学校编码：10384 分类号______ 密级______

学号：19020141152613 UDC______

厦门大学

硕 士 学 位 论 文

SCAD惩罚混合广义Pareto模型的参数估计及应用

Parameter Estimation and Application of Generalized Pareto Mixtures via SCAD Penalty

蔡 庆 淞

指导教师姓名：谭 忠 教 授

专 业 名 称：概率论与数理统计

论文提交日期：2017 年 05 月

论文答辩日期：2017 年 05 月

学位授予日期：2017 年 06 月

答辩委员会主席：__________

评 阅 人：__________

2017 年 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

本文主要研究混合双参数广义Pareto分布模型(下面简称混合广义Pareto模型)的概率密度估计，其密度函数是：

$$f(x;p,\lambda,\theta)=\sum_{j=1}^{m}p_j\theta_j\frac{\lambda_j^{\theta_j}}{(\lambda_j+x)^{\theta_j+1}},x>0$$

其中$p=(p_1,\cdots,p_m),0<p_j<1$是混合权重系数，且$\sum_{j=1}^{m}p_j=1$.参数$\lambda=(\lambda_1,\cdots,\lambda_m),\theta=(\theta_1,\cdots,\theta_m),\lambda_j,\theta_j>0,j=1,\cdots,m$。

混合模型能够非常好地拟合现实数据。Pareto分布的特点是有厚尾特征，所以在金融风险度量方面受到越来越多的重视。与此同时，在保险、可靠性分析等方面应用十分普遍。实际运用中经常有需要用到多成分混合广义pareto模型的情况。

本文选取实际运用最广泛的一类广义pareto分布进行研究。在采取传统的矩估计和极大似然估计对混合分布进行参数估计时，发现它们确实在理论上很好实现，但是在实际计算中却特别繁琐。EM算法是常用的估计参数隐变量的利器，是一种迭代算法，经常被用来解决极大似然估计。它能够从非完整数据集中对参数进行极大似然估计，计算出后验密度函数，算法的最大优点是简单和稳定，不过容易陷入局部最优，对初值的选取十分依赖。我们利用K-means聚类来选取迭代初值。但是，如何确定聚类个数即混合分布模型的分支数m成为了一个难点。本文引入SCAD 惩罚，通过惩罚权重参数，去掉多余的分支密度函数，从而解决分支个数未知及迭代初值选取的问题。最后我们采用K-S两独立样本同质检验检验模型。

本文主要做了以下3点贡献：

(1)在混合广义Pareto模型中创新性地引入SCAD惩罚，证明了混合广义Pareto模型的SCAD惩罚似然估计是一致的。

(2)计算出m成分混合广义Pareto模型参数估计迭代公式，并引入SCAD惩罚，克服如何确定混合序m这一难题，最后通过MATLAB和R软件进行数值模拟验证了方法的可

行性。

(3)利用混合广义Pareto模型分析2016年中国各省及自治区人口密度分布和2016年胡润中国富豪排行榜。证明了SCAD惩罚EM算法估计混合广义Pareto模型在各种样本容量大小的情况下都能快速实现，并且效果很好。

关键词：SCAD惩罚；混合广义pareto分布；EM算法

Abstract

In this paper, we consider the estimation of a two parameters generalized Pareto mixture model(Hereinafter referred to as generalized Pareto mixture model) with density:

$$f(x|p,\lambda,\theta) = \sum_{j=1}^{m} p_j\theta_j \frac{\lambda_j^{\theta_j}}{(\lambda_j + x)^{\theta_j+1}}, x > 0$$

where $p = (p_1, \cdots, p_m), 0 < p_j < 1$ is a mixed weight coefficient with $\sum_{j=1}^{m} p_j = 1$. and $\lambda = (\lambda_1, \cdots, \lambda_m), \theta = (\theta_1, \cdots, \theta_m), \lambda_j, \theta_j > 0, j = 1, \cdots, m$.

