Regularized Gaussian Mixture Model for High-Dimensional Clustering

Yang Zhao

摘要

寻找高维数据集的低维表示是各种应用中的一项重要任务。数据集通常包含嵌入在不同子空间中的簇这一事实对这项任务构成了障碍。由于需要能够同时进行聚类和查找每个簇的固有子空间的方法，在本文中, 我们提出了一种用于聚类的正则化高斯混合模型 (GMM)。尽管 GMM 具有概率解释和对观察噪声的鲁棒性等优点，但 GMM 的传统最大似然估计在高维设置中表现出令人失望的性能。提出的正则化方法找到分量协方差矩阵的低维表示，从而更好地估计局部特征相关性。正则化问题可以合并到期望最大化算法中，以最大化 GMM 的似然函数，修改 M 步以合并正则化。M 步涉及行列式最大化问题，可以有效地解决。使用多个模拟数据集证明了所提出方法的性能。我们还使用四个真实数据集说明了所提出方法在应用中的潜在价值。

关键词：降维；局部特征相关性；正则化；无监督学习

1 引言

我们的数据生成能力在过去十年中有了极大的提高。在生物医学工程、制造业、结构健康监测、计算机视觉等应用领域，复杂高维结构的海量数据出现了爆炸式增长。这些技术中的许多都依赖于观察到数据集的固有维数比环境空间的维数小得多，即高维数据存在低维表示。具有代表性的技术是主成分分析（PCA）。然而，如果数据集中不存在线性相关特征，PCA 的性能就会很差。特征选择技术也被提出来去除不相关的特征。然而，在实践中，数据点可以从不同的子空间中提取，特征的相关性可能随着子空间的不同而不同。因此，我们需要能够同时识别数据中的聚类并为每个聚类找到低维表示的方法。尽管高维数据在许多应用中都很有效，但它给基于模型的聚类方法带来了计算上的挑战。这些方法，包括 GMM，通常在高维数据上表现不佳，因为需要用相对较少的观测来估计大量参数。随着维数的增加，正则化高斯混合模型（GMM）中的参数数量呈二次增长，极大似然估计（MLE）问题很快变得不适定。由于克服这一困难的需要和低维子空间在一些应用中存在的事实，已经提出了一些方法。在本文中，我们提出了一种新的正则化方法来估计 GMM 的参数，它可以同时聚类和寻找每个聚类的子空间。

2 相关工作

2.1 构造生成数据的模型

在模拟研究中，我们从具有特定协方差结构的高斯分布的混合中生成样本。为了检验所提方法在不同维数设置下的性能，维数 D 在 10 到 100 之间变化。在我们的模拟中考虑了以下四个协方差矩阵结构。

1. 采用矩阵 Σ 的部分稀疏模型，矩阵的元素 σ_{ij}^2 由 $\sigma_{ij}^2 = 2{\times}0.8^{|i-j|}$ 决定。

2. 稠密模型定义为 $\Sigma_{(D)} = F_{(D)}F_{(D)}^T$，其中 D×D 对称矩阵为 $F_{(D)} = (1-\rho_{(D)})I_{(D)} + \rho_{(D)}J_{(D)}$。其中 J(D) 是 D×D 矩阵，$\rho_{(D)} \in (0,1)$ 依赖于 D。这是一种产生特征值的机制。第一个特征值比其他特征值大一个数量级。如果 $\rho_{(D)} \in (0,1)$ 是一个固定常数或逐渐减小到 0，那么 $\rho_{(D)} \gg D^{-1/2}$，这样第一个 PC 是一致的。这个模型是子空间基本假设成立的一个例子。在实验中，$\rho_{(D)}$ 作为常数固定在 0.3。

3. 块模型定义为 $\Sigma_{(D)} = F_{(D)}F_{(D)}^T$，是 2D×2D 的矩阵，$F_{(D)}$ 是对角矩阵。其中 $F_{(1,D)} = (1-\rho_{(1,D)}I_{(D)}) + \rho_{(1,D)}J_{(D)}$，$F_{(2,D)} = (1-\rho_{(2,D)}I_{(D)}) + \rho_{(2,D)}J_{(D)}$。假设 $0 < \rho_{(2,D)} < \rho_{(1,D)} < 1$，并且 $\rho_{(1,D)} \gg \rho_{(2,D)} \gg D^{-1/2}$，则前两个 PC 值是一致的。在实验中，$\rho_{(1,D)}$ 和 $\rho_{(2,D)}$ 分别由 $D^{-1/8}$ 和 $D^{-1/4}$ 决定。

