Hacettepe Journal of Mathematics and Statistics
Volume 43 (5) (2014), 885 – 890

Estimating the parameters of hidden binomial trials by the EM algorithm

Degang Zhu*

Received 02 : 09 : 2013 : Accepted 02 : 12 : 2013

Abstract

The EM algorithm has become a popular efficient iterative procedure to compute the maximum likelihood (ML) estimate in the presence of incomplete data. Each iteration of the EM algorithm involves two steps called expectation step (E-step) and maximization step (M-step). Complexity of statistical model usually makes the iteration of maximization step difficult. An identity on which the derivation of the EM algorithm is based is presented. It is showed that deriving iteration formula of parameter of hidden binomial trials based on the identity is much simpler than that in common M-step.

2000 AMS Classification: 62F10,62-07,62P10.

Keywords: Maximum likelihood; EM algorithm; incomplete data; hidden binomial trials; identity.

1. Introduction

The EM (Expectation-Maximization) algorithm has become a popular tool in statistical estimation problems involving incomplete data, or in problems which can be posed in a similar form, such as mixture estimation [7, 6]. This idea has been around for a long time. However, the seminal reference that formalized EM and provided a proof of convergence is the paper by Dempster, Laird, and Rubin [1]. Although, some special cases of the EM algorithm were investigated long time ago [2], it was proposed formally by [1]. At present, there are a number of popular and very useful references about the EM algorithm [6, 10, 8, 9, 5]. In this paper, we first introduce an identity in the EM algorithm and then show that derivations of the iterative formula for the parameters estimation in hidden binomial trials based on this identity are simpler and easier than that in the common procedure.

*School of Statistics and Management Shanghai University of Finance and Economics, Shanghai 200433, China,
Department of Applied Mathematics Nanjing Forestry University, Nanjing 210037, China,
Email: dgzhu@njfu.edu.cn

First we give a brief introduction to the EM algorithm. Suppose Y are the unobserved complete data in some statistical trial, Y_{obs} are the observed incomplete data, Y_{mis} are the missing data. So we can denote $Y = (Y_{obs}, Y_{mis})$. A general idea is to choose Y so that maximum likelihood becomes trivial for the complete data. Suppose Y has the probability density $f(Y|\theta)$, and θ is the parameter vector. As mentioned above, each iteration of the EM algorithm involves two steps: E-step and M-step. In the E-step, we calculate the conditional expectation

$$Q(\theta|\theta^{(t)}) = E[\log f(Y|\theta)|Y_{obs}, \theta^{(t)}].$$

where, $\theta^{(t)}$ denote the current estimate for θ after the tth iteration. In the M-step, we maximize the $Q(\theta|\theta^{(t)})$ with respect to θ, denote the update value as $\theta^{(t+1)}$, that is $\theta^{(t+1)} = \arg\max_\theta Q(\theta|\theta^{(t)})$ be the new parameter estimate, which commonly gives an iterative formula. Repeat the two steps until convergence occurs. Comprehensive discussions about the convergence of the EM algorithm can be found in [6, 11]. The key idea of the EM algorithm is to obtain the point series $\{\theta^{(t)}\}$ which increase the log-likelihood of the observed data and then view the limit value of the series $\{\theta^{(t)}\}$ as the estimate of the parameters. It is particularly suited to missing data problems, as the very fact that there are missing data can sometimes make calculations cumbersome. It seems that the two steps are easy and clear. However, the M-step sometimes can not be handled easily. So it is desirable to find some procedures to make the M-step easier. We introduce an identity first.

2. Methods

2.1. Proposition [4] Suppose the complete data Y possesses a probability density $f(Y|\theta)$, where θ is the unknown parameter vector. Sometimes Y are unobserved and can be denoted by $Y = (Y_{obs}, Y_{mis})$, where Y_{obs} are observed incomplete data, and Y_{mis} are missing data. The EM algorithm for estimating the parameter θ can be formulated based on the following identity

$$\frac{\partial}{\partial\theta} Q(\theta|\theta^{(t)}) \mid_{\theta=\theta^{(t)}} \;=\; \frac{\partial}{\partial\theta} l(\theta|Y_{obs}) \mid_{\theta=\theta^{(t)}} . \tag{2.1}$$

where $\theta^{(t)}$ is a any interior point in the parameter space. $l(\theta|Y_{obs})$ is the log-likelihood of the observed data.

