Variational Inference

@一又七分之四

2012/11/30

【关键字】平均场理论，变分法，贝叶斯推断，EM 算法，KL 散度，变分估计，变分消息传递

引言

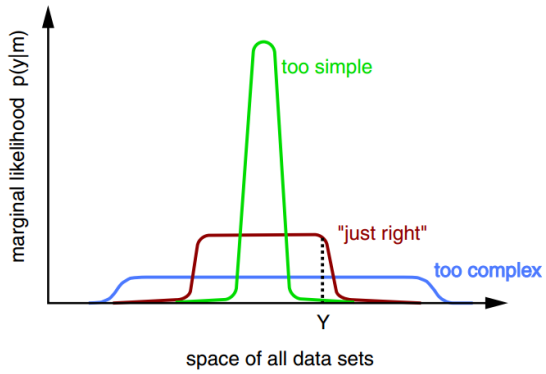
- ### 从贝叶斯推断说起

Question：如果我们有一组观测数据 D，如何推断产生这些数据的模型 m?

（不严格）考虑参数化模型由 1）模型的类别 ξ（如高斯分布，伽马分布，多项式分布等）与 2）模型的参数 Θ 共同决定，即 $m(\xi,\Theta)$.

模型的选择

- 假设 M 为所有可能的模型集合（包括不同类别），那么选择 $m=\arg\max\{p(m(\xi,\Theta)\,|\,D), m\in M\}$
- 如何计算 p(m | D)?
  - 通常情况很难直接计算 p(m | D)，根据贝叶斯公式有 $p(m\,|\,D)=\dfrac{p(m)\,p(D\,|\,m)}{p(D)}$ ，p(m)表示模型的先验，p(D | m)表示证据；
  - 先验：贝叶斯规则倾向于选择能解释数据的最简单模型：Occam 剃刀原理。因为简单模型只在有限范围内做预测，复杂模型（如有更多自由参数）能对更宽范围做预测。



  - 那么如何计算证据（evidence） $p(D\,|\,m)=\int_{\Theta} p(\theta\,|\,m)\,p(D\,|\,\theta,m)\,d\theta$ ？
    - 参数 θ 的后验概率为 $p(\theta\,|\,D,m)=\dfrac{p(\theta\,|\,m)\,p(D\,|\,\theta,m)}{p(D\,|\,m)}$
    - 证据 p(D | m)通常会在最可能的参数 $\hat{\theta}$ 附近有一个很强的峰。
    - 以一维参数为例：利用 Laplace 方法近似，即用被积函数 $p(\hat{\theta}\,|\,m)\,p(D\,|\,\hat{\theta},m)$ 乘以其宽度 $\sigma_{\theta|D}$ 。即 $p(D\,|\,m)\approx p(\hat{\theta}\,|\,m)\,p(D\,|\,\hat{\theta},m)\sigma_{\theta|D}$ 。
    - 此处不在深究 Occam 因子。
- 从模型的选择可以看出参数的估计非常重要。

考虑同一个类别的模型。由于任何模型（函数）都可以由统一的数学形式给出，比如拉格朗日展开，傅里叶极数，高斯混合模型（GMM）等，因而通常我们更关心一个模型的参数 Θ。换句话说，给出一组观测数据 D，我们总是能够通过估计参数来推测模型，即 $m = \arg\max\{p(m(\Theta)\,|\,D), m \in M_\xi\}$。或者更简单的形式 $m = \arg\max\{p(\Theta\,|\,D), m(\Theta) \in M_\xi\}$。

后验概率的估计

通常情况，取后验概率最大的参数值为估计值。根据贝叶斯公式，参数 θ 后验概率为

$$p(\theta\,|\,D) = \frac{p(\theta)\,p(D\,|\,\theta)}{p(D)} = \frac{p(\theta)\,p(D\,|\,\theta)}{\int_\Theta p(\theta)\,p(D\,|\,\theta)d\theta} \propto p(\theta)\,p(D\,|\,\theta)\,,$$

其中 p(D)为归一化常数（normalizing constant）。

- 从经典的统计学角度看，概率是相对频率的，是真实世界的客观属性。因而每个模型被选择的概率是一样的，因而 p(θ) =constant。此时问题转化为：$\theta_{ML} = \arg\max\{p(D\,|\,\theta)\}$，这便是极大似然法（ML, Maximum Likelihood）。
- 从贝叶斯学派的角度看，每一个模型都有一个先验概率 p(θ)，但先验概率需事先给定。此时问题转化为：$\theta_{MAP} = \arg\max\{p(\theta)\,p(D\,|\,\theta)\}$，这便是极大后验估计（MAP, Maximum A Posteriori）。

另一方面，许多科学问题的基本部分是计算一个目标函数的积分 $I = \int_\Omega f(x)dx$。Ω 通常是高维空间中的一个区域，一般情况下 f(x)稍微复杂一些，积分就难以计算。如果 f(x)能被分解成一个函数 g(x)与一个概率密度函数 π(x)的乘积，那么上述积分可看做是 g(x)在密度 π(x)下的期望。

$$I = \int_\Omega f(x)dx = \int_\Omega g(x)\pi(x)dx = E_\pi[g(x)]$$

比如，

（1） 计算后验概率：$p(\theta_1\,|\,D) \propto \int_{\Theta_2^*} p(\theta_1, \theta_2\,|\,D)d\theta_2$

