Vol. 36 ( 2016 )
No. 6

数 学 杂 志
J. of Math. (PRC)

# 厄兰极值混合模型的有效估计及其在保险中的应用

殷崔红 [1], 林小东 [1,2], 袁海丽 [3]
(1. 厦门大学数学科学学院, 福建 厦门 361005)
(2. 多伦多大学统计系, 加拿大 多伦多 M5S 3G3)
(3. 武汉大学数学与统计学院, 湖北 武汉 430072)

**摘 要**: 本文研究了 Erlang 混合分布和广义帕累托分布混合模型的估计问题. 通过引入 iSCAD 惩罚函数, 利用 EM 算法极大化 iSCAD 惩罚似然函数的方法, 获得了混合序和参数的估计值, 计算出有效的度量风险指标 value-at-risk(VaR) 和 tail-VaR(TVaR), 通过模拟实验和实际数据说明了模型和算法的有效性. 推广了有限 Erlang 极值混合模型在保险数据拟合中的应用.

**关键词**: 极值理论; 极值混合模型; iSCAD 惩罚; EM 算法; 似然函数
MR(2010) **主题分类号**: 62E15; 62F10 **中图分类号**: O212.1
**文献标识码**: A **文章编号**: 0255-7797(2016)06-1315-13

## 1 引言

Erlang 混合分布广泛应用于保险损失数据的建模, 在保险破产理论和保险损失数据的拟合中都有良好的表现. 保险破产理论中, 当利用混合 Erlang 分布对保险损失的严重程度建模时, 通常关注的一些指标将有明确的解析式, 比如无限破产概率, 随机破产时刻的拉普拉斯变换等, 这方面的研究可参考文献 [3, 17, 21, 29]; 近几年, 学者更多关注于将 Erlang 混合分布用于拟合保险实际损失数据, 得到了很多令人满意的分布性质, 比如分布函数和矩都有解析式, 使得相关的风险测度 value-at-risk (VaR) 和 tail VaR (TVaR) 比较容易计算. Verbelen [30] 等将双边截断引入 Erlang 混合分布, 计算了再保险合同的纯保费. 类似研究见文献 [9, 10, 19, 26, 30] 等. Lee 和 Lin [20] 提出 Erlang 混合分布的多元形式, 多元混合 Erlang 分布保留了一元 Erlang 混合分布的大部分有用的分布性质, 同时建模相依性, 与 copula 方法相呼应. 关于多元 Erlang 混合分布的研究见文献 [2, 16, 31, 32] 等.

混合模型的首要问题是混合序的确定, Lee 和 Lin [19,30] 等都利用 BIC 来确定 Erlang 混合分布的序, Yin 和 Lin [33] 提出了一种新的 iSCAD 惩罚函数, 建立惩罚似然函数, 运用 EM 算法给出参数的估计, 同时给出了混合序的估计. 然而值得注意的是: Erlang 分布是轻尾的, 用它来拟合重尾数据时可能很难达到预期效果. 其次, 尾部数据相应的权重一般都很小, 公式 (3.11) 可以看出, 权重小于阈值 $\lambda$ 的相应 Erlang 分布都被删除, 这不利用保留拟合尾部数据的 Erlang 分布. 为解决这些问题, 本文引入极值理论拟合尾部数据, 建立 Erlang 极值混合模型.

极值混合模型广泛应用于各领域的数据分析中, 尤其在保险、金融、水文和环境科学等领域. 在保险领域, 大额索赔在保险公司的风险管理和产品定价, 尤其是再保险产品的定价方

*收稿日期: 2016-04-09 接收日期: 2016-06-28
基金项目: 国家自然科学基金资助 (11201352).
作者简介: 殷崔红 (1982–), 女, 山东潍坊, 博士, 主要研究方向: 非寿险精算.

面, 有不可忽略的意义. 文献 [4, 12, 13, 23, 27] 等将极值理论引入到保险的风险管理中. 为使数据的主体和尾部都拟合的很好, Behrens [5] 提出单一参数分布与一个极值分布的混合模型, Carreau 和 Bengio [7] 讨论混合参数分布与极值分布的混合模型, 类似的文献 [5, 6, 15, 22, 24] 等给出多种极值混合模型. Lee [18] 等最早将极值混合模型引入到保险数据中, 但是所有这些混合模型都没有考虑混合序的确定.

本文建立 Erlang 混合分布与广义帕累托 (GPD) 分布的混合模型, 广义帕累托 (GPD) 分布用于拟合数据的尾部, 而 Erlang 混合分布用于拟合数据的主体, 这样即有 Erlang 混合分布的优点, 同时保留了极值理论的长处. 引入 iSCAD 惩罚来估计混合 Erlang 分布的参数, Yin 和 Lin [33] 已证明参数和混合序的估计都有一致性.

## 2 Erlang 极值混合模型

首先给出 Erlang 分布的密度函数为

$$f(x;\gamma,\theta)=\frac{x^{\gamma-1}e^{-x/\theta}}{\theta^{\gamma}(\gamma-1)!},$$

其中 $\gamma$ 是取值为正整数的形状参数 (shape parameter), $\theta>0$ 是尺度参数 (scale parameter).

将 $m$ 个不同的 Erlang 分布以权重 $\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_m)$ 混合, 则 Erlang 混合分布的密度函数为

$$f(x;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)=\sum_{j=1}^{m}\alpha_j f(x;\gamma_j,\theta).$$

相应的分布函数为 $F(x;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)$, 其中权重参数 $\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_m)$ 满足 $\alpha_j\geq 0$ 和 $\sum\limits_{j=1}^{m}\alpha_j=1$, $\boldsymbol{\gamma}=(\gamma_1,\cdots,\gamma_m)$ 是形状参数, 为了可识别性的说明, 一般有 $\gamma_1\leq\cdots\leq\gamma_m$, 而 $\theta>0$ 是共有的尺度参数.

