2022-07-17~2022-08-21

9 Mixture Models and EM 和 10 章 Variational Inference 部分的总结。

1 EM 与 GEM

第 9 章介绍 EM 算法，涉及 3 种变量：已知的可观测变量 X，隐变量 Z，参数 $\boldsymbol{\theta}$。核心公式是如下分解：

$$\ln p(\mathrm{X}|\boldsymbol{\theta}) = \mathcal{L}(q,\boldsymbol{\theta}) + \mathrm{KL}(q||p)$$

$$\mathcal{L}(q,\boldsymbol{\theta}) = \sum_{\mathrm{Z}} q(\mathrm{Z}) \ln \frac{p(\mathrm{X},\mathrm{Z}|\boldsymbol{\theta})}{q(\mathrm{Z})}$$

$$\mathrm{KL}(q||p) = -\sum_{\mathrm{Z}} q(\mathrm{Z}) \ln \frac{p(\mathrm{Z}|\mathrm{X},\boldsymbol{\theta})}{q(\mathrm{Z})}$$

我们强调，这里的分布都是 condition on $\boldsymbol{\theta}$，注意与第 10 章介绍变分推断时作对比。这里假设隐变量是离散的，因此使用的是求和符号，对于连续变量，使用积分符号即可。

变分下界 $\mathcal{L}(q,\boldsymbol{\theta})$： a **functional (泛函)** of the distribution $q(\mathrm{Z})$, a **function (函数)** of the parameters $\boldsymbol{\theta}$。

EM 算法: a two-stage iterative optimization technique of finding maximum likelihood solutions. 对变分下界进行坐标上升迭代优化：

1. 固定 $\boldsymbol{\theta}$ 优化 $q(\mathrm{Z})$，易得 $q(\mathrm{Z})$ 应该取 $p(\mathrm{Z}|\mathrm{X},\boldsymbol{\theta})$；进一步，为了给下面对 $\boldsymbol{\theta}$ 的优化做准备，我们将 $q(\mathrm{Z}) = p(\mathrm{Z}|\mathrm{X},\boldsymbol{\theta})$ 代入变分下界中，计算式中关于 Z 的积分，也就是 $\ln \frac{p(\mathrm{X},\mathrm{Z}|\boldsymbol{\theta})}{q(\mathrm{Z})}$ 关于 $q(\mathrm{Z})$ 的期望。这一步虽然被称为 E 步 (期望步)，但本质并不是求期望，仍然和下面的 M 步一样，关键点仍然是优化；
2. 优化参数 $\boldsymbol{\theta}$ (固定 $q(\mathrm{Z})$ 为 $p(\mathrm{Z}|\mathrm{X},\boldsymbol{\theta}^{old})$，优化的是 $p(\mathrm{X},\mathrm{Z}|\boldsymbol{\theta})$ 中的 $\boldsymbol{\theta}$)，这一步被称为 M 步。

可以看到，EM 算法就是变分下界的坐标上升优化算法，E步、M 步都是在进行优化，只不过E步针对 $q(\mathrm{Z})$，M步针对 $\boldsymbol{\theta}$。

EM 算法的一个例子是 GMM，套用上述公式即可，不再赘述。另一个例子是用在 Bayesian linear regression 中，通过 max model evidence $p(\mathbf{t}|\alpha,\beta)$ 用于优化超参数 α,β。此时，数据 $\mathbf{t}$ 对应可观测变量 X，系数 w 对应隐变量 Z，超参数 α,β 对应参数 $\boldsymbol{\theta}$，见 9.3.4 EM for Bayesian linear regression 小节。

