

文章编号:1006-2467(2013)07-1119-07

混合高斯分布的变分贝叶斯学习参数估计

徐定杰, 沈 忱, 沈 锋

(哈尔滨工程大学 自动化学院, 哈尔滨 150001)

摘 要: 针对常用于非高斯信号或系统建模的包含隐变量的混合高斯分布模型, 提出利用一种变分贝叶斯学习算法进行模型的参数估计. 该方法采用一个形式较为简单的自由分布, 通过不断最大化边缘似然函数的下界, 迭代地更新变分参数, 直至近似分布足够逼近参数真实的后验分布, 从而实现混合高斯分布的参数估计. 文中推导了该方法对混合高斯模型参数学习过程. 实验表明, 变分贝叶斯学习可以有效实现高斯混合模型的多参数估计, 相比采样方法更有工程应用前景.

关键词: 参数估计; 混合模型; 高斯分布; 变分贝叶斯

中图分类号: TN 911.23

文献标志码: A

Variational Bayesian Learning for Parameter Estimation of Mixture of Gaussians

XU Ding-jie, SHEN Chen, SHEN Feng

(College of Automation, Harbin Engineering University, Harbin 150001, China)

Abstract: Non-Gaussian signals or systems are usually modeled by mixture of Gaussians (MoG) models containing hidden variables. A variational Bayesian learning algorithm was suggested to infer the parameters of MoG. The algorithm estimated the parameters of MoG by iteratively maximizing the lower bound of the marginal likelihood and updating the variational parameters until the free-form distribution was sufficiently close to the true posterior. The detailed learning of variational Bayes for MoG was derived and explained. The experiments show that this method can estimate the parameters of MoG favorably with sampling method from the engineering view.

Key words: parameters estimation; mixture model; Gaussian distribution; variational Bayes

在通信、导航、声呐、雷达及生物医学等领域中都存在着典型的非高斯噪声环境, 如低频大气噪声、水声信号、生物医学中的心电信号^[1-3]. 基于高斯噪声模型的信号处理方法在非高斯环境下工作性能会遭受很大损失, 甚至无法工作. 如果能够辨识非高斯噪声的统计特性并加以利用, 则能大幅提高信号处理性能^[4]. 针对非高斯噪声的建模, 目前普遍采用混

合分布模型, 常用的主要包括混合高斯分布、混合伽马分布、混合韦伯分布、混合泊松分布等, 其中尤以混合高斯分布的研究与应用最为广泛^[5-7].

目前针对混合高斯分布参数估计方法主要有最大似然估计 (Maximum Likelihood, ML)、期望最大 (Expectation Maximization, EM)、马尔可夫链蒙特卡洛 (Markov Chain Monte Carlo, MCMC) 等^[8-9].

收稿日期: 2012-07-13

基金项目: 国家自然科学基金资助项目 (61102107), 中国博士后基金项目 (20100480979), 黑龙江省博士后基金项目 (LBH-Z10193)

作者简介: 徐定杰 (1966-), 男, 浙江省瑞安市人, 教授, 现主要从事无线电及卫星导航技术、雷达抗干扰技术方向研究.

电话 (Tel.): 0451-82569651; E-mail: xdjyzyxy@163.com.

ML 通过对似然函数的求解估计参数的后验分布,但在混合模型参数估计中涉及复杂的非线性函数,求解存在困难;EM 作为一种观测数据为不完全数据时求解最大似然函数的迭代算法,有效降低了多参数 ML 求解复杂度,但 EM 与 ML 一样,都容易出现数据过拟合情形,对模型阶数的估计也存在一定困难.同时,EM 与 ML 都无法利用已有的先验信息,而先验信息是贝叶斯推断理论的基础和出发点^[10].MCMC 利用了已有的先验信息,但它收敛速度慢,且马尔可夫链的收敛性也难以保证.而且,对于多参数的节点模型结构,MCMC 计算量巨大,难以在实际工程中应用和推广.

近年来,一种能适应高维数观测环境,以较少运算量进行参数估计的变分近似方法逐渐被人们接受. Attias^[11]最早提出了将变分近似思想应用于图像模型的参数推断中,成为了变分贝叶斯方法的雏形.人们将此方法用于随机系统的参数估计^[12],以及分别借鉴该方法将其应用于盲源分离^[13]和语音增强领域^[14]都取得了不错的成果.相比同样作为逼近方法进行参数估计的 MCMC,变分贝叶斯学习在较好估计精度的前提下有更快的估计速度,更适用于有实时性要求的工程应用领域.鉴于广泛存在于工程领域中的非高斯噪声,近年来也逐渐成为信号处理领域的研究热点,而含有隐变量的高斯混合模型能较好地描述这类信号.本文提出一种基于变分贝叶斯学习的高斯混合模型参数估计方法,利用变分贝叶斯对含有隐变量的混合高斯模型进行参数估计,并对整个过程进行了详细推导,以此为变分贝叶斯学习算法今后在信号处理领域应用与扩展提供一定的理论支持.

