18.465, Revised December 6, 2012

THE EM ALGORITHM

The EM algorithm in general was made popular by a paper of Dempster, Laird and Rubin (1977). Special cases of it had been known earlier.

1. The problem

Suppose we have a parametric family given by $f(x, \theta)$ where x is in a sample space χ and θ in a parameter space Θ. One has not observed x, but rather a function $y(x)$, taking values in another sample space $\mathcal{Y}$, where $y(\cdot)$ is many-to-one: for each possible value η of $y(x)$ the set $y^{-1}(\eta) := \{x \in \chi : \ y(x) = \eta\}$ can contain more than one element, and in the continuous case, typically will have dimension ≥ 1. Often, χ may be a set of vectors $(X_1, ..., X_n)$. For each θ, for which x has a likelihood function (probability density or mass function) $f(x, \theta)$, $y = y(x)$ will have its own likelihood function $g(y, \theta)$. The problem is to find a maximum likelihood estimate (MLE) of θ given $Y = y(x)$, in cases where it may be difficult to do that directly, typically more difficult than if we had observed x itself.

2. The algorithm

Given θ and Y, define a conditional expectation of the log likelihood function $\log f(x, \theta')$ by

$$Q(\theta'|\theta) = E(\log f(x, \theta')|Y, \theta). \tag{1}$$

In a recursive procedure, start with some guess θ_0 for the unknown MLE $\widehat{\theta}$ of θ. For $k = 0, 1, 2, ...,$, the "E-step" is to evaluate $Q(\theta'|\theta_k)$ as a function of θ'. Then the "M-step" is to find θ_{k+1} as $\theta_{k+1} = \arg\max_\theta Q(\theta|\theta_k)$. It is proposed that as k becomes large, under some conditions, θ_k will converge to $\widehat{\theta}$, and indeed it does in a lot of cases.

Example. This example is given on the second page of Dempster, Laird and Rubin (1977). Let $X = (x_1, x_2, x_3, x_4, x_5)$ be multinomial $(\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4})$. The full likelihood, if we could observe it, equals factors not depending on θ times

$$\theta^{x_2+x_5}(1-\theta)^{x_3+x_4}.$$

For maximimization with respect to θ we may as well assume that $f(x, \theta)$ equals this function. The likelihood is of binomial form, and

the MLE of θ given X would be $\frac{x_2+x_5}{x_2+x_3+x_4+x_5}$, easily. But suppose we only observe

$$Y = (y_1, y_2, y_3, y_4) = (x_1 + x_2, x_3, x_4, x_5),$$

specifically with values $(125, 18, 20, 34)$, so that $n = 197$. Suppose we start with some θ_0 and after k steps we have θ_k. For the E-step we need $\log f(x, \theta')$ which equals

$$(x_2 + x_5)\log\theta' + (x_3 + x_4)\log(1 - \theta') = (x_2 + 34)\log\theta' + 38\log(1 - \theta').$$

Then

$$Q(\theta'|\theta_k) = E(\log f(x, \theta')|Y, \theta_k) = 38\log(1-\theta')+(\log\theta')[34+E(x_2|Y, \theta_k)].$$

From a binomial expectation,

$$E(x_2|Y, \theta_k) = y_1\left(\frac{\theta_k/4}{(1/2) + (\theta_k/4)}\right) = y_1\left(\frac{\theta_k}{2 + \theta_k}\right),$$

and therefore

$$\begin{aligned} Q(\theta'|\theta_k) &= 38\log(1 - \theta') + (\log\theta')\left[34 + \frac{125\theta_k}{2 + \theta_k}\right] \\ &= 38\log(1 - \theta') + (\log\theta')\cdot\left[\frac{68 + 159\theta_k}{2 + \theta_k}\right]. \end{aligned}$$

That completes the E-step. For the M-step, θ_{k+1} is the value of θ' maximizing the last expression, which again is equivalent to maximizing a likelihood of binomial form, and gives

$$\theta_{k+1} = \frac{68 + 159\theta_k}{144 + 197\theta_k}.$$

This is a continuous function of θ_k. If θ_k converges to some θ_∞ as $k \to \infty$, we will have

$$\theta_\infty = \frac{68 + 159\theta_\infty}{144 + 197\theta_\infty}.$$

This gives a quadratic equation in θ_∞ having two roots, one being negative, the other being 0.6268215, so this is the MLE of θ given Y by the EM method, as $0 \le \theta \le 1$ by assumption.

One can also find the MLE given Y directly. The likelihood function in terms of $Y = (y_1, y_2, y_3, y_4)$, which is multinomial

$$\left(\frac{1}{2} + \frac{\theta}{4}, \frac{1 - \theta}{4}, \frac{1 - \theta}{4}, \frac{\theta}{4}\right),$$

equals factors not depending on θ times

$$(2 + \theta)^{y_1}(1 - \theta)^{y_2+y_3}\theta^{y_4},$$

and looking for critical points of the log likelihood gives the same quadratic equation. It is not claimed that in this case the EM algorithm is easier, rather the example illustrates how the EM method works, and we do see that it gives the right answer, giving us at least hope that the EM algorithm can work even in cases where we have no so such direct way of finding an MLE given Y. But in general, of course, one has to anticipate doing iterations $\theta_0 \to \theta_1 \to \theta_2 \to \cdots$ and not being able to short-circuit them as in the example.

3. Properties of the EM procedure

Dempster, Laird, and Rubin showed that the values of the log likelihood $L(\theta_k)$ are non-decreasing in k. However, for the proof by Dempster, Laird and Rubin (1977) of convergence of the sequence $\{\theta_k\}$ to at least a local maximum of $L(\theta)$ given Y under some conditions, the conditions given were not sufficient. In fact G. D. Murray (1977), in the discussion published following the Dempster et al. paper, pointed out that θ_k could converge to a saddle point of $L(\cdot)$. C. F. J. Wu, (1983, Theorem 3) gave sufficient conditions under which $L(\theta_k)$ converge to $L(\theta^*)$ for some local maximum θ^* of L. (Wu, as far as I see, did not state that θ_k converged to such a θ^*, which might in general not be unique.) Wu's assumptions regarding continuity, differentiability etc. of $L(\cdot)$ are given in his (5), (6), (7), (10), and (11).

4. Methods

If the expectation in (1) cannot be done in closed form, one may approximate it via Markov chain Monte Carlo. Similarly, if the M-step doesn't have a closed form solution, one could approach it by simulated annealing.

As the main problem itself is a maximization, if $g(Y, \theta)$ can be computed reasonably for a fixed Y as a function of θ, then one could try simulated annealing directly instead of the EM algorithm. In some studies mentioned and given by Ingrassia (1992), solution of the entire problem via simulated annealing was slower, but sometimes preferable, to a solution via the EM algorithm.

REFERENCES

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc.* **39**, 1–38, followed by a Discussion by commentators. Had been cited over 31,000 times according to Google Scholar some time in November 2012.

Ingrassia, S. (1992). A comparison between the simulated annealing and the EM algorithms in normal mixture decompositions. *Statistics and Computing* **2**, 203–211. (20 citations per Google Scholar.)

Wu, C. F. Jeff (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11**, 95–103. Cited 1879 times per Google Scholar, November 2012.