EM 算法除了可以进行 MLE，也可以进行 MAP。MAP 时，E步不变，M步在优化 $\boldsymbol{\theta}$ 时优化目标加上 $\ln p(\boldsymbol{\theta})$ 即可，推导如下：

$$
\begin{aligned}
\ln p(\mathrm{X}|\boldsymbol{\theta}) &= \mathcal{L}(q,\boldsymbol{\theta}) + \mathrm{KL}(q||p) \\
\ln p(\boldsymbol{\theta}|\mathrm{X}) &= \ln p(\boldsymbol{\theta},\mathrm{X}) - \ln p(\mathrm{X}) \\
&= \ln p(\mathrm{X}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - \ln p(\mathrm{X}) \\
&= \mathcal{L}(q,\boldsymbol{\theta}) + \mathrm{KL}(q||p) + \ln p(\boldsymbol{\theta}) - \ln p(\mathrm{X}) \\
&\geqslant \mathcal{L}(q,\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - \ln p(\mathrm{X}) \\
\max_{\boldsymbol{\theta}} \ln p(\boldsymbol{\theta}|\mathrm{X}) &\Leftrightarrow \max_{q,\boldsymbol{\theta}} \mathcal{L}(q,\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})
\end{aligned}
$$

$\ln p(\mathrm{X})$ 对 $q, \boldsymbol{\theta}$ 而言是常数，可忽略；相比前面介绍的 MLE 的情况，目标函数就多了一项 $\ln p(\boldsymbol{\theta})$，这对 E 步优化 $q(\mathrm{Z})$ 时没有影响，在 M 步优化 $\boldsymbol{\theta}$ 时加上这一项即可。

EM 算法的另一个变化 generalized EM (GEM) 实际上就是不强求E步、M步的每次优化都求得最优解，而是放宽要求，只要不断上升就行。

2 Variational Inference / Variational Bayes

推断任务：A central task in the application of probabilistic models is the evaluation of the posterior distribution $p(\mathrm{Z}|\mathrm{X})$ of the latent variables Z given the observed (visible) data variables X, and the evaluation of expectations computed with respect to this distribution.

推断分为：

1. 精确推断；
2. 近似推断：
    1. 随机近似：MCMC 等；
    2. 确定性近似：Laplace 近似，变分推断等。

隐变量和参数的辩证关系：The probabilistic model might also contain some deterministic parameters, which we will leave implicit for the moment, or it may be a fully Bayesian model in which any unknown parameters are given prior distributions and are absorbed into the set of latent variables denoted by the vector Z. **这里参数 $\boldsymbol{\theta}$ 不再显示给出，而是将其打包进 Z 一起进行分析。**因此，相比第 9 章介绍的 EM 算法进行 MLE 时 condition on $\boldsymbol{\theta}$，这里是一种贝叶斯方法。虽然 EM 算法也可以进行 MAP 但实现方式和这里并不相同，两种做法并不等价。This differs from our discussion of EM only in that the parameter vector $\boldsymbol{\theta}$ no longer appears, because the parameters are now stochastic variables and are absorbed into Z. 在概率图模型中，通常将可观测变量记为 x_n，下标 n 表示第 n 个样本，将数量随数据集大小变化的不可观测变量记为隐变量 z_n，而将数量不随数据集大小变化的不可观测变量记为参数 $\boldsymbol{\theta}$。但是，从数学角度看，隐变量与参数本质上并没有区别，因此，在变分推断中，我们将它们统统打包进 Z 中进行分析。This example (指10.2 小节 variational mixture of Gaussians) provides a nice illustration of the distinction between latent variables and parameters. Variables such as z_n that appear inside the plate are regarded as latent variables because the number of such variables grows with the size of the data set. By contrast, variables such as $\boldsymbol{\mu}$ that are outside the plate are fifixed in number independently of the size of the data set, and so are regarded as parameters. From the perspective of graphical models, however, there is really no fundamental difference between them.

