学校编码：10384　　分类号_____ 密级_____

学号：19020141152622　　UDC_____

厦门大学

硕士学位论文

# 粗糙惩罚混合Erlang模型在密度估计上的应用

## Fitting mixed Erlang densities under Roughness Penalty

张鑫梅

指导教师姓名：黄荣坦 副教授

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 6 月

学位授予日期：2017 年 7 月

答辩委员会主席：_____

评　　阅　　人：_____

2017 年 5 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

在统计学中，如何较好地拟合一组给定数据的密度函数并给出密度曲线的参数形式一直备受关注。尤其是实际问题中经常遇见的删失数据和多峰数据的拟合，不但要求密度曲线具有高度的灵活性，而且不能出现过度拟合现象，这就大大增加了拟合难度。其中，对于混合模型的研究越来越受到人们的重视，是因为混合模型是一种介于参数方法与非参数方法间的半参数模型，这种半参数模型的优越在于它既避免了参数模型对数据结构的拟合偏离问题又具有分布函数可知等非参数模型无法具备的拟合性质。继 Tijims[2]给出证明，在弱收敛的意义下，具有相同尺度参数的混合Erlang 模型可以无限逼近任意分布后，利用混合Erlang 模型解决金融、保险等行业的数据建模问题有了更加广泛和深入的应用。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，通常拟合都有良好的表现。

混合Erlang 分布的密度如下:

$$f(x;\alpha,m,\theta)=\sum_{k=1}^{K}\alpha_k\frac{x^{m_k-1}e^{-\frac{x}{\theta}}}{\theta^{m_k}(m_k-1)!},$$

其中 $\theta$ 是这个混合分布公用的尺度参数。$\alpha$ 表示各个Erlang分支在混合分布中所占权重，是个权重向量，满足$0<\alpha_k\leq 1, k\in 1,...,K$, 及 $\sum_{k=1}^{K}\alpha_k=1$. $m=(m_1,...,m_K)$是分布的形状参数向量，每个分量均为正整数. K是混合模型的序，即分支分布的个数。

Lee & Lin[3,4]将Expectation-Maximization(EM) 算法引入混合 Erlang模型的参数估计中，其本质就是利用迭代的EM 算法估计出模型的参数。由于迭代算法对初值的选取依赖很强，不同的初值选取方法对拟合结果会产生不同的影响。Gui等[7,8]利用事先确定的混合个数对数据进行聚类得到初值，再给出BIC 准则选出最佳模型的方法来避免序的过度拟合。在 Lee & Lin[3]和 Yin & Lin[9] 文章中，EM 算法的尺度参数初值来自于一个非常大的备择空间，通过不断迭代来将表现较差的尺度参数估计剔除。这种方法事先所选参数范围较大，尤其是混合模型的序，很容易出现过度拟合现象。因此本文参考 Gui等[7,8]方法来确定迭代初值。

混合 Erlang模型的线性结构很好的实现了异质性，但是产生了不可避免的问题：混合个数（序）的确定。很多学者讨论过正态混合模型序的确定，主要包括最小距离法，假设检验法，惩罚似然法等。Fan & Li[10] 提出应用于线性回归模型的SCAD惩

罚函数，通过惩罚回归系数，实现变量选择和回归系数的估计，由于混合模型的线性结构与线性回归结构类似，在 Yin & Lin[9] 中，类似 SCAD惩罚函数，作者提出一种新的关于混合权重向量估计的阈值惩罚函数，iSCAD 惩罚。通过惩罚混合权重来确定混合分支的个数，即模型的序选择。并且作者给出了估计量满足稀疏性、连续性和无偏性的证明。但在实际数据的密度函数估计中，该惩罚算法收敛性质受到多个因子的影响，可能会出现收敛速度较慢的现象。基于以上，本文提出另一种惩罚似然函数的观点。受到传统的粗糙惩罚定义的启发，将连续的随机变量密度函数三阶导平方作为惩罚项，粗糙惩罚定义如下如下：

$$PEN_3 = \int_0^\infty [f'''(x|\alpha, m, \theta)]^2 dx = \int_0^\infty ([\sum_{k=1}^{K} \alpha_k \frac{x^{m_k-1} e^{-\frac{x}{\theta}}}{\theta^{m_k}(m_k-1)!}]''')^2 dx.$$

在本文第四章中，我们通过模拟数据的例子和实际数据密度估计来证实粗糙惩罚后的序估计要优于未惩罚的序估计。

**关键词：** 混合Erlang模型；EM算法；惩罚似然估计

# Abstract

In statistics, the problem of how to give the density function of given data well by mathematical expression has drawn too much attention. Especially when it comes to the censored data or multi-mode data, the density curve needs not only be of high flexibility but also can get rid of overfit, which greatly increase the difficulty of fitting.Among them, the development of mixed model has received people's attention, because the hybrid model is a semi-parametric model instead parametric model or non parametric model. It avoids the deviation problem between fitting curve and the raw data structure and provides mathematical expression of the fitting curve directly. After Tijims[2] proving that, in the sense of weak convergence, mixed Erlang model with a common scale parameter can converge to any positive distributions, which means mixed Erlang can fit any positive distribution with arbitrary precisions, the performance of mixed Erlang models is widely used in modeling financial and insurance data.