1 变分贝叶斯学习算法

变分贝叶斯学习是在传统贝叶斯推断与 EM 迭代估计算法的基础上引入变分近似理论而提出的.用 Y 代表观测数据矢量, θ, Z 分别表示各个未知模型参数和以及隐变量.其总体思想是避开高维数多变量环境下对边缘似然函数 $p(Y)$ 复杂的多重多维积分运算,并考虑参数的先验信息 $p(\theta, Z)$,利用一族较易处理的分布 $q(\theta, Z)$ 去逼近参数的真实后验分布 $p(\theta, Z|Y)$.变分贝叶斯通过最大化变分参数的对数边缘似然函数的下界(目标函数)求得模型参数,并利用统计物理学中均值域理论^[15],将多变量的联合概率分布近似为各变量边缘概率分布的乘积,使得对多变量的联合估计方便地转化为对这些变量边缘分布的迭代估计,计算复杂度显著降低,运

算效率得到提高.贝叶斯模型的对数边缘似然函数可等价表示为

$$\ln p(Y) = \iint q(\theta, Z) \ln \frac{p(Y, \theta, Z)}{q(\theta, Z)} d\theta dZ + \iint q(\theta, Z) \ln \frac{q(\theta, Z)}{p(\theta, Z|Y)} d\theta dZ = F(q(\theta, Z)) + \text{KL}(q(\theta, Z) \| p(\theta, Z|Y)) \quad (1)$$

式中: $\text{KL}(q(\theta, Z) \| p(\theta, Z|Y)) \geq 0$ 为 $q(\theta, Z)$ 与 $p(\theta, Z|Y)$ 之间的 Kullback-Leibler (KL) 散度; $F(q(\theta, Z))$ 为变分自由能量.当 $q(\theta, Z)$ 与 $p(\theta, Z|Y)$ 同分布时等号成立,此时散度最小, $F(q(\theta, Z))$ 达到最大值.因此,从几何意义上看, $F(q(\theta, Z))$ 是 $\ln p(Y)$ 的下界.最大化变分自由能量等价于最小化 $q(\theta, Z)$ 与 $p(\theta, Z|Y)$ 的 KL 散度,当 KL 散度为 0,即 $q(\theta, Z) \approx p(\theta, Z|Y)$ 时,近似分布可认为等价于原分布,此时 $\ln p(Y)$ 的下界 $F(q(\theta, Z))$ 为最大.变分贝叶斯学习通过 $q(\theta, Z)$ 的迭代实现 $F(q(\theta, Z))$ 的最大化.令 $q(\theta, Z) = q(\theta)q(Z)$,根据变分学泛函理论,分别对 $q(\theta)$ 与 $q(Z)$ 求偏导,可得相应的通解解析表达式:

$$q(\theta_i) = \frac{\exp[E_{q(\theta_{k \neq i})}[\ln p(Y, \theta, Z)]]}{\int \exp[E_{q(\theta_{k \neq i})}[\ln p(Y, \theta, Z)]] d\theta_i} \quad (2)$$

式(2)分母是归一化因子常数,而每一个参数的分布 $q(\theta_i)$ 都需要涉及对其他分布 $q(\theta_k)$ 期望的计算,因此,与 EM 算法中迭代理论类似,归纳以 $F(q(\theta, Z))$ 为目标函数的迭代过程如图 1 所示.即初始化 $q(\theta, Z)$ 中的超参数,再进行循环迭代参数更新,每一循环步骤计算得:

$$q_m(Z) \propto \exp[E_{q(\theta)}[\ln p(Y, Z, \theta)]] \quad (3)$$

$$q_m(\theta) \propto \exp[E_{q(Z)}[\ln p(Y, Z, \theta)]] \quad (4)$$

直至 $\Delta F = |F_M(q(\theta)) - F_{M-1}(q(\theta))| < t$, t 为设定的下界阈值,数量级较低,用于判断收敛, M 代表循

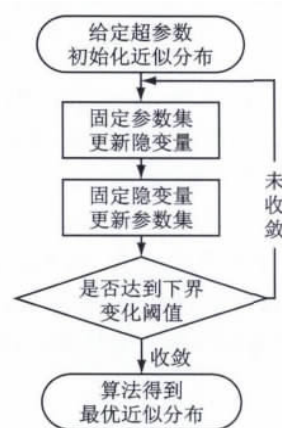


图1 变分贝叶斯学习算法流程图

环次数. 可以发现, 算法中涉及最大化似然函数与求解期望 2 个步骤, 参数更新准则类似于 EM 算法, 因此, 变分贝叶斯学习算法通常也称为变分贝叶斯期望最大 (Variational Bayesian Expectation Maximization, VBEM) 算法^[9], 是广义化的 EM 算法.

