EM 算法是炼金术吗？

鲁晨光

人工智能很火， 人工智能大神很火。大神们的神器是什么？ 有人说找到了， 就是 EM 算法。 请看这篇：

EM 算法的九层境界：Hinton 和 Jordan 理解的 EM 算法
http://mp.weixin.qq.com/s/NbM4sY93kaG5qshzgZzZIQ

但是最近网上引人关注的另一传闻是，一位人工智能论文获奖者在获奖感言中说深度学习方法是炼金术， 从而引起大神家族成员反驳。 报道见：
http://baijiahao.baidu.com/s?id=1586237001216079684&wfr=spider&for=pc

看到上面两篇， 使我想到：EM 算法是炼金术码？

我近两年碰巧在研究用以改进 EM 算法的新算法：http://survivor99.com/lcg/CM/Recent.html ，对 EM 算法存在的问题比较清楚。我的初步结论是：EM 算法虽然在理论上有问题， 但是确实炼出金子了。 反过来也可以说，虽然 EM 算法练出金子了，但是收敛很不可靠，流行的解释 EM 算法的收敛理由更是似是而非。有人使之神秘化，使得它是有炼金术之嫌。论据何在？下面我首先以混合模型为例，简单介绍 EM 算法， 并证明流行的 EM 算法收敛证明是错的(没说算法是错的)。

假设 n 个高斯分布函数是：

$$P(X|\theta_j)=K\exp[-(X-c_j)^2/(2d_j^2)],\quad j=1,2,\ldots,n$$

其中 K 是系数，c_j 是中心，d_j 是标准差。假设一个数据分布 $P(X)$是两个高斯分布的混合

$$P(X)=P^*(y_1)P(X|\theta^*_1)+P(y_2)P(X|\theta^*_2)$$

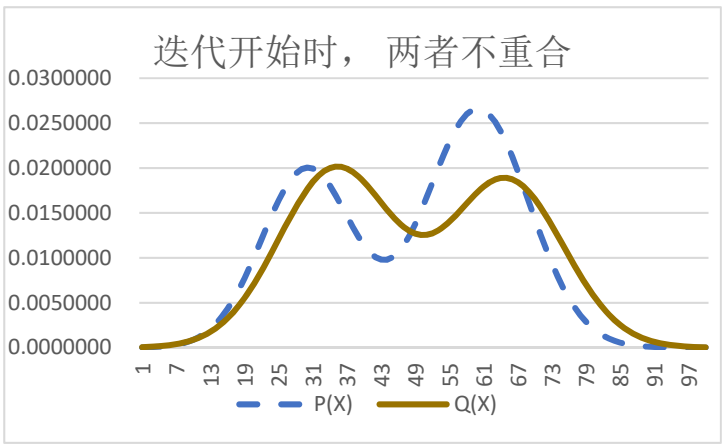
其中 $P^*(y_1), P^*(y_2)$是真的混合比例，θ^*_1 和 θ^*_2 表示真的模型参数。我们只知道模型是高斯分布且 n=2. 现在我们猜测 5 个参数 $P(y_1)$, c_1, c_2, d_1, d_2. 不是 6 个参数，是因为 $p(y_2)$=1-$P(y_1)$. 根据猜测得到的分布是

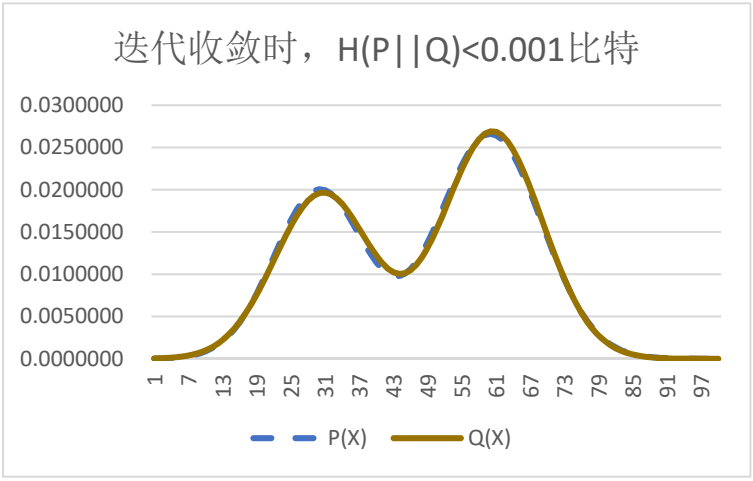
$$Q(X)=P(y_1)P(X|\theta_1)+P(y_2)P(X|\theta_2)$$