（2） 点估计：$\bar{\theta} = \int \theta\, p(\theta\,|\,D)d\theta$

（3） 训练样本预测将来的数据的概率密度：假设 D’与 D 条件独立，$p(D'\,|\,D) = \int p(\theta\,|\,D)p(D'\,|\,\theta)d\theta\,.$

（4） 新观测样本 D’的隐藏变量(hidden variable) x’的后验分布：$p(x'\,|\,D', D) \propto \int p(\theta\,|\,D)\,p(x', D'\,|\,\theta)d\theta$

上述积分最简单的近似方法就是通过估计参数 θ 来估计单点积分值，比如上述贝叶斯选择模型中极大后验估计（MAP）。由于 ML、MAP 只是估计概率密度而不是概率分布，因而省去了积分过程。ML, MAP 估计最常用也最基本的方法是期望最大化算法（Expectation Maximization，EM）。

此外，可以通过蒙特卡洛方法（Monte Carlo），或马氏链蒙特卡洛法（Markov Chain Monte Carlo, MCMC）来模拟积分。此类方法具有较高的精度，但需要大量的计算。

本文介绍一种变分方法来近似积分。其主要思想是，对一个特定模型，构造一种简单（tractable）的数学形式来近似未观测变量的后验分布，同时给出观测数据的边缘似然（或者称证据，evidence）的下界（lower bound）。而积分过程转化为求下界的最优值问题。

变分贝叶斯

前言

变分贝叶斯方法在 Matthew J.Beal 的博士论文《Variational Algorithms for Approximate Bayesian Inference》有了比较充分的讨论，作者将其应用于隐马尔科夫模型，混合因子分析，非线性动力学，图模型等。变分贝叶斯是一类用于贝叶斯估计和机器学习领域中近似计算复杂（intractable）积分的技术。它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(observed variables, data)，未知参数（parameters）和潜变量（latent variables）。在贝叶斯推断中，参数和潜变量统称为不可观测变量(unobserved variables)。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 Data 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，变分贝叶斯可以看做是 EM 算法的扩展，因为它也是采用极大后验估计(MAP)，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（mutually dependent）的等式进行不断的迭代来获得最优解。

问题描述

现在重新考虑一个问题：1）有一组观测数据 D，并且已知模型的形式，求参数与潜变量（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(tractable)的数学形式 Q(Z)来近似 P(Z|D),即 $P(Z\mid D)\approx Q(Z)$ 。

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（dissimilarity）？
（2） 如何得到简单的 Q(Z)?

对于问题一，幸运的是，我们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距离的度量，即相对熵，或者称为 Kullback-Leibler 散度。

对于问题二，显然我们可以自主决定 Q(Z)的分布，只要它足够简单，且与 P(Z|D)接近。然而不可能每次都手工给出一个与 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？在“模型的选择”部分，曾提到 Occam's razor，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。

幸运的是，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，并发展了平均场理论。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。于是我们可以作出后验条件独立（posterior independence）的假设。即， $\forall i, p(Z\mid D)=p(Z_i\mid D)p(Z_{-i}\mid D)$ 。

- **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 $D_{KL}(p\|q)=\sum_{x\in\aleph}p(x)\log\frac{p(x)}{q(x)}$

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；

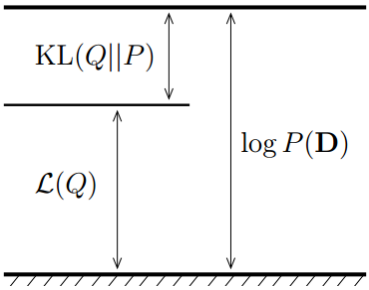
（3） 不满足三角不等式。

Q 分布与 P 分布的 KL 散度为： $D_{KL}(Q\|P)=\sum_Z Q(Z)\log\frac{Q(Z)}{P(Z\mid D)}=\sum_Z Q(Z)\log\frac{Q(Z)}{P(Z,D)}+\log P(D)$

或者 $\log P(D)=D_{KL}(Q\|P)-\sum_Z Q(Z)\log\frac{Q(Z)}{P(Z,D)}=D_{KL}(Q\|P)+L(Q)$

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）：

$$L(Q)=\sum_Z Q(Z)\log P(Z,D)-\sum_Z Q(Z)\log Q(Z)=E_Q[\log P(Z,D)]+H(Q)$$



- **平均场理论（Mean Field Method）**

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。

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（partition）因式分解，比如将 Z 划分为 $Z_1\ldots Z_M$,

$$Q(Z)=\prod_{i=1}^{M}q(Z_i\mid D)$$

注意这里并非一个不可观测变量一个划分，而应该根据实际情况做决定。当然你也可以这么做，但是有时候，将几个潜变量放在一起会更容易处理。

- **平均场方法的合理性**

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他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。

同样在变分分布 Q(Z)这个系统中，我们也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做是其自身的作用。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 与它的互斥集 Z_{-I}(或者更进一步，马尔科夫毯(Markov blanket), mb(Z_i))具有一个简单的关系：

