Statistical Methods in Medical Research 1997; **6**: 3–23

The EM algorithm and medical studies: a historical link

Xiao-Li Meng Department of Statistics, University of Chicago, Illinois, USA

Anderson Gray McKendrick's 1926 paper, 'Applications of mathematics to medical problems', was the earliest reference cited in Dempster *et al.*'s 1977 paper that defined and popularized the EM algorithm. McKendrick's paper was prominently featured by Joseph Oscar Irwin in his 1962 inaugural address as the President of the Royal Statistical Society (in the UK), entitled 'The place of mathematics in medical and biological statistics'. The link of McKendrick's work to the EM algorithm is due to an improvement made by Irwin on a novel method McKendrick used for estimating an infection rate when the observed data do not distinguish between those individuals who are not susceptible to the infection and those who are susceptible, but do not develop symptoms. This article examines this link, along the way illustrating the central ideas underlying the EM algorithm as well as its properties; the examination also suggests a profiling strategy for speeding up EM, which may be worthy of general investigation. McKendrick's data on an epidemic of cholera are used for illustration and to compare EM with Irwin's method as well as the Newton–Raphson algorithm. Issues beyond computation are also discussed whenever appropriate.

1 McKendrick's and Irwin's approaches to an epidemic study

On 15 January 1926, AG McKendrick of the Royal College of Physicians at Edinburgh presented a paper[1] to the Edinburgh Mathematical Society which began:

> In the majority of the processes with which one is concerned in the study of the medical sciences, one has to deal with assemblages of individuals, be they living or be they dead, which become affected according to some characteristic. They may meet and exchange ideas, the meeting may result in the transference of some infectious disease, and so forth. The life of each individual consists of a train of such incidents, one following the other. From another point of view each member of the human community consists of an assemblage of cells. These cells react and interact amongst each other, and each individual lives a life which may be again considered as a succession of events, one following the other. If one thinks these individuals, be they human beings or be they cells, as moving in all sorts of dimensions, reversibly or irreversibly, continuously or discontinuously, by unit stages or *per saltum*, then the method of their movement becomes a study in kinetics, and can be approached by the methods ordinarily adopted of such systems.
>
> It is the objective of this communication to approach this field in a systematic manner, to find solutions for some of the variations which may arise, and to illustrate certain of these by examples.

McKendrick first considered the simplest movement of an infection process (i.e. one-dimensional and irreversible), for which he derived (by solving a differential equation) the negative binomial as a distribution of the number of individuals (e.g. human beings or 'cells') who have experienced x attacks, with the Poisson distribution as a limiting case. He then applied the Poisson model to several data sets, including one on an epidemic of cholera in an Indian village. The data are given in the first two rows of Table 1, where x represents the number of cases an individual house has

Address for correspondence: Professor Xiao-Li Meng, Department of Statistics, University of Chicago, IL 60637, USA. Email: meng@paolu.uchicago.edu

0962-2802(97)SM123RA

Table 1 Data and fitted values for McKendrick's problem

x	0	1	2	3	4	$\geq$5	Total
n_x	168	32	16	6	1	0	223
Simple Poisson fit	151.64	58.48	11.28	1.45	0.00	0.01	223
McKendrick's fit	36.47	33.92	15.78	4.89	1.14	0.25	92.45
MLE fit	33.46	32.53	15.81	5.12	1.25	0.29	88.46

experienced (McKendrick did not define what he meant by 'cases', which can affect the analysis), and n_x is the corresponding observed number of such houses (i.e. houses serve as 'cells' for this data set).

To some, fitting a Poisson model to this data set means to estimate the Poisson mean, λ, by the sample average

$$\hat{\lambda} = \sum_x x n_x / \sum_x n_x = 0.386$$

The fitted counts are then calculated as

$$\left(\sum_x n_x\right) \hat{\lambda}^x \exp(-\hat{\lambda})/x!, \quad x = 0, 1, \ldots$$

which are given in the third row of Table 1. The lack of fit is so clearly evident that we can reject the Poisson model without a formal model checking procedure. Indeed, anyone who has some familiarity with the shape of the Poisson distribution would reject the model before carrying out the arithmetic: there are relatively too many houses with 0 cases (i.e. uninfected) for the Poisson model to be appropriate.

McKendrick, of course, knew this. He used a different fitting procedure which at first might seem to be mysterious. He first calculated $s_1 = \sum_x x n_x = 86$ and $s_2 = \sum_x x^2 n_x = 166$, and

$$\hat{n} = \frac{s_1^2}{s_2 - s_1} = 92.45 \approx 93 \tag{1.1}$$

He then calculated the expected counts as $\hat{n}\tilde{\lambda}^x \exp(-\tilde{\lambda})/x!$, where $\tilde{\lambda} = s_1/\hat{n} = 0.930$. The results are given in the fourth row of Table 1, and provide a good fit to the observed counts for $x \geq 1$. Regarding the large discrepancy between the observed and fitted count for $x = 0$, McKendrick[1] (p. 101) wrote

> This suggests that the disease was probably water borne, that there were a number of wells, and that the inhabitants of 93 out of 223 houses drank from one well which was infected. On further local investigation it was found that there was one particular infected well from which a certain section of the community drank.

This quotation makes it clear that McKendrick attributed the excessive number of uninfected houses to the existence of an *unsusceptible* (or *nonexposed*) population of houses that were 'immune' to the infection, possibly because the wells from which they drank did not carry the disease. Consequently, McKendrick's Poisson model for the *susceptible* population was subject to the problem of the so-called 'zero-class missing': without further information, it is not possible to tell whether an uninfected individual (i.e. $x = 0$) was immune to the infection or was susceptible but did not develop symptoms. This distinction is of crucial importance for inference. For instance, for McKendrick's data the simple Poisson analysis would suggest that a house which uses polluted water has about a 68% $(= \mathrm{e}^{-\hat{\lambda}})$ chance of not being infected (e.g. no individual in the house is affected), while McKendrick's analysis indicates that there is only about a 40% $(= \mathrm{e}^{-\tilde{\lambda}})$ chance of not being infected if exposed. Although McKendrick's analysis is not without problems, it is certainly much more sensible than the simple Poisson analysis, and the difference is substantial in real terms (e.g. as measured by medical resources needed for dealing with an outbreak).

Technically speaking, McKendrick's model is a *zero-truncated* Poisson model, i.e. with n_0 missing. For the cholera data set, the observed zero-class count 168 is not n_0, but rather $n_0 + \tilde{n}_0$, where $\tilde{n}_0$ is the number of unsusceptible houses. (McKendrick did not distinguish between sample and population quantities, an issue that will be discussed in Section 4.) The key idea of McKendrick's approach is to ignore the observed 168, since it tells us little about n_0, and to use the posited Poisson model to infer (i.e. impute) n_0. (Of course if the model is far from adequate, so will be the resulting inference, as with the simple Poisson analysis.) Since $n_{\mathrm{obs}} \equiv \sum_{x \geq 1} n_x$ is known, imputing n_0 is equivalent to imputing $n = n_0 + n_{\mathrm{obs}}$. McKendrick's imputation of n, given by (1.1), is a moment estimator, because under the Poisson model, if one has n draws from the Poisson distribution, then, conditional on n, $E(s_1) = n\lambda$ and $E(s_2 - s_1) = n\lambda^2$. As Irwin[2,3] pointed out, what makes McKendrick's method work is the fact that both s_1 and s_2 are unaffected by the missing n_0. As Irwin[3] noted, the method, with appropriate modification of the moment-equation underlying (1.1), can be applied for fitting other discrete distributions with zero-class missing.

Irwin also made a key improvement on McKendrick's method with the help of iteration. The central feature of McKendrick's approach is to first impute the unknown n and then estimate λ, treating the imputed n as if it were the true sample size. Irwin observed, however, that once an estimate of λ is obtained, it can be used to update the imputation for n_0 (and thus for n) since the expected number of zeros in a sample of size n is $n\mathrm{e}^{-\lambda}$. On the other hand, once an updated imputation for n is obtained, it can be used to further update the estimate of λ. This process can be continued until the improvement (i.e. the change) in the estimates become negligible. Letting $n^{(t)}$ be the imputation of n and $\lambda^{(t)}$ be the estimate of λ at the tth iteration $(t = 0, 1, \ldots)$, Irwin suggested updating $n^{(t)}$ via

$$n^{(t+1)} = n^{(t)}\mathrm{e}^{-\lambda^{(t)}} + n_{\mathrm{obs}} \tag{1.2}$$

and then updating $\lambda^{(t)}$ by

$$\lambda^{(t+1)} = \frac{\sum_x x n_x}{n^{(t+1)}} = \frac{n_{\mathrm{obs}}}{n^{(t+1)}}\, \bar{x}_{\mathrm{obs}} \tag{1.3}$$

where $\bar{x}_{\mathrm{obs}} = \sum_x x n_x / n_{\mathrm{obs}}$ (=1.56 for McKendrick's data). One then repeats these two steps until convergence.

