Statistical Science
2010, Vol. 25, No. 4, 450–457
DOI: 10.1214/08-STS270

The EM Algorithm in Genetics, Genomics and Public Health

Nan M. Laird

Abstract. The popularity of the EM algorithm owes much to the 1977 paper by Dempster, Laird and Rubin. That paper gave the algorithm its name, identified the general form and some key properties of the algorithm and established its broad applicability in scientific research. This review gives a nontechnical introduction to the algorithm for a general scientific audience, and presents a few examples characteristic of its application.

Key words and phrases: Incomplete data, maximum likelihood, gene counting, linkage analysis, finding regulatory motifs, diffusion batteries, particle size distributions.

1. INTRODUCTION

Incomplete data arise in many different settings in the empirical sciences. An obvious example of incomplete data is missing data, where multiple measurements are made on each subject, but some subjects are not observed on all measurements. Many applications are more subtle and consist of problems where the observed data only shed light on "hidden" or "latent" traits which are of primary interest. This frequently occurs in the engineering setting with reconstruction or indirect measurement, such as medical imaging with emission and transmission tomography. The example from public health that is diagrammed in Box 4 is another example of indirect measurement. Many applications involve locating clusters of observations with similar features; distinguishing features are functions of the observed data, but cluster membership must be inferred. Such problems occur prominently in bioinformatics and speech recognition. Finally, some statistical models, such as variance components or random effects, can be reformulated as missing data problems simply to make computations easier, even though no data are missing.

Here we discuss the EM algorithm (Dempster, Laird and Rubin, 1977, henceforth DLR), which is designed for computations in a broadly defined incomplete data setting; it is widely used in many different areas in the empirical sciences. There are numerous applications (Smith, 1957; Hasselblad, 1966; Baum et al., 1970; Orchard and Woodbury, 1972, to mention a few) of the algorithm that predate the DLR paper, but that paper described some key features of the algorithm that underlie its widespread popularity. First was the recognition of the generality of the algorithm, and a probabilistic definition of incomplete data that can be applied very broadly in different settings. Secondly, the paper provided a simple and intuitive description of the algorithm and named it "Expectation–Maximization," or EM for short, to reflect the two steps that comprise its essential nature. There are many technical descriptions of the algorithm and its properties (see, e.g., McLachlan and Krishnan, 1997), and a variety of generalizations. The purpose of this note is to provide an intuitive description of how the algorithm works and give four short examples from genetics, genomics and public health.

We first give a heuristic characterization of the algorithm; the remainder of the introduction discusses its formulation in more detail. The essential idea of the EM is to postulate the availability of additional data (e.g., the values of the missing measurements in the missing data setting) that make the estimation problem easy. The EM then proceeds by alternating between two steps: "fill in the additional data (E-step)" and "estimate the parameters using the filled in data (M-step)." This two-step process is repeated until convergence. DLR put the algorithm on a rigorous foundation by spelling out how the two steps should be implemented

Nan M. Laird is Professor, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, Massachusetts 02115, USA (e-mail: laird@hsph.harvard.edu).

in a general setting and showing that it maximizes an objective function.

To describe the algorithm, the first task is to identify a "complete data" version of the problem. This is often the most creative part of the application, since once a complete data analogue has been identified for the observed data, the application of the EM is straightforward. There generally will not be a unique representation for the complete data, but often there is an "obvious" one. For example, in the case of missing data, the complete data is best described as the observed data, plus the missing observations; in this way, the complete data specifies a complete set of measurements for each subject. In other settings, the complete data will consist of the data that are observed plus some additional information that would make the problem easy; the bioinformatics example we describe in Box 3 and clustering examples in general fall into this category. At one extreme, the complete data is sometimes best described as just the missing data, because the observed data provides no additional information when the missing data are available. As an example, consider our first example (Box 1) of estimating gene frequencies. It would be straightforward to estimate the frequencies of different gene variants (termed alleles), if we directly observed individual genotypes; in the absence of genotype information, we use data on traits related to the genotype. Here the complete data are the observed genetic traits *and* the genotypes of individuals in the sample (Box 1), but if we know the genotypes, the observed traits contribute no additional information about gene frequencies.

