混合因子分析的重新抽样方法

岳 博,焦李成
(西安电子科技大学雷达信号处理重点实验室,陕西西安 710071)

摘 要: 混合因子分析是一种对具有复杂结构的多维数据建立模型的方法.本文提出了一种进行混合因子分析的重新抽样方法.当给定一组数据样本时,我们首先建立样本概率分布的混合高斯模型,然后为每一个高斯混合项重新抽取新的数据样本,在新的样本上再对每一个高斯混合项进行因子分析.与已有的算法相比较,避免了计算各个高斯混合项在每个样本值之下的后验概率,又减少了进行因子分析时参与计算的数据样本的数量.

关键词: 因子分析; 混合高斯模型; EM 算法; 抽样
中图分类号: O212.4 **文献标识码:** A **文章编号:** 0372-2112 (2002) 12-1873-03

The Resampling Method for Mixtures of Factor Analyzers

YUE Bo, JIAO Li-cheng
(*Key Laboratory for Radar Signal Processing, Xidian University, Xi'an, Shaanxi* 710071, *China*)

Abstract: The mixtures of factor analyzers are able to model complex data structures through a combination of the factor analysis model and the Gaussian mixture model. In this paper, a resampling method for the mixtures of factor analyzers is proposed. After approximating the probability distribution density of the data by the Gaussian mixture model, we draw new samples for each of the component Gaussians with its own parameters separately, then on the new samples the factor analysis is performed for each component Gaussians. We also implement this method with the EM algorithm and the good performance of the method is illustrated by an example.

Key words: factor analysis; Gaussian mixture model; EM algorithm; sampling

1 引言

因子分析(factor analysis, FA)[1]是一种对多维数据之间的相关性进行建模的方法,它可以实现数据的降维,因此在机器学习和模式识别中得到了广泛地应用.然而由于因子分析方法仅仅利用了随机样本的一阶矩(均值)和二阶矩(协方差),所以事实上我们总是假设数据样本是服从高斯分布的,并且其生成模型中的隐变量也是服从高斯分布的.为了能够处理非高斯分布的数据样本,文[4]中提出了一种混合因子分析模型(mixtures of factor analyzers),其本质上是首先建立一个混合高斯模型(Gaussian mixture model),然后再对每一个高斯混合项进行因子分析.文[5]给出了使用 EM 算法[3]的混合因子分析的实现.

在混合因子分析模型的学习算法中,当我们完成了混合高斯模型的参数学习之后,对每一个高斯混合项进行因子分析时,每一个原有的样本点只是以一定的概率部分地属于各个高斯混合项,因此每一个高斯混合项实际的等价样本数是较少的,这样如果我们为这些高斯混合项重新抽取较少数目的数据样本,在新的数据样本上进行因子分析将大大减少参加运算的数据量,这样就有了本文提出的重新抽样方法,即在混合高斯模型的学习完成之后,我们为每一个高斯混合项重新抽取新的数据样本,然后再对每个高斯混合项在各自的样本之上进行因子分析.

收稿日期:2001-04-16;修回日期:2002-07-01
基金项目:国家自然科学基金(No.60073053)

2 混合因子分析模型

2.1 因子分析

在因子分析模型中,我们假设 p 维实值观察数据向量 $\boldsymbol{x}$ 是 $k(k<p)$ 维的实值因子向量 $\boldsymbol{z}$ 的线性叠加,即

$$\boldsymbol{x}=\boldsymbol{\Lambda}\boldsymbol{z}+\boldsymbol{u} \tag{1}$$

上式中的 $\boldsymbol{\Lambda}$ 称为因子载荷矩阵(factor loading matrix),$\boldsymbol{u}$ 是随机噪声.通常我们假设公共因子 $\boldsymbol{z}$ 的各个分量之间是不相关的,并且每个分量都服从 0 均值单位方差的高斯分布,即 $\boldsymbol{z}\sim N(0,\boldsymbol{I})$,其中的 $\boldsymbol{I}$ 是单位矩阵.同时假设噪声 $\boldsymbol{u}$ 与 $\boldsymbol{z}$ 不相关,服从高斯分布 $N(0,\boldsymbol{\Psi})$,并且 $\boldsymbol{\Psi}$ 是对角矩阵,这表明在给定因子 $\boldsymbol{z}$ 的值之后,观察数据 $\boldsymbol{x}$ 的各个分量之间是相互独立的.

