

The EM Algorithm and Extensions

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice,
Iain M. Johnstone, Geert Molenberghs, David W. Scott, Adrian F. M. Smith,
Ruey S. Tsay, Sanford Weisberg*

Editors Emeriti: *Vic Barnett, J. Stuart Hunter, David G. Kendall, Jozef L. Teugels*

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*



A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2008 by John Wiley & Sons, Inc. All rights reserved.

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

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic format. For information about Wiley products, visit our web site at www.wiley.com.

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	xix
PREFACE TO THE FIRST EDITION	xxi
LIST OF EXAMPLES	xxv
1 GENERAL INTRODUCTION	1
1.1 Introduction	1
1.2 Maximum Likelihood Estimation	3
1.3 Newton-Type Methods	5
1.3.1 Introduction	5
1.3.2 Newton-Raphson Method	5
1.3.3 Quasi-Newton Methods	6
1.3.4 Modified Newton Methods	6
1.4 Introductory Examples	8
1.4.1 Introduction	8
1.4.2 Example 1.1: A Multinomial Example	8
	vii

1.4.3	Example 1.2: Estimation of Mixing Proportions	13
1.5	Formulation of the EM Algorithm	18
1.5.1	EM Algorithm	18
1.5.2	Example 1.3: Censored Exponentially Distributed Survival Times	20
1.5.3	E- and M-Steps for the Regular Exponential Family	22
1.5.4	Example 1.4: Censored Exponentially Distributed Survival Times (<i>Example 1.3 Continued</i>)	23
1.5.5	Generalized EM Algorithm	24
1.5.6	GEM Algorithm Based on One Newton-Raphson Step	24
1.5.7	EM Gradient Algorithm	25
1.5.8	EM Mapping	26
1.6	EM Algorithm for MAP and MPL Estimation	26
1.6.1	Maximum <i>a Posteriori</i> Estimation	26
1.6.2	Example 1.5: A Multinomial Example (<i>Example 1.1 Continued</i>)	27
1.6.3	Maximum Penalized Estimation	27
1.7	Brief Summary of the Properties of the EM Algorithm	28
1.8	History of the EM Algorithm	29
1.8.1	Early EM History	29
1.8.2	Work Before Dempster, Laird, and Rubin (1977)	29
1.8.3	EM Examples and Applications Since Dempster, Laird, and Rubin (1977)	31
1.8.4	Two Interpretations of EM	32
1.8.5	Developments in EM Theory, Methodology, and Applications	33
1.9	Overview of the Book	36
1.10	Notations	37

2 EXAMPLES OF THE EM ALGORITHM 41

2.1	Introduction	41
2.2	Multivariate Data with Missing Values	42
2.2.1	Example 2.1: Bivariate Normal Data with Missing Values	42
2.2.2	Numerical Illustration	45
2.2.3	Multivariate Data: Buck's Method	45
2.3	Least Squares with Missing Data	47
2.3.1	Healy–Westmacott Procedure	47

2.3.2	Example 2.2: Linear Regression with Missing Dependent Values	47
2.3.3	Example 2.3: Missing Values in a Latin Square Design	49
2.3.4	Healy–Westmacott Procedure as an EM Algorithm	49
2.4	Example 2.4: Multinomial with Complex Cell Structure	51
2.5	Example 2.5: Analysis of PET and SPECT Data	54
2.6	Example 2.6: Multivariate t -Distribution (Known D.F.)	58
2.6.1	ML Estimation of Multivariate t -Distribution	58
2.6.2	Numerical Example: Stack Loss Data	61
2.7	Finite Normal Mixtures	61
2.7.1	Example 2.7: Univariate Component Densities	61
2.7.2	Example 2.8: Multivariate Component Densities	64
2.7.3	Numerical Example: Red Blood Cell Volume Data	65
2.8	Example 2.9: Grouped and Truncated Data	66
2.8.1	Introduction	66
2.8.2	Specification of Complete Data	66
2.8.3	E-Step	69
2.8.4	M-Step	70
2.8.5	Confirmation of Incomplete-Data Score Statistic	70
2.8.6	M-Step for Grouped Normal Data	71
2.8.7	Numerical Example: Grouped Log Normal Data	72
2.9	Example 2.10: A Hidden Markov AR(1) model	73