2 混合高斯分布模型的参数估计

混合高斯模型可以表示为 K 个高斯分布的线性组合, 即

$$p(Y|\pi, \mu, \Lambda) = \sum_{k=1}^K \pi_k N(Y|\mu_k, \Lambda_k^{-1}) \quad (5)$$

式中: Y 为 D 维随机观测变量集合; $\pi = [\pi_1, \pi_2, \dots, \pi_K]$ 为混合参数; μ_k, Λ_k 分别为第 k 个高斯分布的均值矢量和精度矩阵. 当 $K=1$ 且 $\pi_k=1$ 时, 混合高斯模型就退化成了一般的单变量高斯模型. 假设共有 N 个观测样本, 引入隐变量 Z 以表征每个 D 维的观测样本 y_n 来源于某个特定的高斯分布, 可将 K 个高斯分布的隐变量 Z 表示为 $N \times K$ 矩阵. 其中, 元素 z_{nk} 取 0 或 1, $z_{nk}=1$ 表示 y_n 对应第 k 个高斯分量. 特别注意, 每个样本 y_n 只能对应一个分量. 给定混合参数 π 以表征每个高斯分量在混合模型中所占权重, 因此, 隐变量 Z 的条件概率密度函数可以表示为^[16]

$$p(Z|\pi) = \prod_{n=1}^N \prod_{k=1}^K \pi_k^{z_{nk}} \quad (6)$$

共轭指数分布是常用的一类参数先验分布, 其思想基础是先验和后验规律具有一致性, 即先验分布和后验分布属于同一类分布族, 而共轭指数域内分布的性质能够保证这种一致性^[17]. 对于混合参数 π , 其先验为狄里克利分布, 即

$$p(\pi) = \text{Dir}(\pi|\alpha_0) = C(\alpha_0) \prod_{k=1}^K \pi_k^{\alpha_k-1} \quad (7)$$

为方便起见, 本文取 K 个混合参数具有相同的先验分布, 即 $\alpha_1 = \alpha_2 = \dots = \alpha_K = \alpha_0$, 因此, α_0 是具有 K 个元素的矢量, $\alpha_0 = [\alpha_0 \dots \alpha_0]^T$, 则带有隐变量且每个观测样本独立同分布的混合高斯模型概率密度可以表示为

$$p(Y|Z, \pi, \mu, \Lambda) = \prod_{n=1}^N \prod_{k=1}^K \left\{ \frac{1}{(2\pi)^{D/2}} |\Lambda_k|^{1/2} \times \exp \left[-\frac{1}{2} (y_n - \mu_k)^T \Lambda_k (y_n - \mu_k) \right] \right\}^{z_{nk}} \quad (8)$$

均值矢量 μ 和精度矩阵 Λ 的共轭先验分布可表示为条件高斯分布与威沙特分布的乘积, 即

$$p(\mu, \Lambda) = p(\mu|\Lambda) p(\Lambda) =$$

$$\left(\prod_{k=1}^K N(\mu_k|m_0, (\beta_0 \Lambda_k)^{-1}) W(\Lambda_k|W_0, \nu_0) \right) \quad (9)$$

式中: W_0 为正定阵; ν_0 为自由度参数.

$p(Y)$ 的计算需要对包括隐变量在内的 Z, π, Λ 和 μ 4 个参数进行积分, 且积分项较为复杂, 根据变分近似的思想, 引入形式简单的分布 $q(\theta)$. 由于有共轭指数域先验与后验分布一致性的保证, 使得通过在共轭指数域内选取分布函数能够得到闭合的解析表达式, 并通过循环迭代方法求出最优的超参数解. 假设 $q(Z, \pi, \Lambda, \mu)$ 为近似解的联合密度函数, 隐变量与各个参数之间相互独立, 建议将分布因子化, 即

$$q(Z, \pi, \Lambda, \mu) = q(Z) q(\pi) \prod_{k=1}^K q(\mu_k|\Lambda_k) q(\Lambda_k) \quad (10)$$

值得注意, Λ 和 μ 2 个变量不是相互独立的, 根据先验知识, μ 是 Λ 条件下的高斯分布. 将变分贝叶斯学习运用于混合高斯模型的参数估计, 下面由式 (3) 计算隐变量 Z 的近似后验分布:

$$q(Z) \propto \exp \left\{ \sum_{n=1}^N \sum_{k=1}^K z_{nk} \left[\frac{1}{2} E_{\Lambda} \ln |\Lambda_k| - \frac{D}{2} \ln 2\pi - \frac{1}{2} E_{\Lambda, \mu} [(y_n - \mu_k)^T \Lambda_k (y_n - \mu_k)] + E_{\pi} [\ln \pi_k] \right] \right\} = \prod_{n=1}^N \prod_{k=1}^K r_{nk}^{z_{nk}} \quad (11)$$

式中, r_{nk} 需要进行归一化处理.

同理, 由式 (4) 推导其他模型参数的近似联合后验分布:

$$q(\pi, \Lambda, \mu) \propto \exp \left[\sum_{n=1}^N \sum_{k=1}^K E_{z_{nk}} [z_{nk}] (\ln N(y_n|\mu_k, \Lambda_k^{-1}) + \ln \pi_k) + \ln \text{Dir}(\pi|\alpha_0) + \sum_{k=1}^K \ln (N(\mu_k|m_0, (\beta_0 \Lambda_k)^{-1}) W(\Lambda_k|W_0, \nu_0)) \right] \quad (12)$$

由于 z_{nk} 是离散型变量, 只在 0 和 1 之间取值, 式 (11) 已给出了 Z 的近似分布, 因此, 根据概率密度函数形式, 每个 z_{nk} 的期望为

$$E_{z_{nk}} [z_{nk}] = 0 \times p(z_{nk} = 0) + 1 \times p(z_{nk} = 1) = 1 \times r_{nk} = r_{nk} \quad (13)$$

注意到式 (12) 混合系数、精度矩阵和均值向量 3 个参数是相互独立的, 因此, 可分别推导其近似分布表达式. 对于混合系数 π , 可以将其当作另一个狄里克利分布:

$$q(\pi) = C \prod_{k=1}^K \pi_k^{\alpha_0 + \sum_{n=1}^N r_{nk} - 1} = \text{Dir}(\pi|\alpha) \quad (14)$$

式中: C 为常量; $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_K)^T$.