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

（为保持文章的连贯性，此处先不证明，下文将详细说明）

于是，对于某个划分 Z_i ,我们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。相同步骤应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。

具体更新边缘概率（VB-marginal）步骤如下：

（1） 初始化 $Q^{(1)}(Z_i)$，可随机取；

（2） 在第 k 步，计算 Z_{-i} 的边缘密度 $Q^{[k]}(Z_{-i} \mid D) \propto \exp \int_{Z_i^*} Q^{[k-1]}(Z_i \mid D) \log P(Z_i, Z_{-i}, D) dZ_i$

（3） 计算 Z_i 的边缘密度 $Q^{[k]}(Z_i \mid D) \propto \exp \int_{Z_{-i}^*} Q^{[k]}(Z_{-i} \mid D) \log P(Z_i, Z_{-i}, D) dZ_{-i}$

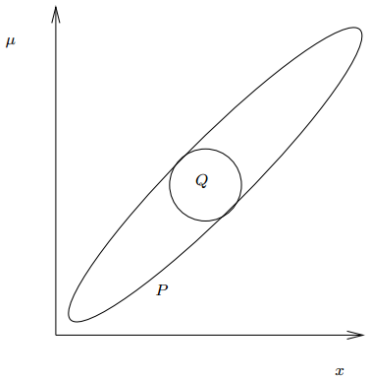
（4） 理论上 $Q^{[\infty]}(Z_i \mid D)$ 将会收敛，则反复执行(2), (3)直到 $Q(Z_i)$, $Q(Z_{-i})$稳定，或稳定在某个小范围内。

（5） 最后，得 $Q(Z) = Q(Z_i \mid D) Q(Z_{-i} \mid D)$

- **平均场估计下边缘概率的无意义性（VB-marginals）**

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。不应该用 $Q_i(Z_i)$来估计真实的边缘密度，比如在一个贝叶斯网络中，你不应该用它来推测某个节点的状态。而这其实是很糟糕的，相比于其他能够使用节点状态信息来进行局部推测的算法，变分贝叶斯方法更不利于调试。

比如一个标准的高斯联合分布 P(μ,x)和最优的平均场高斯估计 Q(μ,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，完全与 $P_x(x)$不同。



- ## 边缘密度（VB-marginal）公式的证明

上文已经提到我们要找到一个更加简单的函数 D(Z)来近似 P(Z|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。我们先介绍一下什么是泛函，以及泛函取得极值的必要条件。

- **泛函的概念**

【泛函】设对于(某一函数集合内的)任意一个函数 y(x)，有另一个数 J[y]与之对应，则称 J [y]为 y(x)的泛函。泛函可以看成是函数概念的推广。这里的函数集合，即泛函的定义域，通常要求 y(x) 满足一定的边界条件，并且具有连续的二阶导数．这样的 y(x)称为可取函数。

泛函不同于复合函数，例如 g=g(f(x)); 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(x)，才能得到一个泛函值 J[y]。(定义在同一区间上的)函数不同，泛函值当然不同，为了强调泛函值 J[y]与函数 y(x)之间的依赖关系，常常又把函数 y(x)称为变量函数。

泛函的形式多种多样，通常可以积分形式： $J[y]=\int_{x_0}^{x_1}F(x,y,y')dx$

- **泛函取极值的必要条件**

泛函的极值

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

所谓函数 y(x) + δy(x)在另一个函数 y(x)的“附近”，指的是：

1. | δy(x) | < ε;
2. 有时还要求| (δy)’(x) | < ε.

这里的 δy(x)称为函数 y(x)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[y]取到极大值的必要条件是一级变分 δJ[y]为 0，其微分形式一般为二阶常微分方程，即 Euler-Largange 方程：

$$\frac{\partial F}{\partial y}-\frac{d}{dx}\frac{\partial F}{\partial y'}=0$$

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[y]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，

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

- **问题求解**

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(Energy)，第二项看做是信息熵(Shannon entropy)。我们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，

$$L(Q(Z))=\int(\prod_i Q_i(Z_i))\ln(Z,D)dZ-\int(\prod_k Q_k(Z_k))\sum_i \ln Q_i(Z_i)dZ$$

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)（第一项），

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

其中定义 $Q_i^*(Z_i)=\frac{1}{C}\exp\left\langle\ln(Z,D)\right\rangle_{Q_{-i}(Z_{-i})}$，C 为 $Q_i^*(Z_i)$ 的归一化常数。再考虑熵量(entropy)（第二项），

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

此时得到泛函，

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

注意到 L(Q(Z))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。为了使得 L(Q(Z))达到最大值，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，

