# An Interval Analysis Approach to the EM Algorithm

Kevin Wright and William J. Kennedy

The EM algorithm is widely used in incomplete-data problems (and some complete-data problems) for parameter estimation. One limitation of the EM algorithm is that upon termination, it is not always near a global optimum. As reported by Wu (1982), when several stationary points exist, convergence to a particular stationary point depends on the choice of starting point. Furthermore, convergence to a saddle point or local minimum is also possible. In the EM algorithm, although the log-likelihood is unknown, an interval containing the gradient of the EM $q$ function can be computed at individual points using interval analysis methods. By using interval analysis to enclose the gradient of the EM $q$ function (and, consequently, the log-likelihood), an algorithm is developed that is able to locate all stationary points of the log-likelihood within any designated region of the parameter space. The algorithm is applied to several examples. In one example involving the $t$ distribution, the algorithm successfully locates (all) seven stationary points of the log-likelihood.

**Key Words:** Interval arithmetic, Optimization, Interval EM, Maximum likelihood

## 1 INTRODUCTION

This article explores a variation of the EM algorithm which uses techniques of interval analysis to locate multiple stationary points of a log-likelihood.

Interval analysis can be used to compute an interval which encloses the range of a function over a given domain. By using interval analysis to compute an enclosure of the gradient of the log-likelihood over specific regions, those regions where the enclosure of the gradient does not contain 0 can be ruled out from containing any stationary points. The algorithm locates stationary points by repeatedly dividing into smaller regions precisely those regions that have not been ruled out.

The structure of this article proceeds as follows. Section 2 presents an introduction to interval analysis sufficient to understand this article. Some of the differences between calculations with real numbers and interval numbers are noted, along with some comments about performing interval arithmetic on digital computers. Section 3 briefly states the traditional EM algorithm and presents a computational example using both scalars and intervals. Section 4 introduces a new approach to the EM algorithm using interval analysis. Section 5 presents several examples of the algorithm applied to different problems. These examples demonstrate both the accuracy which interval arithmetic can provide and the ability of the algorithm to locate multiple stationary points. Section 6 provides some conclusions.

## 2 INTERVAL ANALYSIS

A good introduction to interval analysis can be found in monographs by Hansen (1992) and Moore (1979). Some of the fundamental concepts of interval analysis are now presented.

[1] Kevin Wright is Senior Research Associate, Pioneer Hi-Bred International, Inc., 7300 NW 62nd Avenue, Johnston, IA 50131-1004 (E-mail: Kevin.Wright@pioneer.com). William J. Kennedy is Professor, Department of Statistics, Iowa State University, 117 Snedecor Hall, Ames, IA 50011-1210 (E-mail: wjk@iastate.edu).

*Journal of Computational and Graphical Statistics, Volume 9, Number 2, Pages 1–16*

In this article, intervals will be indicated by superscript $I$ and vectors will be denoted by boldface. An interval $x^I = [\underline{x}, \overline{x}]$ is a closed and bounded set of real numbers. For two intervals $x^I$ and $y^I$, interval arithmetic operators are defined in the following manner:

$$x^I \circ y^I = \{x \circ y : x \in x^I, y \in y^I\}$$

where $\circ \in \{+, -, *, /\}$ and division is undefined for $0 \in y^I$. For these four interval arithmetic operators, closed-form expressions can be obtained for direct calculation of results of the operations. For example, if $x^I = [\underline{x}, \overline{x}]$ and $y^I = [\underline{y}, \overline{y}]$, then $x^I + y^I = [\underline{x} + \underline{y}, \overline{x} + \overline{y}]$. The *Hull* of a set of intervals $x_1^I, \ldots, x_n^I$ is the smallest interval containing $x_1^I, \ldots, x_n^I$; that is,
$\text{Hull}(x_1^I, \ldots, x_n^I) = [\inf\{x : x \in x_i^I, i = 1, \ldots, n\}, \sup\{x : x \in x_i^I, i = 1, \ldots, n\}]$. An *interval vector* or *box* is simply a vector of intervals. An *interval function* is an interval-valued function of one or more interval arguments. In this article, capital letters are used to denote interval functions. An interval function $F(x_1^I, \ldots x_n^I)$ is said to be an *interval extension* or *interval enclosure* of $f(x_1, \ldots, x_n)$ if $F([x_1, x_1], \ldots, [x_n, x_n]) = f(x_1, \ldots, x_n)$ for all $x_i, i = 1, \ldots, n$. An interval function $F$ is said to be *inclusion monotonic* if $F(x^I) \subset F(y^I)$ whenever $x^I \subset y^I$. A fundamental property of interval analysis is that rational interval functions are inclusion monotonic.

In this article, the *natural interval extension* of a real function is used. This is an interval extension in which intervals and interval operations are substituted for scalars and scalar operations. The value of any interval extension of a function is dependent on the form of the real function. For example, let $f_1(x) = (x - 1)(x + 1)$ and $f_2(x) = xx - 1$. Let $F_1$ and $F_2$ be the corresponding natural interval extensions and let $x^I = [-2, 1]$. Then $F_1(x^I) = [-6, 3]$ and $F_2(x^I) = [-3, 3]$ which both contain $[-1, 3]$, the true range of $f_1$ and $f_2$ over $x^I$. This feature of interval computations to sometimes overestimate the range of a function is referred to as *interval dependency.* Attention must be given to the exact expression of an interval function to reduce the effect of interval dependency. Hansen (1997, 1992) presented some results regarding this topic.

When implementing interval arithmetic calculations on computers, care must be taken to ensure that rounding errors do not invalidate the inclusion monotonicity of interval results. One way to achieve this is through the use of directed rounding modes in the floating-point calculations. When calculating the lower endpoint of an interval result, the floating-point processor is set to round all results *down.* For calculation of the upper endpoint of an interval result, all calculations are rounded *up.* Using the symbols $\bigtriangledown$ and $\bigtriangleup$ to denote downward and upward rounding respectively, the actual computer implementation of interval addition is $x^I + y^I = [\bigtriangledown(\underline{x} + \underline{y}), \bigtriangleup(\overline{x} + \overline{y})]$. Correct use of the rounding modes guarantees that the computed result contains the true interval answer. There exist programming languages and software packages which are able to work with interval data types and interval operators. Kearfott (1996) compares some of these packages including INTLIB_90, C-XSC, BIAS/PROFIL, and others. There are several ways in which these packages implement interval versions of common functions (for example, exp, log) depending on the architecture of the underlying hardware and software. In some cases hardware may support directed rounding and it may be sufficient to set the rounding mode before calling a built-in function. In other cases it may be necessary to use techniques like a Taylor series approximation with bounds on the truncation error.

