学校编码：10384　　分类号___密级____
学号：19020141152615　　UDC____

厦门大学

硕 士 学 位 论 文

混合厄朗模型的变分贝叶斯学习

Variational Bayesian Learning of Mixed Erlang Model

林 琴

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
论文答辩时间：2017 年 5 月
学位授予日期：

答辩委员会主席：
评阅人：

2017 年 06 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

摘 要

在本文中，我们关注混合厄朗模型（Mixed Erlang Model, MER模型）的贝叶斯变分方法应用。混合厄朗模型（MER）的多变量混合形式构成了一种通用且易于分析的分布，除此之外，该分布在在弱收敛意义上的正连续多元分布的空间中是稠密的。这些良好的性质使得混合厄朗分布适用于多变量密度估计。因此，我们为混合厄朗模型提出了一种灵活有效的拟合算法，称为CMM-VBEM算法。该算法分为两个部分，第一部分是初始值的选取CMM算法。利用K-Means聚类的方法对参数的初始值进行估计，事实证明该方法大大的提高了初始值选取的有效性；此外，应用贝叶斯信息准则（BIC）选择模型的混合个数；第二部分是VBEM算法，该过程以贝叶斯变分法为基础，同时迭代地使用EM算法，为形状参数、混合权重等参数引入计算有效的估计调整策略。

与传统的 EM 算法等混合模型常用的方法相比，我们的方法有几个优点：首先，防止过度拟合的问题，特别是当数据量较少的时候，我们的方法优越性就更明显；此外，CMM-VBEM 算法成功的避免了局部最优问题。本文，通过模拟数据和实际的数据两个方面，对提出的 CMM-VBEM 算法进行验证。通过图形验证（包括经验直方图、QQ 图、PP 图、等高线图）以及多种检验方法（包括 Kolmogorov-Smirnov 检验，Anderson-Darling 检验和 Cramer-von Mises 检验），从多方面进行验证，充分说明我们的算法在数据拟合效果良好。

关键词：混合厄朗分布；变分法；BIC 检验

ABSTRACT

In this paper, we focus on the application of Bayesian variational methods for mixed Erlang Model. Mixed Erlang Model forms a versatile, yet analytically tractable, class of distributions making them suitable for multivariate density estimation. In addition, the distribution in the weak convergence of the continuous distribution of the space is dense. The above good properties make the multicomponent Erlang model suitable for multivariate density estimation. On this basis, we propose a flexible and effective fitting process for mixed Erlang model, called CMM-VBEM algorithm. The algorithm is divided into two parts. The first part is the selection of initial value. Using the K-Means clustering method to estimate the initial value of the parameter, it is proved that the method greatly improves the validity of the initial value selection. In addition, the Bayesian Information Criterion (BIC) is used to select the number of mixing. The second is the VBEM algorithm. The process is based on the Bayesian variational method and the iterative use of the EM algorithm to introduce the effective estimation and adjustment strategy for the parameters such as shape parameters and mixed weights.

Compared with the traditional EM algorithm and other mixed models commonly used methods, our method has several advantages. First, it can prevent over-fitting problems, especially when the amount of data is less time, our method is more obvious advantages; In addition, the CMM-VBEM algorithm succeeds in avoiding local optimal problems. In this paper, the proposed CMM-VBEM algorithm is validated by simulating the data and the actual data. A variety of test methods (including Kolmogorov-Smirnov test, Anderson-Darling test and Cramer-von Mises test) were validated by graphic verification (including empirical histogram, QQ plot, PP plot, contour map), and verified from various aspects, Full description of our algorithm in the data fitting effect is good. The effectiveness of the proposed algorithm is demonstrated on simulated as well as real data sets. These all proved that our algorithm perform well in data fitting.