为表述方便, 定义

$$\sum_{n=1}^N r_{nk} = N_k, \quad \bar{\mathbf{y}}_k = \frac{1}{N_k} \sum_{n=1}^N r_{nk} \mathbf{y}_n$$

$$S_k = \frac{1}{N_k} \sum_{n=1}^N r_{nk} (\mathbf{y}_n - \bar{\mathbf{y}}_k)(\mathbf{y}_n - \bar{\mathbf{y}}_k)^T$$

则

$$\alpha_k = \alpha_0 + N_k \quad (15)$$

下面考虑形式较为复杂的关于精度矩阵与样本均值的近似分布. 参照式(10), 分别处理 $q(\boldsymbol{\mu}_k | \boldsymbol{\Lambda}_k)$ 与 $q(\boldsymbol{\Lambda}_k)$:

$$q(\boldsymbol{\mu}_k | \boldsymbol{\Lambda}_k) \propto \exp \left[-\frac{1}{2} (\boldsymbol{\mu}_k^T \beta_k \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k - \mathbf{m}_k^T \beta_k \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k - \boldsymbol{\mu}_k^T \beta_k \boldsymbol{\Lambda}_k \mathbf{m}_k + \mathbf{m}_k^T \beta_k \boldsymbol{\Lambda}_k \mathbf{m}_k) \right] =$$

$$N(\boldsymbol{\mu}_k | \mathbf{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}) \quad (16)$$

式中:

$$\beta_k = \beta_0 + \sum_{n=1}^N r_{nk} = \beta_0 + N_k \quad (17)$$

$$\mathbf{m}_k = \frac{\beta_0 \mathbf{m}_0 + N_k \bar{\mathbf{y}}_k}{\beta_0 + N_k} \quad (18)$$

为了得到与先验一致的分布形式, 通过整理得到新的威沙特分布为

$$q(\boldsymbol{\Lambda}_k) \propto \exp \left\{ -\frac{1}{2} \text{Tr} \left[(\mathbf{W}_0^{-1} + \beta_0 (\boldsymbol{\mu}_k - \mathbf{m}_0) \times (\boldsymbol{\mu}_k - \mathbf{m}_0)^T + \sum_{n=1}^N E_{z_{nk}} [z_{nk}] (\mathbf{y}_n - \boldsymbol{\mu}_k) \times (\mathbf{y}_n - \boldsymbol{\mu}_k)^T - \beta_k (\boldsymbol{\mu}_k - \mathbf{m}_k) (\boldsymbol{\mu}_k - \mathbf{m}_k)^T \boldsymbol{\Lambda}_k] \right] + \frac{1}{2} \left[\sum_{n=1}^N E_{z_{nk}} [z_{nk}] + v_0 - D - 1 \right] \ln |\boldsymbol{\Lambda}_k| \right\} =$$

$$W(\boldsymbol{\Lambda}_k | v_k, \mathbf{W}_k) \quad (19)$$

式中:

$$v_k = \sum_{n=1}^N E_{z_{nk}} [z_{nk}] + v_0 =$$

$$\sum_{n=1}^N r_{nk} + v_0 = v_0 + N_k \quad (20)$$

$$\mathbf{W}_k^{-1} = \mathbf{W}_0^{-1} + \beta_0 \mathbf{m}_0^T \mathbf{m}_0 - \beta_k \mathbf{m}_k^T \mathbf{m}_k +$$

$$\sum_{n=1}^N E_{z_{nk}} [z_{nk}] (\mathbf{y}_n \mathbf{y}_n^T) = \mathbf{W}_0^{-1} + N_k \mathbf{S}_k +$$

$$\frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{\mathbf{y}}_k - \mathbf{m}_0) (\bar{\mathbf{y}}_k - \mathbf{m}_0)^T \quad (21)$$

由高斯分布、狄里克利分布和威沙特分布的期望性质, 可得式(11)中的各个期望如下:

$$E_{\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k} [(\mathbf{y}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\mathbf{y}_n - \boldsymbol{\mu}_k)] =$$

$$v_k (\mathbf{y}_n - \mathbf{m}_k)^T \mathbf{W}_k (\mathbf{y}_n - \mathbf{m}_k) + D \beta_k^{-1} \quad (22)$$

$$E_{\boldsymbol{\Lambda}_k} [\ln |\boldsymbol{\Lambda}_k|] = \sum_{i=1}^D \Psi \left(\frac{v_k + 1 - i}{2} \right) +$$

$$D \ln 2 + \ln |\mathbf{W}_k| \quad (23)$$

$$E_{\pi_k} [\ln \pi_k] = \Psi(\alpha_k) - \Psi \left(\sum_{k=1}^K \alpha_k \right) \quad (24)$$

式中: $\Psi(\boldsymbol{\Theta})$ 为双伽马函数, 为伽马函数 $\Gamma(\boldsymbol{\Theta})$ 取自然对数后关于 $\boldsymbol{\Theta}$ 的一阶导数.

