''(\psi^{-1}(u) + \psi^{-1}(v))}{\psi'(\psi^{-1}(u))\psi'(\psi^{-1}(v))}. \tag{10}$$

For the GGS copula, based on (8),

$$\psi'(x) = -\Gamma(\alpha - \beta, x^{1/\beta})/\Gamma(\alpha);$$
$$\psi''(x) = x^{\alpha/\beta-2}\exp\{-x^{1/\beta}\}/(\beta\Gamma(\alpha)).$$

Then we can derive the density function of the copula in (10), where the inverse ψ^{-1} can be obtained by a numerical method. Since here ψ is a strict monotone function, a numerical method usually works very well for deriving the inverse ψ^{-1}. In order to avoid calculating $\Gamma(\alpha)$ for a large α (e.g., $\Gamma(\alpha)$ cannot be calculated in the R software (R Development Core Team, 2012) when $\alpha \geq 172$ on a Windows 7 32-bit operating system), we use the following Eq. (11) for calculating $\Gamma(\alpha, x)/\Gamma(\alpha)$ so that α can be $\geq$172. The calculation actually can be done for $\alpha \in \Re$ by noting that

$$\Gamma(\alpha, x) = (\alpha - 1)\Gamma(\alpha - 1, x) + x^{\alpha-1}e^{-x};$$
$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1);$$
$$\gamma(\alpha, x) = (\alpha - 1)\gamma(\alpha - 1, x) - x^{\alpha-1}e^{-x},$$

where $\alpha \in \Re$ and $\alpha \notin \{-1, -2, \ldots\}$; we refer to Fisher et al. (2003) for a relevant discussion for the cases where α can be a negative integer.

The following are approaches to dealing with numerical issues such as large values of parameters. Letting $[\alpha] := \max\{z \text{ integer} : z < \alpha\}$, and $\xi := \alpha - [\alpha]$,

$$\begin{aligned}\frac{\Gamma(\alpha, x)}{\Gamma(\alpha)} &= \frac{\Gamma(\xi, x)}{\Gamma(\xi)} \\ &\quad + \frac{e^{-x}}{\Gamma(\xi)} \times \frac{x^{\xi}}{\xi} \times \left(1 + \frac{x}{1+\xi}\left(1 + \frac{x}{2+\xi}\right.\right. \\ &\quad \times \left.\left.\left(1 + \cdots + \frac{x}{[\alpha]-2+\xi}\left(1 + \frac{x}{[\alpha]-1+\xi}\right)\right)\right)\right).\end{aligned} \tag{11}$$

To calculate $\Gamma(\alpha - \beta, x)/\Gamma(\alpha)$, we can write

$$\begin{aligned}\frac{\Gamma(\alpha-\beta, x)}{\Gamma(\alpha)} &= \frac{\Gamma(\alpha-\beta, x)}{\Gamma(\alpha-\beta)} \times \frac{\Gamma(\alpha-\beta)}{\Gamma(\alpha)} = \frac{\Gamma(\alpha-\beta, x)}{\Gamma(\alpha-\beta)} \\ &\times \frac{(\alpha-\beta-1)\times\cdots\times(\alpha-\beta-[\alpha-\beta])\times\Gamma(\alpha-\beta-[\alpha-\beta])}{(\alpha-1)\times\cdots\times(\alpha-[\alpha])\times\Gamma(\alpha-[\alpha])},\end{aligned} \tag{12}$$

and then the first term in (12) can be calculated using Eq. (11) again by replacing α by $\alpha - \beta$.

For calculating $\psi''(x)$, when x is too large, there could be numerical errors for calculating $x^{\alpha/\beta-2}$. So, we need to find $\ln(\psi''(x))$ first; that is,

$$\ln(\psi''(x)) = (\alpha/\beta - 2)\ln(x) - x^{1/\beta} - \ln(\beta) - \ln(\Gamma(\alpha)).$$

The following Fig. 6 shows the unsatisfactory performance of zero-truncated Poisson and zero-truncated negative binomial models for the loss frequency.

