Copula函数在非寿险费率厘定中的应用

郭莲丽[1]，李建勋[2](副教授)

(1.陕西广播电视大学工商管理系，西安 710119；2.西安理工大学经济与管理学院，西安 710048)

【摘要】针对离散边际分布的Copula连接问题，采用扰动量将离散随机变量转化为连续随机变量，给出n元离散Copula函数和连续Copula函数之间的关系，建立连续Copula函数对离散边际分布的连接，并以Clayton Copula为例，结合最大似然估计，给出回归模型及实证分析，结果表明连接后回归模型具有和D-Vine相近的拟合效果，并且可方便地使用零膨胀思想来提高多零情况下的拟合优良性。

【关键词】Copula函数；离散边际分布；联合分布

一、引言

近年来，将Copula函数应用到金融领域已受到学者们的普遍关注。Copula函数在解决如何描述市场相关性问题的同时，也为构建多元函数联合分布提供了一种可行方法，并且可以把边际分布函数连接成联合分布函数。Sklar(1958)和Nelsen(2006)对Copula函数进行了详细的理论总结后，学者们广泛地将Copula函数应用在投资组合、风险管理、资产定价、金融市场相关性问题等方面。目前，随着相关性测度、尾部相关性理论被相继引入，椭圆Copula函数族中的Gaussian Copula、t-Copula，Archimedean Copula函数族的Gumbel Copula、Clayton Copula、Frank Copula以及极值Copula得到了更深层次的应用。Jondeau和Rockinger(2006)将Copula函数应用在股票市场，研究了股票回报率之间的依赖模型；Hofert和Scherer(2011)利用Archimedean Copula函数建立了CDO定价模型；Eling和Toplek(2009)将Copula函数应用到动态的财务分析之中，构建了内部风险模型。在高维数据建模方面，Pair-Copula(也称Vine Copula或藤Copula)则更为灵活。Bedford(2001)和Aas(2009)等曾对藤Copula构建及其参数估计和数据模拟方法分别做了详细介绍，文献还提出了一种新型的藤Copula创建方法，以构造出更多的n维藤Copula，有力地推动了Copula理论的发展。在国内，早在2002年就对Copula相关理论进行了介绍，韦艳华(2003)等还讨论了Copula理论与面向均值方差和线性相关的建模方法的不同，通过Copula理论建立了多变量金融时间序列模型来代替GARCH向量模型。史道济则引入了变量之间的Spearmah′s Ro秩、Kendall秩等相关性测度。另有周孝华(2012)、谢赤(2013)、陆静(2013)等众多科研人员还在Copula应用方面进行了研究工作。

纵观现有研究，经过多年的发展，连续Copula函数的探索已经臻于成熟。而离散Copula函数因其并非唯一，且在多元情况下复杂度更高，也难以直接利用连续Copula函数的成果进行离散数据的分析，故而限定了适用范围。

例如，在非寿险风险管理中，索赔数据往往是过离散的，这时连续Copula函数对此问题往往束手无策。为此本文在利用Denuit和Lambert(2005)开展相关性探讨时所采用的变换方法的基础上，采用扰动量将离散随机变量转化为连续随机变量，通过演算建立了离散Copula函数和连续Copula函数之间的映射，实现连续Copula函数对离散边际分布的连接，改变以往使用D-Vine分解方式和图模式的思路，降低离散Copula求解复杂度。并以Clayton Copula为例，结合实证数据，开展最大似然估计，给出回归模型及结果比较。结果表明连接后回归模型具有和D-Vine相近的拟合效果，更加接近了实际索赔中的零次索赔数量过多以及多种费率因子共同作用的客观情况，提高了拟合分析效果，可方便地使用零膨胀思想来降低多零情况下回归结果的AIC(赤池弘次信息量准则Akaike Information Criteria)和BIC(贝叶斯信息准则Bayesian Information Criteria)。

二、离散Copula函数

Copula研究的是如何构造适当的多变量联合分布函数，使其具有指定的单变量边缘分布函数以及多个变量之间的相关性。若采用二重差分方式进行描述，n元连续Copula函数可定义为$C:[0,1]^n\to[0,1]$，并满足：

(1)$\forall u\in[0,1]^n$，若$\vec{u}$中至少有一个分量为0，则$C(\vec{u})=0$；若$\vec{u}$中除u_k外的分量均为1，则$C(\vec{u})=u_k$；

(2)$\forall \vec{a},\vec{b}\in[0,1]^n$，若$\vec{a}\leqslant\vec{b}$，则二重差分$V_c([\vec{a},\vec{b}])\geqslant 0$，其中：$\Delta_{a_k}^{b_k}C(\vec{t})=C(t_1,\cdots,t_{k-1},b_k,t_{k+1},\cdots t_n)-C(t_1,\cdots,t_{k-1},a_k,t_{k+1},\cdots t_n)$，$V_c([\vec{a},\vec{b}])=\Delta_{\vec{a}}^{\vec{b}}C(\vec{t})=$

$\triangle_{a_n}^{b_n}\triangle_{a_{n-1}}^{b_{n-1}}\cdots\triangle_{a_2}^{b_2}\triangle_{a_1}^{b_1}C(\vec{t})$。