2.1 变分推断之平均场模型

下面为了便于表述，有时候会像书中那样统称为 Z，有时会将 $\mathrm{Z}, \boldsymbol{\theta}$ 分开表述，注意区分相应的场景。变分下界：

$$\mathcal{L}(q)=\int q(\mathrm{Z})\ln\left\{\frac{p(\mathrm{X},\mathrm{Z})}{q(\mathrm{Z})}\right\}\mathrm{dZ}=\ln p(\mathrm{X})-\mathrm{KL}(q(\mathrm{Z})\|p(\mathrm{Z}|\mathrm{X}))$$

我们强调 EM 中的变分下界是关于 $q(\mathrm{Z})$ 的泛函和参数 $\boldsymbol{\theta}$ 的函数，而变分推断中，变分下界只是关于 $q(\mathrm{Z})$ 的泛函，参数 $\boldsymbol{\theta}$ 已经打包到 Z 中。因此，虽然都叫变分下界，但二者并不相同。通过优化变分下界，EM 优化的是 $p(\mathrm{X}|\boldsymbol{\theta})=\int p(\mathrm{X},\mathrm{Z}|\boldsymbol{\theta})\mathrm{dZ}$ 或 $p(\boldsymbol{\theta}|\mathrm{X})$，而变分推断优化的是 $p(\mathrm{X})=\int p(\mathrm{X},\mathrm{Z},\boldsymbol{\theta})\mathrm{dZd}\boldsymbol{\theta}$，并不是 MLE 或是 MAP，而是直接计算 $q(\mathrm{Z},\boldsymbol{\theta})$，它是 $p(\mathrm{Z},\boldsymbol{\theta}|\mathrm{X})$ 的近似分布，是一种贝叶斯方法。

此外，对于概率分布 p，无论 EM 还是变分推断，它的函数形式均已知，因此 $\mathrm{Z},\boldsymbol{\theta}$ 等后验分布均可以直接计算出来，那么为何还需要多此一举，进行变分推断计算近似后验 $q(\mathrm{Z},\boldsymbol{\theta})$ 呢，其意义何在，毕竟 EM 算法基于 p 的函数形式，完成了对参数 $\boldsymbol{\theta}$ 的点估计，是有存在的必要的 (事实上，除了 EM 算法，还可以用梯度下降算法)。我们指出，进行变分推断的原因是 $p(\mathrm{Z},\boldsymbol{\theta}|\mathrm{X})$ 可能非常复杂，以致无法使用，因此我们试图通过优化变分下界，寻找一个近似后验 $q(\mathrm{Z},\boldsymbol{\theta})$。

平均场模型只引入了如下的假设 restrict the family of distributions $q(\mathrm{Z})$：Suppose we partition the elements of Z into disjoint groups that we denote by Z_i where $i=1,\cdots,M$. We then assume that the q distribution factorizes with respect to these groups, so that

$$q(\mathrm{Z})=\prod_{i=1}^{M}q_i(\mathrm{Z_i})$$

接着对 $q_i(\mathrm{Z}_i)$ 进行坐标上升，得到书中式 (10.9) 迭代计算格式。

首先，我们从优化的角度考察变分推断。可以看到，$\ln p(\mathrm{X})$ 相对 $q(\mathrm{Z})$ 而言是常数，变分推断实际上就是在最小化 KL 散度 $\mathrm{KL}(q(\mathrm{Z})\|\mathrm{p}(\mathrm{Z}|\mathrm{X}))$：

$$\begin{gathered}\max_q\mathcal{L}(q)=\ln p(\mathrm{X})-\mathrm{KL}(q(\mathrm{Z})\|p(\mathrm{Z}|\mathrm{X}))\\\Updownarrow\\\min_q\mathrm{KL}(q(\mathrm{Z})\|p(\mathrm{Z}|\mathrm{X}))\end{gathered}$$

