Estimation of multiple parameters using algorithms expectation maximization

R Ospina¹, G Correa², and P P Cárdenas²
¹ Universidad Industrial de Santander, Bucaramanga, Colombia
² Universidad Tecnológica de Pereira, Pereira, Colombia

E-mail: rospinao@saber.uis.edu.co

Abstract. In this paper the strategy of the expectation maximization algorithm is developed in order to solve the problem of maximization of the likelihood which becomes a complex problem when it is expected to estimate multiple parameters at the same time. All this will be achieved with the introduction of basic concepts of statistical estimators and will be verified by simulations carried out in the MATLAB software.

1. Introduction
Statistical inference is a type of reasoning that proceeds from the concrete to the general: it tries to draw conclusions about the parameters of a population from the information contained in the statistics of a sample of that population [1]. Statistical inference assumes that samples of a system are available and that we want to know the characteristics that describe it. These values of the characteristics are known as parameters, which can be calculated taking all possible samples of the system analyzed, a task that is inefficient in practice. Another alternative to determine these parameters is to make use of the estimation of them through a sample, a task that is commonly used in everyday life although it may arrive to present errors since these assumptions do not necessarily coincide with the real parameters of the system. Therefore, there are different parameter estimation techniques which can be classified into two groups, point estimation and interval estimation, the difference between both is that the first one delivers a point value of the parameter while the second offers a range of values of this same. The estimation of these parameters is achieved by means of mathematical functions that allow obtaining an approximation of the real value of the parameter from the data of the samples. The mathematical function that meets the following conditions is considered the best estimator.

Absence of bias: An estimator is unbiased when the average of the estimates obtained in different samples is exactly the value of the parameter that is intended to be estimated. However, in practice it is impossible to obtain a direct measurement of said value, since the actual value of the parameter is not available.

Efficiency: It can be considered that an estimator is efficient in absolute sense if it has the lowest mean square error (MSE). The efficiency of the estimator is related to the minimum variance, from which the lower limit can be obtained using the Cramer-Rao criterion [2]. Estimators that meet these two conditions are known as unbiased least variance estimators or minimum variance unbiased (MVU). In general, these estimators do not have to exist.

The mathematical functions mentioned above follow a probability density function (PDF) given as a function of the data set \( x = \{x[n]\}_{n=0}^{N-1} \) and its parameters \( \theta \) in the following way, \( p(x \mid \theta) \). Among the
most commonly used linear estimation methods in practice are the method of moments, the least squares method, the analogy method, the maximum likelihood method, among others. In this document, the maximum likelihood method will be treated, where the basic idea will be to find the value of the parameters that maximizes the likelihood. Therefore, optimization algorithms must be applied, among which the descending gradient algorithm is used as an alternative. Raphson, maximum likelihood estimator (MLE) scoring and the EM algorithm. The latter will be the algorithm by which the optimal values of the parameters that maximize the likelihood will be searched.

2. Estimators of maximum plausibility
In the problem of parameter estimation, as mentioned in the previous section, one of the methods that allow us to find the MVU estimator is the MLE. This estimation technique is one of the most common that allows obtaining practical estimators and that can also be used in complex estimation problems or where the MVU does not exist or cannot be found. Its characteristics compared to other estimators are desirable if the data set that is available is very large, these are: it is asymptotically efficient, is consistent and it is invariant before re-parameterizations. In many cases it is not possible to find a direct formula of the MLE estimator, so it must be searched using numerical optimization methods.

2.1. Modeling the data
The specification of the PDF is key to the determination of the correct estimator, it is to clarify, that in the real practical problems the PDF of the data is unknown, so it must be chosen in a way that is consistent with the restrictions of the problem, reflect prior knowledge of the data and is mathematically treatable. It is common to choose a PDF standard with white gaussian noise (WGN), which makes the PDF mathematically treatable that can lead to estimators whose formula is expressed in some cases directly or closed. The performance of the estimator is strongly dependent on the assumptions of the PDF of the data. When selecting a PDF, the estimator is expected to be robust, that is, small changes in the PDF of the data do not affect the estimator’s performance too much.

3. Unbiased estimators
An estimator of a certain unknown parameter is unbiased if, on average, it leads to the true value of the parameter (Equation 1), that is, if \( \hat{\theta} \) is an estimator of parameter \( \theta \in (a, b) \) it is unbiased if:

\[
E(\hat{\theta}) = \theta, \forall \theta \in (a, b)
\]

The restriction of Equation (1) is very important, it means that if \( \hat{\theta} = g(x) \), with \( x = [x[0], x[1], ..., x[N - 1]]^{T} \) it has to fulfil that (Equation 2),

\[
E(\hat{\theta}) = \int_{-\infty}^{\infty} g(x)p(x; \theta)dx = \theta \forall \theta
\]

This condition could occur only for some values of \( \theta \) but not for others. On the other hand, an estimator that is unbiased does not mean that it is a good estimator. It only indicates that on average it reaches the true value of the parameter. The biased estimators introduce a systematic error in the estimate, and their performance is low, however, in some cases the biased estimators have less variability.