Proof. Since $Y = (Y_{obs}, Y_{mis})$, we have

$$\begin{aligned} f(Y|\theta) &= f(Y_{obs}, Y_{mis}|\theta) \\ &= f(Y_{obs}|\theta) f(Y_{mis}|Y_{obs}, \theta), \end{aligned} \tag{2.2}$$

Taking the logarithm on both sides of (2.2) and making arrangement, we get

$$l(\theta|Y_{obs}) = l(\theta|Y) - \log f(Y_{mis}|Y_{obs}, \theta), \tag{2.3}$$

Taking conditional expectation for given Y_{obs} and $\theta^{(t)}$ on both sides, we thus obtain the log-likelihood of the observed data as

$$l(\theta|Y_{obs}) = Q(\theta|\theta^{(t)}) - q(\theta|\theta^{(t)}), \tag{2.4}$$

where

$$Q(\theta|\theta^{(t)}) = E[l(\theta|Y)|Y_{obs}, \theta^{(t)}],$$
$$q(\theta|\theta^{(t)}) = E[\log f(Y_{mis}|Y_{obs}, \theta)|Y_{obs}, \theta^{(t)}].$$

According to the Entropy inequality [4], we know that $q(\theta|\theta^{(t)})$ attains its maximum when $\theta = \theta^{(t)}$ under certain regularity conditions. Thus, the first-order derivative of $q(\theta|\theta^{(t)})$ takes value 0 at $\theta = \theta^{(t)}$. Taking derivative with respect to θ on both sides of (2.4), we have the following identity

$$\begin{aligned} \frac{\partial}{\partial\theta} l(\theta|Y_{obs}) \mid_{\theta=\theta^{(t)}} &= \frac{\partial}{\partial\theta} Q(\theta|\theta^{(t)}) \mid_{\theta=\theta^{(t)}} + \ 0 \\ &= \frac{\partial}{\partial\theta} Q(\theta|\theta^{(t)}) \mid_{\theta=\theta^{(t)}} . \end{aligned}$$

proving the Proposition 2.1.

The value of Proposition 2.1 is that one can replace the unobserved complete data likelihood with the observed incomplete data likelihood so that the maximization process is simplified (as compared with a direct maximization of $Q(\theta|\theta^{(t)})$). Following this idea, a similar identity aimed to simplify the M-step in some special cases can be formulated as follows.

2.2. Proposition [4] Consider a statistical model involving N binomial trials with success probability θ. Suppose S trials result in success. Let Y_{obs} denote the observed incomplete data. Binomial experiments usually involve constant number of trials. Now the number N is allowed to vary for more generality. Then in the implementation of the EM algorithm, we have the following identity

$$\theta^{(t+1)} = \frac{E[S|Y_{obs}, \theta^{(t)}]}{E[N|Y_{obs}, \theta^{(t)}]}. \tag{2.5}$$

Proof. Obviously, the complete data likelihood can be expressed as

$$\frac{N!}{S!(N-S)!}\theta^S(1-\theta)^{N-S} = k\theta^S(1-\theta)^{N-S}$$

where $k = \frac{N!}{S!(N-S)!}$ is the irrelevant constant. The E-step is to calculate the expected log-likelihood. Specially, it is to compute

$$Q(\theta|\theta^{(t)}) = \log(k) + \log(\theta) \cdot E[S|Y_{obs}, \theta^{(t)}] + \log(1-\theta) \cdot E[N-S|Y_{obs}, \theta^{(t)}] \tag{2.6}$$

The derivative with respect to θ, is

$$\frac{\partial}{\partial\theta} Q(\theta|\theta^{(t)}) = \frac{1}{\theta} E[S|Y_{obs}, \theta^{(t)}] - \frac{1}{1-\theta} E[N-S|Y_{obs}, \theta^{(t)}] \tag{2.7}$$

Setting the derivative equal to 0 and solving yields the iterative formula (2.5), which completes the proof of Proposition 2.2.

In order to examine the relationship between $\theta^{(t+1)}$ and $\theta^{(t)}$, we use Proposition 2.1 to give another specific identity. Specifically, from (2.7) and (2.1), we get

$$l'(\theta^{(t)}|Y_{obs}) = \frac{1}{\theta^{(t)}} E[S|Y_{obs}, \theta^{(t)}] - \frac{1}{1-\theta^{(t)}} E[N-S|Y_{obs}, \theta^{(t)}] \tag{2.8}$$