As we shall see in the next section, Irwin's iteration in fact converges to the maximum likelihood estimate (MLE) of λ under the zero-truncated Poisson model. The MLE is $\lambda_{\mathrm{MLE}} = 0.97218$, and the corresponding fitted counts are given in the fifth row of Table 1; the fit is a slight improvement over McKendrick's fit. Furthermore, as Irwin[2] remarked, his iterative fitting scheme is similar to Hartley's[4] general approach for computing MLEs with incomplete data. Hartley's approach is a version of what is now widely known as the EM algorithm, the key structure of which is to iterate between imputing the 'missing data' (e.g. (1.2); the necessity of the quotation marks around 'missing data' will be explained later) and fitting the model parameters (e.g. (1.3)). The next section reviews the general framework for EM and how to derive an EM iterative scheme to compute MLEs with truncated data. In Section 3, we compare four iterative schemes for fitting McKendrick's truncated model to illustrate the advantages and disadvantages of the EM approach. In particular, we note that the cholera data set can be modelled by the so-called binomial/Poisson mixture model, which yields the same analysis as the zero-truncated Poisson model but leads to a different EM implementation.

2 Maximum likelihood estimation via the EM algorithm

To describe the EM algorithm in its general form, let Y_{obs} be the data we observe (e.g. $\{n_x, x \geq 1\}$), Y_{mis} be the missing data (e.g. n_0), and $Y = (Y_{\text{obs}}, Y_{\text{mis}})$ be the *completed data* or *augmented data*. (The notation $Y = (Y_{\text{obs}}, Y_{\text{mis}})$ is convenient but not adequate for some problems since the observed data may be a general function of Y, not necessarily a subset of Y; see the binomial/Poisson mixture example in Section 3.) Suppose we have a model for Y_{obs} (e.g. a zero-truncated Poisson model), denoted by $f(Y_{\text{obs}}|\theta)$. Our goal is to compute the MLE of θ by maximizing the log-likelihood function $\ell(\theta|Y_{\text{obs}}) = \log f(Y_{\text{obs}}|\theta)$. This maximization, however, may not have a closed-form solution. For example, for the zero-truncated Poisson model discussed in Section 1, $\theta = \lambda$ and

$$\ell(\lambda|Y_{\text{obs}}) = n_{\text{obs}}[\bar{x}_{\text{obs}} \log \lambda - \lambda - \log(1 - \mathrm{e}^{-\lambda})] \tag{2.1}$$

where the maximizer is the unique (nonzero) solution of the following equation for λ

$$\lambda = \bar{x}_{\text{obs}}(1 - \mathrm{e}^{-\lambda}) \tag{2.2}$$

However, had n_0 been observed, the MLE of λ would be trivial because the complete-data log-likelihood

$$\ell(\lambda|Y) = \left(\sum_x x n_x\right) \log \lambda - n\lambda = n_{\text{obs}}\bar{x}_{\text{obs}} \log \lambda - (n_0 + n_{\text{obs}})\lambda \tag{2.3}$$

is maximized by the sample average

$$\hat{\lambda} = \frac{\sum_x x n_x}{n} = \frac{n_{\text{obs}}}{n_0 + n_{\text{obs}}}\,\bar{x}_{\text{obs}} \tag{2.4}$$

where n_0 denotes the number of zeros from the Poisson distribution. Of course, we cannot directly use (2.4) to estimate λ since n_0 is unknown. As with Irwin's procedure, the idea is to somehow impute n_0 based on the model and then iterate until (2.4) reaches equilibrium (i.e. the 'fixed-point' solution when we view n_0 as a function of λ

through imputation). The EM algorithm specifies a particular approach for this imputation which leads to a sequence of iterates that converge to the desired solution (under some regularity conditions) in such a manner that each iterate has higher likelihood value than any of its predecessors has.

Specifically, starting from an initial value $\theta^{(0)}$ inside the parameter space Θ, we carry out an expectation (E) step and a maximization (M) step within each iteration of EM. At the $(t+1)$st E-step, we find the conditional expectation of the complete-data log-likelihood function (e.g. (2.3)) given the observed data and the estimate of θ from the tth iteration

$$Q(\theta|\theta^{(t)}) = E[\ell(\theta|Y)|\theta^{(t)}, Y_{\text{obs}}]$$

We then carry out the $(t+1)$st M-step which maximizes $Q(\theta|\theta^{(t)})$ as a function of θ to determine $\theta^{(t+1)}$; that is, we find $\theta^{(t+1)}$ such that

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)}), \quad \text{for all } \theta \in \Theta$$

As a consequence of an information inequality, it can be shown (see Dempster *et al.*[5]) that $\ell(\theta^{(t+1)}|Y_{\text{obs}}) \geq \ell(\theta^{(t)}|Y_{\text{obs}})$ for all $t \geq 0$; that is, EM iterates always improve the estimate in the sense that each iterate is more 'likely' than all its predecessors. It is possible for the iterates to converge to a local mode (in theory, EM-type algorithms can also converge to a so called 'saddle' point, an issue that is typically not of concern in practice; see Dempster *et al.*[5] and the discussion by Murray[6]), and it is always wise to start the algorithm with several very different initial values especially if one does not have a clear idea about the behaviour (e.g. the number of modes) of the likelihood being maximized. The issue here is not much about finding a global mode of the likelihood (the real MLE in its mathematical sense), but rather to learn how many modes the likelihood has and their corresponding likelihood values (as well as the curvature at each mode). In the presence of several nontrivial modes (measured by their likelihood values), it is not adequate and often misleading to conduct our inference based solely on a point estimator, be it a local MLE or a global MLE. This is a key difference between statistical computation and pure numerical optimization; the latter typically concentrates on finding the globally optimal solution.

The EM algorithm is particularly useful when the complete-data model $f(Y|\theta)$ is from an exponential family, in which case $\ell(\theta|Y)$ is a linear function of some complete-data sufficient statistic $S(Y)$ (which may be a vector), and thus the E-step reduces to the calculation of the conditional expectation of $S(Y)$. The M-step then is the same as that for calculating the complete-data MLE except with $S(Y)$, which depends on the unobserved Y_{mis}, replaced by its conditional expectation found in the E-step. For the zero-truncated Poisson model, (2.4) yields the $(t+1)$st M-step if we rewrite it as

$$\lambda^{(t+1)} = \frac{n_{\text{obs}}}{n_0^{(t+1)} + n_{\text{obs}}} \bar{x}_{\text{obs}}, \quad t = 0, 1, \ldots \tag{2.5}$$

where $n_0^{(t+1)}$ is the output of the $(t+1)$st E-step. It is worthwhile emphasizing that in this problem we can directly impute the missing data, n_0, because the complete-data log-likelihood function is linear in $Y_{\text{mis}} = \{n_0\}$, as can be seen from (2.3). In general, one must impute the complete-data sufficient statistics $S(Y)$ (or even the complete-data log-likelihood function itself, in which case EM is generally complicated and thus loses some of its advantages), which can be a nonlinear function of Y_{mis}. Imputing the missing data directly will generally not lead to MLEs and in fact often produces inconsistent estimates. This is a key difference between EM and its various *ad hoc* predecessors and one of the two reasons that the phrase 'missing data' is in quotation marks in the last paragraph of Section 1.