3 BASIC THEORY OF THE EM ALGORITHM

77

3.1	Introduction	77
3.2	Monotonicity of the EM Algorithm	78
3.3	Monotonicity of a Generalized EM Algorithm	79
3.4	Convergence of an EM Sequence to a Stationary Value	79
3.4.1	Introduction	79
3.4.2	Regularity Conditions of Wu (1983)	80
3.4.3	Main Convergence Theorem for a Generalized EM Sequence	81
3.4.4	A Convergence Theorem for an EM Sequence	82
3.5	Convergence of an EM Sequence of Iterates	83
3.5.1	Introduction	83
3.5.2	Two Convergence Theorems of Wu (1983)	83
3.5.3	Convergence of an EM Sequence to a Unique Maximum Likelihood Estimate	84

3.5.4	Constrained Parameter Spaces	84
3.6	Examples of Nontypical Behavior of an EM (GEM) Sequence	85
3.6.1	Example 3.1: Convergence to a Saddle Point	85
3.6.2	Example 3.2: Convergence to a Local Minimum	88
3.6.3	Example 3.3: Nonconvergence of a Generalized EM Sequence	90
3.6.4	Example 3.4: Some E-Step Pathologies	93
3.7	Score Statistic	95
3.8	Missing Information	95
3.8.1	Missing Information Principle	95
3.8.2	Example 3.5: Censored Exponentially Distributed Survival Times (<i>Example 1.3 Continued</i>)	96
3.9	Rate of Convergence of the EM Algorithm	99
3.9.1	Rate Matrix for Linear Convergence	99
3.9.2	Measuring the Linear Rate of Convergence	100
3.9.3	Rate Matrix in Terms of Information Matrices	101
3.9.4	Rate Matrix for Maximum <i>a Posteriori</i> Estimation	102
3.9.5	Derivation of Rate Matrix in Terms of Information Matrices	102
3.9.6	Example 3.6: Censored Exponentially Distributed Survival Times (<i>Example 1.3 Continued</i>)	103

4 STANDARD ERRORS AND SPEEDING UP CONVERGENCE 105

4.1	Introduction	105
4.2	Observed Information Matrix	106
4.2.1	Direct Evaluation	106
4.2.2	Extraction of Observed Information Matrix in Terms of the Complete-Data Log Likelihood	106
4.2.3	Regular Case	108
4.2.4	Evaluation of the Conditional Expected Complete-Data Information Matrix	108
4.2.5	Examples	109
4.3	Approximations to Observed Information Matrix: i.i.d. Case	114
4.4	Observed Information Matrix for Grouped Data	116
4.4.1	Approximation Based on Empirical Information	116
4.4.2	Example 4.3: Grouped Data from an Exponential Distribution	117
4.5	Supplemented EM Algorithm	120

4.5.1	Definition	120
4.5.2	Calculation of $J(\hat{\Psi})$ via Numerical Differentiation	122
4.5.3	Stability	123
4.5.4	Monitoring Convergence	124
4.5.5	Difficulties of the SEM Algorithm	124
4.5.6	Example 4.4: Univariate Contaminated Normal Data	125
4.5.7	Example 4.5: Bivariate Normal Data with Missing Values	128
4.6	Bootstrap Approach to Standard Error Approximation	130
4.7	Baker's, Louis', and Oakes' Methods for Standard Error Computation	131
4.7.1	Baker's Method for Standard Error Computation	131
4.7.2	Louis' Method of Standard Error Computation	132
4.7.3	Oakes' Formula for Standard Error Computation	133
4.7.4	Example 4.6: Oakes' Standard Error for Example 1.1	134
4.7.5	Example 4.7: Louis' Method for Example 2.4	134
4.7.6	Baker's Method for Standard Error for Categorical Data	135
4.7.7	Example 4.8: Baker's Method for Example 2.4	136
4.8	Acceleration of the EM Algorithm via Aitken's Method	137
4.8.1	Aitken's Acceleration Method	137
4.8.2	Louis' Method	137
4.8.3	Example 4.9: Multinomial Data	138
4.8.4	Example 4.10: Geometric Mixture	139
4.8.5	Example 4.11: Grouped and Truncated Data. (<i>Example 2.8 Continued</i>)	142
4.9	An Aitken Acceleration-Based Stopping Criterion	142
4.10	Conjugate Gradient Acceleration of EM Algorithm	144
4.10.1	Conjugate Gradient Method	144
4.10.2	A Generalized Conjugate Gradient Algorithm	144
4.10.3	Accelerating the EM Algorithm	145
4.11	Hybrid Methods for Finding the MLE	146
4.11.1	Introduction	146
4.11.2	Combined EM and Modified Newton-Raphson Algorithm	146
4.12	A GEM Algorithm Based on One Newton-Raphson Step	148
4.12.1	Derivation of a Condition to be a Generalized EM Sequence	148
4.12.2	Simulation Experiment	149
4.13	EM Gradient Algorithm	149
4.14	A Quasi-Newton Acceleration of the EM Algorithm	151
4.14.1	The Method	151