由模型(1),可得到 $\boldsymbol{x}$ 的概率密度函数为

$$p(\boldsymbol{x})=|2\pi(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}+\boldsymbol{\Psi})|^{-1/2}\exp\{-\boldsymbol{x}^{\mathrm{T}}(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}+\boldsymbol{\Psi})^{-1}\boldsymbol{x}/2\} \tag{2}$$

即 $\boldsymbol{x}$ 服从高斯分布 $N(0,\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}+\boldsymbol{\Psi})$.由 Bayes 定理,已知数据 $\boldsymbol{x}$ 时,因子 $\boldsymbol{z}$ 的后验分布也是高斯分布

$$p(z|x)=\frac{p(x|z)p(x)}{p(z)}=|2\pi M|^{-1/2}\exp\{-(z-M\Lambda^{T}\Psi^{-1}x)^{T}M^{-1}(z-M\Lambda^{T}\Psi^{-1}x)/2\} \quad (3)$$

其中 $M=(\Lambda^{T}\Psi^{-1}\Lambda+I)^{-1}$.

学习因子分析模型就是要寻找模型参数 $\{\Lambda,\Psi\}$,使之能够最佳地重构样本数据 $\{x^n, n=1\cdots N\}$ 的协方差矩阵 cov($\{x^n\}$). 如果使用最大似然方法,就是寻找使得数据样本的似然函数 $L=\log p(x^1,\cdots,x^N)=\sum_{n=1}^{N}\log p(x^n)$ 最大的 $\{\Lambda,\Psi\}$. 当将因子 z 看作是隐变量时,就可以使用 EM 算法进行学习[6].

算法 1 使用 EM 算法的因子分析学习算法

(1)初始化 Λ 和 Ψ;

(2)E 步:对每一个数据样本 x^n, $n=1\cdots N$,,计算 $E(z|x^n)$ 和 $E(zz^{T}|x^n)$;

(3)M 步:更新 $\{\Lambda,\Psi\}$

$$\Lambda=\left(\sum_{n=1}^{N}x^{n}E(z|x^{n})^{T}\right)\left(\sum_{n=1}^{N}E(zz^{T}|x^{n})\right)^{-1}$$

$$\Psi=\frac{1}{N}\mathrm{diag}\left(\sum_{n=1}^{N}x^{n}(x^{n})^{T}-\Lambda E(z|x^{n})(x^{n})^{T}\right)$$

(4)若满足条件,停止;否则,转第(2)步.

其中 diag(·)表示将一个矩阵的非对角线元素置 0. 由式(3),有后验期望 $E(z|x)=(\Lambda^{T}\Psi^{-1}\Lambda+I)^{-1}\Lambda^{T}\Psi^{-1}x$,后验协方差 $\mathrm{cov}(z|x)=(\Lambda^{T}\Psi^{-1}\Lambda+I)^{-1}$,而 $E(zz^{T}|x)=\mathrm{cov}(z|x)+E(z|x)E(z|x)^{T}$.

2.2 混合高斯模型

在这里,以生成模型(隐变量模型)的形式来描述混合高斯模型. 对一个 p 维实空间中的随机变量 $x\in R^{p}$,如果使用混合高斯模型来表示其概率密度,那么需要首先引入一个隐变量 q,它是一个离散随机变量,其值域为 $\{1,\cdots,m\}$,并且有分布 $P(q=i)=\pi_i$. 同时,在给定 q 的取值时,x 服从均值为 μ_i,协方差矩阵为 Σ_i 的高斯分布 $N(x|\mu_i,\Sigma_i)$,

$$p(x|q=i)=N(x|\mu_i,\Sigma_i) \quad (4)$$

这样 x 的分布为

$$p(x)=\sum_{i=1}^{m}P(q=i)p(x|q=i)=\sum_{i=1}^{m}\pi_i N(x|\mu_i,\Sigma_i) \quad (5)$$

图 1 给出了混合高斯模型的图模型表示形式. 因为混合高斯模型中的每一个混合项都是一个高斯概率密度函数,这在分析时是非常方便的,所以混合高斯模型得到了广泛的应用,例如在径向基函数网络中等.

图 1 混合高斯模型的生成模型

由式(5),已知 x 时隐变量 q 的后验概率为

$$P(q=i|x)=\frac{P(q=i)N(x|\mu_i,\Sigma_i)}{\sum_{i=1}^{m}P(q=i)N(x|\mu_i,\Sigma_i)}=\frac{\pi_i N(x|\mu_i,\Sigma_i)}{\sum_{i=1}^{m}\pi_i N(x|\mu_i,\Sigma_i)} \quad (6)$$

如果给定一组 $i.i.d.$ 的随机样本 $\{x^n, n=1\cdots N\}$,要使用混合高斯模型来近似其概率分布,那么我们只需要估计模型(5)中的参数 $\theta=\{\pi_i,\mu_i,\Sigma_i\}_{i=1}^{m}$. 由式(5),样本的似然函数为

$$l(\theta)=\log\left[\prod_{n=1}^{N}p(x^n)\right]=\sum_{n=1}^{N}\log\sum_{i=1}^{m}\pi_i N(x^n|\mu_i,\Sigma_i) \quad (7)$$

同样,可使用 EM 算法学习得到参数 θ 的最大似然估计[2].