由Sklar定理知，若H是一个联合分布函数，且其边缘分布函数和概率密度分别为$F_i(x_i)$、$f_i(x_i)$，i=1，2，…，n，那么存在一个n维Copula函数$C(t_1,t_2,\cdots,t_n)$，使得对所有的x_i有：

$$H(x;\eta)=C(F_1(x_1),F_2(x_2),\cdots,F_n(x_n)) \quad (1)$$

其中，η为参数集合。

这表明，任何一个多元连续分布函数都可以分解为它的连续边际分布和相关结构，并且通过连续Copula函数就可以把多维随机变量的联合分布用一维边际分布连接起来，从而实现了多种边际分布的集成。另外，还可以验证Copula函数本质上是一个n维分布函数，其概率密度为：

$$c(t_1,t_2,\cdots,t_n)=\frac{\partial C(t_1,t_2,\cdots,t_n)}{\partial t_1\partial t_t\cdots\partial t_n} \quad (2)$$

进而还可以得到联合分布H所对应的概率密度为：

$$h(x;\eta)=c(F_1(x_1),F_2(x_2),\cdots,F_n(x_n))\prod_{i=1}^{n}f_i(x_i) \quad (3)$$

当前，人们已经证明了当$F_i(x_i)$是连续函数时，Copula函数C可唯一确定。然而对于离散的Copula函数来说，情况更为复杂。例如二元离散Copula函数$C_{n,m}:I_n\times I_m\rightarrow[0,1]$，$i\in\{1,\cdots,n\}$，$j\in\{1,\cdots,m\}$应满足$C_{n,m}(i/n,j/m)-C_{n,m}:((i-1)/n,j/m)\leqslant 1/n$，$C_{n,m}(i/n,j/m)-C_{n,m}:(i/n,(j-1)/m)\leqslant 1/m$。这是一个并不唯一的映射过程，难以获得一个具体的离散Copula函数的描述。从另一个角度来看，若记$M_{n\times m}$为所有n×m阶矩阵$A=(a_{ij})_{n\times m}$的集合，其中要求$a_{ij}\geqslant 0$，$\sum_{i-1}^{n}a_{ij}=1/m$，$\sum_{j=1}^{m}a_{ij}=1/n$，则二元离散Copula被表示为如下矩阵：

$$C_{n,m}(i/n,j/m)=\begin{cases}0 & i\cdot j=0\\ \sum_{k\leqslant ir\leqslant j}\sum a_{kr} & i\cdot j\neq 0\end{cases}$$

可见离散Copula函数的直接求解内在地隐含了矩阵运算，在n元情况下矩阵阶数增高，演算过程将十分繁琐。

因此，越来越多的学者逐渐关注离散Copula的研究。在理论上，Kolesarova（2006）研究了定义在I_n^2上的离散Copula，表明每一个定义在I_n^2上的离散Copula都有一个双随机矩阵与之对应，并用矩阵的乘积定义了离散Copula的乘积。Aguilo、Suner、Torrens（2008，2010）专门介绍了离散Copula的矩阵表示问题。Mayor（2008）等主要从聚合算子的角度研究了离散Copula，对不可约离散Copula也有所研究，最后处理了离散Copula的延拓的某些方面。Alsina（1993）首次介绍了Quasi-Copula的概念，而Cuculescu等则给出n维Quasi Copula的详细阐述，Nelsen（1993）还基于格论指出所有二元Copula所构成的集合不是格，而所有二元Quasi Copula所构成的集合是一个完备格。在应用方面，Leon（2010）使用Copula建立了二元离散连续混合结构，并给出了基于Gaussian Copula的回归模型。Anastasios（2011）建立了面向离散数据的藤Copula结构PCCs和D-Vine。在Christian（2007）探讨了离散计数数据的Copula函数之后，Nikoloulopoulos（2009）和Avramidis（2009）研究了如何使用Normal Copula和Gaussian Copula对无限定离散分布的连接问题及相关性问题。Adrian（2011）则依靠低算法复杂度的蒙特卡罗经验最大值方法，建立了离散数据的Copula高斯图模式。考虑到大于二元的离散边际分布的Copula函数难以估计问题，Michael（2012）还采用马尔可夫链和蒙特卡罗方法，建立了16维D-Vine Copulas进行了实证分析。在现有的离散Copula方法中，人们大多开展的是二元离散Copula和不可约离散Copula的理论研究，仅有部分依靠Gaussian Copula、Normal Copula等建立了回归模型，而应用则主要以D-Vine分解方式或图模式来实现，该方式需要2n(n-1)甚至2^n次二元Copula函数的估计计算，限制了离散Copula函数的应用。

为了降低离散Copula函数求解的复杂度，本文建立了基于扰动的离散随机变量至连续随机变量的转换关系，依靠连续Copula函数间接地获得了离散边际分布的联合分布函数。

