边际概率推断

1. 基于概率图模型定义的联合概率分布，能够对目标变量的边际分布进行推断。概率图模型的推断方法可以分为两大类：
   - 精确推断方法。该方法的缺点：通常情况下此类方法的计算复杂度随着极大团规模的增长呈现指数级增长，因此适用范围有限。
   - 近似推断方法。此类方法在实际任务中更常用。

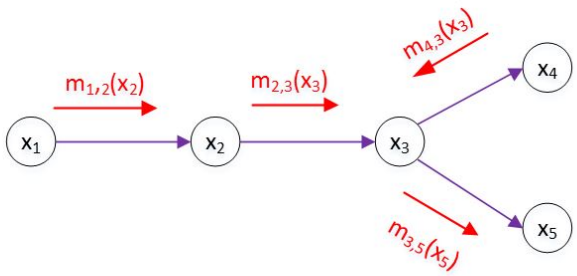
一、精确推断

1. 精确推断的实质是一类动态规划算法。它利用图模型所描述的条件独立性来降低计算量。

1.1 变量消去法

1. 变量消去法是最直观的精确推断算法。

   以下图为例来介绍变量消去法的工作流程。假设推断目标是计算边际概率 $P(X_5)$



贝叶斯网络结构

为了计算 $P(X_5)$，只需要通过加法消去变量 $\{X_1, X_2, X_3, X_4\}$ 即可。

$$P(X_5) = \sum_{X_1}\sum_{X_2}\sum_{X_3}\sum_{X_4} P(X_1, X_2, X_3, X_4, X_5)$$

其中联合概率分布 $P(X_1, X_2, X_3, X_4, X_5)$ 是已知的。

- 根据条件独立性有：
  $P(X_1, X_2, X_3, X_4, X_5) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)P(X_4 \mid X_3)P(X_5 \mid X_3)$ 。
  代入上式并重新安排 $\sum$ 的位置有：

$$P(X_5) = \sum_{X_3} P(X_5 \mid X_3) \sum_{X_4} P(X_4 \mid X_3) \sum_{X_2} P(X_3 \mid X_2) \sum_{X_1} P(X_1)P(X_2 \mid X_1)$$

- 定义：

$$m_{1,2}(X_2) = \sum_{X_1} P(X_1)P(X_2 \mid X_1)$$
$$m_{2,3}(X_3) = \sum_{X_2} P(X_3 \mid X_2)m_{1,2}(X_2)$$
$$m_{4,3}(X_3) = \sum_{X_4} P(X_4 \mid X_3)$$
$$m_{3,5}(X_5) = \sum_{X_3} P(X_5 \mid X_3)m_{2,3}(X_3)m_{4,3}(X_3)$$

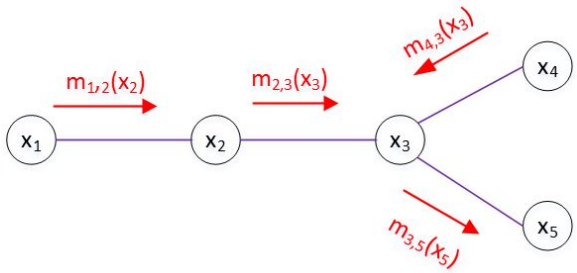
其中：$m_{1,2}(X_2)$ 仅仅是 X_2 的函数；$m_{2,3}(X_3), m_{4,3}(X_3)$ 仅仅是 X_3 的函数；$m_{3,5}(X_5)$ 仅仅是 X_5 的函数。

于是有：

$$\begin{aligned}
P(X_5) &= \sum_{X_3} P(X_5 \mid X_3) \sum_{X_4} P(X_4 \mid X_3) \sum_{X_2} P(X_3 \mid X_2) \sum_{X_1} P(X_1)P(X_2 \mid X_1) \\
&= \sum_{X_3} P(X_5 \mid X_3) \sum_{X_4} P(X_4 \mid X_3) \sum_{X_2} P(X_3 \mid X_2)m_{1,2}(X_2) \\
&= \sum_{X_3} P(X_5 \mid X_3) \sum_{X_4} P(X_4 \mid X_3)m_{2,3}(X_3) \\
&= \sum_{X_3} P(X_5 \mid X_3)m_{2,3}(X_3) \sum_{X_4} P(X_4 \mid X_3) \\
&= \sum_{X_3} P(X_5 \mid X_3)m_{2,3}(X_3)m_{4,3}(X_3) \\
&= m_{3,5}(X_5)
\end{aligned}$$

