学校编码：10384　　分类号_____ 密级_____

学号：19020130154168　　UDC_____

博 士 学 位 论 文

Erlang混合模型在保险精算中的应用

Applications of Erlang Mixtures in Insurance

桂文永

指导教师姓名：林 小 东 教授

专 业 名 称：概率论与数理统计

论文提交日期：2017 年 4 月

论文答辩日期：2017 年 5 月

学位授予日期：2017 年 月

答辩委员会主席：_____

评 阅 人：_____

2017 年 4 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）　1、经厦门大学保密委员会审查核定的保密学位论文，于　　年　月　日解密，解密后适用上述授权。

（　　）　2、不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

中文摘要

等尺度参数的Erlang混合分布具有很多良好的性质，广泛应用于保险精算数据中。对于混合分布的参数估计，最常用的方法是期望最大化(Expectation Maximization，简写EM)算法。Erlang混合分布的形状参数限制为正整数，这使得形状参数的估计比较困难。在本文中我们对标准的EM算法进行修改，提出广义期望最大化(Generalized Expectation-Maximization, GEM) 算法。在EM算法的M步中对形状参数的估计不是使相应的Q函数达到最大，而是调整形状参数保证Q函数的函数值增加即可。最后为了避免出现过拟合的问题，本文用BIC(Bayesian Information Criterion)准则确定混合模型的序。参数个数和模型的序成正比关系，所以混合模型参数个数比较多，尤其是模型的序比较大的时候。本文采用前向选择的方法从两个混合模型开始选择得到最优的混合模型的序，相比后向选择方法大大的缩短了参数估计的时间。

EM算法是一种迭代算法，算法的收敛性高度依赖参数的初始值。本文从一个新的角度出发，并提出新的方法估计参数初始值。通过对观测数据进行聚类，然后用矩估计方法对参数进行初始化，我们将这种方法称为CMM(Clusterized Method of Moment)算法。通过对模拟数据以及实际数据拟合，我们可以看到这种CMM初始化方法可以得到高质量的参数初始值。本文采用K-means算法处理此分类问题，并且对该分类的方法选取给出了一个合理解释。

在实际应用中，数据经常会出现截断或删失的情形，比如保险数据中有免赔额的保单或设定赔付上限的保单。这些情况对保险公司运营非常重要，我们不能忽略这种情况。因此，本文进一步将CMM-GEM算法推广到截断和删失数据的情形，使其应用更加广泛，能够处理各种情形的数据。

在弱收敛意义下，多维Erlang混合在多维正的连续分布空间中是稠密的。对多维Erlang混合分布，很多重要结果都有具体的解析表达式，比如各阶矩、Laplace变换和边际分布密度函数等等。Copula方法是一种常用的处理多维数据的方法，相比copula方法，多维Erlang混合分布能够更灵活的获取变量的相关结构。而且，对一些常用的相关关系度量如Kendal's tau和Spearman's rho都能够得到精确的表达式。

对Erlang混合分布的参数估计常常使拟合曲线不够光滑，为了使拟合的密度曲线更加光滑，我们考虑粗糙惩罚函数并修改前面提出的CMM-GEM算法。关于光滑性度量，对多维数据我们考虑所有分量之和的密度函数曲线的二阶导函数的积分，然后提出合适的惩罚函数。通过研究各参数对曲线光滑性的影响，发现多维Erlang混合模型的

混合权重和形状参数对光滑性影响较小，所以只需要对尺度参数进行惩罚。最后通过拟合模拟的数据和实际数据，对该算法的效果进行验证。

最后，我们用多维Erlang混合描述$(n-k+1)$-out-of-n系统中各部件寿命的联合分布。在这个领域的研究中中，常常假设各部件之间是相互独立的。本文我们假设各部件的联合寿命服从多维Erlang混合分布，然后研究其剩余寿命以及随机序性质。因为多维Erlang混合能抓住多个变量之间的相关结构，所以它是一种非常有用的模型。本文的结果可应用于部件之间具有强相关性且部件数量较多的系统中。

