The EM Algorithm: Selected Readings*

Ted Pedersen
Department of Computer Science
University of Minnesota Duluth
Duluth, MN 55812 USA

tpederse@d.umn.edu
http://www.d.umn.edu/~tpederse

Abstract

The literature of the Expectation Maximization (EM) algorithm is vast, and can be bewildering to the unsuspecting researcher who simply hopes to become familiar with this approach. This note offers a very brief introduction to this literature in the hopes that the interested reader will find a few useful sources to help get them started with EM.

1 Introduction

The seminal paper introducing the EM algorithm is (Dempster et al., 1977), hereafter referred to as DLR. While sections of this paper are rather theoretical, it also incorporates a variety of practical examples. A reader confronting DLR for the first time should seize upon a familiar distribution or application and use that as a toe–hold from which to spring into the rest of the paper. DLR are sometimes credited as the inventors of EM, but they are careful to point out that they are instead attempting to unify and generalize a body of existing work. They take pains to lay this historical foundation, and this makes for interesting reading. What is most intriguing is that many of the authors of this previous work are included in a discussion at the end of the paper. Their comments show that the concerns of today were clearly recognized at the time EM was introduced. These include the rate of convergence, the merits of iterative numerical alternatives such as Newton–Raphson, and the much broader issue regarding the general soundness of Maximum Likelihood Estimation.

DLR point out that there were numerous specialized instances of the EM algorithm to be found in the literature prior to 1977. Their general contribution is to have brought together a large and diverse pool of previous work, and unify it under the banner of the EM algorithm. A specific contribution of DLR is their proof of convergence, which shows that each iteration of the EM algorithm will never decrease the likelihood function, and thereby will arrive at a maximum. They acknowledge that there is no guarantee of finding a global maximum, and the prospect of converging at a local rather than global maximum is a well understood limitation of the EM algorithm. However, (Wu, 1983) shows that there is an error in this proof, and that the EM algorithm can really only guarantee convergence to a stationary point: a local or global maximum, or a saddle point. Wu is somewhat technical reading, and the conditions under which EM will converge at a saddle point are not common. However, it is important to understand this additional pitfall.

The interested reader should not disregard publications prior to 1977, since DLR did not discredit previous research. Rather, they unified a wide range of work into a generic framework for which they showed a number of interesting properties. Most of the previous work remains valid, and if it happens to be similar to the task at hand it may well provide a strong toe–hold into the more generalized EM methodology. For example, to handle multinomial distributions that suffer from missing data, (Hartley, 1958) presents a special case of the EM algorithm that fits neatly within the more general framework developed by DLR. (Baum et al., 1970) stands firm as a seminal work for estimating parameters in a Hidden Markov Model. The method presented there is the Baum–Welch (aka forward–backward) algorithm, one of the better known special cases of the EM algorithm developed prior to DLR.

2 Secondary Sources

The EM algorithm is discussed in textbooks and other secondary sources, often for the benefit of readers whose primary area of expertise is not statistics. Given the generality of the EM algorithm, such sources usually focus on particular applications that fall within the domain of interest.

For example, (Mitchell, 1997) is a Machine Learning textbook. It presents the EM algorithm as a tool for clustering via an example of estimating the parameters of a model that is a mixture of Normal/Gaussian distributions. Then it provides a general formulation of the EM algorithm, and concludes by working backwards from that derivation to show

* Unpublished notes to accompany the panel discussion on the EM algorithm at EMNLP, June 2001

that the mixture model example is an instance of EM.

(Rabiner and Juang, 1993) and (Jelinek, 1997) are speech recognition textbooks, while (Jurafsky and Martin, 2000) is a natural language processing textbook that provides significant coverage of speech recognition. Given this shared focus, it is not surprising that all discuss the Baum–Welch algorithm, and make it clear that this is a special case of the EM algorithm. Of the three, (Jelinek, 1997) describes the relationship between the Baum–Welch algorithm and the EM algorithm in the most detail.

(Charniak, 1993) and (Manning and Schütze, 1999) are textbooks that cover statistical natural language processing. Both include worked examples with discussion showing the application of Baum–Welch for estimating the parameters of Hidden Markov Models. The latter also includes discussion of unsupervised clustering by estimating the parameters of a mixture of normal distributions using the EM algorithm.

A reader interested in finding a single text to take them through the history, derivation, and applications of the EM algorithm will want to consider (McLachlan and Krishnan, 1997). While the intended reader is someone with formal training in statistics, even a relative newcomer can benefit from this text. For example, the historical overview is more extensive than that of DLR and goes back to the late 1800's, drawing attention to (Newcomb, 1887), who estimated the parameters of a mixture model via an EM–like algorithm. DLR only went back as far as (McKendrick, 1926) in their historical review. This text also includes an extensive set of example applications, and provides a detailed comparison of EM with iterative numerical techniques such as Newton–Raphson. It also introduces the use of EM as a Markov Chain Monte Carlo method.

If a reader is primarily interested in EM as a tool for handling missing data, (Little and Rubin, 1987) offers a complete discussion. For those who plan to use the EM algorithm with mixture models, (Titterington et al., 1985) goes well beyond the usual discussion of estimating the parameters of a mixture of normals.

Prior to DLR, researchers did not just give up when confronted with problems of missing data, estimating the parameters of mixture models, etc. In addition to developing instances of EM for special cases, they employed iterative numerical techniques such as the Newton–Raphson algorithm and Fisher scoring. These methods continue to have a strong following, and are certainly worth considering for some of the same kinds of problems for which EM is intended. (Thisted, 1988) provides an overview of these techniques (and the EM algorithm) from a computational point of view.