The expression of mixed Erlang distributions is,

$$f(x;\alpha,m,\theta)=\sum_{k=1}^{K}\alpha_k\frac{x^{m_k-1}e^{-\frac{x}{\theta}}}{\theta^{m_k}(m_k-1)!},$$

where $\alpha$ represents the weight of each Erlang branch in the mixed distribution, say a weight vector, and satisfies $0<\alpha_k\leq 1, k\in 1,...,K$, and $\sum_{k=1}^{K}\alpha_k=1$. $\theta$ is a common scale parameter used by all Erlang distribution branches. $m=(m_1,...,m_K)$ is a set of shape parameters, each component of which is a positive integer. K is the order of the mixed model, that is, the number of Erlang distribution branches.

In Lee & Lin's[3,4], EM algorithm was introduced into the estimation of the mixed Erlang model. In fact, EM algorithm is a kind of iterative algorithms.Because of the strong dependence on the initial values, different initials may influence the fitting results a lot. In [7,8], the initial values of scale parameter came from a very large alternative space, through continuous iteration, the scale parameter with poor performance will be deleted. However, due to a large range of parameters, especially the order of the mixed model, it is prone to overfit. In this paper, CMM[7] initial method is used to determine the initial values of EM algorithm.

The linear structure of the hybrid Erlang model is very good at heterogeneity, but it has an inevitable problem: the determination of the mixed number. Many scholars have discussed the determination of the mixtures of Gaussian models, which includes the minimum distance method, the hypothesis test method, the penalized likelihood method and so on. Fan & Li[10] proposed SCAD penalty on linear regression model to realize variable selection and regression coefficient determination. In Yin & Lin's paper[9], similarly to SCAD penalty function, the author proposed a new threshold penalty function, iSCAD penalty to select the order of the model, and the estimators also satisfy sparsity, continuity and unbiased properties. But in practice, the penalty algorithm converge slowly. Based on the above, this paper proposed another penalized likelihood function, the penalty function here is inspired by the traditional roughness penalty. It's three squared derivative of the continuous density function. The roughness penalty function is defined as follows:

$$PEN_3 = \int_0^\infty [f'''(x|\alpha, m, \theta)]^2 dx = \int_0^\infty ([\sum_{k=1}^{K} \alpha_k \frac{x^{m_k-1}e^{-\frac{x}{\theta}}}{\theta^{m_k}(m_k-1)!}]''')^2 dx.$$

We illustrate the performance of the proposed method by some simulations and real data studies in chapter 4, and the results reveal a better performance compared with non-penalized method generally.

**Key words:** Mixed Erlang; EM algorithm; Penalized Likelihood Estimations.

# 目 录

# Contents

# 第一章 引言

本章首先概述本学位论文所研究问题的相关背景及国内外研究现状，然后简单介绍本文拟解决的问题、处理方法及论文的结构安排。

## 1.1 相关文献综述

由给定样本点集合求解随机变量的概率密度函数问题长期以来都是统计学的基本问题之一。通常来说，解决这一问题的方法主要有以下两类：参数估计和非参数估计。如果我们通过对观测样本进行大致考察评估后而假设数据来自于一个已知分布，这种方法属于参数估计。在参数估计中，人们预先假定给定的数据分布一定符合某种特定分布的性质。如此一来，给定数据的密度估计问题就转化为在直观的目标函数族中寻找特定解的问题，即确定模型中有限个的未知参数。但是经验和理论说明，这种基本假定与实际的模型之间常常存在较大的差距，这种方法并非总能取得令人满意的结果。由于参数估计的上述缺陷，一些非参数的密度估计方法在实际中被提出，并得到广泛应用。例如核密度估计方法。由于核密度估计方法不利用有关数据分布的先验知识，对数据分布不附加任何假定，是一种从数据样本本身出发研究数据分布特征的方法，因而，在统计学理论和应用领域均受到高度的重视。其余非参数的方法还包括K 近邻法（K-nearest-neighbour）、K-means、copula等等。当需要拟合一组数据的密度函数时，采用非参数估计算法是一种对于经典统计技巧的具有吸引力的改动，并且较为容易理解。但是对于一个显示的统计问题，非参数的方法由于没有给出拟合结构的解析表达，导致很难得到最优解，最终模型对于数据的解释能力不足。