3.1. Minimum variance criterion
In the search for optimal estimators it is necessary to adopt some optimality criterion. One widely used in practice is the minimization of MSE, defined as Equation (3).

\[
mse(\theta) = E[(\hat{\theta} - \theta)^{2}]
\]
By developing Equation (3) we get:

$$\text{mse}(\hat{\theta}) = \text{var}(\hat{\theta}) + b^2(\theta)$$  \hspace{1cm} (4)

From the Equation (4) we can see that it is not possible to find an estimator that minimizes the one since it depends on the unknown parameter that is expected to estimate. Therefore, this approach of minimizing it must be abandoned since it leads to unrealizable estimators. The alternative is to restrict to unbiased estimators and minimize variance.

### 3.2. Minimum variance unbiased estimators

It is said that there is an MVU estimator if there is an estimator of less variance than the rest of the possible estimators for all value of $\theta$. In general, the MVU estimator does not have to exist or it may be the case that there is not a single MVU. In this case the search for it will not make sense since there is no fallible rule to find it, however, there are some search approaches:

- Determine the lower bound of Cramér-Rao lower bound (CRLB), and test if any estimator satisfies it for all values of $\theta$.
- Look for sufficient statistics and apply the Rao-Blackwell-Lehmann-Scheffé (RBLS) theorem. There may be an MVU that does not reach the CRLB.
- Restrict the class of estimators not only to the unbiased, but also to the linear unbiased with the data and find the MVU in this class. It must be taken into account that this estimator will not be optimal unless the MVU estimator is linear in that particular problem.

The likelihood function is the PDF of the parameter setting the value of the data. Conceptually, it indicates the probability of the unknown parameter after observing the data. If the PDF of the data is $x \sim p(x; \theta)$ then the MLE estimator is the Equation (5).

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta \in D_{\theta}} \ln p(x; \theta)$$ \hspace{1cm} (5)

The MLE estimator is defined as the value of $\theta$ that maximizes the logarithm of $p(x; \theta)$ by setting $x$. By performing the optimization procedure for Equation (5) we have the Equation (6):

$$\frac{\partial \ln p(x; \theta)}{\partial \theta} = I(\theta)(g(x) - \theta)$$ \hspace{1cm} (6)

By the Cramer-Rao theorem, if there is an efficient estimator, there are the functions $I(\theta)$ and $g(x)$, however, in most cases it is very difficult to reach this factorization, so it is necessary to use numerical methods that allow to find the optimal value of $\theta$.

### 4. Algorithm expectation maximization

The expectation maximization (EM) algorithm has become a very common technique in practice to obtain groupings in large databases. There are two main applications of the EM algorithm. The first is when the data have missing values derived from the observation process and the second is given by the possibility of pattern estimation. This algorithm was first presented by Dempster, Laird & Rubin in 1977 [3]. Due to its implementation, the EM algorithm can be seen as a metaheuristic method, ensuring that it is, with some probability, a local maximum of the registry of the data. EM provides a statistical approach to the problem by looking for the most probable “cluster” number given the data.

The basis of this type of grouping is found in a statistical model called a mixture of distributions, where each distribution represents the probability that an object has a particular set of attribute-value pairs. The EM algorithm is an iterative procedure that calculates the maximum probability (MP)
estimated in the presence of missing or hidden data. Each iteration of the EM algorithm consists of two processes: The E-step and the M-step. In the expectation, or the E-step, the missing data are calculated with the observed data and the current estimation of the parameters model. This archiving is done using the conditional expectation. In the M-step, the probability function is maximized under the assumption that the missing data are known. The estimate of the missing data from the E-step is used in place of the actual missing data. The algorithm stops the iterations when convergence occurs, this being achieved when calculating the likelihood (relationship between elements), this is less than a predefined value [3,4].