4.14.2	Example 4.12: Dirichlet Distribution	153
4.15	Ikeda Acceleration	157
5	EXTENSIONS OF THE EM ALGORITHM	159
5.1	Introduction	159
5.2	ECM Algorithm	160
5.2.1	Motivation	160
5.2.2	Formal Definition	160
5.2.3	Convergence Properties	162
5.2.4	Speed of Convergence	162
5.2.5	Convergence Rates of EM and ECM	163
5.2.6	Example 5.1: ECM Algorithm for Hidden Markov AR(1) Model	164
5.2.7	Discussion	164
5.3	Multicycle ECM Algorithm	165
5.4	Example 5.2: Normal Mixtures with Equal Correlations	166
5.4.1	Normal Components with Equal Correlations	166
5.4.2	Application of ECM Algorithm	166
5.4.3	Fisher's <i>Iris</i> Data	168
5.5	Example 5.3: Mixture Models for Survival Data	168
5.5.1	Competing Risks in Survival Analysis	168
5.5.2	A Two-Component Mixture Regression Model	169
5.5.3	Observed Data	169
5.5.4	Application of EM Algorithm	170
5.5.5	M-Step for Gompertz Components	171
5.5.6	Application of a Multicycle ECM Algorithm	172
5.5.7	Other Examples of EM Algorithm in Survival Analysis	173
5.6	Example 5.4: Contingency Tables with Incomplete Data	174
5.7	ECME Algorithm	175
5.8	Example 5.5: MLE of t -Distribution with Unknown D.F.	176
5.8.1	Application of the EM Algorithm	176
5.8.2	M-Step	177
5.8.3	Application of ECM Algorithm	177
5.8.4	Application of ECME Algorithm	178
5.8.5	Some Standard Results	178
5.8.6	Missing Data	179
5.8.7	Numerical Examples	181

5.8.8	Theoretical Results on the Rate of Convergence	181
5.9	Example 5.6: Variance Components	182
5.9.1	A Variance Components Model	182
5.9.2	E-Step	183
5.9.3	M-Step	184
5.9.4	Application of Two Versions of ECME Algorithm	185
5.9.5	Numerical Example	185
5.10	Linear Mixed Models	186
5.10.1	Introduction	186
5.10.2	General Form of Linear Mixed Model	187
5.10.3	REML Estimation	188
5.10.4	Example 5.7: REML Estimation in a Hierarchical Random Effects Model	188
5.10.5	Some Other EM-Related Approaches to Mixed Model Estimation	191
5.10.6	Generalized Linear Mixed Models	191
5.11	Example 5.8: Factor Analysis	193
5.11.1	EM Algorithm for Factor Analysis	193
5.11.2	ECME Algorithm for Factor Analysis	196
5.11.3	Numerical Example	196
5.11.4	EM Algorithm in Principal Component Analysis	196
5.12	Efficient Data Augmentation	198
5.12.1	Motivation	198
5.12.2	Maximum Likelihood Estimation of t -Distribution	198
5.12.3	Variance Components Model	202
5.13	Alternating ECM Algorithm	202
5.14	Example 5.9: Mixtures of Factor Analyzers	204
5.14.1	Normal Component Factor Analyzers	205
5.14.2	E-step	205
5.14.3	CM-steps	206
5.14.4	t -Component Factor Analyzers	207
5.14.5	E-step	210
5.14.6	CM-steps	211
5.15	Parameter-Expanded EM (PX-EM) Algorithm	212
5.16	EMS Algorithm	213
5.17	One-Step-Late Algorithm	213
5.18	Variance Estimation for Penalized EM and OSL Algorithms	214