实际上图中有 $m_{4,3}(X_3) = \sum_{X_4} P(X_4 \mid X_3) = 1$，最终 $P(X_5) = \sum_{X_3} P(X_5 \mid X_3)m_{2,3}(X_3)$

2. 如果是无向图模型，上述方法同样适用。



马尔可夫随机场

- 根据马尔可夫随机场的联合概率分布有：

$$P(X_1, X_2, X_3, X_4, X_5) = \frac{1}{Z}\psi_{1,2}(X_1, X_2)\psi_{2,3}(X_2, X_3)\psi_{3,4}(X_3, X_4)\psi_{3,5}(X_3, X_5)$$

边际分布：

$$P(X_5)=\sum_{X_1}\sum_{X_2}\sum_{X_3}\sum_{X_4}P(X_1,X_2,X_3,X_4,X_5)$$
$$=\sum_{X_1}\sum_{X_2}\sum_{X_3}\sum_{X_4}\frac{1}{Z}\psi_{1,2}(X_1,X_2)\psi_{2,3}(X_2,X_3)\psi_{3,4}(X_3,X_4)\psi_{3,5}(X_3,X_5)$$

- 定义：

$$m_{1,2}(X_2)=\sum_{X_1}\psi_{1,2}(X_1,X_2)$$
$$m_{2,3}(X_3)=\sum_{X_2}\psi_{2,3}(X_2,X_3)m_{1,2}(X_2)$$
$$m_{4,3}(X_3)=\sum_{X_4}\psi_{3,4}(X_3,X_4)$$
$$m_{3,5}(X_5)=\sum_{X_3}\psi_{3,5}(X_3,X_5)m_{2,3}(X_3)m_{4,3}(X_3)$$

其中：$m_{1,2}(X_2)$ 仅仅是 X_2 的函数；$m_{2,3}(X_3),m_{4,3}(X_3)$ 仅仅是 X_3 的函数；$m_{3,5}(X_5)$ 仅仅是 X_5 的函数。

于是有：

$$P(X_5)=\sum_{X_1}\sum_{X_2}\sum_{X_3}\sum_{X_4}\frac{1}{Z}\psi_{1,2}(X_1,X_2)\psi_{2,3}(X_2,X_3)\psi_{3,4}(X_3,X_4)\psi_{3,5}(X_3,X_5)$$
$$=\frac{1}{Z}\sum_{X_3}\psi_{3,5}(X_3,X_5)\sum_{X_4}\psi_{3,4}(X_3,X_4)\sum_{X_2}\psi_{2,3}(X_2,X_3)\sum_{X_1}\psi_{1,2}(X_1,X_2)$$
$$=\frac{1}{Z}\sum_{X_3}\psi_{3,5}(X_3,X_5)\sum_{X_4}\psi_{3,4}(X_3,X_4)\sum_{X_2}\psi_{2,3}(X_2,X_3)m_{1,2}(X_2)$$
$$=\frac{1}{Z}\sum_{X_3}\psi_{3,5}(X_3,X_5)\sum_{X_4}\psi_{3,4}(X_3,X_4)m_{2,3}(X_3)$$
$$=\frac{1}{Z}\sum_{X_3}\psi_{3,5}(X_3,X_5)m_{4,3}(X_3)m_{2,3}(X_3)$$
$$=\frac{1}{Z}m_{3,5}(X_5)$$

  - 其中的 Z 为归一化常量，使得 $P(X_5)$ 满足概率的定义。
  - 这里 $m_{4,3}(X_3)\neq 1$，这就是无向概率图模型和有向概率图模型的一个重要区别。

3. 变量消去法通过利用乘法对加法的分配律，将多个变量的积的求和转化为对部分变量交替进行求积与求和的问题。
   - 优点：这种转化使得每次的求和与求积运算限制在局部，仅与部分变量有关，从而简化了计算。
   - 缺点：若需要计算多个边际分布，重复使用变量消去法将会造成大量的冗余计算。

