在现代统计信号处理中，ML估计是比较常用的一种关于概率密度函数的参数估计方法，但是实际中许多的参数估计问题并不能直接利用观测值来估计概率密度函数中的参数，因为有一些数据缺失的情况，或存在隐含变量（underlying variables）。举一个常见的例子：假设某校男生身高服从高斯分布，随机抽取100个男生（假设100个男生的身高都是相互独立的），我们可以根据ML参数估计（可参考上篇文章）来估计出男生身高概率分布的参数；同理，假设女生身高服从高斯分布，随机抽取100个女生，可以根据ML参数估计来估计出女生身高概率分布的参数。但是如果我们在该校抽取200个学生的身高作为样本，并不知道这200个样本的性别，此时我们用ML的方法估计参数往往就不可行了，因为并不知道单个样本服从什么分布了（可能是男生的身高分布，也可能是女生的身高分布），此时单个样本的性别就是上面所说的隐含变量（缺失的数据），也就无法直接求似然函数的最大值。

我们将上面的例子归纳一下，可以形成这样一个问题：假设我们有一个包含个样本的样本集，但每个样本的类别是未知的（缺失数据），即包含隐含变量，故实际上我们要估计的是概率模型的参数。根据ML估计参数得到的似然函数为：



（1）

由于隐含变量的存在，似然函数很难求解，但是如果能知道，就是一个纯粹的ML参数估计问题了，所以一个最直接的想法就是找到合适的和使得最大。



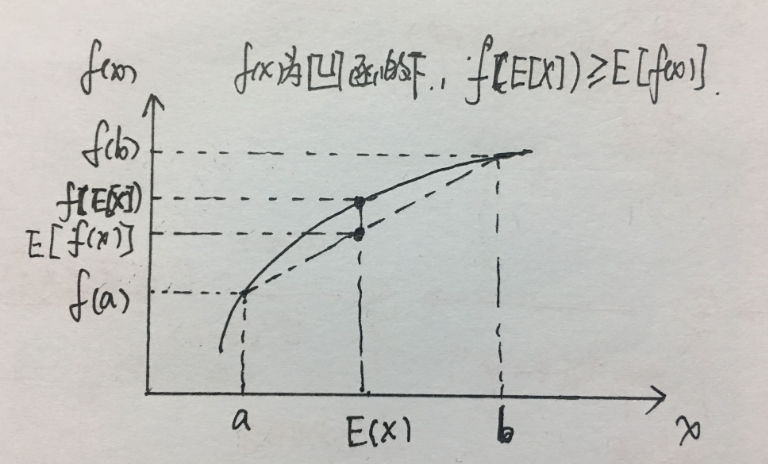
（2）

是关于的概率密度函数，具体表达式未知，直接求偏导，再令偏导为0的方法也是行不通的（无解析解）。

通过观察上式，可以发现，是的期望，又为严格凹函数，则根据Jensen不等式：

1. 如果是凸函数，是随机变量，则有：；
2. 如果是严格凸函数，当且仅当是常量时，等式成立。

凹函数下的Jensen不等式示意图如下图所示：



**利用Jensen不等式确定下界**：



（3）

这样我们就建立了的下界，假设我们给定了一个值（实际上在迭代的时候一定会给个初始值的），的值就取决于和，通过调整这两个概率分布函数来使下界不断上升，直到等式成立则认为调整好了。根据Jensen不等式可知，等式成立的条件是随机变量变成常数值，则有：



（4）

又有 ，则上式可化简为：



（5）



将（5）式代入（4）可得：



由此可知，当隐含变量的概率密度函数为后验概率时等号成立。

**优化下界：**

在给定隐含变量的概率密度函数后，调整参数，极大化的下界：



（6）

此时，根据（6）式估计出新的参数，假设为，若，即最大似然估计若单调增加，则最终会收敛到最大似然估计的最大值。

证明：

在（6）中极大化得到新的参数，则有：



（7）

其中，。

根据（3）式可知，（7）式的左边，即：



1. 式的左边，即有：



（8）

由此可知，最大似然估计单调增加，最终会收敛到最大似然估计的最大值。我们也可以从另一个角度看EM算法：

1）令为：



（9）

2）设KL散度：



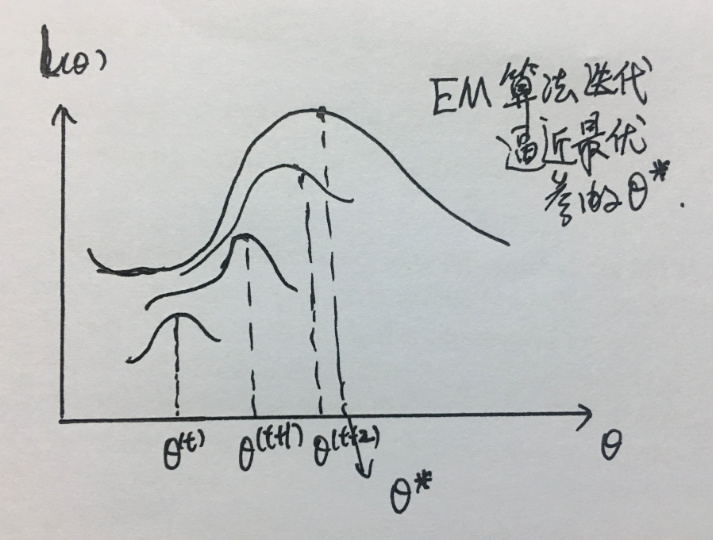
（10）

根据KL散度性质可知，，结合（3）、（9）和（10）式可知，似然函数可写为：

（11）



可用来衡量隐含变量的概率密度函数与的相似度，当与相等时，，此时，（9）式与（3）式相等。假设给定参数，在E-step中求使下界最大的，此时：；在M-step中，利用ML估计得出一个新的参数，使得，同时由于，导致，即（当时等号成立）。



至此，EM算法的迭代步骤如下：

1. 初始化参数；
2. E-step：根据初始参数或上一次迭代的模型参数，计算；
3. M-step：固定E-step中计算出的，最大化似然函数得到新的参数。



依次迭代2）和3）得到收敛的值。

至此，本文最开始的含有隐含变量的参数估计问题就可以通过EM方法解决了，后续有时间再补充案例和Python代码。

参考文献：

1. TODD K. MOON,”The Expectation-Maximization Algorithm” IEEE signal processing magazine,1996
2. Dimitris G. Tzikas, Aristidis C. Likas, and Nikolaos P. Galatsanos,”The Variational Approximation for Bayesian Inference” IEEE signal processing magazine,2008