Key words: Mixed Erlang Model； Variational Bayesian Learning；BIC

目录

Contents

第一章 背景介绍

近年来，科学界一直强调使用混合模型来分析复杂的现象。混合分布假设最早由 Clark 提出，经过数学科学家们如:Epps、Harris 和 Tauchen 不断地完善,理论逐渐趋于成熟。但混合分布的参数估计仍然是一个较为困难的问题，潜变量（Latent Variable）概念的引入就是为了更方便快捷地解决混合模型的参数估计问题。如果定义关于观测变量和潜在变量的联合分布，观测变量单独的相应分布通过边缘化获得。这允许相对复杂的边缘分布与观察到的变量在被观察和潜在变量的扩展空间上的更易处理。潜在变量的引入使得复杂的分布分解为简单分布的组合。常见的混合分布包括：混合高斯分布，混合贝塔分布，混合厄朗分布,有限 Dirichlet 混合模型（Finite Dirichlet Mixture Models）等。

本文引用的混合厄朗模型（Mixed Erlang Model, 称为MER）。混合厄朗模型的提出，是为了更有效地处理非负随机变量问题。混合厄朗模型被定义为具有不同参数的厄朗分布的凸组合，并且被应用于在许多实际情况中出现的数据拟合问题。多变量混合厄朗模型（Multivariate Mixed Erlang，称为MME）的概念最初在[1]中提出。MME模型享有Joe所列的多变量模型的许多理想特性，参见文章[1]。MME形成高度灵活的分布类，因为它们在弱收敛意义上的正连续多元分布的空间中密集，[2]就单变量类（k=1）扩展了这个属性。混合厄朗分布的分析和分布性质的概述可以在文章[3,4,5]中找到。[6]提出了单变量（k=1）情况下的参数估计方法，[7]则扩展到能够处理随机截尾和固定截断数据。

混合厄朗分布在精算科学领域受到最多的关注。[8]使用 Erlangs 作为边缘的单变量混合模型与 Farlie-Gumbel-Morgenstern（FGM）Copula 模型联合分布依赖风险组合。[9]研究了二元变量的下限和上限价值风险，并使用 MME 作为例证。[10]研究了 MME 类的分析属性。MME 的使用应被认为是多变量密度估计技术，而不是作为一种基于模型的聚类。MME 模型可以看作是半参数的，因为混合分量具有特定的参数形式，而混合权重可以具有非参数性质，并且是使用 copulas 的有趣替代方案，其是用于对多变量数据进行建模的主要选择两阶段过程，将依赖性结构与边缘分布分离（参见 [11,12]）。相比之下，MME 能够直接在原始规模上对多元数据建模[13]。

关于 MER 模型的参数估计，在最大似然（ML）估计方面，EM（期望最大化）算法在计算工作方面是一种有效的估计算法。在实践中，在网络流量分析的分析中应用了 MER 分布的 EM 算法。

传统的参数估计大多运用ML算法，但对于混合厄朗模型，当观测数据量较少时，容易出现过拟合（overfitting）的问题。而贝叶斯估计的最严重的缺点是从实际的角度来看，是计算成本。由于MER模型是厄朗分布的凸组合，即使我们应用共轭先验，后验分布也不是由任何封闭形式给出的。此外，由于MER模型包括许多参数，高维积分的数值方法应用于计算后验分布。马尔可夫链蒙特卡洛（MCMC）在普遍的贝叶斯估计框架中非常受欢迎。MCMC是通过使用基于蒙特卡洛积分原理的模拟样本来计算后验分布的通用方法，因此对于许多参数的后验分布是有效的。然而，MCMC通常需要长期的模拟过程才能获得准确的估计[14]。

常用的参数估计法同时也包含EM算法，Dempster最早介绍了EM 算法的详细说明，当似然函数最大值点不易计算时，可用迭代算法得到最大值点。王继霞等人将EM 算法用于有限混合Laplace分布的估计。Panchenko和Thümmler; Thümmler等人在网络流量分析的分析中应用了MER分布的EM算法。在EM算法中，我们需要评估关于潜在变量的后验分布的完全数据对数似然性的期望。针对 EM 算法，主要有两个缺陷。一为迭代部分的估计对初值的选取较为敏感，初值的选取将影响算法的收敛速度。[6]和[7]所用的Tijms近似法并不理想。二是容易产生过拟合的问题，特别是当数据量较少的时候。为此我们找到了相应的解决方法变分贝叶斯方法（Variational Bayesian）。