由于投保人的性别、车型、驾车经验和熟悉程度等的不同, 使得索赔数据一般有明显的异质性, 单一的 Erlang 分布可能很难给出好的拟合效果, 因此数据的主体部分本文仍然选用 Erlang 混合分布来拟合, 而尾部采用极值分布. 故本文采用左右双边截断的 Erlang 混合分布, 大部分保险损失数据都是已知截断值, 比如保险中的免赔额和赔偿限额. 以 $l$ 和 $\mu$ 分别表示左右截断值 (免赔额 $l$ 已知), 双边截断的 Erlang 混合分布的密度函数是

$$\begin{aligned}
f(x;l,\mu,\boldsymbol{\alpha},\boldsymbol{\gamma},\theta) &= \sum_{j=1}^{m}\alpha_j\frac{f(x;\gamma_j,\theta)}{F(\mu;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)-F(l;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)} &(2.1)\\
&= \sum_{j=1}^{m}\alpha_j\frac{F(\mu;\gamma_j,\theta)-F(l;\gamma_j,\theta)}{F(\mu;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)-F(l;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)}\frac{f(x;\gamma_j,\theta)}{F(\mu;\gamma_j,\theta)-F(l;\gamma_j,\theta)}\\
&= \sum_{j=1}^{m}\pi_j f(x;l,\mu,\gamma_j,\theta) &(2.2)\\
&\triangleq f(x;l,\mu,\boldsymbol{\pi},\boldsymbol{\gamma},\theta),
\end{aligned}$$

其中

$$\pi_j=\alpha_j\frac{F(\mu;\gamma_j,\theta)-F(l;\gamma_j,\theta)}{F(\mu;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)-F(l;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)}. \tag{2.3}$$

显然, (2.2) 式是左右截断点为 $l$ 和 $\mu$ 的 Erlang 分布 $f(x;l,\mu,\gamma_j,\theta)$ 的混合模型, 混合权重为 $\boldsymbol{\pi}=(\pi_1,\cdots,\pi_m)$, 满足 $\pi_j\geq 0$ 和 $\sum\limits_{j=1}^{m}\pi_j=1$. 密度函数 (2.1) 相应的分布函数为 $F(x;l,\mu,\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)$.

在统计中, 广义帕累托分布 (Generalized Pareto Distribution, i.e., GPD) 经常被用于拟合其他分布或实际数据的尾部, 本文选用 GPD 拟合数据的尾部, 其密度函数是

$$g_\mu(x;\xi,\sigma)=\begin{cases}\frac{1}{\sigma}\{1+\frac{\xi(x-\mu)}{\sigma}\}^{-\frac{1}{\xi}-1}, & x\in(\mu,\infty)\text{ if }\xi>0,\\ \frac{1}{\sigma}exp\{-\frac{\xi(x-\mu)}{\sigma}\}, & x\in(\mu,\infty)\text{ if }\xi=0,\\ \frac{1}{\sigma}\{1+\frac{\xi(x-\mu)}{\sigma}\}^{-\frac{1}{\xi}-1}, & x\in(\mu,\mu-\frac{\sigma}{\xi})\text{ if }\xi<0.\end{cases}\tag{2.4}$$

广义帕累托分布 (GPD) 的生存函数为

$$\overline{G}_\mu(x;\xi,\sigma)=\begin{cases}\{1+\frac{\xi(x-\mu)}{\sigma}\}^{-\frac{1}{\xi}}, & x\in(\mu,\infty)\text{ if }\xi>0,\\ exp\{-\frac{\xi(x-\mu)}{\sigma}\}, & x\in(\mu,\infty)\text{ if }\xi=0,\\ \{1+\frac{\xi(x-\mu)}{\sigma}\}^{-\frac{1}{\xi}}, & x\in(\mu,\mu-\frac{\sigma}{\xi})\text{ if }\xi<0.\end{cases}\tag{2.5}$$

结合 (2.1) 和 (2.4) 式, 为弥补引言中提过的 Erlang 混合模型的不足, 本文建立的 Erlang 极值混合模型的密度函数为

$$h(x;l,\mu,\boldsymbol{\alpha},\boldsymbol{\gamma},\theta,\xi,\sigma)=\begin{cases}(1-\psi_\mu)f(x;l,\mu,\boldsymbol{\alpha},\boldsymbol{\gamma},\theta), & \text{if } x\leq\mu,\\ \psi_\mu g_\mu(x;\xi,\sigma), & \text{if } x>\mu,\end{cases}\tag{2.6}$$

其中 $\mu$ 为阈值, $X$ 是服从 $h(x;l,\mu,\boldsymbol{\alpha},\boldsymbol{\gamma},\theta,\xi,\sigma)$ 分布的随机变量, 令 $\psi_\mu=P(X>\mu)$, 其一般由大于 $\mu$ 的样本比例来估计.

相应的生存函数为

$$\overline{H}(x;l,\mu,\boldsymbol{\alpha},\boldsymbol{\gamma},\theta,\xi,\sigma)=\begin{cases}\psi_\mu+(1-\psi_\mu)(F(u;l,\mu,\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)-F(x;l,\mu,\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)), & \text{if } x\leq\mu,\\ \psi_\mu\overline{G}_\mu(x;\xi,\sigma), & \text{if } x>\mu.\end{cases}\tag{2.7}$$

风险测度就是各种风险度量指标的总称. 现行的国际标准风险管理工具 VaR 最初由 Morgan 针对银行业务风险的需要提出的, 并很快被推广成为了一种产业标准. 风险价值 VaR 是指在正常的市场条件、给定的置信水平以及给定的持有期间内, 投资组合所面临的潜在最大损失. VaR 是一种分位数风险测度, 一般给定置信水平 $p$, 典型的 $p=95\%$ 或者 $99\%$. 但是, VaR 作为风险测度只考虑了概率为 $p$ 的事件的最大损失 $\text{VaR}_p$, 高于 $\text{VaR}_p$ 的损失并没有纳入风险测度, 为克服这个缺陷, Tail Value at Risk (or TVaR) 被提出来. 在给定置信水平 $p$ 下, TVaR 就是损失落入最糟的 $1-p$ 部分的平均损失. 下面给出 Erlang 极值混合模型关于风险指标 VaR 和 TVaR 的计算.