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

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

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

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

例子:高斯混合模型（GMM）

Question2 假设现在有独立同分布（i.d.d.）的训练样本 X 符合下列混合高斯分布

$$p(x) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \sum_k)$$

如何求解高斯混合分布的三组参数 $\pi_k, \mu_k, \sum_k$？

- **步骤一：选择无信息先验分布（non-informative prior）**

先验分布的不是随便取的，一般可以根据共轭分布方法，Jefferys 原则，最大熵原则等来确定。一般要求先验分布应取共轭分布（conjugate distribution）才合适，即先验分布 h(θ)与后验分布 h(θ|x)属于同一分布类型。本文不展开讨论，直接给出

$$\pi_{i=1,\ldots,k} \sim SymDir(K, \alpha_0)$$
$$\Lambda_{i=1,\ldots,k} \sim \mathrm{W}(w_0, \upsilon_0)$$
$$\mu_{i=1,\ldots,k} \sim N(m_0, (\beta_0 \Lambda_i)^{-1})$$
$$z_{i=1,\ldots,N} \sim Mult(1, \pi)$$
$$X_{i=1,\ldots,N} \sim N(\mu_z)$$

k 表示单高斯分布的个数，N 表示样本个数，每个分布的解释，

- SymDir() 表示 K 维对称 Dirichlet 分布；它是卡方分布（categorical）或多项式分布（multinomial）的共轭先验分布。
- W() 表示 Wishart 分布；对一个多元高斯分布（multivariate Gaussian distribution），它是 Precision 矩阵（逆协方差矩阵）的共轭先验。
- Mult() 表示多项分布（此处也称卡方分布）；多项式分布是二项式分布的推广，表示在一个 K 维向量中只有一项为 1，其它都为 0.
- N() 为高斯分布，在这里特别指多元高斯分布。

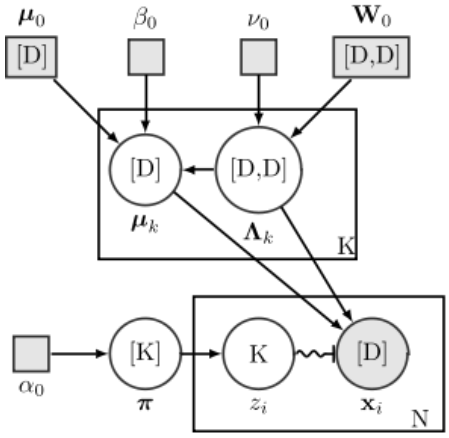
对变量的解释

- $X = \{x_1, \ldots, x_N\}$ 是 N 个训练样本，每一项都是服从多元高斯分布的 K 维向量。
- $Z = \{z_1, \ldots, z_N\}$ 是一组潜在变量，每一项 $z_k = \{z_{1k,\ldots}, z_{nk}\}$ 用于表示对应的样本 x_k 属于哪个混合部分（mixture component）。
- $\pi = \{\pi_1, \ldots, \pi_K\}$ 表示每个单高斯分布混合比例（mix proportion）
- $\mu_{i=1,\ldots,k}$ 和 $\Lambda_{i=1,\ldots,k}$ 分别表示每个单高斯分布参数的均值（mean）和精度（precision）

另外，为了区分联合分布的参数，以上分布的参数如 $K,\alpha_0,\beta_0,w_0,\upsilon_0,m_0$称为超参数（*hyperparameter*），并且都是已知量。又称为超参数（hyperparameter），并且都是已知量。

- **步骤二：写出联合概率密度函数**

用“盘子表示法”（plate notation）表示贝叶斯多元高斯混合模型，如图所示。



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 categorical 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)。

假设各参数与潜在变量条件独立，则联合概率密度函数可以表示为

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

每个因子为：

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

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

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

$$p(\mu\mid\Lambda)=N(\mu_k\mid m_0,(\beta_0\Lambda_k)^{-1})$$

$$p(\Lambda)=W(\Lambda_k\mid w_0,\nu_0)$$

其中，

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

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

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

D 为各观测点的维度。

- **步骤三：计算边缘密度(VB- marginal)**

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

两边分别取对数可得，

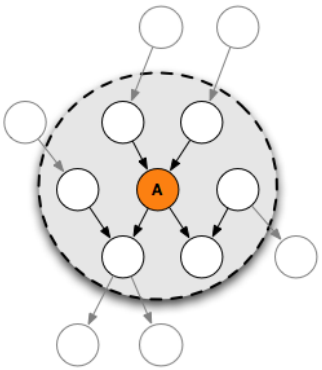
$$q^*(Z)\propto\prod_{n=1}^{N}\prod_{k=1}^{K}\rho_{nk}^{z_{nk}}$$

归一化，即对于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

$q^*(Z)=\prod_{n=1}^{N}\prod_{k=1}^{K}r_{nk}^{z_{nk}}$ ，其中 $r_{nk}=\frac{\rho_{nk}}{\sum_{j=1}^{K}\rho_{nj}}$

可见 $q^*(Z)$ 是多个单观测多项式分布(single-observation multinomial distribution)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 categorical 分布，有 $E[z_{nk}]=r_{nk}$.

另外，需特别注意的是，在求期望的过程中，由于联合密度可以写成几个因子乘积的形式，而方程是关于 Z 的函数，因此对于不包含 Z 的密度函数可以当做常数处理。我们可以用马尔科夫毯（Markov blanket）描述，在一个贝叶斯网络中，它表示一个节点的父节点(parents)，子节点和子节点的父节点（co-parents）,如图所示。我们将在后文中详细说明。



（2） 计算 π 的概率密度， $q(\pi,\mu,\Lambda)=q(\pi)\prod_{k=1}^{K}q(\mu_k,\Lambda_k)$

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

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 $\alpha = \alpha_0 + N_k$ ， $N_k = \sum_{n=1}^{N} r_{nk}$.

