应用概率统计　第二十四卷
第五期　2008年10月

Chinese Journal of Applied Probability
and Statistics Vol.24 No.5 Oct. 2008

基于最大惩罚似然的高斯混合模型无监督分类研究

余　鹏

(北京大学数学科学学院, 北京, 1000871;　国家基础地理信息中心, 北京, 100044)

童行伟*

(北京师范大学数学科学学院, 北京, 100875)

封举富

(北京大学信息科学学院信息科学中心、视觉与听觉信息处理国家重点实验室, 北京, 100871)

摘　要

本文提出了一个基于高斯混合模型的无监督分类算法. 考虑到利用EM算法求解高斯混合模型的参数参数估计问题容易陷入局部最优解, 我们引入逆Wishart分布来代替传统的Jeffery先验. 几个实验数据的结果表明, 采用该方法估计无监督分类的成分数, 无论是估计的正确率, 还是运算速度, 都有较大提高.

关键词:　高斯混合模型, 无监督分类, 最大惩罚似然, EM算法, 逆Wishart分布.

学科分类号:　O212.7.

§1. 引　　言

高斯混合模型是一个强有力的概率建模工具, 它能应用到模式识别、机器学习、计算机视觉等众多领域. 高斯混合模型的研究最早应追溯到1894年, Pearson研究了两个成分的高斯混合模型的参数估计问题. 1969年, Day[1]研究了高斯混合模型的矩估计、最小χ^2估计、贝叶斯估计以及最大似然估计, 发现最大似然估计要优于其它几类估计. 1977年, Dempster[2]发明了EM算法, 随后, 经Redner和Walker[3]等的研究, EM算法成为高斯混合模型最大似然估计的主要算法.

EM算法相对于MCMC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, 有如下两个问题需要考虑: 第一是初始值的选取; 第二是协方差矩阵的奇异性问题[4]. 对于前者, 一般有如下三个方面的解决方案: 一种方法是从多个初始值开始进行计算, 然后选取似然函数最高的估计; 另一种方法是对数据预先进行聚类; 第三种方法就是近来提出的分割与合并的EM算法(SMEM算法)[5]. 对于后者, 一般采用最大惩罚似然和贝叶斯方法, 但贝叶斯方法在高维情况下的执行效果并不理想[6].

*通讯作者, E-mail: xweitong@bnu.edu.cn.

本文2004年7月22日收到.

另外, 对无监督分类而言, 成分数的确定也是一个没有解决的问题. 一般有如下三种方法: 逼近贝叶斯规则、信息和编码理论、分类似然方法等[4]. 文献[7]通过实验指出, MML要优于其它方法.

本文主要结合了文献[4]和[6]的工作成果, 给出了一个基于最大惩罚似然的高斯混合模型无监督分类方法. 通过比较, 发现在分类精度和运行时间上都有很大改善. 本文组织如下: 第二部分简单介绍了高斯混合模型的无监督方法; 第三部分分析了基于逆Wishart先验的最大惩罚似然方法; 第四部分则是MML规则的介绍以及基于逆Wishart先验的一些修改; 第五部分是结合CEM^{2}估计成分数的具体算法; 第六部分是一些分类实验结果; 最后, 总结全文.

§2. 高斯混合模型和基于EM算法的最大似然估计

2.1 高斯混合模型

设$Y=[Y_1,\cdots,Y_d]^T$是d维的随机变量, $y=[y_1,\cdots,y_d]^T$表示Y的一个的实例. 如果它的概率密度函数能写成k个成分分布的和:

$$p(y|\theta)=\sum_{m=1}^{k}\alpha_m p(y|\theta_m), \tag{2.1}$$

则说Y服从有限混合分布, 其对应的模型就为有限混合模型. 其中, $\alpha_1,\cdots,\alpha_k$是各个成分分布混合的概率; θ_m是第m个成分分布的参数; $\theta\equiv\{\theta_1,\cdots,\theta_k,\alpha_1,\cdots,\alpha_k\}$是所有参数的集合; 同时$\alpha_m$必须满足如下条件:

$$\alpha_m\geq 0,\qquad m=1,\cdots,k \text{ 且 } \sum_{m=1}^{k}\alpha_m=1. \tag{2.2}$$

如果假设$p(y|\theta_m)=2\pi^{-d/2}|\Sigma_m|^{-1/2}\exp\{-(1/2)\cdot(y-\mu_m)^T\Sigma_m^{-1}(y-\mu_m)\}$, 则所对应的模型即为高斯混合模型. d维的高斯混合模型的参数θ_m由均值向量μ_m和方差矩阵Σ_m所决定.