如果 $Q(X)$非常接近 $P(X)$, 相对熵或 Kullback-leibler 距离

$$H(Q||P)=\sum_i P(x_i)\log[P(x_i)/P(x_i|\theta)]$$

就接近 0， 比如小于 0.001 比特，那么就算我们猜对了。参看下图：





混合模型问题之所以重要， 是因为它是限制分布类型而不是分布范围的模糊聚类，是无监督学习的一个典型。

EM 算法起源于这篇文章：Dempster, A. P., Laird, N. M., Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B 39, 1–38 (1977). 通过这个网站 http://www.sciencedirect.com/ 搜索可见，光是标题有 EM 算法的英文文章就有 6.8 万篇(有似然度的文章将近 76 万篇)，可见研究和应用 EM 算法的人何其多。

Wiki 百科上的 EM 算法介绍见这里：
https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm
一篇中文介绍见这里：
http://www.cnblogs.com/mindpuzzle/archive/2013/04/05/2998746.html

EM 算法的基本思想是：

目的是改变预测模型参数求似然度 $\log P(X^N|\theta)$或 $\log P(\mathbf{X}|\theta)$达最大($N$ 表示有 N 个样本点，黑体 $\mathbf{X}$ 表示矢量)，样本和 θ 之间的似然度就是负的预测熵(或交叉熵，广义熵的一种)：

$$H_\theta'(X) = \sum_i P(x_i)\log P(x_i|\theta)$$

其中 $P(x_i|\theta)$就是上面的曲线 $Q(X)$上的一个点(X 是变量，x_i 是常量)，即 $P(x_i|\theta)=Q(x_i)$. 我们用 X 的概率分布 $P(X)$取代 X 序列. 则 EM 算法的基本公式如下(下面 y 就是 wiki 百科中的 z)：

$$\begin{aligned}
&\log P(X^N|\theta) = N\sum_i P(x_i)\log P(x_i|\theta) = N\sum_i P(x_i)\log P(x_i|\theta) \\
&\geq L = N\sum_i\sum_j P(x_i)P(y_i|x_i)\log\frac{P(x_i,y_j|\theta)}{P(y_i|x_i)} \\
&= N\sum_i\sum_j P(x_i)P(y_i|x_i)\log P(x_i,y_j|\theta) \\
&-N\sum_i\sum_j P(x_i)P(y_i|x_i)\log P(y_i|x_i) \\
&= Q(\theta|\theta^t) - H
\end{aligned} \tag{1}$$

其中 θ^t 表示 M 步优化前的 θ。这里的 Q 和上面的 $Q(X)$中的 Q 含义不同，下面我

们用 $Q(.)$表示这里的 $Q(\theta|\theta^t)$。

从语义信息论(http://survivor99.com/lcg/books/GIT/)看，我们可以得到

$$[Q(\theta|\theta^t)-H]/N=H_\theta'(X,Y)-H_\theta'(Y|X)= -H_\theta(X,Y)+H_\theta(Y|X)$$

右边是两个负的广义熵或负的交叉熵，和参数有关。为了讨论方便， 后面忽略左边的 N。

EM 算法分两步：

E-step： 写出预测的 y_j 的条件概率

$$P(y_j \mid X) = P(y_j)P(X \mid y_j,\theta) / P(X \mid \theta)$$
$$P(X \mid \theta) = \sum_j P(y_j)P(X \mid y_j,\theta) \qquad (2)$$