Multiplying (2.8) by $\frac{\theta^{(t)}(1-\theta^{(t)})}{E[N-S|Y_{obs},\theta^{(t)}]}$, we have

$$\begin{aligned}
& \frac{\theta^{(t)}(1-\theta^{(t)})}{E[N-S|Y_{obs},\theta^{(t)}]} l'(\theta^{(t)}|Y_{obs}) \\
= & \frac{(1-\theta^{(t)})E[S|Y_{obs},\theta^{(t)}] - \theta^{(t)}E[N-S|Y_{obs},\theta^{(t)}]}{E[N-S|Y_{obs},\theta^{(t)}]} \\
= & \frac{E[S|Y_{obs},\theta^{(t)}] - \theta^{(t)}E[N|Y_{obs},\theta^{(t)}]}{E[N|Y_{obs},\theta^{(t)}]} \\
= & \frac{E[S|Y_{obs},\theta^{(t)}]}{E[N|Y_{obs},\theta^{(t)}]} - \theta^{(t)} \\
= & \theta^{(t+1)} - \theta^{(t)}
\end{aligned}$$

where the last equality is true because of (2.5). Hence we obtain another identity

$$\theta^{(t+1)} = \theta^{(t)} + \frac{\theta^{(t)}(1-\theta^{(t)})}{E[N-S|Y_{obs},\theta^{(t)}]} l'(\theta^{(t)}|Y_{obs}). \tag{2.9}$$

Starting with some initial value of the parameters $\theta^{(0)}$, one cycles between the E and M-steps until $\theta^{(t)}$ converges to a local maxima. As far as the problem of choosing initial values is concerned, many methods have been proposed for the reason that the choice of initial values can heavily influence the speed of convergence of the algorithm and its ability to locate the global maximum [3]. We argue that the concrete problem need a concrete analysis. A good way to construct initial guesses for the unknown parameters depends on the special cases.

3. Results

In this section we make a application of Proposition 2.2 to the twins pairs research in statistical genetics [4]. Consider random sampling from a twin pairs population, let y be the number of female pairs, x be the number of male pairs, z be the number of opposite sex pairs. Suppose that monozygotic twins occur with probability p and dizygotic twins with probability $1-p$. Suppose in monozygotic twins, boys were born with probability q and girls with probability $1-q$. Suppose there are x_1 monozygotic twins and x_2 dizygotic twins in x male pairs ($x_1 + x_2 = x$), there are y_1 monozygotic twins and y_2 dizygotic twins in y female pairs($y_1 + y_2 = y$) . Because the data x_1, x_2, y_1, y_2 are unobserved, we call this problem hidden binomial trials. The goal is to estimate the parameter $\theta = (p, q)$. Next, we propose two different ways to get the iterate formula for estimating the parameter.

3.1. Theorem The iterate formulae for estimating p and q can be expressed by the following two identities

$$p^{(t+1)} = \frac{x^{(t)} + y^{(t)}}{x + y + z}. \tag{3.1}$$

$$q^{(t+1)} = \frac{x^{(t)} + 2(x - x^{(t)}) + z}{x^{(t)} + 2(x - x^{(t)}) + y^{(t)} + 2(y - y^{(t)}) + 2z}. \tag{3.2}$$

where

$$\begin{aligned} x^{(t)} &= E(x_1|x, p^{(t)}, q^{(t)}) = \frac{xp^{(t)}q^{(t)}}{p^{(t)}q^{(t)} + (1-p^{(t)})(q^{(t)})^2}, \\ y^{(t)} &= E(y_1|y, p^{(t)}, q^{(t)}) = \frac{yp^{(t)}(1-q^{(t)})}{p^{(t)}(1-q^{(t)}) + (1-p^{(t)})(1-q^{(t)})^2}. \end{aligned} \tag{3.3}$$

We shall give two proofs for the Theorem 3.1. Proof 1 is based on the common E-step and M-step, while Proof 2 is based on the proposition 2.2.

Proof 1. According to the genetics laws, the probability of x male pairs with x_1 monozygotic twins and x_2 dizygotic twins is

$$l_1(p, q|x_1, x_2) = \frac{x!}{x_1!x_2!}(pq)^{x_1}[(1-p)q^2]^{x_2}. \tag{3.4}$$

Similarly, the probability of y male pairs with y_1 monozygotic twins and y_2 dizygotic twins is

$$l_2(p, q|y_1, y_2) = \frac{y!}{y_1!y_2!}[p(1-q)]^{y_1}[(1-p)(1-q)^2]^{y_2}. \tag{3.5}$$

The probability for z opposite sex pairs is

$$l_3(p, q|z) = [(1-p)2q(1-q)]^z. \tag{3.6}$$

Thus we can get the likelihood of the complete data

$$L(p, q|x_1, x_2, y_1, y_2, z) = \frac{(x+y+z)!}{x!y!z!} l_1(p, q|x_1, x_2) l_2(p, q|y_1, y_2) l_3(p, q|z). \tag{3.7}$$