为便于风险指标 $\text{VaR}_p$ 和 $\text{TVaR}_p$ 的计算, 当 $X\leq\mu$ 时, 将生存函数 (2.7) 用 Erlang 密度函数分别表达为

$$\begin{aligned}
\overline{H}(x) &= \int_x^{\mu}(1-\psi_\mu)f(x;l,\mu,\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)dx + \int_{\mu}^{\infty}\psi_\mu g_\mu(x;\xi,\sigma)dx \\
&= \psi_\mu + (1-\psi_\mu)\int_x^{\mu}\sum_{j=1}^{m}\alpha_j\cdot\frac{1}{F(\mu;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)-F(l;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)}\cdot\frac{x^{j-1}e^{-x/\theta}}{\theta^j(j-1)!}dx \\
&= \psi_\mu + C\sum_{j=1}^{m}\frac{\alpha_j}{(j-1)!}\Big(\sum_{k=0}^{j-1}(\frac{x}{\theta})^k\frac{(j-1)!}{k!}e^{-x/\theta} - \sum_{k=0}^{j-1}(\frac{\mu}{\theta})^k\frac{(j-1)!}{k!}e^{-\mu/\theta}\Big) \\
&= \psi_\mu + C\sum_{j=1}^{m}\alpha_j\Big(\sum_{k=1}^{j}(\frac{x}{\theta})^{k-1}\frac{1}{(k-1)!}e^{-x/\theta} - \sum_{k=1}^{j}(\frac{\mu}{\theta})^{k-1}\frac{1}{(k-1)!}e^{-\mu/\theta}\Big) \\
&= \psi_\mu + C\sum_{k=1}^{m}\sum_{j=k}^{m}\alpha_j\Big((\frac{x}{\theta})^{k-1}\frac{1}{(k-1)!}e^{-x/\theta} - \sum_{j=k}^{m}(\frac{\mu}{\theta})^{k-1}\frac{1}{(k-1)!}e^{-\mu/\theta}\Big) \\
&= \psi_\mu + C\sum_{j=1}^{m}\sum_{k=j}^{m}\alpha_k\Big((\frac{x}{\theta})^{j-1}\frac{1}{(j-1)!}e^{-x/\theta} - \sum_{k=j}^{m}(\frac{\mu}{\theta})^{j-1}\frac{1}{(j-1)!}e^{-\mu/\theta}\Big) \\
&= \psi_\mu + C\theta\sum_{j=1}^{m}Q_j(f(x;j,\theta)-f(\mu;j,\theta)),
\end{aligned}$$

其中 $C=\frac{(1-\psi_\mu)}{F(\mu;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)-F(l;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)}$, $Q_j=\sum\limits_{k=j}^{m}\alpha_k^*$, $j=1,\cdots,m$, 其中当 $k=\gamma_j$, $\alpha_k^*=\alpha_k$, 否则 $\alpha_k^*=0$. 生存函数 (2.7) 也可以表示为

$$\overline{H}(x;l,\mu,\boldsymbol{\alpha},\boldsymbol{\gamma},\theta,\xi,\sigma)=\begin{cases}\psi_\mu+(1-\psi_\mu)\frac{\theta\sum\limits_{j=1}^{m}Q_j(f(x;j,\theta)-f(\mu;j,\theta))}{F(\mu;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)-F(l;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)}, & \text{if } x\le\mu,\\ \psi_\mu\overline{G}_\mu(x;\xi,\sigma), & \text{if } x>\mu.\end{cases}\tag{2.8}$$

假设损失随机变量 $X$ 服从 Erlang 极值混合分布 (2.6), 给定置信水平 $p$, 有

$$\overline{H}(x;l,\mu,\boldsymbol{\alpha},\boldsymbol{\gamma},\theta,\xi,\sigma)=1-p.\tag{2.9}$$

方程 (2.9) 的解即置信水平为 $p$ 的 $\mathrm{VaR}_p$.

计算 $\mathrm{TVaR}_p$ 之前, 首先研究自付责任额为 $R(>l)$ 的再保险的纯保费, 当 $R\le\mu$ 时,

$$\begin{aligned}
\prod(R) =& E((x-R)_+) = \int_R^{\mu}(1-\psi_\mu)(x-R)\cdot f(x;l,\mu,\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)dx + \int_{\mu}^{\infty}\psi_\mu(x-R)\cdot g_\mu(x;\xi,\sigma)dx \\
=&(1-\psi_\mu)[\int_R^{\infty}(x-R)\cdot f(x;l,\mu,\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)dx - \int_{\mu}^{\infty}(x-\mu)\cdot f(x;l,\mu,\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)dx \\
&-\int_{\mu}^{\infty}(\mu-R)\cdot f(x;l,\mu,\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)dx] + \psi_\mu\int_{\mu}^{\infty}(x-R)\cdot g_\mu(x;\xi,\sigma)dx \\
=&\frac{(1-\psi_\mu)}{F(\mu;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)-F(l;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)}[\theta^2\sum_{j=1}^{m}Q_j^*(f(R;j,\theta)-f(\mu;j,\theta)) - (\mu-R)\sum_{j=1}^{m}\alpha_j\overline{F}(\mu;j,\theta)] \\
&+\psi_\mu[(\mu+\frac{\sigma}{1-\xi})I(0<\xi<1)-R],
\end{aligned}$$

其中 $I(\cdot)$ 是示性函数, $Q_j^*=\sum\limits_{k=j}^{m}Q_k, j=1,\cdots,m.$

当 $R>\mu$ 时,

$$\begin{aligned}\prod(R)&=E((x-R)_+)=\int_R^\infty \psi_\mu(x-R)\cdot g_\mu(x;\xi,\sigma)dx\\&=\psi_\mu[\int_R^\infty(1+\frac{\xi(x-\mu)}{\sigma})^{-\frac{1}{\xi}}\frac{1}{\xi}dx+(\mu-R-\frac{\sigma}{\xi})(1+\frac{\xi(R-\mu)}{\sigma})^{-\frac{1}{\xi}}]\\&=\psi_\mu\frac{(R-\mu)\xi+\sigma}{1-\xi}\overline{G}_\mu(R;\xi,\sigma).\end{aligned}$$

综上, 自付责任额为 $R(>l)$ 的再保险的纯保费为

$$E((X-R)_+)=\begin{cases}\frac{1-\psi_\mu}{F(\mu;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)-F(l;\boldsymbol{\alpha},\boldsymbol{\gamma},\theta)}[\theta^2\sum\limits_{j=1}^m Q_j^*\,(f(R;j,\theta)\\-f(\mu;j,\theta))-(\mu-R)\sum\limits_{j=1}^m\alpha_j\overline{F}(\mu,j,\theta)]\\+\psi_\mu[(\mu+\frac{\sigma}{1-\xi})I(0<\xi<1)-R], & \text{if } R\le\mu,\\\frac{\sigma+(R-\mu)\xi}{1-\xi}I(0<\xi<1)\psi_\mu\overline{G}(R;\mu,\sigma,\xi), & \text{if } R>\mu.\end{cases}\tag{2.10}$$

