浙江工业大学

本科毕业设计说明书（论文）

（2013 届）

论文题目　基于变分贝叶斯方法的医学图像分割

作者姓名　　华俊豪

指导教师　　陈胜勇

学科(专业)　计算机+自动化0901

所在学院　　计算机科学与技术学院

提交日期　　2013 年 5 月 25 日

摘要

脑部磁共振图像分割是医学图像处理中的重要内容。在众多分割算法中，广为研究的是基于高斯混合模型的聚类分割技术。本文针对此问题，系统地研究了基于变分贝叶斯的混合模型，并将其应用于脑部图像分割。

变分贝叶斯（VB）作为融入先验信息的确定性推理方法，其基本思想是在平均场假设下，根据变量求和函数的共轭对偶性，把变量求和推理问题转化为关于自由分布的泛函极值问题，并通过近似求解泛函极值计算出目标函数的上下界。

本文第一部分阐述变分贝叶斯方法的数学基础、理论依据以及将贝叶斯网络框架下的变分信息传播。特别地与 EM 算法比较，分析其算法性能，认为作为一种分布估计方法，VB 算法由于融入了先验信息，估参性能更优。

第二部分详细推导了三类混合模型的变分分布以及相应的概率图模型，包括高斯混合模型，学生 t 有限混合模型以及基于 Dirichlet 过程混合的无限混合模型，并通过 MATLAB 仿真将其应用于脑部 MR 图像分割（IBSR 数据集）。实验表明，虽然 VB 没有显著提高精度, 但由于变分推断大幅度减少迭代次数，因此 VB 优于 EM 算法。

为提高分割精度，本文提出一种基于 laplacian 图的变分混合模型。其基本思想是假定产生数据的概率分布在其周围空间中拥有流形子结构。如果两个点在概率分布的流形空间中相接近，那么其条件概率分布也是相似的。用最近邻图构造数据的流形结构，用该流形结构平滑条件后验概率，即将 laplacian 正则项加入到以上各混合模型的损失函数（极大似然函数）中。同样将其应用于脑部 MR 图像分割。结果表明 laplacian 正则化的变分混合模型在精度与稳定性上都表现更优。

关键词： 变分贝叶斯，聚类，混合模型，laplacian 图，脑图像分割

Abstract

Brain magnetic resonance image (MRI) segmentation is one of the most important tasks in medical image processing, due to its key role in the subsequent image analysis process. Among numerous medical image segmentation algorithms, Gaussian mixture model-based clustering segmentation technique is widely studied. This paper systematically studied and proposed the improved mixture model based on variational Bayesian methods and Laplacian graph, and then applied it to the brain magnetic resonance image segmentation.

The variational Bayes (VB) is a prior-considered deterministic method, whose main idea is to transform the sum and inference variables problem into solving extreme value problem of free distribution by approximately calculating functional upper or lower bound of the objective function under the mean field assumption. This paper presents the mathematical foundation and theoretical supports for VB, and then proposes the variational message passing (VMP), a general purpose algorithm for applying variaitonal inference to a Bayesian Network. Furthermore, the performance analysis between the VB methods and the Expectation Maximization is then followed. As a prior-considered distributional approximation method, we conclude that variational Bayes performs better than EM in parameters estimation.

In the second part of this paper, the variational Bayesian method is applied to the parameters estimation of mixture models, including Gaussian mixture, finite student's t-mixture model and infinite student's t-mixture model based on Dirichlet process mixture. All presented models are used in the segmentation of brain MR images. Simulation results on real brain data show that the variational inference method can greatly reduce the number of iterations. But unfortunately, the accuracy is not significantly improved.

In order to improve segment accuracy significantly, we proposed an improved variational mixture model algorithm based on Laplacian graph. We consider the case where the probability distribution that generates the data is supported on a submanifold

of the ambient space. It is natural to assume that if two points are close in the intrinsic geometry of the probability distribution, their conditional probability distributions are similar. Manifold structure of data is modeled by a nearest neighbor graph, and it is incorporated in the maximum likelihood objective function, which is realized by adding the regularized term into the loss function of above models. It was found that the condition posterior probability is smoothed under this manifold structure. Simulation results on brain MR image segmentation show that the variational Bayesian-based laplacian regularized mixture model has a good performance in the segment accuracy and robustness.

Keywords： variational Bayes, clustering, mixture model, Laplacian graph, MR image segmentation

目录

图目录

表目录

第一章 绪论

1.1. 医学图像分割的研究背景和意义

随着工农业，医学等领域自动化和智能化需求的迅猛发展，对图像处理技术的要求也越来越高。图像的自动识别与理解是图像处理技术中的非常重要的任务，而图像分割技术则是图像识别自动化的关键技术。图像分割就是按照一定的规则将一幅图像分成若干部分或子集的过程。

医学图像分割是计算机辅助诊断和治疗计划制定中非常重要的工作，是医学图像处理与分析的一个重要领域，同时也是计算机辅助诊断与治疗的基础。比如在临床手术中需要从三维医学图像得到解剖结构或病理组织的精确三维模型，进行病理或正常组织的量化研究，这就需要分割算法能够从 MRI 或 CT 等图像中分离出解剖结构或把受损组织的位置和形状分割出来。又比如，为了便于对大脑的定量分析，需要将 MR 图像分割成白质、灰质和脑脊髓三种脑组织。由于手工分割对操作者的依赖性很强，费时费力。因此，研究自动或半自动的图像分割算法非常重要。

图像分割是机器视觉及多媒体应用技术中最困难的问题之一。迄今为止，学者们提出多种分割算法，但大多算法都是针对具体问题而言的，尚无一种适合各种图像的通用算法，也不存在判断分割是否正确的客观标准。可见，在未来几年，图像分割问题仍将是研究热点。

基于聚类分析的图像分割方法是中一类应用相当广泛的算法图像分割算法。而聚类算法中，高斯混合模型无疑是最重要的模型之一，一般采用 EM 算法估计参数。变分贝叶斯方法作为一种可以融合先验知识的参数估计方法，近年来受到了广泛关注，其通常可以取得比 EM 算法更好的估计结果。此外，结合流形结构的混合模型也取得一定的成果。本课题正是在此背景下，对 laplacian 正则化的变分混合模型展开研究，并应用于对脑部 MR 图像的分割问题上。

1.2. 图像分割方法综述

1.2.1. 常用的医学图像分割方法

在医学图像分割实现半自动和自动分割的发展过程中，大量图像分割算法涌入到医学图像领域中。常用的图像分割算法大体可分为基于区域和基于边缘检测的分割方法。

（1）基于区域的分割方法

图像分割通常会用到不同对象间特征的不连续性和同一对象内部的特征相似性。基于区域的算法侧重于利用区域内特征的相似性。典型代表包括阈值法，Watershed 算法，Mean-Shift 算法，区域增长算法支持向量机,期望最大化算法等。

阈值法是最常用的并行的直接检测区域方法。P-tile 法[2]是早期的基于灰度直方图的自动阈值选择方法，它假设在亮背景中存在一个暗目标并且已知目标在整幅图像中所占面积比为 p%，该方法选择阈值的原则是依次累计灰度直方图，直到该累计值大于或等于目标物所占面积。另外还有其他经典方法，如双峰法[3]，融入信息论的一维灰度直方图熵法[4]。

区域生长和分裂合并是两种典型的串行区域分割方法。其基本思想是将具有相似性质的像素集合起来构成区域。区域增长方式的优点是计算简单，但缺点是需要人工交互获取种子点。

（2）基于边缘的分割方法

边缘分割法利用边缘灰度、色彩、纹理这些不连续的位置信息，采用边缘检测算子检测图像边缘。为了达到理想的边缘检测分割效果，应将把边缘合并为边缘链，使之与图像边界相对应。

微分算子法是一种利用相邻区域的像素不连续性检测边缘点的方法。比如一阶 Roberts 算子，其边缘定位准，但对噪声敏感，适用于边缘明显且噪声较少的图像分割；Prewitt 算子对噪声有抑制作用，抑制噪声的原理是通过像素平均，但是像素平均相当于对图像的低通滤波。另外，还包括 Sobel 算子，Isotropic Sobel 算子，Laplacian 算子等。近年来还提出了基于曲面拟合的方法，基于边界曲线拟

合方法，基于反应-扩散方程的方法，串行边界查找，基于形变模型的方法等。

此外，随着模式识别，机器学习等的发展，统计方法，人工神经网络，分形和小波变换等也应用到图像分割中。

1.2.2. 基于聚类分析的图像分割算法

在众多的分割算法中，基于聚类分析的图像分割方法是图像分割中一类极其重要和应用相当广泛的算法。聚类分析以相似性为基础，聚类数据由若干模式组成的，通常模式是一个度量的向量，或者是多维空间中的一个点，聚类算法通过特征空间中点的相似度进行迭代划分。

（1）常用聚类算法

针对特定的问题，学者们提出了很多具有代表性的聚类算法，这些算法可以分为基于划分的方法，基于层次的方法和基于密度的方法。

基于划分的聚类算法是相当重要的一中聚类算法，它将数据点集分成k个划分，每个划分为一类，从k个初始划分出发，通过反复执行控制策略使某个准则最优化，最著名的当属 K-means 算法，它用质心来代表聚类中心，而 K-medoids 算法则由该聚类中最靠近中心的一个对象来表示。基于层次的聚类算法主要包括 BIRCH 算法（平衡迭代削减聚类法）和 CURE 算法（使用代表点的聚类方法）等，基于密度的方便包括 DBSCAN 算法，DBCLIQUE 算法（自动子空间聚类算法）。

传统的聚类分析是一种硬划分，它把每个待辨识的对象严格地划分到某个类中。但实际上大多数对象 并没有严格地属性，因而根据 Zadeh 提出模糊集理论，人们开始用模糊的方法处理聚类问题。最著名的是模糊 C-均值聚类(FCM, Fuzzy-C Means)算法，受到了极其广泛的应用。

（2）基于聚类分析的医学图像分割

聚类分析应用于医学图像分割，开始于对核磁共振图像分割，主要是脑脊髓、灰质和脑白质的分割。基本的聚类算法包括 C-means，ISODATA。但其难以处理有噪音或灰度不均匀的图像，鲁棒性比较差，时空效率较低。大多数算法使用图

像灰度特征做聚类，由于特征差异比较小，因而分割易出现重叠区域。

近年来，基于聚类分析广泛应用于针对不同组织器官的图像分割。傅景广等将遗传算法应用到图像分割[12]；廖亮等将 Markov 随机场与改进的模糊核聚类相结合，提出鲁班性更强的图像分割算法[14]；现在越来越多地将描述模型与随机模型相结合的模型，其中比较典型的方法是马尔科夫链蒙特卡洛方法（Markov Chain Monte Carlo）[17]。另外，还包括密度聚类以及基于爬山法[13]的医学图像组织分割。日前，医学图像分割研究有一定的成果，但依然不够完善，无法真正达到对医学图像的分析和理解目的。

（3）基于混合模型的聚类分割

在医学图像处理中，基于聚类分析的高斯混合模型用的相当普遍，实践证明高斯混合模型能够较好地描述医学图像信息，另外还包括采用学生 t 混合模型聚类的，也能取得较好的效果。

早在一百多年以前，人们就开始对混合模型进行研究，皮尔逊在 1894 年用具有两个混合元的正态混合模型对一组数据进行了拟合，用矩估计的方法进行模型参数估计。随后很长一段时间都未得到发展，直到 1972 年 Tan 和 Robertson 等人开始用极大似然法研究混合模型，更重要的是，他们证明了极大似然的优越性[15]。Dempster 等人在 1977 年用 EM 算法对极大似然估计进行计算使得计算机困难得到解决。但该方法只能在混合元个数已知的假定下，才能对参数进行估计，C.E. Rasmussen 提出无限高斯混合模型[17]，通过一种依靠 Gibbs 抽样的自由参数马尔科夫链的方法推断模型。

就基于高斯混合模型的 MR 图像分割，学者们已经提出了许多改进算法。Hayit Greenspan[5]等同时考虑图像的空间信息和灰度信息的 Constrained 高斯混模型(CGMM)，用(I,X,Y,Z)来表示 MR 图像的特征，能比单纯考虑灰度 I 取得更好的分割结果。ZeXuan Ji, Yong Xia[6]将局部模糊(Fuzzy Local)概念引入高斯混合模型中,认为每个体素与其周边的体素点能构成高斯混合模型，该算法通过极小化对象的能量函数来最大化后验概率，从而估计高斯混合模型的各参数，该方法应用到 MR 图像中，能较好地处理噪音，从而使分割具有更高的精度。

1.3. 本文主要研究内容与结构安排

针对高斯混合模型应用在脑部 MR 图像分割上的缺点，对采用变分贝叶斯估参的混合模型进行深入的探索和研究。全文由五章组成，各章内容安排如下：

第一章为绪论部分，介绍医学图像分割背景与发展趋势，特别地介绍基于聚类的图像分割的发展现状与存在的问题。

第二章首先概述变分贝叶斯理论的发展过程，接着详细阐述近似变分推理方法的概念、数学基础，计算方法以及计算复杂性。

第三章在变分推断的理论框架下，推导变分贝叶斯高斯混合模型，变分贝叶斯有限学生 t 混合模型以及基于 Dirichlet 过程的无限混合模型，并将其应用于脑部 MR 图像分割中，分析各模型的优劣。

第四章提出一种基于 laplacian 图正则化的变分混合模型，同样将其运用到脑部 MR 图像分割中，结果表明该明显改善了图像分割效果，提高了图像分割的鲁棒性。

第五章对本文工作进行总结，展望尚待研究的问题。

第二章 变分贝叶斯方法

2.1 变分贝叶斯理论的提出

1763 年英国学者 Thoms Bayes 的一篇《论机会学说中一个问题的求解》打开了贝叶斯统计推理理论的大门。自上世纪中期以来，贝叶斯理论得到迅速发展，英国统计学家 Lindely 称二十一世纪必将是贝叶斯统计的时代。贝叶斯统计理论在计算机科学，生物信息学，经济预测与决策和信号处理等领域有着极其广泛的应用。

贝叶斯推理过程很简单，而且概率形式很优美，但是在很多实际应用中，边缘似然函数的计算需要对所有的参数进行积分，该积分一般是在高维空间中进行的，因而比较困难。最简单的可以用最大似然函数法来进行参数估计。该方法假设参数是一个确定的常量，并且忽略参数先验，直接求取参数的似然函数的最大值，因而得到的是参数的点估计。实际上，贝叶斯要求融入先验信息。于是便有了最大后验概率法(MAP, Maximum a posterior)，这是最简单的贝叶斯方法。但以上两种方法都只是点估计，忽视了积分的作用，容易导致过拟合等现象。可选择的方案是采用边缘似然函数，因为其在计算过程中，对未知参加进行了积分。然而，由于该积分过程同样是高维且复杂的，很难解析求出。就计算边缘似然函数的近似方法成为人们研究的热点课题。

可以在参数的最大后验估计点处作一个高斯近似，这便是拉普拉斯近似方法[18]。但如果数据集较小，高斯假设的近似结果会比较差。蒙特卡洛方法是另一种经典的近似计算方法，该方法在信号处理，统计学，计算机视觉，机器学习和物理学等领域应用十分广泛。但该方法的准确性建立在大数据样本的基础上，因而计算量会很大。特别是当参数的先验概率分布比较复杂，参数维数比较高时，传统的蒙特卡洛方法计算速度很慢，采样困难。于是人们提出了马尔科夫链蒙特卡洛（MCMC）方法[19,20]，即通过构造马尔科夫链的极限不变分布对高维积分计算

进行模拟。

此外还有一种近似方法，可以对所有种类的复杂计算进行简化，那便是变分近似方法。其基本思想是用一个 tractable 的概率分布来近似真实分布。通过调整变分参数 λ 使得简化问题与原问题的距离函数最小化，一般采用 KL 散度来描述概率分布的距离特性，然后通过变分法能得到严格地上界或下界。早在 20 世纪 70 年代 Rustagi 便将变分法应用到统计学中。随后该思想被应用到神经网络中，其中 Mackay 又把这个思想应用于贝叶斯参数估计和模型比较过程中[21]。把变分近似应用到贝叶斯推理学习过程中，于是就形成了变分贝叶斯学习，也称集成学习[22]。

变分推理方法是一种确定性推理方法，其基本思想是用一个简单模型替代真实模型，并建立两个模型的相异度模型，通过泛函求极值最小化相异度模型，通过得到观测模型的上下界[23,24]。变分推理框架下，精确变分问题一般很难解，可以通过约束自由分布结构的变分策略，如平均场方法，使得问题得到简化。另外，可采用变分消息传播（Variational Message Passing）算法求解变分优化问题[25,26]。

2.2 变分贝叶斯数学推导与平均场估计

变分贝叶斯主要应用于复杂的统计模型中，这种模型一般包括三类变量：观测变量，未知参数和隐变量。在贝叶斯推断中，参数和隐变量统称为不可观测变量。变分贝叶斯方法主要是两个目的：

（1）近似不可观测变量的后验概率，以便通过这些变量作出统计推断。

（2）对一个特定的模型，给出观测变量的边缘似然函数（或称为证据，evidence）的下界。主要用于模型的选择，认为模型的边缘似然值越高，则模型对数据拟合程度越好，该模型产生数据的概率也越高。

对于第一个目的，蒙特卡洛模拟，特别是用 Gibbs 取样的 MCMC 方法，可以近似计算复杂的后验分布，能很好地应用到贝叶斯统计推断。此方法通过大量的样本估计真实的后验，因而近似结果带有一定的随机性。与此不同的是，变分贝叶斯方法提供一种局部最优，但具有确定解的近似后验方法。

从某种角度看，变分贝叶斯可以看作是 EM 算法的扩展，因为它也是采用极大后验估计，即用单个最有可能的参数值来代替完全贝叶斯估计。另外，变分贝叶斯也通过一组相互依然的等式进行不断的迭代来获得最优解。

2.2.1 问题描述与分析

考虑以下问题：已知一组观测数据 D，估计某个模型的参数与潜变量（或不可观测变量）$Z=\{Z_1,...,Z_n\}$ 的后验分布 $P(Z|D)$。

正如上文所描述的后验概率的形式通常是很复杂(Intractable)的，对于一种算法如果不能在多项式时间内求解，往往不是我们所考虑的。因而我们想能不能在误差允许的范围内，用更简单、容易理解的数学形式 $Q(Z)$ 来近似 $P(Z|D)$，即 $P(Z|D)\approx Q(Z)$。

由此引出如下两个问题：

（1）假设存在这样的 $Q(Z)$,那么如何度量 $Q(Z)$ 与 $P(Z|D)$ 之间的差异性？

（2）如何得到简单的 $Q(Z)$？

对于问题一，幸运的是不需要重新定义一个度量指标。在信息论中，已经存在描述两个随机分布之间离的度量，即相对熵，或者称为 Kullback-Leibler 散度[32]。

对于问题二，显然可以自主决定 $Q(Z)$ 的分布，只要它足够简单，且与 $P(Z|D)$ 相似。然而不可能每次都人工给出一个与 $P(Z|D)$ 接近且简单的 $Q(Z)$，其方法本身已经不具备可操作性。所以需要一种通用的形式帮助简化问题。那么数学形式复杂的原因是什么？ Occam's razor 理论认为一个模型的参数个数越多，那么模型复杂的概率越大。此外，如果参数之间具有相互依赖关系(mutually dependent)，那么通常很难对参数的边缘概率精确求解。那么如何解决呢？其实统计物理学界很早就关注了高维概率函数与它的简单形式，并形成了平均场理论。简单讲就是：系统中个体的局部相互作用可以产生宏观层面较为稳定的行为。于是我们可以做出后验条件独立（posterior independence）的假设。即，

$$\forall i, p(Z|D)=p(Z_i|D)p(Z_{-i}|D) \qquad (2\text{-}1)$$

2.2.2 理论基础

（1）Kullback-Leibler 散度

在统计学中，相对熵对应的是似然比的对数期望，相对熵 $D(p\|q)$ 度量当真实分布为 p 而假定分布为 q 时的无效性。

定义 两个概率密度函数为 $p(x)$ 和 $q(x)$ 之间的相对熵定义为

$$D_{KL}(p\|q)=\sum_{x\in\mathbb{N}}p(x)\log\frac{p(x)}{q(x)} \tag{2-2}$$

KL 散度有如下性质：

（1） $D_{KL}(p\|q)\neq D_{KL}(q\|p)$；

（2） $D_{KL}(p\|q)\geq 0$，当且仅当 $p=q$ 时为零；

（3）不满足三角不等式。

（2）平均场理论

平均场理论最早由统计物理学家提出。在量子多体问题中，用一个（单体）有效场来代替电子所受到的其他电子的库仑相互作用。这个有效场包含所有其它电受到的其他电子的库仑相互作用。这个有效场包含了所有其他电子对该电子的相互作用。利用有效场取代电子之间的库仑相互作用之后，每一个电子在一个有效场中运动，电子与电子之间的运动是独立的(除了需要考虑泡利不相容原理)，原来的多体问题转化为单体问题。

数学上说，平均场的适用范围只能是完全图，或者说系统结构是 well-mixed，在这种情况下，系统中的任何一个个体以等可能接触其他个体。反观物理，平均场与其说是一种方法，不如说是一种思想。其实统计物理的研究目的就是期望对宏观的热力学现象给予合理的微观理论。物理学家坚信，即便不满足完全图的假设，但既然这种“局部”到“整体”的作用得以实现，那么个体之间的局部作用相较于“全局”的作用是可以忽略不计的。

（3）变分法

变分法是处理泛函的数学领域，和处理函数的普通微积分相对。变分法最终

寻求的是极值函数：它们使得泛函取得极大或极小值。

泛函可以看成是函数概念的推广。泛函不同于复合函数，例如 $g=g\left(f\left(x\right)\right)$；对于后者，给定一个 x 值，仍然是有一个 g 值与之对应；对于前者，则必须给出某一区间上的函数 $y\left(x\right)$，才能得到一个泛函值 $J\left[y\right]$。

泛函取极值的必要条件

“当变量函数为 $y\left(x\right)$ 时，泛函 $J\left[y\right]$ 取极大值”的含义就是：对于极值函数 $y\left(x\right)$ 及其“附近”的变量函数 $y\left(x\right)+\delta y\left(x\right)$ ，恒有 $J\left[y+\delta y\right]\le J[y]$;

所谓函数 $y\left(x\right)+\delta y\left(x\right)$ 在另一个函数 $y\left(x\right)$ 的“附近”，指的是：

1. $\left|\delta y\left(x\right)\right|<\varepsilon$;
2. 有时还要求 $\left|\left(\delta y\right)'\left(x\right)\right|<\varepsilon$.