Before we discuss the calculation of $n_0^{(t+1)}$, let us first review a general strategy for constructing an EM algorithm for truncated-data problems. This strategy was proposed in Dempster *et al.*,[5] the seminal paper on EM methodology and applications. A truncated-data problem is typically described by a region A such that values outside A will be truncated (e.g. A consists of all positive integers for the zero-truncated Poisson model), and a model $f(x|\theta)$ for the original untruncated random variable (e.g. a Poisson variable). The likelihood of θ given an independent and identically distributed (i.i.d.) truncated sample $x_1, \ldots, x_n$ is then given by

$$L(\theta|x_1, \ldots, x_n) = \frac{\prod_{i=1}^{n} f(x_i|\theta)}{[\Pr(A|\theta)]^n}$$

which can be difficult to maximize due to the presence of $\Pr(A|\theta)$, the probability of A under $f(x|\theta)$. The EM algorithm deals with this problem by augmenting the observed data $Y_{\text{obs}} = \{x_1, \ldots, x_n\}$ to include the sample values that were truncated out (i.e. the values that are outside A) and the number of such values: $Y_{\text{mis}} = \{x_{n+1}, \ldots, x_{n+m}; m\}$; here m is a random variable, just like the zero-class number n_0 in the zero-truncated Poisson model. Given $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, the likelihood of θ is

$$L(\theta|x_1, \ldots, x_{n+m}; m) = \frac{\prod_{i=1}^{n+m} f(x_i|\theta)}{[\Pr(A|\theta)]^n [1 - \Pr(A|\theta)]^m} P(m|\theta; Y_{\text{obs}}), \tag{2.6}$$

where $P(m|\theta, Y_{\text{obs}})$ is a (conditional) probability function to be determined. In order to make our resulting algorithm simple, which is a main feature of EM, we would like to specify $P(m|\theta, Y_{\text{obs}})$ in such a way that the augmented likelihood $L(\theta|x_1, \ldots, x_{n+m}; m)$ in (2.6) is easy to maximize. An obvious choice is to have $P(m|\theta; Y_{\text{obs}})$ proportional to $[\Pr(A|\theta)]^n [1 - \Pr(A)]^m$, which will make $L(\theta|x_1, \ldots, x_{n+m}; m)$ have the same form as the likelihood from the original untruncated sample $x_1, \ldots, x_{n+m}$, $\prod_{i=1}^{n+m} f(x_i|\theta)$. This is typically easier to maximize than the direct maximization of the likelihood based on the truncated sample. This can be accomplished if we choose $P(m|\theta; Y_{\text{obs}})$ to be a negative binomial distribution:

$$P(m|\theta; Y_{\text{obs}}) = \binom{m + n - 1}{m} [\Pr(A|\theta)]^n [1 - \Pr(A|\theta)]^m$$

Under this distribution

$$E(m|\theta; Y_{\text{obs}}) = \frac{1 - \Pr(A|\theta)}{\Pr(A|\theta)} \, n \tag{2.7}$$

which is quite intuitive since it says that the ratio of the expected size of the sample outside A to the size of the sample inside A is given by the ratio of the corresponding probabilities of these two regions.

With θ replaced by $\theta^{(t)}$, (2.7) yields a part of the E-step when $\ell(\theta|Y) = \log L(\theta|Y)$ is linear in the random size m (which is typically the case in practice). In fact, for the zero-truncated Poisson model, this is the E-step since $m = n_0$ is the only missing value in $\ell(\theta|Y)$ (see (2.3)). For the zero-truncated Poisson model, $\Pr(A|\lambda) = 1 - \mathrm{e}^{-\lambda}$, so that we have

$$n_0^{(t+1)} = E(n_0|\lambda^{(t)}; Y_{\text{obs}}) = \frac{\mathrm{e}^{-\lambda^{(t)}}}{1 - \mathrm{e}^{-\lambda^{(t)}}}\, n_{\text{obs}} \tag{2.8}$$

Combining (2.8) with (2.5) defines the EM iteration for the zero-truncated Poisson model

$$\lambda^{(t+1)} = \bar{x}_{\text{obs}}(1 - \mathrm{e}^{-\lambda^{(t)}}), \quad t = 0, 1, \ldots \tag{2.9}$$

It is easy to show that the iteration defined by (2.9) will converge to the unique (nonzero) solution of (2.2) from any starting value $\lambda^{(0)} > 0$. Indeed, we could have directly defined (2.9) according to (2.2), but knowing this iteration is in fact an EM sequence ensures us that it will always increase the log-likelihood function given in (2.1). The monotonicity contributes to the superior stability of EM (e.g. less sensitive to starting values) over other algorithms such as the well-known Newton–Raphson algorithm. (It also allows us to recognize programming errors by checking the log-likelihood values for the sequence of the iterates.) This and other issues will be discussed in the next section.

3 Comparing various iterative schemes

3.1 Irwin's method and two EM implementations

To compare the EM iteration (2.9) with Irwin's iteration given in (1.2)–(1.3) for their performance with McKendrick's problem, both algorithms were implemented. An iteration was stopped whenever the absolute difference in iterates, $|\lambda^{(t+1)} - \lambda^{(t)}|$, was less than 0.0001. Such a criterion is suitable for the current problem given the magnitude of λ_{MLE}, and provides essentially fair comparisons of the methods. (Whether one uses absolute difference or relative difference in general depends on the nature and interpretation of the parameter. A potential problem with a criterion like $|\lambda^{(t+1)} - \lambda^{(t)}| \leq 0.0001$ is that it could stop an iteration prematurely for a very slow algorithm, a problem that did not happen with our example.) Two initial values were used: McKendrick's estimate $\tilde{\lambda} = 0.93$, which should be a good starting value, and another arbitrarily chosen more distant value, 0.4. With $\lambda^{(0)} = 0.93$, EM took 12 iterations to converge and Irwin's iteration took 24, and with $\lambda^{(0)} = 0.4$, EM took 19 and Irwin's took 25.

To visualize how each iterative scheme approaches its limit, Figure 1 plots the first nine (or fewer) iterates along the log-likelihood curve given by (2.1), where the numbers on the curve correspond to the iteration index, with 0 indexes the initial value. The left column corresponds to $\lambda^{(0)} = 0.93$ and the right column corresponds to