（3） 最后同时考虑 μ, Λ，对于每一个单高斯分布有，

$$\ln q^*(\mu_k, \Lambda_k) = E_{Z,\pi,\mu_{i\neq k},\Lambda_{i\neq k}}[\ln p(X \mid Z, \mu_k, \Lambda_k) p(\mu_k, \Lambda_k)]$$

$$= \ln p(\mu_k, \Lambda_k) + \sum_{n=1}^{N} E[z_{nk}] \ln N(x_n \mid \mu_k, \Lambda_k^{-1}) + const$$

经过一系列重组化解将得到 Gaussian-Wishart 分布，

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

其中定义，

$$\begin{cases} \beta_k = \beta_0 + N_k, \\ m_k = \dfrac{1}{\beta_k}(\beta_0 m_0 + N_k \bar{x}_k), \\ w_k^{-1} = w_0^{-1} + N_k S_k + \dfrac{\beta_0 N_k}{\beta_0 + N_k}(\bar{x}_k - m_0)(\bar{x}_k - m_0)^T, \\ \nu_k = \nu_0 + N_k, \\ \bar{x}_k = \dfrac{1}{N_k}\sum_{n=1}^{N} r_{nk} x_n, \\ S_k = \dfrac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\bar{x}_k - x_k)(\bar{x}_k - x_k)^T. \end{cases}$$

- **步骤四：迭代收敛**

最后，注意到对 π,μ,Λ 的边缘概率都需要且只需要 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：

$$\begin{cases} \ln \tilde{\pi}_k \equiv E[\ln |\pi_k|] = \psi(\alpha_k) - \psi\left(\sum_{i=1}^{K} \alpha_i\right) \\ \ln \tilde{\Lambda}_k \equiv E[\ln |\Lambda_k|] = \sum_{i=1}^{D} \psi\left(\dfrac{\nu_k + 1 - i}{2}\right) + D \ln 2 + \ln |\Lambda_k| \\ E_{\mu_k,\Lambda_k}[(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)] = D\beta_k^{-1} + \nu_k (x_n - m_k)^T W_k (x_n - m_k) \end{cases}$$

这些结果能够导出： $r_{nk} \propto \tilde{\pi}_k \tilde{\Lambda}_k^{1/2} \exp\left\{-\dfrac{D}{2\beta_k} - \dfrac{\nu_k}{2}(x_n - m_k)^T W_k (x_n - m_k)\right\}$ 。由于需要归一化使得 $\sum_{k=1}^{K} r_{nk} = 1$，这样可以从线性相对值转化为绝对值。

再次分析各参数，

1. 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} .
2. 参数变量 π 更新方程中的超参数 $\alpha_{1...K}$ 都依赖于统计量 N_k ， 即 r_{nk} .
3. 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 $w_k, \nu_k, \alpha_{1...K}$ 通过 $\tilde{\pi}_k, \tilde{\Lambda}_k$ 有间接的依赖关系。