三、基于扰动量的转换关系

为了解决离散边际分布使用连续Copula函数进行连接的问题，本文借鉴Olivier（2012）从概率分布到随机矢量的映射及其反操作的思路，采用加入扰动量方式进行转化，流程如图1所示：①设$(X_1,X_2,\cdots,X_n)$是离散n维随机矢量，具有分布函数H和边际分布$F_1,F_2,\cdots,F_n$。分布函数H可以通过$F_1,F_2,\cdots,F_n$使用离散Copula函数C连接形成；②对$X_1,X_2,\cdots,X_n$分别加入相互独立的连续随机变量$T_1,T_2,\cdots,T_n$，从而构成连续随机变量$X^*_1,X^*_2,\cdots,X^*_n,T_1,T_2,\cdots,T_n$相互独立且与$X_1,X_2,\cdots,X_n$独立，是一定范围内的扰动量；③依靠②中的转换关系，得到$X^*_1,X^*_2,\cdots,X^*_n$所对应的连续边际分布函数$F^*_1,F^*_2,\cdots,F^*_n$，进而根据Sklar定理通过连续Copula函数C^*建立联合分布H^*；④将C^*看作分布函数求得其概率密度c^*并与$F^*_1,F^*_2,\cdots,F^*_n$结合获得h^*（或者直接通过H^*得到h^*），开展极大似然估计，得到参数值；⑤建立$C^*\rightarrow C$或$c^*\rightarrow c$的转换关系，得到离散的Copula函数C，进而通过$F_1,F_2,\cdots,F_n$获得联合分布函数H，这一过程亦可通过建立H^*与H之间关系来获得H。

为了便于分析，设X_i是离散随机变量，取值为以0为起始的顺序自然数（如0、1、2），其概率密度和分布函数分别为$f_i(x_i)$和$F_i(x_i)$。根据上述转化的思想以及Denuit和Lambert（2005）的方法，构造一个连续随机变量$X^*_i=X_i-$

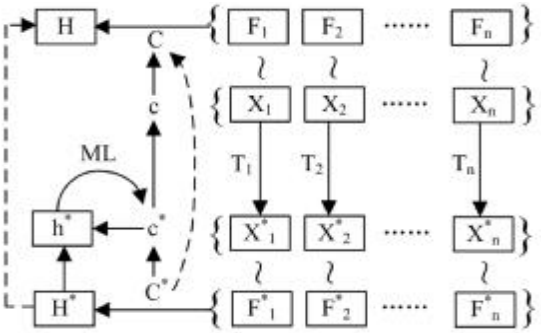

图 1 基于扰动量的转换关系

$\frac{U_i}{b_i-a_i}$，U_i是一个(a_i,b_i)范围内的连续分布，其独立于X_i，且U_i和$U_j(i\neq j)$互相独立，分别具有概率密度$f_{U_i}(u_i)$和分布函数$F_{U_i}(u_i)$，$F_{U_i}(u_i)$在(a_i,b_i)内和X_i的分布函数不具有相同的参数。此处，$T_i=U_i/(b_i-a_i)$是辅助的扰动量，用来将离散随机变量X_i构造为连续随机变量X^*_i，其也可以看作是使用一种连续随机过程来填补离散随机变量之间的“空白”。若记连续随机变量X^*_i的分布函数为$F^*_i(x_i)$，则：

$$F^*_i(x_i)=P(X^*_i\leqslant x_i)=P(X_i-T_i\leqslant x_i)=P(X_i-T_i\leqslant x_i,X_i\leqslant x_i)+P(X_i-T_i\leqslant x_i,X_i>x_i) \quad (4)$$

对于公式(4)$F^*_i(x_i)$表达式中的$P(X_i-T_i\leqslant x_i,X_i\leqslant x_i)$来说，由于$T_i\in(0,1)$，因此$X_i-T_i<X_i$，所以$P(X_i-T_i\leqslant x_i,X_i\leqslant x_i)=P(X_i\leqslant|x_i|)=F_i(|x_i|)$，其中$|x_i|$是$x_i$的整数部分。对于公式(4)中的$P(X_i-T_i\leqslant x_i,X_i>x_i)$来说，由于$X_i$为非负整数，故仅当取值为$|x_i+1|$才能同时满足$X_i-T_i\leqslant x_i$和$X_i>x_i$(如果为大于等于$|x_i+2|$的值，则由于$T_i\in(0,1)$，因而$X_i-T_i\geqslant|x_i+2|-T_i>|x_i+1|>x_i$，此时$P(X_i-T_i\leqslant x_i,X_i>x_i)=0$)，因此直接令$x_i=|x_i+1|$，故有：

$$P(X_i-T_i\leqslant x_i,X_i>x_i)=P(X_i-T_i\leqslant x_i,X_i=|x_i+1|)=P(U_i\geqslant(|x_i+1|-x_i)(b_i-a_i))P(X_i=|x_i+1|)=[1-F_{U_i}((|x_i+1|-x_i)(b_i-a_i))]f_i(|x_i+1|)$$

。综上，可得到连续随机变量X^*_i的分布函数为：

$$F^*_i(x_i)=F_i(|x_i|)+[1-F_{U_i}((|x_i+1|-x_i)(b_i-a_i))]f_i(|x_i+1|) \quad (5)$$

