Finite Mixture Models
WILEY SERIES IN PROBABILITY AND STATISTICS
APPLIED PROBABILITY AND STATISTICS SECTION

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.
Finite Mixture Models

GEOFFREY McLACHLAN
DAVID PEEL
Department of Mathematics
The University of Queensland

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto
To
Beryl, Jonathan, and Robbie

Samantha
Contents

Preface xix

1 General Introduction 1

1.1 Introduction 1

1.1.1 Flexible Method of Modeling 1

1.1.2 Initial Approach to Mixture Analysis 2

1.1.3 Impact of EM Algorithm 3

1.2 Overview of Book 4

1.3 Basic Definition 6

1.4 Interpretation of Mixture Models 7

1.5 Shapes of Some Univariate Normal Mixtures 9

1.5.1 Mixtures of Two Normal Homoscedastic Components 9

1.5.2 Mixtures of Univariate Normal Heteroscedastic Components 11

1.6 Modeling of Asymmetrical Data 14

1.7 Normal Scale Mixture Model 17

1.8 Spurious Clusters 17

1.9 Incomplete-Data Structure of Mixture Problem 19

1.10 Sampling Designs for Classified Data 21

1.11 Parametric Formulation of Mixture Model 22
1.12 Nonparametric ML Estimation of a Mixing Distribution 23
1.13 Estimation of Mixture Distributions 24
1.14 Identifiability of Mixture Distributions 26
1.15 Clustering of Data via Mixture Models 29
  1.15.1 Mixture Likelihood Approach to Clustering 29
  1.15.2 Decision-Theoretic Approach 30
  1.15.3 Clustering of I.I.D. Data 31
  1.15.4 Image Segmentation or Restoration 32
1.16 Hidden Markov Models 33
1.17 Testing for the Number of Components in Mixture Models 34
1.18 Brief History of Finite Mixture Models 35
1.19 Notation 37