4. 第一个特征值大于其他特征值的对角线模型，Σ 其中的第一个元素 σ_{11}^2 由 $(D\times\rho+1-\rho)^2$ 决定，其他元素 $\sigma_{ii}^2(i \neq 1)$ 由 $(1-\rho)^2$，在实验中，ρ 固定为 0.3。

在部分稀疏情况下，当我们远离对角线时，元素趋于零。在密集场景中，所有元素都是非零的。在区块模型中，第一个区块的协方差大于第二个区块。在对角线模型中，只有对角线元素是非零的，并且第一个元素的大小比其他元素大得多。

2.2 对模型进行评估

基于隐式场表达的三维重建方法.....

单协方差矩阵估计: 我们首先考虑这样一种情况，即底层分布中只有一个分量，在这种情况下，估计问题被简化为估计单个高斯分布。我们将所提出的方法（正则化估计）与无约束 MLE 以及稀疏估计进行了比较。考虑了两种具有代表性的稀疏估计：glasso 和 YL。每个模型的性能通过 KL 损失进行评估：$KL = -log|\hat{C}| + tr(\hat{C}\Sigma) - (-log|\Sigma^{-1}| + D)$。在整个模拟过程中，平均值固定为 O(D)，即 D 维空间中所有元素均为 0 的向量。每个模型生成 50 个实例。通过贝叶斯信息准则 (BIC) 对比较方法中涉及的参数进行调优。该方法使用 Kmeans 聚类算法给出的结果进行初始化。实验在 MATLAB 中进行，所有的结果都是超过 20 次运行的平均值。

多个高斯分布的混合: 采用以下四种指标来评估

1. 平均谱范数：$SL = (1/K)\Sigma_{k=1}^K \|\hat{\Sigma}_k - \Sigma_k\|$，其中 $\|A\|$ 是矩阵 A 的最大奇异值。

2. 平均 Frobenius 范数：$FL = (1/K)\Sigma_{k=1}^K \sqrt{\Sigma_{i,j}(\hat{\Sigma}_k(i,j) - \Sigma_k(i,j))^2}$

3. KL 散度: 假设 $P(X)$ 是 GMM 的真实分布，$Q(X)$ 是拟合分布。$KL(Q,P) = E\Big[logP(X) - logQ(X)\Big] = E\Big[log\Sigma_{k=1}^K \pi_k N(x|\mu_k,\Sigma_k) - log\Sigma_{k=1}^K \hat{\pi}_k N(x|\hat{\mu}_k,\hat{\Sigma}_k)\Big]$。由于 $KL(Q,P)$ 不存在封闭形式，我们使用 100 个自举样本来近似实验中的期望。

4. 通过匹配原始聚类的关系和聚类结果计算分类误差。

3 本文方法

3.1 本文方法概述

设 X 为数据集的协方差矩阵，Z 为每个实例对应的潜在变量。先验概率 $P(Z=j) = \pi_j, j \in \{1,...,K\}$。一个完整的数据集可以由 $X, Z = X_1, z_1, ..., X_i, z_i, ..., X_N, z_N$ 表示，其中 N 为数据点的数量。给定 $Z_i = j$，$X_i(i = 1,...,N) \in N(\mu_j, \Sigma_j)$，其中均值为 μ_j，协方差矩阵为 Σ_j。完整数据集的似

然函数采用 $P(X,Z|\pi,\mu,\Sigma)$。目标是最大化完整数据集的对数似然结果。引入指示变量：

$$Z_{ij}\begin{cases}1,if \quad z_i=j\\ 0,otherwise\end{cases}$$

对数似然函数为

$$logP(X,Z|\pi,\mu\,\Sigma)=\Sigma_{i=1}^{N}\Sigma_{j=1}^{K}z_{ij}\Big[log\pi_j+logN(X_i|\mu_j,\Sigma_j)\Big] \tag{1}$$

因为 Z 未知，所以使用 EM 算法，求对数似然的最大化。

3.2 正则化协方差矩阵估计

我们首先考虑单协方差矩阵估计问题。当对数似然函数 1中 $K=1$ 时，简化为单高斯分布的情况，关于 μ 和 Σ^{-1} 的似然函数可以表示为

$$logP(X|\mu,\Sigma)=Nlog|C|-\Sigma_{i=1}^{N}(X_i-\mu)^TC(X_i-\mu) \tag{2}$$

其中，$C=\Sigma^{-1}$，是精度矩阵。Σ 的最大似然估计 MLE 是 $A:(1/N)\Sigma_{i=1}^{N}(X_i-\overline{X})^T(X_i-\overline{X})=(1/N)(X-e\overline{X}^T)^T(X-e\overline{X}^T)$，其中 e 是每个元素都是 1 的列向量。然而，随着维度的增加，这种估计就变得不合理。