在公式(2.2)约束下, 求(2.1)式参数的解析解非常复杂, 一般采用迭代方法. 即先建立样本的最大似然方程, 然后采用EM算法对类参数及混合参数进行估计.

最大似然估计的基本假设是所有N个样本的集合$\mathbf{Y}=\{y^{(1)},\cdots,y^{(N)}\}$是独立的, 其对应的对数似然函数的公司定义如下:

$$\log p(\mathbf{Y}|\theta)=\sum_{n=1}^{N}\ln\sum_{m=1}^{k}\alpha_m p(y|\theta_m). \tag{2.3}$$

所谓最大似然估计, 就是要找到使(2.3)式最大的θ的估计值:

$$\widehat{\theta}_{\mathrm{ML}}=\max_{\theta}\{\log p(\mathbf{Y}|\theta)\}. \tag{2.4}$$

而最大惩罚似然(也即最大后验规则)则为:

$$\widehat{\theta}_{\mathrm{MAP}}=\max_{\theta}\{\log p(\mathbf{Y}|\theta)+\log p(\theta)\}. \tag{2.5}$$

2.2 基于EM算法的最大似然估计

参数估计的EM算法分为E-步和M-步. E-步计算对数似然函数的期望—Q函数, M-步选择使期望最大的参数, 然后将选择的参数代入E-步, 计算期望, 如此反复. 对最大似然估计而言, 高斯混合模型的迭代公式如下:

E-步: 首先初始化参数μ_m, Σ_m和α_m, 计算样本n属于第m类的后验概率:

$$R_{mn} = \alpha_m p(y|\theta_m) \Big/ \Big[\sum_{m=1}^{k} \alpha_m p(y|\theta_m) \Big]. \tag{2.6}$$

M-步: 在约束(2.2)下, 参数值$\widetilde{\mu}_m, \widetilde{\Sigma}_m$和$\widetilde{\alpha}_m$的最大似然估计迭代计算公式如下:

$$\widetilde{\alpha}_m = \Big(\sum_{n=1}^{N} R_{mn} \Big) / N, \tag{2.7}$$

$$\widetilde{\mu}_m = \Big(\sum_{n=1}^{N} R_{mn} y_n \Big) / (N\widetilde{\alpha}_m), \tag{2.8}$$

$$\widetilde{\Sigma}_m = \Big[\sum_{n=1}^{N} R_{mn} (y_n - \widetilde{\mu}_m)(y_n - \widetilde{\mu}_m)^T \Big] / (N\widetilde{\alpha}_m), \tag{2.9}$$

其中, N为样本个数, d为数据的维数.

§3. 基于逆Wishart先验的高斯混合模型最大惩罚似然估计

但是, 以上算法有一个致命的弱点, 就是其解容易产生退化, 很多作者都注意到这个现象[1][8]. 一个解决办法是在对数似然函数中添加一个惩罚项, 如果将对数先验作为惩罚项, 就是最大后验参数估计(参见公式(2.5)). 在单变量的情况下, Ridolfi和Idier[9]提出使用逆Gamma先验来作为惩罚项. Ormoneit和Tresp[6]以及Snoussi和M-Djafari[10]则使用逆Wishart先验作为惩罚项来避免解的退化. 各个参数的先验假设如下:

$$p(\theta) = D(\alpha|\gamma) \prod_{m=1}^{k} p(\mu_m, \Sigma_m) = D(\alpha|\gamma) \prod_{m=1}^{k} N(\mu_m|v_m, \eta_m^{-1}\Sigma_m) IW(\Sigma_m^{-1}|\alpha_m, \beta_m). \tag{3.1}$$