当自付责任额 $R=\mathrm{VaR}_p$ 时, 置信水平为 $p$ 的 $\mathrm{TVaR}_p$ 为

$$\begin{aligned}\mathrm{TVaR}_p&=E(X|X>\mathrm{VaR}_p)=E((X-\mathrm{VaR}_p)|X>\mathrm{VaR}_p)+\mathrm{VaR}_p\\&=\frac{E((X-R)_+)}{1-p}+\mathrm{VaR}_p.\end{aligned}\tag{2.11}$$

## 3 EM 算法

文献 [33] 针对每一个分量权重参数 $\pi_j$, $j=1,\cdots,m$, 提出的 iSCAD 惩罚函数为

$$\begin{aligned}P_{\varepsilon,\lambda}(\pi_j)=&\lambda\{\log\frac{a\lambda+\varepsilon}{\varepsilon}+\frac{a^2\lambda^2}{2}-\frac{a\lambda}{a\lambda+\varepsilon}\}I(\pi_j>a\lambda)\\&+\lambda\{\log\frac{\pi_j+\varepsilon}{\varepsilon}-\frac{\pi_j^2}{2}+(a\lambda-\frac{1}{a\lambda+\varepsilon})\pi_j\}I(\pi_j\le a\lambda),\end{aligned}\tag{3.1}$$

其中 $I(\cdot)$ 是示性函数. 本文建立的 Erlang 极值混合分布中 Eralng 混合分布的参数估计与新引入的极值分布的参数估计互不影响, 因此关于 Eralng 混合分布的极大惩罚似然估计仍然是一致的.

Expectation–Maximization (EM) 算法最早由 Dempster [11] 给出比较详细的说明, 当似然函数的最大值点不能直接得到时, EM 算法通过迭代的方法找到最大值点. EM 算法需引入隐变量, 隐变量可以是未知参数, 丢失的数据或者任何可以使模型简化的未观测数据量. EM 算法分为 $E$-step 和 $M$-step 两步, 其中 $E$-step 计算目标函数关于隐变量 $\mathbf{Z}$ 的条件期望, $M$-step 是最大化目标函数, 求得参数的极大似然估计. 王继霞等 [1] 将 EM 算法用于有限混合 Laplace 分布的估计.

Erlang 极值混合模型的所有待估参数是: 拟合数据主体部分的 Erlang 混合分布的序 $m$, 形状参数 $\boldsymbol{\gamma}=(\gamma_1,\cdots,\gamma_m)$, 相应的权重参数 $\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_m)$, 所有 Erlang 分布共用的尺

度参数 $\theta$, 拟合数据尾部的广义帕累托分布的阈值 $\mu$, 尺度参数 $\sigma$, 形状参数 $\xi$, 下面逐一介绍它们的估计.

由公式 (2.2) 知, 密度函数 (2.6) 也可以由新权重参数 $\boldsymbol{\pi}$ 表示为

$$h(x;l,\mu,\boldsymbol{\pi},\boldsymbol{\gamma},\theta,\xi,\sigma)=\begin{cases}(1-\psi_\mu)f(x;l,\mu,\boldsymbol{\pi},\boldsymbol{\gamma},\theta), & \text{if } x\le\mu,\\ \psi_\mu g_\mu(x;\xi,\sigma), & \text{if } x>\mu.\end{cases}\tag{3.2}$$

假设 $\boldsymbol{X}=(X_1,\cdots,X_n)$ 是独立同分布的随机变量, 服从密度函数 $h(x;l,\mu,\boldsymbol{\pi},\boldsymbol{\gamma},\theta,\xi,\sigma)$, 即 (3.2), 样本观测值为 $\boldsymbol{x}=(x_1,\cdots,x_n)$, 相应有序样本观测值为 $x_{(1)}\le\cdots\le x_{(n)}$, 记

$$k=\sum_{i=1}^{n}I(x_{(i)}>\mu).\tag{3.3}$$

Pickands [25] 给出与阈值 $\mu$ 相应的 $k$ 的选择方法, 从 1 开始依次增加, 最大值为 $[n/4]$, 而 $\mu=x_{(n-k)}$, 本文最终由似然函数的大小选出 $\mu$. 为方便后面的说明, 重新表示 $n'=n-k$ 和 $x'=(x_{(1)},\cdots,x_{(n')})$.

形状参数的估计采用 Yin 和 Lin [33] 类似的方法, 即预先给定一个大的混合序 $M$, 形状参数的所有可能取值是 $\boldsymbol{\gamma}^0=(\gamma_1^0,\cdots,\gamma_M^0)$, 通过估计相应的权重参数, 来实现混合序的估计和形状参数的选择.

Erlang 极值混合分布的密度函数 $h(x;l,\mu,\boldsymbol{\pi},\boldsymbol{\gamma}^0,\theta,\xi,\sigma)$ 中的部分未知参数记为 $\boldsymbol{\phi}=(\pi_1,\cdots,\pi_M,\theta)$, 本文采用 EM 算法来估计 $\boldsymbol{\phi}$.

样本 $\boldsymbol{x}=(x_1,\cdots,x_n)$ 的对数似然函数为

$$\begin{aligned}\ell_n(\boldsymbol{\phi},\xi,\sigma;\boldsymbol{x})&=\sum_{i=1}^{n}\ln h(x;l,\mu,\boldsymbol{\pi},\boldsymbol{\gamma}^0,\theta,\xi,\sigma)\\&=\sum_{i=1}^{n}\ln[(1-\psi_\mu)\sum_{j=1}^{M}\pi_j f(x_i;l,\mu,\gamma_j^0,\theta)I(x_i\le\mu)+\psi_\mu g_\mu(x_i;\xi,\sigma)I(x_i>\mu)]\\&=\sum_{i=1}^{n'}\ln[(1-\psi_\mu)\sum_{j=1}^{M}\pi_j f(x_{(i)};l,\mu,\gamma_j^0,\theta)]+\sum_{i=n'+1}^{n}\ln[\psi_\mu g_\mu(x_{(i)};\xi,\sigma)]\\&=\ell_{n'}(\boldsymbol{\phi};\boldsymbol{x}')+\sum_{i=n'+1}^{n}\ln[\psi_\mu g_\mu(x_{(i)};\xi,\sigma)].\end{aligned}$$

样本 $\boldsymbol{x}=(x_1,\cdots,x_n)$ 的 iSCAD 惩罚对数似然函数, 其中与参数 $\boldsymbol{\phi}=(\pi_1,\cdots,\pi_M,\theta)$ 有关的部分是

$$\ell_{n',P}(\boldsymbol{\phi};\boldsymbol{x})=\ell_{n'}(\boldsymbol{\phi};\boldsymbol{x}')-n'\sum_{j=1}^{M}P_{\varepsilon,\lambda}(\pi_j).\tag{3.4}$$