For the research in this article, the computations were done using the BIAS/PROFIL package in C++ developed by Knüppel (1993).

## 3 THE EM ALGORITHM

The present-day version of the EM algorithm first appeared in a landmark article by Dempster et al. (1977). The EM algorithm is a general iterative algorithm for maximum likelihood estimation in incomplete-data problems. The EM algorithm has not only been successfully applied in obvious incomplete-data problems, but also in many situations where the data appears to be complete, but can be viewed as incomplete by introducing latent variables. The intuitive idea behind the EM algorithm is to iterate the following two steps:

Expectation step: Replace missing values (sufficient statistics) by estimated values.

Maximization step: Estimate parameters as if no data were missing.

Formally, starting with a parameter estimate $\boldsymbol{\phi}_p$, the E-step calculates the conditional expectation of the complete-data log-likelihood, $\log L_c(\boldsymbol{\phi})$, as $q(\boldsymbol{\phi}|\boldsymbol{\phi}_p) = E_{\boldsymbol{\phi}_p}\{\log L_c(\boldsymbol{\phi})\}$ and then the M-step chooses $\boldsymbol{\phi}_{p+1}$ to be any value of $\boldsymbol{\phi} \in \boldsymbol{\Omega}$ that maximizes $q(\boldsymbol{\phi}|\boldsymbol{\phi}_p)$; that is, $q(\boldsymbol{\phi}_{p+1}|\boldsymbol{\phi}_p) \geq q(\boldsymbol{\phi}|\boldsymbol{\phi}_p)$ for any $\boldsymbol{\phi} \in \boldsymbol{\Omega}$.

### 3.1 An Example Application of the EM Algorithm

The following example from Dempster et al. (1977) is frequently used to introduce the EM algorithm. Consider a set of 197 animals which are classified into four categories. The observed classification counts are $\mathbf{y} = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$. The classification of the random variable $Y$ is modeled as following a multinomial distribution:

$$Y \sim Multinomial\left(197, \frac{1}{2} + \frac{1}{4}p, \frac{1}{4} - \frac{1}{4}p, \frac{1}{4} - \frac{1}{4}p, \frac{1}{4}p\right)$$

where $p$ is unknown and to be estimated. There is no missing data in this problem and $p$ is easily estimated by a maximum likelihood approach. For illustration purposes, the problem is reformulated with missing data. Suppose the first classification category $Y_1$ is split into two categories and a new random variable $X$ is modeled:

$$X \sim Multinomial\left(197, \frac{1}{2}, \frac{1}{4}p, \frac{1}{4} - \frac{1}{4}p, \frac{1}{4} - \frac{1}{4}p, \frac{1}{4}p\right).$$

The incomplete data vector $\mathbf{x}$ is $(x_1, x_2, x_3, x_4, x_5)$ and thus $\mathbf{y}$ can be written $\mathbf{y}(\mathbf{x}) = (x_1 + x_2, x_3, x_4, x_5)$.

For the incomplete-data problem, it is easy to show that

$$(X_2|X_1 + X_2 = 125) \sim Binomial\left(125, \frac{\frac{1}{4}p}{\frac{1}{2} + \frac{1}{4}p}\right),$$

and the E-Step in the usual scalar EM algorithm becomes $x_{1,k} = 125(\frac{1}{2})/(\frac{1}{2} + \frac{1}{4}p)$ and $x_{2,k} = 125(\frac{1}{4}p)/(\frac{1}{2} + \frac{1}{4}p)$. From the complete-data likelihood of $p$,

$$f(p|\mathbf{x}) \propto \left(\frac{1}{2}\right)^{x_{1,k}} \left(\frac{1}{4}p\right)^{x_{2,k}} \left(\frac{1}{4} - \frac{1}{4}p\right)^{x_3+x_4} \left(\frac{1}{4}p\right)^{x_5}$$

the M-Step in the usual scalar EM algorithm is:

$$p_k = \frac{x_{2,k} + x_4}{x_{2,k} + x_3 + x_4 + x_5} = \frac{x_{2,k} + 34}{x_{2,k} + 72}.$$

Using $p_0 = 0.5$ as a starting value and using a convergence tolerance of $\epsilon = 10^{-7}$, the (scalar real) EM algorithm yields:

```
Epsilon:    1e-07
Initial p: 0.5
k       x2        p
1     25         0.608247
2     29.1502   0.624321
3     29.7373   0.626489
4     29.8159   0.626777
5     29.8263   0.626816
6     29.8277   0.626821
7     29.8279   0.626821
8     29.8279   0.626821
```

The algorithm converges at the specified tolerance after eight iterations. In this case, the starting value of 0.5 for $p$ was chosen simply because 0.5 lies exactly halfway between 0 and 1, which define the bounds for possible starting values. A questioning user may well wonder what results would be obtained for different starting values and how the steps of convergence might change. Interval analysis can be used to answer those questions.

### 3.2 Traditional EM with Intervals as Computing Elements

The foregoing example can easily be programmed to use intervals as the computing elements, though with a slight modification. Because of the dependency problem, narrower interval enclosures of computed values are more likely to be obtained if each variable appears only once in a calculation. The iterates in the EM algorithm for this particular example are therefore written equivalently as

$$x_{2,k}^I = \frac{125}{2/p_k^I + 1} \quad \text{and} \quad p_k^I = 1 - \frac{38}{x_{2,k}^I + 72}.$$

The convergence tolerance remains the same as before, but now $p_0^I = [\delta, 1]$ where $\delta$ is a small machine number greater than zero. The scalar EM algorithm using interval arithmetic produces the following output:

```
Epsilon:   1e-07
Initial p: [4.94066e-324,1]
k       x2                    p
1   [0,41.6667]         [0.472222,0.665689]
2   [23.8764,31.2156]   [0.603656,0.631839]
3   [28.9812,30.0094]   [0.623692,0.627485]
4   [29.7144,29.852]    [0.626405,0.626910]
5   [29.8129,29.8311]   [0.626766,0.626833]
6   [29.8259,29.8284]   [0.626814,0.626823]
7   [29.8277,29.828]    [0.626821,0.626822]
8   [29.8279,29.828]    [0.626821,0.626822]
9   [29.8279,29.8279]   [0.626821,0.626822]
```

It is now easy to see that all scalar starting values of $p_0$ in the scalar EM algorithm will lead to the same point of convergence, and furthermore the number of iterations to convergence is not highly dependent on the starting value of $p$. The use of interval arithmetic has allowed the user to consider all possible values of the input parameter simultaneously.