其中,

$$D(\alpha|\gamma) = b(\gamma) \prod_{m=1}^{k} \alpha_m^{\gamma_m - 1}, \qquad \alpha_m \geq 0, \ \sum_{m=1}^{k} \alpha_m = 1, \tag{3.2}$$

$$N(\mu_m|\nu_m, \eta_m^{-1}\Sigma_m) = (2\pi)^{-d/2} |\eta_m^{-1}\Sigma_m|^{1/2} \exp\Big[-\frac{\eta_m}{2} (\mu_m - \nu_m)^T \Sigma_m^{-1} (\mu_m - \nu_m) \Big], \tag{3.3}$$

$$IW(\Sigma_m^{-1}|\delta_m, \beta_m) = c(\delta_m, \beta_m) |\Sigma_m^{-1}|^{\delta_m - (d+1)/2} \exp[-\mathrm{tr}(\beta_m \Sigma_m^{-1})]. \tag{3.4}$$

其中, $\delta_m > (d-1)/2$, $b(\gamma)$和$c(\delta_m, \beta_m)$是标准化因子, $\mathrm{tr}(\cdot)$是迹运算.

将公式(3.2)(3.3)(3.4)代入(2.5)式, 我们有:

$$\theta_{\rm MAP} = \max_{\theta}\Big\{\sum_{n=1}^{N}\sum_{m=1}^{k} R_{mn}[\log\alpha_m + \log N(x_n|\mu_m,\Sigma_m)] + \log D(\alpha_m|\gamma) + \sum_{m=1}^{k}[\log N(\mu_m|\nu_m,\eta_m^{-1}\Sigma_m) + \log IW(\Sigma_m^{-1}|\delta_m,\beta_m)]\Big\}. \tag{3.5}$$

而具体的迭代公式如下:

E-步: 后验概率的迭代公式与公式(2.6)一致, 其方程如下:

$$R_{mn} = \alpha_m p(y|\theta_m)\Big/\Big[\sum_{m=1}^{k}\alpha_m p(y|\theta_m)\Big]. \tag{3.6}$$

M-步: 各参数值$\widetilde{\mu}_m, \widetilde{\Sigma}_m$和$\widetilde{\alpha}_m$的最大惩罚似然估计迭代公式如下:

$$\widetilde{\alpha}_m = \Big(\sum_{n=1}^{N} R_{mn} + \gamma_m - 1\Big)\Big/\Big(N + \sum_{m=1}^{k}\gamma_m - k\Big), \tag{3.7}$$

$$\widetilde{\mu}_m = \Big(\sum_{n=1}^{N} R_{mn}x_n + \eta_m\nu_m\Big)/(N\widetilde{\alpha}_m + \eta_m), \tag{3.8}$$

$$\widetilde{\Sigma}_m = \frac{\sum_{n=1}^{N} R_{mn}(x_n-\widetilde{\mu}_m)(x_n-\widetilde{\mu}_m)^T + \eta_m(\widetilde{\mu}_m - v_m)(\widetilde{\mu}_m - v_m)^T + 2\beta_m}{N\widetilde{\alpha}_m + 2\delta_m - d}, \tag{3.9}$$

其中, γ,η,ν,δ都为常数向量, β为常数矩阵, N为样本个数, d为数据的维数.

§4. 成分数估计的MML规则

最小信息长度(minimum message length, "MML")准则是由Wallace和Freeman[11]于1987年提出来的. 它从编码过程来考虑统计估计问题, 其基本假设是通过该模型所导致的编码长度的变短能够补偿为发现该估计值所带来的变长. 这样自然在模型的复杂性与拟合的准确性之间达到平衡. 其信息长度的公式如下:

$$\text{MessLen} \approx -\log p(\theta) + \frac{1}{2}\log|I(\theta)| - \log p(\mathbf{Y}|\theta) + \frac{c}{2}\log\kappa_c + \frac{c}{2}. \tag{4.1}$$

其中, $p(\theta)$是参数的先验分布; $I(\theta)$是期望的Fisher信息矩阵, 即$I(\theta) = -\mathsf{E}[\partial^2\log p(\mathbf{Y}|\theta)/\partial\theta^2]$; $p(\mathbf{Y}|\theta)$是混合的似然函数; c是被估计的参数数目; κ_c是c最优的量化阁网常数, 其中$\kappa_1 = 1/12$, $\kappa_2 = 5/(36\sqrt{3})$, 其值可参考Conwan和Sloane[12]的列表.