M-step: 最大化 $Q(\theta|\theta^t)$, 也就是最大化负的联合熵 $H_\theta'(X,Y)$， 或最小化联合交叉熵 $H_\theta(X,Y)$。

为什么这样能收敛呢？wiki 百科这样说：

1）Expectation-maximization works to improve $Q(\theta|\theta^t)$ rather than directly improving $\log(\mathbf{X}|\theta)$. Here is shown that improvements to the former imply improvements to the latter.[13]

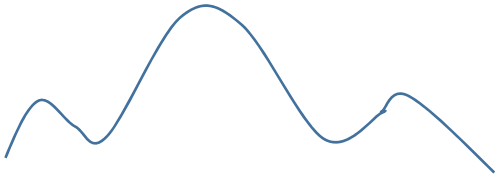
这就是说，只要 $Q(.)$达最大，$\log(\mathbf{X}|\theta)$就达到最大。

2）M-step 可以增大 $Q(.)$, E-step 也不减少 $Q(.)$, 所以反复迭代就能收敛。

这篇证明文章比较出名：Wu, C. F. J.: On the Convergence Properties of the EM Algorithm. Annals of Statistics 11, 95–10 (1983).

这等于说， 真模型在山顶上， 爬山人每一步只上不下就能到达山顶。M 步只上，E 步不下， 所以能到达山顶。

但是，使用 EM 算法时，往往在预测分布 $P(X|\theta)$和实际分布 $P(X)$不重合就停下来了。流行的解释是：那叫局部收敛(其实就是收错了地方);因为大山周围有一些小山，出发点不对就上到小山顶上了。所以起始点很重要。于是有人专门研究如何通过测试找到较好的出发点，比如随机选多点测试看哪个好。



EM 有时虽然能收敛，但是常常收敛很慢。为了提高速度，于是又有很多人提出很多改进办法。其中比较出名的一个就是上面《九层境界》中提到的 VBEM 算法(详见 Neal, R., Hinton, G.: A view of the EM algorithm that justifies incremental, sparse, and other variants. in: Michael I. Jordan (ed.) *Learning in Graphical Models*, *pp 355–368. MIT Press, Cambridge, MA (1999)*)，就是用

$$F(\theta, q)=Q(\theta|\theta^t)+H(Y)$$

取代 $Q(\theta|\theta^t)$(上面忽略了系数 N)，不断最大化 $F(\theta, q)$($q=P(Y)$). 在 M 步最大化 $F(.)$，在 E 步也最大化 $F(.)$. 据说这样收敛更快。但是 VBEM 的收敛证明是不是一样有问题呢？我的结论是：算法好一些，但是收敛证明问题还是存在。

首先我们看 EM 算法(包括 VBEM 算法(的收敛证明错在哪里。

在 Shannon 信息论中有公式：

$$H(X,Y)=H(X)+H(Y|X)$$

由于引进似然函数，Shannon 熵变成预测熵或交叉熵， 但是基本关系还是成立

$$H_\theta(X,Y)=H_\theta(X)+H_\theta(Y|X)= -\log P(\mathbf{X}|\theta)+ H_\theta(Y|X)$$

写成负熵的形式是：

$$H'_\theta(X,Y)=\log P(\mathbf{X}|\theta)+ H'_\theta(Y|X)$$

后面这一项 $H_\theta(Y|X)$和 Y 的熵有关，$P(y_1)=P(y_2)=0.5$ 的时候 $H(Y)$最大，负熵就最小，$H'_\theta(Y|X)$也比较小。如果真的比例 $P^*(Y)$是接近等概率的，起步时 $P(y_1)=0.1$, $P(y_2)=0.9$，Y 的负熵较大，我们不断最大化 $H'_\theta(X,Y)$，就会阻止 $P(Y)$向真比例 $P^*(Y)$靠近。这是 EM 算法收敛慢甚至不收敛的一个原因。这也是为什么 VBEM 方法要用 $F(.)$取代 $Q(.)$. 上式两边加上 $H(Y)$以后就有