显然，用 q 近似后验分布，我们这里选择的是最小化 $\mathrm{KL}(q\|p)$，理论上当然也可以 $\mathrm{KL}(p\|q)$，而这种形式的 KL 散度 is used in an alternative approximate inference framework called **expectation propagation**.。书中对比了二者所得结果的区别，并给出了例子见Figure 10.2 和 Figure 10.3。总地来说，最小化 $\mathrm{KL}(q\|p)$ 得到的 q 会试图拟合 p 分布多个modal中的某一个，而其余modal处接近 0，也就是说拟合是局部的；而最小化 $\mathrm{KL}(p\|q)$ 得到的 q 则会试图拟合 p 整体，也就是会尝试去拟合多个 modal，在 p 有较大的值的区域，q 也一定保持有值而不是接近 0。

$$\mathrm{KL}(q\|p)=-\int q\ln\frac{p}{q}\mathrm{dZ}$$

为了最小化上述 KL 散度，$-q\ln p$ 这一项是关键。为此，p 靠近 0 的区域 q 一定也要靠近 0。而对

$$\mathrm{KL}(p\|q)=-\int p\ln\frac{q}{p}\mathrm{dZ}$$

只有 $-p\ln q$ 这一项起作用，$p\ln q$ 可视为常数，因为 p 是已知的。为了最小化上述 KL散度，我们需要保证，对于 p 不为 0 比较大的区域，q 也不能接近 0。

事实上，这两种形式的 KL 散度的统一形式是 alpha family of divergences:

$$\mathrm{D}_\alpha(p\|q)=\frac{4}{1-\alpha^2}\left(1-\int p(x)^{(1+\alpha)/2}q(x)^{(1-\alpha)/2}\right)$$

where where $-\infty < \alpha < \infty$ is a continuous parameter. The Kullback-Leibler divergence $\mathrm{KL}(p\|q)$ corresponds to the limit $\alpha \to 1$, whereas $\mathrm{KL}(q\|p)$ corresponds to the limit $\alpha \to -1$.

下面，我们从逻辑和计算两个角度考察变分推断与 EM 算法的关系。

首先，**从逻辑上看**变分推断不再 condition on $\boldsymbol{\theta}$，参数和隐变量 Z 一起以泛函的形式出现。因此对比 EM 算法，变分推断的 E 步退化消失了，不需要再对参数 $\boldsymbol{\theta}$ 进行优化，全部的优化都在 M 步中，而且 $\boldsymbol{\theta}$ 的优化不再以变量，而是以泛函的形式进行。对于 M 步，针对平均场假设，变分推断依然是使用坐标上升对泛函进行优化，得到迭代优化公式。

可以看到，变分推断所有的优化均在 M 步进行，若有平均场假设 $q(\mathrm{Z}, \boldsymbol{\theta}) = q(\mathrm{Z})q(\boldsymbol{\theta})$，则迭代优化关于泛函 $q(\boldsymbol{\theta})$ 的部分**在计算上**又可以与 EM 算法的 E 步相对应，只不过优化的是泛函而不是变量 (变分推断是一种贝叶斯方法，没有对参数做点估计)；而优化关于 $q(\mathrm{Z})$ 的部分则在计算上与 M 步对应。

上面梳理了《PRML》中关于 EM、GEM、VI 的定义和内容。可以看到 《PRML》还是严谨呐！其他文献就会对这些概念混着用，让人搞不懂为什么一会是 EM，一会是变分推断，一会又是广义 EM。不过，上述内容的核心是一致的，就是对变分下界的坐标上升优化，因此在使用过程中可以很灵活，不必拘泥于某一概念或定义。比如，我们可以 condition on $\boldsymbol{\theta}$，然后，在 M 步优化时又限制 $q(\mathrm{Z})$ 的函数空间 (平均场假设)，这样 $q(\mathrm{Z})$ 就取不到 $p(\mathrm{Z}|\mathrm{X}, \boldsymbol{\theta})$，再在 M 步内进行泛函优化，此时，M 步本身又包含了一套坐标上升的迭代计算。上述做法相当于是 EM 算法和变分推断的结合。进一步，E 步和 M 步内的优化又可以不用等到收敛，而是只要有所提高就停止进行接下来的 E 步或 M 步，这又有了 GEM 的影子。