正如式(3)、(4)所示变量之间的依赖性, 发现计算式(11)中需要计算期望(22)~(24), 而三者的计算均有赖于(15)、(17)、(18)、(20)与(21), 但这 5 式中需要计算 N_k 、 $\bar{\mathbf{y}}_k$ 与 S_k , 而三者的计算又必须已知式(11)即 r_{nk} . 对照图 1 流程, 在初始化超参数 α_0 , β_0 , v_0 , \mathbf{m}_0 与 \mathbf{W}_0 后, 需要像 EM 算法一样开始进行循环迭代计算来得到最优解: 固定分布的参数 α_k , β_k , v_k , \mathbf{m}_k 与 \mathbf{W}_k , 计算式(17)中 3 个期望(22)~(24)来得到隐变量的分布参数 r_{nk} ; 最新得到的 r_{nk} 保持不变, 重新计算 N_k 、 $\bar{\mathbf{y}}_k$ 与 S_k , 从而更新参数 α_k , β_k , v_k , \mathbf{m}_k 与 \mathbf{W}_k ; 如此迭代计算, 至变分自由能量 $F(q(\boldsymbol{\theta}, \mathbf{Z}))$ 最大, 即对数边缘似然函数的下界最大, 可得混合高斯模型的变分贝叶斯学习参数估计.

下面给出判断收敛性的下界 $F(q(\boldsymbol{\theta}, \mathbf{Z}))$ 的确定方法. 将式(10)代入 $F(q(\boldsymbol{\theta}, \mathbf{Z}))$:

$$F(q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})) = E[\ln p(\mathbf{Y} | \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) + \ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) + \ln p(\mathbf{Z} | \boldsymbol{\pi}) + \ln p(\boldsymbol{\pi})] -$$

$$E[\ln q(\boldsymbol{\pi}) + \ln q(\mathbf{Z}) + \ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})] \quad (25)$$

利用各个分布性质以及式(22)~(24), 计算式(25)中各项期望. 其中, 对于 $E[\ln q(\mathbf{Z})]$, $E[\ln q(\boldsymbol{\pi})]$, $E[\ln q(\boldsymbol{\mu} | \boldsymbol{\Lambda})]$ 以及 $E[\ln q(\boldsymbol{\Lambda})]$, 只需要将其代入每个近似分布的负熵值计算式即可. 熵计算公式为

$$H(p(\boldsymbol{\theta})) = - \int p(\boldsymbol{\theta}) \ln p(\boldsymbol{\theta}) d\boldsymbol{\theta} \quad (26)$$

其他期望分别通过直接计算得到:

$$E[\ln p(\mathbf{Z} | \boldsymbol{\pi})] = \sum_{n=1}^N \sum_{k=1}^K r_{nk} E_{\pi_k} [\ln \pi_k] \quad (27)$$

$$E[\ln p(\boldsymbol{\pi})] =$$

$$\sum_{k=1}^K (\alpha_0 - 1) E_{\pi_k} [\ln \pi_k] + \ln C(\boldsymbol{\alpha}_0) \quad (28)$$

$$E[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda})] = \frac{1}{2} \sum_{k=1}^K \left\{ D \ln \frac{\beta_0}{2\pi} + E_{\Lambda_k} [\ln |\boldsymbol{\Lambda}_k|] - \frac{D\beta_0}{\beta_k} - \beta_0 v_k (\mathbf{m}_k - \mathbf{m}_0)^T \mathbf{W}_k (\mathbf{m}_k - \mathbf{m}_0) \right\} -$$

$$\frac{1}{2} v_k \text{Tr}(\mathbf{W}_0^{-1} \mathbf{W}_k) + K \ln B(\mathbf{W}_0, v_0) +$$

$$\frac{v_0 - D - 1}{2} \sum_{k=1}^K E_{\Lambda_k} [\ln |\boldsymbol{\Lambda}_k|] \quad (29)$$

$$E[\ln p(\mathbf{Y} | \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] = \frac{1}{2} \sum_{k=1}^K \left\{ N_k E_{\Lambda_k} [\ln |\boldsymbol{\Lambda}_k|] - D/\beta_k - v_k \text{Tr}(\mathbf{S}_k \mathbf{W}_k) - v_k (\bar{\mathbf{y}}_k - \mathbf{m}_k)^T \times \right.$$

$$\left. \mathbf{W}_k (\bar{\mathbf{y}}_k - \mathbf{m}_k) - D \ln 2\pi \right\} \quad (30)$$

每次迭代之后计算下界的变化值, 记作 ΔF , 当该值

低于预先设定的近似小量时,可以认定该算法已经趋于收敛,从而得到足够逼近原分布的近似分布。

3 实验结果及分析

为了验证本文提出的混合高斯模型参数估计方法的性能,实验中用程序实现并产生了可以指定维数、分量个数的服从混合高斯分布的样本数据,且样本容量可变。下面用变分贝叶斯学习方法进行模型参数估计。