     如：如果要同时计算 $P(X_4),P(X_5)$，则变量消去法会重复计算 $m_{1,2}(X_2),m_{2,3}(X_3)$。

1.2 信念传播

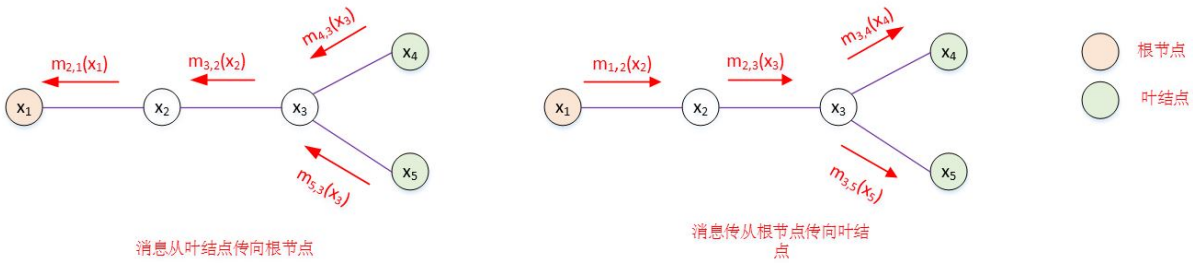
1. 信念传播`Belief Propagation`算法将变量消去法中的求和操作看作一个消息传递过程，解决了求解多个边际分布时的重复计算问题。
2. 在变量消去法中，求和操作为：$m_{i,j}(X_j)=\sum_{X_i}\psi(X_i,X_j)\prod_{k\in n(i),k\neq j}m_{k,i}(X_i)$。其中 $n(i)$ 表示结点 X_i 的相邻结点的下标集合。

在信念传播算法中，这个操作被看作从 X_i 向 X_j 传递了一个消息 $m_{i,j}(X_j)$ 。这样，变量消去的过程就可以描述为消息传递过程。

每次消息传递操作仅与 X_i 及其邻接结点直接相关，消息传递相关的计算被限制在图的局部进行。

3. 在信念传播算法中：
   - 一个结点仅在接收到来自其他所有结点的消息后才能向另一个结点发送消息。
   - 结点的边际分布正比于它所接收的消息的乘积：$P(X_i) \propto \prod_{k \in n(i)} m_{k,i}(X_i)$ 。
4. 如果图结构中没有环，则信念传播算法经过两个步骤即可完成所有消息传递，进而计算所有变量上的边际分布：
   - 指定一个根节点，从所有叶结点开始向根节点传递消息，直到根节点收到所有邻接结点的消息。
   - 从根节点开始向叶结点传递消息，直到所有叶结点均收到消息。

> 此时每条边上都有方向不同的两条消息



二、近似推断

1. 精确推断方法通常需要很大的计算开销，因此在现实应用中近似推断方法更为常用。

   近似推断方法可以分作两类：
   - 采样 `sampling` 。通过使用随机化方法完成近似。
   - 使用确定性近似完成近似推断，典型代表为变分推断 `variantional inference` 。

2.1 MCMC 采样

1. `MCMC` 采样是一种常见的采样方法，可以用于概率图模型的近似推断。其原理部分参考数学基础部分的 `蒙特卡洛方法与 MCMC 采样` 。

2.2 变分推断

1. 变分推断通过使用已知简单分布来逼近需要推断的复杂分布，并通过限制近似分布的类型，从而得到一种局部最优、但具有确定解的近似后验分布。
2. 给定多维随机变量 $\vec{\mathbf{x}} = (x_1, x_2, \cdots, x_d)^T$ ，其中每个分量都依赖于随机变量 $\vec{\mathbf{z}}$ 。假定 $\vec{\mathbf{x}}$ 是观测变量， $\vec{\mathbf{z}}$ 是隐含变量。

   推断任务是：由观察到的随机变量 $\vec{\mathbf{x}}$ 来估计隐变量 $\vec{\mathbf{z}}$ 和分布参数变量 Θ ， 即求解 $p(\vec{\mathbf{z}} \mid \vec{\mathbf{x}}; \Theta)$ 和 Θ 。