2.2 Variational Mixture of Gaussians

10.2 Illustration: Variational Mixture of Gaussians 以高斯混合模型为例，给出了变分法应用的全过程。

首先，我们手头上只有可观测数据 X，我们试图通过 GMM 对真实分布进行建模 (也就是说 $p(\mathrm{X}, \mathrm{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ 的函数形式已知，并确定为 GMM)，这是第一层近似；而使用变分法引入分布 $q(\mathrm{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ 近似 GMM 的后验分布 $p(\mathrm{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathrm{X})$ (注意，不是 $p(\mathrm{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$) 则是第二层近似。我们并没有假设 $q(\mathrm{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ 的函数形式，而只是取平均场假设：

$$q(\mathrm{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\mathrm{Z})q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$$

接着，套用前面基于坐标上升法得到的平均场模型的迭代优化公式，自然而然地就可以推得 $q(\mathrm{Z}), q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ 的函数形式，这就是 10.2.1 小节的内容。和前文类似，我们同样可以分别从逻辑和计算的角度与 GMM 的 EM 算法进行对比。此外，GMM 的变分推断是一种贝叶斯方法，因此会有正则化效应，比如文中给出的 the rescaled Old Faithful data set 的例子。10.2.2 小节给出了 GMM 变分下界中各项的详细计算公式；10.2.3 小节解决了预测问题；10.2.4 小节给出了模型选择的两种方案，一种类似于 AIC，BIC，KIC 等模型选择指标，另一种则类似于相关向量机，通过学习自动进行模型选择；10.2.5 小节则讨论了 induced factorizations。所谓 induced factorizations 就是基于平均场假设和模型本身的条件独立性而额外引入的 factorizations，比如 GMM 中我们只假设了因子分解 $q(\mathrm{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\mathrm{Z})q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$，但在此基础上结合 GMM 本身的条件独立性，我们可推得最优的 $q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ 可作进一步进行分解，即 $q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{\pi})q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$。induced factorizations 可以简化计算，文中也给出了找 induced factorizations 的方法。

下面，我们对一些细节进行说明。

首先是 GMM 及其变分下界，我们只以一个样本点 x 为例进行说明：

$$p(\mathrm{x}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})=\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathrm{x}|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1})$$

$$p(\mathrm{x}|\mathrm{z},\boldsymbol{\mu},\boldsymbol{\Lambda})=\sum_{k=1}^{K}z_k\mathcal{N}(\mathrm{x}|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1})=\prod_{k=1}^{K}\mathcal{N}^{z_k}(\mathrm{x}|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1})$$

事实上，$p(\mathrm{x}|\mathrm{z},\boldsymbol{\mu},\boldsymbol{\Lambda})=p(\mathrm{x}|\mathrm{z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})$。虽然概率的记号都是 p，但：(1) p 不是一个通用的概率记号，而是表示 GMM；(2) 在 GMM 的大背景下，$p(\cdot)$ 的函数表达式又根据括号中的具体变量而定。此外，

$$p(\mathrm{x},\mathrm{z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})=p(\mathrm{x}|\mathrm{z},\boldsymbol{\mu},\boldsymbol{\Lambda})p(\mathrm{z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu},\boldsymbol{\Lambda})$$
$$p(\mathrm{x},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})=p(\mathrm{x}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})p(\boldsymbol{\pi})p(\boldsymbol{\mu},\boldsymbol{\Lambda})$$

二者之间是积分的关系，这里 z 为离散变量，即为求和关系：

$$p(\mathrm{x},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})=\sum_{\mathrm{z}}p(\mathrm{x},\mathrm{z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})$$