5.18.1	Penalized EM Algorithm	214
5.18.2	OSL Algorithm	215
5.18.3	Example 5.9: Variance of MPLE for the Multinomial (<i>Examples 1.1 and 4.1 Continued</i>)	215
5.19	Incremental EM	216
5.20	Linear Inverse Problems	217

6 MONTE CARLO VERSIONS OF THE EM ALGORITHM 219

6.1	Introduction	219
6.2	Monte Carlo Techniques	220
6.2.1	Integration and Optimization	220
6.2.2	Example 6.1: Monte Carlo Integration	221
6.3	Monte Carlo EM	221
6.3.1	Introduction	221
6.3.2	Example 6.2: Monte Carlo EM for Censored Data from Normal	223
6.3.3	Example 6.3: MCEM for a Two-Parameter Multinomial (<i>Example 2.4 Continued</i>)	224
6.3.4	MCEM in Generalized Linear Mixed Models	224
6.3.5	Estimation of Standard Error with MCEM	225
6.3.6	Example 6.4: MCEM Estimate of Standard Error for One-Parameter Multinomial (<i>Example 1.1 Continued</i>)	226
6.3.7	Stochastic EM Algorithm	227
6.4	Data Augmentation	228
6.4.1	The Algorithm	228
6.4.2	Example 6.5: Data Augmentation in the Multinomial (<i>Examples 1.1, 1.5 Continued</i>)	229
6.5	Bayesian EM	230
6.5.1	Posterior Mode by EM	230
6.5.2	Example 6.6: Bayesian EM for Normal with Semi-Conjugate Prior	231
6.6	I.I.D. Monte Carlo Algorithms	232
6.6.1	Introduction	232
6.6.2	Rejection Sampling Methods	233
6.6.3	Importance Sampling	234
6.7	Markov Chain Monte Carlo Algorithms	236
6.7.1	Introduction	236

6.7.2	Essence of MCMC	238
6.7.3	Metropolis–Hastings Algorithms	239
6.8	Gibbs Sampling	241
6.8.1	Introduction	241
6.8.2	Rao–Blackwellized Estimates with Gibbs Samples	242
6.8.3	Example 6.7: Why Does Gibbs Sampling Work?	243
6.9	Examples of MCMC Algorithms	245
6.9.1	Example 6.8: M-H Algorithm for Bayesian Probit Regression	245
6.9.2	Monte Carlo EM with MCMC	246
6.9.3	Example 6.9: Gibbs Sampling for the Mixture Problem	249
6.9.4	Example 6.10: Bayesian Probit Analysis with Data Augmentation	250
6.9.5	Example 6.11: Gibbs Sampling for Censored Normal	251
6.10	Relationship of EM to Gibbs Sampling	254
6.10.1	EM–Gibbs Sampling Connection	254
6.10.2	Example 6.12: EM–Gibbs Connection for Censored Data from Normal (<i>Example 6.11 Continued</i>)	256
6.10.3	Example 6.13: EM–Gibbs Connection for Normal Mixtures	257
6.10.4	Rate of Convergence of Gibbs Sampling and EM	257
6.11	Data Augmentation and Gibbs Sampling	258
6.11.1	Introduction	258
6.11.2	Example 6.14: Data Augmentation and Gibbs Sampling for Censored Normal (<i>Example 6.12 Continued</i>)	259
6.11.3	Example 6.15: Gibbs Sampling for a Complex Multinomial (<i>Example 2.4 Continued</i>)	260
6.11.4	Gibbs Sampling Analogs of ECM and ECME Algorithms	261
6.12	Empirical Bayes and EM	263
6.13	Multiple Imputation	264
6.14	Missing-Data Mechanism, Ignorability, and EM Algorithm	265