Although this particular example works quite well, it can be shown that in a situation where two stationary points exist, the algorithm may become stuck in a loop with each endpoint of an interval at a stationary point. A more general technique is needed.

## 4 AN INTERVAL EM ALGORITHM

A method is now presented which uses certain properties of the EM algorithm and of interval arithmetic to locate all stationary points of the likelihood inside of a given region of the parameter space. Briefly, from the EM algorithm it is known that the $q$ function has a gradient which is equal to the gradient of the log-likelihood at stationary points of the log-likelihood. Using interval arithmetic, it is possible to derive interval vectors which enclose values of the gradient of the $q$ function even over regions which do not contain a stationary point.

The complete method is presented below, followed by a summary outline and additional comments. Some numerical results are presented in Section 5.

### 4.1 Enclosing the Gradient of the Log-Likelihood

The fundamental task for the method being proposed will be to eliminate regions of the parameter space where it can be determined that a stationary point of the likelihood does not exist. This can be accomplished by finding a box which encloses the range of the gradient of the log-likelihood over a region. If, for example, the interval enclosure of the set of all values of the gradient of the log-likelihood $\ell(\boldsymbol{\phi})$ over the box $\boldsymbol{\phi}^I$,

$$\left\{ \left.\frac{\partial \ell(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}_p} : \boldsymbol{\phi}_p \in \boldsymbol{\phi}^I \right\},$$

does not contain 0 in one or more of its coordinates, then the gradient of the log-likelihood is nonzero over $\boldsymbol{\phi}^I$ and $\ell(\boldsymbol{\phi})$ does not contain a stationary point inside the box $\boldsymbol{\phi}^I$. A more thorough explanation of how this is accomplished is now presented by deriving an interval enclosure for the gradient of the log-likelihood. The first part of this derivation is similar to the development in Dempster et al. (1977).

Denote the complete data (which includes missing values) by $\mathbf{x}$ and the observed (incomplete) data by $\mathbf{y}$, where $\mathbf{y} = \mathbf{y}(\mathbf{x})$. Let the density function of $\mathbf{x}$ be $f(\mathbf{x}|\boldsymbol{\phi})$, where $\boldsymbol{\phi} \in \boldsymbol{\Omega}$. From this, the density function for $\mathbf{y}$ is

$$g(\mathbf{y}|\boldsymbol{\phi}) = \int_{\mathbf{x}(\mathbf{y})} f(\mathbf{x}|\boldsymbol{\phi}) d\mathbf{x}.$$

For simplicity and tractability, the maximization step would ideally be accomplished over $\boldsymbol{\phi}$ in $\log f(\mathbf{x}|\boldsymbol{\phi})$. However, since $\mathbf{x}$ is unobservable, replace $\log f(\mathbf{x}|\boldsymbol{\phi})$ by its conditional expectation. To that end, let $k(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi}) = f(\mathbf{x}|\boldsymbol{\phi})/g(\mathbf{y}|\boldsymbol{\phi})$ be the conditional density of $\mathbf{x}$ given $\mathbf{y}$ and $\boldsymbol{\phi}$. Using this, the log-likelihood can be written

$$\ell(\boldsymbol{\phi}) = \log g(\mathbf{y}|\boldsymbol{\phi}) = \log f(\mathbf{x}|\boldsymbol{\phi}) - \log k(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi}).$$

Taking the conditional expectation (using $\boldsymbol{\phi}_p$ as an estimate for $\boldsymbol{\phi}$),

$$\ell(\boldsymbol{\phi}) = E_{\boldsymbol{\phi}_p}\left[\log f(\mathbf{x}|\boldsymbol{\phi})|\mathbf{y}\right] - E_{\boldsymbol{\phi}_p}\left[\log k(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi})|\mathbf{y}\right].$$

For simplicity, this is often written $\ell(\boldsymbol{\phi}) = q(\boldsymbol{\phi}|\boldsymbol{\phi}_p) - h(\boldsymbol{\phi}|\boldsymbol{\phi}_p)$. To find values of $\boldsymbol{\phi} \in \boldsymbol{\Omega}$ which maximize $\ell(\boldsymbol{\phi})$, solutions to

$$\frac{\partial \ell(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \frac{\partial q(\boldsymbol{\phi}|\boldsymbol{\phi}_p)}{\partial \boldsymbol{\phi}} - \frac{\partial h(\boldsymbol{\phi}|\boldsymbol{\phi}_p)}{\partial \boldsymbol{\phi}} = 0$$

are needed.

Now, it is easy to show that $h(\boldsymbol{\phi}|\boldsymbol{\phi}_p) \le h(\boldsymbol{\phi}_p|\boldsymbol{\phi}_p)$ for any $\boldsymbol{\phi} \in \boldsymbol{\Omega}$; that is, $\boldsymbol{\phi}_p$ maximizes $h(\boldsymbol{\phi}|\boldsymbol{\phi}_p)$ with respect to $\boldsymbol{\phi}$, and so $\frac{\partial h(\boldsymbol{\phi}|\boldsymbol{\phi}_p)}{\partial \boldsymbol{\phi}}\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}_p} = 0$. It is therefore sufficient when searching for maxima of $\ell(\boldsymbol{\phi})$ to limit consideration to $\frac{\partial q(\boldsymbol{\phi}|\boldsymbol{\phi}_p)}{\partial \boldsymbol{\phi}}$, specifically, to an enclosure of the gradient of the $q$ function over the box $\boldsymbol{\phi}^I$,

$$\left\{ \frac{\partial q(\boldsymbol{\phi}|\boldsymbol{\phi}_p)}{\partial \boldsymbol{\phi}}\bigg|_{\boldsymbol{\phi}=\boldsymbol{\phi}_p} : \boldsymbol{\phi}_p \in \boldsymbol{\phi}^I \right\}$$

for arbitrary $\boldsymbol{\phi}^I \in \boldsymbol{\Omega}$. Let $Q'(\boldsymbol{\phi}|\boldsymbol{\phi}^I) = [\underline{Q}'(\boldsymbol{\phi}|\boldsymbol{\phi}^I), \overline{Q}'(\boldsymbol{\phi}|\boldsymbol{\phi}^I)]$ be an interval extension of $\frac{\partial q(\boldsymbol{\phi}|\boldsymbol{\phi}_p)}{\partial \boldsymbol{\phi}}$ for interval $\boldsymbol{\phi}^I$ and $\boldsymbol{\phi}_p \in \boldsymbol{\phi}^I$. Note that $Q'(\boldsymbol{\phi}|\boldsymbol{\phi}^I)$ is *not* $\frac{\partial Q(\boldsymbol{\phi}|\boldsymbol{\phi}^I)}{\partial \boldsymbol{\phi}}$ where $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^I)$ is the interval extension of $q(\boldsymbol{\phi}|\boldsymbol{\phi}_p)$. Also, let $Q_2'(\boldsymbol{\phi}^I|\boldsymbol{\phi}^I)$ be an interval function which contains $Q'(\boldsymbol{\phi}|\boldsymbol{\phi}^I)$ for all $\boldsymbol{\phi} \in \boldsymbol{\phi}^I$.