这里的 $\delta y\left(x\right)$ 称为函数 $y\left(x\right)$ 的变分。

Euler–Lagrange 方程

可以仿造函数极值必要条件的导出办法，导出泛函取极值的必要条件，这里不做严格的证明，直接给出。泛函 $J\left[y\right]$ 取到极大值的必要条件是一级变分 $\delta J\left[y\right]$ 为 0，其微分形式一般为二阶常微分方程，即 Euler-Largange 方程：

$$\frac{\partial F}{\partial y}-\frac{d}{dx}\frac{\partial F}{\partial y'}=0 \qquad (2\text{-}3)$$

泛函的条件极值

在约束条件 $J_0[y]=\int_{x_0}^{x_1}G(x,y,y')dx=C$ 下求函数 $J\left[y\right]$ 的极值，可以引入 Largange 乘子 λ，从而定义一个新的泛函，$\tilde{J}[y]=J[y]-\lambda J_0[y]$。仍将 δy 看成是独立的，则泛函 $\tilde{J}[y]$ 在边界条件下取极值的必要条件就是，

$$(\frac{\partial}{\partial y}-\frac{d}{dx}\frac{\partial}{\partial y'})(F-\lambda G)=0 \qquad (2\text{-}4)$$

2.2.3 问题求解

（1）目标函数

用简单分布Q来替代真实分布P，用 KL 散度来表示Q分布与P分布之间的距离：

$$D_{KL}(Q\|P)=\sum_{Z}Q(Z)\log\frac{Q(Z)}{P(Z|D)}=\sum_{Z}Q(Z)\log\frac{Q(Z)}{P(Z,D)}+\log P(D) \quad (2\text{-}5)$$

或者

$$\log P(D)=D_{KL}(Q\|P)-\sum_{Z}Q(Z)\log\frac{Q(Z)}{P(Z,D)}=D_{KL}(Q\|P)+L(Q) \quad (2\text{-}6)$$

由于对数证据$logP(D)$被相应的Q所固定，为了使 KL 散度最小，则只要极大化$L(Q)$。通过选择合适的Q使$L(Q)$便于计算和求极值。这样就可以得到后验$P(Z|D)$的近似解析表达式和证据（log evidence）的下界$L(Q)$，又称为变分自由能（variational free energy）。

于是得到目标函数：

$$\max\{L(Q)\}=\sum_{Z}Q(Z)\log P(Z,D)-\sum_{Z}Q(Z)\log Q(Z)=E_{Q}[\log P(Z,D)]+H(Q) \quad (2\text{-}7)$$

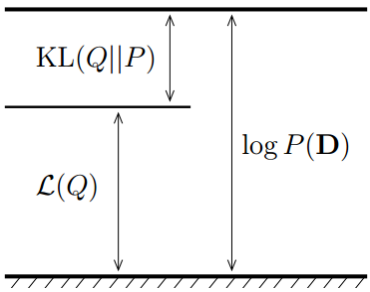


图 2-1 变分自由能与对数证据之间的关系

（2）平均场估计

根据平均场理论，变分分布$Q(Z)$可以通过参数和潜在变量的划分因式分解，比如将 Z 划分为$Z_1 \ldots Z_M$,则变分分布可写成，

$$Q(Z)=\prod_{i=1}^{M}q(Z_i) \quad (2\text{-}8)$$

由于$q(Z_i)$是估计分布，因而满足

$$\forall i.\int q(Z_i)dZ_i=1 \qquad (2\text{-}9)$$

在变分分布$Q(Z)$这个系统中，我们可以将每一个潜变量划分看成是一个单体，其他划分对其的影响都可以用一个看作是其自身的作用。采用的办法是迭代(Iterative VB(IVB) algorithm)。这是由于当变分自由能取得最大值的时候，划分Z_i的边缘密度与它的互斥集Z_{-i} (或者更进一步，马尔科夫毯，$mb(Z_i)$)具有一个简单的关系（将于下一节证明）：

$$Q(Z_i)\propto\frac{1}{C}\exp\left\langle\ln P(Z_i,Z_{-i},D)\right\rangle_{Q(Z_{-i})orQ(mb(Z_i))} \qquad (2\text{-}10)$$

于是，对于某个划分Z_i ,可以先保持其他划分Z_{-i}不变，然后用以上关系式更新Z_i。相同步骤作用于其他划分的更新，使得每个划分之间充分相互作用，最终达到稳定值。

另外，需要注意到，$Q(Z)$估计的是联合概率密度，而对于每一个$Q_i(Z_i)$，其与真实的边缘概率密度$P_i(Z_i)$的差别可能是很大的。比如一个标准的高斯联合分布$P(\mu,x)$和最优的平均场高斯估计$Q(\mu,x)$。Q选择了在它自己作用域中的高斯分布，因而变得很窄。此时边缘密度$Q_x(x)$变得非常小，完全$P_x(x)$与不同。

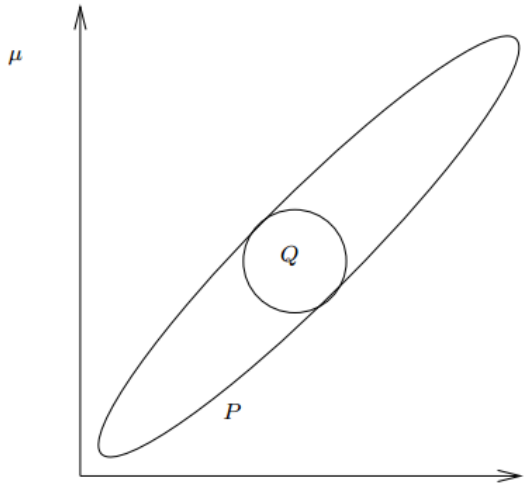


图 2-2 真实联合分布 P 与 VB 估计 Q

（3）变分贝叶斯算法描述

表 2-1 变分贝叶斯算法步骤

Step1：初始化$Q^{(1)}(Z_i)$，可随机取；

Step2：在第 k 步，计算Z_{-i}的边缘密度，

$$Q^{[k]}(Z_{-i}\mid D)\propto \exp\int_{Z_i^*} Q^{[k-1]}(Z_i\mid D)\log P(Z_i,Z_{-i},D)dZ_i \qquad (2\text{-}11)$$

Step3：计算Z_i的边缘密度

$$Q^{[k]}(Z_i\mid D)\propto \exp\int_{Z_{-i}^*} Q^{[k]}(Z_{-i}\mid D)\log P(Z_i,Z_{-i},D)dZ_{-i}\ ;$$

Step4：理论上$Q^{[\infty]}(Z_i\mid D)$将会收敛，则反复执行步骤二、三，直到$Q(Z_i),Q(Z_{-i})$稳定，或稳定在某个小范围内；

Step5：最后得$Q(Z)=Q(Z_i\mid D)Q(Z_{-i}\mid D)$。

2.2.4 边缘密度公式的推导

上文已经提到要找到一个更加简单分布$D(Z)$来近似$P(Z\mid D)$，同时问题转化为求解证据$logP(Z)$的下界$L(Q)$，或者$L(Q(Z))$。应该注意到$L(Q)$并非普通的函数，而是以整个函数为自变量的函数，这便是泛函。在上文已经说明什么是泛函以及泛函取得极值的必要条件。下面在平均场假设(2-8),(2-9)的基础上求解目标函数(2-7);

对于$L(Q(Z))=E_{Q(Z)}[\ln P(Z,D)]+H(Q(Z))$，将右式第一项定义为能量，第二项看作信息熵(Shannon entropy)。只考虑自然对数的形式，因为对于任何底数的对数总是可以通过换底公式将其写成自然对数与一个常量的乘积形式。另外根据平均场假设可以得到如下积分形式，

$$L(Q(Z))=\int(\prod_i Q_i(Z_i))\ln(Z,D)dZ-\int(\prod_k Q_k(Z_k))\sum_i \ln Q_i(Z_i)dZ \qquad (2\text{-}12)$$

其中 $Q(Z)=\prod_i Q_i(Z_i)$，且满足 $\forall i.\int Q_i(Z_i)dZ_i=1$

考虑划分 $Z=\{Z_i,\ Z_{-i}\}$，其中 $Z_{-i}=Z\setminus Z_i$ ，先考虑能量项(Energy)，

$$\begin{aligned}E_{Q(Z)}[\ln P(Z,D)]&=\int(\prod_i Q_i(Z_i))\ln(Z,D)dZ\\&=\int Q_i(Z_i)dZ_i\int Q_{-i}(Z_{-i})\ln(Z,D)dZ_{-i}\\&=\int Q_i(Z_i)\left\langle\ln(Z,D)\right\rangle_{Q_{-i}(Z_{-i})}dZ_i\\&=\int Q_i(Z_i)\ln\exp\left\langle\ln(Z,D)\right\rangle_{Q_{-i}(Z_{-i})}dZ_i\\&=\int Q_i(Z_i)\ln Q_i^*(Z_i)dZ_i+\ln C\end{aligned}$$

其中定义

$$Q_i^*(Z_i)=\frac{1}{C}\exp\left\langle\ln(Z,D)\right\rangle_{Q_{-i}(Z_{-i})} \qquad (2\text{-}13)$$

C 为 $Q_i^*(Z_i)$ 的归一化常数。再考虑熵量，

$$\begin{aligned}H(Q(Z))&=\sum_i\int(\prod_k Q_k(Z_k))\ln Q_i(Z_i)dZ\\&=\sum_i\iint Q_i(Z_i)Q_{-i}(Z_{-i})\ln Q_i(Z_i)dZ_idZ_{-i}\\&=\sum_i\left\langle\int Q_i(Z_i)\ln Q_i(Z_i)dZ_i\right\rangle_{Q_{-i}(Z_{-i})}\\&=\sum_i\int Q_i(Z_i)\ln Q_i(Z_i)dZ_i\end{aligned}$$

得到泛函，

$$\begin{aligned}\mathrm{L(Q(Z))}&=\int Q_i(Z_i)\ln Q_i^*(Z_i)dZ_i+\sum_i\int Q_i(Z_i)\ln Q_i(Z_i)dZ_i+\ln C\\&=(\int Q_i(Z_i)\ln Q_i^*(Z_i)dZ_i-\int Q_i(Z_i)\ln Q_i(Z_i)dZ_i)+\sum_{k\neq i}\int Q_k(Z_k)\ln Q_k(Z_k)dZ_k+\ln C\\&=\int Q_i(Z_i)\ln\frac{Q_i^*(Z_i)}{Q_i(Z_i)}dZ_i+\sum_{k\neq i}\int Q_k(Z_k)\ln Q_k(Z_k)dZ_k+\ln C\\&=-\mathrm{D}_{KL}(Q_i(Z_i)\|Q_i^*(Z_i))+H[Q_{-i}(Z_{-i})]+\ln C\end{aligned}$$

注意到 $L\left(Q\left(Z\right)\right)$ 并非只有一个等式，如果不可观测变量有 M 个划分，那么将有 M 个方程。为了最大化 $L\left(Q\left(Z\right)\right)$，同时注意到约束条件 $\forall i.\int Q_i(Z_i)dZ_i=1$，根据泛函求条件极值的必要条件，得，

$$\forall i.\ \frac{\partial}{\partial Q_i(Z_i)}\{-\mathrm{D}_{KL}[Q_i(Z_i)\|Q_i^*(Z_i)]-\lambda_i(\int Q_i(Z_i)dZ_i-1)\}:=0 \qquad (2\text{-}14)$$

直接求解将得到 Gibbs 分布，略显复杂。实际上，注意到 KL 散度，我们可以直接得到 KL 散度等于 0 的时候，L(D)达到最大值，最终得到

$$Q_i(Z_i)=Q_i^*(Z_i)=\frac{1}{C}\exp\left\langle \ln(Z_i,Z_{-i},D)\right\rangle_{Q_{-i}(Z_{-i})} \qquad (2\text{-}15)$$

C 为归一化常数 $C=\int \exp\left\langle \ln(Z_i,Z_{-i},D)\right\rangle_{Q_{-i}(Z_{-i})}dZ_{-i}$， $Q(Z_i)$ 为联合概率函数在除 Z_i 本身外的其他划分下的对数期望。又可以写为

$$\ln Q_i(Z_i)=\left\langle \ln(Z_i,Z_{-i},D)\right\rangle_{Q_{-i}(Z_{-i})}+\text{const} \qquad (2\text{-}16)$$

2.3 变分消息传播

传统的变分贝叶斯方法对模型的推导是繁琐而复杂的。J. Winn, Bishop[25,26] 考虑了贝叶斯网络中的共轭指数网络（conjugate-exponential networks）提出变分消息传播（VMP, Variational Message Passing）。这种方法使得充分统计量与自然参数都有一个标准形式，现在该方法已经取代了手工推导，成为标准的变分贝叶斯推断方法。而对于非共轭指数网络（比如混合模型），也能通过进一步的近似转化为标准形式。

2.3.1 理论基础

（1）贝叶斯网络

变分信息传播方法是建立在贝叶斯网络[33]上的，如图所示，对于一个节点 H_j，它的父节点为 pa_j，子节点为 ch_j，子节点 x_k 的父节点为 $cp_k^{(j)}\equiv pa_k\setminus H_j$。所有节点统称为 H_j 的马尔科夫毯，对于变分贝叶斯推理，我们只需要关心这个模型，H 为参数或潜在变量，其父节点为它的超参数，子节点为数据样本, co-parents 为其他参数或潜在变量。

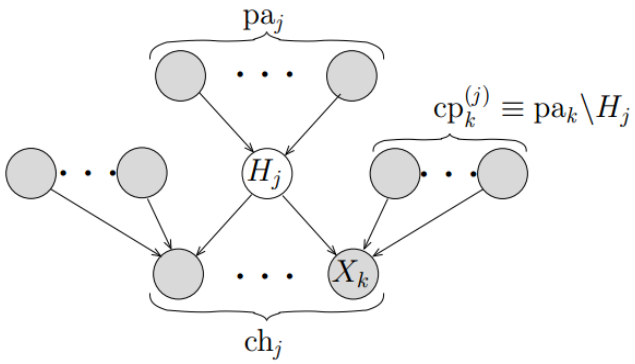


图 2-3 贝叶斯网络（马尔科夫毯）

（2）指数分布族

设$(X,B|\,p_\theta:\theta\in\Theta|)$是可控参数统计结构，其密度函数可表示为如下形式：

$$p_\theta(x)=c(\theta)\exp\{\sum_{i=1}^{k}c_j(\theta)T_j(x)\}h(x) \qquad (2\text{-}17)$$

并且它的支撑$\{x:p_\theta(x)>0\}$ 不依赖于 θ，则称此结构为指数型的统计结构，简称指数结构，其中的分布族为指数分布族。$0<c(\theta),c_1(\theta),...,c_k(\theta)<\infty,T_j(x)$ 都与 θ 无关，且取有限值的 B 可测函数，k 为正整数，$h(x)>0$，常见指数分布族，如二项分布，二元正态分布，伽马分布。

对于一个条件分布，如果它能写成如下形式，则称它属于指数分布族，

$$P(X\,|\,Y)=\exp[\phi(Y)^T u(X)+f(X)+g(Y)] \qquad (2\text{-}18)$$

其中$\phi(Y)$称为自然参数（natural parameter）向量，$u(X)$称为自然统计（natural statistic）向量。$g(Y)$作为归一化函数使得对于任意的 Y 都能整合到统一的形式。指数分布族的好处是它的对数形式是可计算的并且它的状态可以用自然参数向量所概括。

（3）共轭指数模型

当变量 X 关于父节点 Y 的条件概率分布 P(X|Y)为指数分布族，且为父节点分布 P(Y)的共轭先验，那么称这样的模型是共轭指数模型（Conjugate-Exponential Model）。考虑共轭指数模型，其后验的每个因子与它的先验都有相同的形式，因而只需要关心参数的变化，而无需整个函数。所谓相同的形式是指属于同样的分

布，比如都属于正态分布，伽马分布，多项式分布等。

（4）自然统计量的期望

如果知道自然参数向量$\phi(Y)$，那么就能找到自然统计量的期望。重写指数分布族，用ϕ作为参数，g重新参数化为$\tilde{g}$则，

$$P(X|\phi)=\exp[\phi^T u(X)+f(X)+\tilde{g}(\phi)]$$

对X积分有，

$$\int_X \exp[\phi^T u(X)+f(X)+\tilde{g}(\phi)]dX=\int_X P(X|\phi)dX=1$$

然后对ϕ微分，

$$\int_X \frac{d}{d\phi}\exp[\phi^T u(X)+f(X)+\tilde{g}(\phi)]dX=\frac{d}{d\phi}(1)=0$$

$$\int_X P(X|\phi)\left[u(X)+\frac{d\tilde{g}(\phi)}{d\phi}\right]dX=0$$

得自然统计量的期望，

$$\langle u(X)\rangle_{P(X|\phi)}=-\frac{d\tilde{g}(\phi)}{d\phi} \tag{2-19}$$

2.3.2 变分分布与下界的变分消息传播模型

（1）变分分布 Q 在共轭指数模型下的最优化

不失一般性，考虑变分分布的一个因子$Q(Y),Y$为马尔科夫毯上一个节点，子节点为X，如图 2-4 所示

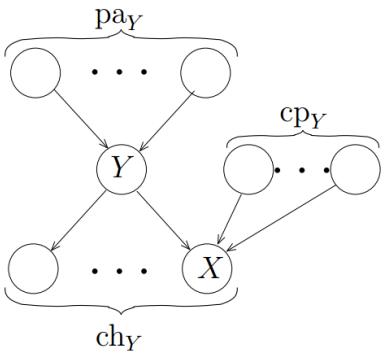


图 2-4 马尔科夫毯

根据指数族条件分布的一般形式，则变量 Y 关于父节点的条件概率为，

$$\ln P(Y \mid pa_Y) = \phi_Y(pa_Y)^T u_Y(Y) + f_Y(Y) + g_Y(pa_Y). \tag{2-20}$$

ϕ_Y, u_Y, f_Y, g_Y 的下标 Y 用于区分不同节点对数条件概率中的各成员。考虑 Y 的子节点 $X \in ch_Y$，则 X 的关于其父节点的条件概率为,

$$\ln P(X \mid Y, cp_Y) = \phi_X(Y, cp_Y)^T u_X(X) + f_X(X) + g_X(Y, cp_Y). \tag{2-21}$$

可以将 $P(Y \mid pa_Y)$ 看出 Y 的先验，$P(X \mid Y, cp_Y)$ 作为 Y 的似然函数。共轭的要求是这两个条件分布具有关于 Y 相同的函数形式，因而可以通过定义 ϕ_{XY} 和 λ 函数将后者改写成

$$\ln P(X \mid Y, cp_Y) = \phi_{XY}(X, cp_Y)^T u_Y(Y) + \lambda(X, cp_Y). \tag{2-22}$$

为了更新 Q(Y),需要找到(2-20),(2-21)关于除 Y 外其他因子的期望。对任何指数族的自然统计量 u 的期望都可以用自然参数向量 ϕ 带入（2-19）式得到。即对于任何变量 A，都可以找到 $\langle u_A(A) \rangle_Q$。特别的，当 A 为被观测量时，我们能直接计算得 $\langle u_A(A) \rangle_Q = u_A(A)$。

从(2-21)，(2-22)式可以看出 $\ln P(X \mid Y, cp_Y)$ 与 $u_X(X), u_Y(Y)$ 分布成线性关系。而共轭要求对数条件分布也会与所有的 $u_Z(Z)$ 成线性，$Z \in cp_Y$。因而看得出 $\ln P(X \mid Y, cp_Y)$ 是一个关于 u 的多线性函数。

考虑 Y 的变分更新方程，

$$\begin{aligned}
\ln Q_Y^*(Y) &= \left\langle \phi_Y(pa_Y)^T u_Y(Y) + f_Y(Y) + g_Y(pa_Y) \right\rangle_{\sim Q(Y)} + \\
&\quad \sum_{k \in ch_Y} \left\langle \phi_{XY}(X, cp_Y)^T u_Y(Y) + \lambda(X, cp_Y) \right\rangle_{\sim Q(Y)} + const. \\
&= \left[\left\langle \phi_Y(pa_Y) \right\rangle_{\sim Q(Y)} + \sum_{k \in ch_Y} \left\langle \phi_{XY}(X, cp_Y) \right\rangle_{\sim Q(Y)} \right]^T u_Y(Y) + f_Y(Y) + const. \\
&= \left[\phi_Y^* \right]^T u_Y(Y) + f_Y(Y) + const.
\end{aligned}$$

其中，

$$\phi_Y^* = \left\langle \phi_Y(pa_Y)^T \right\rangle_{\sim Q(Y)} + \sum_{k \in ch_Y} \left\langle \phi_{XY}(X, cp_Y)^T \right\rangle_{\sim Q(Y)} \tag{2-23}$$

正如以上所解释的，ϕ_Y 和 ϕ_{XY} 的期望都是相应的自然统计向量期望的多线性函数。因而有可能将以上期望重新参数化为

$$\tilde{\phi}_Y\left(\left\{\langle u_i\rangle\right\}_{i\in pa_Y}\right)=\langle\phi_Y(pa_Y)\rangle \tag{2-24}$$

$$\tilde{\phi}_{XY}\left(\langle u_X\rangle,\left\{\langle u_j\rangle\right\}_{j\in cp_Y}\right)=\langle\phi_{XY}(X,cp_Y)\rangle \tag{2-25}$$