7 SOME GENERALIZATIONS OF THE EM ALGORITHM **269**

7.1	Introduction	269
7.2	Estimating Equations and Estimating Functions	270
7.3	Quasi-Score and the Projection-Solution Algorithm	270
7.4	Expectation-Solution (ES) Algorithm	273
7.4.1	Introduction	273

7.4.2	Computational and Asymptotic Properties of the ES Algorithm	274
7.4.3	Example 7.1: Multinomial Example by ES Algorithm (<i>Example 1.1 Continued</i>)	274
7.5	Other Generalizations	275
7.6	Variational Bayesian EM Algorithm	276
7.7	MM Algorithm	278
7.7.1	Introduction	278
7.7.2	Methods for Constructing Majorizing/Minorizing Functions	279
7.7.3	Example 7.2: MM Algorithm for the Complex Multinomial (<i>Example 1.1 Continued</i>)	280
7.8	Lower Bound Maximization	281
7.9	Interval EM Algorithm	283
7.9.1	The Algorithm	283
7.9.2	Example 7.3: Interval-EM Algorithm for the Complex Multinomial (<i>Example 2.4 Continued</i>)	283
7.10	Competing Methods and Some Comparisons with EM	284
7.10.1	Introduction	284
7.10.2	Simulated Annealing	284
7.10.3	Comparison of SA and EM Algorithm for Normal Mixtures	285
7.11	The Delta Algorithm	286
7.12	Image Space Reconstruction Algorithm	287

8 FURTHER APPLICATIONS OF THE EM ALGORITHM 289

8.1	Introduction	289
8.2	Hidden Markov Models	290
8.3	AIDS Epidemiology	293
8.4	Neural Networks	295
8.4.1	Introduction	295
8.4.2	EM Framework for NNs	296
8.4.3	Training Multi-Layer Perceptron Networks	297
8.4.4	Intractibility of the Exact E-Step for MLPs	300
8.4.5	An Integration of the Methodology Related to EM Training of RBF Networks	300
8.4.6	Mixture of Experts	301
8.4.7	Simulation Experiment	305
8.4.8	Normalized Mixtures of Experts	306

8.4.9	Hierarchical Mixture of Experts	307
8.4.10	Boltzmann Machine	308
8.5	Data Mining	309
8.6	Bioinformatics	310
REFERENCES		311
AUTHOR INDEX		339
SUBJECT INDEX		347

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
Brisbane

March 2008

T. Krishnan
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 exam-

ples 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 $\hat{\Psi}$ of the parameter vector Ψ , 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

Example Number	Title	Section	Page Number
1.1	A Multinomial Example	1.4.2	8
1.2	Estimation of Mixing Proportions	1.4.3	13
1.3	Censored Exponentially Distributed Survival Times	1.5.2	20
1.4	Censored Exponentially Distributed Survival Times (<i>Example 1.3 Continued</i>)	1.5.4	23
1.5	A Multinomial Example (<i>Example 1.1 Continued</i>)	1.6.2	27
2.1	Bivariate Normal Data with Missing Values	2.2.1	42
2.2	Linear Regression with Missing Dependent Values	2.3.2	47
2.3	Missing Values in a Latin Square Design	2.3.3	49
2.4	Multinomial with Complex Cell Structure	2.4	51
2.5	Analysis of PET and SPECT Data	2.5	54
2.6	Multivariate t-Distribution with Known Degrees of Freedom	2.6	58
2.7	(Finite Normal Mixtures) Univariate Component Densities	2.7.1	61
2.8	(Finite Normal Mixtures) Multivariate Component Densities	2.7.2	64
2.9	Grouped and Truncated Data	2.8	66
2.10	A hidden Markov AR(1) model	2.9	73
3.1	Convergence to a Saddle Point	3.6.1	85
3.2	Convergence to a Local Minimum	3.6.2	88
3.3	Nonconvergence of a Generalized EM Sequence	3.6.3	90
3.4	Some E-Step Pathologies	3.6.4	93