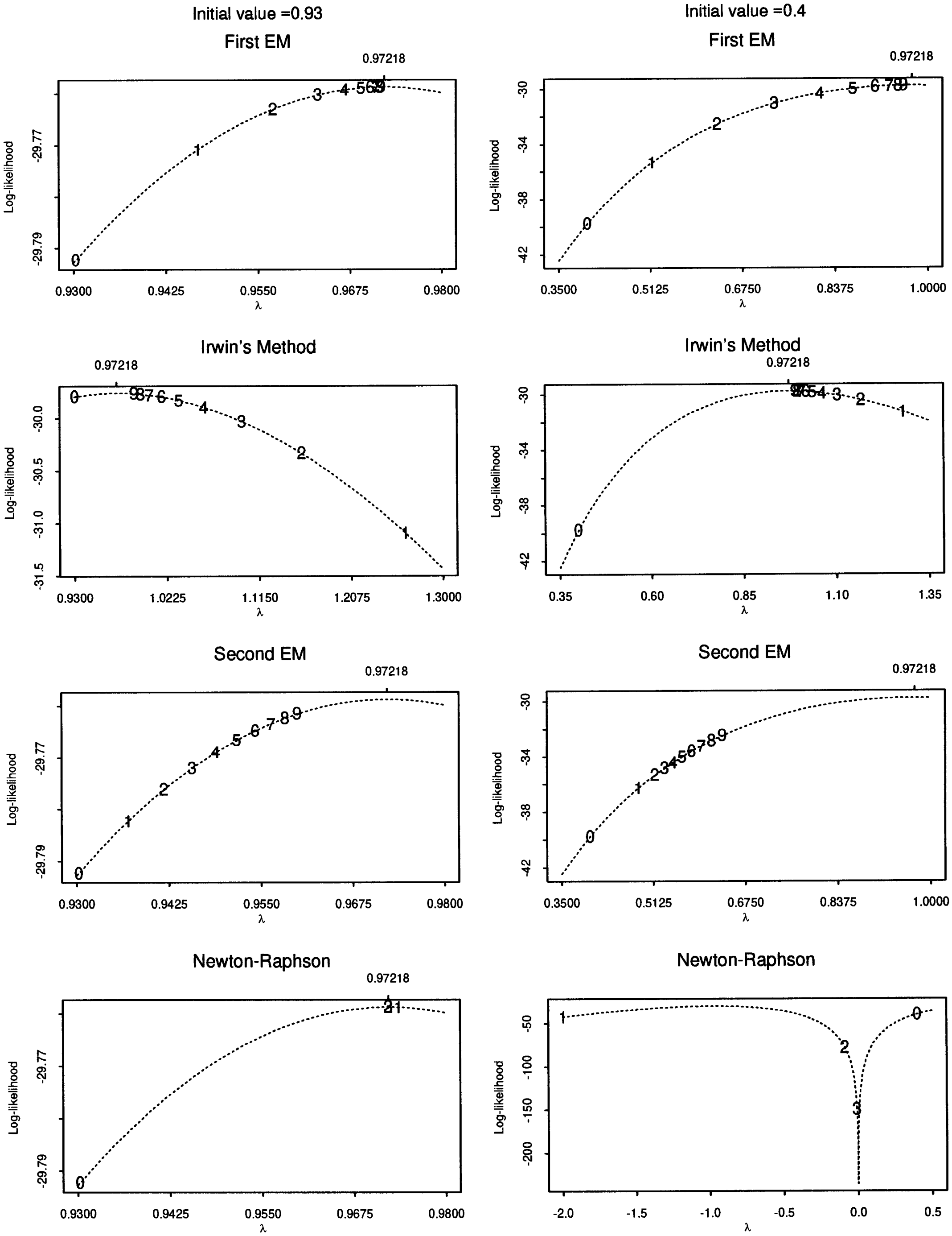

Figure 1 How iterates move along the log-likelihood surface

$\lambda^{(0)} = 0.4$. Each row corresponds to an iterative scheme; the last two will be discussed shortly. For EM (labelled First EM), the iterates climbed the log-likelihood surface monotonically from one side of $\lambda_{\text{MLE}} = 0.97218$, ticked at the top. For Irwin's iteration, the behavior of the iterates are rather different. In both cases, Irwin's method 'overshot', that is, it jumped to the other side of λ_{MLE} before coming back. In the first case (i.e. when $\lambda^{(0)} = 0.93$), it even substantially decreased the log-likelihood before coming back. This cannot happen with any EM iteration because EM cannot decrease the log-likelihood value along its iterates. It is interesting to observe that Irwin's method does not have the monotonicity property, even though it is intrinsically related to EM. In fact, if we change $n^{(t)}$ in the right-hand side of (1.2) to $n^{(t+1)}$, then (1.2) becomes $n^{(t+1)} = n_{\text{obs}}/(1 - \mathrm{e}^{-\lambda^{(t)}})$ and thus Irwin's iteration will be the same as the EM iteration (2.9). In other words, Irwin's iteration is what Green[7] called a 'one-step-late' variation of EM. This illustrates that an EM 'lookalike' can easily be nonmonotone.

When an iterative algorithm is not monotone we know it cannot be an EM algorithm. On other hand, for a particular problem there can be infinitely many EM iterative schemes, most of which are essentially useless as they are difficult or even impossible to implement. The key to obtaining a useful EM implementation is to seek an efficient *data augmentation* scheme, that is, the construction of Y_{mis}, such that both the E-step and M-step are easy to implement and at the same time the amount of augmentation (measured by relative Fisher information) is kept as small as possible. The reason we want to keep the augmentation small is because the speed of convergence of EM is directly governed by the relative Fisher information, or the so-called 'fraction of missing information', one of the key results in Dempster *et al.*[5] To illustrate the impact of different augmentation schemes, let us consider a different EM implementation (in fact, a different model) for McKendrick's problem.

As we discussed in Section 1, McKendrick dealt with the problem of excessive zeros in the cholera data set with the zero-truncated Poisson model. An alternative model, perhaps more direct for many modern statisticians, is the binomial/Poisson mixture model. This model says that a random variable X has probability p of being from a Poisson distribution with parameter $\lambda > 0$, and has probability $1 - p$ of being zero, where $0 \leq p \leq 1$ is a parameter. Mathematically, this means

$$\Pr(X = k) = \begin{cases} 1 - p + p\,\mathrm{e}^{-\lambda} & \text{if } k = 0 \\ p\ \lambda^k \dfrac{\mathrm{e}^{-\lambda}}{k!} & \text{if } k = 1, 2, \ldots \end{cases} \tag{3.1}$$

It is not difficult to speculate that this is what McKendrick had in mind, if we view p as the percentage of houses that used the polluted well or in general, as the probability of being susceptible. With this model, we do not ignore the observed zero-class count; in fact, it is this count that allows us to estimate p.

To fit the binomial/Poisson mixture model, we first write down the log-likelihood for $\theta = (p, \lambda)$ given the data $Y_{\text{obs}} = \{n_0; n_x, x \geq 1\}$, where n_0 is now a part of the observed data. This differs from the notation in Section 2 where n_0 was used to denote zero-class count from the susceptible population only; we will retain the notation $n_{\text{obs}} = \sum_{x \geq 1} n_x$ and $n = n_0 + n_{\text{obs}}$

$$\ell(p, \lambda | Y_{\text{obs}}) = n_0 \log\left(1 - p + pe^{-\lambda}\right) + n_{\text{obs}}[\bar{x}_{\text{obs}} \log \lambda - \lambda + \log p] \tag{3.2}$$

Differentiating this function, we find that for any given $\lambda \geq \log(n/n_0)$, $\ell(p, \lambda|Y_{\rm obs})$ is maximized by

$$p = \hat{p}(\lambda) = \frac{n_{\rm obs}}{n(1 - {\rm e}^{-\lambda})} \tag{3.3}$$

and for given p, $\ell(p, \lambda|Y_{\rm obs})$ is maximized by the λ that satisfies the following equation

$$\lambda = \bar{x}_{\rm obs} \frac{1 - p + p\,{\rm e}^{-\lambda}}{1 - p + (n/n_{\rm obs})p\,{\rm e}^{-\lambda}} \tag{3.4}$$

Substituting (3.3) into (3.4) leads to (2.2), that is, $\lambda_{\rm MLE}$ under the binomial/Poisson mixture model is the same as that from the zero-truncated Poisson as long as $\lambda_{\rm MLE} \geq \log(n/n_0)$ (which is true for McKendrick's data set). This is intuitive and is not a coincidence. In fact the profile log-likelihood for λ, $\ell(\hat{p}(\lambda), \lambda|Y_{\rm obs})$, is identical to the log-likelihood under the zero-truncated Poisson model (2.1), when $\lambda \geq \log\ (n/n_0)$. The restriction $\lambda \geq \log\ (n/n_0)$, or equivalently $n\,{\rm e}^{-\lambda} \leq n_0$, is also intuitive because when n_0 is too small compared to what is expected from a Poisson model (i.e. when $p = 1$), the binomial/Poisson mixture model can be rejected formally. The zero-truncated Poisson model, however, cannot be rejected formally on that ground because n_0 does not enter the model. (Of course, in most applications of the zero-truncated Poisson model, n_0 is not over-reported, but rather unobserved or under-reported, e.g. when susceptible but symptom-free individuals are more likely to refuse to be tested and such untested individuals are excluded from the study, in which case the binomial/Poisson mixture model is not appropriate or relevant.)

Although the binomial/Poisson mixture model provides the same estimate for λ as the zero-truncated Poisson model for McKendrick's data set, it leads to a different EM implementation for computing this estimate. At first glance, one may wonder how one can implement EM for the model in (3.1), because there are no missing data in the usual sense. This highlights a key message conveyed in Dempster *et al.*[5]: EM methodology is much more generally applicable than meets the eye since we can *deliberately* create 'missing' data (i.e. data augmentation). By doing so, we turn a difficult maximization problem into a sequence of easier maximizations. This is the second reason that the phrase 'missing data' is in quotation marks because the 'missing data' used in an EM construction can be a purely computational device which is not even hypothetically observable.

For mixture models in general, we can treat the subpopulation memberships (i.e. the mixture indicator) as the missing data.[5] For the binomial/Poisson mixture model, this means that we can treat the information for distinguishing between the two types of zeros (e.g. which type of well a house used) as the missing data. In particular, we can construct $Y = \{(y_j, z_j),\ j = 1, \ldots, n\}$ as the augmented data, where $\{y_j,\ j = 1, \ldots, n\}$ are i.i.d. samples from the Poisson model with parameter λ, and $\{z_j,\ j = 1, \ldots, n\}$ are i.i.d. Bernoulli trials with success probability p. Here z_j is a subpopulation indicator, that is, the jth individual is from the susceptible population (e.g. the Poisson distribution) if $z_j = 1$, and from the unsusceptible population if $z_j = 0$. We do not fully observe Y. What we observe is $Y_{\rm obs} = \{x_i \equiv y_i z_i,\ i = 1, \ldots, n\}$ – we have $x_i = 0$ (e.g. no symptoms) either when $z_i = 0$ (e.g. unsusceptible) or when $z_i = 1$ but $y_i = 0$ (e.g. susceptible but did not develop symptoms). When y and z are independent, it is easy to see that $x = yz$ follows the binomial/Poisson mixture distribution given in (3.1).