方程(4.1)中的先验分布采用公式(3.1)中的$p(\theta)$, 但对于混合模型, 不能写出Fisher信息矩阵$I(\theta)$的表达式. 为避免这个困难, 在[4]中使用了$I(\theta)$的上边界$I_c(\theta)$.

$$\widehat{c}_{\rm MDL} = \arg\min_{c}\Big\{-\sum_{m=1}^{k}\Big[(\gamma_m-1)\log\alpha_m - \frac{1}{2}\log|\Sigma_m| - \frac{\eta_m}{2}(\mu_m-v_m)^T\Sigma_m^{-1}(\mu_m-v_m) + \Big(\delta_m - \frac{d+1}{2}\Big)\log|\Sigma_m^{-1}| - \mathrm{tr}(\beta_m\Sigma_m^{-1})\Big] - n\sum_{m=1}^{k}\log(I_{d\times d}+K)(\Sigma_m\otimes\Sigma_m) - \log p(\mathbf{Y}|\theta) + \frac{c}{2}\log\kappa_c + \frac{c}{2}\Big\}, \tag{4.2}$$

其中, $\gamma, \eta, \nu, \delta$都为常数向量, β为常数矩阵, N为样本个数, d为数据的维数, $I_{d\times d}$为$d \times d$的单位阵, $K = \sum\limits_{ij} H_{ij} \otimes H_{ij}$, 而$H_{ij}$是$d \times d$的矩阵, 仅第$(i, j)$个元素为1, 其它为0.

但是, 在具体计算时, 由于有公式(2.2)的约束, 必须满足如下方程:

$$\widehat{\alpha}_m(t+1) = \max\left\{0, \left(\sum_{i=1}^{n} R_m^{(i)}\right) - \frac{N}{2}\right\} \Big/ \sum_{m=1}^{k} \max\left\{0, \left(\sum_{i=1}^{n} R_m^{(i)}\right) - \frac{N}{2}\right\}. \tag{4.3}$$

如果α_m为零, 则必须将$\widehat{\alpha}_m(t+1) = 0$所对应的成分删除. 为方便删除为零成分, 我们采用了CEM^{2}算法[13]. 该算法在每次更新完α_1和θ_1后, 重新计算后验概率; 更新完α_2和θ_2后, 再重新计算后验概率, 直到收敛. 由于可以任意删除成分, 可将类别k初始化为任意大小.

§5. 算　　法

根据第3节和第4节的分析, 我们给出无监督分类的具体算法如下:

1. 初始化最大和最小的类别数: $k_{\min}, k_{\max}$, 设置迭代终止的阈值ε, 初始化混合比例α_m、均值向量μ_m和方差矩阵Σ_m.

2. 令$t = 0$, $k = k_{\max}$, $L_{\min} = +\infty$.

3. 如果$k > k_{\min}$, 则:

   内循环计算每次迭代的最小信息长度(MML).

   (1) $t = t + 1$

   利用CEM^{2}算法, 根据公式(3.6)-(3.9)进行计算.

   - 设成分数comp $= 1$.
   - 当comp $< k$, 循环计算下面两步.
   - 计算R_{mn}, $\widetilde{\mu}_m$, $\widetilde{\Sigma}_m$和$\widetilde{\alpha}_m$.
   - 如果$\widehat{\alpha}_m(t+1) = 0$, 删除当前成分(重新分配$\widetilde{\mu}_m$, $\widetilde{\Sigma}_m$和$\widetilde{\alpha}_m$), 并令$k = k - 1$.

   (2) 利用公式(4.2), 计算描述程度$L(t)$.

   (3) 如果$|L(t-1) - L(t)| \geq \varepsilon |L(t)|$, 重复(1).

4. 如果$L(t) \leq L_{\min}$, 则$L_{\min} = L(t)$, $k_{\text{best}} = k$.