   Θ 的估计可以使用 `EM` 算法： （设数据集 $\mathbb{D} = \{\vec{\mathbf{x}}_1, \cdots, \vec{\mathbf{x}}_N\}$ ）
   - 在 `E` 步：根据 t 时刻的参数 $\Theta^{<t>}$ ，计算 Q 函数：

$$Q(\Theta; \Theta^{<t>}) = \sum_{i=1}^{N} \sum_{\vec{\mathbf{z}}} \ln p(\vec{\mathbf{x}} = \vec{\mathbf{x}}_i, \vec{\mathbf{z}}; \Theta) p(\vec{\mathbf{z}} \mid \vec{\mathbf{x}} = \vec{\mathbf{x}}_i; \Theta^{<t>})$$

- 在`M`步：基于`E`步的结果进行最大化寻优：$\Theta^{<t+1>}=\arg\max_\Theta Q(\Theta;\Theta^{<t>})$。

3. 根据`EM`算法的原理知道，$p(\vec{\mathbf{z}}\mid\vec{\mathbf{x}};\Theta^{<t>})$是隐变量$\vec{\mathbf{z}}$的一个近似后验分布。

   事实上我们可以人工构造一个概率分布$q(\vec{\mathbf{z}};\lambda)$来近似后验分布$p(\vec{\mathbf{z}}\mid\vec{\mathbf{x}})$，其中$\lambda$为参数。

   如：$q(\vec{\mathbf{z}};\lambda)=c_1\mathcal{N}(\vec{\mu}_1,\sigma_1^2\mathbf{I})+c_2\mathcal{N}(\vec{\mu}_2,\sigma_2^2\mathbf{I})$，其中$\lambda=(c_1,c_2,\vec{\mu}_1,\vec{\mu}_2,\sigma_1^2,\sigma_2^2)$为参数，$\mathcal{N}$表示正态分布。

   - 这样构造的$q(\vec{\mathbf{z}};\lambda)$与$p(\vec{\mathbf{z}}\mid\vec{\mathbf{x}};\Theta^{<t>})$的作用相同，它们都是对$p(\vec{\mathbf{z}}\mid\vec{\mathbf{x}})$的一个近似。

     但是选择构造$q(\vec{\mathbf{z}};\lambda)$的优势是：可以选择一些性质较好的分布。
   - 根据后验概率的定义，对于每个$\vec{\mathbf{x}}=\vec{\mathbf{x}}_i$都需要构造对应的$q_i(\vec{\mathbf{z}};\lambda)$。

4. 根据$p(\vec{\mathbf{x}})=p(\vec{\mathbf{x}},\vec{\mathbf{z}})/p(\vec{\mathbf{z}}\mid\vec{\mathbf{x}})=\frac{p(\vec{\mathbf{x}},\vec{\mathbf{z}})}{q(\vec{\mathbf{z}};\lambda)}/\frac{p(\vec{\mathbf{z}}|\vec{\mathbf{x}})}{q(\vec{\mathbf{z}};\lambda)}$，两边同时取对数有：

$$\log p(\vec{\mathbf{x}})=\log\frac{p(\vec{\mathbf{x}},\vec{\mathbf{z}})}{q(\vec{\mathbf{z}};\lambda)}-\log\frac{p(\vec{\mathbf{z}}\mid\vec{\mathbf{x}})}{q(\vec{\mathbf{z}};\lambda)}$$

   同时对两边对分布$q(\vec{\mathbf{z}};\lambda)$求期望，由于$\log p(\vec{\mathbf{x}})$与$\vec{\mathbf{z}}$无关，因此有：

$$\begin{aligned}\log p(\vec{\mathbf{x}})&=\mathbb{E}_q\left[\log\frac{p(\vec{\mathbf{x}},\vec{\mathbf{z}})}{q(\vec{\mathbf{z}};\lambda)}\right]-\mathbb{E}_q\left[\log\frac{p(\vec{\mathbf{z}}\mid\vec{\mathbf{x}})}{q(\vec{\mathbf{z}};\lambda)}\right]\\&=\mathbb{E}_q\left[\log\frac{p(\vec{\mathbf{x}},\vec{\mathbf{z}})}{q(\vec{\mathbf{z}};\lambda)}\right]+KL(q(\vec{\mathbf{z}};\lambda)||p(\vec{\mathbf{z}}\mid\vec{\mathbf{x}}))\end{aligned}$$