在 Ramsay & Silverman[1]一书中，作者提出了一种用基函数线性组合完成对于函数型数据密度估计的解决方案。函数型数据是指随着某一连续体（时间、空间

等）变化的数据，可以是曲线、平面或三维图像等。但就其本质而言，是给定一列 $y_{ij}, \quad i = 1...N, \quad j = 1...n, \quad y_{i1}...y_{in}$ 是光滑函数 $x_i$ 在自变量 $t$ 的 n 个值 $t_{i1}...t_{in}$ 上的记录（观测）。这种用基函数估计密度的方法好处是拟合函数具有解析表达式，且拟合函数的形式为基函数的线性组合，结构简单。此外基函数的备择空间广泛，可以根据函数型数据的观测值性质来选择基函数的形式。例如利用傅里叶基的线性组合来拟合具有周期性的数据，用 B样条基函数或小波函数基函数来拟合一般性无特殊限制的函数。此外，密度函数在节点的光滑度可以得到很好控制。

除以上所提及的密度估计方法之外，作为一个非常灵活而强有力的概率建模工具，有限混合分布模型也已在理论和实践中得到了极为广泛的应用，尤其是在多峰分布领域，有限混合分布的灵活性得到很好的应用。究其原因，在于以下优势：

有限混合提供了用简单结构模拟复杂分布的一个有效方法。我们知道，正态分布在实践中应用最广，并且它有理论上的支持，其形式也比较简单。已经证明，利用混合正态分布这样的简单结构，经由混合而成的混合分布可以逼近任一个光滑分布，只要项数 K足够大。因此有限混合分布模型可以用于描述复杂现象。

有限混合模型提供了模拟同质性和异质性的一个自然框架。当 m=1时，混合模型其实是一个单一分布。因而数据具有相同的性质；当 m>1时，混合模型就反映了混合数据的异质性，它综合了参数模型的解析优势和非参数模型的灵活性，因而具有更多的建模优势。

除了传统的混合正态模型，近年来混合 Erlang模型也备受关注。继 Tijms[2] 给出了任意正的分布都可由一组混合 Erlang 弱依分布收敛的证明后，Lee & Lin[3,4]给出了混合 Erlang 模型在删失数据密度估计方面的应用。Verbelen等[5,6] 就给出了利用混合 Erlang 模型对于删失数据和截断数据的拟合，表现都要优于 K-M 估计（Kaplan-Meier Estimate）。随后，Gui等[7] 给出 GEM-CMM 算法在拟合删失数据和截断数据的密度估计。Yin & Lin[9] 在 Fan & Li[10] 的基础上构造了一种新的 iSCAD 惩罚函数来实现变量选择和混合模型权重参数的估计，应用于保险数据的密度估计

上。

到目前为止，解决上面问题的方法仅局限在参数估计及计算方面，也就是说各分支的分布形式已知，只是含有未知的参数，关于参数的估计归纳起来有以下方法和算法：

矩估计法：Pearson最早考虑了混合模型的估计问题，他用矩估计法估计两个正态分支的参数，经过相当大的代数运算后，矩方程的求解最后简化为先求一个9阶多项式方程的负实根。由负实根可给出所有参数的矩估计。后来许多学者用矩方法研究了其他情况，但是这种方法的计算量太大，使得该方法的进一步发展受到了很大局限。

极大似然估计法：20世纪60年代后期，随着计算机的广泛应用和数值方法的不断发展。人们开始把目光转向具有更多优势的极大似然估计法。极大似然估计是一种常用的参数估计的方法，它是以观测值出现的概率最大作为准则。由于极大似然估计的渐进最优性质，它已经成为参数估计的一种常用方法，并且已经在众多领域中广泛得到应用。但是，对于似然函数方程的求解却没有一般的理论方法，似然函数最大化往往要求似然函数有较好的性质，但在大多数情况下很难满足这种要求。正是在这种情况下，针对传统的极大似然参数估计方法解决实际问题的不足，1977年，Dempster[11] 发明了 EM 算法，EM算法主要用于非完全数据参数估计，通过假设隐变量的存在，极大地简化了似然函数方程，从而解决了方程求解问题。对于一些特殊的参数估计问题，利用 EM算法也可以较容易地实现。