若记连续随机变量X^*_i的概率密度为$f^*_i(x_i)$，则$f^*_i(x_i)=P(X^*_i=x_i)=P(X_i-T_i=x_i)$，要使得$X_i-T_i=x_i$，由于$X_i$为非负整数且$T_i\in(0,1)$，必须满足$X_i=|x_i+1|$(否则在$X_i>|x_i+1|$或$X_i<|x_i+1|$时必然出现$X_i-T_i>x_i$或$X_i-T_i<x_i$)，故$P(X_i-T_i=x_i)=P(U_i=(|x_i+1|-x_i)(b_i-a_i))P(X_i=|x_i+1|)=f_{U_i}((|x_i+1|-x_i)(b_i-a_i))f_i(|x_i+1|)$。所以：

$$f^*_i(x_i)=f_{U_i}((|x_i+1|-x_i)(b_i-a_i))f_i(|x_i+1|) \quad (6)$$

根据公式(5)和(6)，显然在离散随机变量的情况下，则$F_i(x_i)=F^*_i(x_i)$，$f^*_i(x_i)=\int_{s-1}^{s}f^*_i(x_i)dx_i$。然而，此处的$F^*_i(x_i)$和$f^*_i(x_i)$并不满足分布函数和概率密度中$F^*_i(+\infty)=1$和$\int_{-\infty}^{+\infty}f^*_i(x_i)dx_i=1$的要求，故可补充定义$f_i(\max(x_i)+1)=f_i(0)$，并令$U_i$为$(a_i,a_i+1)$范围内的均匀分布，则公式(5)和(6)可修正并简化为：

$$F^*_i(x_i)=F_i(|x_i|)+[1-F_{U_i}(|x_i+1|-x_i)]f_i(|x_i+1|)=F_i(|x_i|)+(x_i-|x_i|)f_i(|x_i+1|)$$

$$f^*_i(x_i)=f_{U_i}(|x_i+1|-x_i)f_i(|x_i+1|)=f_i(|x_i+1|) \quad (7)$$

据此，$f_i(x_i)$和$f^*_i(x_i)$的关系可如图2所示，其中空心圈为离散随机变量的概率值，而直线则为连续随机变量的概率密度。在该图中，P_9为补充定义，$P_0\sim P_8$之和与图中阴影面积相等，均为1，并且每个线段下的面积则等于前一点的离散值即$A_i=P_{i+1}$，i=0，1，…，8。

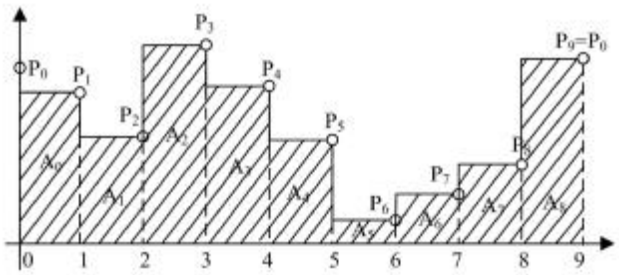

图 2 概率密度函数转换关系示意图

通过公式(7)的推导，实际上给出了离散随机变量分布列及其分布函数对连续随机变量概率密度和分布函数的映射，但要实现对离散边际分布的连接，还需引入连续 Copula 函数。设$H^*(x_1,x_2,\cdots,x_n)$为n个连续边际分布$F^*_i(x_i)$的联合分布函数，且连续随机变量X^*_i的分布函数$F^*_i(x_i)$由离散随机变量X_i的概率分布函数$F_i(x_i)$根据公式(7)转化而来，由 Sklar 定理可知，必然存在一个连续的n元 Copula 函数$C^*(t_1,t_2,\cdots,t_n)$满足$H^*(x_1,x_2,\cdots,x_n)=C^*(F^*_1(x_1),F^*_2(x_2),\cdots,F^*_n(x_n))$，又根据联合分布的定义$H^*(x_1,x_2,\cdots,x_n)=P(X^*_1\leqslant x_1,X^*_2\leqslant x_2,\cdots,X^*_n\leqslant x_n)$，有：

$$H^*(x_1,x_2,\cdots,x_n)=Q(1,2,\cdots,n)+\sum_{1\leqslant k_1\leqslant n}Q(k_1,\cdots,n)G(k_j)+\sum_{1<k_1<k_2\leqslant n}Q(k_1,k_2,\cdots,n)\prod_{j=1}^{2}G(k_j)+\sum_{1\leqslant k_1<k_2<k_3\leqslant n}Q(k_1,k_2,k_3,\cdots,n)\prod_{j=1}^{3}G(k_j)+\cdots+Q(k_1,k_2,\cdots,k_n)\prod_{j=1}^{n}G(k_j) \quad (8)$$

其中，$G(i)=1-F_{U_i}(|x_i+1|-x_i)$，$Q(k1,k2,kj\cdots,n)=P(X\leqslant|x_{k1}|+1,X\leqslant|x_{k1}|+1,X\leqslant|x_{kj}|+1,\cdots\cdots,X_n\leqslant x_n)$。$Q(k_1,k_2,k_j\cdots,n)$是一个概率值，且$k_j$位置的约束条件为$X\leqslant|x_{kj}|+1$，其他位置约束条件为$X_i\leqslant x_i$。如果设离散随机变量$X_i$通过离散 Copula 函数$C(t_1,t_2,\cdots,t_n)$进行连接后的联合分布为$H(x_1,x_2,\cdots,x_n)=P(X_1\leqslant x_1,X_2\leqslant x_2,\cdots,X_n\leqslant x_n)=C(F_1(x_1),F_2(x_2),\cdots,F_n(x_n))$，则可以得到：$Q(k_1,k_2,kj\cdots,n)=C(F_{k1}(|x_{k1}|+1),F_{k2}(|x_{k2}|+1),F_{kj}(|x_{kj}|+1),\cdots,F_n(x_n))$。又因为，当$x_i$取整数值时则$x_i=|x_i|$、$G(i)=0$，由公式(8)可以得到：