样本数据 1 为服从混合高斯分布的一维数据,即 $D=1$,样本容量 $N=3\ 000$,如图 2(a)所示。该混合高斯分布由 3 个高斯分布单元构成,即 $K=3$,混合参数 $\pi=[0.25\ 0.35\ 0.40]$,各高斯分布的均值矢量值分别为 $\mu_1=-1.5, \mu_2=1.2, \mu_3=0.5$,各基本单元方差均为 0.05。

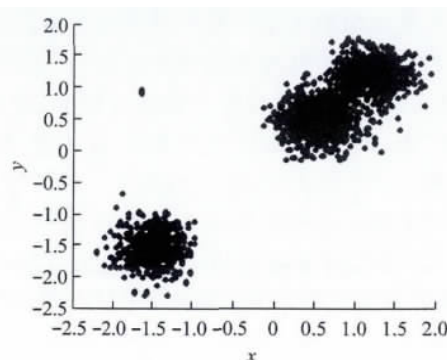
样本数据 2 为服从混合高斯分布的二维数据,即 $D=2$,样本容量 $N=5\ 000$,如图 2(b)所示。该混合高斯分布由 4 个高斯分布单元构成,即 $K=4$,混合参数 $\pi=[0.15\ 0.20\ 0.30\ 0.35]$,各高斯分布的均值矢量 $\mu_1=[0\ 0]^T, \mu_2=[0.3\ 0.3]^T, \mu_3=[-0.3\ -0.3]^T, \mu_4=[0.3\ -0.3]^T$,协方差矩阵均为 $\begin{bmatrix} 0.03 & 0 \\ 0 & 0.03 \end{bmatrix}$ 。

运用变分贝叶斯学习方法进行参数估计。作为循环迭代的开始,首先需要初始化各未知超参数。本文针对 2 组数据取相同先验,分别为 $\alpha_0=1, \beta_0=1, \nu_0=2, m_0=0, W_0=2I$ 。并且对于每个高斯分布单元 $k=1, 2, \dots, K$,初始值取 $K=8, m_k=0, \alpha_k=1, \beta_k=10, \nu_k=2, W_k=2I$ 。仿真中,取相邻 2 次的变分自由能量之差 $\Delta F < 10^{-8} N$ 时停止迭代。

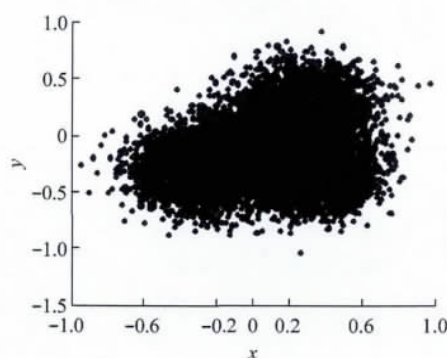
图 3 给出了样本数据 1 变分贝叶斯参数学习后描述的混合高斯分布概率密度曲线与真实概率密度曲线的对比。由图可见,利用变分贝叶斯学习方法估计的分布与真实分布基本重合。

图 4 给出了变分自由能量 $F=-2\ 970.652$ 时样本数据 2 参数估计后不同高斯基本分布单元的二维等值线图。为了便于观察,对样本数据进行了弱化显示。虽然在算法初始化时给定的高斯分量个数为 8,但由于设置了隐变量,变分贝叶斯在每个迭代循环中更新完超参数后,都会重新对各个分量的权值进行计算,当隐变量所指示分量在加权求和时权重过小,可忽略该分量的存在,即认为 $z_{nk}=0$ 。本实验中,程序设置权值小于 1% 时忽略该分量。

综合图 3、4 可知,算法最终收敛后通过模型选择,样本 1 得到 3 个有效分量,样本 2 为 4 个。



(a) 一维高斯样本



(b) 二维高斯样本

图 2 一、二维高斯样本

Fig. 2 1D, 2D Gaussian samples

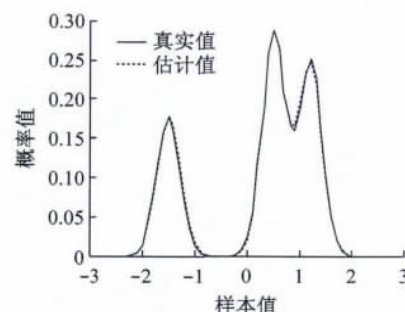


图 3 样本 1 估计概率密度与真实概率密度

Fig. 3 Estimated pdf and true pdf of Sample 1

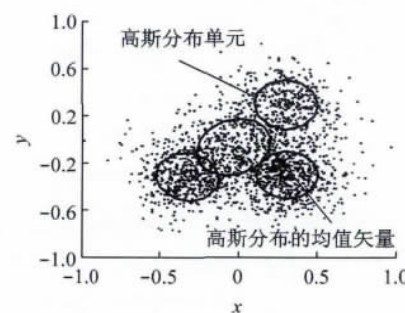


图 4 样本 2 高斯分量

Fig. 4 Gaussian components of Sample 2

表 1 量化地给出了变分贝叶斯学习方法对于上述两组服从混合高斯分布的样本数据的参数学习结果. 为进一步体现该学习算法的估计精度, 表中还给出了 MCMC 对于 2 组样本数据的估计值, 样本容量分别取 5 000 与 10 000. 通过对比 2 种算法发现, 变分贝叶斯估计精度总体优于样本容量为 5 000 的 MCMC, 而样本容量增加后, MCMC 估计精度得到提升, 略优于变分贝叶斯. 虽然 MCMC 估计精度跟变分贝叶斯相当, 都有较小的估计误差, 甚至精度还略高于变分贝叶斯, 但本实验中, MCMC 为了保证算法必须收敛, 同时达到较好的估计精度, 使用了容

量为 10 000 样本数据, 设定循环次数为 500 次, 耗时约 16 200 s (实验环境为一台主频 2.66 GHz 的双核 Pentium 4 处理器, 搭载 Windows XP 系统).