EM算法相对于其他方法而言，其时间复杂度要低许多，这是它应用较广的主要原因。但是EM算法得到的只是最大似然意义上的局部最优解，需要考虑初始值的选取方法来克服这个问题。一般来说有三种解决方案，一种方法是预先确定一个较大范围的初始值备择空间，然后选取使得似然函数达到最大的初始值估计，这种方法计算速度慢且计算量过大，删除权重参数最小的混合分布过程中，一次只删除一个，删除后需要重新进行 EM 算法，这可能造成繁重的计算，运行时间也会相应增加，

过度拟合的风险大。第二种方法是近来提出的分割与合并的EM算法（SMEM算法），一般采用最大惩罚似然和贝叶斯方法，不过这种方法在高维情况下的执行效果并不理想。第三种方法对数据预先进行聚类，获得迭代初始值。本文采用的是第三种初值选取的方法，即CMM初值选取[7]的方法，先对于数据进行聚类，然后采用多重矩估计拟合的方法得到EM算法的迭代初值。

回归分析中，尤其是稀疏的广义线性模型中，通常采用惩罚似然函数来实现变量选择。基于似然惩罚的变量选择方法是一种应用广泛的统计方法，其中常用的凸惩罚如岭回归，lasso，以及 Fan & Li[10]所提出的惩罚回归系数的 smoothly clipped absolutely deviation （SCAD）函数，速度快，解连续，不过多数的估计是有偏的。而非凸惩罚在凹度过大时，由于多个局部最优值的出现而导致相对稳定性不够，且不容易找到好的算法进行局部最优质的选择，或者最终算法不收敛，或者收敛到不是严格局部最优的稳定点、或者多个局部最优值不能渐进到一点，从而相对稳定性较差。在 Ramsay & Silverman[1]一书所提到的对于样条基函数粗糙惩罚的方法，很好地保留了混合模型的灵活性的优势，并且这种惩罚方法的理解直观简单，适用范围更广，估计效果更优。

基于以上工作成果，本文给出了在粗糙惩罚下，最大惩罚似然的 Erlang混合模型拟合概率密度函数的方法。

在本文中，为了优化拟合结果，我们加入一个粗糙惩罚(roughness penalty)工具，并通过广义交叉验证(genaral cross-validation)选出最优的平滑参数。我们采用粗糙惩罚混合 Erlang分布来估计密度函数，用 EM算法来估计参数，并将我们的估计结果与没有考虑粗糙惩罚的 GEM-CMM[7]算法加以比较。

对于惩罚性系数的选取，Ahn[17]提出了一个得到最优平滑参数 $\lambda$ 的方法：使广义交叉验证(generalized cross-validation) 最小化的平滑参数就是最优的平滑参数。交叉验证（Cross Validation）简称CV。CV是用来验证数据模型的性能的一种统计分析方法，基本思想是把在某种意义下将原始数据（data set）进行分组，一部分作为训

练集（training set），另一部分作为验证集（validation set），首先用训练集对数据模型进行训练，再利用验证集来测试训练得到的模型，以此来作为评价数据模型的性能指标，常用的 CV 方法为 Leave-One-Out Cross Validation （记为 LOO-CV）：如果设原始数据有N个样本，那么 LOO-CV就是 N-CV，即每个样本单独作为验证集，其余的 N-1个样本作为训练集，所以 LOO-CV会得到 N个模型，用这 N个模型最终的验证集的分类准确率的平均数作为此 LOO-CV分类器的性能指标。利用交叉验证方法选择模型思路是：使用训练集（training set）数据所有候选模型进行参数估计，使用验证集（validation set）为检验样本，然后计算预测均方误差，比较各个模型的预测均方误差，选择预测均方误差最小的拟合模型为选择模型。在本文中，通过对混合模型建立粗糙惩罚，我们将此项作为惩罚系数。并在后续的数据检验中，证实了此项的确起到调节惩罚程度的作用。

## 1.2 本文结构安排

本文内容主要分为四章

第一章概述本文所研究问题的相关背景以及国内外现状，并简要介绍了本文拟解决的问题、处理方法和论文的结构安排。

第二章介绍混合 Erlang分布，并给出粗糙惩罚的定义，介绍其相关性质。

第三章给出粗糙惩罚 Erlang混合模型的混合权重和共用尺度参数的估计。并在完全数据的参数估计基础上，给出基于删失数据与截断数据的粗糙惩罚GEM参数估计。

第四章通过数值模拟试验和实际数据，比较两种方法（GEM-CMM 的方法1，以及本文引入的方法2）在估计结果方面的表现。