变分贝叶斯方法最早由 Matthew J.Beal 在他的博士论文中提出并在机器学习方面应用广泛。例如 Waterhouse 等人提出通过将变分近似技术合并到贝叶斯推理中来避免过拟合。VB 方法可以比拉普拉斯近似更准确，因为它不假定后验的高斯分布。此外，由于有确定性的算法，它比马尔科夫链蒙特卡罗（MCMC）方法更有效，计算成本更低。近年来，统计学在亦开始广泛应用 VB 方法，例如混合模型,隐藏马尔可夫链建模，隐马尔可夫随机场建模的空间数据分析,广义线性混合模型，有限混合 Dirichlet 模型的变分学习[15]。在基因工程方面亦有不俗的成绩，例如功能磁共振图像数据的分析，遗传学建模和人类移动模式建模 [16]。

由于 VB 是后验分布的分析近似方法，它可以降低计算成本。VB 背后的基本思想是将 KL 散度（Kullback-Leibler divergence）从近似后验分布到后验分布作为变分问题最小化。[16]提出了 MER 模型的 VB-EM 算法。随后[13]又对 VB-EM 算法进行了完善，同时提出了估计形状参数的算法，详细地提供了 VB 公式的推导及其对 MER 分布的原理[14]。然而，Yamaguchi 等人只对单变量（k=1）的混合厄朗模型进行变分贝叶斯估计方法的研究，不够全面；其次，他们对形状参数的调整不够理想，利用相位法（phase-type, PH）调整得到的形状参数值都太小，并不适合将此方法应用到多变量（k>1 的情况）的混合厄朗分布。

本文总共分为七个部分。第二部分是介绍混合厄朗模型的性质；第三部分是参数估计的介绍，主要介绍了最大似然估计，贝叶斯估计和变分贝叶斯估计；第四部分是混合厄朗模型在变分贝叶斯方法上的应用，主要从先验分布、变分法应用和后验分布三个部分进行详细介绍。第五部分是 CMM-VBEM 算法，包括初始值的选取、形状参数的调整和混合个数的选择，VBEM 算法。第六部分是模拟数据与示例，说明算法拟合的优良性。最后部分是结论。

第二章 混合厄朗模型

在本节中，我们回想一下具有共同速率参数的混合厄朗分布（Mixed Erlang Model, MER 分布）的定义和一些重要性质。

[1]定义了一个k变量混合厄朗分布，使得每个混合成分是k个具有共同速率参数（Rate Parameter）$\beta>0$的独立的厄朗分布的联合分布。相互独立是通过不同的厄朗分布在每个维度中的正整数形状参数的组合来获得。我们用矩阵$m_u=(m_{u1},\dots,m_{uk}),u=1,2,\dots,d$表示混合厄朗分布中共同独立的正整数形状参数，记混合权重参数为$\alpha=(\alpha_1,\alpha_2,\dots,\alpha_d)$，参数族$\{\alpha_u\}$满足条件$0\le\alpha_u\le1.$

$u=(1,2,\dots d)$，$\sum_{u=1}^{d}\alpha_u=1.$则参数$\beta$的k元混合厄朗分布的概率密度函数为：

$$f(x\mid m,\beta)=\sum_{u=1}^{d}\alpha_u\prod_{j=1}^{k}f(x_j\mid m_{uj},\beta)=\sum_{u=1}^{d}\alpha_u\prod_{j=1}^{k}\frac{\beta^{m_{uj}}x_u^{m_{uj}-1}e^{-\beta x_u}}{(m_{uj}-1)!}. \quad (2.1)$$

其中，$m_u=(m_{u1},\dots,m_{uk}),u=1,2,\dots,d$，

$X=\{x_1,\dots,x_k\},x_j>0,j=1,2,\dots,k$，

$\Phi=\{\alpha_u,m_{uj},\beta;u=1,2,\dots,d,j=1,2,\dots,k\}.$

为书写方便，记

$$f(x\mid m_u,\beta)=\prod_{j=1}^{k}\frac{\beta^{m_{uj}}x_u^{m_{uj}-1}e^{-\beta x_u}}{(m_{uj}-1)!}, \quad (2.2)$$

