The EM Algorithm and Extensions

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

The EM Algorithm and Extensions

Second Edition

Geoffrey J. McLachlan
The University of Queensland
Department of Mathematics and Institute for Molecular Bioscience
St. Lucia, Australia

Thriyambakam Krishnan
Cranes Software International Limited
Bangalore, India

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

McLachlan, Geoffrey J., 1946–
The EM algorithm and extensions / Geoffrey J. McLachlan,
Thriyambakam Krishnan. — 2nd ed.
p. cm.
ISBN 978-0-471-20170-0 (cloth)
1. Expectation-maximization algorithms. 2. Estimation theory. 3.
Missing observations (Statistics) I. Krishnan, T.
(Thriyambakam), 193– II. Title.
QA276.8.M394 2007
519.5'44—dc22 2007017908

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To
Beryl, Jonathan, and Robbie

CONTENTS

PREFACE TO THE SECOND EDITION

The second edition attempts to capture significant developments in EM methodology in the ten years since the publication of the first edition. The basic EM algorithm has two main drawbacks—slow convergence and lack of an in-built procedure to compute the covariance matrix of parameter estimates. Moreover, some complex problems lead to intractable E-steps, for which Monte Carlo methods have been shown to provide efficient solutions. There are many parallels and connections between the EM algorithm and Markov chain Monte Carlo algorithms, especially EM with data augmentation and Gibbs sampling. Furthermore, the key idea of the EM algorithm where a surrogate function of the log likelihood is maximized in a iterative procedure occurs in quite a few other optimization procedures as well, leading to a more general way of looking at EM as an optimization procedure.

Capturing the above developments in the second edition has led to updated, revised, and expanded versions of many sections of the first edition, and to the addition of two new chapters, one on Monte Carlo Versions of the EM Algorithm (Chapter 6) and another on Generalizations of the EM Algorithm (Chapter 7). These revisions and additions have necessitated the recasting of the first edition's final (sixth) chapter, some sections of which have gone into the new chapters in different forms. The remaining sections with some additions form the last chapter with the modified title of "Further Applications of the EM Algorithm."

The first edition of this book appeared twenty years after the publication of the seminal paper of Dempster, Laird, and Rubin (1977). This second edition appears just over ten years after the first edition. Meng (2007) in an article entitled "Thirty Years of EM and Much More" points out how EM and MCMC are intimately related, and that both have been "workhorses for statistical computing". The chapter on Monte Carlo Versions of the EM Algorithm attempts to bring out this EM–MCMC connection.

In this revised edition, we have drawn on material from Athreya, Delampady, and Krishnan (2003), Ng, Krishnan, and McLachlan (2004), and Krishnan (2004). Thanks are thus due to K.B. Athreya, M. Delampady, and Angus Ng.

We owe debts of gratitude to a number of other people for helping us prepare this edition: Ravindra Jore for the computations for the Linear Mixed Model Example; Mangalmurti Badgujar for carrying out Markov chain Monte Carlo computations with WinBugs, R, and SYSTAT; Arnab Chakraborty for reading the new chapters, pointing out errors and inadequacies, and giving valuable comments and suggestions; Ian Wood for reading Chapter 6 and providing us with valuable comments and suggestions; N.R. Chaganty for reading a draft of sections of Chapter 7 and giving useful comments; and David Hunter for reading sections of Chapters 3 and 7 and giving valuable comments and suggestions.

Lloyd Flack, Ian Wood, Vivien Challis, Sam Wang, and Richard Bean provided us with a great deal of LaTeX advice at various stages of the typesetting of the manuscript, and we owe them a great sense of gratitude. We thank too Devish Bhat for his assistance with the preparation of some of the figures.

The first author was supported by the Australian Research Council. Thanks are also due to the authors and owners of copyrighted material for permission to reproduce data, tables and figures. These are acknowledged in the appropriate pages in the text.

The web address for further information related to this book is:
`http://www.maths.uq.edu.au/~gjm/em2ed/`.

G.J. McLachlan		T. Krishnan
Brisbane	March 2008	Bangalore

PREFACE TO THE FIRST EDITION

This book deals with the Expectation–Maximization algorithm, popularly known as the EM algorithm. This is a general-purpose algorithm for maximum likelihood estimation in a wide variety of situations best described as *incomplete-data* problems. The name EM algorithm was given by Dempster, Laird, and Rubin in a celebrated paper read before the Royal Statistical Society in 1976 and published in its journal in 1977. In this paper, a general formulation of the EM algorithm was presented, its basic properties established, and many examples and applications of it provided. The idea behind the EM algorithm is intuitive and natural and so algorithms like it were formulated and applied in a variety of problems even before this paper. However, it was in this seminal paper that the ideas in the earlier papers were synthesized, a general formulation and a theory developed, and a host of traditional and non-traditional applications indicated. Since then, the EM algorithm has become a standard piece in the statistician's repertoire. The incomplete-data situations where the EM algorithm has been successfully applied include not only evidently incomplete-data situations, where there are missing data, truncated distributions, censored or grouped observations, but also a whole variety of situations where the incompleteness of the data is not natural or evident. Thus, in some situations, it requires a certain amount of ingenuity on the part of the statistician to formulate the incompleteness in a suitable manner to facilitate the application of the EM algorithm in a computationally profitable manner. Following the paper of Dempster, Laird, and Rubin (1977), a spate of applications of the algorithm have appeared in the literature.