$$H'_\theta(X,Y)+H(Y)=\log P(\mathbf{X}|\theta)+ H'_\theta(Y|X)+H(Y)$$

$$H(Y)-H_\theta(X,Y)= -H_\theta(X)+\mathrm{H(Y)}-H_\theta(Y|X)$$

近似得到(后面解释近似)：

$$-H_\theta(X|Y)= -H_\theta(X)+I_\theta(X;Y)$$

也就是

$$F(\theta,q)= - H_\theta(X|Y)= -H_\theta(X)+I_\theta(X;Y) \qquad (3)$$

可见，$F(\theta,q)$就是负的后验熵。两边加上 $P(X)$, 左边就是预测互信息或语义互信息(我早在 1993 年的《广义信息论》一书中就提出)：

$$F(\theta,q)+H(X)=H(X) - H_\theta(X|Y)= I(X;\Theta) \qquad (4)$$

从上面两个公式可以得到

$$H(Q||P)=H(X|\theta)-H(X)= I_\theta(X;Y)- I(X;\Theta) \qquad (5)$$

我们可以粗略理解，相对熵=Shannon 互信息-预测互信息。后面我们将介绍这个公式的更加严格形式。

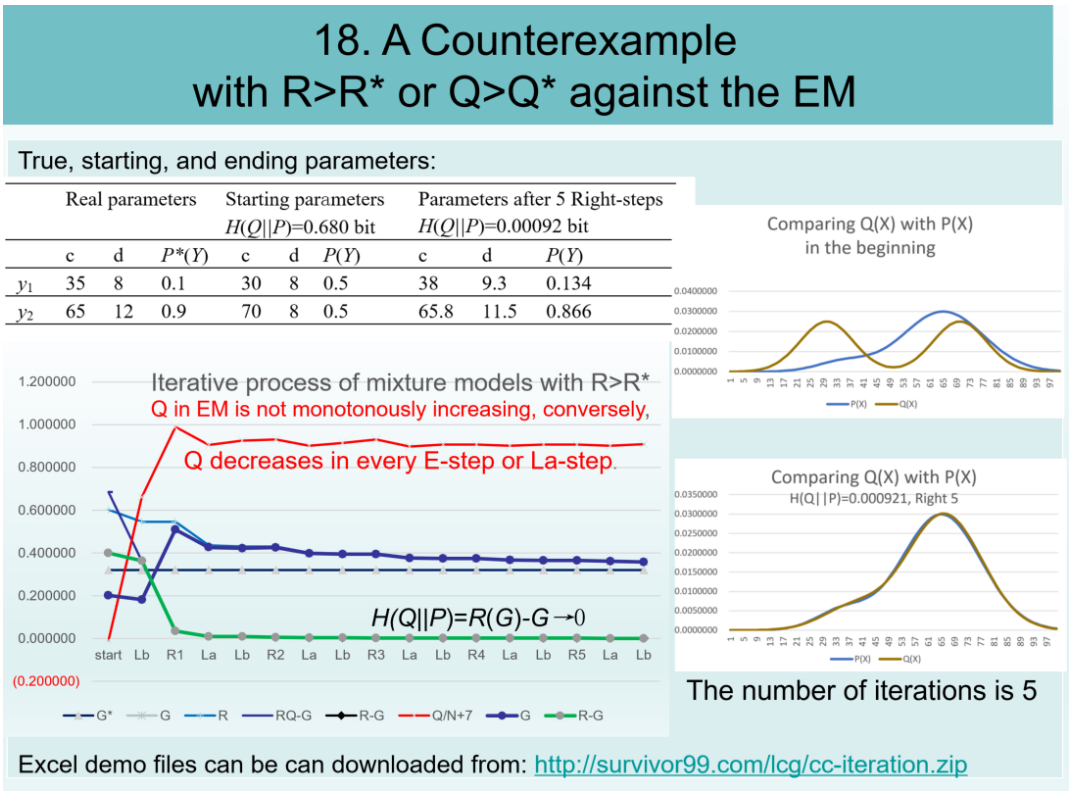
因为 $H(X)$和模型无关，最大化 $F(.)$就是最大化预测互信息， 避免误改 $P(Y)$。这样就好理解为什么 VBEM 比 EM 好。