为了能够在高维空间中估计协方差矩阵，我们采用了正则化的思想。我们首先旋转坐标系统，以消除特征依赖性，并找到主成分（PCs）。因此，新坐标系中的协方差矩阵是对角线的。然后，我们将这个新的协方差矩阵的逆正则化，将其约束为正定，并迫使前 q 个 pc 解释大部分的变化。为了找到 pc，对中心输入矩阵进行奇异值分解（SVD）。样本协方差矩阵为 $\widehat{\Sigma}=U(S^TS)U^T$，其中 U 是原坐标系的线性变换。所以在旋转坐标系中最大对数似然是：

$$min-log|\widetilde{C}+tr(\widetilde{C}(S^TS))| \tag{3}$$

其中，$tr()$ 是矩阵的迹，$\widetilde{C}$ 是新坐标系下的逆协方差矩阵。有以下两个正则化限制：

$$tr(\widetilde{C}\leqslant t_1);\Sigma_{i=q+1}^{D}\widetilde{c}_{ii}-\Sigma_{i=1}^{q}\widetilde{c}_{ii}\geqslant t_2$$

其中，D 是数据矩阵 X 的维度，$\widetilde{C}ii$ 是 $\widetilde{C}$ 的第 i 个对角元素，q 是 PCs 的数量，表示每个集群低维嵌入的维度，t_1、t_2 是正常数。第一个式子确保了正定矩阵，两个式子控制了前 q 个主成分 PC 中保存的信息量。

3.3 改进 GMM 的 EM 算法

正则化协方差估计 (3) 的对数似然函数 (1) 中的参数可以用 EM 算法[1][2]逼近。完整数据对数似然的期望是：

$$\begin{aligned}Q(\theta;\theta^{old})&=E_{z|X,\theta^{old}}\Big[logP(X,Z)\Big]\\&=\Sigma_{i=1}^{N}\Sigma_{j=1}^{K}E_{z|X,\theta^{old}}\Big[z_{ij}\Big]\Big[logpi_j+logp_j\Big(X_i|\theta_j\Big)\Big]\end{aligned} \tag{4}$$

其中，θ 是所有模型变量的集合，p_j 是参数 θ_j 的混合分量 j 的高斯密度函数。在 E-step 中，对 z_{ik}

的估计：

$$\begin{aligned}\widehat{z}_i k &= E_{z|X,\theta^{old}}\Big[z_{ij}\Big] \\ &= P(z_{ij}=1|X,\theta^{old}) = \frac{pi_k p_k\Big(X_i|\theta_k\Big)}{\Sigma_{j=1}^{K}\pi_j p_j\Big(X_i|\theta_j\Big)}\end{aligned} \tag{5}$$

在 M-step 中，先验概率 pi、平均值 μ 和协方差矩阵 Σ 分别为：

$$\pi_k^{(t+1)} = \frac{1}{N}\Sigma_{i=1}^{N} z_{ik}^{(t)} \tag{6}$$

$$\mu_k^{(t+1)} = \frac{\Sigma_{i=1}^{N} z_{ik}^{(t)} X_i}{\Sigma_{i=1}^{N} z_{ik}^{(t)}} \tag{7}$$

$$\Sigma_k^{(t+1)} = U_k^{(t)}\Big(C_k^{(t+1)}\Big)^{-1} U_k^{(t)^T} \tag{8}$$

其中，

$$C_k^{(t+1)} = \underset{\widetilde{C}\in\Omega}{\operatorname{argmin}}\left[-\log|\widetilde{C}| + \frac{1}{N}]tr\Big(\widetilde{C}\Big(S_k^{(t)^T} S_k^{(t)}\Big)\Big)\right],$$

$$\Omega = \Big\{\widetilde{C}|tr\Big(\widetilde{C}\Big) \leqslant t_1, \sum_{i=q+1}^{D}\widetilde{c}_{ii} - \sum_{i=1}^{q}\widetilde{c}_{ii} \geqslant t_2\Big\},$$

$$\frac{1}{N}S_k^{(t)^T} S_k^{(t)} = U_k^{(t)}\left[\Sigma_{i=1}^{N}\left(\frac{z_{ik}^{(t)}}{\Sigma_{j=1}^{N} z_{ik}^{(t)}}\right)\times\Big(X_i - \mu_k^{(t+1)}\Big)\Big(X_i - \mu_k^{(t+1)}\Big)^T\right] U_k^{(t)^T}$$