continued

Example Number	Title	Section	Page Number
3.5	Censored Exponentially Distributed Survival Times (<i>Example 1.3 Continued</i>)	3.8.2	96
3.6	Censored Exponentially Distributed Survival Times (<i>Example 1.3 Continued</i>)	3.9.6	103
4.1	Information Matrix for the Multinomial Example (<i>Example 1.1 Continued</i>)	4.2.5	109
4.2	(Information Matrix) Mixture of Two Univariate Normals with Known Common Variance	4.2.5	111
4.3	Grouped Data from an Exponential Distribution	4.4.2	117
4.4	(SEM) Univariate Contaminated Normal Data	4.5.6	125
4.5	(SEM) Bivariate Normal Data with Missing Values	4.5.7	128
4.6	Oakes' Standard Error for Example 1.1	4.7.4	134
4.7	Louis' Method for Example 2.4	4.7.5	134
4.8	Baker's Method for Example 2.4	4.7.7	136
4.9	(Aitken Acceleration) Multinomial Data	4.8.3	138
4.10	(Aitken Acceleration) Geometric Mixture	4.8.4	139
4.11	(Aitken Acceleration) Grouped and Truncated Data (<i>Example 2.8 Continued</i>)	4.8.5	142
4.12	(Quasi-Newton Acceleration) Dirichlet Distribution	4.14.2	153
5.1	ECM Algorithm for Hidden Markov AR(1) Model	5.2.6	164
5.2	(ECM) Normal Mixtures with Equal Correlations	5.4	166
5.3	(ECM) Mixture Models for Survival Data	5.5	168
5.4	(ECM) Contingency Tables with Incomplete Data	5.6	174
5.5	MLE of t -Distribution with Unknown D.F.	5.8	176
5.6	(ECME Algorithm) Variance Components	5.9	182
5.7	REML Estimation in a Hierarchical Random Effects Model	5.10.4	188
5.8	(ECM Algorithm) Factor Analysis	5.11	192
5.9	Mixtures of Factor Analyzers	5.14	204
5.10	Variance of MPLE for the Multinomial (<i>Examples 1.1 and 4.1 Continued</i>)	5.18.3	215

continued

Example Number	Title	Section	Page Number
6.1	Monte Carlo Integration	6.2.2	221
6.2	Monte Carlo EM for Censored Data from Normal	6.3.2	223
6.3	MCEM for a Two-Parameter Multinomial (<i>Example 2.4 Continued</i>)	6.3.3	224
6.4	MCEM Estimate of Standard Error for One-Parameter Multinomial (<i>Example 1.1 Continued</i>)	6.3.6	226
6.5	Data Augmentation in the Multinomial (<i>Examples 1.1, 1.5 Continued</i>)	6.4.2	229
6.6	Bayesian EM for Normal with Semi-Conjugate Prior	6.5.2	231
6.7	Why Does Gibbs Sampling Work?	6.8.3	243
6.8	M-H Algorithm for Bayesian Probit Regression	6.9.1	245
6.9	Gibbs Sampling for the Mixture Problem	6.9.3	249
6.10	Bayesian Probit Analysis with Data Augmentation	6.9.4	250
6.11	Gibbs Sampling for Censored Normal	6.9.5	251
6.12	EM–Gibbs Connection for Censored Data from Normal (<i>Example 6.11 Continued</i>)	6.10.2	256
6.13	EM-Gibbs Connection for Normal Mixtures	6.10.3	257
6.14	Data Augmentation and Gibbs Sampling for Censored Normal (<i>Example 6.12 Continued</i>)	6.11.2	259
6.15	Gibbs Sampling for a Complex Multinomial (<i>Example 2.4 Continued</i>)	6.11.3	260
7.1	Multinomial Example by ES Algorithm (<i>Example 1.1 Continued</i>)	7.4.3	274
7.2	MM Algorithm for the Complex Multinomial (<i>Example 1.1 Continued</i>)	7.7.3	280
7.3	Interval-EM Algorithm for the Complex Multinomial (<i>Example 2.4 Continued</i>)	7.9.2	283