Thus $Q_2'(\boldsymbol{\phi}^I|\boldsymbol{\phi}^I)$ is an interval-valued function which encloses the union of the ranges of a class of interval functions $q(\boldsymbol{\phi}|\boldsymbol{\phi}^I)$ indexed by $\boldsymbol{\phi} \in \boldsymbol{\phi}^I$. At each $\boldsymbol{\phi}_p \in \boldsymbol{\phi}^I$, the enclosure of the gradient of the log-likelihood can be obtained by

$$\frac{\partial \ell(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\bigg|_{\boldsymbol{\phi}=\boldsymbol{\phi}_p} = \frac{\partial q(\boldsymbol{\phi}|\boldsymbol{\phi}_p)}{\partial \boldsymbol{\phi}}\bigg|_{\boldsymbol{\phi}=\boldsymbol{\phi}_p} \in [\underline{Q}'(\boldsymbol{\phi}_p|\boldsymbol{\phi}^I), \overline{Q}'(\boldsymbol{\phi}_p|\boldsymbol{\phi}^I)]$$

and

$$\left\{ \frac{\partial \ell(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\bigg|_{\boldsymbol{\phi}=\boldsymbol{\phi}_p} : \boldsymbol{\phi}_p \in \boldsymbol{\phi}^I \right\} \subset Q_2'(\boldsymbol{\phi}^I|\boldsymbol{\phi}^I).$$

If 0 is not contained in $Q_2'(\boldsymbol{\phi}^I|\boldsymbol{\phi}^I)$, then the box $\boldsymbol{\phi}^I$ cannot contain a local maximizer of $\ell(\boldsymbol{\phi})$ and may therefore be excluded from further consideration.

After a user of this method specifies an initial box $\boldsymbol{\phi}^I \in \boldsymbol{\Omega}$, locating optima of the log-likelihood proceeds by conducting a bisection search by dividing $\boldsymbol{\phi}^I$ into successively smaller boxes and evaluating the enclosure of the gradient of the log-likelihood over each box. Boxes which do not contain a stationary point are discarded. The initial box $\boldsymbol{\phi}^I$ will frequently be quite large so as to (we hope) enclose all stationary points of $\ell(\boldsymbol{\phi})$. At a certain point in this process, typically when the box size becomes smaller than a specified size, the subdividing stops and a list, $\mathcal{G}$, of boxes from the search, is output along with the enclosure of the gradient and the enclosure of the range of $q$ functions over each box. These boxes contain all the stationary points of $\ell(\boldsymbol{\phi})$ that exist within the initial interval box $\boldsymbol{\phi}^I$.

### 4.2 Definitions for Interval EM

In this section an interval EM algorithm is presented. A few necessary definitions are stated and then utilized in the interval EM method being presented.

**Definition 1**. An *interval EM algorithm* on an interval vector $\boldsymbol{\Phi}$ in a parameter space $\boldsymbol{\Omega}$ is an iterative method which employs a sequences of intervals $\boldsymbol{\phi}_0^I \rightarrow \boldsymbol{\phi}_1^I \rightarrow \cdots \boldsymbol{\phi}_p^I \rightarrow$ with respect to interval enclosures $Q(\boldsymbol{\phi}|\boldsymbol{\phi}_0^I)$, $Q(\boldsymbol{\phi}|\boldsymbol{\phi}_1^I), \cdots, Q(\boldsymbol{\phi}|\boldsymbol{\phi}_p^I)$ of sets of functions $q(\boldsymbol{\phi}|\boldsymbol{\phi}_0)$, $q(\boldsymbol{\phi}|\boldsymbol{\phi}_1), \cdots, q(\boldsymbol{\phi}|\boldsymbol{\phi}_p)$ so that $q(\boldsymbol{\phi}|\boldsymbol{\phi}_p) \in Q(\boldsymbol{\phi}|\boldsymbol{\phi}_p^I)$ where $\boldsymbol{\phi}_p \in \boldsymbol{\phi}_p^I \subset \boldsymbol{\Phi}$ for each $p$. The interval $\boldsymbol{\phi}_{p+1}^I$ contains at least one value of $\boldsymbol{\phi}_{p+1}$ which maximizes a $q(\boldsymbol{\phi}|\boldsymbol{\phi}_p)$ for at least one $\boldsymbol{\phi}_p \in \boldsymbol{\phi}_p^I \subset \boldsymbol{\Phi}$. Moving from $\boldsymbol{\phi}_p^I$ to $\boldsymbol{\phi}_{p+1}^I$ is referred to as an *interval EM step.*

**Definition 2**. An *interval GEM algorithm* is an interval EM algorithm except instead of maximizing $q(\boldsymbol{\phi}|\boldsymbol{\phi}_p^I)$ with respect to $\boldsymbol{\phi}$, the interval $\boldsymbol{\phi}_{p+1}^I$ contains as least one value $\boldsymbol{\phi}_{p+1}$ such that $q(\boldsymbol{\phi}_{p+1}|\boldsymbol{\phi}_p) \geq q(\boldsymbol{\phi}_p|\boldsymbol{\phi}_p)$, where $\boldsymbol{\phi}_{p+1} \in \boldsymbol{\phi}_{p+1}^I$, $\boldsymbol{\phi}_p \in \boldsymbol{\phi}_p^I$. Moving from $\boldsymbol{\phi}_p^I$ to $\boldsymbol{\phi}_{p+1}^I$ is referred to as an *interval GEM step.*