Note that when $z_j = 0$ (and thus $x_j = 0$) we can define y_j arbitrarily. We choose to define y_j the same way regardless of the value of z_j because it results in a very simple augmented-data log-likelihood of $\theta = (p, \lambda)$

$$\ell(p, \lambda|Y) = \left(\sum_{i=1}^{n} y_i\right) \log \lambda - n\lambda + \left(\sum_{i=1}^{n} z_i\right) \log p + \left(n - \sum_{i=1}^{n} z_i\right) \log (1 - p) \qquad (3.5)$$

which is maximized when

$$\lambda = \frac{\sum_{j=1}^{n} y_j}{n} \quad \text{and} \quad p = \frac{\sum_{j=1}^{n} z_j}{n} \qquad (3.6)$$

This defines the M-step of EM. To perform the E-step, we need to compute the conditional expectations of the sufficient statistics from Y, $\sum_{j=1}^{n} y_j$ and $\sum_{j=1}^{n} z_j$, given Y_{obs} and θ. This calculation is straightforward once we observe that if $x_j > 0$ then $y_j = x_j$ and $z_i = 1$, and conditional on $x_j = 0$, z_j is a Bernoulli trial with success probability $p\ \mathrm{e}^{-\lambda}/(1 - p + p\ \mathrm{e}^{-\lambda})$ and y_j has conditional mean $(1 - p)\lambda/(1 - p + p\,\mathrm{e}^{-\lambda})$. Replacing the numerators in (3.6) by their corresponding conditional expectations yields our second EM for calculating λ_{MLE} (which also calculates p_{MLE})

$$\lambda^{(t+1)} = \frac{n_{\text{obs}}}{n} \bar{x}_{\text{obs}} + \frac{n_0}{n} \frac{(1 - p^{(t)})\lambda^{(t)}}{1 - p^{(t)} + p^{(t)}\ \mathrm{e}^{-\lambda^{(t)}}} \qquad (3.7)$$

$$p^{(t+1)} = \frac{n_{\text{obs}}}{n} + \frac{n_0}{n} \frac{p^{(t)}\ \mathrm{e}^{-\lambda^{(t)}}}{1 - p^{(t)} + p^{(t)}\mathrm{e}^{-\lambda^{(t)}}} \qquad (3.8)$$

This second EM took 32 iterations to converge with $\lambda^{(0)} = 0.93$ and 63 iterations with $\lambda^{(0)} = 0.4$, essentially tripling the number of iterations needed by the first EM given in (2.9). The slower convergences can be visualized from the third row of Figure 1, in comparison with the first EM implementation plotted in the first row. (Since the second EM also requires $p^{(0)}$, $p^{(0)}$ was calculated according to $p^{(0)} = \min\{n_{\text{obs}}/(n(1 - \mathrm{e}^{-\lambda^{(0)}})), 0.9\}$, which prevents $p^{(0)} \geq 1$ and makes the comparison approximately fair.) A key reason for this slower convergence is that there is more augmentation for the second EM than for the first EM. (The analytic calculation of the amount of augmentation is omitted since it is a bit too technical; examples of such calculations can be found in Meng and van Dyk.[8])

3.2 The Newton–Raphson algorithm

For a simple log-likelihood like (2.1), the well-known Newton–Raphson algorithm can be very effective for obtaining the maximizer once we have a good idea of the region where the solution lies. (This is not a strong requirement for one- or two-dimensional problems, since we can almost always plot the log-likelihood to be maximized.) The general formula for Newton–Raphson iteration for solving $g(\lambda) = 0$

is given by

$$\lambda^{(t+1)} = \lambda^{(t)} - \frac{g(\lambda^{(t)})}{g'(\lambda^{(t)})}, \quad t = 0, 1, \ldots \tag{3.9}$$

where $g'(\lambda)$ is the derivative of g. In contrast, EM does not require derivative calculations unless such calculations are needed for implementing its M-step. There are many variations of (3.9); see Thisted,[9] chapter 4, in particular, section 4.3.5.1, which discusses the binomial/Poisson mixture. Note that Thisted's ξ is our $1-p$ and his N is our n, and thus his (4.3.14) and (4.3.15) can be simplified to (3.3) and (3.4) after correcting a typographical error in (4.3.15) (the n_0 in (4.3.15) should be $N - n_0$). Note also that the restriction $\lambda_{\text{MLE}} \geq \log(N/n_0)$ is needed for (4.3.14) – when it is not satisfied, $p_{\text{MLE}} = 1$ or $\xi_{\text{MLE}} = 0$.

For the current problem, $g(\lambda) = \lambda - \bar{x}_{\text{obs}}(1 - \mathrm{e}^{-\lambda})$, as in (2.2), $g'(\lambda) = 1 - \bar{x}_{\text{obs}}\,\mathrm{e}^{-\lambda}$, and the Newton–Raphson iteration is thus given by

$$\lambda^{(t+1)} = \frac{(1 - \mathrm{e}^{-\lambda^{(t)}} - \lambda^{(t)}\mathrm{e}^{-\lambda^{(t)}})\bar{x}_{\text{obs}}}{1 - \mathrm{e}^{-\lambda^{(t)}}\bar{x}_{\text{obs}}} \tag{3.10}$$

This iterative scheme converged in four iterations with $\lambda^{(0)} = 0.93$ and in six iterations with $\lambda^{(0)} = 0.4$, much faster than any of the previous iterative schemes. However, when $\lambda^{(0)} = 0.4$, it converged to $\lambda = 0$, which is a solution of (2.2) but corresponds to the minimizer of (2.1). This phenomenon cannot happen with any EM iterations as long as the initial value is inside the parameter space (e.g. $\lambda^{(0)} > 0$ with (2.1)), since it cannot decrease the likelihood.

The right plot in the fourth row of Figure 1 illustrates how the Newton–Raphson iterates converged to the wrong limit. Because all the iterates are outside the parameter space (i.e. $\lambda^{(t)} < 0$ for $t \geq 1$), a mirror image (with respect to $\lambda = 0$) of the log-likelihood surface was created to plot the iterates. In contrast, EM iterates can never escape from the parameter space as long as the initial value is inside the space. It is worthwhile to point out that the (original) log-likelihood surface here is as simple and smooth as it can be, and $\lambda^{(0)} = 0.4$ is not an impossible choice of starting value in practice for such problems. For more complicated problems, especially multi-dimensional ones, the Newton–Raphson algorithm can be very sensitive to the starting value, and sometimes fail to converge (which is less harmful then converging to a wrong limit). Of course, when using (3.10), a careful user will not choose $\lambda^{(0)} \leq \log(\bar{x}_{\text{obs}}) = 0.447$, which makes its denominator negative (or even zero), the reason for negative iterates and convergence to the wrong limit. (For a careful user, (3.10) is not even a Newton–Raphson iteration once iterates move outside the space where the original log-likelihood was defined. In that sense, the Newton–Raphson iteration failed at the first iteration when $\lambda^{(0)} = 0.4$.) In general, however, it may not be a trivial task to detect such a problem before running the Newton–Raphson algorithm. The general point is that the fast convergence of the Newton–Raphson algorithm often comes at the expense of more human investment in terms of delicate choice of the starting value and careful monitoring of convergence.

To summarize, Table 2 gives the number of iterations needed for all four iterative schemes with different choices of $\lambda^{(0)}$, some of which are quite extreme for the purpose

Table 2 Number of iterations needed by the four algorithms

$\lambda^{(0)}$	First EM	Irwin's method	Second EM	Newton–Raphson
100	17	26	70	6
1.56	16	25	49	5
0.93	12	24	32	4
0.4	19	25	63	6*
0.1	23	25	76	5*
0.01	29	26	77	4*

of illustration. (Comparing the number of iterations is informative for such simple iterations; in general, comparisons of the actual computation time are more informative and appropriate.) An asterisk indicates cases where the iterations converged to the wrong limit, $\lambda = 0$, which occurs for Newton–Raphson whenever $\lambda^{(0)} < \log(\bar{x}_{\rm obs}) = 0.447$. Both EM iterations are somewhat sensitive to $\lambda^{(0)}$ in terms of the speed, but not in terms of where they converge to – for a unimodal likelihood EM is guaranteed to converge to the unique mode under quite weak regularity conditions – see Wu.[10] (This type of sensitivity cannot be studied using the standard theoretical rate of convergence, e.g. see Dempster *et al.*[5] and Meng[11], which does not depend on the starting value.) It is interesting to see that Irwin's method is not sensitive to $\lambda^{(0)}$. This is an advantage because it prevents the iteration from being affected by a bad choice of $\lambda^{(0)}$ (e.g. a $\lambda^{(0)}$ that is close to zero). Indeed, when $\lambda^{(0)} = 0$, the EM given by (2.9) will stay on the boundary (i.e. $\lambda^{(t)} = 0$ for all t), a typical consequence of starting an EM iteration from a boundary point (which should be avoided), while Irwin's method will converge properly to $\lambda_{\rm MLE}$. But it is also a disadvantage, as revealed clearly in the left plot in the second row of Figure 1, for the iterates there were forced to move to the far right even if the initial value was already quite close to $\lambda_{\rm MLE}$. This is a computational example of the common tradeoff between 'efficiency' and 'robustness', a central issue in statistical inference.