图 5 给出了自由变分能量(下界)迭代收敛图. 图中, 横坐标 t 为变分贝叶斯学习算法运行时间. 对比图中 $K=8$ 曲线, 约在 1 900 s 左右收敛. 由此可见, MCMC 由于计算量巨大, 耗时较长, 为了提高有限的估计精度而大幅增加样本容量与循环次数, 代价过大, 在工程应用上较为困难, 而变分贝叶斯算法在保证运算量较小且运算速度较快的情况下, 有着较为满意的估计精度.

表 1 混合高斯模型参数估计值
Tab. 1 Estimated parameters of MoG

样本	基本单元	混合参数 π_k			均值矢量 μ_k			协方差矩阵 Δ_k^{-1}		
		MCMC			MCMC			MCMC		
		变分 算法	容量 5 000	容量 10 000	变分 算法	容量 5 000	容量 10 000	变分 算法	容量 5 000	容量 10 000
1	1	0.254	0.242	0.251	-1.486	-1.537	-1.501	0.051	0.054	0.051
	2	0.360	0.372	0.356	1.183	1.182	1.199	0.055	0.055	0.051
	3	0.385	0.387	0.393	0.494	0.488	0.499	0.047	0.054	0.051
2	1	0.143	0.142	0.148	$\begin{bmatrix} -0.030 \\ -0.064 \end{bmatrix}$	$\begin{bmatrix} -0.042 \\ -0.065 \end{bmatrix}$	$\begin{bmatrix} -0.003 \\ -0.002 \end{bmatrix}$	$\begin{bmatrix} 0.029 & 0.001 \\ -0.001 & 0.032 \end{bmatrix}$	$\begin{bmatrix} 0.031 & -0.000 \\ -0.000 & 0.033 \end{bmatrix}$	$\begin{bmatrix} 0.031 & -0.000 \\ -0.000 & 0.031 \end{bmatrix}$
	2	0.219	0.219	0.206	$\begin{bmatrix} 0.304 \\ 0.297 \end{bmatrix}$	$\begin{bmatrix} 0.316 \\ 0.307 \end{bmatrix}$	$\begin{bmatrix} 0.301 \\ 0.302 \end{bmatrix}$	$\begin{bmatrix} 0.038 & 0.007 \\ 0.007 & 0.040 \end{bmatrix}$	$\begin{bmatrix} 0.035 & -0.005 \\ -0.005 & 0.037 \end{bmatrix}$	$\begin{bmatrix} 0.029 & -0.001 \\ -0.001 & 0.031 \end{bmatrix}$
	3	0.286	0.288	0.299	$\begin{bmatrix} -0.324 \\ -0.317 \end{bmatrix}$	$\begin{bmatrix} -0.319 \\ -0.280 \end{bmatrix}$	$\begin{bmatrix} -0.301 \\ -0.300 \end{bmatrix}$	$\begin{bmatrix} 0.028 & 0.003 \\ -0.003 & 0.030 \end{bmatrix}$	$\begin{bmatrix} 0.033 & -0.001 \\ -0.001 & 0.031 \end{bmatrix}$	$\begin{bmatrix} 0.031 & -0.001 \\ -0.001 & 0.031 \end{bmatrix}$
	4	0.351	0.351	0.348	$\begin{bmatrix} 0.311 \\ -0.301 \end{bmatrix}$	$\begin{bmatrix} 0.317 \\ -0.311 \end{bmatrix}$	$\begin{bmatrix} 0.303 \\ -0.301 \end{bmatrix}$	$\begin{bmatrix} 0.028 & -0.000 \\ -0.000 & 0.030 \end{bmatrix}$	$\begin{bmatrix} 0.032 & -0.000 \\ -0.000 & 0.031 \end{bmatrix}$	$\begin{bmatrix} 0.030 & -0.000 \\ -0.000 & 0.030 \end{bmatrix}$

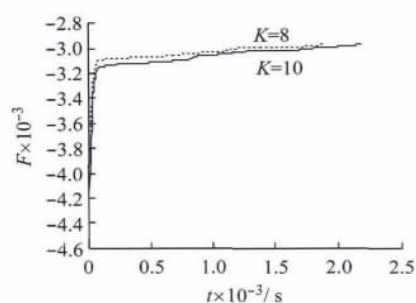


图 5 自由能量变化图

Fig. 5 Diagram of free energy

为验证参数设置对变分贝叶斯学习方法估计性能的影响, 本文对样本数据 2 给定不同基本单元个数的初始值, 即 $K=10$, 其余初始参数同上, 并取相同的先验. 对比 $K=8$, $K=10$ 前期收敛速度变化幅度略小, 最终约在 2 200 s 左右算法收敛, 速度偏慢.