Having identified the complete data, we can now delineate the E- and M-steps of the algorithm. Although logically the E-step precedes the M-step in the computations, hence EM, conceptually it is easier to define the M-step first.

M-step: The way we define the complete data determines the M-step of the algorithm. At the M-step, we obtain our estimates of the parameters of interest assuming we have observed the possibly hypothetical "complete data." Formally the M-step (for Maximum likelihood) uses the complete data to obtain maximum likelihood estimates of the parameters. In many instances, this will be simple, familiar statistics, that is, means, variances and covariances, or proportions; each of the four examples we will discuss is based on a complete data multinomial likelihood and just requires estimating probabilities from sample frequencies at the M-step. Exactly how this M-step is carried out depends upon the application, but it is worth noting that the ease of computations depends in large measure on defining the complete data so that performing the M-step is easy.

E-step: Once the M-step is done, we have interim estimates of the relevant parameters which can now be used, along with the observed data, to calculate expected values for the "missing data," or technically, computing the expected log-likelihood of the complete data. Again, the exact nature of the E-step (for computing the Expected log-likelihood) is application dependent; in each of the examples we discuss, the E-step involves the computation of conditional probabilities. These two steps are iterated until convergence. Although the algorithm is not guaranteed to maximize the likelihood function, it has some attractive numerical properties. These include increasing the likelihood at each iteration and a guarantee that the parameter estimates will remain in the boundary space, that is, probabilities will always be between 0 and 1, and variance–covariance matrices will be positive semi-definite.

2. EXAMPLES FROM GENETICS

The genetics literature is replete with examples of the EM. Before genotypes were readily available via modern technologies, the EM was often used to estimate gene frequencies from data on associated Mendelian traits. An individual's genotype consists of a pair of alleles, one inherited from each parent. Even when the genotypes of individuals are observed, there are still many important estimation problems with naturally occurring incompleteness, especially if it is important to determine which allele is inherited from which parent. One example is the reconstruction of haplotypes, that is, the set of alleles at different loci all lying on the same chromosome, from pairs of alleles at the different loci. A second example that we will discuss is estimating allele sharing in a pair of affected siblings.

Gene Counting

Box 1 illustrates using the EM algorithm for estimating the three allele frequencies at the ABO hemoglobin locus. When the genotype data are directly observed, estimation is referred to as gene counting, because one simply counts the number of alleles of each type, and divides by the total number of alleles. Gene counting with the observed genotype data is shown in the M-step of Box 1, where the number of individuals (possibly unobserved) with each genotype is denoted as g_{AA},

Box 1 **Gene Counting**: Estimating the allele frequencies at the ABO hemoglobin locus

Missing Data: Genotype counts g_{AA} g_{AO} g_{BB} g_{BO} g_{AB} g_{OO}

Observed Data: Blood Type Counts t_A t_B t_{AB} t_O

Complete Data: Genotypes and blood type counts.

M-Step: Estimate allele frequencies using complete data

$$f_A = (g_{AO} + 2g_{AA} + g_{AB})/2n,$$
$$f_B = (g_{BO} + 2g_{BA} + g_{AB})/2n,$$
$$f_O + (g_{AO} + 2g_{BO} + g_{OO})/2n,$$

where n is the total number of people, and $2n$ is the number of alleles.

E-Step: Compute expected values for the unobserved genotypes

$$E(g_{AA}) = t_A\ P(AA \mid AA\ or\ AO)$$
$$E(g_{AO}) = t_A\ P(AO \mid AA\ or\ AO)$$
$$E(g_{BB}) = t_B\ P(BB \mid BB\ or\ BO)$$
$$E(g_{AO}) = t_B\ P(BO \mid BB\ or\ BO),$$

where expectations are not needed for g_{AB} and g_{OO} since they are uniquely determined by blood type. See Box 1B for details.