   其中$KL(\cdot)$为`KL`散度(`Kullback-Leibler divergence`)，其定义为：

$$KL(p||q)=\int_{-\infty}^{+\infty}p(x)\log\frac{p(x)}{q(x)}dx$$

   我们的目标是使得$q(\vec{\mathbf{z}};\lambda)$尽可能靠近$p(\vec{\mathbf{z}}\mid\vec{\mathbf{x}})$，即：$\min_\lambda KL(q(\vec{\mathbf{z}};\lambda)||p(\vec{\mathbf{z}}\mid\vec{\mathbf{x}}))$。

   考虑到$\log p(\vec{\mathbf{x}})$与$\vec{\mathbf{z}}$无关，因此上述目标等价于：

$$\max_\lambda\mathbb{E}_q\left[\log\frac{p(\vec{\mathbf{x}},\vec{\mathbf{z}})}{q(\vec{\mathbf{z}};\lambda)}\right]$$

   称$\mathbb{E}_q[\log p(\vec{\mathbf{x}},\vec{\mathbf{z}})-\log q(\vec{\mathbf{z}};\lambda)]$为`ELBO:Evidence Lower Bound`。

   > $p(\vec{\mathbf{x}})$是观测变量的概率，一般被称作`evidence`。因为$KL(q(\vec{\mathbf{z}};\lambda)||p(\vec{\mathbf{z}}\mid\vec{\mathbf{x}}))>0$，所以有：$\log p(\vec{\mathbf{x}})\geq\mathbb{E}_q[\log p(\vec{\mathbf{x}},\vec{\mathbf{z}})-\log q(\vec{\mathbf{z}};\lambda)]$。因此它被称作`Evidence Lower Bound`。

5. 考虑`ELBO`：

$$ELBO=\mathbb{E}_q[\log p(\vec{\mathbf{x}},\vec{\mathbf{z}})-\log q(\vec{\mathbf{z}};\lambda)]=\mathbb{E}_q\log p(\vec{\mathbf{x}},\vec{\mathbf{z}})-H(q)$$

   - 第一项称作能量函数。为了使得`ELBO`最大，则它倾向于在$p(\vec{\mathbf{x}},\vec{\mathbf{z}})$较大的地方$q(\vec{\mathbf{z}};\lambda)$也较大。
   - 第二项为$q(\vec{\mathbf{z}};\lambda)$分布的熵。为了使得`ELBO`最大，则它倾向于$q(\vec{\mathbf{z}};\lambda)$为均匀分布。

6. 假设$\vec{\mathbf{z}}$可以拆解为一系列相互独立的子变量$\vec{\mathbf{z}}_k$，则有：$q(\vec{\mathbf{z}};\lambda)=\prod_{k=1}^K q_k(\vec{\mathbf{z}}_k;\lambda_k)$。这被称作平均场`mean field approximation`。

   此时`ELBO`为：

$$ELBO = \int_{\vec{\mathbf{z}}_1} \int_{\vec{\mathbf{z}}_2} \cdots \int_{\vec{\mathbf{z}}_K} \prod_{k=1}^{K} q_k(\vec{\mathbf{z}}_k;\lambda_k) \log p(\vec{\mathbf{x}},\vec{\mathbf{z}}_1,\vec{\mathbf{z}}_2,\cdots,\vec{\mathbf{z}}_K) d\vec{\mathbf{z}}_1 d\vec{\mathbf{z}}_2 \cdots d\vec{\mathbf{z}}_K$$
$$- \int_{\vec{\mathbf{z}}_1} \int_{\vec{\mathbf{z}}_2} \cdots \int_{\vec{\mathbf{z}}_K} \prod_{k=1}^{K} q_k(\vec{\mathbf{z}}_k;\lambda_k) \sum_{k=1}^{K} \log q_k(\vec{\mathbf{z}}_k;\lambda_k) d\vec{\mathbf{z}}_1 d\vec{\mathbf{z}}_2 \cdots d\vec{\mathbf{z}}_K$$

定义 $\bar{\vec{\mathbf{z}}}_k = (\vec{\mathbf{z}}_1, \cdots, \vec{\mathbf{z}}_{k-1}, \vec{\mathbf{z}}_{k+1}, \cdots, \vec{\mathbf{z}}_K)$，它就是 $\vec{\mathbf{z}}$ 中去掉 $\vec{\mathbf{z}}_k$ 的剩余部分。定义 $\bar{\lambda}_k$ 为 λ 中去掉 λ_k 的剩余部分。