关键词：Erlang混合模型；初始值；CMM-GEM算法；相关性；波动性

Abstract

The class of Erlang mixtures with a common scale parameter has many desirable properties and it is widely used in insurance. A common method to estimate the parameters of mixture models is the Expectation-Maximization (EM) algorithm. The constrain that the shape parameters of an Erlang mixture must be positive integers makes the problem of estimation for shape parameters very difficult. In this thesis, we modify the standard EM algorithm and propose a generalized Expectation-Maximization (GEM) algorithm. In the Maximization step of an EM algorithm, we adjust the shape parameters by increasing the corresponding Q-function value rather than maximizing the function. In order to avoid the issue of overfitting, the BIC (Bayesian Information Criterion) is used to choose the order of an Erlang mxiture. Since the number of parameters of the model is proportional to the order of the model, the number of parameters is large, especially when the order of the model is high. We adopt a forward method to determine the order of the model with starting from a 2 order Erlang mixture and hence the running time will be less compared with backward methods.

As an iterative algorithm, the convergency of an EM algorithm highly depends on the initial values. In this thesis, we propose a new method to initialize the parameters from another perspective. We cluster the data first and then initialize the parameters using the method of moments, which we call the Clusterized Method of Moments (CMM). The method is demonstrated through stimulation studies and applied to real data, and show that we can obtain high quality initial parameters. We adopt the K-means method to deal with the clustering issue and provide an explanation about why we choose the K-means algorithm.

The data often appear to be truncated and/or censored in insurance practice. For example, this phenomenon appears when the policies have deductable and/or policy limit. These cases are important to the insurance company. We extend the CMM-GEM algorithm to truncated and/or censored case so that the mixture model can be used more widely.

Modelling dependency among insurance losses is of practical importance. One way is to use a multivariate Erlang mixture instead of a copula. The class of multivariate Erlang

mixtures is dense in the space of positive continuous multivariate distributions in the sense of weak convergence. There are explicit expressions of many distributional quantities such as the moments, the Laplace transform and marginal distributions. Compared with copulas, a popular tool to deal with multivariate data, the multivariate Erlang mixtures can capture the dependence structure more flexibly. Moreover, some dependence measures such as the Kendal's tau and Spearman's rho also have analytic expressions.

We find that the fitted curve for observed data is often not smooth enough. To smooth the fitted curve, we add a roughness penalty and modify the CMM-GEM algorithm mentioned above. In this case, the integrated squared second derivative of the density function of aggregate data is used to quantify the smoothness with a properly defined penalty function. Through the studies about the effects of the parameters on the smoothness, we find that both the mixing weights and the shape parameters have little effects on the smoothness and hence we only add a roughness penalty function on the scale parameter. We test the performance of the method through simulation studies and real data.

Finally, we use a multivariate Erlang mixture to describe the joint distribution of the lifetimes of the components of an $(n-k+1)$-out-of-n system. The research in this area is always under the assumption that the components of a system are independent. In this thesis, we study the conditional mean residual lifetime function and stochastic ordering properties of an $(n-k+1)$-out-of-n system with assumption that the lifetimes have a multivariate Erlang mixture. The multivariate Erlang mixture is a useful model as it can capture the dependence structure of a large number of multiple variables well. The results in this thesis may be useful when the components of a system are of strong dependency and the number of components is high.

Key words: Erlang mixtures; initial values; CMM-GEM algorithm; dependence structure; volatility.

目 录

Contents

第 一 章　绪论

1.1　问题研究背景及意义

混合模型广泛应用于数据分类、图形分割以及拟合具有分类特性的数据等问题。一般的保险公司保单损失数据类型多样、来源广泛，所以用混合模型建模是很合理的选择。统计学中，常用的混合模型是混合Gaussian 模型，如[4,17,31,43]等。但是，保险损失数据通常是非负的，所以混合Gaussian模型在我们的问题当中显然是不合适的。保险应用中一个较早期的模型见[18]，文中考虑的是混合指数模型或称为超指数模型，该模型众数在支集边界上且变异系数大于或等于1。指数分布的组合可以看成超指数分布的推广，在弱收敛意义下可以近似任意的正分布。但是在统计估计方面有很大的挑战，因为其权重的估计非常敏感和不稳定，相关文献见[9,11]。