举例：如果 X 服从 $N(Y,\beta^{-1})$，那么

$$\begin{aligned}\ln P(X\mid Y,\beta)&=\begin{bmatrix}\beta Y\\-\beta/2\end{bmatrix}^T\begin{bmatrix}X\\X^2\end{bmatrix}+\frac{1}{2}(\ln\beta-\beta Y^2-\ln 2\pi)\\&=\begin{bmatrix}\beta X\\-\beta/2\end{bmatrix}^T\begin{bmatrix}Y\\Y^2\end{bmatrix}+\frac{1}{2}(\ln\beta-\beta X^2-\ln 2\pi)\\&=\begin{bmatrix}-\frac{1}{2}(X-Y)^2\\\frac{1}{2}\end{bmatrix}^T\begin{bmatrix}\beta\\\ln\beta\end{bmatrix}-\frac{1}{2}\ln 2\pi.\end{aligned}$$

其中 $u_X(X)=\begin{bmatrix}X\\X^2\end{bmatrix},u_Y(Y)=\begin{bmatrix}Y\\Y^2\end{bmatrix},u_\beta(\beta)=\begin{bmatrix}\beta\\\ln\beta\end{bmatrix}.$

$\phi_{XY}(X,\beta)=\begin{bmatrix}\beta X\\-\beta/2\end{bmatrix}$ 可以重参数化为 $\tilde{\phi}_{XY}(\langle u_X\rangle,\langle u_\beta\rangle)=\begin{bmatrix}\langle u_\beta\rangle\langle u_X\rangle\\-\langle u_\beta\rangle/2\end{bmatrix}$

（2）下界 L(Q)的变分消息传播模型

在贝叶斯网络中,由于 Q 可因式分解，则有

$$\begin{aligned}L(Q)&=\langle\ln P(H,V)\rangle-\langle Q(H)\rangle\\&=\sum_i\langle\ln P(X_i\mid pa_i)\rangle-\sum_{i\in H}\langle\ln Q_i(H_i)\rangle\\&\overset{def}{=}\sum_i L_i\end{aligned} \tag{2-26}$$

$L(Q)$ 被分解为每一个节点上的贡献值 $\{L_i\}$，如节点 H_j 的贡献值为

$$\begin{aligned} L_j &= \left\langle \ln P(H_i \mid pa_j) \right\rangle - \left\langle \ln Q_i(H_i) \right\rangle \\ &= \left\langle \phi_j(pa_j)^T \right\rangle \left\langle u_j(H_j) \right\rangle + \left\langle f_j(H_j) \right\rangle + \left\langle g_j(pa_j) \right\rangle - \left[\phi_j^{*T} \left\langle u_j(H_j) \right\rangle + \left\langle f_j(H_j) \right\rangle + \tilde{g}_j(\phi_j^*) \right] \\ &= \left(\left\langle \phi_j(pa_j) \right\rangle - \phi_j^* \right)^T \left\langle u_j(H_j) \right\rangle + \left\langle g_j(pa_j) \right\rangle - \tilde{g}_j(\phi_j^*) \end{aligned} \tag{2-27}$$

注意到$\left\langle \phi_j(pa_j) \right\rangle$和$\phi_j^*$在求$H_j$的后验分布时就已经计算了；$\left\langle u_j(H_j) \right\rangle$在$H_j$传出消息的时候也已经计算了，这样降低了下界的计算成本。

特别地，对于每个观测变量V_k对下界的贡献值则更简单，

$$\begin{aligned} L_k &= \left\langle \ln P(V_k \mid pa_k) \right\rangle \\ &= \left\langle \phi_j(pa_j) \right\rangle^T u_k(V_k) + f_k(V_k) + \tilde{g}_k\left(\left\langle \phi_j(pa_j) \right\rangle \right) \end{aligned} \tag{2-28}$$

2.3.3 变分消息传播算法

（1）变分消息的定义

来自父节点的消息（Message from parents）：父节点传播给子节点的消息只是自然统计量的期望：

$$m_{Y \to X} = \left\langle u_Y \right\rangle. \tag{2-29}$$

消息传播给父节点（Message to parents）：依赖于 X 之前从 Y 的 co-parents 接收到的消息；对任何节点 A，如果 A 是被观测量，那么$\left\langle u_A \right\rangle = u_A$，

$$m_{X \to Y} = \tilde{\phi}_{XY}\left(\left\langle u_X \right\rangle, \left\{ m_{i \to X} \right\}_{i \in cp_Y} \right) \tag{2-30}$$

用 Y 接收来自父节点与子节点的消息来计算ϕ_Y^*，然后我们就能通过计算更新后的自然参数向量ϕ_Y^* 来找到 Y 的更新后的后验分布Q_Y^*，ϕ_Y^*的计算公式如下，

$$\phi_Y^* = \tilde{\phi}_Y\left(\left\{ m_{i \to Y} \right\}_{i \in pa_Y} \right) + \sum_{j \in ch_Y} m_{j \to Y} \tag{2-31}$$

该式与（2-23）式一致。从（2-19）式可以看出自然统计量的期望$\left\langle u_Y \right\rangle_{Q_Y^*}$是$Q_Y^*$的单一函数，这样我们就可以用它来计算期望的新值。变分消息传播算法通过迭代的消息传播来最优化变分分布Q.

（2）算法描述

表 2-2 变分消息传播步骤

Step1. 通过初始化相关的矩向量$\langle u_j(X_j)\rangle$来初始化每个因子分布Q_j.
Step2. 对于每一个节点X_j，
• 从父节点和子节点接收 (2-29),(2-30)式所定义的消息。前提是子节点已经从X_j的 co-parents 接收到消息。
• 使用（2-31）式更新自然参数向量ϕ_j^*；
• 根据新的参数向量更新距向量$\langle u_j(X_j)\rangle$；
Step3. 计算新的下界$L(Q)$；
Step4. 如果经过数次迭代已经无法增加下界值，或者各边缘分布达到稳定值，则结束；否则回到 Step2。

举例：对于单一高斯模型消息传播过程如下图

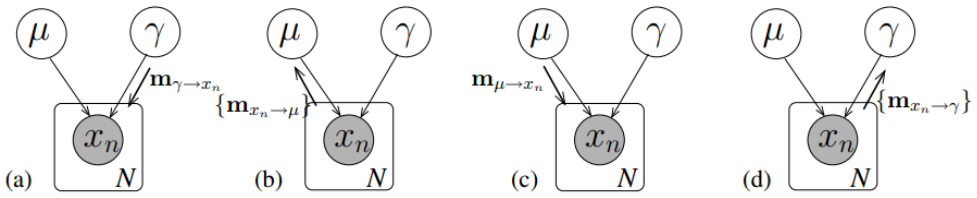


图 2-5 单一高斯模型消息传播过程[25]

2.3.4 混合模型

到目前为止只考虑了来自指数族的分布。而通常来讲，混合模型并非来自指数族，比如高斯混合模型，通常需要将混合分布转化为指数族分布形式。

考虑高斯混合模型，通常有如下形式，

$$P(X\mid\{\pi_k\},\{\theta_k\})=\sum_{k=1}^{K}\pi_k P_k(X\mid\theta_k) \qquad (2\text{-}32)$$

可以引入一个离散型潜在变量 λ,表示每个观测点是属于哪个单高斯分布。重写分布函数为：

$$P(X \mid \lambda,\{\theta_k\}) = \sum_{k=1}^{K} P_k(X \mid \theta_k)^{\delta_{\lambda k}} \quad (2\text{-}33)$$

加入该 λ 变量后该分布属于指数分布族，可写成

$$\ln P(X \mid \lambda,\{\theta_k\}) = \sum_{k} \delta(\lambda,k)\left[\phi_k(\theta_k)^T u_k(X) + f_k(X) + g_k(\theta_k)\right] \quad (2\text{-}34)$$

如果 X 有子节点 Z，那么共轭条件要求每一个成分都有相同的自然统计向量，统一定义为 $u_1(X) = u_2(X) = ... = u_K(X) \overset{def}{=} u_X(X)$。另外，我们可能要使模型的其他部分也有相同的形式，虽然不要求共轭，即 $f_1 = f_2 = ... = f_K \overset{def}{=} f_X$。在这种情况下，混合模型的每个成分都有相同的形式，可写成，

$$\begin{aligned}\ln P(X \mid \lambda,\{\theta_k\}) &= \left[\sum_{k} \delta(\lambda,k)\phi_k(\theta_k)\right]^T u_X(X) + f_X(X) + \sum_{k} \delta(\lambda,k) g_k(\theta_k) \\ &= \phi_X(\lambda,\{\theta_k\})^T u_X(X) + f_X(X) + \tilde{g}_X(\phi_X(\lambda,\{\theta_k\}))\end{aligned} \quad (2\text{-}35)$$

其中定义 $\phi_X = \sum_{k} \delta(\lambda,k)\phi_k(\theta_k)$。这样对于每个成分来说条件分布都有了与指数分布族一样的形式，便可以应用变分消息传播算法。

从某个节点 X 传播个子节点的消息为 $\langle u_X(X) \rangle$，而这是通过混合参数向量 $\phi_X(\lambda,\{\theta_k\})$ 计算的。相似地，节点 X 到父亲节点 θ_k 的消息是那些以它为父节点的子节点发出的，而节点 X 中哪些属于 θ_k 是由指标变量 $Q(\lambda = k)$ 的后验确定的。最后，从 X 到 λ 的消息是一个 K 维向量，其中第 k 个元素为 $\langle \ln P_k(X \mid \theta_k) \rangle$.

2.4 算法分析

2.4.1 VB 算法与 EM 算法比较

EM 算法计算随机变量（或归类于参数）后验分布的点估计，但估计隐变量的真实后验分布。用这些参数的众数作为点估计，无任何其他信息。而在 VB 算法作为一个分布估计（Distributional Approximation）方法，计算所有变量的真实后验分布的估计，包括参数和隐变量。在贝叶斯推断中，计算点估计一般使用常

用的均值而非众数。与此同时，应该注意的是计算参数在 VB 中与 EM 有不同的意义。EM 算法计算贝叶斯网络本身的参数的最优值。而 VB 计算用于近似参数和隐变量的贝叶斯网络的参数最佳值，VB 会先找一个合适的参数分布，通常是一个先验分布的形式，然后计算这个分布的参数值，更准确说是超参数，最后得到联合分布的各参数的分布。

2.4.2 算法复杂性

变分贝叶斯估计方法是众多概率函数估计技术之一。还有许多其他被广泛使用的估计算法，一般分为确定性（deterministic）和随机性（stochastic）的方法，比如基于点估计的极大似然估计、极大后验概率估计，基于局部估计的 Laplace 估计，基于 spline 估计的 B-样条估计，还有经验性估计，利用随机采用的如 MCMC 方法。变分贝叶斯方法作为平均场估计，能够在计算复杂度和精度之间保持一个良好的关系，如图（2-6）所示。变分贝叶斯方法主要的计算压力在于它的 IVB 算法——一系列为求取变分边缘概率相关的矩估计而进行的迭代。如果只关心计算代价而对精度要求不高，那么可以用简单的估计方法来代替变分边缘概率，或者减少估计迭代的次数，这样变分估计的路径将沿着虚线往下。

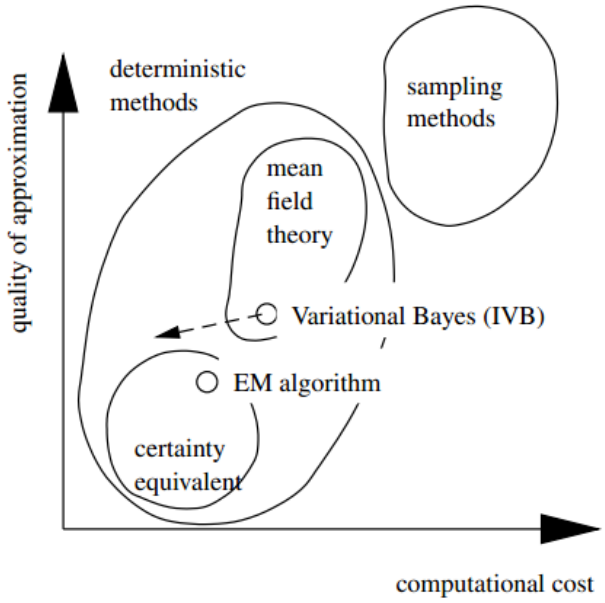


图 2-6 VB 方法的精度与复杂性之间的关系[23]

2.5 本章小结

本章阐述了变分贝叶斯方法，包括理论提出的背景，数学基础，基本原理，变分消息传播以及算法的复杂度。从数学上对变分贝叶斯方法进行了严格地推导，分析了平均场估计的合理性。对基于贝叶斯网络的变分消息传播方法从理论基本到算法流程展开论述。最后特别地与 EM 算法进行比较，分析了变分贝叶斯方法的算法复杂性。为下文基于变分贝叶斯的脑图像分割做好理论基础。

第三章 基于变分混合模型的脑图像分割

混合模型是一类重要的图像分割建模工具，特别高斯混合模型在医学图像分割中应用非常广泛。一般其参数用期望最大化算法(Expectation Maximum, EM)估计，变分贝叶斯作为一种分布估计方法，融入先验信息，不直接求取参数值而用参数边缘分布代替之，通常能有更好的效果。本章将用变分贝叶斯推导高斯混合模型以及学生 t 混合模型的参数估计算法，并将其应用到脑图像分割中。

3.1 变分高斯混合模型及其参数估计

3.1.1 高斯混合模型

设 d 维空间 R^d 中的有独立同分布的数据集 $X=\{x_1,x_2,...,x_n\}$，如果该数据集的分布近似椭圆球体，则可采用单一高斯分布密度函数描述这些数据点。但大多数情况下，这些数据集是不符合椭圆体的，那么就可以采用多个单一高斯函数的加权平均表示数据集的分布，这便是高斯混合模型（Gaussian Mixture Model，简称 GMM）。可以表示为：

$$p(x)=\sum_{k=1}^{K}\pi_k N(x\mid\mu_k,\Lambda_k^{-1}) \tag{3-1}$$

其中 π_k 表示数据点属于第 k 个高斯分布的概率，即混合模型的混合比例，满足

$$\sum_{k=1}^{K}\pi_k=1,0\le\pi_k\le 1; \tag{3-2}$$

$N(x\mid\mu_k,\Lambda_k^{-1})$ 表示第 k 个高斯分布，μ_k,Λ_k 分别为它的期望和逆协方差矩阵。如果高斯个数足够多，它能够逼近任意的连续分布，但考虑到模型的复杂度，k 值通常不会太大。对于一组数据，如果已经知道或假定了高斯成分的个数，那么接下来的任务就是估计模型的三组参数 π_k,μ_k,Λ_k。

3.1.2 基于变分贝叶斯的高斯混合参数估计

变分贝叶斯方法需要先选取各参数的先验分布，一般要求是无信息先验。然后建立图模型，推导出各参数的边缘概率密度表达式。通过变分消息传播机制迭代计算得各参数分布的超参数。

式（3-1）的高斯混合模型并非指数分布形式，因而无法应用变分消息传播机制。需要先将其转化为（2-33）的形式才能成为指数分布族，即

$$p(X \mid Z,\mu,\Lambda)=\prod_{n=1}^{N}\prod_{k=1}^{K}N(x_n \mid \mu_k,\Lambda_k^{-1})^{z_{nk}} \tag{3-3}$$

其中 k 表示单高斯分布的个数，N 表示样本个数，$Z=\{z_1,...,z_N\}$是一组隐变量，每一项$z_k=\{z_{1k,...},z_{nk}\}$表示对应的样本$x_k$属于哪个混合成分。

（1）确定无信息先验分布

先验分布一般可以根据共轭分布方法，Jefferys 原则，最大熵原则等来确定。要求先验分布应取共轭分布（conjugate distribution）才合适，即先验分布$h(\theta)$与后验分布$h(\theta \mid x)$属于同一分布类型。本文不展开讨论，直接给出：

1）混合模型混合比例$\pi_{i=1,...,k}$后验分布属于多项式分布，则它的共轭先验分布为 K 维对称 Dirichlet 分布(附录 A.8)，即$\pi_{i=1,...,k} \sim SymDir(K,\alpha_0)$

$$p(\pi)=\frac{\Gamma(K\alpha_0)}{\Gamma(\alpha_0)^K}\prod_{k=1}^{K}\pi_k^{\alpha_0-1} \tag{3-4}$$

2）高斯分布的期望$\mu_{i=1,...,k}$的后验分布为高斯分布，其共轭先验仍为高斯分布(附件 A.1)：

$$p(\mu_k \mid \Lambda_k)=N(\mu_k \mid m_0,(\beta_0\Lambda_k)^{-1}) \tag{3-5}$$

3）多元高斯分布的 Precision 矩阵（逆协方差矩阵）$\Lambda_{i=1,...,k}$，其共轭先验分布为 Wishart 分布（附录 A.2），即,

$$p(\Lambda_k)=W(\Lambda_k \mid w_0,\nu_0) \tag{3-6}$$

4）隐变量$Z=\{z_1,...,z_N\}$的共轭先验为多项分布；多项式分布是二项式分布的推广，在一个 K 维向量中只有一项为 1，其它都为 0。

$$p(Z|\pi)=\prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}} \tag{3-7}$$

另外，为了区分联合分布的参数，以上分布的参数$K,\alpha_0,\beta_0,w_0,\upsilon_0,m_0$ 又称为超参数，下标 0 表示其为超参数的初始值。

（2）建立概率图模型

用“盘子表示法”（plate notation）表示贝叶斯多元高斯混合模型，如图 3-1 所示。小正方形表示不变的超参数，如β_0 ,ν_0 ,α_0 ,μ_0 ,W_0；圆圈表示随机变量，如$\pi,z_i,x_i,\mu_k,\Lambda_k$；圆圈内的值为已知量。其中[K],[D]表示 K、D 维的向量，[D,D]表示$D\times D$的矩阵，单个 K 表示一个有 K 个值的多项式变量；波浪线和一个开关表示变量x_i通过一个 K 维向量z_i来选择其他传入的变量(μ_k,Λ_k)。

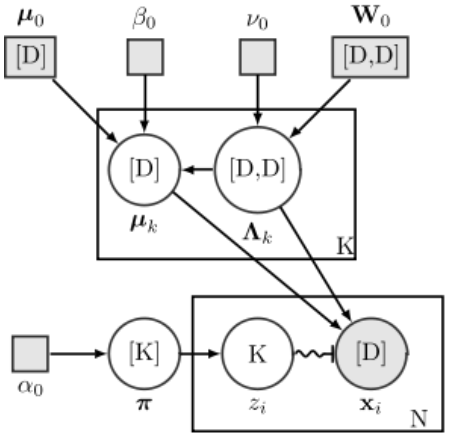


图 3-1 多元高斯混合模型的盘子表示法

假设各参数与潜在变量条件独立，则联合分布可以表示为

$$p(X,Z,\pi,\mu,\Lambda)=p(X|Z,\mu,\Lambda)p(Z|\pi)p(\pi)p(\mu|\Lambda)p(\Lambda) \tag{3-8}$$

（3）变分边缘分布

1）计算 Z 的边缘密度，根据平均场假设，$q(Z,\pi,\mu,\Lambda)=q(Z)q(\pi,\mu,\Lambda)$，则

$$\begin{aligned}\ln q^*(Z) &= E_{\pi,\mu,\Lambda}[\ln p(X,Z,\pi,\mu,\Lambda)]+\text{const}\\ &= E_{\pi,\mu,\Lambda}[\ln p(X\mid Z,\mu,\Lambda)\,p(Z\mid\pi)\,p(\pi)\,p(\mu\mid\Lambda)\,p(\Lambda)]+\text{const}\\ &= E_{\pi}[\ln p(Z\mid\pi)]+E_{\mu,\Lambda}[\ln p(X\mid Z,\mu,\Lambda)]+\text{const}\\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\ln\rho_{nk}+\text{const}\end{aligned}$$

其中，

$$\ln\rho_{nk} = E[\ln\pi_k]+\frac{1}{2}E[\ln|\Lambda_k|]-\frac{D}{2}\ln(2\pi)-\frac{1}{2}E_{\mu_k,\Lambda_k}[(x_n-\mu_k)^T\Lambda_k(x_n-\mu_k)] \qquad (3\text{-}9)$$

两边分别取对数可得，

$$q^*(Z)\propto\prod_{n=1}^{N}\prod_{k=1}^{K}\rho_{nk}^{z_{nk}} \qquad (3\text{-}10)$$

归一化 $\rho_{nk}^{z_{nk}}$，即观测变量的属于某个单高斯分布的概率相加应等于 1,则有

$$q^*(Z)=\prod_{n=1}^{N}\prod_{k=1}^{K}r_{nk}^{z_{nk}} \qquad (3\text{-}11)$$

其中，

$$r_{nk}=\frac{\rho_{nk}}{\sum_{j=1}^{K}\rho_{nj}} \qquad (3\text{-}12)$$

可见 $q^*(Z)$ 是多个单观测多项式分布的乘积，可以因式分解成一个个以 $r_{nk},(k=1...K)$ 为参数的单观测多项式分布 z_n。更进一步，根据多项式分布，有

$$E[z_{nk}]=r_{nk} \qquad (3\text{-}13)$$

2）计算 π 的边缘密度，根据平均场假设有，$q(\pi,\mu,\Lambda)=q(\pi)\prod_{k=1}^{K}q(\mu_k,\Lambda_k)$

$$\begin{aligned}\ln q^*(\pi) &= E_{Z,\mu,\Lambda}[p(X\mid Z,\pi,\mu,\Lambda)]+const\\ &=\ln p(\pi)+E_Z[\ln p(Z\mid\pi)]+const\\ &=(\alpha_0\text{-}1)\sum_{k=1}^{K}\ln\pi_k+\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\ln\pi_k+const\end{aligned}$$