Box 1B **Gene Counting: A Numerical Example**

Observed Blood Types: $t_A = 200,\ t_B = 50,\ t_{AB} = 40,\ t_O = 300$

Initialization: $f_A = 1/3, f_B = 1/3, f_O = 1/3$

Iteration	Step	g_{AA}	g_{AO}	g_{BB}	g_{BO}[1]	f_A	f_B	f_O
initialize						.3333	.3333	.3333
1	E	66.6667	133.3333	20	40[2]			
	M					.2556[3]	.1000	.6444
2	E	33.0935	166.9065	4.3200	55.6800			
	M					.2276	.0869	.6855
3	E	28.4762	171.5238	3.5766	56.4234			
	M					.2237	.0863	.6900
4	E	27.8980	172.1020	3.5325	56.4675			
	M					.2237	.0863	.6900

Notes:

1. Note that $g_{AB} = t_{AB}$ and $g_{OO} = t_O$ are observed and do not change.
2. The genotypes are computed at the E-step using the frequencies from the preceding M-step (or initialization), i.e.

$$g_{BO} = t_B\{p(BO)/[p(BO) + p(BB)]\}$$
$$40 = 60\{2(1/3)^2/[2(1/3)^2 + (1/3)^2]\} = 60(2/3)$$

3. The allele frequencies are estimated at the M-step using the expected genotype counts computed at the preceding E-step and formulas given in Box 1, i.e.

$$.2556 = (66.6667 + 2\times 133.333 + 40)/1200$$

g_{AB}, etc., and the observed number of individuals with each blood type are denoted as t_A, t_B, etc. For an autosomal chromosome, each person contributes two alleles, hence the denominator of each estimated frequency is $2n$, where n is the total number of subjects.

The E-step takes into account the known relationships between blood type and genotype given in the top of Box 1, namely a person with blood type A must be either AA or AO, and similarly for blood type B, but blood types AB or O identify genotypes uniquely. For the E-step, we assume the allele frequencies are known and fixed at the values estimated at the previous M-step, and use them to calculate the P(genotype|blood type).

The EM algorithm consists of cycling through the two steps, alternately estimating the allele frequencies assuming the allele counts are observed, then updating the expected allele counts, assuming that the frequencies are known. We can start with either the E- or the M-step, depending upon whether it is easier to start with a guess about the frequencies or with a guess about the genotype counts. Note that in the iterations, the genotypes used at the M-step are expected, computed at the previous E-step; the allele frequencies used to compute the conditional expectations at the E-step are likewise those updated at the previous M-step.

Box 1B provides a numerical example to illustrate the algorithm, using hypothetical, but not unrealistic blood counts from a sample of 600 subjects. We start the iterations by setting the frequencies to be $1/3$ each. Computing the expected genotype frequencies at the E-step uses the assumption of Hardy Weinberg equilibrium to calculate the probability of genotype frequencies, given allele frequencies. Hardy Weinberg assumes that allele frequency is the same for everyone in the population and random mating, hence the two alleles of an individual are independent. Simple inspection of the observed data suggests that our initial frequencies are not very good estimates, but the algorithm converges in only a few iterations. For actual data examples and further discussion of using the EM for gene counting, see Lange (2002), Chapter 12.

Linkage Example

Linkage analysis is widely used to find the chromosomal location of a hypothesized gene affecting some trait of interest. Simply put, linkage refers to the relative position of two genetic locations; if they are physically "close" on the same chromosome, the two loci are said to be linked. If they are on the same chromosome, but distant, or if they are on different chromosomes, they are said to be unlinked. The EM has many applications in linkage analysis; here we consider its use in estimating allele sharing of affected siblings. When two loci are unlinked, Mendel's laws of inheritance holds

independently for the two loci, and one can easily calculate the probability of allele sharing between two siblings at a genetic locus. In particular, the probability that both siblings inherit the exact same alleles from both their parents (allele sharing is 2) is $\frac{1}{4}$. The probability that they each inherit two different alleles from both parents (or allele sharing is zero) is also $\frac{1}{4}$, and the probability that they inherit the same allele from exactly one parent (allele sharing is 1) is $\frac{1}{2}$. This type of allele sharing is called identity-by-descent (IBD) to indicate that the alleles are shared because the sibs have obtained the same copy from a common parent.