4 Discussion and bibliographical notes regarding McKendrick and Irwin

The beauty of McKendrick's fitting method is its simplicity. There is no iteration involved, which perhaps was especially important in 1920s. It also worked very well for the cholera data. In fact, McKendrick's estimate of λ has the log-likelihood value -29.7923, very close to the log-likelihood value of the MLE, -29.7587. It does not, however, always work this well. Irwin[2] gave an example where McKendrick's method yielded a negative estimate of n_0, a common problem with moment estimators. Also, it appears that McKendrick did not appreciate the distinction between sample and population quantities and consequently the uncertainty in his estimates. This is evident from his conclusion, based on his fitting given in Table 1, '... that the inhabitants of 93 out of 223 houses drank from one well which was infected'. Even if there was no sampling variability because the whole village was included in the study, which perhaps was true, and the posited model was indeed correct (the mean number of infected cases per house is defined irrespective of the model assumption), there are still uncertainties in the estimates of λ and n_0. From the calculations in Section 3.1, the MLE of the percentage of susceptible houses is $p_{\mathrm{MLE}} = 40\%$, which yields 88 as the MLE of the number susceptible houses with an approximately 95% confidence interval (63, 117), which is fairly wide. (Formulae for all the confidence limits are given in the Appendix.)

The uncertainty in estimating λ is particularly important when we want to use the results from this study to estimate, say, the medical resources needed for a future outbreak in a *similar susceptible* community. Our MLE for the percentage of houses that will be infected is 62% ($= 1 - \mathrm{e}^{-\lambda_{\mathrm{MLE}}}$), but because an (approximate) 95% interval for λ is (0.69, 1.36), the corresponding 95% confidence interval for this percentage is (47%, 73%). Failing to appreciate such a large uncertainty could conceivably lead to serious mis-estimation of the needed medical resources. There are two reasons for this large uncertainty despite the fact that there were 223 houses in the study, which seems to be quite large. First, only the 55 infected houses carried information about λ; this is why the inference for λ from the binomial/Poisson mixture model is the same as that from the truncated Poison model, once $p = 1$ is ruled out. Second, the 55 observations are not from the Poisson distribution, but rather from the zero-truncated Poisson distribution. As the calculation given in the Appendix shows, for McKendrick's problem, a sample of 55 zero-truncated observations (assuming i.i.d) carries approximately the same amount of information about λ as a sample of 36 (i.i.d) observations from the untruncated Poisson model. In other words, if we treat the information about λ in a single observation from the untruncated Poisson distribution as the baseline, then the *effective sample size* from McKendrick's data set for estimating λ is only about 16% of what it appeared to be. Practitioners should be aware of such deceptively large samples and assess the uncertainties in their estimates using appropriate methods.

There are other issues that could be raised with McKendrick's analysis. For example, the number of cases a house can have is limited by the number of its residents, unless one counts repeated infections to the same individual, which is unlikely for a potentially fatal disease like cholera. I surmise that for the village where McKendrick's data were collected, the sizes of the households were relatively homogeneous and large, in which cases the household size may not be an issue. Being aware of such issues, even if we do not act upon them with the data set in hand, helps to prevent us from making misleading general inferences. For example, if the sizes of households are relevant, then an apparent good fit under the posited Poisson model alone may or may not provide convincing evidence of the correctness of the Poisson model, which itself was taken by McKendrick as an indication for no evidence of infection by contagion or by insect. (Cholera is spread by pollution of water supplies, so the statistical evidence in this example was validated by scientific knowledge.)

Despite these potential issues with McKendrick's analysis, the sophistication displayed in McKendrick's paper is rather remarkable considering the paper was written in the 1920s and that he was not a statistician or mathematician by training (the zero-truncated Poisson model is only a small part, and in fact the simplest case considered, in his paper). This was made clear by Irwin, who reproduced essentially all the major results of McKendrick's 1926 paper as an appendix to his 1962 inaugural address as the President of the Royal Statistical Society (UK). In the introductory part of the appendix, Irwin[3] (p. 18) wrote:

> AG McKendrick was in earlier life a lieutenant-colonel in the Indian Medical Service, and later became Curator of the College of Physicians at Edinburgh. Though an amateur, he was a brilliant mathematician, with a far greater insight than many professionals. The joint work of Kermack and McKendrick on epidemiological theory is well known; it was done after the publication of the paper I am about to discuss and the approach was entirely deterministic.
>
> Why this paper and an earlier one (1914) in the *Proceedings of the London Mathematical Society*, which gave the first solution of the general homogeneous birth process, attracted so little notice at that time is something of mystery. It was known to Karl Pearson in late twenties for he once mentioned it to me; it was also known to Greenwood and Yule, but none of them, I think, realized its importance. I gave a reference to its place of publication, which unfortunately was not quite correct, in a discussion on a paper on accidents by Chambers and Yule, published in our *Journal* in 1941. About seven years ago I had some copies of it made and circulated these to a number of interested people, among them Professor Feller. He wrote back to say that he was amazed at how much McKendrick had done in those early days.

A more detailed account of McKendrick's life (1874–1943) can be found in the obituary by Harvey.[12]

Irwin's effort certainly helped in attracting more attention to McKendrick's work, but perhaps not as much as he had hoped. In 1982, Gani[13] made another attempt. After summarizing McKendrick's 1914[14] and 1926[1] papers, Gani wrote (p. 267)

> McKendrick's work still bears reading: it is rich in ideas, contains many unexpected results, and is full of practical good sense. He remains an inspiring model on which modern mathematical epidemiologists could well pattern themselves.

McKendrick's work was also briefly discussed in Lancaster's[15] recent book on historical development of quantitative methods in biological and medical studies.

As for Irwin himself, he was also a man with many accomplishments especially in the field of mathematical and medical statistics, as MG Kendall, in his move of the vote of thanks to Irwin's address, put it: 'Dr Irwin's outstanding work in the field of mathematical and medical statistics confers distinction on any office he assumes'. Indeed, it is not by accident that Irwin chose 'The Place of Mathematics in Medical and Biological Statistics' as the topic for his inaugural address. The bibliographical footnote of his address stated that

> JOSEPH OSCAR IRWIN (*b.* 1898) was educated at the City of London School and Christ's College, Cambridge, where he completed his Mathematics Tripos in 1921. He then went to the Galton Laboratory, at University College, London, where he remained as a member of Karl Pearson's staff until 1926. Two years later he went to Rothamsted Experimental Station and worked there, in RA Fisher's Department, until 1931. Since 1931 he has been a member of the Medical Research Council's scientific staff at the London School of Hygiene and Tropical Medicine.

A more detailed account of Irwin's life (1898–1982) can be found in an obituary by Armitage.[16]

The following quotation from Irwin's speech (p. 2) provides a good indication of his long interest in and deep concerns about medical and biological statistics:

> For more than 30 years I have been in the service of the Medical Research Council; I have seen the continual growth of the importance attached to statistical ideas in medical and biological research. I have seen a great increase in the understanding of these ideas among research workers in all applied

> sciences. Thirty years ago one had continually to emphasize the mere fact that biological variation was universal and demanded an understanding of statistical ideas and techniques to deal with it; that these ideas and techniques involved probability considerations and that probability considerations involved mathematics. Today this seems much more widely understood; on the other hand, it seems to me that it is necessary as ever to urge caution against too hasty acceptance of the adequacy of statistical models for specific purposes and to urge care in the conclusions drawn from analyses based on them.

Although another 35 years have passed, Irwin's urge of caution is still very much needed today, especially with the ever growing usage of statistical arguments in substantiating findings in medical, biological, and especially epidemiological studies, as well as in general scientific studies. While algorithms like EM are very useful for computing MLEs, they cannot correct flaws in the posited model or add information to the study. This seems to be an obvious point, yet from time to time I find myself having to disappoint someone by telling them that the fancy computational methods I could provide could not make their studies scientifically more meaningful. It is indeed very admirable to also note McKendrick's emphasis on exercising such caution in those early days. After noting that a model he developed failed to fit the data collected by the Australian government on epidemics which had occurred in incoming ships during the great epidemic of influenza in 1918, McKendrick[1] (p. 116) wrote

> Now influenza is a peculiarly difficult disease to diagnose in the individual case; consequently epidemics of less than 10 cases are in ordinary circumstances seldom recognised and reported. The conclusion is suggested that as our limited experience of epidemic influenza is based upon statistics which may relate only to small selected minority of the total number of epidemics, it may be in no sense representative, and may even be misleading.