2 ML Fitting of Mixture Models 40
  2.1 Introduction 40
  2.2 ML Estimation 40
  2.3 Information Matrices 41
  2.4 Asymptotic Covariance Matrix of MLE 42
  2.5 Properties of MLEs for Mixture Models 42
  2.6 Choice of Root 44
  2.7 Test for a Consistent Root 44
    2.7.1 Basis of Test 44
    2.7.2 Example 2.1: Likelihood Function with Two Maximizers 45
    2.7.3 Formulation of Test Statistic 45
  2.8 Application of EM Algorithm for Mixture Models 47
    2.8.1 Direct Approach 47
    2.8.2 Formulation as an Incomplete-Data Problem 48
    2.8.3 E-Step 48
    2.8.4 M-Step 49
    2.8.5 Assessing the Implied Error Rates 50
  2.9 Fitting Mixtures of Mixtures 51
  2.10 Maximum a Posteriori Estimation 52
  2.11 An Aitken Acceleration-Based Stopping Criterion 52
  2.12 Starting Values for EM Algorithm 54
    2.12.1 Specification of an Initial Parameter Value 54
    2.12.2 Random Starting Values 55
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5 Assessment of Model Fit</td>
<td>84</td>
</tr>
<tr>
<td>3.6 Examples of Univariate Normal Mixtures</td>
<td>85</td>
</tr>
<tr>
<td>3.6.1 Basic Model in Genetics</td>
<td>85</td>
</tr>
<tr>
<td>3.6.2 Example 3.1: PTC Sensitivity Data</td>
<td>86</td>
</tr>
<tr>
<td>3.6.3 Example 3.2: Screening for Hemochromatosis</td>
<td>87</td>
</tr>
<tr>
<td>3.6.4 Example 3.3: Diagnostic Criteria for Diabetes</td>
<td>89</td>
</tr>
<tr>
<td>3.7 Examples of Multivariate Normal Mixtures</td>
<td>90</td>
</tr>
<tr>
<td>3.7.1 Example 3.4: Crab Data</td>
<td>90</td>
</tr>
<tr>
<td>3.7.2 Example 3.5: Hemophilia Data</td>
<td>92</td>
</tr>
<tr>
<td>3.8 Properties of MLE for Normal Components</td>
<td>94</td>
</tr>
<tr>
<td>3.8.1 Heteroscedastic Components</td>
<td>94</td>
</tr>
<tr>
<td>3.8.2 Homoscedastic Components</td>
<td>96</td>
</tr>
<tr>
<td>3.9 Options</td>
<td>97</td>
</tr>
<tr>
<td>3.9.1 Choice of Local Maximizer</td>
<td>97</td>
</tr>
<tr>
<td>3.9.2 Choice of Model for Component-Covariance Matrices</td>
<td>97</td>
</tr>
<tr>
<td>3.9.3 Starting Points</td>
<td>98</td>
</tr>
<tr>
<td>3.10 Spurious Local Maximizers</td>
<td>99</td>
</tr>
<tr>
<td>3.10.1 Introduction</td>
<td>99</td>
</tr>
<tr>
<td>3.10.2 Example 3.6: Synthetic Data Set 2</td>
<td>100</td>
</tr>
<tr>
<td>3.10.3 Example 3.7: Synthetic Data Set 3</td>
<td>102</td>
</tr>
<tr>
<td>3.10.4 Example 3.8: Modeling Hemophilia Data under Heteroscedasticity</td>
<td>103</td>
</tr>
<tr>
<td>3.10.5 Detection of Spurious Local Maximizers</td>
<td>103</td>
</tr>
<tr>
<td>3.10.6 Example 3.9: Galaxy Data Set</td>
<td>104</td>
</tr>
<tr>
<td>3.11 Example 3.10: Prevalence of Local Maximizers</td>
<td>105</td>
</tr>
<tr>
<td>3.12 Alternative Models for Component-Covariance Matrices</td>
<td>109</td>
</tr>
<tr>
<td>3.12.1 Spectral Representation</td>
<td>109</td>
</tr>
<tr>
<td>3.12.2 Example 3.11: Minefield Data Set</td>
<td>110</td>
</tr>
<tr>
<td>3.13 Some Other Models</td>
<td>112</td>
</tr>
<tr>
<td>3.13.1 Clustering of Treatment Means in ANOVA</td>
<td>112</td>
</tr>
<tr>
<td>3.13.2 Three-Way Models</td>
<td>114</td>
</tr>
<tr>
<td>3.13.3 Example 3.12: Consumer Data on Cat Food</td>
<td>114</td>
</tr>
<tr>
<td>3.13.4 Errors-In-Variables Model</td>
<td>116</td>
</tr>
<tr>
<td>4 Bayesian Approach to Mixture Analysis</td>
<td>117</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>117</td>
</tr>
<tr>
<td>4.2 Estimation for Proper Priors</td>
<td>119</td>
</tr>
</tbody>
</table>
5.7 Poisson Regression Model
  5.7.1 Some Standard Modifications for Overdispersed Data 150
  5.7.2 Gamma-Poisson Mixture Model 151
  5.7.3 Multiplicative Random Effects Model 153
  5.7.4 Additive Random Effects Model 153
5.8 Finite Mixture of Poisson Regression Models 154
  5.8.1 Mean and Variance 154
  5.8.2 Identifiability 155
  5.8.3 Example 5.2: Fabric Faults Data Set 155
  5.8.4 Components and Mixing Proportions Without Covariates 157
  5.8.5 Algorithms for NPMLE of a Mixing Distribution 158
  5.8.6 Disease Mapping 158
5.9 Count Data with Excess Zeros 159
  5.9.1 History of Problem 160
  5.9.2 Zero-Inflated Poisson Regression 160
5.10 Logistic Regression Model 160
5.11 Finite Mixtures of Logistic Regressions 162
  5.11.1 Mean and Variance 162
  5.11.2 Mixing at the Binary Level 163
  5.11.3 Identifiability 164
  5.11.4 Example 5.3: Beta-Blockers Data Set 165
5.12 Latent Class Models 166
5.13 Hierarchical Mixtures-of-Experts Model 167
  5.13.1 Mixtures-of-Experts Model 167
  5.13.2 Hierarchical Mixtures-of-Experts 169
  5.13.3 Application of EM Algorithm to HME Model 171
  5.13.4 Example 5.4: Speech Recognition Problem 172
  5.13.5 Pruning HME Tree Structures 174
6 Assessing the Number of Components in Mixture Models 175
  6.1 Introduction 175
  6.1.1 Some Practical Issues 175
  6.1.2 Order of a Mixture Model 176
  6.1.3 Example 6.1: Adjusting for Effect of Skewness on the LRT 177
  6.2 Example 6.2: 1872 Hidalgo Stamp Issue of Mexico 179
CONTENTS