两边取对数 $q^*(\pi)\sim\prod_{n=1}^{K}\pi_k^{\sum_{n=1}^{N}r_{nk}+\alpha_0-1}$，可见 $q^*(\pi)$ 是 Dirichlet 分布，$q^*(\pi)\sim Dir(\alpha)$

其中，

$$\alpha = \alpha_0 + N_k \tag{3-14}$$

$$N_k = \sum_{n=1}^{N} r_{nk} \tag{3-15}$$

3）最后同时考虑 μ,Λ，对于每一个单高斯分布有，

$$\begin{aligned}\ln q^*(\mu_k,\Lambda_k) &= E_{Z,\pi,\mu_{i\neq k},\Lambda_{i\neq k}}[\ln p(X\mid Z,\mu_k,\Lambda_k)p(\mu_k,\Lambda_k)] \\ &= \ln p(\mu_k,\Lambda_k) + \sum_{n=1}^{N} E[z_{nk}]\ln N(x_n\mid \mu_k,\Lambda_k^{-1}) + const\end{aligned}$$

经过一系列化简将得到 Gaussian-Wishart 分布，

$$q^*(\mu_k,\Lambda_k) = N(\mu_k\mid m_k,(\beta_k\Lambda_k)^{-1})W(\Lambda_k\mid w_k,\nu_k) \tag{3-16}$$

其中，

$$\beta_k = \beta_0 + N_k \tag{3-17}$$

$$m_k = \frac{1}{\beta_k}(\beta_0 m_0 + N_k\bar{x}_k) \tag{3-18}$$

$$w_k^{-1} = w_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{x}_k - m_0)(\bar{x}_k - m_0)^T \tag{3-19}$$

$$\nu_k = \nu_0 + N_k \tag{3-20}$$

$$\bar{x}_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}x_n \tag{3-21}$$

$$S_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\bar{x}_k - x_n)(\bar{x}_k - x_n)^T \tag{3-22}$$

注意到 π,μ,Λ 边缘概率公式都需要且只与 r_{nk}；另一方面，r_{nk} 的计算需要 ρ_{nk}，而这又是基于 $E[\ln\pi_k], E[\ln|\Lambda_k|]$，$E_{\mu_k,\Lambda_k}[(x_n-\mu_k)^T\Lambda_k(x_n-\mu_k)]$，即需要知道 π,μ,Λ 的值。不难确定这三个期望的一般表达式为：

$$\ln\tilde{\pi}_k \equiv E[\ln|\pi_k|] = \psi(\alpha_k) - \psi\left(\sum_{i=1}^{K}\alpha_i\right) \tag{3-23}$$

$$\ln\tilde{\Lambda}_k \equiv E[\ln|\Lambda_k|] = \sum_{i=1}^{D}\psi\left(\frac{\nu_k+1-i}{2}\right) + D\ln 2 + \ln|\Lambda_k| \tag{3-24}$$

$$E_{\mu_k,\Lambda_k}[(x_n-\mu_k)^T\Lambda_k(x_n-\mu_k)] = D\beta_k^{-1} + \nu_k(x_n-m_k)^T W_k(x_n-m_k) \tag{3-25}$$

这些结果能够导出：

$$r_{nk} \propto \tilde{\pi}_k \tilde{\Lambda}_k^{1/2} \exp\left\{-\frac{D}{2\beta_k}-\frac{\nu_k}{2}(x_n-m_k)^T W_k(x_n-m_k)\right\} \tag{3-26}$$

（4）计算变分下界

由定义可得变分高斯混合模型的下界为：

$$\begin{aligned} L &= \sum_Z \iiint q(Z,\pi,\mu,\Lambda)\ln\left\{\frac{p(X,Z,\pi,\mu,\Lambda)}{q(Z,\pi,\mu,\Lambda)}\right\} d\pi d\mu d\Lambda \\ &= E[\ln p(X,Z,\pi,\mu,\Lambda)] - E[\ln q(Z,\pi,\mu,\Lambda)] \\ &= E[\ln p(X,Z,\pi,\mu,\Lambda)] + E[\ln p(Z\mid\pi)] + E[\ln p(\pi)] + E[\ln p(\mu,\Lambda)] \\ &\quad -E[\ln q(Z)] - E[\ln q(\pi)] - E[\ln q(\mu,\Lambda)] \end{aligned} \tag{3-27}$$

其中 $E[\cdot]$表示期望，相关值列于附录（B.1-B.7）

（5）算法步骤

参数变量 μ_k,Λ_k 更新方程中的超参数 $\beta_k,m_k,w_k,\upsilon_k$ 都依赖与统计量 $N_k,\bar{x}_k,S_k$，而这些统计量又依赖于 r_{nk}.参数变量 π 更新方程中的超参数 $\alpha_{1\cdots K}$ 依赖于统计量 N_k，即 r_{nk}。潜在变量 r_{nk} 的更新方程对超变量 $\beta_k,m_k,w_k,\upsilon_k$ 有直接的依赖关系，同时对 $w_k,\nu_k,\alpha_{1\cdots K}$ 通过 $\tilde{\pi}_k,\tilde{\Lambda}_k$ 有间接的依赖关系。

这样算法步骤总结为：

表 3-1 变分贝叶斯高斯混合模型算法步骤

Step1 确定高斯混合个数 K, 设定先验分布超参数 $\alpha_0,\beta_0,w_0,\upsilon_0,m_0$， 设较小值;
Step2 初始化隐变量 r_{nk}， 可以用 K-means 算法初始化;
Step3(VBE-Step) 用参数 μ_k,Λ_k,π 和超参数 $\beta_k,m_k,w_k,\upsilon_k,\alpha_k$ 计算隐变量 r_{nk}；
Step4(VBM-Step) 用隐变量 r_{nk} 计算参数 μ_k,Λ_k,π 和超参数 $\beta_k,m_k,w_k,\upsilon_k,\alpha_k$ 的新值;
Step5 计算变分下界 L,若 $L(t)-L(t-1)<\delta$ 则结束， 否则继续执行 Step3,Step4;

3.1.3 与 EM 算法的比较

VBEM 算法与 EM 算法用 ML 或 MAP 估计高斯混合模型参数相似。在 E-step 中，隐变量 r_{nk} 对应于隐变量关于数据样本的后验概率；$N_k, \bar{x}_k, S_k$ 统计量对应于 EM 算法中“soft-count”统计量；然而用这些统计量去计算参数的新值与 EM 算法中用“soft-count”计算新参数值一致。

虽然如此，VB 算法与 EM 算法还有很多不同之处的。比如迭代中，逼近最优值的过程是不一样的。如图 3-2,受约束 EM 算法极大似然值是动态变化的。刚开始与当前最优值相差一个 KLD。在 E 步骤，下界逼近最大似然值；然后在 M 步骤中，根据参数新值重新计算似然值。反复迭代直到收敛。而在 VBEM 算法中，极大似然值是不变的。VBE 与 VBM 步骤，都是逼近极大似然值的过程，如图 3-3。

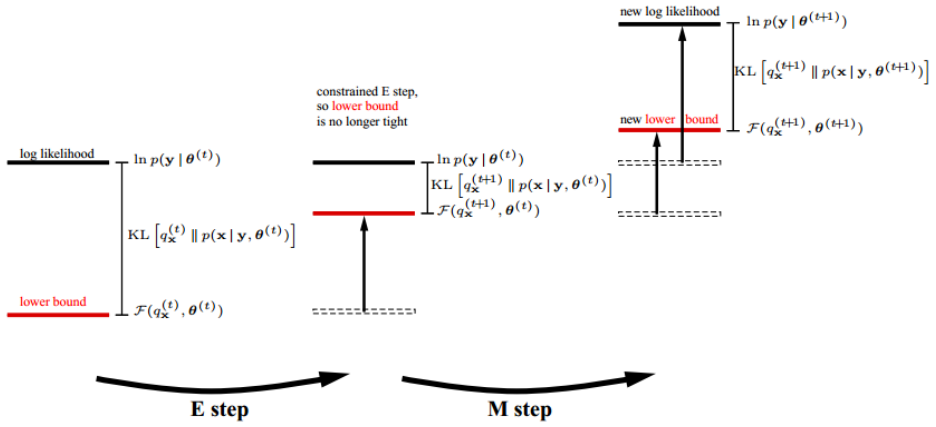


图 3-2 受约束的 EM 算法

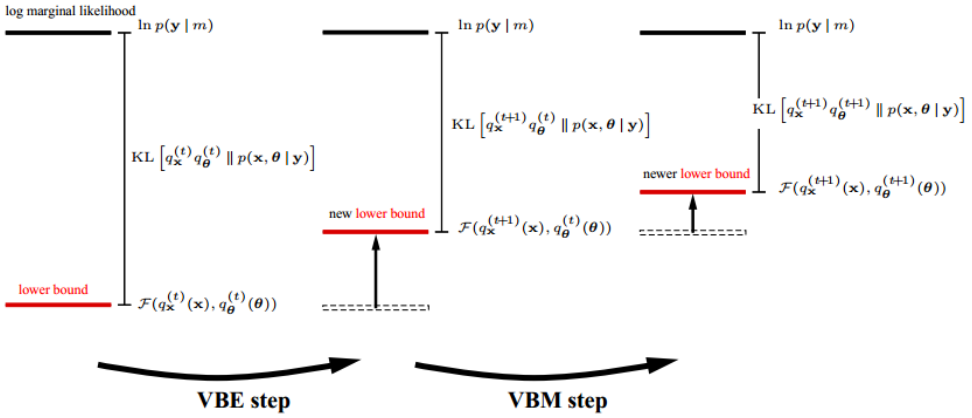


图 3-3 VBEM 算法

3.2 基于变分贝叶斯的学生 t 混合模型

高斯混合模型最主要的问题在于它很难处理离群点，因为高斯分布的尾部呈指数下降。相比于高斯分布，学生 t 分布具有更重的尾部，因而鲁棒性更强。传统的学生 t 混合模型也由 EM 算法估计参数，本节将说明如何推断学生 t 混合模型参数的变分后验分布。又由于有限混合模型需要知道混合成分的个数，本节末尾将说明无限学生 t 混合模型及相应的变分推断。

3.2.1 学生 t 混合模型

均值为 μ ,逆协方差矩阵为 Λ，自由度为 v 的 d 维学生 t 分布（也称 t 分布）的概率密度函数为：

$$St(x\mid\mu,\Lambda,v)=\frac{\Gamma(v/2+d/2)\mid\Lambda\mid^{1/2}}{\Gamma(v/2)(\pi v)^{d/2}}\left(1+\frac{\Delta^2}{v}\right)^{-(v+d)/2} \tag{3-28}$$

其中

$$\Gamma(y)=\int_0^\infty z^{y-1}e^{-z}dz \tag{3-29}$$

为 gamma 函数，

$$\Delta^2=(x-\mu)^T\Lambda(x-\mu) \tag{3-30}$$

为 x 到 μ 的 Mahalanobis 平方距离。随着自由度 $v\to\infty$，学生 t 分布逐渐退化为有相同均值 μ 和逆协方差矩阵 Λ 的高斯分布。如图 3-所示。

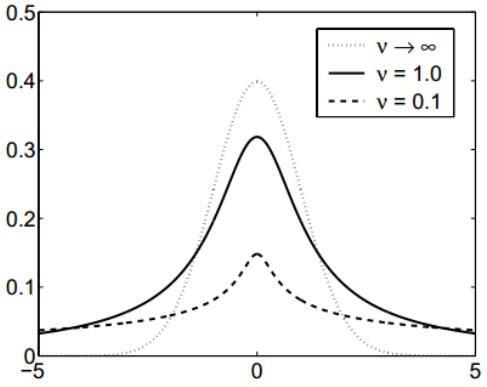


图 3-4 估计学生 t 分布随着自由度 $v\to\infty$ 的变化关系

与高斯模型不同的是无法通过极大似然函数的方法得到学生分布的闭合解。然而，可以将学生分布表示成无限混合的比例高斯模型，即如下形式，

$$St(x|\mu,\Lambda,v)=\int_0^\infty N(x|\mu,u\Lambda)\varsigma(u|v/2,v/2)du \tag{3-31}$$

其中 $N(x|\mu,u\Lambda)$ 为高斯分布，$\varsigma(u|v/2,v/2)$ 为 gamma 分布（附件 A.3）。

现在考虑学生混合模型，

$$p(x_n|\{\mu,\Lambda,v\},\pi)=\sum_{m=1}^{M}\pi_m St(x_n|\mu_m,\Lambda_m,v_m) \tag{3-32}$$

其中混合系数 $\pi=(\pi_1,...,\pi_M)^T$ 满足 $\pi_m\geq 0$ 且 $\sum_m \pi_m=1$。

为了得到学生混合模型的变分形式，与高斯混合模型一样，引入隐变量 s，每一项 $s_m=\{s_{1m},...,s_{nm}\}$ 表示对应的样本 x_m 属于哪个混合成分,得

$$p(x_n|s,\{\mu,\Lambda,v\})=\sum_{m=1}^{M}St(x_n|\mu_m,\Lambda_m,v_m)^{s_m} \tag{3-33}$$

同样的，隐变量 s 的的先验为多项式分布

$$p(S|\pi)=\prod_{n=1}^{N}\prod_{m=1}^{M}\pi_m^{s_{nm}} \tag{3-34}$$

另外，均值 μ 的先验为正态分布，即 $p(\mu_m|\Lambda_m)\sim N(\mu_m|m_0,(\rho_0\Lambda_m)^{-1})$；逆协方差矩阵的先验为 Wishart 分布，即 $p(\Lambda_m)\sim W(\Lambda_m|W_0,\eta_0)$，混合系数的先验为 Dirichlet 分布，即 $\pi\sim Dir(K,\alpha_0)$。

所有变量的联合分布可以表示成有向图模型，如图 3-5。

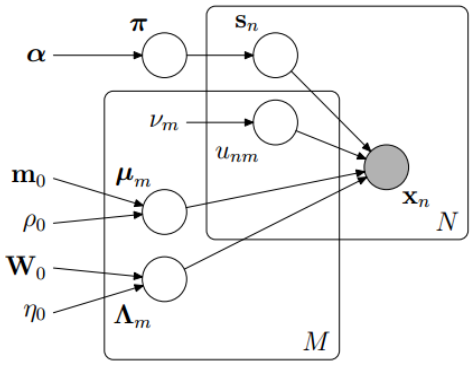


图 3-5 学生 t 混合模型的概率图模型

其中 M-plate 表示 M 个混合成分，N-plate 表示 N 个独立同分布的观测量 x_n。圆圈节点表示随机变量，如 μ_m, Λ_m，其他无节点的标号表示常量，如 m_0, ρ_0，或可调节参数（v_m）。有阴影的节点 x_n 表示该变量是可观测的，其他（无阴影）的节点是隐藏的。需要注意到 $\{u_{nm}\}$ 同时属于两个盘子，这说明该变量对每个混合成分和每个观测变量相关。

特别需要说明的是，v 是一个非随机（non-stochastic）参数，因为它没有共轭先验。然而，由于在每个混合成分中只有一个这样的参数，因此在变分推断过程中，可以通过一些最优化方法得到该值，下文将会讨论。

3.2.2 变分推断

上诉贝叶斯模型的准确推断是不可计算的。然而，可以通过选择共轭指数分布找到优雅的变分形式。为此考虑观测量 X 的对数似然函数，

$$\ln p(X) = L(q) + KL(q \| p) \tag{3-35}$$

其中

$$L(q) = \int q(\theta) \ln \left\{ \frac{p(X,\theta)}{q(\theta)} \right\} d\theta \tag{3-36}$$

$KL(q \| p)$ 为 q 与 p 分布之间的 KL 散度。

与高斯混合模型参数估计的方法一样，在平均场假设下，极小化 KL 散度，即极大化式（3-36）。将变分分布因式分解为，

$$q(\theta) = q(\{\mu_m\}) q(\{\Lambda_m\}) q(\pi) q(\{S_n\}) q(\{u_n\}) \tag{3-37}$$

然后对每一项展开计算，由于在对高斯混合模型的变分推断已经有比较详细的论述，不作详细推断，所有相关公式见附录 2 中 VB-SMM 变分推断公式(附录 B.8-B.22)。此处只给出一个例子。对于变分后验分布 $q(\{\mu_m\})$ 有，

$$\ln q(\Lambda_m) = \left\langle \ln p(X \mid \theta) p(\theta \mid m_0, \rho_0, W_0, \eta_0, v, \alpha) \right\rangle_{\mu,u,s} + const. \tag{3-38}$$

$$= \frac{N}{2} \ln |\Lambda_m| - \frac{1}{2} Tr \left[\sum_n^N \left\langle (x_n - \mu_m)(x_n - \mu_m)^T u_{nm} s_{nm} \right\rangle_{\mu,u,s} \Lambda_m \right]$$

$$+\frac{\eta_0-d-1}{2}\ln|\Lambda_m|-\frac{1}{2}Tr[W_0^{-1}\Lambda_m]+const. \tag{3-39}$$

式（3-38）是根据（2-16）来的，对比（3-39）与（A.2）,可看出，

$$q(\Lambda_m)\sim W(\Lambda_m|W_m,\eta_m) \tag{3-40}$$

其中，

$$\begin{aligned}W_m^{-1}&=W_0^{-1}+\sum_n^N\left\langle(x_n-\mu_m)(x_n-\mu_m)^T u_{nm}s_{nm}\right\rangle_{\mu,u,s}\\&=W_0^{-1}+\sum_n^N\langle u_{nm}\rangle\langle s_{nm}\rangle\left(x_nx_n^T-2x_n\langle\mu_m\rangle^T+\left\langle\mu_m\mu_m^T\right\rangle\right)\end{aligned} \tag{3-41}$$

且

$$\eta_m=\eta_0+\sum_n^N\langle s_{nm}\rangle \tag{3-42}$$

这样因子$q(\Lambda_m)$的最优值依赖于一些矩，这里是$\langle u_{nm}\rangle,\langle s_{nm}\rangle,\langle\mu_m\rangle$和$\left\langle\mu_m\mu_m^T\right\rangle$，而这些矩是在其他因子在求变分后验分布的求得的。可见变分分布是通过各因子之相关迭代计算的。

而对于各混合成分的自由度，无先验的参数v_m，可以通过置下界L的梯度为0，然后（独立）求解以下非线性方程，

$$1+\frac{1}{\tilde{s}_m}\sum_n^N\langle s_{nm}\rangle\left[\langle\ln u_{nm}\rangle-\langle u_{nm}\rangle\right]+\ln\frac{v_m}{2}-\Phi(\frac{v_m}{2})=0 \tag{3-43}$$

其中，

$$\tilde{s}_m=\sum_n^N\langle s_{nm}\rangle \tag{3-44}$$

且$\Phi(\cdot)$为 di-gamma 函数。在 MATLAB 中可以用 *fzero* 函数求解（3-43）。

3.2.3 变分下界

变分学生 t 混合模型的下界可以写成，

$$L(q)=\langle \ln p(x\mid \mu,\Lambda,u,s)\rangle+\sum_{m}^{M}\langle \ln p(\mu_m\mid m_0,\rho_0)\rangle$$
$$+\sum_{m}^{M}\langle \ln p(\Lambda_m\mid W_0,\eta_0)\rangle+\langle \ln p(u\mid v)\rangle+\langle \ln p(\pi\mid \alpha)\rangle+\langle \ln p(s\mid \pi)\rangle$$
$$-\sum_{m}^{M}\langle \ln q(\mu_m)\rangle-\sum_{m}^{M}\langle \ln q(\Lambda_m)\rangle-\langle \ln q(u)\rangle-\langle \ln q(\pi)\rangle-\langle \ln q(s)\rangle \qquad (3\text{-}45)$$

这样，下界的每个部分可以通过变分后验因子的每个矩所得到，这些矩在变分推断过程中已经被计算过了，因而能够被高效地计算。具体细节见附录 2 中 VB-SMM 变分推断公式 (B.23-B.33)部分。

3.2.4 算法步骤

算法步骤总结为：

表 3-2 变分贝叶斯学生 t 混合模型算法步骤

Step1 确定高斯混合个数 K, 设定先验分布超参数$\alpha,W_0,\eta_0,m_0,\rho_0$，设较小值;
Step2 初始化隐变量$\{r_{nm}\}$，可以用 K-means 算法初始化；初始化隐变量$\{u_{nm}\}$，由于只需要知道u的矩，因而可只计算$E[u]$和$E[\ln u]$，可用$E[r_{nm}]$估计;
Step3(VBM-Step) 用隐变量$\{r_{nm}\},E[u],E[\ln u]$计算参数μ_m,Λ_m,π和超参数$\alpha,W_m,\eta_m,m_m,\rho_m$的新值，并独立解非线性方程求$v_m$;
Step4(VBE-Step) 用参数μ_k,Λ_k,π和超参数$\alpha,W_m,\eta_m,m_m,\rho_m,v_m$计算隐变量$\{r_{nm}\}$和$\{u_{nm}\}$;
Step5 计算变分下界 L,若$L(t)-L(t-1)<\delta$则结束，否则继续执行 Step3,Step4;

3.2.5 无限混合模型

学生 t 混合模型的一个不足之处在于需要知道混合成分的个数，这是所有有限混合模型都存在的问题。如果个数选择不当，模型将会欠拟合或者过度拟合数据。在贝叶斯方式中，最常用的方法是基于模型选择规则的方法，如贝叶斯信息准则（BIC, Bayesian Information Criterion）[34]，伪似然信息准则（PLIC,

Pseudo-likelihood information criterion）[35]。然而使用这些准则通常需要经过模型的训练。

针对成分个数选择问题，还有一类高效的非参数贝叶斯统计方法。该方法的基本思想是用单个能调节其复杂性的模型拟合数据，而不是将所有不同模型拟合数据，然后比较它们之间的复杂性。在所有非参数化贝叶斯统计方法中，Dirichlet 过程混合模型（DPM, Dirichlet process mixture）[30]受到了最为广泛的关注。