但是，$F(\theta,q)$最大化和 $\log P(\mathbf{X}|\theta)$最大化是否总是一致的？结论是否定的。证明很简单：

假设有一个真模型 θ^*和子模型比例 $P^*(Y)$， 它们产生 $P(X)$。 同时相应的 Shannon 联合熵是 $H^*(X,Y)$, X 的后验熵是 $H^*(X|Y)$, 互信息是 $I^*(X;Y)$. 那么改变 X 和 Y 的概率分布，上面三个量都会变。

我们猜测的时候， 如果联合概率分布 $P(Y,X|\theta)$比 $H^*(X,Y)$更加集中， 负熵 $\log(\mathbf{X},Y|\theta)=H'_\theta(X,Y)$ 就会大于真模型的负熵$-H^*(X,Y)$， 继续增大 $H'_\theta(X,Y)$就会南辕北辙。

比如， 下图例子中， 第一轮优化参数前就可能有 $H'_\theta(X,Y)> -H^*(X,Y)$。它对于 EM 算法收敛证明来说就是反例。 图中 $R=I(X;Y)$， $R^*=I^*(X;Y)$。 其中真实的 $Q^*(.)$和互信息 $I^*(X;Y)$比较小。



这个例子中迭代用的是信道匹配算法，和 VBEM 算法比，主要差别是，在 E 步把继续增大 $F(.)$改为调整 $P(Y)$。其中红线就是 EM 算法中 $Q(\theta|\theta^t)$的变化轨迹，第一个 E 步之后，$Q(.)$就大于真模型的 $Q^*(.)$. 如果起始参数是随机的，那么它可能导致 $Q(.)$出现在红线的任何位置，从而大于 $Q^*(.)$。

$F(.)$有类似情况，它和预测互信息(图示是 G)走势完全一致， 也是不断下降的。原来影响交叉熵有三个因素：1) 预测的分布和样本的分布是否符合， 如果更符合，负熵 $Q(.)$和 $F(.)$就更大；2) X 和 Y 的分布是否集中，如果更集中负熵就更大； 3) X 和 Y 是否相关，如果更相关负熵就更大。流行的收敛证明只考虑了第一条。

到此有人会问，如果终点不在山顶上，起点很可能高于终点， 那么为什么 EM 算法在大多数情况下是收敛的？

回答是：EM 算法收敛证明中第二条，$Q(.)$只增不减也错了！原来在 E 步，$Q(.)$和 $F(.)$是可能减小的！一般情况下都会使 $Q(.)$向真模型的 $Q^*=-H^*(X,Y)$靠拢，使 $F(.)$向$-H^*(X|Y)$靠拢。如果调整 $P(Y)$, 收敛就更加可靠，EM 算法就变为 CM 算法。

下面提供 CM 算法的粗略数学证明。

先看为什么需要调整 $P(Y)$。在 E 步的公式(2)(计算 Shannon 信道的公式)中，$P(y_j|X)$和 $P(X)$及四个参数可能不匹配， 导致

$$P^{+1}(y_j) \neq \sum_i P(x_j)P(y_j \mid x_i) = \sum_i P(x_i)P(x_i \mid y_j,\theta)P(y_j)/P(x_i \mid \theta)$$

那样，上面计算出的 Shannon 信道就是一个不称职的 Shannon 信道。调整方法很简单， 就是不断用左边的 $P^{+1}(y_j)$代替右边的 $P(y_j)$, 直到两者相等。