Gani[13] attributed the great insights McKendrick displayed to his familiarity with the medical science and his facility in passing from mathematical theory to practical applications. Indeed, understanding the substantive aspects of a problem and being able to apply relevant theory or methodology are the two keys in ensuring valid statistical inferences.

5 A suggestion and very incomplete bibliographical notes on EM

In Section 3, we observed that the second EM, which operates on the joint likelihood of (p, λ), converged much slower than the first EM, which operates on the profiled likelihood of λ. This suggests that sometimes it is worthwhile to consider profiling a likelihood before implementing the EM algorithm. To be specific, suppose we need to maximize a log-likelihood $\ell(\theta_1, \theta_2 | Y_{\text{obs}})$ and suppose given θ_1 it is easy to find the conditional maximizer of $\ell(\theta_1, \theta_2 | Y_{\text{obs}})$ in the form of $\theta_2 = \phi(\theta_1)$. Then, just as with McKendrick's problem, we may have a choice between constructing an EM directly for $\ell(\theta_1, \theta_2 | Y_{\text{obs}})$ or constructing an EM for the profile log-likelihood $\ell(\theta_1, \phi(\theta_1) | Y_{\text{obs}})$. The latter has the advantage of working with a lower dimensional parameter, which can provide substantial reduction in computational complexity and time, especially for supplemented EM-type algorithms for computing information matrices; see Meng and Rubin[17] and van Dyk *et al.*[18]

This profiling strategy is in the same spirit as the expectation/conditional maximize either (ECME) algorithm of Liu and Rubin,[19] where it is used with a *conditional* *maximization* (CM) step inside the ECME iteration. A CM-step is the same as an M-step except that the maximization is with respect to part of the parameter (e.g. θ_2) with the rest part of the parameter (e.g. θ_1) held as constant (i.e. being conditioned upon). This conditional maximization idea is the base for the expectation/conditional maximization (ECM) algorithm of Meng and Rubin,[20] which was designed to reduce the complexity of the original M-step in some problems because it is often easier to seek maximizers in lower dimensional problems. The key observation made by Liu and Rubin[19] is that, with the flexibility provided by the conditional maximization scheme, one sometimes has the choice of maximizing *either* the augmented log-likelihood (e.g. (3.5)) or the actual log-likelihood (e.g. (3.2)) in the CM-steps. The profiling strategy suggests that sometimes it can be beneficial to perform the conditional maximization for the actual log-likelihood *before* implementing any EM-type algorithm. The usefulness of this strategy, of course, depends on whether one can find an *efficient* EM-type algorithm (or any other algorithms), in terms of both human and computer time, for the resulting profile log-likelihood (e.g. $\ell(\theta_1, \phi(\theta_1)|Y_{\text{obs}})$). An investigation of this strategy in the context of general EM-type algorithms may be worthwhile.

Since one of the main purposes of this article is to provide a general-level tutorial, I'd like to conclude it with a highly biased selection of recommended readings. For those interested in a systematic treatment of the EM methodology, the recently published book by McLachlan and Krishnan[21] should be an ideal textbook. The book by Little and Rubin[22] on statistical analysis of missing data also contains a substantial amount of material on EM and many examples. Tanner's[23] book is another choice, especially if one is also interested in various stochastic counterparts of the EM-type algorithms (e.g. Gibbs sampler). The recent book by Gelman *et al.*[24] contains applications of EM and its extensions (e.g. ECM) in the context of Bayesian data analysis, where these algorithms are used to find modes of posterior densities.

For those whose main interest is to acquire some general knowledge of EM and learn about its historical development, two recent encyclopedia entries are good places to start. Laird's[25] entry should be of particular interest to those whose main interest is in medical or biological applications. It contains an easily readable summary and several illustrations, with discussions on applications in medical image, molecular biology, and genetics, among others. My own entry[26] was written at a more general level with emphasis on the recent methodological development of EM-type algorithms. Both entries also contain a fairly long list of references. There are also two bibliographies available. Meng and Pedlow[27] was an earlier attempt, but soon we realized that it was essentially a hopeless task because there were simply too many EM-related papers to work with (we found over 1000 papers throughout 1991 in nearly 300 journals and volumes, only about 15% of which are statistical), not to mention the difficulties in tracking down papers that used EM but without citation of any kind. Krishnan and McLachlan have compiled a more recent (unpublished) bibliography.

For those who want to see how EM-type algorithms are actually applied in real-life medical or biological studies, the following three papers may be of interest. Wanek *et al.*[28] applied the ECM algorithm for computing MLEs to fit a multistage Markov model for progressive diseases such as melanoma. Broman *et al.*[29] applied the EM algorithm with a normal/Poisson mixture model to estimate antigen-responsive T-cell frequencies among peripheral blood mononuclear cells from human subjects. They also applied the supplemented EM (SEM) algorithm (see Meng and Rubin[17]) to compute the large-sample standard errors of their parameter estimates. In Tu *et al*.,[30,31] we applied the EM algorithm for computing the MLEs under a discrete proportional hazards model in a survival analysis of AIDS patients using the surveillance data from the Centers of Diseases Control.

Finally, for those interested in methodological and theoretical research on EM-type algorithms, the special theme issue 'EM and related algorithms' in *Statistica Sinica* (No. 1, 1995) is a good place to sample the current state-of-the-art. Besides the papers in that issue and the papers that have been cited in the previous sections, other recent methodological papers, in general or with particular models, include Wei and Tanner,[32] Baker,[33] Vardi and Lee,[34] Jamshidian and Jennrich,[35] Fesslor and Hero,[36] Lange,[37] Heyde and Morton[38] and Meng and Schilling.[39] An exciting current finding is that EM-type algorithms can be made considerably faster than previously thought possible while maintaining their stability and simplicity. The advance was made through new methods of constructing efficient data augmentation schemes. A detailed description of these methods, including the alternating expectation/conditional maximization (AECM) algorithm, is provided in Meng and van Dyk.[8]

Acknowledgements

The author thanks G McLachlan for inviting this article for publication in *Statistical Methods in Medical Research,* S Stigler for help in the literature search, and A Gelman, W Rosenberger, D van Dyk and A Zaslavsky for helpful comments. The research was supported in part by NSA Grant MDA 904-96-1-0007 and in part by NSF Grant DMS-9626691. The manuscript was prepared using computer facilities supported in part by several NSF grants awarded to the Department of Statistics at The University of Chicago, and by The University of Chicago Block Fund.

References

1 McKendrick AG. Applications of mathematics to medical problems. *Proceedings Edinburgh Mathematics Society* 1926; **44**: 98–130.

2 Irwin JO. On the estimation of the mean of a Poisson distribution with the zero class missing. *Biometrics* 1959; **15**: 324–26.

3 Irwin JO. The place of mathematics in medical and biological statistics. *Journal of the Royal Statistical Society, Series A* 1963; **126**: 1–45.

4 Hartley HO. Maximum likelihood estimation from incomplete-data. *Biometrics* 1958; **14**: 174–94.

5 Dempster AP, Laird NM, Rubin DB. Maximum likelihood estimation from incomplete-data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* 1977; **39**: 1–38.

6 Murray GD. Discussion of 'Maximum likelihood estimation from incomplete-data via the EM algorithm' by Dempster, Laird, and Rubin. *Journal of the Royal Statistical Society, Series B* 1977; **39**: 27–28.

7 Green PJ. On use of the EM algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society, Series B* 1990; **52**: 443–52.

8 Meng XL, van Dyk DA. The EM algorithm – an old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society, Series B* 1997 (in press).

9 Thisted RA. *Elements of statistical computing. Numerical computation.* New York: Chapman & Hall, 1988.

10 Wu CFJ. On the convergence properties of the EM algorithm. *Annals of Statisitics* 1983; **11**: 95–103.

11 Meng XL. On the rate of convergence of the ECM algorithm. *Annals of Statisitics* 1994; **22**: 326–39.

12 Harvey WF. Anderson Gray McKendrick (1874–1943). *Edinburgh Medical Journal* 1943; **50**: 500–504.

13 Gani J. The early use of stochastic methods: a historical note on McKendrick's pioneering papers. In: Kallianpur, Krishnaiah PR, Ghosh JK eds, *Statistics and probability: essays in honor CR Rao*. Amsterdam: North-Holland, 1982: 263–68.

14 McKendrick AG. Studies on the theory of continuous probabilities with special reference to its bearing on natural phenomena of a progressive nature. *Proceedings of the London Mathemtical Society* 1914; **13**: 401–16.

15 Lancaster HO. *Quantitative methods in biological and medical sciences.* New York: Springer, 1994.

16 Armitage P. Joseph Oscar Irwin, 1989–1982. *Journal of the Royal Statistical Society, Series A* 1982; **145**: 526–28.