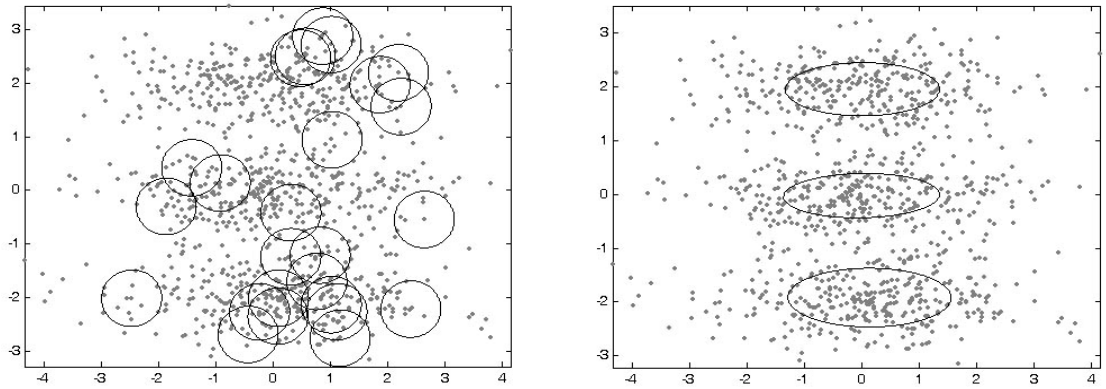
5. $k = k - 1$, 重复3.

其中, 终止的阈值ε的选取一般为10^{-5}; 混合比例为类别数的均匀分布; 均值向量μ的选取一般为数据均值; 方差矩阵Σ为数据方差的对角化.

§6. 实　　验

所有实验数据采用[4]中相同的数据, 以便将我们的实验结果与他们的实验结果进行对比. 第一个实验数据共900个, 是一个两维数据, 由3个成分混合而成. 其混合比例为: $\alpha_1 = \alpha_2 = \alpha_3 = 1/3$; 均值向量为: $\mu_1 = [0, -2]$, $\mu_2 = [0, 0]$, $\mu_3 = [0, +2]$; 方差矩阵为:

$C_1 = C_2 = C_3 = \mathrm{diag}\{2, 0.2\}$. 选定初始类别数为25, 迭代约80步后, 该算法收敛到了类别数3, 并给出了良好的分类结果(见图1).



(a) (b)

图1 (a) 初始25个类别; (b) 自动确定分类数3, 并给出分类结果

将该算法与Figueiredo和Jain的算法相比较, 我们的算法在估计分类数方面, 要更稳定. 下图是对实验数据运算50次后, 它们各自学习的分类数结果对比. 采用Jeffrey先验, 其确定的分类数正确率为98%, 而我们算法的正确率为100%.

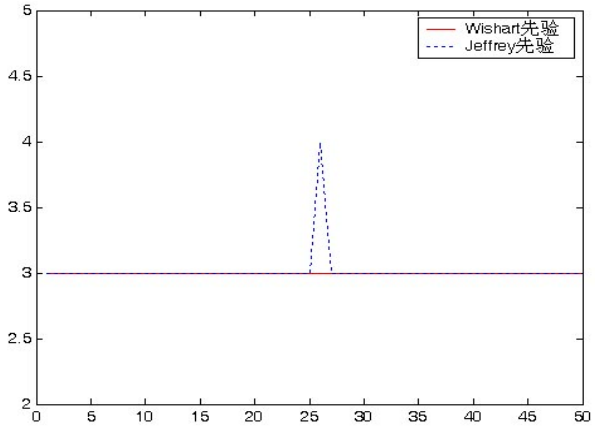


图2 分类数估计对比

第二个实验是考虑混合成分相互重叠的情况, 该实验数据共1000个样本, 也是一个两维数据, 但由4个成分混合而成, 其中有两个成分有相同的均值. 混合比例为:

$$\alpha_1 = \alpha_2 = \alpha_3 = 0.3, \qquad \alpha_4 = 0.1.$$

均值为:

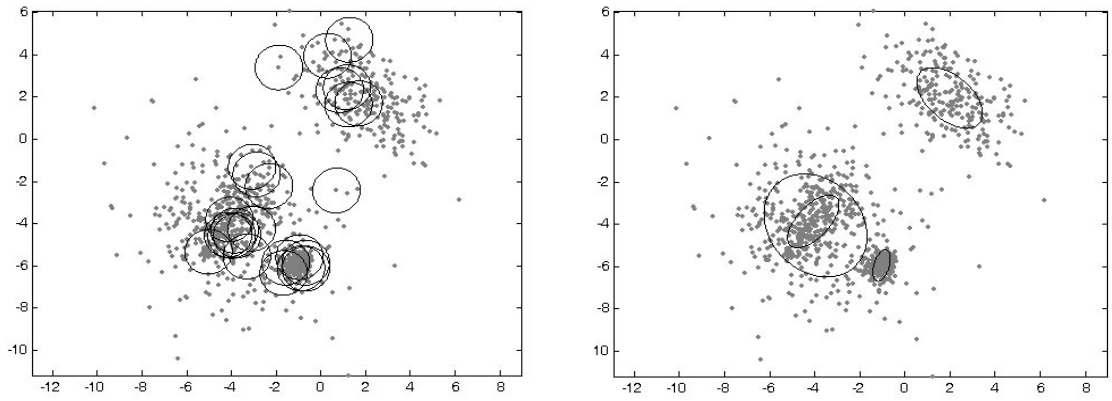
$$\mu_1 = \mu_2 = [-4, -4], \qquad \mu_3 = [2, 2], \qquad \mu_4 = [-1, -6].$$

协方差矩阵为:

$$C_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \qquad C_2 = \begin{bmatrix} 6 & -2 \\ -2 & 6 \end{bmatrix},$$

$$C_3 = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \qquad C_4 = \begin{bmatrix} 0.125 & 0 \\ 0 & 0.125 \end{bmatrix}.$$

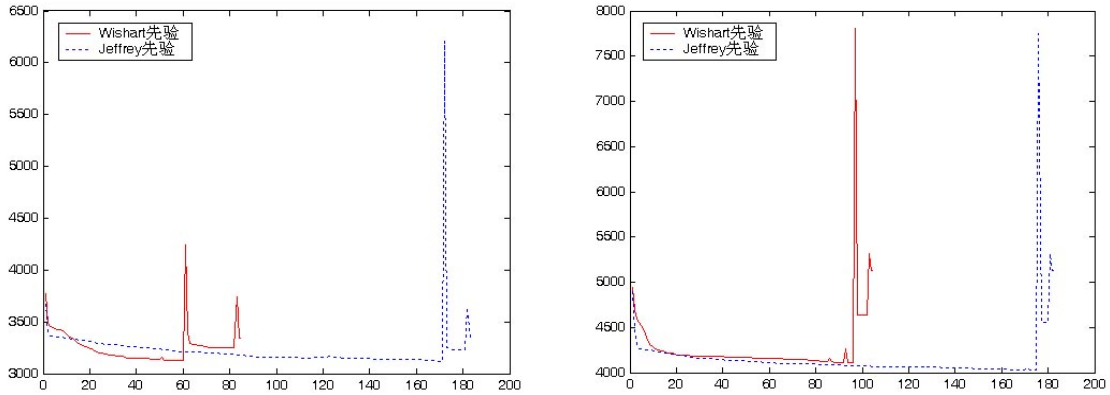
选定初始类别数为25, 迭代约100步后, 该算法收敛到了类别数4, 并给出了良好的分类结果(见图3).



(a)　　(b)

图3　(a) 初始25个类别; (b) 自动确定分类数4, 并给出分类结果

我们的算法除稳定性优于Figueiredo和Jain的算法外, 其计算效率也提高近1半. 下图是迭代次数的对比, 纵轴表示信息长度下降的过程.



(a)　　(b)

图4　(a) 第一个实验数据的迭代次数对比; (b) 第二个实验数据的迭代次数对比

通过以上的一些实验, 我们认为采用Wishart先验要比Jeffrey先验优越, 不仅能保证算法的稳定性, 而且在计算效率上也有很大提高.