Dirichlet 过程是基于 Dirichlet 分布而生产，既为一随机过程，又是一个随机概率测度，可表示为 $DP(\alpha, G_0)$。该过程应用于聚类问题时，能够自动确定聚类数目和生成聚类中心的分布参数。限于篇幅，本文不展开介绍。直接给出在学生 t 混合模型中的应用结果。

无限学生 t 混合模型（iSMM, infinite Students' t-Mixture Model）的构造基于 Stick-Breaking 先验[36]，Stick-Breaking 先验是一种能明确表示 DPM 模型中 G 的方法。考虑两个独立随机变量的无限集 $V=\{V_j\}_{j=1}^{\infty}$ 和 $\{\Theta_j\}_{j=1}^{\infty}$，其中 $V_j \sim Beta(1,\alpha)$（Beta 分布见 A.5），$\Theta_j \sim G_0$。Stick-Breaking 法可以表示 G 为，

$$G=\sum_{j=1}^{\infty}\pi_j(V)\delta_{\Theta_j} \tag{3-46}$$

其中，

$$\pi_j(V)=V_j\prod_{i=1}^{j-1}(1-V_i) \tag{3-47}$$

满足 $\sum_{j=1}^{\infty}\pi_j(V)=1$。在 iSMM 模型中，让 $\{\pi_j(V)\}_{j=1}^{\infty}$ 表示混合系数的无限维向量，$\{\Theta_j\}_{j=1}^{\infty}=\{\mu_j,\Lambda_j,v_j\}_{j=1}^{\infty}$ 表示相关成分的参数，这样可以将概率密度函数写成，

$$p(X)=\prod_{n=1}^{N}\sum_{j=1}^{\infty}\pi_j(V)\cdot St(x_n\mid\mu_j,\Lambda_j,v_j) \tag{3-48}$$

另外，引入连续型隐变量 u_{nj} 表示精度的尺度，以及隐变量 z_n 表示观测量 x_n 属于哪个成分。下面说明各参数的共轭先验分布。与有限学生混合模型类似，均值 μ 的先验为高斯分布，逆协方差矩阵的先验为 Wishart 分布，合写成 Gaussian -

Wishart 分布，即

$$p(\mu_j,\Lambda_j)=N(\mu_j\mid m_j,\lambda_j\Lambda_j)W(\Lambda_j\mid W_j,\rho_j) \tag{3-49}$$

其中$\{m_j,\lambda_j,W_j,\rho_j\}$为超参数。另外，需要额外说明的是$V_j$的分布中的参数$\alpha$，可以通过 gamma 先验分布描述，即

$$p(\alpha)=Gam(\alpha\mid\eta_1,\eta_2) \tag{3-50}$$

其中超参数$\{\eta_1,\eta_2\}$同样通过变分推断得到，见附录 B. 34– B.52。

最后，可以用概率图模型来表示相关变量之间的关系，如图 3-6 所示。

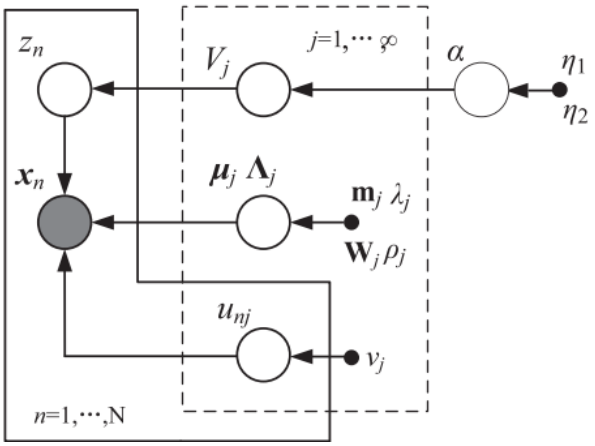


图 3-6 无限学生 t 混合模型的概率图模型

这样可以写出观测量 X 与所有随机变量$\Phi=\{Z,\mu,\Lambda,u,V,\alpha\}$的联合分布，

$$P(X,\Phi)=p(X\mid Z,\mu,\Lambda,u)p(\mu,\Lambda)p(u)p(Z\mid V)p(V\mid\alpha)p(\alpha) \tag{3-51}$$

相关的超参数$\Omega=\{m_j,\lambda_j,W_j,\rho_j,\eta_1,\eta_2\}$可以通过变分推断计算其值，其过程与有限学生 t 混合模型一致，详见附录 2 的 iSMM 模型相关公式（B.-B.）。

同样的，可以写出 iSMM 模型的下界为，

$$\begin{aligned}L=&E[\ln p(X\mid Z,\mu,\Lambda,u)]+E[\ln p(\mu,\Lambda)]+E[\ln p(u)]\\&+E[\ln p(Z\mid V)]+E[\ln p(V\mid\alpha)]+E[\ln p(\alpha)]-E[\ln q(\mu,\Lambda)]\\&-E[\ln q(u)]-E[\ln q(Z\mid V)]-E[\ln q(V\mid\alpha)]-E[\ln q(\alpha)]\end{aligned} \tag{3-52}$$

其中$E[\cdot]$的每个表达式见附录 2（B.53-B.63）.

无限学生混合模型与 SMM 算法步骤相似，但在 VBM-Step 不是计算参数π及

其超参数α，而是计算参数V_j及其超参数$\{\beta_{j1},\beta_{j2}\}$，参数$\alpha$及其超参数$\{\tilde{\eta}_1,\tilde{\eta}_2\}$；然后在 VBE-Step，用以上所求参数与超参数计算隐变量$\{r_{nm}\}$和$\{u_{nm}\}$，反复迭代直到收敛。

3.3 混合模型应用于图像分割

理论上讲，具有无限个混合成分的模型能够对任何数据进行建模。用混合模型对图像数据建模，在分割任务中，将每一个成分作为一类。对混合模型参数估计的过程便是对数据点进行聚类的过程。即通过估计参数，得到数据点属于某一类的条件概率，通过最大化后验概率，得到聚类结果。而对于脑图像，针对的是体素级的图像分割，即将体素聚类为灰质，白质和脑脊髓三类。这样聚类的结果便是图像分割的结果。

3.4 脑部 MR 仿真实验

脑图像分割的主要任务是将 MR 图像分割成灰质(GM, gray matter)，白质(WM, white matter)，脑脊髓(CSF, cerebrospinal fluid)。然而在实际应用中，由于射频场的不均匀性等因素，导致脑 MR 图像的灰度均匀性变差；其次，脑 MR 图像成像过程中由于受仪器设备等物理原因影响，使得图像中经常含有噪声，影响分割精度。本节将用上述变分混合模型对脑 MR 图像进行分割。

3.4.1 实验数据

仿真的脑部 MR 图像数据来自 Internet Brain Segmentation Repository(IBSR)[38]，包括 20 个低分辨率的临床 T1 加权脑部 MR 图像。这些 MR 数据是通过两个不同的图像系统扫描得来的。在 1.5 特斯拉的西门子核磁共振系统（Iselin, NJ）中，用 10 个 FLASH 扫描 4 名男性和 4 名女性。其参数设置为 TR = 40msec, TE = 8mesc, flip angle = 50 度，Field of view = 30cm, slice thickness = 3.1mm，图像像素为 256x256，平均值为 1。通过设定密度阈值为高于或低于 99.99%的不同密度值数目，将其伸缩到[0,255]，这样所有的图像从 16bit 转化为 8bit.

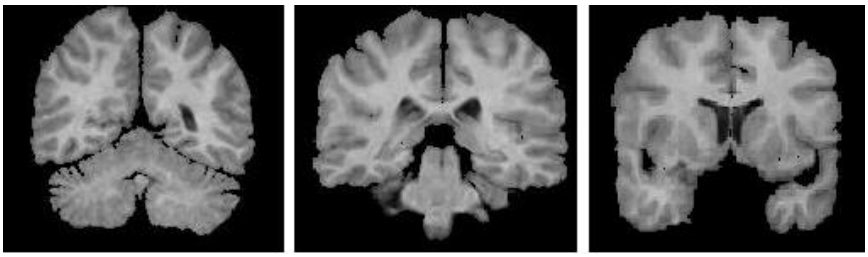


图 3-7 (a)(b)(c)分别为 IBSR 中下标为 12-3 的第 18,25,38 个 slice

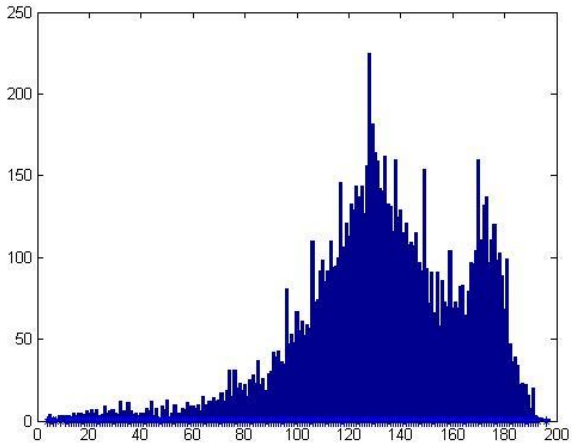


图 3-8 下标为 12-3-38 的 T1 加权 MR 脑部图像的体素直方图

3.4.2 评价指标

IBSR 不仅提供了脑图像，而且还提供了专家人工分割 MR 图像结果，可以认为是真实值（GroundTruth）。这样可以用 Tanimoto 系数或者 Jaccard 指数来度量脑中的白质(WM)，灰质(GM)与脑脊髓(CSF)的分割效果。

Jaccard 指数（Jaccard Index），也称为 Jaccard 相似性系数（Jaccard similarity coefficient），是用于比较两个集合相似性的统计量。定义为两个集合的交集个数比并集个数，即

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|} \qquad (3\text{-}53)$$

其值越大表示越相似。或者采用 Jaccard 距离表示两个采样集合的相异性（dissimilarity）定义两个集合不同元素个数比上集合总数，表示为，

$$J_{\delta}(A,B)=1-J(A,B)=\frac{|A\cup B|-|A\cap B|}{|A\cup B|} \tag{3-54}$$

在 matlab 中，可以用 pdist 函数求 jaccard 距离。

3.4.3 实验结果

先用 12-3-38 这一个 silce 做实验，图 3-9(a)为该切面的真实值（GroundTruth），其中黑色为背景，蓝色为灰质（GM），绿色为白质（WM），红色为脑脊髓（CSF）。图 3-9(b)是基于 EM 的高斯混合模型，图 3-9(c)(d)(e)分别是基于变分贝叶斯的高斯混合模型，无限和无限学生 t 混合模型。

从结果上看，四个算法与 groundTruth 都有一定的差距，主要的问题在于去噪性不好，分割后的图像受噪音影响比较严重。但从算法的时间空间效率上看，基于变分贝叶斯的算法的参数估计方法比基于期望最大化参数方法的迭代次数要少。也就是说，虽然一次迭代用变分贝叶斯估计参数的空间代价和时间代价都比 EM 大一些，但是由于每次迭代 VB 方法的梯度下降得比 EM 快，因此收敛地较快，迭代次数要少很多。

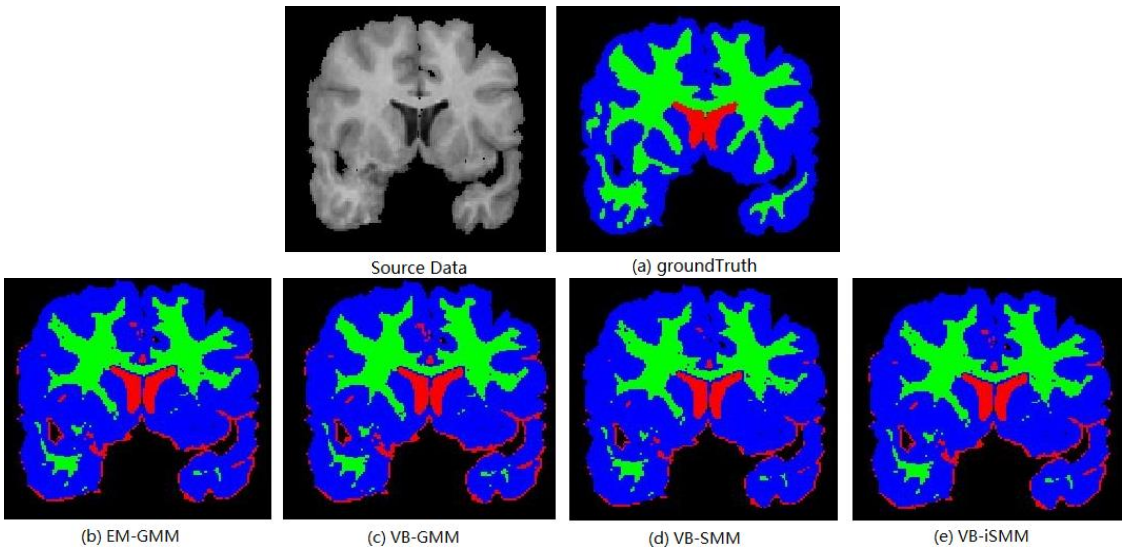
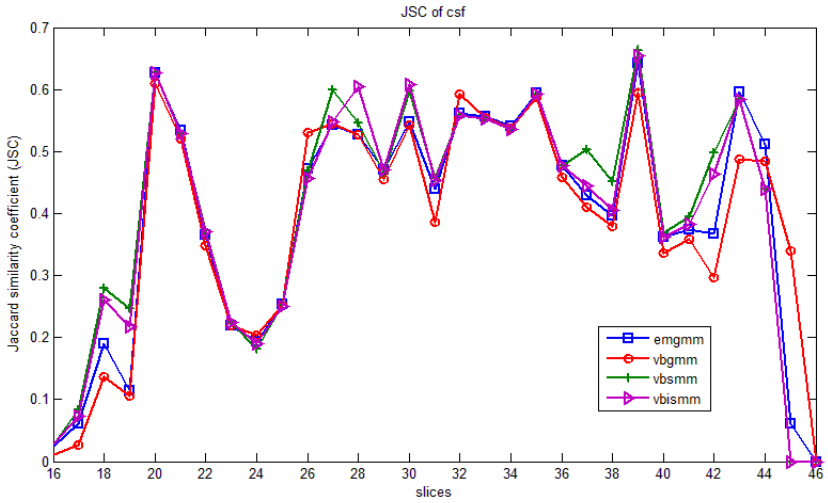


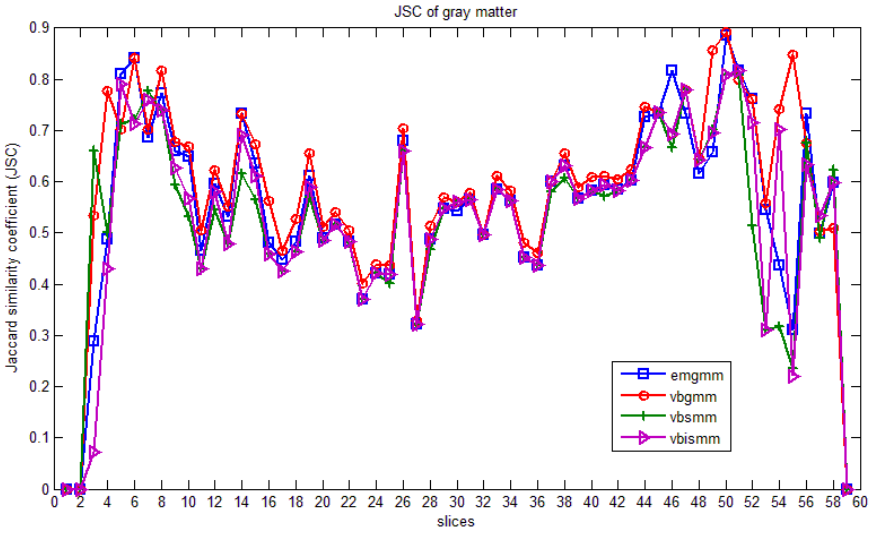
图 3-9 (a)为 12-3-38 的专家人工分割真实值(groundTruth)；(b)(c)(d)(e)分别为 EM-GMM, VB-GMM,VB-SMM,VB-iSMM 算法的分割结果

图 3- (c)(d)(e)都基于变分推断，从理论上讲，用 t 分布应该比高斯分布的鲁棒性更强，能更好地处理离群点。然而，在这个 MR 脑图像例子中，有限/无限 t 混合模型并没有发挥出它的优势。其可能的原因是混合模型已经无法准确刻画脑图像的灰度分布，从图 3-8 的体素直方图可明显看出两个峰，即两个成分，且这两个成分明显有较多的重叠部分。为此可采纳的预处理技术是图像增强（image enhancement）或者直方图均衡（histogram equalization），增强图像的全局对比度，通过这种方法，亮度可以更好地在直方图上分布。另外，也可以考虑用混合的混合（mixture of mixture）来处理，即认为每一类都由一个高斯混合模型构成而不是单个高斯。

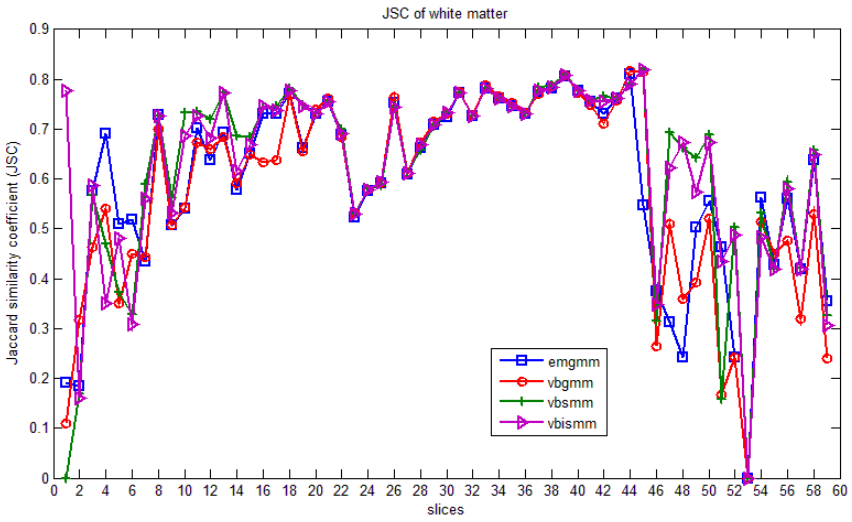
接下来考虑一次 Run/Scan，即下标为 12-3 的扫描数据，共包括 65 个 slices。图 3-10 为 emgmm,vbgmm,vbsmm 和 vbismm 算法的分割结果，横坐标表示每一个 slice，纵坐标为 Jaccard 相似度，值越大越好。其中需要注意的是脑脊髓的横坐标，前一段和后一段都截断了，这是因为核磁共振扫描仪还没扫到脑脊髓，其值可不考虑。单单从结果看，并没有特别突出的模型，四个算法的分割结果相差不大。由于上述算法都是基于模型(model-based)的聚类，而不像基于相似性（similarity-based）的算法考虑空间一致性，这是需要改进的地方。在下一章，我们将讨论基于流形学习的方法，即考虑条件概率分布的流形结构，或许能够通过流形结构的局部一致性达到更好的聚类分割结果。



(a) 脑脊髓(csf)的 JSC



(b) 灰质的 JSC



(c) 白质的 JSC

图 3-10 (a)(b)(c)分别为脑脊髓（CSF），灰质（GM）和白质（WM）用各算法分割 12-3 的 T1 加权像的精度（Jaccard 相似度）

3.5 VB 与 EM 算法效率比较

从脑图像分割效果看，基于 VB 与基于 EM 的混合模型相比并没有体现出明显的优势。但就算法效率看，变分贝叶斯显著优于 EM 算法。再次考虑下标为 12-3 的 Scan，共包括 65 个 slices 的扫描数据。分别比较了 VB 与 EM 算法分割每个 slices 所需的迭代次数，一次迭代所需时间以及总共消耗时间，如图 3-11(a)(b)(c)所示。

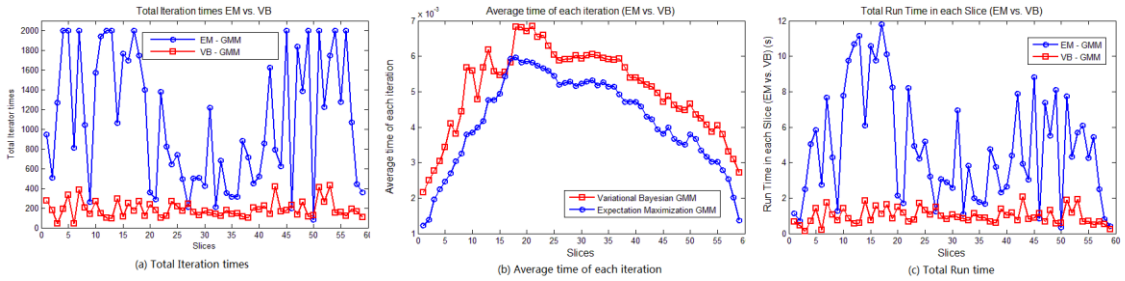


图 3-11 (a)(b)(c)分别为比较 EM 与 VB 算法（蓝色圆圈表示 EM 算法，红色方块表示 VB 算法）的迭代次数，一次迭代所需时间和分割每个 slice 所需总时间

结果表明，变分贝叶斯能够极大程度地降低似然函数收敛所需要的迭代次数，只需要 EM 算法的 1/10~1/2。且对不同的 slice，即不同的观测数据，其 VB 迭代次数都稳定在 150~350 次左右，而 EM 算法则在 200~2000 次不等，可见变分贝叶斯鲁棒性更强。然而一次迭代所需时间，变分贝叶斯由于其算法本身复杂性，比 EM 算法略长。但从总运行时间看，VB 显著短于 EM 算法。可见采用高斯混合模型的脑图像分割而言，变分贝叶斯显著优于期望最大化。