直接关于 $\ell_{n',P}(\boldsymbol{\phi};\boldsymbol{x})$ 求极大似然估计是困难的, 本文使用 EM 算法, 引入隐变量, 即 $\boldsymbol{Z}=(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n)$, 其中 $\boldsymbol{Z}_i=(Z_{ij}|i=1,\cdots,n,j=1,\cdots,M)$,

$$Z_{ij}=\begin{cases}1, & \text{如果观测值 } x_i \text{ 来自第 } j \text{ 个分量分布 } f(x_i;l,\mu,\gamma_j^0,\theta),\\ 0, & \text{否则.}\end{cases}\tag{3.5}$$

那么完整样本 $(\boldsymbol{x}, \boldsymbol{Z})$ 的似然函数为

$$\begin{aligned} L_n(\boldsymbol{\phi}; \boldsymbol{x}, \boldsymbol{Z}) =& \prod_{i=1}^{n}\prod_{j=1}^{M}\{[\pi_j(1-\psi_\mu)f(x_i; l, \mu, \gamma_j^0, \theta)I(x_i \le \mu)]^{z_{ij}} \\ &+ \psi_\mu g_\mu(x_i; \xi, \sigma)I(x_i > \mu)\} \\ =& \prod_{j=1}^{M}\prod_{i=1}^{n'}[\pi_j(1-\psi_\mu)f(x_{(i)}; l, \mu, \gamma_j^0, \theta)]^{z_{ij}} \prod_{i=n'+1}^{n}[\psi_\mu g_\mu(x_{(i)}; \xi, \sigma)]. \end{aligned} \tag{3.6}$$

相应完整样本 $(\boldsymbol{x}, \boldsymbol{Z})$ 的对数似然函数为

$$\begin{aligned} \ell_n(\boldsymbol{\phi}; \boldsymbol{x}, \boldsymbol{Z}) =& \sum_{j=1}^{M}\sum_{i=1}^{n'} z_{ij}\{\ln(\pi_j) + \ln[(1-\psi_\mu)f(x_{(i)}; l, \mu, \gamma_j^0, \theta)]\} \\ &+ \sum_{j=1}^{M}\sum_{i=n'+1}^{n}\ln[\psi_\mu g_\mu(x_{(i)}, \xi, \sigma)]. \end{aligned} \tag{3.7}$$

相应的完整样本 $(\boldsymbol{x}, \boldsymbol{Z})$ 的 iSCAD 惩罚对数似然函数为

$$\ell_{n,P}(\boldsymbol{\phi}; \boldsymbol{x}, \boldsymbol{Z}) = \ell_n(\boldsymbol{\phi}; \boldsymbol{x}, \boldsymbol{Z}) - n'\sum_{j=1}^{M} P_{\varepsilon,\lambda}(\pi_j). \tag{3.8}$$

EM 算法是利用迭代过程来估计参数的方法, 假设已经完成第 $k$ 次迭代, 获得的当前估计是 $\boldsymbol{\phi}^{(k)} = (\pi_1^{(k)}, \cdots, \pi_M^{(k)}, \theta^{(k)})$, EM 算法的 $E$-step 和 $M$-step 分别为

**$E$-step** $\ell_{n,P}(\boldsymbol{\phi}; \boldsymbol{x}, \boldsymbol{Z})$ 关于隐变量 $\boldsymbol{Z}$ 求条件期望, 得到关于可观测样本 $\boldsymbol{x}$ 的边际似然函数, 即

$$\begin{aligned} Q(\boldsymbol{\phi} \mid \boldsymbol{\phi}^{(k)}) =& \sum_{j=1}^{M}\sum_{i=1}^{n'} q(\gamma_j^0 \mid x_{(i)}, \boldsymbol{\phi}^{(k)})\{\ln(\pi_j) + \ln[(1-\psi_\mu)f(x_{(i)}; l, \mu, \gamma_j^0, \theta)]\} \\ &+ \sum_{j=1}^{M}\sum_{i=n'+1}^{n}\ln[\psi_\mu g_\mu(x_{(i)}, \xi, \sigma)] - n'\sum_{j=1}^{M} P_{\varepsilon,\lambda}(\pi_j), \end{aligned} \tag{3.9}$$

其中 $q(\gamma_j^0 \mid x_{(i)}, \boldsymbol{\phi}^{(k)})$ 是观测值 $x_{(i)}$ $(i = 1, \cdots, n')$ 来自第 $j$ 个分量分布的概率,

$$q(\gamma_j^0 \mid x_{(i)}, \boldsymbol{\phi}^{(k)}) = \frac{\pi_j^{(k)}[(1-\psi_\mu)f(x_{(i)}; l, \mu, \gamma_j^0, \theta)]}{\sum_{j=1}^{M}\pi_j^{(k)}[(1-\psi_\mu)f(x_{(i)}; l, \mu, \gamma_j^0, \theta)]}. \tag{3.10}$$

**$M$-step** (3.9) 式是权重参数 $\pi_j$ $(j = 1, \cdots, M)$ 和尺度参数 $\theta$ 的函数, 求函数 (3.9) 的极

大估计, 即

$$\begin{aligned}\hat{\boldsymbol{\phi}}^{(k+1)} &= \arg\max_{\boldsymbol{\phi}\in\Phi} Q(\boldsymbol{\phi} \mid \boldsymbol{\phi}^{(k)}) \\ &= \arg\max_{\boldsymbol{\phi}\in\Phi}\{\sum_{j=1}^{M}\sum_{i=1}^{n'} q(\gamma_j^0 \mid x_{(i)}, \boldsymbol{\phi}^{(k)})\ln(\pi_j) + \sum_{j=1}^{M}\sum_{i=1}^{n'} q(\gamma_j^0 \mid x_{(i)}, \boldsymbol{\phi}^{(k)}) \\ &\quad \cdot [\ln(1-\psi_\mu) - x_{(i)}/\theta - \gamma_j^0\ln(\theta) - \ln(F(\mu;\gamma_j^0,\theta) - F(l;\gamma_j^0,\theta))] \\ &\quad + \sum_{i=n'+1}^{n}\sum_{j=1}^{M}[\ln(\psi_\mu) + \ln g_\mu(x_{(i)};\xi,\sigma)] - n'\sum_{j=1}^{M} P_{\varepsilon,\lambda}(\pi_j)\}.\end{aligned}$$

权重参数 $\pi_j$ 的第 $(k+1)$ 次迭代的估计为

$$\hat{\pi}_j^{(k+1)} = \bar{q}_j^{(k)} I(\bar{q}_j^{(k)} > a\lambda) + \frac{M}{\lambda}(\bar{q}_j^{(k)} - \lambda)_+ I(\bar{q}_j^{(k)} \le a\lambda), \tag{3.11}$$

其中 $\bar{q}_j^{(k)} \triangleq \frac{\sum_{i=1}^{n'} q(\gamma_j^0 | x_{(i)}, \boldsymbol{\phi}^{(k)})}{n'}$.