4.1. Properties
In order to give a quick idea of the potential of the algorithm as a useful tool in statistical estimation problems, we will summarize below the reasons for its attractiveness. We will also mention the inconveniences that it presents. Some of the advantages of the EM algorithm are the following:

- It is numerically stable with each iteration EM, that is to say, in each iteration the likelihood increases (except arrived at the point of convergence) [5].
- Under fairly general conditions, we can trust that the EM algorithm will reach convergence. That is, starting from an arbitrary point $T$ in the space of the parameters, we will end up in most cases in a local maximum, except for the very inadequate choices of $T$ or some local pathology of the likelihood function [6].
- The EM algorithm requires little storage space and can usually be carried out on a simple computer. For example, you do not have to store the information matrix or its inverse in any iteration.
- Since the problem of complete data is more or less standard, the M-step can be carried out frequently using standard statistical packages in situations where the maximum likelihood estimate of complete data does not exist in a closed form. In other situations of this type, extensions of the EM algorithm such as Greedy EM (GEM) and the conditional expectation maximization algorithms (ECM) usually allow the M-step to be implemented in a relatively simple manner. In addition, these extensions share the stable monotonic convergence of the EM algorithm [7,8].
- The EM algorithm can be used to provide estimates of the values of the lost data.

Some of the disadvantages suffered by the EM algorithm are the following:

- Does not have an included procedure to provide an estimate of the covariance matrix of the parameter estimates. In any case, this disadvantage can easily be eliminated by employing a suitable methodology associated with the EM algorithm [9].
- It can converge in a desperately slow way on problems where there is too much incomplete information and even on seemingly innocuous problems.

Suppose a complete set of data $Z$ will assume that it can be decomposed into two components, the observed component $Z^0$ and the missing component $Z^m$. With $p(Z|\theta)$, the probability density function that generated the data is specified. Therefore, you can write the Equation (7):

$$p(Z|\theta) = p(Z^0, Z^m|\theta) = p(Z^m, Z^0|\theta)p(Z^0|\theta)$$

(7)

If you take the logarithm of the likelihood, we have the Equation (8).

$$\ln p(Z|\theta) = \ln p(Z^m, Z^0|\theta) + \ln p(Z^0|\theta)$$

(8)
The interest is to optimize the parameters with respect to the observed part, therefore (Equation (9)).

\[ \ln p(Z^0|\theta) = \ln p(Z|\theta) - \ln p(Z^m, Z^0|\theta) \]  

(9)

Taking into account that \( L(\theta|Z) = p(Z|\theta) \) then Equation (10).

\[ l(\theta|Z^0) = \ln p(Z|\theta) - \ln p(Z^m, Z^0|\theta) \]  

(10)

The EM algorithm tries to find a value of \( \theta \) that maximizes \( p(Z^0|\theta) \) given an observation \( Z^0 \). The function \( p(Z^m, Z^0|\theta) \) plays an important role, since it relates the probability function of the complete set. In addition, fixed the parameters allows to estimate a value of \( Z^m \), and fixed \( Z^m \) allows to verify how well the parameters are adjusted. Since the values of the lost component are not known, it will be necessary to calculate expectations for the functions as the Equation (11) [10]:

\[ l(\theta|Z^0) = \int \ln p(Z^0, Z^m|\theta) p(Z^m|Z^0, \theta) dZ^m - \int \ln p(Z^m|Z^0, \theta) p(Z^m|Z^0, \theta) dZ^m \]  

(11)

The functions \( G(\theta|\theta^k) \) and \( H(\theta|\theta^k) \) are introduced to give greater clarity the Equation (12).

\[ l(\theta|Z^0) = Q(\theta|\theta^k) - H(\theta|\theta^k) \]  

(12)

The four steps of the EM algorithm are the following [11,12]: Set \( k = 0 \) and initialize \( \theta \) with an arbitrary value \( \theta^k \); step E: calculate \( Q(\theta^{k+1}|\theta^k) \); step M: find \( \theta^{k+1} \) such that \( G(\theta|\theta^k) \) is maximum; if \( \theta^{k+1} = \theta^k \) increase \( k \) and go back to step 2.

5. Results

In this section, we will apply the concept of maximum likelihood using the algorithm for the following model (Equation (13)).

\[ X[n] = \sum_{i=1}^{P} \cos(2\pi f_i n) + w[n], \text{con } w[n] \sim N(0, \sigma^2) \]  

(13)

Where \( f_i \) are the \( P \) parameters to estimate, then calculating the logarithm of the likelihood we have the Equation (14) [10].

\[ \ln(p(X, \theta)) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - \sum_{i=1}^{P} \cos(2\pi f_i n))^2 \]  

(14)

The EM algorithm simplifies the problem through a complete set \( y_i[n] = \cos(2\pi f_i n) + w_i[n] \). Each part of the complete set must be independent of the others, that is \( y_i[n] \perp y_j[n] \). There must be a function \( g(\cdot) \) that allows you to relate the complete set to the incomplete set \( x[n] \), as follows (Equation (15) to Equation (17)):

\[ x[n] = g(y_1[n], y_2[n], ..., y_P[n]) = \sum_{i=1}^{P} y_i[n] \]  