其中，$m_u=(m_{u1},\dots,m_{uk}),u=1,2,\dots,d$。

则有

$$f(x\mid\Phi)=\sum_{u=1}^{d}\alpha_u f(x\mid m_u,\beta) \quad (2.3)$$

观察上式可发现，当k = 1时，混合厄朗分布实际上是伽马分布（Gamma Distribution），其形状参数是正整数，因此可以看作是独立同分布的指数随机变量。k = 1的混合厄朗分布的概率密度函数如下：

$$f(x\mid m,\beta)=\sum_{u=1}^{d}\alpha_u\frac{\beta^{m_u}x^{m-1}e^{-\beta x}}{(m_u-1)!}. \quad (2.4)$$

其中，$\alpha=(\alpha_1,\alpha_2,...,\alpha_d)$为第$u$个混合厄朗分布的非负权重，$\beta>0$为共同速率参数（Rate Parameter）。

Tijms 表明，这种分布类在正连续密度空间中是密集的。对于任何给定的正分布函数*F(x)*，令

$$\tilde{f}(x\mid\beta)=\sum_{i=1}^{\infty}(F(i/\beta)-F((i-1)/\beta))p_i(x) \tag{2.5}$$

其中，

$$p_i(x)=\frac{\beta^i x^{i-1}e^{-\beta x}}{(i-1)!}. \tag{2.6}$$

那么，如 Tijms 所示，对于所有连续点x，$\lim_{\beta\to\infty}\tilde{F}(x\mid\beta)=F(x)$。其中$\tilde{F}(x\mid\beta)$是$\tilde{f}(x\mid\beta)$的分布函数，因此，式（2.1）中的密度函数可以趋近于任何正的连续分布。理论上，可以通过增加k和β来提高精度[11]。

下面的性质指出，对于任何正的多变量分布，存在弱收敛到目标分布的多变量厄朗分布序列。[1]的附录中给出了详细证明。

性质 2.1：Lee 和 Lin（2012）.式（2.1）的混合厄朗混合分布的类在弱收敛意义上的正连续多变量分布的空间是密集的。换句话说，令f(X)是 k 变量正随机变量的密度函数，F(X)为累积分布函数。对于任何给定的$\beta>0$，定义以下 k 变量混合厄朗分布：

$$f(x\mid\Phi)=\sum_{u_1}^{\infty}\cdots\sum_{u_k=1}^{\infty}\alpha_u\prod_{j=1}^{k}f(x\mid m_{uj},\beta) \tag{2.7}$$

其中，混合权重参数满足：

$$\alpha_u(\beta)=\int_{(u_1-1)/\beta}^{u_1/\beta}\cdots\int_{(u_k-1)/\beta}^{u_k/\beta}f(x)dx \tag{2.8}$$

则在 F 连续的每个 x 上，有$\lim_{\beta\to\infty}F(x,\beta)=F(x).$

在性质 2.1 中，对于任何给定的共同速率参数（Rate Parameter）$\beta>0$，在式（2.7）中的无限 MME（Multivariate Mixtures of Erlang Distributions）被认为

是在每个边缘维度中使用从1到无穷大的形状参数m的组合。通过将形状参数m乘以公共尺度参数形成的k维网格的相应的k维矩形上的密度进行积分来限定混合分布中式（2.8）中的权重。当共同速率参数（Rate Parameter）$\beta > 0$的值减小时，该网格变得更精细，并且混合厄朗分布的序列收敛到潜在的累积分布函数。

另外，考虑到灵活性的因素，[1]已证明出由于厄朗分布在每个混合组分内的独立结构，使得使用这类分布进行分析工作更为简便。同时，独立性使得许多分布量的能够有明确的表达式，例如特征函数，联合距和关联的双变量度量（Kendall's tau and Spearman's rho）。除此之外，作者进一步揭示了闭合属性，例如每个p变量边际或$p \le k$的条件分布与可以再次写成一个k变量混合厄朗分布。同样的属性适用于多元超额损失（精算科学背景）或多元剩余寿命（生存分析环境）的分配。此外，MME 分布随机变量的分量随机变量之和的分布是单变量厄朗混合分布。

文献[10]考虑了 MME 类的扩展，允许每个维度中的不同尺度参数。然而，在性质 2.1 中，他们展示了具有不同比例参数的一个 MME 分布如何可以被重写为具有比所有原始尺度小的公共比例参数的 MME 分布。因此，我们专注于具有共同速率参数β的模型[11]。