尺度参数 $\theta$ 的第 $(k+1)$ 次迭代的估计为

$$\hat{\theta}^{(k+1)} = \frac{\frac{1}{n'}\sum_{i=1}^{n'} x_{(i)} - t^{(k)}}{\sum_{j=1}^{M}\gamma_j^0\bar{q}_j^{(k)}}, \tag{3.12}$$

其中

$$t^{(k)} = \sum_{j=1}^{M}\bar{q}_j^{(k)} \left.\frac{l^{\gamma_j^0}e^{-l/\theta} - \mu^{\gamma_j^0}e^{-\mu/\theta}}{\theta^{\gamma_j^0-1}(\gamma_j^0-1)![F(\mu;\gamma_j^0,\theta) - F(l;\gamma_j^0,\theta)]}\right|_{\theta=\theta^{(k)}}. \tag{3.13}$$

迭代过程一直持续到 $|Q(\phi^{(k)}) - Q(\phi^{(k-1)})|$ 小于某个既定的误差界. 分别以 $\hat{\boldsymbol{\pi}} = \{\hat{\pi}_j | \hat{\pi}_j \neq 0, j = 1, \cdots, M\}$ 和 $\hat{\theta}$ 表示 EM 迭代的最终结果. 混合模型序的估计是

$$\hat{m} = \#\{\hat{\pi}_j | \hat{\pi}_j \neq 0, j = 1, \cdots, M\}.$$

为便于说明, 重新将 $\hat{\boldsymbol{\gamma}} = \{\gamma_j^0 | \hat{\pi}_j \neq 0, j = 1, \cdots, M\}$ 表示为 $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \cdots, \hat{\gamma}_{\hat{m}})$, 对应的权重参数记为 $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \cdots, \hat{\pi}_{\hat{m}})$. 原权重参数的估计记为 $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_{\hat{m}})$, 由方程 (2.3) 可得

$$\hat{\alpha}_j = c\frac{\hat{\pi}_j}{F(\mu;\hat{\gamma}_j,\hat{\theta}) - F(l;\hat{\gamma}_j,\hat{\theta})}, \tag{3.14}$$

其中 $c$ 是常数, 选择合适的 $c$ 进行标准化, 满足 $\sum_{j=1}^{\hat{m}}\hat{\alpha}_j = 1$.

最后, 关于广义帕累托分布 (GPD) 的尺度参数 $\sigma$ 和形状参数 $\xi$ 的极大似然估计, Coles [8] 已经详细讨论过, 本文就不再作重复说明.

本文利用 R 软件进行计算, 基于 Yin 和 Lin [33] 关于 Erlang 混合分布的 R 程序和软件包“ismev”, 编写本文 Erlang 混合分布和 GPD 分布混合模型的 R 程序, 完成模拟实验和实际数据中模型参数的估计.

## 4 模拟实验

为验证模型和估计的有效性, 本文给出一个模拟实验, 从密度函数 (2.6) 中随机抽取了 2500 个随机数, 其中 (2.6) 式中的所有参数见表 1 中的真实参数.

表 1: 真实参数与参数估计值的对比

| - | m | $\boldsymbol{\gamma}$ | $\theta$ | $\boldsymbol{\alpha}$ | $\mu$ | $\sigma$ | $\xi$ | $\psi_\mu$ |
|---|---|---|---|---|---|---|---|---|
| 真实参数 | 2 | (2, 7) | 1 | (0.5, 0.5) | 10 | 3 | 0.4 | 0.1 |
| 估计参数 | 2 | (2, 7) | 1.0135 | (0.501, 0.499) | 10.571 | 3.0142 | 0.466 | 0.083 |

参数的初始化主要参考文献 [19, 28]. 事先给定 $M = 10$, 形状参数的备择范围即 $\boldsymbol{\gamma} = (1, \cdots, 10)$, 以 Tijms [28] 的方法初始化, 公式 (3.11) 给出极大惩罚似然的权重参数估计, 其稀疏性实现了在形状参数备择范围 $\boldsymbol{\gamma} = (1, \cdots, 10)$ 中进行合理选择. 从表 1 可以看出, 形状参数最终仅选中 $\hat{\boldsymbol{\gamma}} = (2, 7)$, 只有这两个形状参数对应的权重参数估计为非零的, 即 $(\hat{\alpha}_2, \hat{\alpha}_7) = (0.501, 0.499)$, 其它形状参数相应的权重参数估计均为零, 即 $\hat{\alpha}_j = 0, j = 1, 3, 4, 5, 6, 8, 9, 10$. 显然, 混合模型序的估计 $\hat{m} = 2$.

由本实验可以看出, 引入 iSCAD 惩罚的优势所在: 通过对权重参数的估计, 同时实现了对形状参数的选择和混合模型序的估计. 表 1 列出的所有参数估计值与真实值都很接近, 说明模型和算法都很有效, 能够反映出数据的特征. 图 1 很好的反应了这一点, 图 1 中的真实曲线和拟合曲线几乎是重合的.

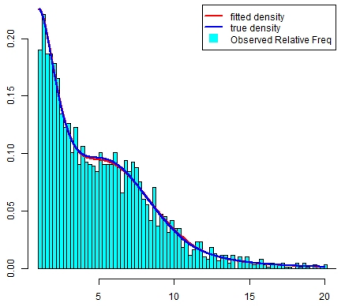


图 1: 模拟数据的直方图, 真实曲线与拟合曲线

## 5 丹麦火灾数据

丹麦火灾赔偿数据有 2167 个观测值, Embrechts [14] 和 Mendes [24] 等都用极值理论研究过这组数据的尾部, 本文采用 Erlang 极值混合模型从总体上研究这组数据, 不再仅仅限于研究其尾部特征.

文献 [33] 讨论了带左截断点 $l$ 的 Erlang 混合分布, 本文在其基础上提出了 Erlang 极值混合分布, 在本例中将利用这两种不同的分布分别拟合丹麦火灾赔偿数据, 比较两种分布的优劣.