下面介绍用信道匹配算法(CM 算法)用到的公式：

$$H(Q\|P)=R_Q-G=R+H(Y\|Y^{+1})-G$$

其中

$$G=I(X;\Theta)=\sum_i\sum_j P(x_i)\frac{P(x_i|\theta_j)}{Q(x_i)}P(y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}$$

$$R_Q=I_Q(X;Y)=\sum_i\sum_j P(x_i)\frac{P(x_i|\theta_j)}{Q(x_i)}P(y_j)\log\frac{P(x_i|\theta_j)}{Q(x_i)}$$

$$R=I(X;Y)=\sum_i\sum_j P(x_i)\frac{P(x_i|\theta_j)}{Q(x_i)}P(y_j)\log\frac{P(y_j|x_i)}{P^{+1}(y_j)}=R_Q-H(Y\|Y^{+1})$$

$$H(Y\|Y^{+1})=\sum_j P^{+1}(y_j)\log[P^{+1}(y_j)/P(y_j)] \quad (6)$$

上面第一行公式就是公式(5)的更加严格形式。其中用到子模型 θ_j, $P(X|\theta_j)$就等于前面方法中的 $P(X|y_j, \theta)$. R_Q 就近似于前面方法中 $I_\theta(X;Y)$. 为什么说近似呢， 因为这里用到 Y 的广义熵或交叉熵 $H(Y^{+1})=-\sum_j P^{+1}(y_j)\log P(y_j)$和相对熵 $H(Y^{+1}||Y)$。联合熵减去这个熵才是预测后验熵。而在 VBEM 方法中， 增加的 $H(Y)$是 Shannon 熵。

有了上面公式(6), 我们就可以采用下面三步减小相对熵 $H(Q||P)$——保证每一步都是减小的：

I：写出 Shannon 信道表达式，等于 EM 算法中的 E 步;

II：调整 $P(Y)$,使得 $P(Y^{+1})=P(Y)$; 如果 $H(Q||P)<0.001$, 结束。

III: 改变参数最大化语义互信息 G. 转到 I.

详细讨论见这里：http://survivor99.com/lcg/CM.html

如果 EM 算法在 M 步先调整 $P(Y)$， 固定 $P(Y)$再优化参数，EM 算法就和 CM 算法相同。如果在 VBEM 算法中添加的 $H(Y)$改为交叉熵，E 步调整 $P(Y)$而不是最大化 $F(.)$, VBEM 算法就和 CM 算法相同。

从语义互信息公式看

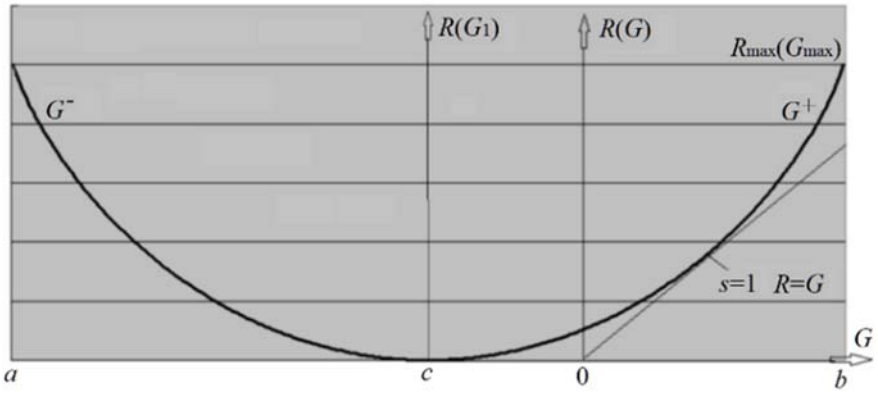
$$G=I(X;\Theta)=\sum_i\sum_j P(x_i)\frac{P(x_i|\theta_j)}{Q(x_i)}P(y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}$$