- 考虑第一项：

$$\int_{\vec{\mathbf{z}}_1} \int_{\vec{\mathbf{z}}_2} \cdots \int_{\vec{\mathbf{z}}_K} \prod_{k=1}^{K} q_k(\vec{\mathbf{z}}_k;\lambda_k) \log p(\vec{\mathbf{x}},\vec{\mathbf{z}}_1,\vec{\mathbf{z}}_2,\cdots,\vec{\mathbf{z}}_K) d\vec{\mathbf{z}}_1 d\vec{\mathbf{z}}_2 \cdots d\vec{\mathbf{z}}_K$$
$$= \int_{\vec{\mathbf{z}}_j} q_j(\vec{\mathbf{z}}_j;\lambda_j) \left(\int \prod_{k=1,k\neq j}^{K} q_k(\vec{\mathbf{z}}_k;\lambda_k) \log p(\vec{\mathbf{x}},\vec{\mathbf{z}}_1,\vec{\mathbf{z}}_2,\cdots,\vec{\mathbf{z}}_K) d\vec{\mathbf{z}}_1 d\vec{\mathbf{z}}_2 \cdots d\vec{\mathbf{z}}_{j-1} d\vec{\mathbf{z}}_{j+1} \cdots d\vec{\mathbf{z}}_K \right) d\vec{\mathbf{z}}_j$$

考虑到括号内的内容为：

$$\int \prod_{k=1,k\neq j}^{K} q_k(\vec{\mathbf{z}}_k;\lambda_k) \log p(\vec{\mathbf{x}},\vec{\mathbf{z}}_1,\vec{\mathbf{z}}_2,\cdots,\vec{\mathbf{z}}_K) d\vec{\mathbf{z}}_1 d\vec{\mathbf{z}}_2 \cdots d\vec{\mathbf{z}}_{j-1} d\vec{\mathbf{z}}_{j+1} \cdots d\vec{\mathbf{z}}_K$$
$$= \int q(\bar{\vec{\mathbf{z}}}_j;\bar{\lambda}_j) \log p(\vec{\mathbf{x}},\vec{\mathbf{z}}) d\bar{\vec{\mathbf{z}}}_j = \mathbb{E}_{q(\bar{\vec{\mathbf{z}}}_j;\bar{\lambda}_j)}[\log p(\vec{\mathbf{x}},\vec{\mathbf{z}})]$$

因此第一项为：$\int_{\vec{\mathbf{z}}_j} q_j(\vec{\mathbf{z}}_j;\lambda_j) \mathbb{E}_{q(\bar{\vec{\mathbf{z}}}_j;\bar{\lambda}_j)}[\log p(\vec{\mathbf{x}},\vec{\mathbf{z}})] d\vec{\mathbf{z}}_j$ 。