在 E-step 中，计算后验概率 z_{ik} 需要进行 O(KND) 操作。在 M-step 中，计算参数 π_k 和 μ_k 需要 O(KN) 次操作。协方差矩阵 Σ_k 的估计涉及到 SVD 和行列式最大化的额外计算，需要 O($D^3N + D^2$) 次运算。假设正则化 GMM 迭代 t 次后收敛，则整体计算复杂度为 O($tKD^2(1 + DN)$)。可以看出，维数在计算中占主导地位，特别是对于高维数据集。

4 复现细节

4.1 与已有开源代码对比

没有参考任何代码，伪代码如下：

Procedure 1 Proposed Regularized GMM Algorithm

Input: $\boldsymbol{X}=\{\boldsymbol{X_1},\boldsymbol{X_2},...,\boldsymbol{X_N}\}\subset\mathbb{R}^{\boldsymbol{D}}$, the data set $\boldsymbol{X}$with size $N\times D$; K, the number of clusters; t_1 and t_2 , the regularization parameters; q, the dimensionality of the low-dimensional embedding for each cluster; δ , the termination condition value.

Output: $\pi_1,...,\pi_k$;$\{\boldsymbol{\mu_1},\boldsymbol{\Sigma_1}\},...,\{\boldsymbol{\mu_k},\boldsymbol{\Sigma_k}\}$.

Initialize the parameter $\boldsymbol{\Theta_0}$by K-means clustering algorithm

$t\leftarrow 0$;

repeat E-step

Compute the posterior probabilities, i.e. the ‘membership weight’ of data point x_i in mixture component k;

$$\boldsymbol{P}(\boldsymbol{z_{ik}}=\boldsymbol{1}|\boldsymbol{X},\boldsymbol{\Theta})=\frac{\boldsymbol{\pi_k^{(t)}p_k\left(X_i|\theta_k^{(t)}\right)}}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{K}}\boldsymbol{\pi_j^{(t)}p_j\left(X_i|\theta_j^{(t)}\right)}}$$

M-step

Update the prior probability, mean and covariance matrix for mixture component k;

$$\boldsymbol{\pi_k^{(t+1)}}=\frac{\boldsymbol{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}\boldsymbol{z_{ik}^{(t)}};$$

$$\boldsymbol{\mu_k^{(t+1)}}=\frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}\boldsymbol{z_{ik}^{(t)}X_i}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}\boldsymbol{z_{ik}^{(t)}}};$$

$$\boldsymbol{\Sigma_k^{(t+1)}}=\boldsymbol{U_k^{(t)}}(\boldsymbol{C_k^{(t+1)}})^{-\boldsymbol{1}}\boldsymbol{U_k^{(t)^T}};$$

where

$$\boldsymbol{C_k^{(t+1)}}=\underset{\widetilde{\boldsymbol{C}}\in\boldsymbol{\Omega}}{\textbf{argmin}}\left[-\log|\widetilde{\boldsymbol{C}}|+\frac{\boldsymbol{1}}{\boldsymbol{N}}]\boldsymbol{tr}\left(\widetilde{\boldsymbol{C}}\left(\boldsymbol{S_k^{(t)^T}S_k^{(t)}}\right)\right)\right],$$

$$\boldsymbol{\Omega}=\left\{\widetilde{\boldsymbol{C}}|\boldsymbol{tr}\left(\widetilde{\boldsymbol{C}}\right)\leqslant\boldsymbol{t_1},\sum_{\boldsymbol{i=q+1}}^{\boldsymbol{D}}\widetilde{\boldsymbol{c}}_{\boldsymbol{ii}}-\sum_{\boldsymbol{i=1}}^{\boldsymbol{q}}\widetilde{\boldsymbol{c}}_{\boldsymbol{ii}}\geqslant\boldsymbol{t_2}\right\},$$

Evaluate the regularized log likelihood:

$$\boldsymbol{L}\left(\boldsymbol{\Theta^{(t+1)}}\right)=\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}\log\boldsymbol{p}\left(\boldsymbol{X_i}|\boldsymbol{\Theta^{(t+1)}}\right)=\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}\left(\log\sum_{\boldsymbol{j=1}}^{\boldsymbol{K}}\boldsymbol{\pi_j^{(t+1)}p_j(X_i|\theta_j^{(t+1)})}\right)$$