Hence,

$$\begin{aligned} l &= \log L(p, q|x_1, x_2, y_1, y_2, z) \\ &= \log \frac{(x+y+z)!}{x!y!z!} + x_1[\log p + \log q] + x_2[\log(1-p) + \log(q^2)] \\ &\quad + y_1[\log p + \log(1-q)] + y_2[\log(1-p) + \log(1-q)^2] \\ &\quad + z[\log(1-p) + \log 2q(1-q)]. \end{aligned} \tag{3.8}$$

Applying the common E-step and M-step [1] to the complete data log-likelihood (3.8), we shall get the iterative formula (3.1) and (3.2).

Proof 2. Since p denotes the probability of monozygotic twins, there are $x_1 + y_1$ monozygotic twins in the total $x + y + z$ twins pairs, we can argue that $N = x + y + z$, $S = x_1 + y_1$, $Y_{obs} = (x, y)$ in (2.5). So we get

$$\begin{aligned} p^{(t+1)} &= \frac{E[S|Y_{obs}, \theta^{(t)}]}{E[N|Y_{obs}, \theta^{(t)}]} \\ &= \frac{E[x_1 + y_1|Y_{obs}, \theta^{(t)}]}{E[x+y+z|Y_{obs}, \theta^{(t)}]} \\ &= \frac{E(x_1|x, p^{(t)}, q^{(t)}) + E(y_1|y, p^{(t)}, q^{(t)})}{x+y+z}. \end{aligned} \tag{3.9}$$

While by the properties of the binomial distribution, (3.3) is obvious given the data and current estimate x, y, $p^{(t)}$, $q^{(t)}$. Thus, (3.9) is just (3.1).

Similarly, since q denotes the probability of boys being chosen in monozygotic twins pairs, and there are x_1 male monozygotic twins pairs, $x_2 = x - x_1$ male dizygotic twins pairs, z opposite sex pairs, so totally the number of boys are $x_1+2(x-x_1)+z$, the number of girls are $y_1+2(y-y_1)+z$, the total number is $x_1+2(x-x_1)+z+y_1+2(y-y_1)+z = x_1+2(x-x_1)+y_1+2(y-y_1)+2z$. We can argue that $N = x_1+2(x-x_1)+y_1+2(y-y_1)+2z$, $S = x_1+2(x-x_1)+z$, $Y_{obs} = (x, y, z)$ in (2.5). Then we shall have

$$\begin{aligned} q^{(t+1)} &= \frac{E[S|Y_{obs}, \theta^{(t)}]}{E[N|Y_{obs}, \theta^{(t)}]} \\ &= \frac{E[x_1+2(x-x_1)+z|Y_{obs}, p^{(t)}, q^{(t)}]}{E[x_1+2(x-x_1)+y_1+2(y-y_1)+2z|Y_{obs}, p^{(t)}, q^{(t)}]} \qquad (3.10) \end{aligned}$$

Again according to (3.3), we find that (3.10) is just (3.2).

4. Conclusion and Discussion

According to the two proofs for the Theorem 3.1 shown above, we can draw a conclusion that deriving iteration formula of parameters of hidden binomial trials based on Proposition 2.2 is simpler than that in the common M-step. Whether the iteration formula in Proposition 2.2 can be extended to other expositional families such as hidden Poisson or exponential trials requires further research.

Acknowledgments. The author gratefully acknowledges the anonymous referees and the editor for their valuable comments which substantially improved the paper. This research was supported in part by the Joint PhD Student Program of Shanghai University of Finance and Economics.

References

[1] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[2] HO Hartley. Maximum likelihood estimation from incomplete data. *Biometrics*, 14(2):174–194, 1958.

[3] Dimitris Karlis and Evdokia Xekalaki. Choosing initial values for the em algorithm for finite mixtures. *Computational Statistics & Data Analysis*, 41(3):577–590, 2003.

[4] Kenneth Lange. *Mathematical and statistical methods for genetic analysis*. Springer, 2002.

[5] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 539. Wiley New York, 1987.

[6] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[7] Geoffrey McLachlan and David Peel. *Finite mixture models*. Wiley. com, 2004.

[8] Xiao-Li Meng and Donald B Rubin. Using em to obtain asymptotic variance-covariance matrices: The sem algorithm. *Journal of the American Statistical Association*, 86(416):899–909, 1991.

[9] Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239, 1984.

[10] Martin A Tanner. *Tools for statistical inference: observed data and data augmentation methods*. Springer-Verlag New York, 1991.

[11] CF Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.