(15)

\[ p(X, \theta) = \prod_{i=1}^{P} p(y_i, \theta_i) = \prod_{i=1}^{P} \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2_i}} \exp\left(-\frac{(y_i[n]-\cos(2\pi f_i n))^2}{2\sigma^2_i}\right) \]  

(16)
\begin{align*}
\ln(p(Y; \theta)) &= \sum_{i=1}^{p} \left[ -\frac{N}{2} \ln(2\pi\sigma^2) - \sum_{n=0}^{N-1} \frac{(y_i[n] - \cos(2\pi f_i n))^2}{2\sigma^2} \right] = \\
g(Y) + \sum_{i=1}^{p} \frac{1}{2\sigma_i^2} \sum_{n=0}^{N-1} (y_i[n] \cos(2\pi f_i n) - \frac{1}{2} \cos^2(2\pi f_i n)) \quad (17)
\end{align*}

Where \( g(Y) \) is a function that does not depend on frequencies. Now, doing the approximation \( \sum_{n=0}^{N-1} \cos^2(2\pi f_i n) \approx \frac{N}{2} \) we have Equation (18):

\begin{align*}
\ln(p(Y; \theta)) &= h(Y) + \sum_{i=1}^{p} \frac{1}{2\sigma_i^2} \sum_{n=0}^{N-1} (y_i[n] \cos(2\pi f_i n)) \\
\quad (18)
\end{align*}

Taking \( \mathbf{c}_i = [1 \; \cos(2\pi f_i) \ldots \; \cos(2\pi f_i (N - 1))]^T \) we obtain the Equation (19).

\begin{align*}
\ln(p(Y; \theta)) &= h(Y) + \sum_{i=1}^{p} \frac{1}{2\sigma_i^2} \mathbf{c}_i^T \mathbf{y}_i \\
\quad (19)
\end{align*}

Optimizing the Equation (19) and applying the EM algorithm computationally, we obtain the following results with \( P = 3 \) that is \( f_i = \left[ \frac{1}{20} \; \frac{1}{30} \; \frac{1}{40} \right] \).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{(a) Function \( x[n] \); (b) component \( y_1 \); (c) component \( y_2 \); (d) component \( y_3 \); (e) estimator for the frequency \( f_1 \); (f) estimator for the frequency \( f_2 \); (g) estimator for the frequency \( f_3 \).}
\end{figure}
In Figure 1(a) you can see 3 different functions, the first in red line represents the data without noise, the second black line shows the data, but taking into account the noise and the blue line shows the estimation of said data once each estimate their parameters. As mentioned before, the data were generated with 3 different cosine components which are shown in Figure 1(b) to Figure 1(d) with their respective noise, which are the functions that form part of the complete set. In Figure 1(e) to Figure 1(g) we can see how efficient the EM algorithm [11] is since we obtain the estimated values very close to the real values, with differences of the order of $10^{-3}$.

6. Conclusion
In most cases where it is possible to apply the EM algorithm, it will have better results compared to other alternatives such as the descending gradient or Newton-Raphson since these methods require the calculation of derivatives that can become a problem when finding the optimal solution, in addition in some cases it is necessary to establish parameters for convergence. For this reason, the EM algorithm is a good option to consider when dealing with parameter estimation problems.

References
[1] Aubone A and Wöhler O C 2000 Aplicación del método de máxima verosimilitud a la estimación de parámetros y comparación de curvas de crecimiento de von Bertalanffy INIDEPI Informe Técnico 37 1
[2] Stoica P and Nehorai A 1989 Maximum likelihood, and Cramer-Rao bound IEEE Transactions on Acoustics Speech and Signal Processing 37(5) 720
[3] Hernández Orallo J, Ramirez Quintana M J, Ferri Ramírez C 2004 Introducción a la minería de datos (España: Pearson Education)
[4] 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) 1
[5] Dorfman D D and Alf Jr E 1969 Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals-rating-method data Journal of mathematical psychology 6(3) 487
[6] Guindon S, Dufayard J F, Lefort V, Anisimova M, Hordijk W, and Gascuel O 2010 New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0 Systematic Biology 59(3) 307
[7] Bock R D and Aitkin M 1981 Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm Psychometrika 46(4) 443
[8] Wu C J 1983 On the convergence properties of the EM algorithm The Annals of Statistics 11(1) 95
[9] Parra L and Barrett H H 1998 List-mode likelihood: EM algorithm and image quality estimation demonstrated on 2-D PET IEEE transactions on medical imaging 17(2) 228
[10] Lindstrom M J and Bates D M 1988 Newton—Raphson and EM algorithms for linear mixed-effects models for repeated-measures data Journal of the American Statistical Association 83 1014
[11] Johansen S and Juselius K 1990 Maximum likelihood estimation and inference on cointegration with applications to the demand for money Oxford Bulletin of Economics and Statistics 52 2 169