To implement a linkage analysis, we obtain data on a genetic locus, called marker data, that we hypothesize is close to a genetic locus that affects our disease of interest. If both siblings are affected with the same genetic disease, and the marker and the disease locus are linked, we expect they are likely sharing at least one allele, inherited from the same parent, at the unobserved disease locus. Further, if the locus underlying the disease is linked to the marker, independent transmission at the 2 loci does not hold. Rather, we expect that the allele sharing probabilities at the marker differ from $\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$, in the direction of increased sharing. To test this, we estimate the probabilities of sharing 0, 1 or 2 alleles under H_A, π_0, π_1, π_2 say, from the sample of affected sib pairs and compare them to the allele sharing probabilities under the null hypothesis of no linkage, $\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$.

Estimation of the allele sharing frequencies under the alternative of linkage between the marker and the disease locus is straightforward if we could observe the IBD sharing directly; we would simply count the number of affected sibs pairs sharing 0, 1 or 2 alleles, and divide by the number of affected sib pairs. This is illustrated in the M-step of Box 2 where the complete data IBD counts are denoted as Z_0, Z_1 and Z_2. As illustrated in Box 2, one cannot always infer IBD sharing from data on parental and offspring genotypes; it depends upon the pattern of alleles which are observed. The "complete data" illustrates a setting where we can always deduce IBD sharing, that is, the parents have four distinct alleles. The "observed data" shown in Box 2 illustrates two situations where we cannot. In general, the "observed data" are a pattern of sharing, for example, either 0 or 1, 1 or 2, etc., together with the observed parental genotypes. The M-step is based on a multinomial likelihood, assuming that we observe IBD for each sib pair, and the E-step provides expectations of the multinomial counts, by computing the expected allele sharing for each pair conditional on their observed pattern of sharing, and adding over pairs. The conditional probability of IBD sharing is obtained from Bayes rule

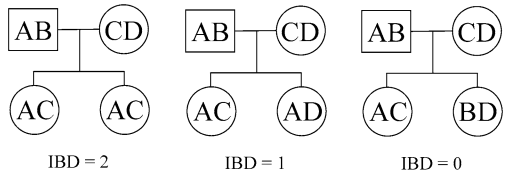
$$P(\text{IBD} = j|\text{observed data}) \propto P(\text{observed data}|\text{IBD} = j)\pi_j \quad \text{for } j = 0, 1, 2,$$

where π_j has been estimated at the M-step; and $P(\text{observed data}|\text{IBD} = j)$ is calculated from the observed data on parents and child's genotypes (Risch, 1990).

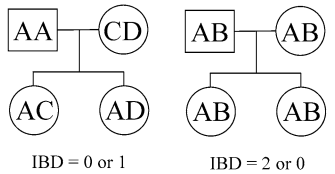
Risch (1990) also extends the EM to the setting where parental data are missing, but now one must first have an estimate of parental allele frequencies. Kruglyak et al. (1995) extended the algorithm to cover multipoint analyses with complex pedigrees as well as additional markers, but the basic idea of using EM to estimate IBD probabilities under H_A remains the same.

Box 2 | Affected Sib Pair Linkage Analysis

Complete Data: Number of alleles shared IBD for each pair of sibs:



Observed Data: Pattern of allele sharing:



M-Step: Estimate IBD sharing, assuming IBD is known for each pair:

$$\pi_0 = Z_0/n, \quad \pi_1 = Z_1/n, \quad \pi_2 = Z_2/n$$

where n is the number of sib pairs, and Z_0, Z_1, Z_2 are counts of pairs sharing 0, 1, or 2 alleles.