迭代就是轮流改变 log 左边(I 和 II)和右边(III)。改变右边是语义信道匹配 Shannon 信道， 改变左边是 Shannon 信道匹配语义信道。 所以该算法叫信道匹配(Channels' Matching)算法或 CM 算法(保留 EM 中的 M)。我们也可以说 CM 算法是 EM 算法的改进版本。但是基本思想是不同的。CM 算法也没用到 Jensen 不等式。

为什么 CM 算法会使 $Q(X)$会收敛到 $P(X)$呢？ 相对熵会不会表现为很多山洼， 有高有低， 我们不巧， 落到一个较高的山洼里呢？

这要从 Shannon 的信息率失真理论谈起。 Shannon 在 1948 年发表经典文章《通信的数学理论之》，11 年之后， 他提出信息率失真函数 $R(D)$——就是给定平均损失上限 D 求互信息 $I(X;Y)$最小值 $R(D)$。我在 1993 年的《广义信息论》中把它改造为 $R(G)$函数。G 是广义互信息或语义互信息(也就是平均 log 标准似然度)的下限。$R(G)$是给定 G 时的 Shannon 互信息 $I(X;Y)$的最小值. 可以证明，所有 $R(G)$函数都是碗状的。$R(D)$函数像是半个碗，也是凹的。证明 G 是碗状的

很简单， 因为 *R(D)*函数处处是凹的(已有结论)， *R(G)*函数是它的自然扩展，也是处处是凹的。



进一步结论：*R(G)*-*G* 也是处处是凹的。所以山洼只有一个。求相对熵最小，就是求 *R(G)*-*G* 最小，就是找 *R*=*G* 的点， 也就是上图中 45 度斜线和碗状曲线相切的点。

像 E 步那样，用公式(2)求 Shannon 信道，并调整 *P(Y)*, 就能得到 *R(G)*。原来求信息率失真函数也用到迭代算法，也用到和公式(2)类似的公式。EM 和 VBEM 算法之所以慢，除了因为没有调整 *P(Y)*, 还和指导思想有关。 因为认为增大负熵就能达到目的， 所以初始参数就想把负熵弄得小一点，比如把两个标准差 d_1 和 d_2 设得大一点，防止漏掉真模型。但是在计算试验中我发现，有时候选择较小的偏差， 收敛反而更快。从文献数字和我的计算实验看， CM 算法迭代 5 次收敛比较常见，而 EM 算法(包括改进的)达到收敛的迭代次数大多超过 10 次.

诚然， CM 算法也不是完美无缺的。在少数极端情况下(比如两个分布重叠较多，两个分量比例相差较大时)，初始值选择不当收敛也很慢。结合已有的 EM 算法研究中关于初始参数的研究成果，应该还能改进。

我上面只是证明了流行的 EM 算法收敛证明是错的， 是否还有其他对的证明？ 我不能肯定。北大的马尽文教授是 EM 算法专家，做过很多推广和应用。他说有， 我很期待。我以为如果有，可能和我的证明异途同归。我和 VBEM 的第一作者 Neal 教授通过信，他说要是 E 步减小 Q(.)或 F(.)，那就太震动了。看来他一直相信流行的负熵只增不减证明。

现在回到“炼金术”话题。

实际上，把深度学习和炼金术联系起来的 Ali Rahimi 教授演讲标题是：《从“炼金术”到“电力”的机器学习》，
http://www.chaoqi.net/xinchao/2017/1206/87303.html
他并没有否定深度学习， 只是说要加强理论研究， 也不排除先有实践再有理论。

根据上面分析， 可以说 EM 算法正在从“炼金术”向“冶金术”过度。如过它在理论上停滞不前，如果我们把它神话，它就真像是炼金术了。反之，正视已有问题，寻找更简洁更有说服力理论，才能使“炼金术”成为可靠“冶金”工具。

2017-12-17