更进一步分析基于 EM 与 VB 的 GMM 算法复杂性。假设有 N 个数据，每一个数据有 d 维，分成 K 个 clusters，需要迭代 t 次。则对于 EM 算法，用 K-means 初始化需要 $O(t_1KN)$ 次操作。在 E-step 中计算协方差的行列式需要 $O(d^3)$ 次操作，需要对每个 clusters 的每个数据点计算，则共需要 $O(KNd^3)$ 次操作；在 M-step 步骤中，对混合系数，均值和协方差的更新分别需要 $O(KN)$，$O(KN)$，$O(KNd)$ 次操作。因此，EM 算法的计算复杂度为 $O(tKNd^3)$。

对于变分贝叶斯方法，在 M-step 中进行 Cholesky 分解需要 $O(d^3)$ 次操作，采用相同方法分析可得其计算复杂度也为 $O(tKNd^3)$。就一次迭代而言，VB 方法复杂度与 EM 算法相同，都为 $O(KNd^3)$。然而实际上，其计算代价并不相同，VB 在 E-step 和计算对数似然函数所需要的计算步骤比 EM 更多，这也解释了图 3-11(b) 的结果，即 VB 比 EM 的平均迭代时间更长，但复杂度相同，因此总的来说两者平均迭代时间呈线性关系。另外由于 VB 的迭代次数 t 大大小于 EM 的迭代次数，因此出现了图 3-11(a),(c)的结果。

3.6 本章小结

本章用变分贝叶斯方法对混合模型进行了参数估计。首先针对高斯混合模型进行了详细的变分推断，并与 EM 算法进行了比较。然后变分推断了学生 t 混合模型。在此基础上，考虑到有限混合模型需要预先知道混合成分的个数，于是用 DPM 的非参数化方法将学生 t 混合模型延伸到无限混合，使得能自动确定混合成分个数。最后将变分混合模型应用到脑部 MR 图像分割中。结果表明变分贝叶斯方法显著优于 EM 算法。

第四章 基于 Laplacian 正则化变分混合模型的脑图像分割

近年来随着流形学习理论的不断完善，一系列基于流形学习的聚类算法得到显著的发展。假定产生数据的概率分布拥有流形结构（Manifold Structure），如果两个点在概率分布的流形空间中比较接近，那么他们的条件概率分布是相似的。基于此，Xiaofei He 等[9]提出基于流形结构的 Laplacian 正则化高斯混合模型，数据的流形结构用最近邻图建模，最终得到在流形几何中平滑变化的条件后验概率，能较好地进行数据聚类。本章将详细介绍 Laplacian 正则化方法，并将其与混合模型结合，应用到脑部 MR 图像分割中。

4.1 Laplacian 正则化

4.1.1 理论基础

如果两个点 $x_1, x_2 \in X$ 在 P_X 的内在空间中相近，那么他们的条件概率密度 $P(y \mid x_1)$ 和 $P(y \mid x_2)$ 是相似的，其中 $y \in \{1,...,K\}$ 表示属于哪个类别。就是说，条件分布 $P(y \mid \cdot)$ 的平滑度随着 P_X 的内在空间测地线的不同而变化的。这正是流形假设（manifold assumption）,该假设能使各种算法降维或者得到半监督学习算法。

假设有 K 个类，令 $f_k(x) = P(y = k \mid x) P_k(x)$ 为条件概率密度函数，用 $\| f_k \|_M$ 度量 f_k 的平滑度，当考虑 P_X 的基是紧子流形（compact submanifold）$M \subset R^d$，那么 $\| f_k \|_M$ 的一个常用选择是，

$$\int_{x \in M} \| \nabla_M f_k \|^2 dP_X(x) \tag{4-1}$$

它等价于

$$\int_{x \in M} L(f_k) f_k dP_X(x) \tag{4-2}$$

其中 ∇_M 表示 f_k 在流形空间中的梯度，L 为 Laplace-Belerami 算子，$Lf = -div\nabla(f)$。通过最小化 $\|\nabla_M f_k\|^2$，就能得到充分光滑的条件概率密度函数。然而，数据的流形结构通常是未知的，因而 $\|\nabla_M f_k\|^2$ 是不可计算的。为了建立 M 的几何结构，可通过构造最近邻图 G。对于每个数据点 x_i，我们可以找到它的 p 个与之最近的邻居（p nearest neighbors），对 x_i 与这些邻居连边。

有许多定义图的权重矩阵 S 的方法。最常用的包括 0-1weighting, Heat kernel weighting 和 Dot-product weighting。定义 $L = D - S$，其中 D 为对角阵，$D_{ii} = \sum_j S_{ij}$，L 称为 Laplacian 图（graph laplacian）。根据图谱理论(Spectral Graph Theory)[37], $\| f_k \|_M^2$ 可写成如下离散形式，

$$\begin{aligned} R_k &= \frac{1}{2}\sum_{i,j=1}^{m}(P(k\mid x_i) - P(k\mid x_j))^2 S_{ij} \\ &= \sum_{i,j=1}^{m} P(k\mid x_i)^2 D_{ii} - \sum_{i,j=1}^{m} P(k\mid x_i)P(k\mid x_j)S_{ij} \\ &= f_k^T D f_k - f_k^T S f_k \\ &= f_k^T L f_k \end{aligned} \tag{4-3}$$

其中,

$$f_k = (f_k(x_1),..., f_k(x_m))^T = (P(k\mid x_1),..., P(k\mid x_m))^T \tag{4-4}$$

通过最小化 R_k 我们可以得到在数据流形空间中充分光滑的 f_k。也就是说最小化 R_k 的过程是保证如果 x_i 和 x_j 比较接近,那么 $f_k(x_i)$ 和 $f_k(x_i)$ 也是接近的。值得强调的是图中的 Laplacian 算式与流形空间中的 Laplace-Beltrami 算子是相似的。

现在定义正则化的对数似然函数如下，

$$L(\Theta) = \log P(X\mid\Theta) - \lambda\sum_{k=1}^{K} R_k \tag{4-5}$$

其中 λ 为用于控制 $f_k(k=1,...,K)$ 平滑度的归一化参数。需要极大化似然函数（4-5），然而对于（4-5）并没有闭合解，因此将左右两项分开考虑。对于第一项依然采用 EM 算法，而对于第二项采用 Netwon-Raphson 法。考虑函数 $f(x)$ 和它的初始值 x_t，牛顿更新方程为，

$$x_{t+1} = x_t - \gamma \frac{f'(x)}{f''(x)} \tag{4-6}$$

其中$0 \le \gamma \le 1$为 step 参数。通过对$f_k^T L f_k$求关于$P(k \mid x_i)$的一阶和二阶导，可得$P(k \mid x_i)$的更新方程为，

$$P(k \mid x_i) \leftarrow (1-\gamma)P(k \mid x_i) + \gamma \frac{\sum_{j=1}^{m} S_{ij} P(k \mid x_j)}{\sum_{j=1}^{m} S_{ij}}, i = 1, ..., m \tag{4-7}$$

4.1.2 算法描述

对于用 EM 算法估计参数的高斯混合模型，我们可以总结如下算法步骤：

表 4-1 Laplacian 正则化的高斯混合模型算法

Step1 置 Step 参数$\gamma = 0.9$；用 K-NN 算法构造最近邻图，计算权重矩阵S；用 K-means 算法初始化参数Θ_0

Step2(E-step) 计算后验概率，$P(k \mid x_i) = \dfrac{\alpha_k^{n-1} p_k(x_i \mid \theta_k^{n-1})}{\sum_{j=1}^{K} \alpha_j^{n-1} p_j(x_i \mid \theta_j^{n-1})}$

Step3(M-step)

1. 用（4-7）式平滑后验概率直到收敛；
2. 估计参数$\alpha_i^n, \mu_i^n, \Sigma_i^n$：

$$\alpha_i^n = \frac{1}{m} \sum_{j=1}^{m} P(i \mid x_j),$$

$$\mu_i^n = \frac{\sum_{j=1}^{m} x_j P(i \mid x_j)}{\sum_{j=1}^{m} P(i \mid x_j)}$$

$$\Sigma_i^n = \frac{\sum_{j=1}^{m} P(i \mid x_j)(x_j - \mu_i)(x_j - \mu_i)^T}{\sum_{j=1}^{m} P(i \mid x_j)}$$

Step4 用(4-5)计算正则化对数似然；如果$L(\Theta^n) < L(\Theta^{n-1})$，$\gamma \leftarrow 0.9\gamma$，转到 **Step3** 如果$L(\Theta^n) - L(\Theta^{n-1}) \le \delta$，则结束，否则转到 Step2。

4.1.3 算法复杂性分析

假设有 N 个数据，每个数据有 d 维，分成 K 个 clusters， p 最近邻图的权重矩阵 S 每一行有 p 个非零数据点，EM 步骤迭代 t_1 次。则（4-7）式需要 $O(KpN^2)$ 次操作，设（4-7）式收敛需要迭代 t_2 次。上章分析过 GMM 每次迭代需要 $O(KNd^3)$，则 lapGMM 的算法复杂度为 $O(t_1KN(d^3+t_2pN))$。当观测数据个数 N 很大时（$N \gg d, N \gg k, N \gg p$），可发现算法效率比一般用 EM 或 VB 的 GMM 要低很多。并且使用 Netwon-Raphson 法迭代平滑后验概率，当迭代次数 t_2 很大时，算法代价也非常高，这时适当降低迭代。

另外，对于高维数据，即 d 很大时， d^3 将成为主要计算代价。这种情况下，可以先用一些数据降维算法（如 PCA）来降低维度。

4.1.4 简单人造实验

用线性不可分实验数据测试 Laplacian 图平滑条件后验概率的有效性，实验结果表示，用 lapGMM 算法能够正确聚类，如图 4-1。

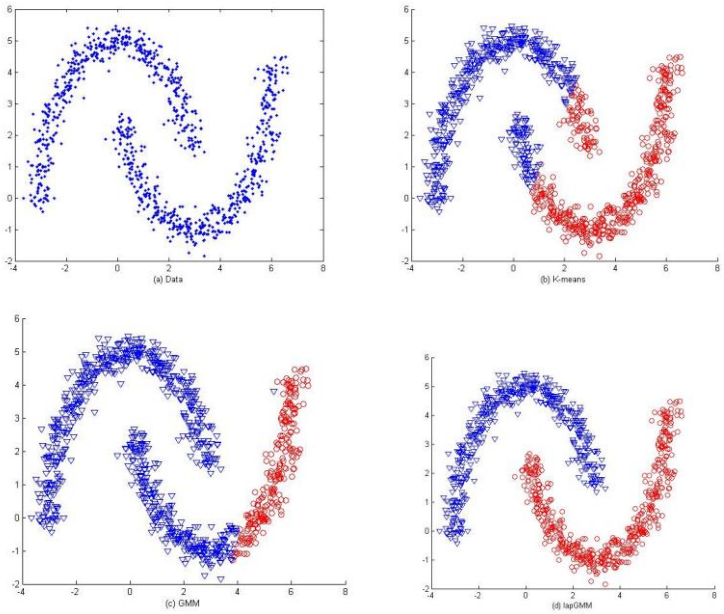


图 4-1 双月环模式聚类 (a)原始数据；(d) K-means 聚类结果；(c) GMM 聚类结果(d) lapGMM 聚类结果

4.2 Laplacian 正则化混合模型

在第三章中，我们用变分贝叶斯方法对高斯混合模型，有限/无限学生 t 混合模型进行了参数估计，实验表明比 EM 算法有更好的效果。接下来，将 Laplacian 正则项加入以上模型中，期望能有更好的分割效果。

分析 4.1 节中将 Laplacian 正则化的高斯混合模型的算法步骤可发现，与普通的 EM 算法步骤的主要区别在于两点：1）多了一个平滑后验概率的步骤。2）计算对数极大似然函数值多加一个 Laplacian 正则项。而 Laplacian 正则项只与后验概率 $P(k|x)$ 和权重矩阵 S 相关。对于一组数据，其权重矩阵 S 是确定的。因此不论是何种混合模型，不论是用哪种方法进行参数估计，若加入 Laplacian 正则项，其主要不同之处在于平滑后验概率。在变分贝叶斯方法中，不是高斯混合，有限或无线学生 t 混合模型，其隐变量 r_{nk} 即为后验概率 $P(k|x)$。

Laplacian 正则化变分高斯混合模型和学生 t 混合模型算法步骤可总结为

表 4 -2 Laplacian 正则化的变分高斯混合模型算法步骤

Step1 确定高斯混合个数 K, 设定先验分布超参数 $\alpha_0,\beta_0,w_0,\upsilon_0,m_0$，设较小值;

Step2 用 K-NN 算法构造最近邻图，计算权重矩阵 S；

Step3 初始化隐变量 r_{nk}，可以用 K-means 算法初始化；

Step4(VBE-Step) 用参数 μ_k,Λ_k,π 和超参数 $\beta_k,m_k,w_k,\upsilon_k,\alpha_k$ 计算隐变量 r_{nk}；

Step5(VBM-Step)

1. 用（4-7）式计算隐变量 r_{nk} 直到收敛；
2. 用隐变量 r_{nk} 计算参数 μ_k,Λ_k,π 和超参数 $\beta_k,m_k,w_k,\upsilon_k,\alpha_k$ 的新值;

Step6 用（4-5）计算正则化对数似然；如果 $L(\Theta^n)<L(\Theta^{n-1})$，$\gamma\leftarrow 0.9\gamma$，转到 **Step5**

如果 $L(\Theta^n)-L(\Theta^{n-1})\le\delta$，则结束，否则转到 **Step4**。

表 4 -3 Laplacian 正则化的变分学生 t 混合模型算法步骤

Step1 确定高斯混合个数 K，设定先验分布超参数 $\alpha, W_0, \eta_0, m_0, \rho_0$，设较小值;

Step2 用 K-NN 算法构造最近邻图，计算权重矩阵 S；

Step3 初始化隐变量 $\{r_{nm}\}$，可以用 K-means 算法初始化；初始化隐变量 $\{u_{nm}\}$，由于只需要知道 u 的矩，因而可只计算 $E[u]$ 和 $E[\ln u]$，可用 $E[r_{nm}]$ 估计;

Step5(VBE-Step)

用参数 μ_k, Λ_k, π 和超参数 $\alpha, W_m, \eta_m, m_m, \rho_m, \nu_m$ 计算隐变量 $\{r_{nm}\}$ 和 $\{u_{nm}\}$；

Step4(VBM-Step)

1. 用（4-7）式隐变量 r_{nk} 直到收敛,并更新 $E[u], E[\ln u]$；
2. 用隐变量 $\{r_{nm}\}, E[u], E[\ln u]$ 计算参数 μ_m, Λ_m, π 和超参数 $\alpha, W_m, \eta_m, m_m, \rho_m$ 的新值，并独立解非线性方程求 ν_m；

Step6 用（4-5）计算正则化对数似然；如果 $L(\Theta^n) < L(\Theta^{n-1})$，$\gamma \leftarrow 0.9\gamma$，转到 **Step5** 如果 $L(\Theta^n) - L(\Theta^{n-1}) \le \delta$，则结束，否则转到 **Step4**。

4.3 模型选择

模型选择在大多数学习问题中都是非常关键的问题。在很多情况下，不同的模型参数值能在很大程度上影响模型性能。同样在 Laplacian 正则化模型，需要确定两个模型参数，即最近邻图的邻居个数 p 和正则化系数 λ。Xiaofei He 等[9]讨论了这两个参数对聚类效果的影响。用 USPS,COIL2 和 TDT2[39]数据集做测试，比较了不同的 p 值和不同 λ 对聚类效果，同时将 laplacian GMM 算法与 GMM, K-Means, PCA+K-Means, NMF 聚类性能进行比较。结果表明，邻居个数 p 在 5-12 之间，正则化系数 λ 在 500-10^5 之间，都能得到较好的分割效果。以下实验选择 $p=10$ 和 $\lambda=1000$。另外，最近邻图采用 0-1 weighting 方法构造权重矩阵 S。

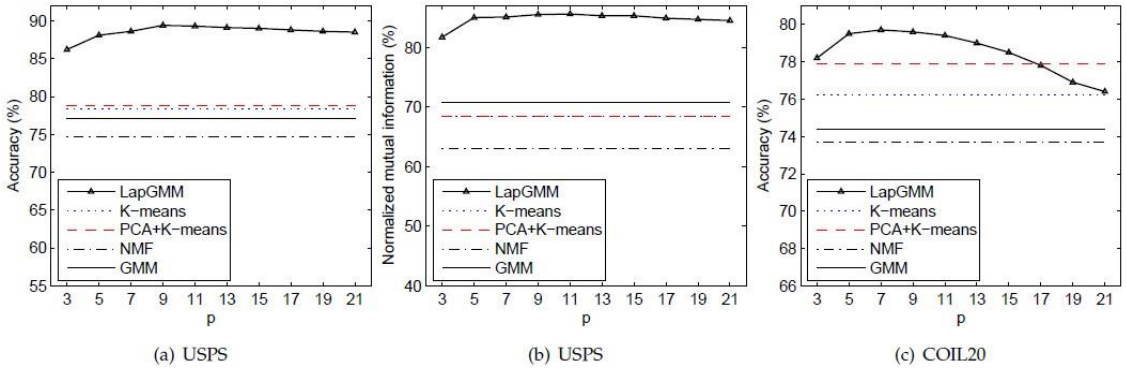


图 4-2 不同的邻居个数 p 下 lapGMM 的聚类精度[9]

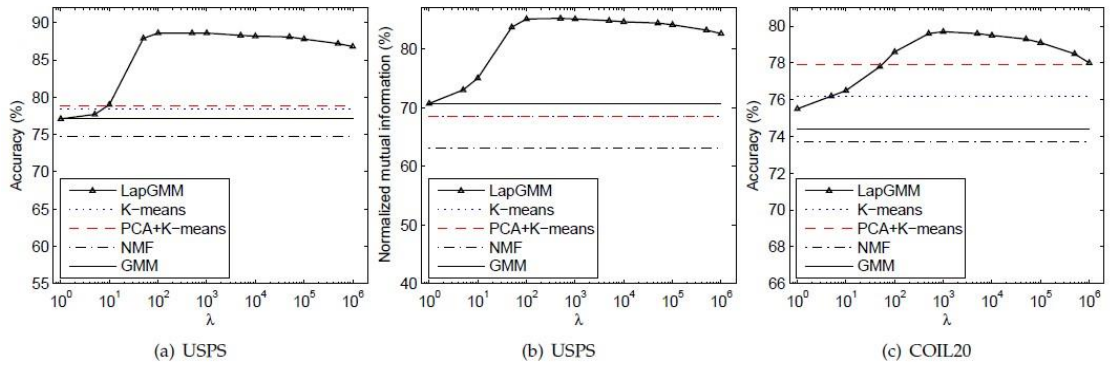


图 4-3 不同的正则化系数 λ 下 lapGMM 的聚类精度[9]

4.4 脑图像分割实验

用与 3.3 节相同的图像数据进行实验。用标号为 12-3-38 的切面进行图像分割，得到图 4-4 的分割结果。其中 EM-GMM 模型加入 laplacian 正则项后，优化效果比较明显,可见由于利用了条件概率分布的流行结构性质，加入一些流形结构空间信息，起到了一定的平滑去噪的作用。但对于其他 VB-lapGMM,VB-lapSMM 和 VB-lapiSMM 虽然有一定的去噪作用，但不够显著。其原因可能在于用牛顿法最小化梯度算子$\|\nabla_M f_k\|^2$。一般情况下牛顿迭代收敛速度非常快，但如果初始值不合适（初始近似值要在根的附近才能保证收敛），则可能无法收敛。鉴于基于 Laplacian 正则化的算法收敛速度相对较慢，因此将每次迭代光滑后验概率的精度降低（如 1e-5）。因为只需要保证每次比上次的对数似然值更大，而不需要每次都是最优值，因此该方案是可行的。

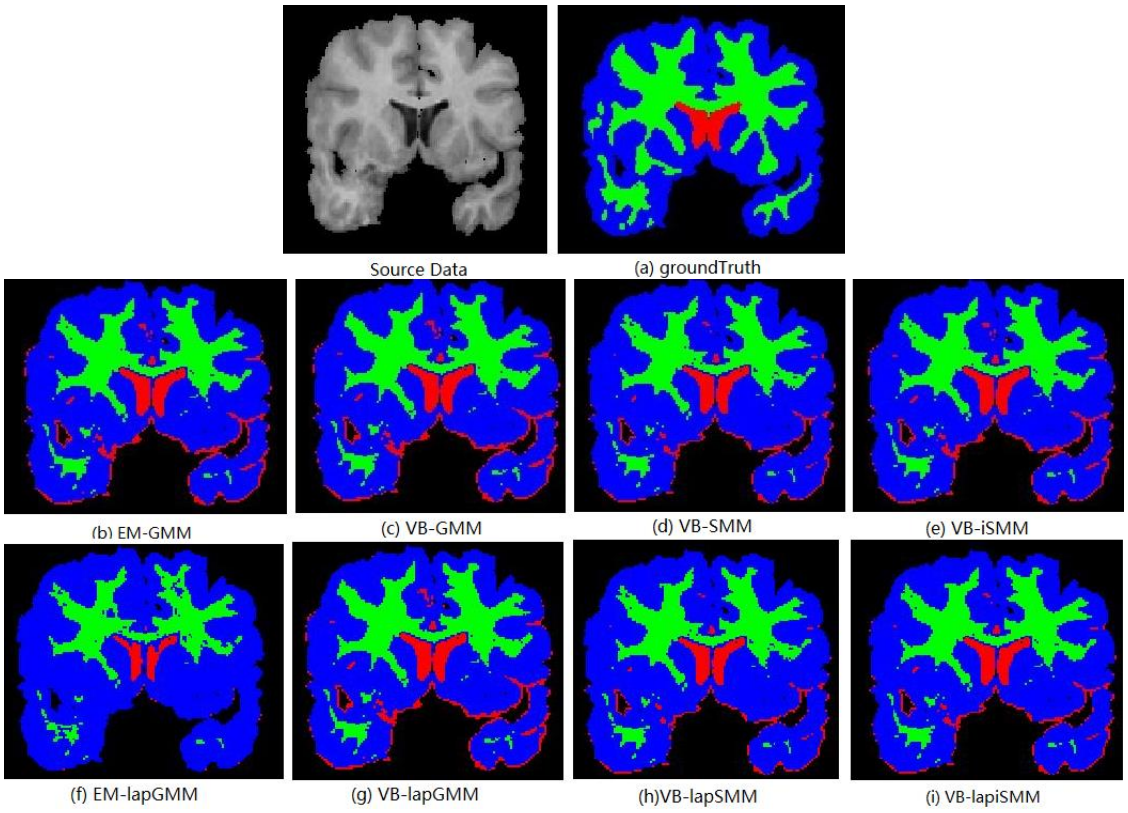


图 4-4 (a)为 12-3-38 的专家人工分割真实值(groundTruth)；第二排(b)(c)(d)(e)分别为 EM-GMM, VB-GMM,VB-SMM,VB-iSMM 算法的分割结果；第三排(f)(g)(h)(i)是加入 Laplacian 项