References

Abramowitz, M., Stegun, I.A., 1964. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Number 1972. Dover Publications.
Bingham, N.H., Goldie, C.M., Teugels, J.L., 1987. Regular Variation. In: Encyclopedia of Mathematics and its Applications, vol. 27. Cambridge University Press, Cambridge.
Charpentier, A., Segers, J., 2009. Tails of multivariate Archimedean copulas. J. Multivariate Anal. 100 (7), 1521–1537.
Czado, C., Kastenmeier, R., Brechmann, E.C., Min, A., 2012. A mixed copula model for insurance claims and claim sizes. Scand. Actuar. J. 2012 (4), 278–305.
Dunn, P.K., Smyth, G.K., 1996. Randomized quantile residuals. J. Comput. Graph. Statist. 5 (3), 236–244.
Embrechts, P., Goldie, C.M., 1980. On closure and factorisation properties of subexponential and related distributions. J. Aust. Math. Soc. Ser. A 29, 243–256.
Embrechts, P., Klüppelberg, C., Mikosch, T., 1997. Modelling Extremal Events, Vol. 33. Springer-Verlag, Berlin.
Ferguson, T., 1973. A bayesian analysis of some nonparametric problems. Ann. Statist. 209–230.
Fisher, B., Jolevsaka-Tuneska, B., KiliÇman, A., 2003. On defining the incomplete gamma function. Integral Transforms Spec. Funct. 14 (4), 293–299.
Frees, E.W., 2010. Regression Modeling with Actuarial and Financial Applications. Cambridge University Press.
Frees, E.W., Gao, J., Rosenberg, M.A., 2011. Predicting the frequency and amount of health care expenditures. N. Am. Actuar. J. 15 (3), 377–392.
Gschlößl, S., Czado, C., 2007. Spatial modelling of claim frequency and claim size in non-life insurance. Scand. Actuar. J. 2007 (3), 202–225.
Hua, L., Joe, H., 2011. Tail order and intermediate tail dependence of multivariate copulas. J. Multivariate Anal. 102, 1454–1471.
Hua, L., Joe, H., 2012. Tail comonotonicity: properties, constructions, and asymptotic additivity of risk measures. Insurance Math. Econom. 51, 492–503.
Hua, L., Joe, H., 2013. Intermediate tail dependence: a review and some new results. In: Li, H., Li, X. (Eds.), Stochastic Orders in Reliability and Risk: In Honor of Professor Moshe Shaked. Springer, pp. 291–311 (chapter 15).
Hua, L., Xia, M., 2014. Assessing high-risk scenarios by full-range tail dependence copulas. N. Am. Actuar. J. 18 (3), 363–378.
Joe, H., 1997. Multivariate Models and Dependence Concepts. In: Monographs on Statistics and Applied Probability, vol. 73. Chapman & Hall, London.
Joe, H., Hu, T., 1996. Multivariate distributions from mixtures of max-infinitely divisible distributions. J. Multivariate Anal. 57 (2), 240–265.
Krämer, N., Brechmann, E.C., Silvestrini, D., Czado, C., 2013. Total loss estimation using copula-based regression models. Insurance Math. Econom. 53 (3), 829–839.
Larsson, M., Nešlehová, J., 2011. Extremal behavior of archimedean copulas. Adv. Appl. Probab. 43 (1), 195–216.
Malov, S., 2001. On finite-dimensional archimedean copulas. In: Asymptotic Methods in Probability and Statistics with Applications. Birkhauser, pp. 19–35.
McNeil, A.J., Nešlehová, J., 2009. Multivariate Archimedean copulas, d-monotone functions and l_1-norm symmetric distributions. Ann. Statist. 37 (5B), 3059–3097.
McNeil, A.J., Nešlehová, J., 2010. From archimedean to liouville copulas. J. Multivariate Anal. 101 (8), 1772–1790.
Nelsen, R.B., 2006. An Introduction to Copulas, second ed. In: Springer Series in Statistics, Springer, New York.
R Development Core Team 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN: 3-900051-07-0.
Tweedie, M., 1984. An index which distinguishes between some important exponential families. In Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference, pp. 579–604.
Xia, M., Gustafson, P., 2014. Bayesian sensitivity analyses for hidden sub-populations in weighted sampling. Canad. J. Statist. 42 (3), 436–450.
Yang, X., Frees, E., Zhang, Z., 2011. A generalized beta copula with applications in modeling multivariate long-tailed data. Insurance Math. Econom. 49, 265–284.
Zipf, G.K., 1932. Selected Studies of The Principle of Relative Frequency in Language. Harvard Univ. Press.