E-Step: Compute the expected counts of allele sharing, e.g.: for $j = 0, 1, 2$

$$E(Z_j) = \sum P(i^{th} \text{ pair shares } j \text{ alleles} \,|\, \text{observed data},)$$

where summation is over all sib pairs and

$$P(i^{th} \text{ pair shares } j \,|\, \text{observed data}) \propto P(\text{observed data} \,|\, \text{sharing } j)\pi_j.$$

3. EXAMPLE FROM COMPUTATIONAL BIOLOGY: FINDING MOTIFS

Computational Biology deals with the analysis of data that comes from sequencing the DNA of humans and other organisms. The specific sequence of the four nucleotides which make up DNA, Adenine, Guanosine, Thymine and Cytosine, or A, G, C, T, for short,

determine the function of genes and location of genes, the process of RNA transcription, and the manufacture of proteins essential for cell function. Understanding many fundamental biological processes requires tools to identify relatively short patterns of these four base pairs embedded in long strings of base pairs (approximately 3 billion in the entire human genome). The problems are challenging and relevant to statisticians because the sequences of interest may not be "exact." For example, the simple sequence consisting of two specific base pairs, CG, is relatively rare in the DNA of many organisms, because CG readily mutates to TG. But mutation is suppressed in regions near specific genes, forming CG rich islands, or stretches of DNA that have more CG pairs than "usual." One approach to identifying GC rich islands is to treat strings of DNA as realizations of Hidden Markov Models, with different (unobserved) states corresponding to GC rich or GC poor regions (Parida, 2008, Chapter 5.5; Jones and Pevzner, 2004, Chapter 11). An EM algorithm to estimate the state and transmission probabilities was developed by Baum et al. (1970).

A related problem in computational biology where EM is used is the identification of regulatory motifs. Motifs are short sequences of base pairs, from 6 to 20 pairs in length, which have a similar pattern of base pairs. Proteins bind to functional motifs located upstream of genes to encourage the process of RNA transcription in the genes. Given a set of known genes, the approximate location of the corresponding functional motifs is known, however their exact sequences vary because the protein binding process does not require an exact sequence of base pairs.

The basic idea is illustrated in the top panel of Box 3. Seven hypothetical DNA fragments are given an input data (Jones and Pevzner, 2004, Chapter 4). The underlined portion of each sequence denotes the actual (unobserved) motif. Here we assume there is only one motif in each input sequence, and it is known to be exactly eight base pairs long. The exact DNA letters vary from fragment to fragment, but two motifs are "more similar" than two randomly selected sequences of eight DNA letters. The problem can be defined probabilistically by assuming that the probability corresponding to a letter in a motif location is the same for every motif, but differs from nonmotif, or "background" DNA. The objective is to characterize the general pattern of DNA for the motifs as a "consensus" sequence.

A purely computational approach to this problem is to pick a metric for measuring similarity between eight letter sequences, such as the number of positions

Box 3 | Motif Finding

Complete Data: Seven DNA fragments, with the motifs underlined:

CGGGGCTATCCAGCTGGGTCGTCACATCCCCTT
TGCCCAATAAGGGCAACTCCAAAGCGGACAAA
GGATGGATCTGATGVVVGTTTGACGACCTA
AAGGAAGCAACCCCAGGAGCCCTTTGCTC
AATTTTCTAAAAAGATTATAATGTCGGTCCTTGGAACTTC
CTGCTGTACAACTGAGATCATGCTGCATGCCATTTCAAC
ACATGATCTTTTGATGGCACTTGGATGAGGGAATGATGC

Observed Data: The sequences without the motifs shown:

CGGGGCTATCCAGCTGGGTCGTCACATCCCCTT
TGCCCAATAAGGGCAACTCCAAAGCGGACAAA
GGATGGATCTGATGVVVGTTTGACGACCTA
AAGGAAGCAACCCCAGGAGCCCTTTGCTC
AATTTTCTAAAAAGATTATAATGTCGGTCCTTGGAACTTC
CTGCTGTACAACTGAGATCATGCTGCATGCCATTTCAAC
ACATGATCTTTTGATGGCACTTGGATGAGGGAATGATGC

M-Step: Estimate the probability of each base pair in each motif position, assuming known starting position, i.e., $P(A \mid \text{position } 1) = 5/7$, etc.,