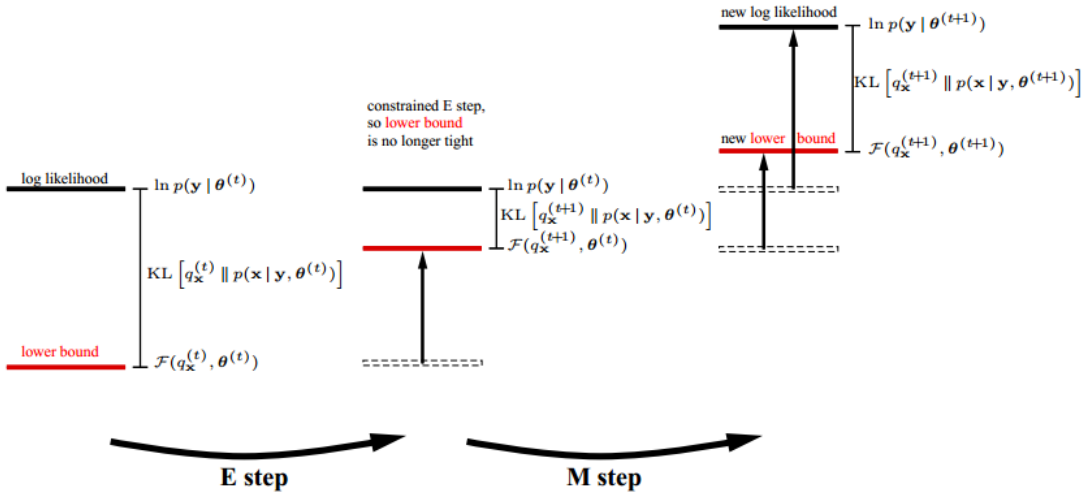
这样迭代过程便很清楚了，可以总结为如下**两个迭代步骤：**

1. 在 VBE-Step，用参数和超参数的旧值计算潜在变量 r_{nk} ；
2. 在 VBM-Step，用潜在变量计算参数和超参数的新值。

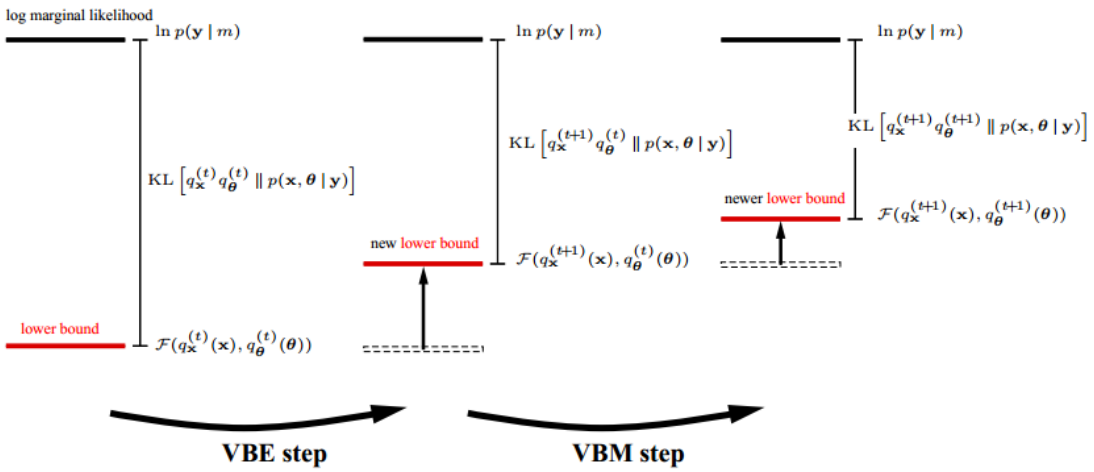
- **与 EM 算法的比较**

以上迭代步骤与 EM 算法用 ML 或 MAP 解决高斯混合模型很相似。在 E-step 中，潜在变量 r_{nk} 对应于潜在变量关于数据样本的后验概率，如 P(Z|X)；$N_k, \bar{x}_k, S_k$ 统计量对应于 EM 算法中“soft-count”统计量；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。

但经过如此，VBEM 算法与 EM 算法还有有很多不同之处的。比如迭代中，逼近最优值的过程是不一样的，如下图所示。在有限制的情况下，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 KL.在 E 步骤，下界逼近最大似然值（或者由于条件限制相差一点）；然后在 E 步骤中，根据新参数重新确定新的似然值。如此往复，直到达到稳定。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。



EM 算法（with constrained）



VBEM 算法

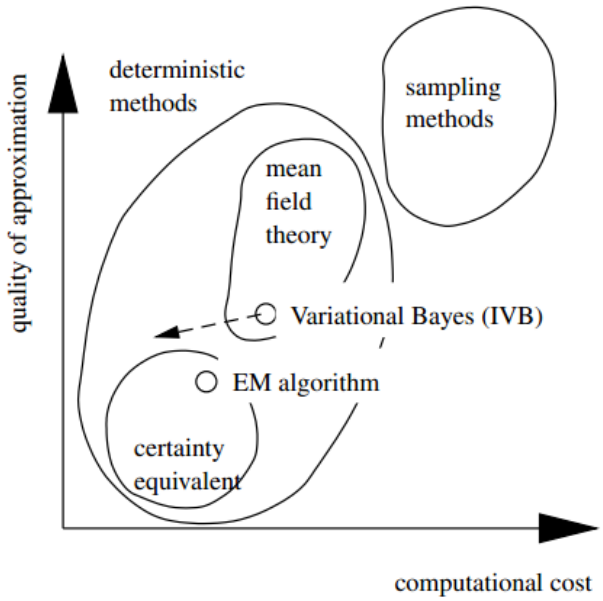
进一步讨论

- ### 再谈 EM 与 VBEM

EM 算法计算随机变量（或归类于参数）后验分布的点估计，但估计潜在变量的真实后验分布（至少在 soft EM 算法，并且经常只当潜在变量是离散化的情况）。这些参数的众数（modes）作为点估计，无任何其他信息。

而在 VB 算法，计算所有变量的真实后验分布的估计，包括参数和潜在变量。计算点估计的过程中，一般使用在贝叶斯推断中常用的均值（mean）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潜在变量的贝叶斯网络的参数最佳值。正如之前的高斯混合模型例子，对于每一个混合部分都需要计算其参数。EM 算法将会直接估计这些参数的最优值；而 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。

- ### 算法复杂度



- ### 信息耦合(Intractable Coupling)

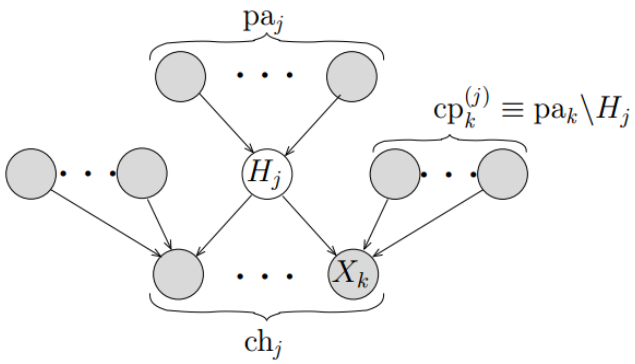
改进：变分消息传递(Variational Message Passing, VMP)