The EM algorithm is not without its limitations, many of which came to light in attempting to apply it in certain complex incomplete-data problems and some even in innocuously simple incomplete-data problems. However, a number of modifications and extensions of the algorithm has been developed to overcome some of these limitations. Thus there is a whole battery of EM-related algorithms and more are still being developed. The current developments are, however, in the direction of iterative simulation techniques or Markov Chain Monte Carlo methods, many of which can be looked upon as simulation-based versions of various EM-type algorithms.

Incomplete-data problems arise in all statistical contexts. Hence in these problems where maximum likelihood estimates usually have to be computed iteratively, there is the scope and need for an EM algorithm to tackle them. Further, even if there are no missing data or other forms of data incompleteness, it is often profitable to express the given problem as an incomplete-data one within an EM framework. For example, in some multiparameter problems like in random effects models, where an averaging over some parameters is to be carried out, an incomplete-data approach via the EM algorithm and its variants has been found useful. No wonder then that the EM algorithm has become an ubiquitous statistical tool, is a part of the entire spectrum of statistical methods, and has found applications in almost all fields where statistical techniques have been applied. The EM algorithm and its variants have been applied in such fields as medical imaging, dairy science, correcting census undercount, and AIDS epidemiology, to mention a few. Articles containing applications of the EM algorithm and even some with some methodological content have appeared in a variety of journals on statistical theory and methodology, statistical computing, and statistical applications in engineering, biology, medicine, social sciences, etc. Meng and Pedlow (1992) list a bibliography of over 1000 items and now there are at least 1700 publications related to the EM algorithm.

It is surprising that despite the obvious importance of the technique and its ramifications, no book on the subject has so far appeared. Indeed, many modern books dealing with some aspect of statistical estimation have at least some EM algorithm content. The books by Little and Rubin (1987), Tanner (1991, 1993), and Schafer (1996) have substantial EM algorithm content. But still, there seems to be a need for a full-fledged book on the subject. In our experience of lecturing to audiences of professional statisticians and to users of statistics, it appears that there is a definite need for a unified and complete treatment of the theory and methodology of the EM algorithm and its extensions, and their applications. The purpose of our writing this book is to fulfill this need. The various extensions of the EM algorithm due to Rubin, Meng, Liu, and others that have appeared in the last few years, have made this need even greater. Many extensions of the EM algorithm in the direction of iterative simulation have also appeared in recent years. Inclusion of these techniques in this book may have resulted in a more even-handed and comprehensive treatment of the EM algorithm and its extensions. However, we decided against it, since this methodology is still evolving and rapid developments in this area may make this material soon obsolete. So we have restricted this book to the EM algorithm and its variants and have only just touched upon the iterative simulation versions of it.

The book is aimed at theoreticians and practitioners of Statistics and its objective is to introduce to them the principles and methodology of the EM algorithm and its tremendous potential for applications. The main parts of the book describing the formulation of the EM algorithm, detailing its methodology, discussing aspects of its implementation, and illustrating its application in many simple statistical contexts, should be comprehensible to graduates with Statistics as their major subject. Throughout the book, the theory and methodology are illustrated with a number of examples. Where relevant, analytical examples are followed up with numerical examples. There are about thirty examples in the book. Some parts of the book, especially examples like factor analysis and variance components analysis, will need basic knowledge of these techniques to comprehend the full impact of the use of the EM algorithm. But our treatment of these examples is self-contained, although brief. However, these examples can be skipped without losing continuity.

Chapter 1 begins with a brief discussion of maximum likelihood (ML) estimation and standard approaches to the calculation of the maximum likelihood estimate (MLE) when it does not exist as a closed form solution of the likelihood equation. This is followed by a few examples of incomplete-data problems for which an heuristic derivation of the EM algorithm is given. The EM algorithm is then formulated and its basic terminology and notation established. The case of the regular exponential family (for the complete-data problem) for which the EM algorithm results in a particularly elegant solution, is specially treated. Throughout the treatment, the Bayesian perspective is also included by showing how the EM algorithm and its variants can be adapted to compute maximum *a posteriori* (MAP) estimates. The use of the EM algorithm and its variants in maximum penalized likelihood estimation (MPLE), a technique by which the MLE is smoothed, is also included.