算法 2 混合高斯模型参数学习的 EM 算法

(1)初始化 π_i,μ_i 和 Σ_i;

(2)E 步:对每一个样本 x^n,按照式(6)计算隐变量 q 的后验分布,并且记 $R_{ni}=P(q=i|x^n)$;

(3)M 步:更新 π_i,μ_i 和 Σ_i

$$\pi_i=\frac{1}{N}\sum_{n=1}^{N}R_{ni};\qquad \mu_i=\frac{\sum_{n=1}^{N}R_{ni}x^{n}}{\sum_{n=1}^{N}R_{ni}};$$

$$\Sigma_i=\frac{\sum_{n=1}^{N}R_{ni}(x^{n}-\mu_i)(x^{n}-\mu_i)^{T}}{\sum_{n=1}^{N}R_{ni}}$$

(4)若满足条件,停止;否则,转第(2)步.

2.3 混合因子分析模型

混合高斯模型与因子分析相结合得到的混合因子分析模型,给我们提供了一种对复杂结构的数据建立模型的有效手段. 它利用了数据不同分量之间的相关性对数据进行建模,减少了表示这组数据时所需的维数. 混合因子分析模型的参数学习算法可以看作是首先进行混合高斯模型的学习,然后再分别对每一个高斯混合项进行因子分析,其过程可描述如下

算法 3 混合因子分析模型的参数学习算法

(1)建立数据样本概率分布的混合高斯模型;

(2)对此模型中的每一个高斯混合项进行因子分析.

当给定一组数据样本时,利用 EM 算法,只需要使用样本的似然函数,就能够完全确定模型的所有参数. 在使用 EM 算法的实现中,上述算法第二步中的因子分析与通常的因子分析是不同的,在这里,对每一个要进行因子分析的高斯混合项,它并没有一个完全属于自己的样本,也就是说,每一个样本点 x^n 只是以一定的概率 R_{ni} 属于第 i 个高斯混合项. 因此在实现因子分析参数学习的 EM 算法的 M 步有

$$\Lambda_i=\left(\sum_{n=1}^{N}R_{ni}x^{n}E(z_i|x^{n})^{T}\right)\left(\sum_{n=1}^{N}R_{ni}E(z_i z_i^{T}|x^{n})\right)^{-1}$$

$$\Psi_i=\frac{1}{\pi_i N}\mathrm{diag}\left(\sum_{n=1}^{N}R_{ni}x^{n}(x^{n})^{T}-R_{ni}\Lambda_i E(z_i|x^{n})(x^{n})^{T}\right) \quad (8)$$

3 混合因子分析的重新抽样方法

在使用混合高斯模型学习得到数据样本的概率密度函数之后,得到了一些不同的高斯混合项,它们有各自的权系数,均值和协方差矩阵,我们需要对每一个高斯混合项进行因子分析. 由式(6)和算法 2 看到,每一个样本点 x^n 是以概率 R_{ni} 部分地属于第 i 个高斯混合项的,这样虽然每一个高斯混合项中的均值和协方差都来自于全部的 N 个样本,但是实际上,第 i 个高斯混合项所属的等价样本容量为 $N\pi_i$. 此外,由算法 1 和式(8),因子分析模型中的参数 Λ_i,Ψ_i 完全取决于使用

后验概率 R_{ni} 进行加权的数据样本的一阶和二阶统计量,而这正是混合高斯模型中各个高斯混合项的均值 μ_i 和协方差 Σ_i,因此如果我们按照 μ_i 和 Σ_i 重新抽取一组新的数据样本,然后在这些新的数据样本上学习因子分析模型中的参数,将会得到同样的结果.这样我们就有下面的使用重新抽样方法的混合因子分析模型的学习算法

算法 4　使用重新抽样方法的混合因子分析模型学习算法

(1)建立数据样本概率分布的混合高斯模型;

(2)对每一个高斯混合项独立地抽取新的样本;

(3)每个高斯混合项各自在新的数据样本上学习因子分析模型的参数.

在重新抽样中,为第 i 个高斯混合项抽取的样本数为 $N\pi_i$ 个,这些新数据样本是"完全"属于此高斯混合项的,而与其它的高斯混合项无关.这样与使用式(8)的算法 3 相比较,避免了对后验概率 R_{ni} 的计算,又减少了进行因子分析时参与计算的数据样本的数量.