§7. 结 论

本文讨论了高斯混合模型的无监督分类问题, 对于无监督分类, 其重点是成分数的确定. 成分数的确定从70年代到现在, 经过了30多年的发展, 到现在为止也没有找到一个普适的方法. 这里, 我们在最大惩罚似然的基础上, 结合MML规则, 给出了一个自动确定成分数的无监督分类方法. 该方法结合文献[4]和[6]的成果, 使用逆Wishart分布的对数代替Jeffery先验的对数作为惩罚项. 实验表明, 该方法在运算的速度和准确性方面比原有方法有较大提高.

参 考 文 献

[1] Day, N.E., Estimating the components of a mixture of normal distributions, *Biometrika*, **56(3)**(1969), 463–474.

[2] Dempster, A., Laird, N. and Rubin, D., Maximum likehood estimation from incomplete data via the EM algorithm, *J. Royal Statistical Soc. B*, **39**(1977), 1–38.

[3] Redner, R.A., Walker, H.F., Mixtures densities, maximum likelihood and the EM algorithm, *SIAM Review*, **26**(1984), 195–239.

[4] Figueiredo, M.A.T., Jain, A.K., Unsupervised learning of finite mixture models, *IEEE-PAMI*, **24(3)**(2002), 381–396.

[5] Ueda, N., Nakano, R., Gharhamani, Z. and Hinton, G., SMEM algorithm for mixture models, *Neural Computation*, **12**(2000), 2109–2128.

[6] Ormoneit, D., Tresp, V., Averaging, maximum penalized likelihood and Bayesian estimation for improving Gaussian mixture probability density estimates, *IEEE Transactions on Neural Networks*, **9(4)**(1998), 639–649.

[7] Oliver, J., Baxter, R. and Wallace, C., *Unsupervised Learning Using MML*, Proc. 13$^{\text{th}}$ Int'l Conf. Machine Learning, 364–372, 1996.

[8] Hathway, R., Another interpretation of the EM algorithm for mixture distributions, *Journal of Statistics & Probability Letters*, **4**(1986), 53–56.

[9] Ridolfi, A., Idier, J., *Penalized Maximum Likelihood Estimation for Normal Mixture Distributions*, Actes 17 Coll. GRETSI, Vannes, France, 259–262, 1999.

[10] Snoussi, H., M-Djafari, A., *Penalized Maximum Likelihood for Multivariate Gaussian Mixture, Bayesian Inference and Maximum Entropy Methods in Science and Engeering*, 21$^{\text{st}}$ International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engeering, Baltimore, Maryland, 77–88, 2001.

[11] Wallace, C.S. and Freeman, P.R., Estimation and inference by compact coding, *Journal of the Royal Statistical B*, **49**(1987), 240–252.

[12] Conway, J.H. and Sloane, N.J.A., *Sphere Packings, Lattices and Groups*, Springer-Verlag, London, 1988.

[13] Celeux, G., Chretien, S., Forbes, F. and Mkhadri, A., A component-wise EM algorithm for mixtures, *Technical Report 3746*, INRIA Rhone-Alpes, France, 1999. Available at http://www.inria.fr/RRRT/RR-3746.html.

Unsupervised Classification Based on Penalized Maximum Likelihood of Gaussian Mixture Models

Yu Peng

(*School of Mathematical Sciences, Peking University, Beijing, 100871;*
National Geomatics Center of China, Beijing, 100044)

Tong Xinwei

(*School of Mathematical Sciences, Beijing Normal University, Beijing, 100875*)

Feng Jufu

(*National Laboratory on Machine Perception, Center for Information Science,*
School of Electronics Engineering and Computer Science, Peking University, Beijing, 100871)

In this paper we propose an unsupervised classification algorithm which is based on Gaussian mixture models. Thinking that EM algorithm will result in a local optimal resolution of Gaussian mixture models in parameter estimations, we substitute invert Wishart distribution for Jeffery prior. Experiments show that this algorithm improves correct rates and decreases time while estimating classifications.

Keywords: Gaussian mixture models, unsupervised Classification, penalized maximum likelihood, EM algorithm, invert Wishart distribution.

AMS Subject Classification: 62G32.