因此, 若高斯分量基本单元个数的初始值偏离真实值越多, 其收敛速度越慢. 由此可见, 参数初始化对于算法的收敛速度有较大影响, 合理准确的初始化可以提高算法速度. 今后可以通过合理初始化参数在速度以及运算量方面继续改进算法.

4 结 语

本文提出了一种基于变分贝叶斯学习的多参数估计方法, 并进行了详细的推导. 仿真结果表明, 该方法可以有效、精确、快速地实现高斯混合模型多参数的估计. 与此同时, 通过混合高斯模型的参数估计方法还可以看出, 应用变分贝叶斯学习, 可以对其他包含隐变量或者其他混合模型进行参数估计, 在工程领域中有较好的实用价值. 今后研究工作将围绕进一步提高算法的实时性、准确性展开.

参考文献:

- [1] Nasri A, Schober R. Performance of BICM-SC and BICM-OFDM systems with diversity reception in non-Gaussian noise and interference [J]. **IEEE Transactions on Communications**, 2009, 57(11): 3316-3327.
- [2] 沈锋, 孙枫. 弱相关非高斯环境下基于局部最佳检测器的伪码捕获方法[J]. **电子与信息学报**, 2010, 32(4): 811-815.
SHEN Feng, SUN Feng. PN code acquisition based on the locally optimum detector in weakly dependent non-Gaussian impulsive channels [J]. **Journal of Electronics & Information Technology**, 2010, 32(4): 811-815.
- [3] He J, Liu Z. Underwater acoustic azimuth and elevation angle estimation using spatial invariance of two identically oriented vector hydrophones at unknown locations in impulsive noise [J]. **Digital Signal Processing**, 2009, 19(3): 452-462.
- [4] 沈锋, 孙枫, 薛冰. 一种弱相关非高斯信道下的伪码捕获方法[J]. **系统工程与电子技术**, 2009, 31(7): 1568-1572.
SHEN Feng, SUN Feng, XUE Bing. Pseudo noise-code acquisition method in weakly dependent non-Gaussian impulsive channels [J]. **Systems Engineering and Electronics**, 2009, 31(7): 1568-1572.
- [5] 王平波, 蔡志明, 刘旺锁. 混合高斯参数估计的动态簇算法[J]. **声学技术**, 2007, 26(4): 741-746.
WANG Ping-bo, CAI Zhi-ming, LIU Wang-suo. Dynamic cluster algorithm for Gaussian mixture parameter estimation [J]. **Technical Acoustics**, 2007, 26(4): 741-746.
- [6] 王永忠, 梁彦, 潘泉. 基于自适应混合高斯模型的时空背景建模[J]. **自动化学报**, 2009, 35(4): 371-378.
WANG Yong-zhong, LIANG Yan, PAN Quan. Spatiotemporal background modeling based on adaptive mixture of Gaussians [J]. **Acta Automatica Sinica**, 2009, 35(4): 371-378.
- [7] Bouguila, N. Bayesian hybrid generative discriminative learning based on finite Liouville mixture models [J]. **Pattern Recognition**, 2011, 44(6): 1183-1200.
- [8] Shen Y, Cornford D. Variational markov chain monte carlo for bayesian smoothing of non-linear diffusions [J]. **Computational Statistics**, 2012, 27(1): 149-176.
- [9] Emtiyaz K M. An expectation-maximization algorithm for learning the latent Gaussian model with Gaussian likelihood [EB/OL]. (2011-04-22) [2012-06-13]. <http://www.cs.ubc.ca/~emtiyaz/Writings/FA1.pdf>.
- [10] 朱慧明, 韩玉启. 贝叶斯多元统计推断理论[M]. 北京: 科学出版社. 2006: 2-5.
- [11] Attias H. A variational Bayesian framework for graphical models [C]// **Advances in Neural Information Processing Systems 12**. Cambridge, MA: [s. n.], 2000: 209-215.
- [12] Vrettas M D, Cornford D, Oppor M. Estimating parameters in stochastic systems: a variational Bayesian approach [J]. **Physica D**, 2011, 240(23): 1877-1900.
- [13] Sun S J, Peng C L, Hou W S. Blind source separation with time series variational Bayes expectation maximization algorithm [J]. **Digital Signal Processing**, 2012, 22(1): 17-33.
- [14] Huang Q H, Yang J, Zhou Y. Variational Bayesian method for speech enhancement [J]. **Neurocomputing**, 2007, 70(16-18): 3063-3067.
- [15] Armagan A, Zaretzki R L. A note on mean-field variational approximations in Bayesian probit models [J]. **Computational Statistics & Data Analysis**, 2011, 55(1): 641-643.
- [16] Ishikawa Y, Takeuchi I, Nakanob R. Multi-directional search from the primitive initial point for Gaussian mixture estimation using variational Bayes method [J]. **Neural Networks**, 2010, 23(3): 356-364.
- [17] Beal M J. Variational algorithms for approximate bayesian inference [D]. London: University College London, 2003.