The methods described in this section may be more easily understood by referring to Figure 1, which graphically illustrates an interval EM step in a hypothetical one-dimensional case. In Figure 1, the dotted lines $q(\boldsymbol{\phi}|\boldsymbol{\phi}_i)$ and $q(\boldsymbol{\phi}|\boldsymbol{\phi}_j)$ are two separate scalar $q$ functions that might be encountered in different iterations of a scalar EM algorithm. The solid lines $\underline{Q}(\phi|\phi_k^I)$ and $\overline{Q}(\phi|\phi_k^I)$ denote the extent of an interval-valued function $Q(\boldsymbol{\phi}|\boldsymbol{\phi}_k^I)$ which encloses all the scalar $q$ functions $q(\boldsymbol{\phi}|\boldsymbol{\phi}_i)$ and $q(\boldsymbol{\phi}|\boldsymbol{\phi}_j)$ indexed by $\boldsymbol{\phi}_i \in \boldsymbol{\phi}_k^I$, $\boldsymbol{\phi}_j \in \boldsymbol{\phi}_k^I$. Finally, the vertical line segments $Q(\boldsymbol{\phi}_{k+1,1}^I|\boldsymbol{\phi}_{k+1,1}^I)$ and $Q(\boldsymbol{\phi}_{k+1,2}^I|\boldsymbol{\phi}_{k+1,2}^I)$ denote enclosures of $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^I)$ evaluated for $\boldsymbol{\phi} \in \boldsymbol{\phi}_{k+1,1}^I$ and $\boldsymbol{\phi} \in \boldsymbol{\phi}_{k+1,2}^I$, respectively.

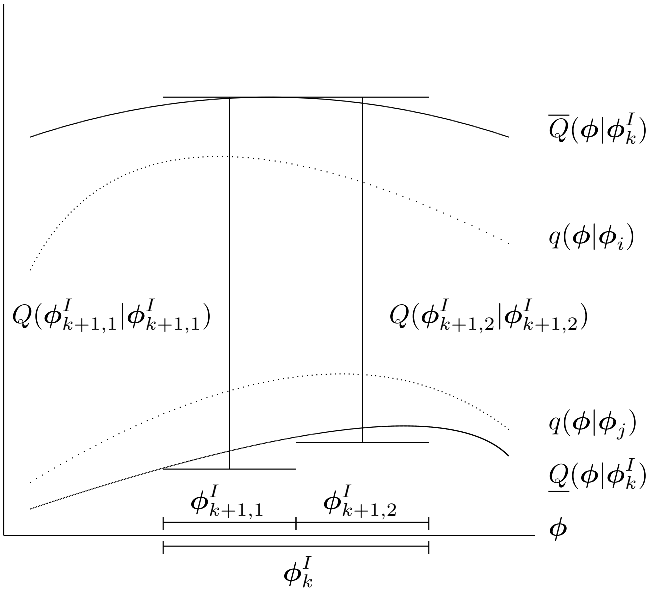
$\overline{Q}(\boldsymbol{\phi}|\boldsymbol{\phi}_k^I)$
$q(\boldsymbol{\phi}|\boldsymbol{\phi}_i)$
$Q(\boldsymbol{\phi}_{k+1,1}^I|\boldsymbol{\phi}_{k+1,1}^I)$
$Q(\boldsymbol{\phi}_{k+1,2}^I|\boldsymbol{\phi}_{k+1,2}^I)$
$q(\boldsymbol{\phi}|\boldsymbol{\phi}_j)$
$\underline{Q}(\boldsymbol{\phi}|\boldsymbol{\phi}_k^I)$
$\boldsymbol{\phi}_{k+1,1}^I$
$\boldsymbol{\phi}_{k+1,2}^I$
$\boldsymbol{\phi}$
$\boldsymbol{\phi}_k^I$

Figure 1: One interval EM step. $\overline{Q}(\phi|\phi_k^I)$ and $\underline{Q}(\phi|\phi_k^I)$ bound the extent of the interval-valued function $Q(\phi|\phi_k^I)$, while $q(\phi|\phi_i)$ and $q(\phi|\phi_j)$ are examples of two of the scalar functions contained within the interval function.

A fundamental difference between the interval EM algorithm and the scalar EM algorithm is that the interval EM algorithm combines the E-step and M-step, and deals directly with the problem of finding the stationary points. It is iterative in the sense that boxes are examined, cut in half, and re-examined; not in the sense of alternating between the E and M steps.

### 4.3 Full Bisection Search

The bisection algorithm starts by putting an initial box $\boldsymbol{\phi}_0^I$ in a list of boxes $\mathcal{G}$. The method proceeds by simply bisecting boxes from $\mathcal{G}$ until no boxes are left or until all boxes have reached a certain size. Let $m$ be the dimension of $\boldsymbol{\phi}$ and initialize $i := 1$. Proceed by removing and bisecting each box of $\mathcal{G}$ in the $i^{th}$ coordinate. Discard any boxes which do not contain 0 in at least one direction of the enclosure of the gradient. Return all remaining boxes to $\mathcal{G}$ and increase $i$ by 1, resetting $i := 1$ when $i > m$. Repeat as necessary until the diameter of every box is small. If at any point $\mathcal{G}$ becomes empty, print a message stating that no stationary points were contained in the initial region $\boldsymbol{\phi}_0^I$.

The bisection algorithm differs from traditional EM in that there are no expectation and maximization steps. The only use of the EM theory is to obtain an enclosure for the gradient of the log-likelihood of $\boldsymbol{\phi}$. Still, the bisection search can in some way be viewed as many simultaneous interval GEM algorithms. In making an interval GEM step from $\boldsymbol{\phi}_k^I$ to $\boldsymbol{\phi}_{k+1,1}^I$ and from $\boldsymbol{\phi}_k^I$ to $\boldsymbol{\phi}_{k+1,2}^I$, there will be a nondecreasing change in the lower bound of the enclosure of the $q$ functions; that is, $\underline{Q}(\boldsymbol{\phi}_k^I|\boldsymbol{\phi}_k^I) \leq \underline{Q}(\boldsymbol{\phi}_{k+1,i}^I|\boldsymbol{\phi}_{k+1,i}^I)$ for $i = 1, 2$. The method is now summarized in the following algorithm.

- ALGORITHM: Bisection Interval EM Search

```
Input an initial interval box $\boldsymbol{\phi}_0^I$ and place it as the only element of the list $\mathcal{G}$.
i := 0
REPEAT
    i := (i + 1) MOD m
    FOR j = 1 TO LENGTH($\mathcal{G}$)
        Remove the first box from $\mathcal{G}$. Call it $\boldsymbol{\phi}^I$
        Bisect $\boldsymbol{\phi}^I$ along the $i^{th}$ direction, creating $\boldsymbol{\phi}_1^I$ and $\boldsymbol{\phi}_2^I$
        If $0 \in Q_2'(\boldsymbol{\phi}_k^I|\boldsymbol{\phi}_k^I)$, append $\boldsymbol{\phi}_k^I$ to $\mathcal{G}$, $k = 1, 2$
    NEXT
UNTIL $\mathcal{G}$ is empty or maximum diameter of boxes $\leq \epsilon$
```

The method described above will not, of course, find any global optima which lie outside of the initial box $\boldsymbol{\phi}_0^I \subset \boldsymbol{\Omega}$. In practice this is often not of concern, primarily because the observed data places practical limitations on the portion of the parameter space of interest. Also, in a manner similar to that observed by Hansen (1992), it is often possible to make the parameter space exceedingly large without significantly increasing the computing time to search for global optima. This can happen when the stationary points are clustered in a small portion of the parameter space (relative to the initial box). When the initial box is bisected, it will often be possible to discard one of the resulting halves. With each iteration of the interval EM algorithm, the parameter space is halved and the length of the list $\mathcal{G}$ remains the same. The univariate $t$ example in Section 5 illustrates this behavior.