$H^*(|x|;\eta)=Q(1,2,\cdots,n)=C(F_1(x_1),F_2(x_2),\cdots,F_n(x_n))=H(x;\eta)$

$C(F_1(x_1),F_2(x_2),\cdots,F_n(x_n))=C^*(F^*_11(x^*_1),F^*_2(x^*_2),\cdots,F^*_n(x^*_n))$ (9)

由上面的分析可知,想要获得离散随机变量X_i所对应边际分布$F_i(x_i)$的离散Copula连接函数或联合分布,可首先将X_i转化为连续的随机变量X^*_i,并借此获得连续边际分布函数的Copula连接函数$C^*(t_1,t_2,\cdots,t_n)$,然后取这个连续联合分布$H^*(x_1,x_2,\cdots,x_n)$所在整数点位的离散值即可获得X_i所对应的离散联合分布函数$H(x_1,x_2,\cdots,x_n)$,最后进行参数估计即可。下面以泊松分布作为离散边际分布,以Archimedean Copula函数族中的Clayton Copula为连接函数,结合实证数据,给出回归模型及分析结果。

四、实证分析

(一)回归模型

给定n个离散随机变量X_i服从参数λ_i的泊松分布$F(x_i;\lambda_i)$,根据公式(7)转化后所得的X^*_i的连续分布为:

$$F^*_i(x_i)=\sum_{k=0}^{|x_i|}\frac{e^{-\lambda_i}\lambda_i^k}{k!}+w_{x_i}\frac{e^{-\lambda_i}\lambda_i^{m_{x_i}}}{m_{x_i}!} \quad (10)$$

其中,$f^*_i(x_i)=\frac{e^{-\lambda_i}\lambda_i^{m_{x_i}}}{m_{x_i}!}$,$w_\xi=\xi-|\xi|$,$m_\xi=|\xi|+1=|\xi+1|$。

又由于n元的连续Copula函数Clayton Copula,其表达式(分布函数)描述为:

$$C^*(\delta_i;\theta)=[1+\sum_{i=1}^n(\delta_i^{-\theta}-1)]^{-1/\theta},\theta\geq-1,\theta\neq0 \quad (11)$$

根据公式(2)对公式(11)求偏导,可得Clayton Copula的密度函数为:

$$c^*(\delta_i;\theta)=\theta^n\prod_{i=0}^{n-1}(\frac{1}{\theta}+i)\cdot\prod_{i=0}^{n}\theta^{-\theta-1}\cdot[1+\sum_{i=1}^n(\delta_i^{-\theta}-1)]^{-n-1/\theta} \quad (12)$$

将公式(10)和公式(11)代入公式(1),得到连续随机变量X^*_i通过Clayton Copula连接后的联合分布函数:

$$H^*(x;\theta,\lambda_i)=\{1+\sum_{i=1}^n[(\sum_{k=0}^{|x_i|}\frac{e^{-\lambda_i}\lambda_i^k}{k!}+w_{x_i}\frac{e^{-\lambda_i}\lambda_i^{m_{x_i}}}{m_{x_i}!})^{-\theta}-1]\}^{-1/\theta} \quad (13)$$

根据公式(9)要获得n个泊松分布$F(x_i;\lambda_i)$的联合分布则只需要对$H^*(x;\theta,\lambda_i)$取整数位的值即可。在参数估计时,采用最大似然方法,先通过公式(12)求得连续的联合概率密度:

$$h^*(x;\theta,\lambda_i)=\theta^n\prod_{i=0}^{n-1}(\frac{1}{\theta}+i)\cdot\prod_{i=0}^{n}F^*_i(x_i)^{-\theta-1}\cdot[1+\sum_{i=1}^n(F^*_i(x_i)^{-\theta}-1)]^{-n-1/\theta}\cdot\prod_{i=1}^n f^*_i(x_i) \quad (14)$$

然后根据式(14)构造对数似然函数$\ln h^*(x;\theta,\lambda_i)$。给定m个样本,对参数$\lambda_i$、$\theta$求偏导得方程组:

$$\frac{\partial\ln h^*(x;\theta,\lambda_i)}{\partial\lambda_i}=\sum\{\frac{(-\theta-1)}{F^*_i(x_i)}\frac{\partial F^*_i(x_i)}{\partial\lambda_i}+\theta(n+1/\theta)\frac{F^*_i(x_i)^{-\theta-1}\frac{\partial F^*_i(x_i)}{\partial\lambda_i}}{1+\sum_{i=1}^n(F^*_i(x_i)^{-\theta}-1)}+\frac{1}{f^*_i(x_i)}\frac{\partial f^*_i(x_i)}{\partial\lambda_i}\}$$

$$\frac{\partial\ln h^*(x;\theta,\lambda_i)}{\partial\theta}=\sum\{\frac{n}{\theta}-\sum_{i=0}^{n-1}\frac{1/\theta^2}{i+1/\theta}-\sum_{i=0}^n\ln F^*_i(x_i)+\frac{1}{\theta^2}\ln[1+\sum_{i=1}^n(F^*_i(x_i)^{-\theta}-1)]+(n+\frac{1}{\theta})\frac{F^*_i(x_i)^{-\theta}\ln F^*_i(x_i)}{1+\sum_{i=1}^n(F^*_i(x_i)^{-\theta}-1)}\} \quad (15)$$