Chapter 1 also gives a summary of the properties of the EM algorithm. Towards the end of Chapter 1, a comprehensive discussion of the history of the algorithm is presented, with a listing of the earlier ideas and examples upon which the general formulation is based. The chapter closes with a summary of the developments in the methodology since the Dempster et al. (1977) paper and with an indication of the range of applications of the algorithm.

In Chapter 2, a variety of examples of the EM algorithm is presented, following the general formulation in Chapter 1. These examples include missing values (in the conventional sense) in various experimental designs, the multinomial distribution with complex cell structure as used in genetics, the multivariate t-distribution for the provision of a robust estimate of a location parameter, Poisson regression models in a computerized image reconstruction process such as SPECT/PET, and the fitting of normal mixture models to grouped and truncated data as in the modeling of the volume of red blood cells.

In Chapter 3, the basic theory of the EM algorithm is systematically presented, and the monotonicity of the algorithm, convergence, and rates of convergence properties are established. The Generalized EM (GEM) algorithm and its properties are also presented. The principles of Missing Information and Self-Consistency are discussed. In this chapter, attention is inevitably given to mathematical details. However, mathematical details and theoretical points are explained and illustrated with the help of earlier and new examples. Readers not interested in the more esoteric aspects of the EM algorithm may only study the examples in this chapter or omit the chapter altogether without losing continuity.

In Chapter 4, two issues which have led to some criticism of the EM algorithm are addressed. The first concerns the provision of standard errors, or the full covariance matrix in multivariate situations, of the MLE obtained via the EM algorithm. One initial criticism of the EM algorithm was that it does not automatically provide an estimate of the covariance matrix of the MLE, as do some other approaches such as Newton-type methods. Hence we consider a number of methods for assessing the covariance matrix of the MLE $\hat{\boldsymbol{\Psi}}$ of the parameter vector $\boldsymbol{\Psi}$, obtained via the EM algorithm. Most of these methods are based on the observed information matrix. A coverage is given of methods such as the Supplemented EM algorithm that allow the observed information matrix to be calculated within the EM framework. The other common criticism that has been leveled at the EM algorithm is that its convergence can be quite slow. We therefore consider some methods that have been proposed for accelerating the convergence of the EM algorithm. They include methods based on Aitken's acceleration procedure and the generalized conjugate gradient approach, and hybrid methods that switch from the EM algorithm after a few iterations to some Newton-type method. We consider also the use of the EM gradient algorithm as a basis of a quasi-Newton approach to accelerate convergence of the EM algorithm. This algorithm approximates the M-step of the EM algorithm by one Newton-Raphson step when the solution does not exist in closed form.

In Chapter 5, further modifications and extensions to the EM algorithm are discussed. The focus is on the Expectation–Conditional Maximum (ECM) algorithm and its extensions, including the Expectation–Conditional Maximum Either (ECME) and Alternating ECM (AECM) algorithms. The ECM algorithm is a natural extension of the EM algorithm in situations where the maximization process on the M-step is relatively simple when conditional on some function of the parameters under estimation. The ECM algorithm therefore replaces the M-step of the EM algorithm by a number of computationally simpler conditional maximization (CM) steps. These extensions of the EM algorithm typically produce an appreciable reduction in total computer time. More importantly, they preserve the appealing convergence properties of the EM algorithm, such as its monotone convergence.

In Chapter 6, very brief overviews are presented of iterative simulation techniques such as the Monte Carlo E-step, Stochastic EM algorithm, Data Augmentation, and the Gibbs sampler and their connections with the various versions of the EM algorithm. Then, a few methods such as Simulated Annealing, which are considered competitors to the EM algorithm, are described and a few examples comparing the performance of the EM algorithm with these competing methods are presented. The book is concluded with a brief account of the applications of the EM algorithm in such topical and interesting areas as Hidden Markov Models, AIDS epidemiology, and Neural Networks.

One of the authors (GJM) would like to acknowledge gratefully financial support from the Australian Research Council. Work on the book by one of us (TK) was facilitated by two visits to the University of Queensland, Brisbane, and a visit to Curtin University of Technology, Perth. One of the visits to Brisbane was under the Ethel Raybould Fellowship scheme. Thanks are due to these two Universities for their hospitality. Thanks are also due to the authors and owners of copyrighted material for permission to reproduce tables and figures. The authors also wish to thank Ramen Kar and Amiya Das of the Indian Statistical Institute, Calcutta, and Pauline Wilson of the University of Queensland, Brisbane, for their help with LaTeX and word processing, and Rudy Blazek of Michigan State University for assistance with the preparation of some of the figures.

G.J. McLachlan
Brisbane

T. Krishnan
Calcutta

November 1995

LIST OF EXAMPLES

continued

continued