3 Extensions to EM

Two main criticisms of the EM algorithm are its tendency to converge very slowly, and to converge at points other than a global maximum. Alternatives and variations to EM that address these issues are of current interest.

A range of approaches to accelerate the convergence of EM are described in (Meng and van Dyk, 1997). This paper was presented to the Royal Statistics Society on the occasion of the twenty–fifth anniversary of DLR and as such offers a historical perspective as well.

When the likelihood function has a particularly difficult form, (Rubin, 1991) suggests sampling values from the distribution of the likelihood function rather than attempting to maximize via iterative computation. This moves EM into the realm of Multiple Imputation and Markov Chain Monte Carlo methods. (Tanner, 1993) provides a more detailed overview of these methods, and includes discussion specific to the EM algorithm.

4 Software

The EM algorithm can be easy to implement for distributions whose likelihood functions do not result in tremendously complex derivatives. There are also a few different sources of freely available software. As would be expected given the generality of the approach, these do not attempt to provide a generic implementation but rather focus on a particular distribution or application.

Clustering approaches to unsupervised learning have been implemented via the EM algorithm. Two examples are AutoClass (Cheeseman and Stutz, 1996) and Weka (Witten and Frank, 2000). Both implement the EM algorithm for finite mixture models. AutoClass is dedicated to clustering, while Weka includes EM as one of a general suite of supervised and unsupervised machine learning methods.

CoCo (Badsberg, 1995) is a statistical package for graphical models that can be represented as contingency tables. It includes support for the EM algorithm in order to handle missing data and latent variables in graphical models.

(I am sure there is more software available than this. I would be delighted to learn of anything you have found useful.)

5 Prerequisite Knowledge

As you must have guessed, I am not a statistician. I'm a computer scientist who sees the world as a collection of discrete countable objects. In order to come to terms with the EM algorithm, I have had to modify my world–view to include certain continuous concepts. The EM algorithm is a method of making maximum likelihood estimates under adverse circumstances, and as such it requires a command of

maximum likelihood estimation in general. Even a brief encounter with MLE leads one back into basic calculus, and a short review may be necessary before fully grasping the method of MLE. In particular, first and second derivatives are handy as are techniques for finding maxima and minima of functions. Simply having reasonable undergraduate calculus and mathematical statistics textbooks at the ready should smooth over any rusty patches. A reader who already has this knowledge or has undertaken a brief review can approach the primary and secondary sources mentioned above with confidence.

6 Conclusion

The simplicity of EM makes it relatively easy to employ in many different settings, whether it is appropriate or not. A firm grasp of the underlying issues will certainly avoid any potential abuse.

I hope this note points you towards a few resources that will help you in your dealings with EM. If you have your own favorites, be they written materials or software, I would be very grateful to learn of them. Also, if there are errors in this presentation I would be most thankful if you would point those out. This is a somewhat hastily drawn first draft (at best) and I plan to continue to refine this into something that will serve as a guide for those of us coming to EM from outside of statistics.

References

J. Badsberg. 1995. *An Environment for Graphical Models.* Ph.D. thesis, Aalborg University.

L. Baum, T. Petrie, G. Soules, and N. Weiss. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41:164–171.

E. Charniak. 1993. *Statistical Language Learning.* The MIT Press, Cambridge, MA.

P. Cheeseman and J. Stutz. 1996. Bayesian classification (AutoClass): Theory and results. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining.* AAAI Press/MIT Press.

A. Dempster, N. Laird, and D. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39(1):1–38.

H.O. Hartley. 1958. Maximum likelihood estimation from incomplete data. *Biometrics*, 14:174–194.

F. Jelinek. 1997. *Statistical Methods for Speech Recognition.* The MIT Press, Cambridge, MA.

D. Jurafsky and J. Martin. 2000. *Speech and Language Processing.* Prentice–Hall, Upper Saddle River, NJ.

R. Little and D. Rubin. 1987. *Statistical Analysis with Missing Data.* Wiley, New York.

C. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing.* The MIT Press, Cambridge, MA.

A. McKendrick. 1926. Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44:98–130.

G. McLachlan and T. Krishnan. 1997. *The EM algorithm and extensions.* Wiley, New York.

X. Meng and D. van Dyk. 1997. The EM algorithm – an old folk song sung to a new fast new tune. *Journal of Royal Statistics Society B*, 59(3):511–567.

T. Mitchell. 1997. *Machine Learning.* McGraw–Hill, Boston, MA.

S. Newcomb. 1887. A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, 8:343–366.

L. Rabiner and B-H. Juang. 1993. *Fundamentals of Speech Recognition.* Prentice–Hall, Englewood Cliffs, NJ.

D. Rubin. 1991. EM and beyond. *Psychometrika*, 56(2):241–254.

M. Tanner. 1993. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions.* Springer–Verlag, New York, NY.

R. Thisted. 1988. *Elements of Statistical Computing: Numerical Computation.* Chapman and Hall, New York.

D. Titterington, A. Smith, and U. Makov. 1985. *Statistical analysis of finite mixture distributions.* Wiley, New York.

I. Witten and E. Frank. 2000. *Data Mining - Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan–Kaufmann, San Francisco, CA.

C.F.J. Wu. 1983. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11(1):95–103.