17 Meng XL, Rubin DB. Using EM to obtain asymptotic variance–covariance matrices: the SEM algorithm. *Journal of the American Statistical Association* 1991; **86**: 899–909.

18 van Dyk DA, Meng XL, Rubin DB. Maximum likelihood estimation via the ECM algorithm: computing the asymptotic variance. *Statistica Sinica* 1995; **5**: 55–76.

19 Liu C, Rubin DB. The ECME algorithm: a simple extension of EM and ECM with fast monotone convergence. *Biometrika* 1994; **81**: 633–48.

20 Meng XL, Rubin DB. Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* 1993; **80**: 267–78.

21 McLachlan GJ, Krishnan T. *The EM algorithm and extensions.* New York: John Wiley, 1997.

22 Little RJA, Rubin DB. *Statistical analysis with missing data.* New York: John Wiley, 1987.

23 Tanner MA. *Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions,* 3rd edition. New York: Springer, 1996.

24 Gelman A, Carlin J, Stern H, Rubin DB. *Bayesian data analysis.* London: Chapman & Hall, 1995.

25 Laird NM. The EM algorithm. In: *Encyclopedia of biostatistics* 1997 (in press).

26 Meng XL. The EM algorithm. In: Kotz S, Read CB, eds. *Supplement volume to encyclopedia of statistical sciences.* New York: John Wiley, 1997.

27 Meng XL, Pedlow S. EM: a bibliographic review with missing articles. *Proceedings of the Statistical Computing Section*. Washington, DC: American Statistical Association: 1992; 24–27.

28 Wanek LA, Goradia TM, Elashoff RM, Makov UK. Multi-stage Markov analysis of progressive disease applied to melanoma. *Biometrics Journal* 1993; **35**: 967–83.

29 Broman K, Speed T, Tigges M. Estimate of antigen-responsive T-cell frequencies in PBMC from human subjects. *Journal of the Immunological Methods* 1997 (in press).

30 Tu XM, Meng XL, Pagano M. The AIDS epidemic: estimating the survival distribution after AIDS diagnosis from surveillance data. *Journal of the American Statistical Association* 1993; **88**: 26–36.

31 Tu XM, Meng XL, Pagano M. Survival differences and trends in patients with AIDS in the United States. *Journal of Acquired Immune Deficiency Syndromes* 1993; **6**: 1150–56.

32 Wei CG, Tanner MA. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* 1990; **85**: 699–704.

33 Baker SG. A simple method for computing the observed information matrix when using the EM algorithm with categorical data. *Journal of Computational and Graphical Statistics* 1992; **1**: 63–76 [correction, 180].

34 Vardi Y, Lee D. From image deblurring to optimal investments: maximum likelihood solutions for positive linear inverse problems (with discussion). *Journal of the Royal Statistical Society, Series B* 1993; **55**: 569–612.

35 Jamshidian M, Jennrich RI. Conjugate gradient acceleration of the EM algorithm. *Journal of the American Statistical Association* 1993; **88**: 221–28.

36 Fessler JA, Hero AO. Space-alternating generalized EM algorithm. *IEEE Transactions on Signal Processing* 1994; **42**: 2664–77.

37 Lange K. A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society, Series B* 1995; **57**: 425–37.

38 Heyde CC, Morton R. Quasi-likelihood and generalizing the EM algorithm. *Journal of the Royal Statistical Society, Series B* 1996; **58**: 317–27.

39 Meng XL, Schilling S. Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association* 1996; **91**: 1254–67.

Appendix: variances and confidence intervals

Under the binomial/Poisson model and the assumption that $0 < p < 1$ (i.e. the mixture is not degenerated), one can show that to the first order of approximation (with respect to n)

$$\mathrm{Var}(p_{\mathrm{MLE}}) = \frac{p(1-p)}{n} + \frac{p}{n(\mathrm{e}^{\lambda} - \lambda - 1)} = \frac{p(1-p)}{n}\left[1 + \frac{1}{(1-p)(\mathrm{e}^{\lambda} - \lambda - 1)}\right] \tag{A.1}$$

and

$$\mathrm{Var}(\lambda_{\mathrm{MLE}}) = \frac{\lambda}{np} + \frac{\lambda^2}{np(\mathrm{e}^{\lambda} - \lambda - 1)} = \frac{\lambda}{np}\left[1 + \frac{\lambda}{(\mathrm{e}^{\lambda} - \lambda - 1)}\right] \tag{A.2}$$

Plugging in $n = 223$ and the MLEs for p and λ, we estimate the variance of p_{MLE} by $\mathrm{var}(p_{\mathrm{MLE}}) = 0.0037$, and the variance of λ_{MLE} by $\mathrm{var}(\lambda_{\mathrm{MLE}}) = 0.0269$.

In the absence of mixing with a Poisson variable, $\mathrm{Var}(p_{\mathrm{MLE}})$ would be $p(1-p)/n$, and thus (A.1) tells us the absolute increase (i.e. the second term in the middle expression) and the relative increase (i.e. the second term inside the brackets) in variance due to mixing. For McKendrick's data set, the MLE of the relative increase is 247%, which implies that the MLE for the *effective sample size* for estimating p is only about $1/(1+2.47) = 29\%$ of $n = 223$, that is, about 65. This substantial reduction in sample size is responsible for the large standard error in p_{MLE} (about 6%), and hence the wide confidence interval for the expected number of susceptible houses, pn.

Similarly, without the zero class being truncated out, $\mathrm{Var}(\lambda)$ would be $\lambda/(np)$ – note that we already have a large reduction in sample size from $n = 223$ to np (whose MLE is $np_{\mathrm{MLE}} = 88$) due to mixing. Thus, (A.2) tells us the absolute and relative increases in variance due to truncation. For McKendrick's data set, the MLE of the relative increase is 145%, implying an effective sample size about $1/(1+1.45) = 41\%$ of $np_{\mathrm{MLE}} = 88$, that is, about 36, which is only 16% of the original size $n = 223$.

To obtain a 95% confidence interval for p or for λ, one can use the standard procedure

$$p_{\mathrm{MLE}} \pm 2\sqrt{\mathrm{var}(p_{\mathrm{MLE}})}$$

or

$$\lambda_{\mathrm{MLE}} \pm 2\sqrt{\mathrm{var}(\lambda_{\mathrm{MLE}})}$$

A better procedure is obtained by first applying an appropriate transformation to p or λ, constructing corresponding confidence intervals, and then transforming back to the original scale. With appropriate choices of transformation, it is well known that such a procedure can yield considerably better finite-sample coverage properties. For the current problem, a good choice of transformation for λ is the log transformation, $\log(\lambda)$, and for p is the logit transformation, $\mathrm{logit}(p) = \log(p/1-p)$. The MLE for $\mathrm{logit}(p)$ is $\mathrm{logit}(p_{\mathrm{MLE}})$, whose standard error can be estimated using the delta method by

$$s_p \equiv \sqrt{\operatorname{var}(\operatorname{logit}(p_{\text{MLE}}))} = \frac{\sqrt{\operatorname{var}(p_{\text{MLE}})}}{p_{\text{MLE}}(1 - p_{\text{MLE}})} = 0.2549$$

The MLE for $\log(\lambda)$ is $\log(\lambda_{\text{MLE}})$, whose standard error can be estimated by

$$s_\lambda \equiv \sqrt{\operatorname{var}(\log(\lambda_{\text{MLE}}))} = \frac{\sqrt{\operatorname{var}(\lambda)}}{\lambda_{\text{MLE}}} = 0.1687$$

The resulting interval for p is

$$\left(\frac{p_{\text{MLE}}\mathrm{e}^{-2s_p}}{1 - p_{\text{MLE}} + p_{\text{MLE}}\mathrm{e}^{-2s_p}},\ \frac{p_{\text{MLE}}\mathrm{e}^{2s_p}}{1 - p_{\text{MLE}} + p_{\text{MLE}}\mathrm{e}^{2s_p}}\right) = (0.2831, 0.5226) \tag{A.3}$$

and for λ

$$\left(\lambda_{\text{MLE}}\mathrm{e}^{-2s_\lambda},\ \lambda_{\text{MLE}}\mathrm{e}^{2s_\lambda}\right) = (0.6938, 1.3623) \tag{A.4}$$

We see that because of the transformations, the confidence intervals are always within the parameter space. The 95% intervals for the percentage of susceptible houses, np, and for the probability of being infected conditional on being susceptible, $1 - \mathrm{e}^{-\lambda}$, follow respectively from (A.3) and (A.4).