表 2 给出 Erlang 混合分布和 Erlang 极值混合分布 (2.6) 拟合火灾损失数据得到的所有参数的估计值, 其中利用 Erlang 极值混合分布得到的结果说明拟合数据的主体部分采用了三个 Erlang 分布, 数据的尾部由广义帕累托分布来拟合, 两部分的阈值点为 4.174, 尾部数据比例为 0.152; 而利用 Erlang 混合分布拟合同一组火灾数据则需要十个不同的 Erlang 分布的混合.

表 2: 参数估计值

| 参数估计 | m | $\boldsymbol{\gamma}$ | $\theta$ | $\boldsymbol{\alpha}$ | $\mu$ | $\sigma$ | $\xi$ | $\psi$ |
|---|---|---|---|---|---|---|---|---|
| Erlang 混合分布 | 10 | (2, 3, 9, 15, 25, 35, 45, 55, 65, 75) | 0.48 | (0.168, 0.596, 0.134, 0.05, 0.022, 0.011, 0.01,0.004, 0.004, 0.002) | - | - | - | - |
| Erlang 极值混合分布 | 3 | (2, 3, 9) | 0.487 | (0.679, 0.224, 0.097) | 4.174 | 3.068 | 0.661 | 0.152 |

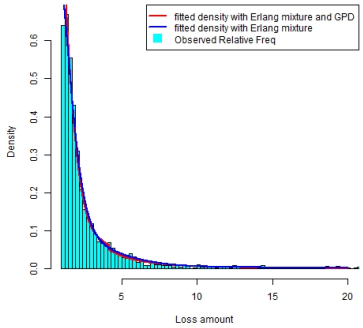


图 2: 丹麦火灾数据的直方图与拟合曲线

图 2 是丹麦火灾数据的直方图、Erlang 混合分布和 Erlang 极值混合分布的拟合曲线, 可以看出拟合效果较好.

图 3 和 4 分别给出 Erlang 混合分布和 Erlang 极值混合分布的 Q-Q 图, 显然 Erlang 极值混合分布在尾部数据的拟合上更优.

本文给出 VaR 的非参数 (nonparametric) 法估计作为标杆, 在置信水平为 $p$ 的条件下, $\mathrm{VaR}_p$ 的非参数估计是方程 $F_n(\mathrm{VaR}_p) = p$ 的解, 其中 $F_n(x) = \frac{\sum_{i=1}^{n} I(x_i \le x)}{n}$.

表 3: 非参数法、Erlang 混合分布和 Erlang 极值混合分布的 $\mathrm{VaR}_p$ 值的比较

| $\mathrm{VaR}_p$ | $p = 0.2$ | $p = 0.15$ | $p = 0.1$ | $p = 0.05$ | $p = 0.01$ |
|---|---|---|---|---|---|
| 非参数法 | 3.478227 | 4.259546 | 5.541526 | 9.972647 | 26.042526 |
| Erlang 混合分布 | 3.763277 | 4.702897 | 6.246157 | 10.398645 | 24.64542 |
| Erlang 极值混合分布 | 3.490014 | 4.22099 | 5.661758 | 9.223786 | 27.61687 |

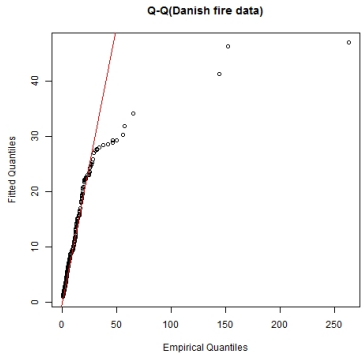


图 3: 丹麦火灾数据的 Q-Q 图 (Erlang 混合分布)

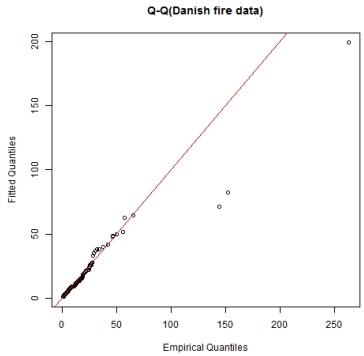


图 4: 丹麦火灾数据的 Q-Q 图 (Erlang 极值混合分布)

表 3 给出三种方法的 $\mathrm{VaR}_p$ 估计值, 表 3 可以看出, Erlang 极限混合分布估计得到的 $\mathrm{VaR}_p$ 与非参数法得到的 $\mathrm{VaR}_p$ 非常接近, 估计效果很好.

表 4: 非参数法、Erlang 混合分布和 Erlang 极值混合分布的 $\mathrm{TVaR}_p$ 值的比较

| $\mathrm{VaR}_p$ | $p=0.2$ | $p=0.15$ | $p=0.1$ | $p=0.05$ | $p=0.01$ |
|---|---|---|---|---|---|
| 非参数法 | 9.976278 | 11.90635 | 15.89804 | 23.31677 | 63.5827 |
| Erlang 混合分布 | 9.064751 | 10.68636 | 13.33043 | 18.8377 | 30.791 |
| Erlang 极值混合分布 | 10.98294 | 13.36598 | 17.61856 | 28.13227 | 82.42143 |

表 4 给出非参数法、Erlang 混合分布和 Erlang 极值混合分布的 $\mathrm{TVaR}_p$ 估计值, 其中 TVaR 的非参数估计为 $\mathrm{TVaR}_p = \frac{\sum_{i=1}^{n}(x_i \cdot I(x_i > \mathrm{VaR}_p))}{\sum_{i=1}^{n} I(x_i > \mathrm{VaR}_p)}$. Erlang 混合分布的 $\mathrm{TVaR}_p$ 比非参数法的结果偏小, 这主要是因为 Erlang 混合分布对火灾损失数据的尾部拟合不足, 见图 3; 而 Erlang 极值混合分布的结果稍大, 而且越到尾部, 这种趋势越明显, 这主要是因为估计得到的 $\hat{\xi} = 0.661 > 0$, 即估计的极值分布为厚尾的, 而实际数据的尾部过于稀疏, 不足以表现这种厚尾性.

## 参 考 文 献

[1] 王继霞, 汪春峰, 苗雨. 有限混合 Laplace 分布回归模型局部估计的 EM 算法 (英文)[J/OL]. 数学杂志, 2016, 36(4): 667–675.

[2] Badescu A L, Lan G, Lin X S, et al. Modeling correlated frequencies with application in operational risk management[J]. J. Oper. Risk, 2015, 10(1): 1–43.