	Background	Motif Position						
Letter	Position	1	2	3	4	5	6	7	8
A	48	5	1	0	0	5	5	0	0
C	43	0	0	1	4	2	0	6	0
G	45	1	1	6	3	0	1	0	0
T	48	1	5	0	0	0	1	1	8
Totals	144	7	7	7	7	7	7	7	7

E-Step: Compute expected counts needed for the M-step:

Expected #A's in position 1
$= \Sigma\Sigma P(\text{A in position 1} \mid \text{motif starting point, observed data})$
Where the summation is over all possible starting positions within a fragment, and over all fragments.

Data modified from Jones and Pevzner (1985, Figure 4.3).

that have the same letter across all fragments, and then search for the set of eight letter sequences that optimizes the metric. This is a time consuming search, since each possible alignment has to be considered, and each fragment of X letters has $X - 7$ possible starting points for the eight letter sequence; additionally it is difficult to measure optimality of the solution.

Alternatively, one may utilize a probability model for the fragment data, and use the EM (Lawrence and Reilly, 1990). The data can be modeled as incomplete because the motif locations are not observed. Conceptually, the missing data are seven indicator vectors, indicating the starting point of each motif in each of the seven fragments. The parameters to be estimated are the frequencies of the four DNA letters in the motif and nonmotif positions. We assume a 4×8 matrix of multinomial probabilities, one column for each position in a motif. Each element of the column gives $P(A)$, $P(C)$, $P(G)$ and $P(T)$ for the DNA letter in each position of the motif. If we knew the location of each motif in each fragment, we would estimate these probability vectors via simple multinomial frequencies for each position of the motif as is illustrated in the M-step of Box 3.

We do not know the starting points of each motif, but given the motif and background probabilities, one can

readily calculate the conditional probability of any possible starting position for each fragment, by assuming a priori that all possible starting values are equally likely and evaluating the probability of each DNA sequence for each assumed starting point. These probabilities of each starting point for each motif are then used to compute the expected multinomial counts needed for the M-step, as shown in the E-step of Box 3.

At the conclusion of the iterations of the EM, we have the matrix of estimated base pair probabilities for each location. These can be used to compute a "consensus sequence" by taking the base pair with the highest frequency in each location of the motif. For example, based on the frequencies computed at the M-step in Box 4, the consensus sequence would be ATGCAACT. At the conclusion of the EM we also have the probability of each alignment (determined by the estimated starting points). This has been used to assign motif locations to specific sequences. While the motif sequence frequencies can be well estimated with a large number of fragments, in this formulation there is no simplifying model for alignment probabilities, and the number of possibilities grows exponentially with the number of fragments. Hence alignments are unlikely to be well estimated. This has led to interest in Bayesian approaches based on Gibbs sampling (Lawrence et al., 1993). The application of the EM described here, as in many other settings, is sensitive to starting values; see Parida (2008), Chapter 8.6. Generalizations of the simple case discussed here which allow multiple types of motifs and as well as multiple numbers per fragment have been given in Cardon and Stormo (1992) and Bailey and Elkan (1995).

4. EXAMPLE FROM PUBLIC HEALTH: MONITORING AIR QUALITY

Many harmful exposures, radon or diesel emissions for example, are characterized by having particles with very small diameters (less than 0.4 micrometers). Particles this small cannot be directly measured, but having estimates of particle sizes are important for monitoring air quality. Several measurement devices have been developed to deal with this problem; here we discuss diffusion batteries which operate on a principle of indirect measurement. The basic principle is the same used for positron emission tomography (PET scans) as well as transmission tomography (Vardi, Shepp and Kaufman, 1985; Lange and Carson, 1984; Kay, 1997).