$t\leftarrow t+1$;

until $\boldsymbol{L(\Theta^{(t+1)})-L(\Theta^t)\leqslant\delta}$

return $\boldsymbol{\Theta^{(t+1)}}$

4.2 实验环境

Matlab 软件，以及 Statistics and Machine Learning Toolbox 工具包

4.3 实现流程

先初始化，用 k-means 聚类算法对样本进行聚类，利用各类的均值作为 μ，初始的协方差矩阵 Σ 来源于所有原始数据的协方差矩阵，π 自定义，保证 $\Sigma_{i=1}^{K}\pi_i=1$。再利用 EM 算法，不断迭代，求出 argmin，进而求出 Σ_k。

5 实验结果分析

将该方法应用于基于乘客时间出行行为的地铁乘客聚类。该数据来源于深圳地铁股份有限公司的自动售票系统，包括 2013 年 8 月 1 日至 11 月 30 日 121 天内 1 万名乘客在地铁站的智能卡交易。其

中，每个乘客的时间出行属性用该时间段内每 10 分钟的 tap-in 或 tap-out 次数表示，该数据集的大小为 10000 × 17424。

我们考虑找到两个隐藏的地铁乘客群，(a) 是第一组乘客。(b) 是第二组乘客。对于每个聚类，我们计算并绘制了深圳地铁站 24 小时内乘客的时间分布。从图 1(a) 和 (b) 中可以看出，本文提出的聚类方法能够有效地找到具有不同出行模式的乘客群体：第 1 组具有更明显的早高峰 (从早上 6:30 到 7:30)。而组 2 的早晚峰分布更均匀。

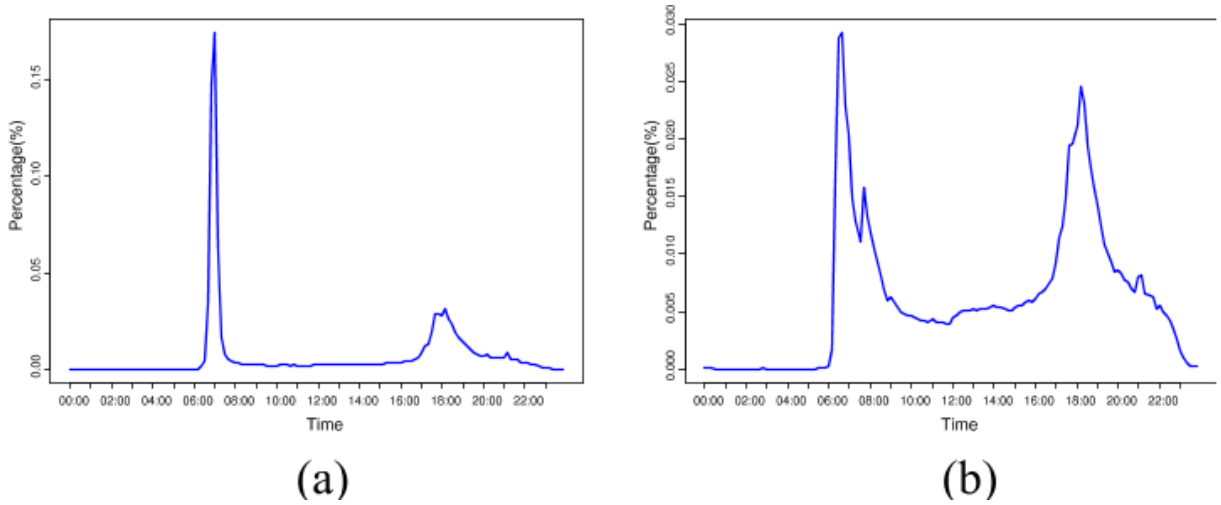

图 1: 实验结果示意

6 总结与展望

在本文中，使用了一种新的正则化方法来估计 GMM 的参数。将正则化问题转化为确定最大化问题后，可以有效地求解正则化问题的最优解；正则化方法可纳入 EM 算法中，使 GMM 参数的对数似然值最大化。我们使用真实数据集来评估所提出方法的性能。结果表明，该方法在高维空间中具有一定的适用性。这么多年以来，许多专业人士提出了各式各样的聚类方法，目的是为了适应数据的复杂性和多样性，针对这方面的工作还需要继续进行。

参考文献

[1] 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 series b-methodological, 1977.

[2] WU C F J. ON THE CONVERGENCE PROPERTIES OF THE EM ALGORITHM[J]. Annals of Statistics, 1983.