- 考虑第二项：

$$\int_{\vec{\mathbf{z}}_1} \int_{\vec{\mathbf{z}}_2} \cdots \int_{\vec{\mathbf{z}}_K} \prod_{k=1}^{K} q_k(\vec{\mathbf{z}}_k;\lambda_k) \sum_{k=1}^{K} \log q_k(\vec{\mathbf{z}}_k;\lambda_k) d\vec{\mathbf{z}}_1 d\vec{\mathbf{z}}_2 \cdots d\vec{\mathbf{z}}_K$$
$$= \int_{\vec{\mathbf{z}}_1} q_1(\vec{\mathbf{z}}_1;\lambda_1) \int_{\vec{\mathbf{z}}_2} q_2(\vec{\mathbf{z}}_2;\lambda_2) \cdots \int_{\vec{\mathbf{z}}_K} q_K(\vec{\mathbf{z}}_K;\lambda_K) (\log q_1(\vec{\mathbf{z}}_1;\lambda_1) + \log q_2(\vec{\mathbf{z}}_2;\lambda_2) + \cdots$$
$$+ \log q_K(\vec{\mathbf{z}}_K;\lambda_K)) d\vec{\mathbf{z}}_1 d\vec{\mathbf{z}}_2 \cdots d\vec{\mathbf{z}}_K$$
$$= \int_{\vec{\mathbf{z}}_1} q_1(\vec{\mathbf{z}}_1;\lambda_1) \int_{\vec{\mathbf{z}}_2} q_2(\vec{\mathbf{z}}_2;\lambda_2) \cdots \int_{\vec{\mathbf{z}}_K} q_K(\vec{\mathbf{z}}_K;\lambda_K) \log q_1(\vec{\mathbf{z}}_1;\lambda_1) d\vec{\mathbf{z}}_1 d\vec{\mathbf{z}}_2 \cdots d\vec{\mathbf{z}}_K$$
$$+ \int_{\vec{\mathbf{z}}_1} q_1(\vec{\mathbf{z}}_1;\lambda_1) \int_{\vec{\mathbf{z}}_2} q_2(\vec{\mathbf{z}}_2;\lambda_2) \cdots \int_{\vec{\mathbf{z}}_K} q_K(\vec{\mathbf{z}}_K;\lambda_K) \log q_2(\vec{\mathbf{z}}_2;\lambda_2) d\vec{\mathbf{z}}_1 d\vec{\mathbf{z}}_2 \cdots d\vec{\mathbf{z}}_K$$
$$+ \cdots + \int_{\vec{\mathbf{z}}_1} q_1(\vec{\mathbf{z}}_1;\lambda_1) \int_{\vec{\mathbf{z}}_2} q_2(\vec{\mathbf{z}}_2;\lambda_2) \cdots \int_{\vec{\mathbf{z}}_K} q_K(\vec{\mathbf{z}}_K;\lambda_K) \log q_K(\vec{\mathbf{z}}_K;\lambda_K) d\vec{\mathbf{z}}_1 d\vec{\mathbf{z}}_2 \cdots d\vec{\mathbf{z}}_K$$

由于 $q_i(\vec{\mathbf{z}}_i), i = 1, 2, \cdots, K$ 构成了一个分布函数，因此：

$$\int \prod_{k=1,k\neq j}^{K} q_k(\vec{\mathbf{z}}_k;\lambda_k) d\vec{\mathbf{z}}_1 d\vec{\mathbf{z}}_2 \cdots d\vec{\mathbf{z}}_{j-1} d\vec{\mathbf{z}}_{j+1} \cdots d\vec{\mathbf{z}}_K = 1$$

则有：

$$\int_{\vec{\mathbf{z}}_1}\int_{\vec{\mathbf{z}}_2}\cdots\int_{\vec{\mathbf{z}}_K}\prod_{k=1}^{K}q_k(\vec{\mathbf{z}}_k;\lambda_k)\sum_{k=1}^{K}\log q_k(\vec{\mathbf{z}}_k;\lambda_k)d\vec{\mathbf{z}}_1 d\vec{\mathbf{z}}_2\cdots d\vec{\mathbf{z}}_K$$
$$=\int_{\vec{\mathbf{z}}_1}q_1(\vec{\mathbf{z}}_1;\lambda_1)\log q_1(\vec{\mathbf{z}}_1;\lambda_1)d\vec{\mathbf{z}}_1+\int_{\vec{\mathbf{z}}_2}q_2(\vec{\mathbf{z}}_2;\lambda_2)\log q_2(\vec{\mathbf{z}}_2;\lambda_2)d\vec{\mathbf{z}}_2+\cdots$$
$$+\int_{\vec{\mathbf{z}}_K}q_K(\vec{\mathbf{z}}_K;\lambda_K)\log q_K(\vec{\mathbf{z}}_K;\lambda_K)d\vec{\mathbf{z}}_K$$
$$=\sum_{k=1}^{K}\int_{\vec{\mathbf{z}}_k}q_k(\vec{\mathbf{z}}_k;\lambda_k)\log q_k(\vec{\mathbf{z}}_k;\lambda_k)d\vec{\mathbf{z}}_k$$

即：

$$ELBO=\int_{\vec{\mathbf{z}}_j}q_j(\vec{\mathbf{z}}_j;\lambda_j)\mathbb{E}_{q(\bar{\vec{\mathbf{z}}}_j;\bar{\lambda}_j)}[\log p(\vec{\mathbf{x}},\vec{\mathbf{z}})]d\vec{\mathbf{z}}_j-\sum_{k=1}^{K}\int_{\vec{\mathbf{z}}_k}q_k(\vec{\mathbf{z}}_k;\lambda_k)\log q_k(\vec{\mathbf{z}}_k;\lambda_k)d\vec{\mathbf{z}}_k$$