从高斯混合模型的例子可以看出，传统的变分贝叶斯方法对模型的推导是繁琐而复杂的。而在推导边缘概率的时候，我们也提到对数联合概率的在一些参数下的期望可以简化，我们只需要关心所求参数的马尔科夫毯上的节点。另外，又认识到许多先验和条件概率属于指数分布族，而对数可以将乘积形式展开为加法。那么，我们是不是可以找到一些简单计算方法或者统一的形式呢？

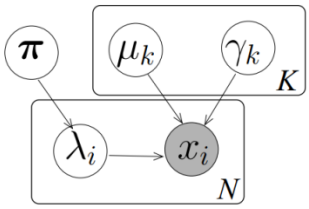
这便是变分消息传递（Variational Message Passing; John Winn, Bishop 2005）他们考虑了贝叶斯网络中的共轭指数网络（conjugate-exponential networks）,这种方法使得充分统计量（sufficient statistics）与自然参数（natural parameter）都有一个标准形式（standard form），现在该方法已经取代了手工推导(derivation by hand),成为标准的变分贝叶斯推断方法。而对于非共轭指数网络（比如混合模型），也能通过进一步的近似转化为标准形式。

- 贝叶斯网络

变分信息传递方法是建立在贝叶斯网络上的，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。



具体的，对于混合高斯模型，我们有如下图模型，



- 指数分布族

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

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

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，这里的 $0 < c(\theta), c_1(\theta), ..., c_k(\theta) < \infty, T_j(x)$ 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，

$$P(X \mid Y) = \exp[\phi(Y)^T u(X) + f(X) + g(Y)]$$

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（be tractable to compute）并且它的状态可以用自然参数向量所概括。

共轭指数模型（Conjugate-Exponential Model）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，那么我们称这样的模型是*共轭指数模型*。

我们考虑共轭指数模型，这样后验的每个因子与它的先验都有相同的形式，因而我们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，后面我们将详细说明。

- ## 自然统计量的期望

如果我们知道自然参数向量 $\phi(Y)$ ，那么我们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(reparameterisation)为 $\tilde{g}$ 则,

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

对 X 积分有，

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

然后对 ϕ 微分，

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

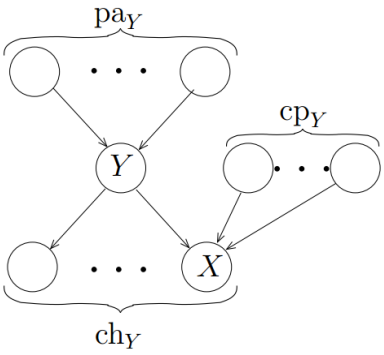
$$\int_X P(X\mid\phi)\left[u(X)+\frac{d\tilde{g}(\phi)}{d\phi}\right]dX=0$$

得自然统计量的期望，

$$\langle u(X)\rangle_{P(X|\phi)}=-\frac{d\tilde{g}(\phi)}{d\phi}$$ （1） 如何计算？

- ## 变分分布 Q 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，如下图所示。



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，

$$\ln P(Y\mid pa_Y)=\phi_Y(pa_Y)^T u_Y(Y)+f_Y(Y)+g_Y(pa_Y).$$ （2）

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，

$$\ln P(X \mid Y, cp_Y) = \phi_{XY}(Y, cp_Y)^T u_Y(X) + \lambda(Y, cp_Y). \quad (3)$$

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

$$\ln P(X \mid Y, cp_Y) = \phi_{XY}(X, cp_Y)^T u_Y(Y) + \lambda(X, cp_Y). \quad (4)$$

为了更新 Q(Y),需要找到(2),(3)式关于除 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（1）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(3)，(4)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(multi-linear function)。

重新考虑 Y 的变分更新方程，

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

其中 $\phi_Y^* = \left\langle \phi_Y(pa_Y)^T \right\rangle_{\sim Q(Y)} + \sum_{k \in ch_Y} \left\langle \phi_{XY}(X, cp_Y)^T \right\rangle_{\sim Q(Y)}$ （5）

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

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

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

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 $\tilde{\phi}_{XY}(\langle u_X\rangle,\langle u_\beta\rangle)=\begin{bmatrix}\langle u_\beta\rangle\langle u_X\rangle\\-\langle u_\beta\rangle/2\end{bmatrix}$

- ## 下界 L(Q)的计算

在贝叶斯网络中,由于 Q 可因式分解，则有

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

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

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

在变分消息传递算法中，$\langle\phi_j(pa_j)\rangle$ 和 ϕ_j^* 在找 H_j 的后验分布时就已经计算了；$\langle u_j(H_j)\rangle$ 在 H_j 传出消息的时候也已经计算了。这样下界节约了计算成本。

特别地，对于每个观测变量 V_k 对下界的贡献值更简单，

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

- ## 变分消息传递算法

现在我们已经准确知道节点之间的消息应该由什么样的形式传递，那么定义变分消息传递算法：

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

  $$m_{Y\to X}=\langle u_Y\rangle.\quad（6）$$

- **消息传递给父节点（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 $\langle u_A\rangle=u_A$，

  $$m_{X\to Y}=\tilde{\phi}_{XY}\left(\langle u_X\rangle,\{m_{i\to X}\}_{i\in cp_Y}\right)\quad（7）$$