其中,$\frac{\partial F^*_i(x_i)}{\partial\lambda_i}=\sum_{k=0}^{|x_i|}\frac{ke^{-\lambda_i}\lambda_i^{k-1}-e^{-\lambda_i}\lambda_i^k}{k!}+w_{x_i}\frac{m_{x_i}e^{-\lambda_i}\lambda_i^{m_{x_i}-1}-e^{-\lambda_i}\lambda_i^{m_{x_i}}}{m_{x_i}!}$,

$\frac{\partial f^*_i(x_i)}{\partial\lambda_i}=\frac{m_{x_i}e^{-\lambda_i}\lambda_i^{m_{x_i}-1}-e^{-\lambda_i}\lambda_i^{m_{x_i}}}{m_{x_i}!}$。

令公式(15)方程组的各方程等于零,即可得到参数的估计值,但难以得到显式解,需要通过数值迭代求解。在迭代过程中,将泊松分布的参数估计值作为本文方法的参数初始值,参数θ初值设置为1。数据的求解使用SAS程序中的非线性混合效应PROC NLMIXED模块,通过(adaptive) Gauss-Hermite quardrature方法来实现,利用Dual Quasi-Newton进行优化。采用赤池弘次信息量准则Akaike Information Criteria(AIC)统计量和贝叶斯信息准则Bayesian Information Criteria(BIC)统计量来衡量统计模型的拟合优良性。AIC统计量的定义为:AIC=-2lnL+2P,其中,lnL为似然函数,P为参数的个数。AIC在理论结构上采用最小限度的假定,其值越小,表明模型的拟合越好。BIC统计量的定义为:BIC=-2lnL+Plog(m),其中m为观测值的个数。BIC值越小,表明模型的拟合越好。

另外,如果结合零膨胀方法进行分析,则只需要根据将零膨胀模型结构引入到公式(14)中,增加零膨胀结构零的比例参数φ,并对修改后的对数似然函数$\ln h^*(x;\theta,\lambda_i,\varphi)$的参数求偏导获得方程组,参看公式(16),进行数据求解即可。然而对于D-Vine方法来说,由于采用了藤结构分解方法,因而无法便捷地接入零膨胀模型。

$$\frac{\partial \ln h^*(x;\theta,\lambda_i,\varphi)}{\partial \lambda_i}=\sum_{h_j^*=0}\frac{\partial \Theta}{\partial \lambda_i}+\sum_{h_j^*>0}\{\frac{(-\theta-1)}{F_i^*(x_i)}\frac{\partial F_i^*(x_i)}{\partial \lambda_i}+$$

$$\theta(n+1/\theta)\frac{F_i^*(x_i)^{-\theta-1}\frac{\partial F_i^*(x_i)}{\partial \lambda_i}}{1+\sum_{i=1}^{n}(F_i^*(x_i)^{-\theta}-1)}+\frac{1}{f_i^*(x_i)}\frac{\partial f_i^*(x_i)}{\partial \lambda_i}\}$$

$$\frac{\partial \ln h^*(x;\theta,\lambda_i,\varphi)}{\partial \theta}=\sum_{h_j^*=0}\frac{\partial \Theta}{\partial \theta}+\sum_{h_j^*>0}\{\frac{n}{\theta}-\sum_{i=0}^{n-1}\frac{1/\theta^2}{i+1/\theta}-$$

$$\sum_{i=0}^{n}\ln F^*{}_i(x_i)+\frac{1}{\theta^2}\ln[1+\sum_{i=1}^{n}(F^*{}_i(x_i)^{-\theta}-1)]+(n+\frac{1}{\theta})$$

$$\frac{F_i^*(x_i)^{-\theta}\ln F_i^*(x_i)}{1+\sum_{i=1}^{n}(F_i^*(x_i)^{-\theta}-1)}\}$$

$$\frac{\partial \ln h^*(x;\theta,\lambda_i,\varphi)}{\partial \varphi}=\sum_{h_j^*=0}(1-\Psi)/[\varphi+(1-\varphi)\Psi]$$

$$+\sum_{h_j^*>0}\frac{-1}{1-\varphi}$$

其中，$\Theta=\ln[\varphi+(1-\varphi)\Psi]$，

$$\Psi=\theta^n\prod_{i=0}^{n-1}(\frac{1}{\theta}+i)\cdot\prod_{i=0}^{n}e^{(\theta+1)\lambda_i}\cdot[1+\sum_{i=1}^{n}(e^{\theta\lambda_i}-1)]^{-n-1/\theta}\cdot\prod_{i=1}^{n}e^{-\lambda_i},$$