近年来，Erlang混合模型受到很大的关注，广泛应用于保险损失数据。Erlang混合分布在弱收敛意义下可以逼近任意的正连续分布，相应的多维Erlang混合可以逼近任意的正连续多维分布。利用一维Erlang混合分布建模时，很多重要的统计量都有精确的表达式，如生存函数、各阶矩、Laplace变换等，因此相关的风险度量也能计算出来，例如VaR (Value-at-Risk)值和TVaR (Tail Value-at-Risk)。类似的性质可以推广到多维Erlang 混合分布，不仅如此，对于多维Erlang混合模型，其联合分布和边际分布的关系处理也非常方便。根据多维Erlang混合的联合密度，其边际分布密度函数、分量之和的密度函数以及分量之间的相关度量都能够算出解析表达式。在处理多维数据时，copula方法是最常用的方法。和copula方法相比，多维Erlang混合分布能够更灵活的反应各个变量之间的相关性。因此，Erlang混合模型是处理保险数据很重要的工具。多维Erlang 混合模型的应用见 [5,6,14,37,41]以及相关文献。

在保险应用中，因为保险公司和投保人对保单进行一些调整，使保险损失数据出现截断或删失情况。例如，保险人和被保险人事先约定损失数额在一定范围之内，被保险自己承担损失，保险人不进行赔付，此时损失数据就为左截断；若保险人和被保险人约定一个最高限额，超过限额的部分仍由被保险人自己承担，此时损失数据就为右删失。这些情况对保险公司运营会产生很大的影响，我们不能将其忽略。所以当我们用模型对数据进行拟合时，也要把截断和删失的情形考虑进去。

1.2 研究现状和动机

Lee和Lin[24]提出等尺度参数的一维Erlang混合模型并给出EM算法对参数进行估计。该模型对保险数据拟合效果非常好，甚至对厚尾分布数据也能达到令人满意的拟合效果。 [25]将该模型推广到多维的情形，表明多维Erlang混合模型在保险精算和风险管理非常适用。文中还研究了多维Erlang混合模型变量之间的相关性度量，证明多维Erlang混合模型可以灵活地反应各个变量之间的相关性。[38]和[37]考虑数据的截断和删失，将上述的EM算法推广到截断和删失情形。为了提高EM 算法效率，Yin和Lin[44]引入一个新的惩罚项并分析其估计的一致性。

混合Erlang分布参数比较多，并且直接用最大似然估计(MLE)方法比较困难。因此，上述文献中提出EM算法对参数进行估计。虽然EM算法估计参数简单有效，也能对数据拟合的很好，但是仍然存在一些共同的不足：

(1) EM算法是迭代算法，对初始值的依赖性很强，初始值的选取对于算法最终的收敛效果影响很大。一方面好的初始值可以使得算法收敛很快，减少算法运行的时间，尤其是对多维数据。另一方面，EM算法容易收敛到局部最优解，这也跟初始值的选取有很大关系。对于初始值的选取，上述文献初始值确定方法都是基于Tijms近似(见[36])给出的，但是效果并不是很理想。尤其是对于形状参数的初始值确定，一直在做改进。在Erlang 混合分布的序M给定时，[24]对形状参数初始值设为$1, 2, \cdots, M$，[38]引进扩展因子s 使初始的形状参数范围更大一些，而 [37]是根据各分量的分位数确定形状参数的初始值，使得初始值的效果有一定的提高。

(2) 在上述文献的EM算法主要是对混合权重和尺度参数进行估计，也就是说在参数估计时假定形状参数是已知的。然后再人为调整形状参数，重新估计混合权重和尺度参数，最后根据似然函数选出最优的参数。这样处理的缺点是，即使给定混合模型的序，仍然需要多次使用EM算法，消耗大量的时间。对于尾部较厚的数据，这种算法也使得尾部拟合的效果不好。

(3) 对于混合模型序的选取，我们使用的信息准则是Bayesian Information Criterion (BIC)或者Akaike Information Criterion (AIC)信息准则。上述文献均采用后向选择法，即初始时选择一个比较大的序M，然后再逐步删除权重较小的组分最终得到合适的序。这种选取办法会使得尾部的信息丢失，造成尾部拟合效果较差。另外因为初始的序M较大，要估计的参数个数比较多，算法时间比较长。

对于上述的不足，本文提出新的初始值估计方法并且对经典EM算法做出修改，提出CMM-GEM算法。EM算法将观测数据视为不完全数据，需要引入潜变量。在估计初始值时，我们采用最常用的矩估计方法。因为潜在变量不可观测，在估计初值之前需要将观测数据分类得到潜变量的值。本文选择K-means估计对观测数据进行分类，并且