The mixture model fits the real data very well. Due to the characteristics of thick tail, the Pareto distribution has been paid more and more attention in the field of financial risk measurement. At the same time, it has been widely used in insurance and reliability analysis. In practice, we often need to use the generalized Pareto mixture model.

In this paper, we select the most widely used class of generalized Pareto distribution. The traditional method of moment estimation and maximum likelihood estimation was used to estimate the parameters of mixture distribution. It is found that the traditional method of moment estimation and maximum likelihood estimation which is very good. EM algorithm is an iterative algorithm to estimate the maximum likelihood which is very common. It can be carried out on the maximum likelihood estimation of parameters from incomplete data, calculated the posterior density function.The biggest advantage of the proposed algorithm is simple and stable, but easy to fall into local optimum, selection is very dependent on the initial value. K-means clustering provides a method for selecting initial value of iteration. However, how to cluster the number of mixed distribution of the number of branches m becomes a problem. In this paper, we introduce the SCAD penalty to remove the redundant branch density function by punishing the weight parameter, so as to solve the problem that the number of branches is unknown and the initial value of iteration is chosen.Finally, we use K-S two-independent sample homogeneity test model.

There are three contributions in this paper :

(1) The SCAD penalty is introduced to prove that the SCAD penalized likelihood estimator for the generalized Pareto mixture model is uniform.

(2) Calculated the iterative formula for parameter estimation of the generalized Pareto mixture model, the introduction of SCAD punishment, and overcome the problem of how to determine the mixing order m, finally through the MATLAB and R software numerical simulation to verify the feasibility of the improved EM algorithm.

(3) The distribution of population density of China's provinces and autonomous regions in 2016 and the Hurun Report of China rich list in the year of 2016 by via the generalized Pareto mixture model.It is proved that EM algorithm via the SCAD penalty can be used to estimate the generalized Pareto mixture model in the case of various sample sizes, and the results are very good.

Key words: SCAD penalty; Generalized Pareto mixture model; EM algorithm.

目　录

Contents

第1章 绪论

本章首先概述本论文研究问题的背景及意义，然后简要介绍本文拟解决的问题及其处理方法，最后简要介绍本文的结构安排。

1.1 问题的背景及意义

在实际生活中，我们需要更好地对各种各样复杂的随机现象进行建模，混合模型是解决该问题的一个很好的工具。近年来，越来越受到人们的重视，在自然科学、工程学、医学、金融学等领域都有了广泛的运用。Pareto分布族的特点是有厚尾特征,在金融分析、寿命分析等领域中都是一个必不可少的统计模型。

本文主要研究混合双参数广义Pareto分布模型(下面简称混合广义Pareto模型)的概率密度估计，其密度函数是：

$$f(x;p,\lambda,\theta)=\sum_{j=1}^{m}p_j\theta_j\frac{\lambda_j^{\theta_j}}{(\lambda_j+x)^{\theta_j+1}},x>0$$

其中$p=(p_1,\cdots,p_m),0<p_j<1$是混合权重系数，且$\sum_{j=1}^{m}p_j=1$.参数$\lambda=(\lambda_1,\cdots,\lambda_m),\theta=(\theta_1,\cdots,\theta_m),\lambda_j,\theta_j>0,j=1,\cdots,m$。参数集$\alpha=(p,\lambda,\theta)$。

我们可以很明显地看到混合广义Pareto模型并不像简单模型参数估计那么便于计算，传统的矩法估计、极大似然估计不能发挥出它们应有的作用。1977年，Rubin[17]等人首先提出了EM算法,大大简化计算极大似然估计。 EM算法的最大优点是简单稳定，对大数据处理有优越性。有部分文献已经尝试用EM算法来估计混合分布问题.上世纪70年代末80年代初，Quandt 和Ramsey[15], Aitkin和Wilson[34] 对该领域作出重大贡献，发表多篇文章。2009年，刘媚和汤银才[36]解决了两成分混合广义Pareto模型的参数估计。