Diffusion batteries are designed to filter out particles of different sizes by passing a volume of aerosol

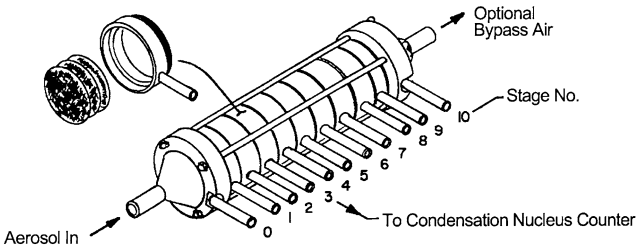


FIG. 1. *Diagram of a diffusion battery. Reproduced from the TSI Instruction Manual for Diffusion Battery Models* 3040/3041.

through a succession of fine wire mesh screens, and counting the number of particles remaining in the aerosol at each stage. Figure 1 illustrates how the diffusion battery works.

A fixed volume of air is drawn in through the entrance port, with only one of the exit ports open. The total number of particles passing through the exit port is counted. Thus the observed data consists of 11 counts of particles. The "zero port" counts the total number in the volume of aerosol regardless of size, since there are no screens before the zero port. The subsequent ports have differing numbers of screens which increase the probability that particles of different sizes are trapped in the screens, and are thus not counted. The smaller particles are more likely to be removed at the early stages, since a particle's diameter determines how fast they move. The smaller particles are moving faster and are more likely to hit a barrier (a screen), and become trapped at the earlier stages. The larger particles are sluggish; they tend to fall through the battery, only becoming trapped at the end stages of the battery where there are many more screens. We estimate the distribution of particle sizes from the total particle counts measured at each port by dividing the particle size distribution into intervals, and estimating the proportions in each interval.

A natural way to formulate an incomplete data problem in this setting is illustrated at the top of Box 4. We define complete data as a 2-way array of counts of particles in size category j exiting at the ith port, Z_{ij}, where $i = 1, 10$ and $j = 1, 8$ in our example. The unobserved Z_{ij} can be modeled as independent Poisson counts with $E(Z_{ij}) = P_0 w_{ij} f_j$, where f_j are the frequencies of the jth size category, and

$$w_{ij} = P(\text{particle of size } j \text{ exits at the } i\text{th port}).$$

The w_{ij} are calculated from the known characteristics of the diffusion battery. The observed counts exiting

Box 4 Estimating Particle Size Distribution from Diffusion Battery Data

Complete Data: Array of size classified counts (Z's) for each port (excluding the 0 port):

Midpoint of Size Interval (in micrometers)

	0.001	0.002	0.005	0.01	0.02	0.043	0.095	0.233	Totals
Port 1	$Z_{1,1}$	$Z_{1,2}$	$Z_{1,3}$	$Z_{1,4}$	$Z_{1,5}$	$Z_{1,6}$	$Z_{1,7}$	$Z_{1,8}$	P_1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Port 10	$Z_{10,1}$	$Z_{10,2}$	$Z_{10,3}$	$Z_{10,4}$	$Z_{10,5}$	$Z_{10,6}$	$Z_{10,7}$	$Z_{10,8}$	P_{10}
Totals	N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_8	

Observed Data: Total counts at each port (including the zero port):

$$P_0, P_1, P_2, \ldots, P_{10}$$

M-Step: Estimate the frequency of particles in each size interval using weighted counts:

$$f_j = N_j / P_0 \times W_j, \quad \text{for } j = 1,\ldots,8$$

where $W_j = P$(particle in interval j is counted at any port); this probability is calculated from the properties of the diffusion battery.

E-Step: Compute expected counts of the complete data array as:

$$E\left(Z_{i,j}\right) = P_i \times \left(w_{i,j} \times f_j\right) \Big/ \left(\sum_{j=1}^{8} w_{ij} \times f_j\right)$$

where

$w_{i,j} = P\left(\text{particle in interval } j \text{ exits at port } i\right)$ and $W_j = \sum_{i=1}^{10} w_{ij}$

Obtain the $E(Nj) = \sum_i E\left(Z_{ij}\right)$

each port are just the row totals of the Z_{ij}, P_i, plus the count at the zero port, P_0. We note that under this set up, the expected values of the observed counts follow a simple linear model:

$$E(P_i) = P_0 \sum_j w_{ij} \times f_j.$$

Thus estimation can be treated as a linear regression problem where the f_j's are the coefficients to be estimated and the w_{ij}'s are the known predictors. Because the f_j's are constrained to be positive, and the errors are not normally distributed, ordinary least squares does not work well. Typically, non-negative least squares or Ridge regression have been used as alternatives, but using the EM to obtain maximum likelihood estimates under the Poisson model described below represents a substantial improvement (Maher and Laird, 1985).