Because the algorithm uses intervals instead of real numbers, measurement error in data and floating-point approximations can immediately be incorporated. For example, one might use $\pi^I = [3.14, 3.15]$ to indicate uncertainty in known constants. Even more useful is the ability to represent data as intervals, e.g. $x_i^I = [x_i - \delta, x_i + \delta]$, where $x_i$ is the observed value and $\delta$ is a bound on the measurement error.

The speed of the algorithm depends on the complexity of the data under consideration. Let $m$ be the dimension of $\boldsymbol{\phi}$. Bisection of just one box from the list $\mathcal{G}$ has the potential to create $2^m$ additional boxes that will be appended to $\mathcal{G}$. This might happen in situations where a region contains many stationary points or where the log-likelihood is relatively flat and the gradient is near zero. Since interval arithmetic sometimes calculates an interval wider than optimal, it may be the case that the gradient is nonzero in every direction, but the enclosure of the gradient contains zero in at least one direction. If some combination of high dimensionality and/or a fairly flat likelihood occurs, the length of $\mathcal{G}$ can grow exponentially, a situation in which the efficiency of the computations becomes important. As indicated by Knüppel (1994), interval calculations in the PROFIL package require slightly more than twice as much time as ordinary floating point calculations (using a standard C compiler). The examples in this article each required less than one second of real time on a DEC Alpha 400 workstation.

# 5 EXAMPLES

Several examples are now presented to illustrate use of the interval EM algorithm described above. Note that the following examples each have an algebraic, real expression for $q(\boldsymbol{\phi}|\boldsymbol{\phi}_k)$. This is consistent with traditional EM notation. Though not shown, a person would determine an expression for the gradient of this function with respect to $\boldsymbol{\phi}$, $q'(\boldsymbol{\phi}|\boldsymbol{\phi}_k)$, and then express $q'(\boldsymbol{\phi}_k|\boldsymbol{\phi}_k)$ in as simple a way as possible. This is coded in the program as $Q_2'(\boldsymbol{\phi}_k|\boldsymbol{\phi}_k)$.

When numerical results are reported, sub/superscript notation will sometimes be used to simplify the representation of an interval, e.g. $[2.33, 2.35] = 2.3_3^5$.

## 5.1 Multinomial (continued)

For the multinomial example in Section 3.1, it can be shown that

$$q(p|p_k) = k(\mathbf{x}) + \left[125\frac{\frac{p_k}{4}}{\frac{1}{2}+\frac{p_k}{4}} + x_5\right]\frac{1}{p} - (x_3 + x_4)\frac{1}{1-p} \tag{1}$$

where $k(\mathbf{x})$ does not depend on $p$ and can be ignored in the maximization step. Figure 2 shows a plot of the corresponding interval extension, $Q(p|p_k^I)$. An accurate interpretation of this interval-valued function can be obtained in this case by actually overlaying plots of $q(p|p_k)$ for various $p \in p_k^I$, in this case $p = 0.1(0.1)0.9$.

The initial interval selected is $p_0^I = [.00001, .99999]$. Although a wider interval can be used, the maximum likelihood estimate of $p$ is certainly contained in $[.00001, .99999]$. Furthermore, the values of $p = 0$ and $p = 1$ are excluded by the gradient of Equation (1). If the user selects an inappropriate value for the initial interval, such as $[0.1, 0.2]$, then the program terminates with the message:

```
Gradient of Q(Phi|Phi_k) = ([152.262,411.414])
Gradient of likelihood does not contain zero.
No stationary point in ([0.1,0.2])
```

The bisection algorithm applied to this problem using an initial interval $p_0^I$ produces a list $\mathcal{G}$ which contains two interval boxes,

$$y_1 = 0.626821497870982_3^4$$

$$y_2 = 0.626821497870982_4^5.$$

Any stationary points of the log-likelihood are guaranteed to be contained in the hull of the boxes in the list $\mathcal{G}$. If a scalar estimate is desired, the midpoint of the hull can be given: $\hat{p} = 0.6268214978709824$.

## 5.2 Univariate t

McLachlan and Krishnan (1997) give an example by Arslan et al. (1993) where the EM algorithm can converge to a local *minimum*. A $p$-dimensional random variable $\mathbf{W}$ is said to have a multivariate $t$-distribution $t_p(\boldsymbol{\mu}, \Sigma, \nu)$ with location $\boldsymbol{\mu}$, positive definite inner product matrix $\Sigma$, and degrees of freedom $\nu$ when the density of $\mathbf{W}$ is given by

$$f_p(\mathbf{w}|\boldsymbol{\mu}, \Sigma, \nu) = \frac{\Gamma(\frac{p+\nu}{2})|\Sigma|^{-1/2}}{(\pi\nu)^{p/2}\Gamma(\frac{\nu}{2})\{1 + (\mathbf{w}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{w}-\boldsymbol{\mu})/\nu\}^{(p+\nu)/2}}. \tag{2}$$

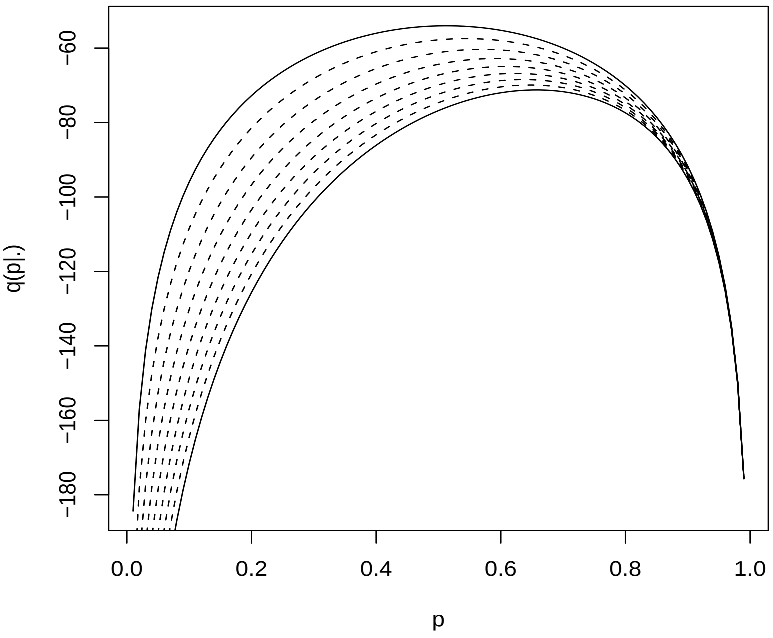

Figure 2: Plot of $Q(p|p_0^I)$ versus $p$ for $p_0^I = [0.1, 0.9]$. Solid lines denote the extent of the interval-valued function $Q(p|p_0^I)$.