将 $\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}$ 统记为 $\boldsymbol{\theta}$，关于变分下界，实际上我们有含 Z 和不含 Z 两种选择：

$$\begin{aligned}L(q(\mathrm{Z},\boldsymbol{\theta}))&=\int q(\mathrm{Z},\boldsymbol{\theta})\ln\frac{p(\mathrm{X},\mathrm{Z},\boldsymbol{\theta})}{q(\mathrm{Z},\boldsymbol{\theta})}\mathrm{d}\mathrm{Z}\mathrm{d}\boldsymbol{\theta}\\&=\log p(\mathrm{X})-KL(q(\mathrm{Z},\boldsymbol{\theta})\|p(\mathrm{Z},\boldsymbol{\theta}|\mathrm{X}))\\\mathcal{L}(q(\boldsymbol{\theta}))&=\int q(\boldsymbol{\theta})\ln\frac{p(\mathrm{X},\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta}\\&=\log p(\mathrm{X})-\mathrm{KL}(q(\boldsymbol{\theta})\|p(\boldsymbol{\theta}|\mathrm{X}))\end{aligned}$$

虽然最大化这两个变分下界均对应着最大化 $p(\mathrm{X})$，但二者显然不是同一个量。再次强调：

- 注意区分最大化 $p(\mathrm{X})$ 与最大化 $p(\mathrm{X}|\boldsymbol{\theta})$，前者是一种贝叶斯方法，需要先验概率的参与，后者是 MLE，无需先验概率参与；
- 虽然 $p(\mathrm{X},\boldsymbol{\theta}),p(\mathrm{X},\mathrm{Z},\boldsymbol{\theta})$ 都是用的记号 p 但它们的含义或者说函数表达式并不相同，对 $q(\boldsymbol{\theta}),q(\mathrm{Z},\boldsymbol{\theta})$ 也同理。书中选择的是带 Z 的情况，但理论上对不带 Z 的变分下界也可以进行处理，只不过可能引入隐变量 Z 后处理起来反而更加方便。

文中提出的类似于 AIC、BIC、KIC 的模型选择指标是变分下界加上 $\ln K!$，即 $\mathcal{L}(q)+\ln K!$。下面，我们结合练习题 10.22 对其做出如下梳理。

首先，高斯混合模型由 K 个分量组成 (也就有 K 个 modal，即 K 个峰)：$p(\mathrm{x}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})=\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathrm{x}|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1})$。可以看到，For any given setting of the parameters in a Gaussian mixture model (except for specifific degenerate settings), there will exist other parameter settings for which the density over the observed variables will be identical. These parameter values differ only through a re-labelling of the components. 这也就意味着，参数 $\boldsymbol{\theta}=\{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}\}$ 的后验分布也是 multimodal，有 $K!$ 个 modal。如前所述，变分推断最小化的 KL 散度的形式为 $KL(q\|p)$，因此变分后验 q 只会试图近似参数后验 p 的 $K!$ 个 modal 中的某一个 modal，这是随机的，与初始设置有关。

下面，我们明确目标，即进行模型选择实际上就是要比较不同 K 值下 $p(\mathrm{X}|K)$ 的大小。而前文通过最大化变分下界而试图优化的 $p(\mathrm{X})$ 实际上就是 $p(\mathrm{X}|K)$ 只不过我们省略了条件项 K。既然如此，那么我们直接比较变分下界不就好了，为何还要加上 $\ln K!$ 呢？原因是，在模型选择的场景下，前面的变分下界对 $p(\mathrm{X}|K)$ 并不是一个很好的近似。正如上面的分析，优化得到的变分后验 q 只能近似参数后验 p 的一个 modal，这对推断来说足够了，因此前面我们说变分推断是一种贝叶斯方法，有一定的正则化效应。但对模型选择这是不够的，此时，我们不能只考虑优化得到的某一个变分后验 q 的变分下界，而应

该考虑优化得到的变分后验 (正如前面所说，每次优化得到的 q 具有随机性) 的平均情况。下面就开始进行推导，也就是练习题 10.22。

推导的前提：

- We can then approximate the full posterior distribution as a mixture of $K!$ such q distributions, once centred on each mode and having equal mixing coeffificients. 也就是说平均变分后验 $\bar{q}=\frac{1}{K!}\sum_{k=1}^{K!}q_k(\boldsymbol{\theta})$；
- we assume negligible overlap between the components of the q_k mixture. 也就是说，q_k 的 probability mass 主要集中在各自拟合的 mode 附近，而其他位置的概率为 0.