对每一个高斯混合项而言,新的数据样本与原数据样本相比较必然分布在较小的空间范围内,因此具有更强的局部性.所以在某种意义上来说,本文提出的混合因子分析的重新抽样方法相当于将原样本空间划分为一些小的子空间,然后在各个子空间上分别进行因子分析.

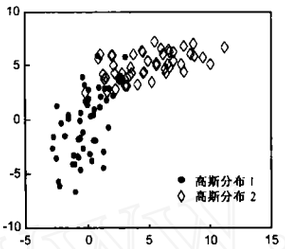


图 2　原始样本的分布

4　实例

我们用两个二维的高斯分布各产生 50 个样本点,然后将它们混合在一起构成一个容量为 100 的原始样本.这两个高斯分布的均值分别为 $[0,0]^T$ 和 $[5,5]^T$,协方差矩阵分别为 $\begin{bmatrix}2 & 2\\ 2 & 6\end{bmatrix}$ 和 $\begin{bmatrix}6 & 2\\ 2 & 2\end{bmatrix}$.图 2 给出了这 100 个样本点在空间的分布情况,并对来自不同高斯分布的样本点做了不同的标记.

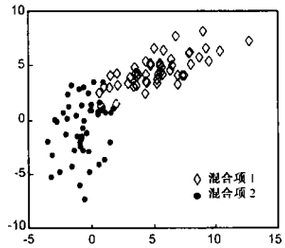


图 3　重新抽取样本的分布

我们首先使用两个高斯混合项,应用算法 2 学习数据样本概率分布的混合高斯模型中的参数 π_i, μ_i, Σ_i, $i = 1,2$.然后按照这些参数为各个高斯混合项抽取新的数据样本,样本数为 $100\pi_i$,这些新的数据样本的空间分布如图 3 所示.接下来分别对每个高斯混合项在各自的新数据样本上使用算法 1 进行因子分析模型的学习,其中公共因子的维数为 1,算法迭代 20 次.图 4 给出了对这两个高斯混合项进行因子分析之后,按照模型(1)重构出的协方差矩阵 $\Lambda\Lambda^T$ 与样本协方差矩阵的误差随迭代步数的变化曲线,可以看到算法的收敛速度是相当快的.在误差的计算中,使用的是矩阵的 ∞ 范数 $|A|_\infty = n \cdot \max_{i,j} |a_{ij}|$.

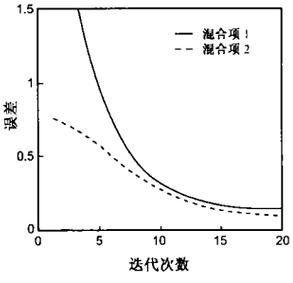


图 4　协方差矩阵的误差曲线

5　小结

混合因子分析是一种对具有复杂结构的多维数据建立模型的方法.本文提出了一种进行混合因子分析的重新抽样方法.当给定一组数据样本时,我们首先建立此样本概率分布的混合高斯模型,然后再为每一个高斯混合项重新抽取新的数据样本,最后在各自新的样本上对每一个高斯混合项进行因子分析.与已有的算法相比较,避免了计算各个高斯混合项在每个样本值之下的后验概率,又减少了进行因子分析时参与计算的数据样本的数量.文中的实例表明了该算法同时具有较好的性能.

参考文献:

[1] Everitt B S. An Introduction to Latent Variable Models[M]. London: Chapman and Hall, 1984.

[2] Redner R A, Walker H F. Mixture densities, maximum likelihood, and the EM algorithm[J]. SIAM Rev, 1984, 26(2): 195 - 239.

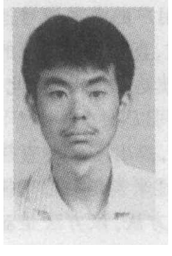
[3] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society, 1977, B-39(1): 1 - 38.

[4] Hinton G E, Dayan P, Revow M. Modeling the manifolds of images of handwritten digits[J]. IEEE Trans on Neural Networks, 1997, 8(1): 65 - 74.

[5] Ghahramani Z, Hinton G E. The EM Algorithm for Mixtures of Factor Analyzers[R]. Technical Report CRG-TR-96-1, Dept of Comp Sci, Univ of Toronto, 1996.

[6] Rubin D, Thayer D. EM algorithms for ML factor analysis[J]. Psychometrika, 1982, 47(1): 69 - 76.

作者简介:



岳　博　男,1970 年生于陕西西安,西安电子科技大学在读博士研究生,主要研究方向为概率模型,Bayes 统计方法等.

焦李成　男,1959 年生于陕西白水,1984 年和 1990 年在西安交通大学分别获硕士和博士学位,现为西安电子科技大学教授,博士生导师,主要研究领域有非线性科学,智能信息处理,神经网络,数据挖掘等.