7. 定义一个概率分布 $q_j^*(\vec{\mathbf{z}}_j,\lambda_j)=\frac{1}{C}\exp\{\mathbb{E}_{q(\bar{\vec{\mathbf{z}}}_j;\bar{\lambda}_j)}[\log p(\vec{\mathbf{x}},\vec{\mathbf{z}})]\}$，其中 C 是与 λ_j 有关、与 $\vec{\mathbf{z}}_j$ 无关的常数项。

   则有：

$$ELBO=\int_{\vec{\mathbf{z}}_j}q_j(\vec{\mathbf{z}}_j;\lambda_j)[\log C+\log q_j^*(\vec{\mathbf{z}}_j,\lambda_j)]d\vec{\mathbf{z}}_j-\sum_{k=1}^{K}\int_{\vec{\mathbf{z}}_k}q_k(\vec{\mathbf{z}}_k;\lambda_k)\log q_k(\vec{\mathbf{z}}_k;\lambda_k)d\vec{\mathbf{z}}_k$$
$$=\log C+\int_{\vec{\mathbf{z}}_j}q_j(\vec{\mathbf{z}}_j;\lambda_j)\log q_j^*(\vec{\mathbf{z}}_j,\lambda_j)-\int_{\vec{\mathbf{z}}_j}q_j(\vec{\mathbf{z}}_j;\lambda_j)\log q_j(\vec{\mathbf{z}}_j;\lambda_j)d\vec{\mathbf{z}}_j$$
$$-\sum_{k=1,k\neq j}^{K}\int_{\vec{\mathbf{z}}_k}q_k(\vec{\mathbf{z}}_k;\lambda_k)\log q_k(\vec{\mathbf{z}}_k;\lambda_k)d\vec{\mathbf{z}}_k$$

   其中 $-\sum_{k=1,k\neq j}^{K}\int_{\vec{\mathbf{z}}_k}q_k(\vec{\mathbf{z}}_k;\lambda_k)\log q_k(\vec{\mathbf{z}}_k;\lambda_k)d\vec{\mathbf{z}}_k=H(q(\bar{\vec{\mathbf{z}}}_j,\bar{\lambda}_j))$，因此有：

$$ELBO=\log C-KL(q_j(\vec{\mathbf{z}}_j;\lambda_j)||q_j^*(\vec{\mathbf{z}}_j;\lambda_j))+H(q(\bar{\vec{\mathbf{z}}}_j,\bar{\lambda}_j))$$

   为求解 $\max_{\lambda_j}ELBO$，则可以看到当 $KL(q_j(\vec{\mathbf{z}}_j;\lambda_j)||q_j^*(\vec{\mathbf{z}}_j;\lambda_j))=0$ 时，$ELBO$ 取最大值。因此得到 $q_j(\vec{\mathbf{z}}_j;\lambda_j)$ 的更新规则：

$$q_1(\vec{\mathbf{z}}_1;\lambda_1)=q_1^*(\vec{\mathbf{z}}_1;\lambda_1)$$
$$q_2(\vec{\mathbf{z}}_2;\lambda_2)=q_2^*(\vec{\mathbf{z}}_2;\lambda_2)$$
$$q_3(\vec{\mathbf{z}}_3;\lambda_3)=q_3^*(\vec{\mathbf{z}}_3;\lambda_3)$$
$$\cdots$$

   根据 $q_j^*(\vec{\mathbf{z}}_j,\lambda_j)$ 可知：在对 q_j 进行更新时，融合了 $\vec{\mathbf{z}}_j$ 之外的其他 $\bar{\vec{\mathbf{z}}}_j$ 的信息。

8. 在实际应用变分法时，最重要的是考虑如何对隐变量 $\vec{\mathbf{z}}$ 进行拆解，以及假设各种变量子集服从何种分布。

   如果隐变量 $\vec{\mathbf{z}}$ 的拆解或者变量子集的分布假设不当，则会导致变分法效率低、效果差。