$$\frac{\partial \Theta}{\partial \lambda_i}=(1-\varphi)\Psi\{\sum_{i=1}^{n}\theta-(n+\frac{1}{\theta})\frac{\sum_{i=1}^{n}e^{\theta\lambda_i}\theta}{[1+\sum_{i=1}^{n}(e^{\theta\lambda_i}-1)]}$$

$$+\sum_{i=1}^{n}(-1)\}/[\varphi+(1-\varphi)\Psi],$$

$$\frac{\partial \Theta}{\partial \theta}=(1-\varphi)\Psi\{\frac{\theta}{n}+\sum_{i=0}^{n-1}\frac{-1}{\theta+i\theta^2}+\sum_{i=1}^{n}\lambda_i+$$

$$\frac{\ln[1+\sum_{i=1}^{n}(e^{\theta\lambda_i}-1)]}{\theta^2}-(n+\frac{1}{\theta})\frac{\sum_{i=1}^{n}e^{\theta\lambda_i}\lambda_i}{[1+\sum_{i=1}^{n}(e^{\theta\lambda_i}-1)]}\}/$$

$/[\varphi+(1-\varphi)\Psi]$。

（二）数据分析

本文数据来自新加坡保险协会连续9年的一般性汽车保险索赔数据，内容包括保单信息、驾驶记录、潜在风险、投保人特征，以及索赔日期、索赔频度、补偿数量等。我们对原始数据进行整理，剔除不完整数据后，并取投保人信息全面的数据资料，获得观测值3 105个索赔记录。如表1所示，零次索赔高达81%以上，具有零膨胀特征，这与汽车保险受到多元化影响的实际情况相符；投保人的数量在逐年变化，体现了数据的不平衡性，客户可能会在不同的保险之间转换，或者出现新投保人加入以及旧客户的离去；各年度内索赔频次越高所发生的次数越低（如第1年度，零次索赔发生次数为369次，1次索赔发生次数为48次），5次及其以上索赔的数量均为0，可见数据具有较好的同质性。

表1　　各年度索赔次数及百分比

Count	Claim Counts by Year									Overall	
	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Number	Percent
0	369	482	311	156	58	182	420	353	321	2 652	85.41
1	48	60	55	32	6	37	63	59	46	406	13.08
2	6	10	4	3	0	2	7	4	3	39	1.26
3	2	0	1	0	1	0	1	1	0	6	0.19
4	0	0	1	0	0	0	0	1	0	2	0.06
5	0	0	0	0	0	0	0	0	0	0	0
Number	425	552	372	191	65	221	491	418	370	3 105	100

在回归模型中，选取了8个费率因子，包括客户性别（“0”为女性；“1”为男性）、客户年龄（“1”为小于等于25岁；“0”为其他）、婚姻状况（“1”为已婚；“0”为未婚）、汽车品牌（“0”、“1”、“2”分别代表三种不同品牌）、汽车颜色（0~5分别代表6种不同颜色）、NCD等级（“1”为0%级；“0”为其他）、行驶时长、汽车价格等，除行驶时长、汽车价格为连续随机变量外其他均为属性变量。其中NCD等级以百分比表示，最大值为50%，投保者一年内无索赔则增加10%，如果NCD为40%或50%，则一年内1次索赔将降低20%或30%，两次以上的索赔则完全丢失当前的NCD等级，如果NCD在30%及其以下，若一年内发生索赔则丧失NCD等级。经过风险分类后，选取Poisson、ZIP、D-Vine、CP、ZICP作为回归模型，其中CP为采用Clayton Copula对5个泊松边际分布进行连接后的联合分布，ZICP为与CP对应的零膨胀模型，回归模型拟合结果如表2所示。

表2　　不同回归模型的拟合结果

参数估计	Poisson	ZIP	D-Vine	CP	ZICP
截距	-1.521	-0.536 2	-0.698 1	-0.671 2	-0.530 4
	(0.102 3)*	(0.081 2)*	(0.075 2)*	(0.070 5)*	(0.063 4)*
客户年龄	0.680 2	0.610 9	0.620 9	0.622 4	0.618 1
	(0.001 1)	(0.000 8)	(0.001 5)*	(0.001 3)*	(0.001 9)*
婚姻状况	-0.123 5	-0.100 1	-0.137 6	-0.135 5	-0.106 5
	(0.002 3)*	(0.007 1)*	(0.009 2)	(0.000 8)*	(0.006 2)*
NCD等级	0.697 0	0.621 6	0.612 3	0.618 0	0.608 8
	(0.088 2)*	(0.080 1)*	(0.098 1)*	(0.080 4)*	(0.058 9)*
行驶时长	-0.056 5	-0.033 2	-0.048 1	-0.047 7	-0.039 9
	(0.011 1)*	(0.009 9)*	(0.003 9)*	(0.002 5)*	(0.001 0)*
客户性别	0.198 1	0.203 8	0.331 2	0.303 0	0.260 0
	(0.051 3)	(0.082 2)	(0.012 3)	(0.031 0)	(0.015 1)
-2L	5 824.9	5 212.9	5 422.7	5 398.4	5 023.1
AIC	5 836.9	5 226.9	5 438.7	5 422.4	5 049.1
BIC	5 873.1	5 269.2	5 487.0	5 494.9	5 127.6