The example considered is a univariate case of the $t$-distribution where $\nu = 0.05$, $\Sigma = 1$, and $\mu$ is taken as unknown. The observed data is $\mathbf{w} = (-20, 1, 2, 3)$. Ignoring additive and multiplicative constants, the log-likelihood is $\log L(\mu) \propto -\sum_i \log\{1 + 20(w_i - \mu)\}$. A plot showing the shape of this log-likelihood appears in Figure 3.

The function has seven stationary points. The most interesting are the local maxima at $\mu_2 = 1.086$, $\mu_3 = 1.997$, and $\mu_4 = 2.906$. In this complete-data problem it is possible to graph the log-likelihood and visually choose starting values that will cause a scalar EM algorithm to converge to each of the local maxima, *and even to a local minimum*. However, the domain of attraction for each stationary point is not necessarily a contiguous region.

Using $\mu_0 = [-1000, 1000]$, the bisection algorithm completes 59 iterations (bisections), during which the length of $\mathcal{G}$ is scarcely longer than the 20 boxes at the final step. These boxes occur in distinct groupings around each of the seven stationary points. Although the algorithm actually outputs the list of boxes from $\mathcal{G}$, for brevity the hull of each group of boxes and the hull of the associated enclosures of the $q$ functions are given in Table 1.

Looking at this table, the nature of each stationary point is not immediately clear. Since this is a univariate case, it would be possible to evaluate the gradient on either side of each $\boldsymbol{\phi}_{S_i}$ and thereby determine which stationary points are local maxima and which are local minima. However, it is immediately clear from the table that $\boldsymbol{\phi}_{S_5}$ gives the largest value of $Q(\boldsymbol{\phi}_{S_i}|\boldsymbol{\phi}_{S_i})$ and contains the global maximum of the log-likelihood as displayed by Figure 3.

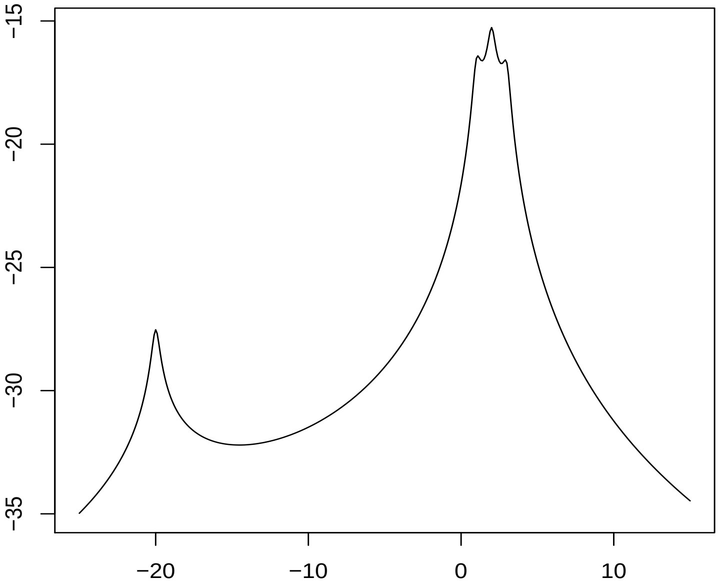

Figure 3: Plot of log-likelihood function $\log L(\mu)$ versus $\mu$. Local maxima occur at $\mu_1 = -19.993$, $\mu_2 = 1.086$, $\mu_3 = 1.997$, and $\mu_4 = 2.906$

Table 1: Enclosures of the stationary points for the univariate $t$ example

| $i$ | $\boldsymbol{\phi}_{S_i}$ | $Q(\boldsymbol{\phi}_{S_i}\vert\boldsymbol{\phi}_{S_i})$ |
|---|---|---|
| 1 | $-19.993164608871^{29}_{30}$ | $-1.575326662795595^{4}_{7}$ |
| 2 | $-14.5161774794253^{0}_{2}$ | $-2.098837787645^{297}_{302}$ |
| 3 | $1.08616780631075^{7}_{0}$ | $-1.606093870388^{397}_{426}$ |
| 4 | $1.3731761015634^{32}_{18}$ | $-1.89224275084^{2981}_{3016}$ |
| 5 | $1.9975126089118^{24}_{17}$ | $-1.525009886703^{386}_{402}$ |
| 6 | $2.6468546770426^{35}_{20}$ | $-1.884158362286^{176}_{208}$ |
| 7 | $2.9056308944679^{85}_{75}$ | $-1.617024174245^{677}_{707}$ |

This suggests one possible way that the algorithm could be accelerated. Suppose that only the stationary point $\boldsymbol{\phi}$ with the largest value of $q(\boldsymbol{\phi}|\boldsymbol{\phi})$ (among all stationary points) was of interest. Let $\boldsymbol{\phi}_j^I$ and $\boldsymbol{\phi}_k^I$ be boxes (remaining to be processed) on the list $\mathcal{G}$. If $\overline{Q}(\boldsymbol{\phi}_k^I|\boldsymbol{\phi}_k^I) < \underline{Q}(\boldsymbol{\phi}_j^I|\boldsymbol{\phi}_j^I)$, then $\boldsymbol{\phi}_k^I$ could be omitted from further consideration.