这里需要仔细分析一下 Laplacian 收敛速度慢的原因。由于需要平滑的条件概率密度函数的个数为 $N \times k$（N 为数据点个数，k 为聚类个数），当要求收敛的精度较高的时候，实际上需要迭代的次数是很多的，因而收敛速度较慢。分析以上算法，当过渡平滑时，有可能使得对数似然值变小，当这种情况发生时，降低 Step 参数 γ 的值。注意此时使用的是牛顿下山法，通过控制 Step 参数保证函数数值稳定下降。当 $\gamma \to 0$,即完全没有迭代作用。而正是由于这些原因，使得算法的收敛速度变得更慢。

另外，需要特别指出的是，用 Laplacian 图优化的算法中，用 EM 比用 VB 的速度要慢点多。这是由于每次迭代用 VB 其梯度下降得比 EM 快，因此迭代次数 VB 比 EM 少得多，这一点在 3.4 节已经分析过。而现在每一次迭代，最花费时间的是平滑后验概率。并且实验表明，对于相同数量级的精度，用 EM 比 VB 估参

迭代次数多得多。而本章的目的是研究结合 laplacian 图变分模糊模型是否有效。鉴于此，以下跑 Run/Scan 的实验没有用耗时非常严重的 EM-lapGMM 模型。

下图为包括 65 个 slices 的分割结果，横坐标表示每一个 slice，纵坐标为 Jaccard 相似度(JSC)，值越大越好。图 4-5~图 4-8 分别为灰质、白质、脑脊髓和总分割结果的 JSC 值。结果表明，

（1）GMM，SMM 和 iSMM 的灰质分割精度都明显提高，并且基于 GMM 的模型精度提高最显著，如图 4-5 所示。这说明平滑后验概率 $P(k \mid x)$ 是有效的，即“像素值接近的点属于某一类的概率相似”对于灰质而言是有效的。考虑到学生 t 分布具有更重的尾部，已经能较好地处理离群点。因而，使得 laplacian 正则化对其影响不如高斯混合模型深。总的来说，通过平滑后验概率能在一定程度上提高分割精度。

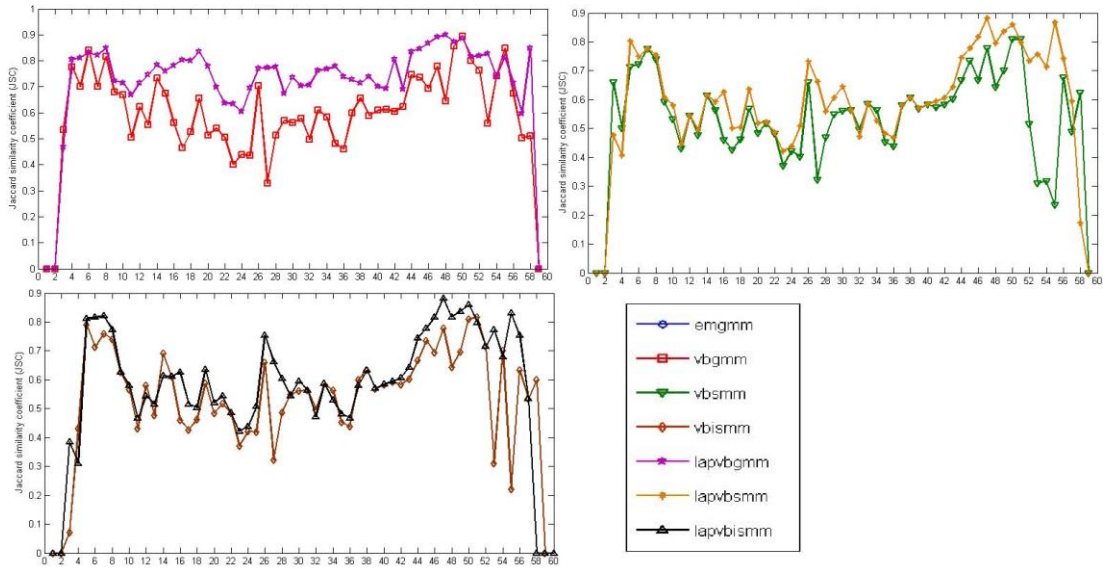


图 4-5 分割 12-3 的 T1 加权像的灰质分割精度，左上角为 VB-GMM 对比 VB-lapGMM,右上角为 VB-SMM 对比 VB-lapSMM，左下角为 VB-iSMM 对比 VB-lapiSMM.

（2）对白质而言，Laplacian 正则化的 GMM 模型分割精度略有提高；而对 SMM 模型， Laplacian 正则化使得分割效果更稳定，尤其在左右两侧（数据点较少的情况），这说明 Laplacian 正则化能较好地处理小数据样本；然而对 iSMM 而言，正则化的作用不大，如图 4-6 所示。

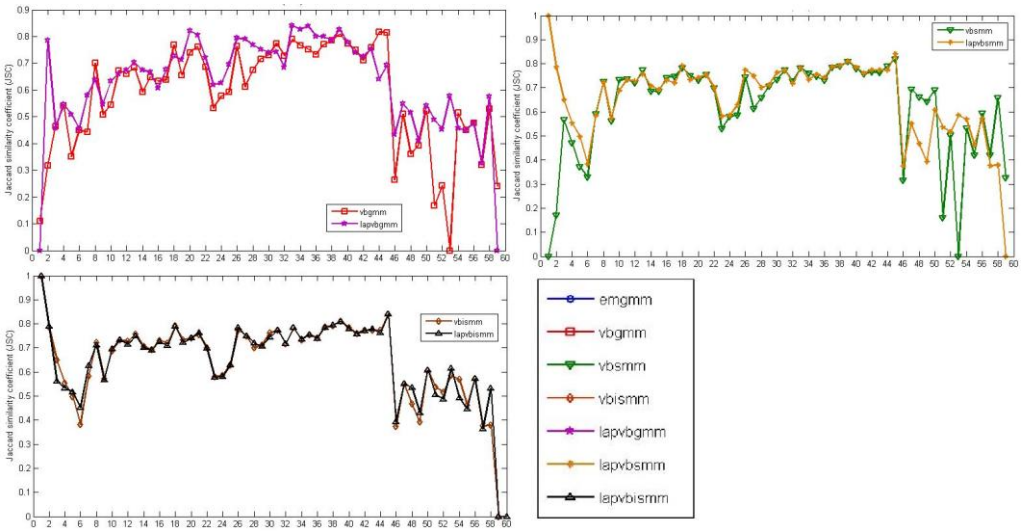


图 4-6 分割 12-3 的 T1 加权像的白质分割精度

（3）从图 4-7 可看出 laplacian 正则化的混合模型同样难以对脑脊髓准确分割，其分割精度不管是哪种模型都不高。甚至在加入正则项后，其分割精度降低了。可能的原因是，在数据预处理过程中，没能消除噪音，而这些噪音的像素值与脑脊髓像素值相接近。这样在流形假设下，laplacian 正则化使得这些噪音更倾向归为脑脊髓这一类，而由于脑脊髓的数据点个数本来就不多，使得 JSC 值较小。一个有效的解决方向是考虑图像的空间结构。虽然如此，注意到脑图像中成分最多的是白质与灰质，在很多情况下，对脑图像仅分割白质与灰质，在这种情况下，该方法依然有效。

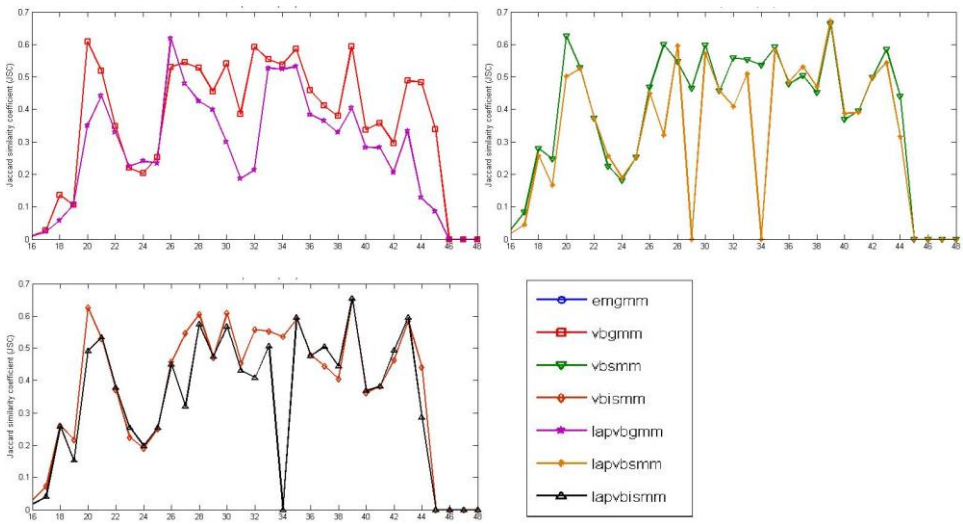


图 4-7 分割 12-3 的 T1 加权像的脑脊髓分割精度

（4）总体来看，laplacian 正则化的 GMM, SMM, iSMM 的分割精度都略有提高，并且不同的数据集其分割精度都能维持在一个较恒定的水平，特别是处理小样本数据，这种优势更明显（即在扫描脑图像的开始于结束阶段时，其切面数据点变少，在无 laplacian 正则项时，其分割精度波动大，且精度低），如图 4-8 所示。可见与一般的 GMM, SMM, iSMM 相比，基于变分贝叶斯的 laplacian 正则化混合模型是一种在精度与稳定性上都表现不错的数据模型。

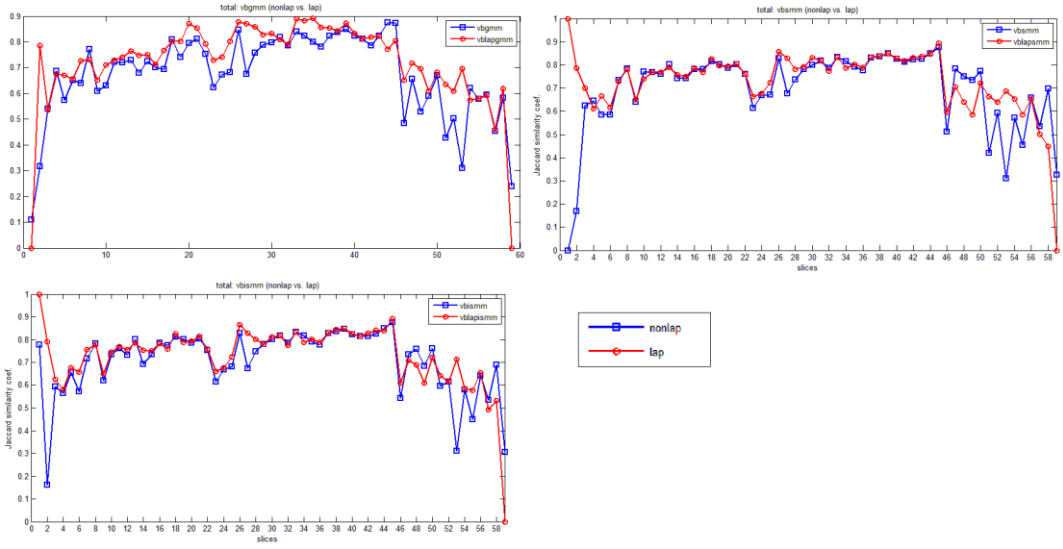


图 4-8 分割 12-3 的 T1 加权像的分割精度

4.5 本章小结

本章首先研究 Laplacian 正则化方法，认为两个点在概率分布的流形空间中接近，那么其条件概率分布是相似的。在此基础上，阐述基于 Laplacian 正则化的混合模型方法，计算复杂度，并用一个简单的线性不可分的例子检验方法的有效性。将其推广到变分混合模型中，并应用到 MR 医学图像中，特别地与第三章中的变分混合模型的分割结果进行了比较。结果表明基于变分贝叶斯的 laplacian 正则化混合模型是一种在精度与稳定性上都表现不错的数据模型。

第五章 总结与展望

5.1 完成的工作

图像分割技术一直是计算机视觉与图像处理过程中的一项关键技术。本文对聚类分割技术展开了研究。特别地对混合模型的参数估计与基于混合模型的聚类分割改进算法进行了较详尽的分析与研究。总结全文，具体包括：

（1）变分贝叶斯方法。从贝叶斯理论开始，对变分推断的数学基础，基本原理以及变分消息传播展开研究。变分贝叶斯作为一类分布估计方法有其特有的优势，也有其固有的不足之处。其优势在于基于平均场假设，利用概率图模型（贝叶斯网络和马尔科夫随机场），使得对变量的分析更加简单且直观。概率图模型中的变分推断和学习方法是一块非常有价值的研究方向。然而美中不足，计算量相对 EM 算法较大。

（2）医学图像分割技术的研究，包括相关分割算法与评价指标的研究。特别地针对功能核磁共振图像（fMRI），理解 fMRI 图像的特点以及图像分割的挑战。研究医学图像数据处理方法，特别是在 MATLAB 环境下的预处理方法，为此专门学习了 SPM8 软件包。

（3）混合模型的研究，包括高斯混合与学生 t 混合模型，有限混合与无线混合模型。混合模型作为一种聚类方法，受到相关广泛的研究，其改进算法不胜举。一般针对某个应用范围进行算法改进，就医学图像分割而言，主要可以分为四类：一是改变模型核，比如将高斯核换成学生 t 核；二是研究其参数估计方法，比如用变分贝叶斯方法替代 EM 算法；三是研究初始化方法，包括 K 均值，遗传算法等；四是从空间结构入手，包括特征空间结构，坐标空间和概率空间，一般从局部一致性和模糊性考虑。另外，也有考虑混合的混合模型。总的来说，各类方法都能在其特定条件下发挥其特有的作用。

（4）流形学习的研究，主要指 Laplacian 正则化方法。将该方法与各种混合模型结合，应用到医学图像分割中，研究其性能。结果表明利用后验概率的流形

结构提高白质和灰质的精度，而对脑脊髓的精度没有提升。

（5）MATLAB 程序设计。就科学计算而言，MATLAB 的程序设计与 C/C++, C#, Java 等程序语言相比要简单得多，其丰富的软件包为解算问题提供了巨大的帮助。但由于其基本数据单位是矩阵，因而将基本运算转化矩阵运算非常必要，因为对 for 运算，MATLAB 的计算与 C/C++相比，要慢得多得多。为此，需要改变惯有的编程思路，同时需要不断地尝试新方法优化问题，在不断尝试中提高 MATLAB 程序设计能力。

5.2 存在的问题及下一步工作

基于变分贝叶斯的混合模型以及基于 Laplacian 正则化的变分混合模型，在脑部 MR 图像分割的应用中，虽然有一定的成果，但仍然存在一定的问题。变分推断与 EM 算法相比，其分割质量未有足够的提升，而基于 Laplacian 正则化的方法其脑脊髓分割精度未显著提高。考虑单个模型来表示一类存在固有的局限性，如果用混合的混合，即用一个混合模型表示一类，或许能提高拟合效果，提高分割精度。考虑到 Laplacian 方法的分割速度比较慢，因此开展快速 Laplacian 迭代方法研究非常必要。另外，对图像像素坐标的局部一致性的研究也非常有意义。

参考文献

[1] Adleman, L. M. (1994). Molecular computation of solutions to combinatorial problems [J], Science-AAAS-Weekly Paper Edition, 266(5187), 1021-1023.

[2] QiOuyang, Kaplun Peter D. DNA Solution of the Maximal Clique Problem [J]. Science, 1997, 278(17): 446-449.

[3] Paun G, Rozenberg G. DNA Computing: New Computing Paradigms [J]. Springer, 1998, 60-63.

[4] Sungchul Ji. The Cell as the Smallest DNA-Dased Molecular Computer [J]. Biosystems, 1999, 52:123-133.

[5] Hayit Greenspan, AmitRuf, Constrained Gaussian Mixture Model Framework for Automatic Segmentation of MR Brain Images [J], IEEE Transactions on Information Technology in Biomedicine, VOL. 25.NO.9,September 2006

[6] Ji, Zexuan, et al. "Fuzzy local Gaussian mixture model for brain MR image segmentation."[J], *Information Technology in Biomedicine, IEEE Transactions on* 16.3 (2012): 339-347.

[7] Tian, GuangJian, et al. "Hybrid genetic and variational expectation-maximization algorithm for Gaussian-mixture-model-based brain MR image segmentation."[J], *Information Technology in Biomedicine, IEEE Transactions on* 15.3 (2011): 373-380.

[8] Nikolaos Nasios, Adrian G. Bors, Variational Learning for Gaussian Mixture Models [J], IEEE Transactions on systems, man, and cybernetics—Part B: cybernetics, Vol. 36, No. 4, august 2006

[9] He, X., Cai, D., Shao, Y., Bao, H., & Han, J. (2011). Laplacian regularized gaussian mixture model for data clustering.[J] *Knowledge and Data Engineering, IEEE Transactions on*, *23*(9), 1406-1418.

[10] J.C. Rajapakse, J.N.Giedd, Statistical approach to segmentation of single-channel cerebral MR images [J], IEEE Trans. on Medical Imaging, 1997, 16:176-186.

[11] W.M.Wells, W.E.L Gimson, Adaptive segmentation of MRI data [J], IEEE Trans.

On Medical Imaging, 1996, 15:429-442.

[12] 傅景广，许刚，基于遗传算法的聚类分析 [J], 计算机工程，2004, 04:122-124

[13] 宋余庆，陈健美，数学医学图像 [M], 北京，清华大学出版社, 2008

[14] 廖亮，林土胜，基于核聚类算法和模糊 Markov 随机场模型的脑部 MR 图像的分割 [J]，中国图象图形学报，2009,09:1732-1738

[15] 张树伟，医学图像的高斯混合模型及聚类研究 [D], 江苏大学，2010:10-14

[16] C.E.Rasmussen, The Infinite Gaussian Mixture Model, in Advances in Neural Information Processing Systems, 2000: 554–560, MIT Press

[17] Tu, Z., & Zhu, S. C. (2002). Image segmentation by data-driven Markov chain Monte Carlo.[J] *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *24*(5), 657-673.

[18] Bishop, Christopher M. *Pattern recognition and machine learning*.[M] Vol. 1. New York: springer, 2006.

[19] Hanson, Kenneth M. "Tutorial on Markov Chain Monte Carlo."[J], *Workshop for Maximum Entropy and Bayesian Methods*. 2000.

[20] RM Neal, Probabilistic inference using Markov chain Monte Carlo methods [R], Technical Report CRG-TR-93-1, University of Toronto, 1993

[21] David MacKay, Developments in probabilistic modelling with neural networks ensemble learning, Neural Networks: Artificial Intelligence and Industrial Applications. Proc. of the 3rd Annual Symposium on Neural Networks, 1995: 191-198

[22] Scholarpedia, Ensemble Learning [Z], http://www.scholarpedia.org/article/Ense-mble_ learning, 2008

[23] Sm ídl, V áclav, and Anthony Quinn. *The Variational Bayes Method in Signal Processing (Signals and Communication Technology).[M]* Springer-Verlag New York, Inc., Secaucus, NJ, 2005.

[24] Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference* (Doctoral dissertation, University of London).

[25] Winn, John, and Christopher M. Bishop. "Variational message passing."[J] *Journal of Machine Learning Research* 6.1 (2006): 661.

[26] Winn, John. "Variational message passing and its applications."*Unpublished doctoral dissertation, Cambridge University* (2003).

[27] Svensén, Markus, and Christopher M. Bishop. "Robust Bayesian mixture modelling." *Neurocomputing* 64 (2005): 235-252.

[28] Wikipedia, Variational Bayesian methods [Z], http://en.wikipedia.org/wiki/Variational_Bayes, 2013

[29] Xin Wei,Cg Li, The infinite Student's t-mixture for robust modeling [J], IEEE Transactions Signal Processing 92(2012) 224-234

[30] R. Neal, Markov chain sampling methods for Dirichlet process mixture models [J], Journal of Computational and Graphical Statistics 9(2) (2000) 249–265.

[31] 亢冬春, 医学超声图像增强技术的研究与设计 [D], 吉林大学, 2007

[32] Cover, Thomas M., and Joy A. Thomas. *Elements of information theory*. Wiley-interscience, 2012.

[33] Michael I. Jordan, An Introduction to Variational Methods for Graphical Models [J], Machine Learning, 1999: 37,183-233

[34] Noboru Murata, S. Yoshizawa, Network information criterion-determining the number of hidden units for an artificial neural network model [J], IEEE Transactions on Neural Networks, Vol. 5, No 6, November 1994

[35] D.C Stanford, A.E Raftery. Approximate Byes factors for image segmentation: the pseudolikelihood information criterion (PLIC) [J], IEEE Transactions on Pattern Analysis and Machine Intelligence 24(11) (2002) 1517-1520.

[36] J. Sethuraman, A constructive definition of the Dirichlet priors [J], Statistica Sinica 2 (1994) 639–650

[37] Chung, Fan RK. "Spectral Graph Theory (CBMS Regional Conference Series in Mathematics, No. 92)." (1997).