注：*表示在水平为5%下是显著的，()表示参数估计值的标准误差。

结果表明:①所有的回归模型结果均表明客户性别、客户年龄、婚姻状况、行驶时长、NCD等级是与索赔次数相关的重要风险因素,且婚姻状况、行驶时长、NCD等级均在水平为5%下显著。②在采用ZIP进行回归分析后,其结构零的比率参数φ=0.481 0,且AIC和BIC明显比泊松回归模型降低,反映了观测数据的零膨胀特点。③在前5个回归模型中,Poisson拟合效果最差;ZIP与D-Vine、CP的结果差异不大,这是由于ZIP引入了零膨胀模型获得较好的拟合优度,同时D-Vine和CP也因为考虑了多种风险因素,从而比单个的泊松模型要更为有效,并且CP和D-Vine拟合优度十分接近。值得注意的是D-Vine由于分解方式复杂,因而难以再结合零膨胀模型对拟合效果进行提升。④ZICP模型对多个边界分布进行连接并加入了零膨胀模型,导致了参数数量的增加,虽然影响了拟合优良性,但其更加接近了实际索赔中的零次索赔数量过多以及多种费率因子共同作用的客观情况,此时AIC和BIC分别为5 049.1和5 127.6,是所有模型中的最小值,具有最佳的回归效果,更加有效地描述了潜在的索赔次数分布,同时表明观测数据中有1 586(3 105×0.510 9)个结构零。⑤6个模型的回归结果还共同说明:男性驾驶者发生事故频次相对女性驾驶者较小;更为年轻的驾驶者出现事故的可能性越大;已婚被保险者驾驶车辆更为谨慎,索赔率较低;行驶时长越高者出现事故的可能性相对较小,但并不明显;NCD等级越低者由于驾驶经验缺乏,因而索赔频次较高。这些结论都符合人们对汽车保险索赔现象的认知,在一定程度上说明了模型的实用性。

五、结论

本文采用扰动量方式将离散随机变量转化为连续随机变量,探讨了离散边际分布对连续Copula函数的利用问题,建立了n元离散Copula函数和连续Copula函数之间的关系,采用Clayton Copula对5个泊松边际分布进行了连接,同时加入零膨胀思想,结合最大似然估计,建立了ZICP回归模型,开展了实证分析,并与Poisson、ZIP、D-Vine进行了对比,结果表明,ZICP更加接近了实际索赔中的零次索赔数量过多以及多种费率因子共同作用的客观情况,提高了拟合分析效果,具有一定的实用性。本文下一步的研究方向是:一方面重点对不同扰动量带来的离散Copula的不同进行探讨,分析连续Copula函数连接边际分布时的唯一性与转化后的离散Copula函数的非唯一性之间的关系,另一方面将本文思想向椭圆Copula函数族中的Gaussian Copula、t-Copula以及极值Copula等方面进行应用,提高方法的适用范围。

主要参考文献

韦艳华,张世英.多元Copula-GARCH模型及其在金融风险分析上的应用[J].数理统计与管理,2007(3).

Jondeau E., Rockinger M.. The copula- GARCH model of conditional dependencies: an international stock market application. [J]Journal of International Money and Finance,2006(5).

Hofert M., Scherer M.. CDO pricing with nested Archimedean copulas[J]. Quantitative Finance,2011(5).

Eling M., Toplek D.. Modeling and management of nonlinear dependencies copulas in dynamic financial analysis. [J]Journal of Risk and Insurance,2009(3).

Bedford T., Cooke R. M.. Probability density decomposition for conditionally dependent random variables modeled by vines[J].Annals of Mathematics and Artificial Intelligence,2001(1).

Bedford T., Cooke R. M.. Vines- a new graphical model for dependent random variables[J].Annals of Statics, 2002(30).

Aas K., Czado C., Frigessi A.. Pair-Copula constructions of multiple dependence [J].Insurance: Mathematics and Economics,2009(2).

张尧庭.连接函数(Copula)技术与金融风险分析[J].统计研究,2002(4).

史道济,关静.沪深股市风险的相关性分析[J].统计研究,2003(10).

周孝华,李强,张保帅.基于Copula-ASV-GPD的我国多元外汇储币组合的风险度量[J].系统工程,2012(11).

谢赤,余聪,罗长青等.基于MRS Copula-GJR-Skewed-t模型的股指期货套期保值研究[J].系统工程学报,2013(1).

陆静,张佳.基于极值理论和多元Copula函数的商业银行操作风险计量研究[J].中国管理科学,2013(3).

Denuit M., Lambert P.. Constraints on concordance measures in bivariate discrete data[J]. Journal of Multivariate Analysis,2005(1).

吴娟.Copula理论与相关性分析[D].武汉:华中科技大学,2009.

Aguilo I., Suner J., Torrens J.. Matrix representation of discrete Quasi-Copulas[J].Fuzzy Sets and Systems,2008 (13).

【基金项目】西安市软科学基金"西安市环境污染责任保险定价策略研究"(项目编号:HJ1111)