### 5.3 Binomial-Poisson Mixture

This example from Thisted (1988) presents a simple multivariate-parameter example dealing with the number of children per widow in a pension fund.

| Children per widow, $i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Number of widows, $n_i$ | 3062 | 587 | 284 | 103 | 33 | 4 | 2 |

Since the number of widows with no children is larger than would be expected for a Poisson distribution, it is assumed that there are actually two underlying populations. The number of children, $Y$, for a widow is modeled as:

$$Y \sim \begin{cases} 0 & \text{with probability } \xi \\ Poisson(\lambda) & \text{with probability } 1-\xi. \end{cases} \tag{3}$$

With $\boldsymbol{\phi} = (\lambda, \xi)$, the function to be maximized in the M-step is:

$$q(\boldsymbol{\phi}|\boldsymbol{\phi}_k) = \frac{n_0 \xi_k}{\xi_k + (1-\xi_k)\exp(-\lambda_k)} \left\{ \log \xi - \log(1-\xi) + \lambda \right\} + \\ N\left\{\log(1-\xi) - \lambda\right\} + \sum_{i=1}^{6} \left\{ i n_i \log \lambda - n_i \log i! \right\}. \tag{4}$$

Based on a visual examination of the data, the starting values of

$$\boldsymbol{\phi}_0 = (\lambda_0^I, \xi_0^I) = ([0.001, 10], [0.001, 0.999])$$

were chosen as being certain to contain the true parameter values.

Applying the Bisection search, after 52 iterations of bisecting $\boldsymbol{\phi}$ in both directions, the list $\mathcal{G}$ contains 82 boxes, the first and last of which are:

$$y_1 = (1.0378390789897_{57}^{60}, 0.615056697573 12_{12}^{14})$$
$$y_{82} = (1.0378390789897_{77}^{80}, 0.61505669757312_{88}^{90}).$$

The hull of the boxes on this list is: $\boldsymbol{\phi}_{S_1} = (1.0373890789897_{57}^{80}, 0.61505669757312_{12}^{90})$.

In this problem, what is important is not the extremely narrow and high degree of accuracy of $\boldsymbol{\phi}_{S_1}$, but the guarantee that considered over the initial parameter space $\boldsymbol{\phi}_0$, the only stationary points of the log-likelihood (if any exist) are guaranteed to be contained in the interval box $\boldsymbol{\phi}_{S_1}$. Moreover, if a scalar EM algorithm converges to some stationary point in $\boldsymbol{\phi}_0$, that point will be inside $\boldsymbol{\phi}_{S_1}$.

### 5.4 Genetic example

This example is also taken from McLachlan and Krishnan (1997). Suppose there are 435 observations from a multinomial distribution as given in Table 2. where $r = 1 - p - q$. The observed data are the cell frequencies of blood types $(n_\mathrm{O}, n_\mathrm{A}, n_\mathrm{B}, n_{AB})$ believed to be determined (genetically) by the unknown parameters $\boldsymbol{\phi} = (p, q)$. As in the multinomial example above, a natural way to introduce missing data is to split the A and B cells across the sum in the cell probability. For this model, the $q$ function is

$$q(\boldsymbol{\phi}|\boldsymbol{\phi}_k) = \left( \frac{182}{1 + 2(1-p_k-q_k)/p_k} + 199 \right) \log(p) + \\ \left( \frac{60}{1 + 2(1-p_k-q_k)/q_k} + 77 \right) \log(q) + \\ \left( 594 - \frac{182}{1 + 2(1-p_k-q_k)/p_k} - \frac{60}{1 + 2(1-p_k-q_k)/q_k} \right) \log(1-p-q). \tag{5}$$

Table 2: Distribution of data in the genetic example

| Cell | Cell<br>Probability | Observed<br>Frequency |
|---|---|---|
| O | $r^2$ | $n_{\mathrm{O}} = 176$ |
| A | $p^2 + 2pr$ | $n_{\mathrm{A}} = 182$ |
| B | $q^2 + 2qr$ | $n_{\mathrm{B}} = 60$ |
| AB | $2pq$ | $n_{\mathrm{AB}} = 17$ |

It is not always possible to search the entire portion of the parameter space with one application of the bisection algorithm. In this example, certain combinations of $p^I$ and $q^I$ cause a division by zero error. Specifically, the gradient does not exist along the lines $p = 0$, $q = 0$, $1 - p - q = 0$, $q = 2 - 2p$, and $2q = 2 - p$. The software can be written to catch *division by zero* errors and mark a box as containing such until further subdivision occurs. Alternatively, the user can specify a smaller initial region. The only stationary point located inside $\boldsymbol{\phi}_0 = (p_0, q_0) = ([0.00001, 0.45], [0.00001, 0.45]$ is found to be located inside $\boldsymbol{\phi}_S = (0.2644443138466^{706}_{694}, 0.0931688118156 8^{200}_{122})$.

## 6 CONCLUSIONS

Interval analysis first gained noticeable development in the 1960s by R. E. Moore. Two monographs suitable for an introduction to the subject are Moore (1966, 1979). Interval analysis has a fairly extensive literature in some areas, e.g. global optimization, but has seen little development in statistical settings. This article takes a step toward expanding the current state of knowledge by using interval analysis together with ideas from the EM algorithm. The resulting method is capable of finding multiple stationary points of a log-likelihood to a high degree of accuracy. The EM algorithm cannot be relied upon to do this. Unlike other algorithms for optimization, the method retains the ability of the EM algorithm to handle missing-data problems.

## ACKNOWLEDGEMENTS

The authors thank the referees and associate editor for helpful comments which led to improvements in the article. This work was partially supported by National Science Foundation grant DMS-9500831.

*[Received January 1999. Revised October 1999.]*

## REFERENCES

Arslan, O., Constable, P. D. L., and Kent, J. T. (1993). Domains of Convergence for the EM Algorithm: A Cautionary Tale in a Location Estimation Problem. *Statistical Computing*, 3:103–108. 9

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood From Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B. Methodological*, 39:1–22. 2, 3, 5

Hansen, E. (1992). *Global Optimization Using Interval Analysis*. Marcel Dekker Inc., New York. 1, 2, 8

——— (1997). Sharpness in Interval Computations. *Reliable Computing*, 3:17–29. 2

Kearfott, R. B. (1996). *Rigorous Global Search: Continuous Problems*. Kluwer Academic Publishers. 2

Knüppel, O. (1993). PROFIL – Programmer's Runtime Optimized Fast Interval Library. Technical Report 93.4, Informationstechnik, Technische Uni. Hamburg–Harburg. 2

——— (1994). PROFIL/BIAS - A Fast Interval Library. *Computing*, 53:277–287. 8

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. John Wiley & Sons. 9, 12

Moore, R. E. (1966). *Interval Analysis*. Prentice-Hall, Englewood Cliffs, N.J. 13

Moore, R. E. (1979). *Methods and Applications of Interval Analysis*. SIAM, Philadelphia. 1, 13

Thisted, R. A. (1988). *Elements of Statistical Computing*. Chapman and Hall. 12

Wu, C. F. J. (1982). On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95–103. 1