[38] Center for Morphometric Analysis at Massachusetts General Hospi-tal, "The Internet Brain Segmentation Repository (IBSR),"http://www.cma.mgh.harvard.edu/ibsr/index.html, Jan. 2009

[39] USPS,http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html#usps;COLI2,http://www1.cs.columbia.edu/CAVE/software/softlib/coil-20.php;

致谢

本论文是在我的导师陈胜勇老师, 黄伟老师和李春光老师的悉心指导下完成的，陈老师严谨的治学态度，执着的学术精神以及朴实高尚的人格为我今后的研究生涯树立的榜样。黄老师，李老师一丝不苟的工作态度，科学的工作方法给予我极大的帮助。谨在此，向我的三个导师致以深深的敬意和衷心的感谢。

其次，感谢课题组刘英老师，沈鹏程，李雨柯，黄松延，吴强，罗奕梁，闫国钰等师兄以及叶艳师姐在生活和学习上的关心和帮助。感谢大家创造了一个团结和谐、互帮互助的科研氛围，使我能更有效的完成科研任务。

最后感谢浙江工业大学为我们创造了良好的学习环境，感谢浙江工业大学所有老师在这四年对我的精心培养。

附录

附录 1　概率分布

多元高斯分布

$$N(x\mid\mu,\Sigma)=\frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)^{T}\Sigma^{-1}(x-\mu)\right\} \tag{A.1}$$

Wishart 分布

$$W(\Lambda\mid w,\nu)=B(w,\nu)\mid\Lambda\mid^{(\nu-D-1)/2}\exp(-\frac{1}{2}Tr(w^{-1}\Lambda)) \tag{A.2}$$

其中 D 为各观测点的维度，$Tr(.)$ 为矩阵的迹，

$$B(w,\nu)=\mid w\mid^{-\nu/2}(2^{\nu D/2}\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma(\frac{\nu+1-i}{2}))^{-1} \tag{A.3}$$

$$E[\ln\mid\Lambda\mid]=\sum_{i=1}^{D}\varphi(\frac{\nu+1-i}{2})+D\ln 2+\ln\mid W\mid \tag{A.4}$$

$$H[\Lambda]=-\ln(W,\nu)-\frac{(\nu-D-1)}{2}E[\ln\mid\Lambda\mid]+\frac{\nu D}{2} \tag{A.5}$$

Gamma 分布

$$\varsigma(u\mid a,b)=\frac{1}{\Gamma(a)}a^{b}u^{a}e^{-bu} \tag{A.6}$$

Beta 分布

$$Beta(\mu\mid a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \tag{A.7}$$

Dirichlet 分布

$$Dir(\mu\mid\alpha)=C(\alpha)\prod_{k=1}^{K}\mu_{k}^{\alpha_{k}-1} \tag{A.8}$$

其中，

$$C(\alpha)=\frac{\Gamma(\hat{\alpha})}{\Gamma(\alpha_{1})\cdots\Gamma(\alpha_{k})},\hat{\alpha}=\sum_{k=1}^{K}\alpha_{k} \tag{A.9}$$

附录 2　变分混合模型公式

VB-GMM 变分推断相关公式

1. 计算下界需要的期望

$$E(\ln p(X \mid Z, \mu, \Lambda)) =$$

$$\frac{1}{2}\sum_{k=1}^{K} N_k \left\{ \ln \tilde{\Lambda}_k - D\beta_k^{-1} - v_k Tr(S_k W_k) - v_k (\bar{x}_k - m_k)^T W_k (\bar{x}_k - m_k) - D\ln(2\pi) \right\} \quad (B.1)$$

$$E[\ln p(Z \mid \pi)] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \ln \tilde{\pi}_k \quad (B.2)$$

$$E[\ln p(\pi)] = \ln C(\alpha_0) + (\alpha_0 - 1)\sum_{k=1}^{K} \ln \tilde{\pi}_k \quad (B.3)$$

$$E[\ln p(\mu, \Lambda)] =$$

$$= \frac{1}{2}\sum_{k=1}^{K} \left\{ D\ln(\beta_0 / 2\pi) + \ln \tilde{\Lambda}_k - \frac{D\beta_0}{\beta_k} - \beta_0 v_k (m_k - m_0)^T W_k (m_k - m_0) \right\} + K \ln B(W_0, v_0) +$$

$$+ \frac{(v_0 - D - 1)}{2}\sum_{k=1}^{K} \ln \tilde{\Lambda}_k - \frac{1}{2}\sum_{k=1}^{K} v_k Tr(W_0^{-1} W_k) \quad (B.4)$$

$$E[\ln q(Z)] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \ln r_{nk} \quad (B.5)$$

$$E[\ln q(\pi)] = \ln C(\alpha) + \sum_{k=1}^{K} (\alpha_k - 1) \ln \tilde{\pi}_k \quad (B.6)$$

$$E[\ln q(\mu, \Lambda)] = \sum_{k=1}^{K} \left\{ \frac{1}{2} \ln \tilde{\Lambda}_k + \frac{D}{2} \ln \left(\frac{\beta_k}{2\pi} \right) - \frac{D}{2} - H[q(\Lambda_k)] \right\} \quad (B.7)$$

其中 $B(W_0, v_0)$ 是 Wishart 分布的系数（A.3），$C(\alpha)$ 为 Dirichlet 分布系数(A.9), $H[q(\Lambda_k)]$为 Wishart 分布的熵。

VB-SMM 变分推断相关公式

1. 变分分布

(1) $q(s)$

$$\ln q(s) \propto \sum_{n,m}^{N,M} s_{nm} \ln r_{nm}$$

其中，

$$r_{nm} = \exp\left(\left\langle \ln \pi_m \right\rangle + \frac{1}{2} \left\langle \ln |\Lambda_m| \right\rangle + \frac{d}{2} \left\langle \ln u_{nm} \right\rangle - \frac{\left\langle u_{nm} \right\rangle \left\langle \Delta_{nm}^2 \right\rangle}{2} - \frac{d}{2} \ln 2\pi \right) \quad \text{（B.8）}$$

$$\left\langle \Delta_{nm}^2 \right\rangle = \left\langle \left\langle x_n - \mu_m \right\rangle^T \Lambda_m \left\langle x_n - \mu_m \right\rangle \right\rangle_{\mu_m, \Lambda_m}$$

$$= x_n^T \left\langle \Lambda_m \right\rangle x_n - 2 x_n^T \left\langle \Lambda_m \right\rangle \left\langle \mu_m \right\rangle + Tr[\left\langle \mu_m \mu_m^T \right\rangle \left\langle \Lambda_m \right\rangle] \quad \text{（B.9）}$$

$$q(s) = \prod_{n,m}^{N,M} p_{nm}^{s_{nm}} \quad \text{（B.10）}$$

$$p_{nm} = \frac{r_{nm}}{\sum_{m'}^{M} r_{nm'}} \quad \text{（B.11）}$$

$$\left\langle s_{nm} \right\rangle = p_{nm} \quad \text{（B.11）}$$

(2) $q(\pi)$

$$\ln q(\pi) \propto \sum_{n,m}^{N,M} (\left\langle s_{nm} \right\rangle + (\alpha_m - 1)) \ln \pi_m$$

可得到，

$$q(\pi) = Dir(\pi | \hat{\alpha}) \quad \text{(B.13)} \quad \text{（B.13）}$$

为 Dirichlet 分布，其参数定义为，

$$\hat{\alpha}_m = \sum_{n}^{N} \left\langle s_{nm} \right\rangle + \alpha_m \quad \text{（B.14）}$$

这样可得到，

$$\left\langle \ln \pi_m \right\rangle = \Phi(\hat{\alpha}_m) - \Phi(\hat{\alpha}_0) \quad \text{（B.15）}$$

其中 $\Phi(\cdot)$ 为 di-gamma 函数。

(3) $q(\Lambda_m)$

$$\left\langle \Lambda_m \right\rangle = \eta_m W_m \quad \text{（B.16）}$$

$$\left\langle \ln |\Lambda_m| \right\rangle = d \ln 2 - \ln |W_m| + \sum_{i}^{d} \Phi(\frac{\eta_m + 1 - i}{2}) \quad \text{（B.17）}$$

(4) $q(\mu_m)$

$$q(\mu_m) = N(\mu_m \mid m_m, R_m)$$

其中，

$$R_m = \langle \Lambda_m \rangle \sum_n^N \langle w_{nm} \rangle + \rho_0 I,$$

$$m_m = R_m^{-1} \left(\langle \Lambda_m \rangle \sum_n^N \langle w_{nm} \rangle x_n + \rho_0 m_0, \right)$$

并且，

$$\langle w_{nm} \rangle = \langle s_{nm} \rangle \langle u_{nm} \rangle \qquad \text{（B.18）}$$

然而可以得到，

$$\langle \mu_m \rangle = m_m \qquad \text{（B.19）}$$

$$\langle \mu_m \mu_m^T \rangle = m_m m_m^T + R_m \qquad \text{（B.20）}$$

(5) $q(u)$

$$q(u_{nm}) = \varsigma(u_{nm} \mid a_{nm}, b_{nm})$$

其中，

$$a_{nm} = \frac{v_m + \langle s_{nm} \rangle d}{2}$$

$$b_{nm} = \frac{v_m + \langle s_{nm} \rangle \langle \Delta_{nm}^2 \rangle}{2}$$

在$\langle \Delta_{nm}^2 \rangle$定义在（B.9）,需要的矩包括

$$\langle u_{nm} \rangle = \frac{a_{nm}}{b_{nm}} \qquad \text{(B.21)}$$

$$\langle \ln u_{nm} \rangle = \Phi(a_{nm}) - \ln b_{nm} \qquad \text{(B.22)}$$

2. 计算下界需要的期望

$$\langle \ln p(X \mid \mu_m, \Lambda_m, u, s) \rangle = \frac{1}{2} \sum_{n,m}^{N,M} \langle s_{nm} \rangle \left(\langle \ln |\Lambda_m| \rangle - d \ln(2\pi) + d \langle \ln u_m \rangle \langle \Delta_{nm}^2 \rangle \right) \qquad \text{(B.23)}$$

$$\left\langle \ln p(\mu_m \mid m_0, \rho_0) \right\rangle = \frac{d}{2}\ln\frac{\rho_0}{2\pi} - \frac{\rho_0}{2}\left\langle \| \mu_m - m_0 \|^2 \right\rangle \tag{B.24}$$

$$\left\langle \ln p(\Lambda_m \mid W_0, \eta_0) \right\rangle = \ln C_W(W_0, \eta_0) + \frac{\eta_0 - d - 1}{2}\left\langle \ln \mid \Lambda_m \mid \right\rangle - \frac{1}{2}Tr[W_0^{-1} \mid \Lambda_m \mid] \tag{B.25}$$

其中$C_W(\cdot)$Dirichlet 分布的系数(A.9)。

$$\left\langle \ln p(u \mid \{v_m\}) \right\rangle = \sum_m^M \left(N\left(\frac{v_m}{2}\ln\left(\frac{v_m}{2}\right) - \ln\Gamma\left(\frac{v_m}{2}\right)\right) + \sum_n^N \left(\left(\frac{v_m}{2} - 1\right)\left\langle \ln u_{nm} \right\rangle - \frac{v_m}{2}\left\langle u_{nm} \right\rangle \right) \right) \tag{B.26}$$

$$\left\langle \ln p(\pi \mid \alpha) \right\rangle = \ln\Gamma(\alpha_0) + \sum_m^M \left((\alpha_m - 1)\left\langle \ln \pi_m \right\rangle - \ln\Gamma(\alpha_m) \right) \tag{B.27}$$

$$\left\langle \ln p(s \mid \pi) \right\rangle = \sum_{n,m}^{N,M} \left\langle s_{nm} \right\rangle \left\langle \ln \pi_m \right\rangle \tag{B.28}$$

然后再考虑变分分布，

$$\left\langle \ln q(\mu_m) \right\rangle = \frac{1}{2}\ln \mid R_m \mid - \frac{d}{2}(1 + \ln(2\pi)) \tag{B.29}$$

$$\left\langle \ln q(\Lambda_m) \right\rangle = \ln C_W(W_m, \eta_m) + \frac{\eta_m - d - 1}{2}\left\langle \ln \mid \Lambda_m \mid \right\rangle - \frac{\eta_m d}{2} \tag{B.30}$$

$$\left\langle \ln q(u) \right\rangle = \sum_{n,m}^{N,M} ((a_{nm} - 1)\Phi(a_{nm}) - a_{nm} - \ln\Gamma(a_{nm}) + \ln b_{nm}) \tag{B.31}$$

$$\left\langle \ln q(\pi) \right\rangle = \ln\Gamma(\hat{\alpha}_0) + \sum_m^M \left((\hat{\alpha}_m - 1)\left\langle \ln \pi_m \right\rangle - \ln\Gamma(\hat{\alpha}_m) \right) \tag{B.32}$$

$$\left\langle \ln p(s) \right\rangle = \sum_{n,m}^{N,M} \left\langle s_{nm} \right\rangle \ln \left\langle s_{nm} \right\rangle \tag{B.33}$$

VB-iSMM 变分推断相关公式

1. 变分分布

（1） $q(z_{nj}) = \tilde{\gamma}_{nj} / \sum_{j=1}^{T} \tilde{\gamma}_{nj}$ （当 $j > T$ 时， $\pi_j(V) = 0$），其中，

$$\tilde{\gamma}_{nj} = \exp\left(\sum_{i=1}^{j-1} \left\langle \log(1 - V_i) \right\rangle + \left\langle \log V_j \right\rangle + \frac{1}{2}\left\langle \ln \mid \Lambda_j \mid \right\rangle + \frac{d}{2}\left\langle \ln u_{nj} \right\rangle - \frac{\left\langle u_{nj} \right\rangle \left\langle (x_n - \mu_j)^T \Lambda_j (x_n - \mu_j) \right\rangle}{2} - \frac{d}{2}\ln 2\pi \right) \tag{B.34}$$

（2） $q(u_{nj})=Gam(u_{nj}\mid\hat{v}_{nj1},\hat{v}_{nj2})$ 其中，

$$\hat{v}_{nj1}=\frac{1}{2}[q(z_{nj})\cdot d+v_j] \tag{B.35}$$

$$\hat{v}_{nj2}=\frac{1}{2}[q(z_{nj})\left\langle(x_n-\mu_j)^T\Lambda_j(x_n-\mu_j)\right\rangle+v_j] \tag{B.36}$$

（3） $q(\mu_j,\Lambda_j)=N(\mu_j\mid\tilde{m}_j,\tilde{\lambda}_j\Lambda_j)W(\Lambda_j\mid\tilde{W}_j,\tilde{\rho}_j)$

定义统计量

$$N_j=\sum_{n=1}^{N}q(z_{nj})\left\langle u_{nj}\right\rangle$$

$$\overline{x}_j=\frac{1}{N_j}\sum_{n=1}^{N}q(z_{nj})\left\langle u_{nj}\right\rangle x_n$$

$$S_j=\frac{1}{N_j}\sum_{n=1}^{N}q(z_{nj})\left\langle u_{nj}\right\rangle(x_n-\overline{x}_j)(x_n-\overline{x}_j)^T$$

然后计算超参数为，

$$\tilde{\lambda}_j=\lambda_j+N_j \tag{B.37}$$

$$\tilde{m}_j=\frac{1}{\tilde{\lambda}_j}(\lambda_j m_j+N_j\overline{x}_j) \tag{B.38}$$

$$\tilde{W}_j^{-1}=W_j^{-1}+N_jS_j+\frac{\lambda_jN_j}{\lambda_j+N_j}(\overline{x}_j-m_j)(\overline{x}_j-m_j)^T \tag{B.39}$$

$$\tilde{\rho}_j=\rho_j+\sum_{n=1}^{N}q(z_{nj}) \tag{B.40}$$

（4） $q(V_j)=Beta(\tilde{\beta}_{j1},\tilde{\beta}_{j2})$ ，其超参数更新式为

$$\tilde{\beta}_{j1}=1+\sum_{n=1}^{N}q(z_{nj}) \tag{B.41}$$

$$\tilde{\beta}_{j2}=\left\langle\alpha\right\rangle+\sum_{n=1}^{N}\sum_{i=j+1}^{T}q(z_{ni}) \tag{B.42}$$

(5) $q(\alpha)=Gam(\tilde{\eta}_1,\tilde{\eta}_2)$ 其超参数更新式为，

$$\tilde{\eta}_1=\eta_1+T-1 \tag{B.43}$$

$$\tilde{\eta}_2 = \eta_2 - \sum_{j=1}^{T-1}\left\langle \log(1-V_j)\right\rangle \tag{B.44}$$

2. 超参数更新式相关期望$\langle\cdot\rangle$

$$\left\langle u_{nj}\right\rangle = \tilde{v}_{nj1} / \tilde{v}_{nj2} \tag{B.45}$$

$$\left\langle \ln u_{nj}\right\rangle = \Phi(\tilde{v}_{nj1}) - \ln \tilde{v}_{nj2} \tag{B.46}$$

$$\left\langle \ln |\Lambda_j|\right\rangle = d\ln 2 + \ln|\tilde{W}_j| + \sum_{i}^{d}\Phi(\frac{\tilde{\rho}_j + 1 - i}{2}) \tag{B.47}$$

$$\left\langle (x_n-\mu_j)^T\Lambda_j(x_n-\mu_j)\right\rangle = \frac{D}{\tilde{\lambda}_j} + \tilde{\rho}_j(x_n-\tilde{m}_j)^T\tilde{W}_j(x_n-\tilde{m}_j) \tag{B.48}$$

$$\left\langle \ln V_j\right\rangle = \Phi(\tilde{\beta}_{j1}) - \Phi(\tilde{\beta}_{j1}+\tilde{\beta}_{j2}) \tag{B.49}$$

$$\left\langle \ln(1-V_j)\right\rangle = \Phi(\tilde{\beta}_{j2}) - \Phi(\tilde{\beta}_{j1}+\tilde{\beta}_{j2}) \tag{B.50}$$

$$\left\langle \alpha\right\rangle = \tilde{\eta}_1 / \tilde{\eta}_2 \tag{B.51}$$

$$\left\langle \ln\alpha\right\rangle = \Phi(\tilde{\eta}_1) - \ln\tilde{\eta}_2 \tag{B.52}$$

3. 计算下界相关期望$E[\cdot]$

$$E[\ln p(X|Z,\mu,\Lambda,u)] = \frac{1}{2}\sum_{j,n}^{T,N}\left\langle z_{nj}\right\rangle\left(\left\langle \ln|\Lambda_j|\right\rangle - d\ln(2\pi)\right.$$
$$\left. + d\left\langle \ln u_{nj}\right\rangle - \left\langle u_{nj}\right\rangle\left\langle (x_n-\mu_j)^T\Lambda_j(x_n-\mu_j)\right\rangle\right) \tag{B.53}$$

$$E[\ln p(u)] = \sum_{j=1}^{T}\left\{ N\left[\frac{v_j}{2}\ln\left(\frac{v_j}{2}\right) - \ln\Gamma\left(\frac{v_j}{2}\right)\right]\right.$$
$$\left. + \sum_{n}^{N}\left(\left(\frac{v_j}{2}-1\right)\left\langle \ln u_{nj}\right\rangle - \frac{v_j}{2}\left\langle u_{nj}\right\rangle\right)\right) \tag{B.54}$$

$$E[\ln p(\mu,\Lambda)] = \frac{1}{2}\sum_{j=1}^{T}\left\{ d\ln(\lambda_j/2\pi) - \lambda_j\left\langle (x_n-\mu_j)^T\Lambda_j(x_n-\mu_j)\right\rangle\right.$$
$$\left. + \frac{\rho_j - d + 1}{2}\left\langle \ln|\Lambda_j|\right\rangle - \frac{1}{2}Tr(W_j^{-1}\tilde{W}_j)\tilde{\rho}_j\right. \tag{B.55}$$

$$E[\ln p(Z|V)] = \sum_{j=1}^{T}\sum_{n=1}^{N} q(z_{nj})\left\{\sum_{i=1}^{j-1}\left\langle \ln(1-V_i)\right\rangle + \left\langle \ln V_j\right\rangle\right\} \tag{B.56}$$

$$E[\ln p(V\mid\alpha)]=(\langle\alpha\rangle-1)\sum_{j=1}^{T}\left\langle\ln(1-V_j)\right\rangle \tag{B.57}$$

$$E[\ln p(\alpha)]=(\eta_1-1)\langle\ln\alpha\rangle-\eta_2\langle\alpha\rangle \tag{B.58}$$

$$E[\ln q(u)]=\sum_{j,n}^{T,N}((\tilde{v}_{nj1}-1)\Phi(\tilde{v}_{nj1})-\tilde{v}_{nj1}-\ln\Gamma(\tilde{v}_{nj1})+\ln\tilde{v}_{nj2}) \tag{B.59}$$

$$\begin{aligned}E[\ln p(\mu,\Lambda)]=\frac{1}{2}\sum_{j=1}^{T}\Big\{&d\ln(\rho_j/2\pi)-2\ln B(\tilde{W}_j,\tilde{\rho}_j)+d\\&+(\tilde{\rho}_j-d)\langle\ln|\Lambda_j|\rangle+\tilde{\rho}_jD\Big\}\end{aligned} \tag{B.60}$$

$$E[\ln q(Z\mid V)]=\sum_{j=1}^{T}\sum_{n=1}^{N}q(z_{nj})\ln q(z_{nj}) \tag{B.61}$$

$$E[\ln q(V\mid\alpha)]=\sum_{j=1}^{T}\left\{(\tilde{\beta}_{j1}-1)\left\langle\ln V_j\right\rangle+(\tilde{\beta}_{j2}-1)\left\langle\ln(1-V_j)\right\rangle\right\} \tag{B.62}$$

$$E[\ln q(\alpha)]=(\tilde{\eta}_1-1)\langle\ln\alpha\rangle-\tilde{\eta}_2\langle\alpha\rangle \tag{B.63}$$

其中 $B(\tilde{W}_j,\tilde{\rho}_j)$ 为 Wishart 分布的归一化参数（A.3）