然而EM算法是一种局部收敛算法，这导致其存在着缺陷，十分依赖于初值选取。矩估计是初值选择的好方法，但是混合模型传统矩估计计算十分繁琐。2007年，

Christopher M.Bishop[22] 采用K-means聚类改进EM 算法。通过K-means聚类对数据进行处理、分类，得到K组数据，对每一组数据利用矩估计计算出每组的参数估计，并把估计值粗略的作为算法的初始值。然而实际运用中，对于混合模型，K-means 聚类k 值很难选择，我们并不知道混合模型有几个分支。

回归分析中，尤其是稀疏的广义线性模型中，经常使用惩罚函数来实现变量的选择。2001年，Fan和Li[12]提出了惩罚回归系数的smoothly clipped absolute deviation (SCAD) 函数，得到的回归系数估计具有稀疏性，实现了变量选择。混合模型和线性模型都具有类似的线性结构，本文参考回归模型中变量选择的思想，将其用于混合模型中分量模型的选择，去掉聚类分析中多余的分支密度函数。

1.2　Pareto分布及其性质

Pareto分布起源于意大利经济学家Pareto对意大利20%的人口拥有80%的财产这一现象的观察,最早研究时视其为一种收入分布，慢慢地发展被概括为帕累托法则，也被称作80/20 法则。最后，被进一步发展为Pareto分布。随后，由于该分布的概率密度函数是递减的一种失效率函数。我们最常见的运用是用它来描述个人收入的分布情况，这是由于个人收入有收入越高,继续提高收入的能力也会提高。它在不同的领域中越来越受到重视。后来人们发现，采用Parcto分布分析研究股票收益分布非常有效，这是因为收益分布有厚尾特征，并且明显偏离正态分布。在医学上，人们观察到针对个人的医疗消费数据也会有偏态、厚尾的特征。因此，Pareto 分布被运用到国家医疗健康数据，用来分析尾部情况。在环境学中，Pareto分布族的运用十分普遍，比如对诸如风雨雷电等各种自然现象的研究。值得关注的是，在精算师常用的八大分布中，有两个分布属于Pareto 分布族。总总表明Pareto 分布族在实际中的运用价值是十分巨大的，同时也间接地说明了我们有必要对Pareto分布族进行深入的研究。

定义 1.1　若X的分布函数为$f(x)=1-(1+x)^{-1}, x>0$,则称X为标准的pareto分布。对应的密度函数为

$$p(x)=\frac{1}{(1+x)^2}$$

Pareto分布族有衍生的多种形式。例如：

(1)$F(x)=1-(\frac{x}{\sigma})^{-\sigma}$

(2)$F(x)=1-(1+\frac{x-\mu}{\sigma})^{-\sigma}, x>\mu, \alpha>1$,该分布又称Lomax分布，常用于对商业失败数据的分析。

(3)$F(x)=1-(1+\frac{x-\mu}{\sigma}^{\frac{1}{\gamma}})^{-1}, x>\mu, \sigma>0, \gamma>0$

(4)$F(x)=1-((1+\frac{x-\mu}{\sigma}^{\frac{1}{\gamma}})^{-1})^{-\alpha}, \sigma>0, \gamma>0, \alpha>0$

$\cdots$

$\cdots$

Parcto分布的优良性质使得其在拟合保险中的损失数据分布有着重要的作用，精算师渐渐发现拟合某些损失数据，如果运用下面要介绍的一种推广的Pareto分布，效果会更好。于是这就促进统计学家去研究这一广义Pareto分布，同时，这也推动了广义Parcto分布发展的进程。

定义 2.1　若X的分布密度函数为

$$p(x)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\lambda^\alpha x^{\beta-1}}{(\lambda+x)^{\alpha+\beta}}, x>0$$

则称X服从广义Pareto分布。

其中，又以双参数广义Pareto分布运用最为广泛，即当$\beta=1$时。此时X的密度函数为

$$p(x)=\frac{\theta\lambda^\theta}{(\lambda+x)^{\theta+1}}, x>0, \lambda>0, \theta>0$$