The use of the EM for this application is illustrated in Box 4. The column totals of the array, N_j, give the number of particles in a given size category which exit at all ports combined. Only the N_j's are needed for the M-step, but taking the full array of counts as complete data simplifies the calculations. The proportions, say f_j, in each size category are estimated from the weighted frequencies $f_j = N_j/P_0 \times W_j$. We use weights,

$$W_j = P(\text{particle of size } j \text{ is counted at any port}) = \sum_{i=1}^{10} w_{ij},$$

because the complete data counts have been filtered according to size.

At the E-step, we compute the expected values of each Z_{ij}, conditioning on the observed row margins and assuming the f_j's are known.

Conclusion

This paper merely skims the surface of the multitude of applications in diverse scientific areas where the EM plays an important role, not just in the computations, but in the conceptualization of the problem as well. This volume provides additional examples from other areas of the empirical sciences.

REFERENCES

BAILEY, T. L. and ELKAN, C. (1995). Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Mach. Learn.* **41** 51–80.

BAUM, L. E., PETRIE, T., SOULES, G. and WEISS, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* **41** 164–171. MR0287613

CARDON, L. R. and STORMO, G. D. (1992). Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *J. Mol. Biol.* **223** 159–170.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

HASSELBLAD, V. (1966). Estimation of parameters for a mixture of normal distributions. *Technometrics* **8** 431–444. MR0196842

JONES, N. C. and PEVZNER, P. A. (2004). *An Introduction to Bioinformatics Algorithms*. MIT Press, Cambridge, MA.

KAY, J. (1997). The EM algorithm in medical imaging. *Stat. Methods Med. Res.* **6** 55–75.

KRUGLYAK, L. and LANDER, E. S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics* **57** 439–454.

LANGE, K. (2002). *Mathematical and Statistical Methods for Genetic Analysis*, 2nd ed. Springer, New York. MR1892279

LANGE, K. and CARSON, R. (1984). EM reconstruction algorithms for emission and transmission tomography. *Journal of Computer Assisted Tomography* **8** 306–316.

LAWRENCE, C. E., ALTSCHUL, S. F., BOGUSKI, M. S., LIU, J. S., NEUWALD, A. F. and WOOTTON, J. C. (1993). Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* **262** 208–214.

LAWRENCE, C. E. and REILLY, A. A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *PROTEINS: Structure, Function, and Genetics* **7** 41–51.

MAHER, E. F. and LAIRD, N. M. (1985). EM algorithm reconstruction of particle size distributions from diffusion battery data. *Journal of Aerosol Science* **16** 557–570.

MCLACHLAN, G. J. and KRISHNAN, T. (1997). *The EM Algorithm and Extensions*. Wiley, New York. MR1417721

ORCHARD, T. and WOODBURY, M. A. (1972). A missing information principle: Theory and applications. In *Proceedings of the Sixth Berkeley Symp. Math. Statist. Probab.* **I** 697–715. Univ. California Press, Berkeley. MR0400516

PARIDA, L. (2008). *Pattern Discovery in Bioinformatics: Theory & Algorithms. CRC Mathematicsal and Computational Biology Series* **16**. Chapman & Hall/CRC, Boca Raton, Florida. MR2327741

RISCH, N. (1990). Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. *American Journal of Human Genetics* **46** 242–253.

SMITH, C. A. (1957). Counting methods in genetical statistics. *Annals of Human Genetics* **21** 254–276. MR0088408

VARDI, Y., SHEPP, L. A. and KAUFMAN, L. (1985). A statistical model for positron emission tomography. *J. Amer. Statist. Assoc.* **80** 8–37. MR0786595