用 Y 接收来自父节点与子节点的所有消息来计算 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，

$$\phi_Y^* = \tilde{\phi}_Y\left(\{m_{i\to Y}\}_{i\in pa_Y}\right) + \sum_{j\in ch_Y} m_{j\to Y} \quad (8)$$

该式与（5）式一致。从（1）式可以看出自然统计量的期望 $\langle u_Y \rangle_{Q_Y^*}$ 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。变分消息传递算法通过反复迭代的消息传递来最优化变分分布。

完整的算法描述如下，

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 .

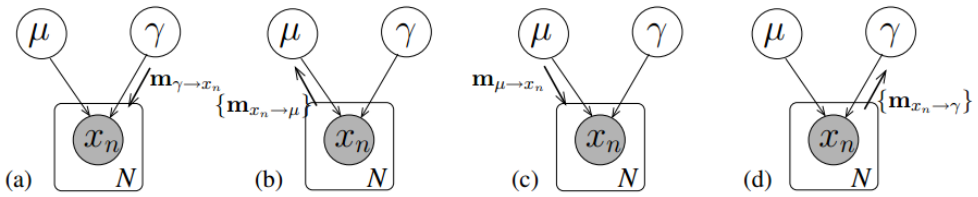
Step2. 对于每一个节点 X_j ，

- 从父节点和子节点接收 (6),(7)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（8）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 $L(Q)$（如果需要）

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；否则回到第二步。

举例：对于单一高斯模型（univariate Gaussian Model）消息传递过程如下图(a)(b)(c)(d)，



- ## 混合模型

到目前为止，我们只考虑了来自指数族的分布。而通常来讲混合分布(Mixture Distribution)并非来自指数族，比如高斯混合模型。那么我们是否能将这些混合分布转化为指数族的分布呢？

考虑高斯混合模型，通常有如下形式，

$$P(X\,|\,\{\pi_k\},\{\theta_k\}) = \sum_{k=1}^{K} \pi_k P_k(X\,|\,\theta_k)$$

我们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。在变分贝叶斯方法中，我们已经举过该例子，故不加以描述。重写分布函数为：

$$P(X|\{\pi_k\},\{\theta_k\}) = \sum_{k=1}^{K} \pi_k P_k(X|\theta_k)$$

加入该 λ 变量后该分布属于指数分布族，可写成

$$\ln P(X|\lambda,\{\theta_k\}) = \sum_k \delta(\lambda,k)\left[\phi_k(\theta_k)^T u_k(X) + f_k(X) + g_k(\theta_k)\right]$$

如果 X 有孩子 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，我们写成，

$$\begin{aligned}\ln P(X|\lambda,\{\theta_k\}) &= \left[\sum_k \delta(\lambda,k)\phi_k(\theta_k)\right]^T u_X(X) + f_X(X) + \sum_k \delta(\lambda,k) g_k(\theta_k) \\ &= \phi_X(\lambda,\{\theta_k\})^T u_X(X) + f_X(X) + \tilde{g}_X(\phi_X(\lambda,\{\theta_k\}))\end{aligned}$$

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。我们便可以应用变分消息传递算法。

从某个节点 X 传递个子节点的消息为 $\langle u_X(X)\rangle$，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 $\langle \ln P_k(X|\theta_k)\rangle$.

应用

- **隐马尔科夫模型(Hidden Markov Model, HMM)**
- **混合因子分析(mixture of factor analysers)**
- **线性动力系统（linear dynamical systems）**
- **图模型（Graphical models）**

总结

参考文献

[1] V. Smidl, A.Quinn(2005), The Variational Bayes Method In Signal Processing, *Signal and Communication Technology*,

[2] Matthew J.Beal(1998), Variational Algorithms for Approximate Bayesian Inference, London, UK: University of Cambridge, *PHD. Thesis*

[3] John M. Winn(2003), Variational Message Passing and its Applications, University of Cambridge , *PHD. Thesis*

[4] John M. Winn, M. Bishop(2004), Variational Message Passing, *Journal of Machine Learning Research*

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

[6] Charles W.Fox, Stephen J.Roberts(2011), A tutorial on variational Bayesian inference, *Artif Intell Rev*

[7] Jason Eisner(2011), High-Level Explanation of Variational Inference, http://www.cs.jhu.edu/~jason/tutorials/variational.html

[8] Michael I. Jordan, Z. Ghahramani(1999), An Introduction to Variational Methods for Graphical Models, *Machine Learning*.

[9] Tommi S. Jaakkola(2001), Tutorial on variational approximation methods, Advanced mean field methods: theory and practice

【注】文本仅为作者快速学习变分贝叶斯方法的读书笔记，不免有误。关于概率图模型，变分推断方法，作者今后 blog 将侧重于模型方法的思想，程序实现，更重要的是自己的心得体会。而对于模型本身，原文献介绍得很详细了，本 blog 不做翻译工作。O(∩_∩)O~