X的分布函数为：

$$F(x)=\int_0^x\frac{\theta\lambda^\theta}{(\lambda+t)^{\theta+1}}dt$$

$$
\begin{aligned}
&= -\lambda^{\theta}\int_0^x d((\lambda+t)^{-\theta}) \\
&= -\lambda^{\theta}(\lambda+t)^{-\theta}|_0^x \\
&= 1-(\frac{\lambda}{\lambda+x})^{-\theta}
\end{aligned}
$$

其k阶原点矩为:

$$
\begin{aligned}
m_k &= E(X^k) \\
&= \int_0^{\infty}\frac{\Gamma(\theta+1)}{\Gamma(\theta)}\frac{\lambda^{\theta}x^k}{(\lambda+x)^{\theta+1}}dx \\
&= \frac{\Gamma(\theta-k)\Gamma(k+1)}{\Gamma(\theta)}\lambda^{\theta}\lambda^{k} \qquad\qquad \theta > k
\end{aligned}
$$

特别的有

$$
\begin{aligned}
&m_1 = \frac{\lambda}{\theta-1} \quad \theta > 1 \\
&m_2 = \frac{2\lambda^2}{(\theta-1)(\theta-2)} \quad \theta > 2 \\
&\cdots
\end{aligned}
$$

所以，分布的期望和方差分别为

$$
\begin{aligned}
E(x) &= m_1 = \frac{\lambda}{\theta-1} \quad \theta > 1 \\
D(x) &= E(x^2)-[E(x)]^2 \\
&= \frac{\lambda^2\theta}{(\theta-1)^2(\theta-2)}
\end{aligned}
$$

根据矩法估计，可列方程

$$
\begin{aligned}
\bar{X} &= \frac{\lambda}{\theta-1} \\
S^2 &= \frac{\lambda^2\theta}{(\theta-1)^2(\theta-2)}
\end{aligned}
$$

反解得

$$\hat{\lambda} = \frac{\bar{X}S^2 + \bar{X}^3}{S^2 - \bar{X}^2}$$
$$\hat{\theta} = \frac{S^2}{S^2 - \bar{X}^2} \tag{1-1}$$

1.3 EM算法

EM算法是估计参数隐变量的利器，它的基本思想是：若参数Θ已知，那么我们可以根据训练数据推断出最优隐变量Z的值，反之，若Z的值已知，则可方便地对参数Θ做极大似然估计。我们事先肯定是不知道模型的真实参数，通过聚类或者其他方法得到初始参数$\theta^{(0)}$，利用当前估计的参数值计算对数似然的期望值，寻找能使当前产生的似然期望最大化的参数值。新得到的参数值重新被作为初始值，重复以上步骤，指导收敛到局部最优解。

我们假定集合$Z=(X,Y)$由观测数据X和缺失数据Y组成。$z=(x,y)$和x分别被称为完整数据和不完整数据。记$f(x,y|\theta)$为z的联合概率密度，其中θ为需要估计的参数。θ的极大似然估计是求能使不完整数据的对数似然函数$ln \quad L(\theta|x)$的最大化的参数值而得到。

EM算法把求极大似然估计的过程分两步走:第一步基于当前参数$\Theta^{(t)}$推断隐变量分布；第二步基于已观测变量X和$Z^{(t)}$对参数Θ做极大似然估计。即分成E步和M步，两个步骤交替计算，直到收敛到最优解。

假设在算法第t次迭代θ，获得的估计记为$\theta^{(t)}$，则在t+1次迭代时：

E步：基于当前观测数据x和第t步估计值$\theta=\theta^{(t)}$来推断隐变量Z的分布，并计算对数似然函数关于Z的期望。记为：

$$Q(\theta|\theta^{(t)},x) = E[L(\theta|z)|x,\theta^{(t)}] \tag{1-2}$$