[3] Bargès M, Loisel S, Venel X. On finite-time ruin probabilities with reinsurance cycles influenced by large claims[J]. Scand. Actua. J., 2013, 2013(3): 163–185.

[4] Beirlant J, Goegebeur Y, Segers J, et al. Statistics of extremes: theory and applications[M]. Ltd England: John Wiley & Sons, 2006.

[5] Behrens C N, Lopes H F, Gamerman D. Bayesian analysis of extreme events with threshold estimation[J]. Stat. Model., 2004, 4(3): 227–244.

[6] Carreau J, Bengio Y. A hybrid Pareto model for asymmetric fat-tail data[R]. Technical Report 1283, Canada: Dept. IRO, Université de Montréal, 2006.

[7] Carreau J, Bengio Y. A hybrid pareto mixture for conditional asymmetric fat-tailed distributions[J]. Neural Net., IEEE Trans., 2009, 20(7): 1087–1101.

[8] Coles S, Bawa J, Trenner L, et al. An introduction to statistical modeling of extreme values[M]. London: Springer, 2001.

[9] Cossette H, Mailhot M, Marceaué. TVaR-based capital allocation for multivariate compound distributions with positive continuous claim amounts[J]. Insur.: Math. Econ., 2012, 50(2): 247–256.

[10] Cossette H, Coté M P, Marceau E, et al. Multivariate distribution defined with Farlie Gumbel Morgenstern copula and mixed Erlang marginals: Aggregation and capital allocation[J]. Insur.: Math. Econ., 2013, 52(3): 560–572.

[11] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm[J]. J. Royal Stat. Soc., Ser. B (Methodological), 1977: 1–38.

[12] Embrechts P, Klüppelberg C, Mikosch T. Modelling extremal events for insurance and finance[J]. Springer, 1997, 71(2): 183–199.

[13] Embrechts P, Resnick S I, Samorodnitsky G. Extreme value theory as a risk management tool[J]. North Amer. Actua. J., 1999, 3(2): 30–41.

[14] Embrechts P, Klüppelberg C, Mikosch T. Modelling extremal events: for insurance and finance[M]. Germany: Springer Sci. Business Media, 2013.

[15] Frigessi A, Haug O, Rue H. A dynamic mixture model for unsupervised tail estimation without threshold selection[J]. Extremes, 2002, 5(3): 219–235.

[16] Hashorva E, Ratovomirija G. On Sarmanov mixed Erlang risks in insurance applications[J]. Astin Bull., 2015, 45(01): 175–205.

[17] Landriault D, Willmot G E. On the joint distributions of the time to ruin, the surplus prior to ruin, and the deficit at ruin in the classical risk model[J]. North Amer. Actua. J., 2009, 13(2): 252–270.

[18] Lee D, Li W K, Wong T S T. Modeling insurance claims via a mixture exponential model combined with peaks-over-threshold approach[J]. Insur.: Math. Econ., 2012, 51(3): 538–550.

[19] Lee S C K, Lin X S. Modeling and evaluating insurance losses via mixtures of Erlang distributions[J]. North Amer. Actua. J., 2010, 14(1): 107–130.

[20] Lee S C K, Lin X S. Modeling dependent risks with multivariate Erlang mixtures[J]. Astin Bull., 2012, 42(01): 153–180.

[21] Lin X S, Willmot G E. The moments of the time of ruin, the surplus before ruin, and the deficit at ruin[J]. Insur.: Math. Econ., 2000, 27(1): 19–44.

[22] MacDonald A, Scarrott C J, Lee D, et al. A flexible extreme value mixture model[J]. Comput. Stat. Data Anal., 2011, 55(6): 2137–2157.

[23] McNeil A J. Estimating the tails of loss severity distributions using extreme value theory[J]. Astin Bull., 1997, 27(01): 117–137.

[24] Melo Mendes B V, Lopes H F. Data driven estimates for mixtures[J]. Comput. Stat. Data Anal., 2004, 47(3): 583–598.

[25] Pickands III J. Statistical inference using extreme order statistics[J]. Ann. Stat., 1975: 119–131.

[26] Porth L, Zhu W, Seng Tan K. A credibility-based Erlang mixture model for pricing crop reinsurance[J]. Agricul. Finance Rev., 2014, 74(2): 162–187.

[27] Resnick S I. Discussion of the Danish data on large fire insurance losses[J]. Astin Bull., 1997, 27(01): 139–151.

[28] Tijms H C. A first course in stochastic models[M]. UK: John Wiley and Sons, 2003.

[29] Tsai C C L, Willmot G E. On the moments of the surplus process perturbed by diffusion[J]. Insur.: Math. Econ., 2002, 31(3): 327–350.

[30] Verbelen R, Gong L, Antonio K, et al. Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm[J]. Astin Bull., 2015, 45(03): 729–758.

[31] Verbelen R, Antonio K, Claeskens G. Multivariate mixtures of Erlangs for density estimation under censoring[J]. Life. Data Anal., 2015: 1–27.

[32] Willmot G E, Woo J K. On some properties of a class of multivariate Erlang mixtures with insurance applications[J]. Astin Bull., 2015, 45(01): 151–173.

[33] Yin C, Lin X S. Efficient estimation of Erlang mixtures using iSCAD penalty with insurance application[J]. Astin Bull., Available on CJO2016, doi:10.1017/asb.2016.14.

# EFFICIENT ESTIMATION OF ERLANG AND GPD MIXTURES USING ISCAD PENALTY WITH INSURANCE APPLICATION

YIN Cui-hong[1], LIN Xiao-dong[1,2], YUAN Hai-li[3]

(*1.School of Mathematical Sciences, Xiamen University, Xiamen 361005, China*)

(*2.Department of Statistical Sciences, University of Toronto, Ontario M5S 3G3, Canada*)

(*3.School of Mathematics and Statistics Sciences, Wuhan University, Wuhan 430072, China*)

**Abstract:** In this paper, we study efficient estimation of Erlang & GPD mixture model. By using a new thresholding penalty function and a corresponding EM algorithm, we estimate model parameters and determine the order of the mixture model. We obtain risk measure including VaR and TVaR and show efficiency of the new mixture model in simulation studies and a real data application, which improve Erlang & extreme value mixture model in modeling insurance losses.

**Keywords:** extreme value theory; mixture model; iSCAD penalty; EM algorithm; likelihood function

**2010 MR Subject Classification:** 62E15; 62F10