由此，我们开始计算 $\bar{q}$ 的变分下界 $\bar{\mathcal{L}}$：

$$
\begin{aligned}
\bar{\mathcal{L}} &= \int \bar{q}(\boldsymbol{\theta}) \ln \frac{p(\mathrm{X},\boldsymbol{\theta})}{\bar{q}(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} \\
&= \int \bar{q}(\boldsymbol{\theta}) \ln p(\mathrm{X},\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\
&\quad - \int \bar{q}(\boldsymbol{\theta}) \ln \bar{q}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\
&= \frac{1}{K!}\sum_{k=1}^{K!} \int q_k(\boldsymbol{\theta}) \ln p(\mathrm{X},\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\
&\quad - \frac{1}{K!}\sum_{k=1}^{K!} \int q_k(\boldsymbol{\theta}) \ln \left(\frac{1}{K!}\sum_{k'=1}^{K!} q_{k'}(\boldsymbol{\theta}) \right) \mathrm{d}\boldsymbol{\theta} \\
&= \int q_k(\boldsymbol{\theta}) \ln p(\mathrm{X},\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\
&\quad - \frac{1}{K!}\sum_{k=1}^{K!} \int q_k(\boldsymbol{\theta}) \left(\ln \frac{1}{K!} + \ln \left(\sum_{k'=1}^{K!} q_{k'}(\boldsymbol{\theta}) \right) \right) \mathrm{d}\boldsymbol{\theta} \\
&= \int q_k(\boldsymbol{\theta}) \ln p(\mathrm{X},\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\
&\quad - \frac{1}{K!}\sum_{k=1}^{K!} \int q_k(\boldsymbol{\theta}) \ln \left(\sum_{k'=1}^{K!} q_{k'}(\boldsymbol{\theta}) \right) \mathrm{d}\boldsymbol{\theta} + \frac{\ln K!}{K!} \\
&\approx \int q_k(\boldsymbol{\theta}) \ln p(\mathrm{X},\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} - \int q_k(\boldsymbol{\theta}) \ln q_k(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} + \frac{\ln K!}{K!} \\
&= \int q(\boldsymbol{\theta}) \ln p(\mathrm{X},\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} - \int q(\boldsymbol{\theta}) \ln q(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} + \frac{\ln K!}{K!} \\
&= \mathcal{L}(q(\boldsymbol{\theta})) + \frac{\ln K!}{K!}
\end{aligned}
$$

其中，基于第 2 个假设，我们有 $\int q_k(\boldsymbol{\theta}) \ln \left(\sum_{k'=1}^{K!} q_{k'}(\boldsymbol{\theta}) \right) \mathrm{d}\boldsymbol{\theta} = \int q_k(\boldsymbol{\theta}) \ln q_k(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$.

可以看到，$\frac{\ln K!}{K!}$ 衡量了模型的复杂度。此外，目前我们有 3 个变分下界 $\bar{\mathcal{L}}, \mathcal{L}(q(\boldsymbol{\theta})), \mathcal{L}(q(\mathrm{Z},\boldsymbol{\theta}))$，前面我们计算的都是变分下界 $\mathcal{L}(q(\mathrm{Z},\boldsymbol{\theta}))$，而这里涉及的是变分下界 $\bar{\mathcal{L}}, \mathcal{L}(q(\boldsymbol{\theta}))$，考虑到最优时 $\mathcal{L}(q(\boldsymbol{\theta})), \mathcal{L}(q(\mathrm{Z},\boldsymbol{\theta}))$ 应该差不多，所以个人认为将 $\mathcal{L}(q(\mathrm{Z},\boldsymbol{\theta})) + \frac{\ln K!}{K!}$ 作为模型的选择指标也是可以的。