6.3 Approaches for Assessing Mixture Order 184
  6.3.1 Main Approaches 184
  6.3.2 Nonparametric Methods 184
  6.3.3 Method of Moments 185
6.4 Likelihood Ratio Test Statistic 185
  6.4.1 Introduction 185
  6.4.2 Example 6.3: Breakdown in Regularity Conditions 186
6.5 Distributional Results for the LRTS 187
  6.5.1 Some Theoretical Results 187
  6.5.2 Some Simulation Results 189
  6.5.3 Mixtures of Two Unrestricted Normal Components 190
  6.5.4 Mixtures of Two Exponentials 191
6.6 Bootstrapping the LRTS 192
  6.6.1 Implementation 192
  6.6.2 Application to Three Real Data Sets 194
  6.6.3 Applications in Astronomy 196
6.7 Effect of Estimates on P-Values of Bootstrapped LRTS 198
  6.7.1 Some Simulation Results 198
  6.7.2 Double Bootstrapping 200
6.8 Information Criteria in Model Selection 202
  6.8.1 Bias Correction of the Log Likelihood 202
  6.8.2 Akaike's Information Criterion 203
  6.8.3 Bootstrap-Based Information Criterion 203
  6.8.4 Cross-Validation-Based Information Criterion 205
  6.8.5 Minimum Information Ratio Criterion 206
  6.8.6 Informational Complexity Criterion 207
6.9 Bayesian-Based Information Criteria 207
  6.9.1 Bayesian Approach 207
  6.9.2 Laplace's Method of Approximation 208
  6.9.3 Bayesian Information Criterion 209
  6.9.4 Laplace-Metropolis Criterion 210
  6.9.5 Laplace-Empirical Criterion 211
  6.9.6 Reversible Jump Method 212
  6.9.7 MML Principle 212
6.10 Classification-Based Information Criteria 212
  6.10.1 Classification Likelihood Criterion 212
  6.10.2 Normalized Entropy Criterion 214
6.10.3 Integrated Classification Likelihood Criterion 215
6.11 An Empirical Comparison of Some Criteria 217
6.11.1 Simulated Set 1 218
6.11.2 Simulated Set 2 218
6.11.3 Simulated Set 3 219
6.11.4 Conclusions from Simulations 220

7 Multivariate t Mixtures 221
7.1 Introduction 221
7.2 Previous Work 222
7.3 Robust Clustering 222
7.4 Multivariate t Distribution 223
7.5 ML Estimation of Mixture of t Distributions 224
7.5.1 Application of EM Algorithm 224
7.5.2 E-Step 225
7.5.3 M-Step 227
7.5.4 Application of ECM Algorithm 229
7.6 Previous Work on M-Estimation of Mixture Components 230
7.7 Example 7.1: Simulated Noisy Data Set 231
7.8 Example 7.2: Crab Data Set 234
7.9 Example 7.3: Old Faithful Geyser Data Set 236

8 Mixtures of Factor Analyzers 238
8.1 Introduction 238
8.2 Principal Component Analysis 239
8.3 Single-Factor Analysis Model 240
8.4 EM Algorithm for a Single-Factor Analyzer 241
8.5 Data Visualization in Latent Space 243
8.6 Mixtures of Factor Analyzers 244
8.7 AECM Algorithm for Fitting Mixtures of Factor Analyzers 245
8.7.1 AECM Framework 245
8.7.2 First Cycle 245
8.7.3 Second Cycle 246
8.7.4 Representation of Original Data 248
8.8 Link of Factor Analysis with Probabilistic PCA 248
8.9 Mixtures of Probabilistic PCAs 250
8.10 Initialization of AECM Algorithm 250
8.11 Example 8.1: Simulated Data 252
8.12 Example 8.2: Wine Data 254

9 Fitting Mixture Models to Binned Data 257
9.1 Introduction 257
9.2 Binned and Truncated Data 258
9.3 Application of EM Algorithm 259
  9.3.1 Missing Data 259
  9.3.2 E-Step 260
  9.3.3 M-Step 261
  9.3.4 M-Step for Normal Components 261
9.4 Practical Implementation of EM Algorithm 262
  9.4.1 Computational Issues 262
  9.4.2 Numerical Integration at Each EM Iteration 262
  9.4.3 Integration over Truncated Regions 263
  9.4.4 EM Algorithm for Binned Multivariate Data 264
9.5 Simulations 264
9.6 Example 9.1: Red Blood Cell Data 265

10 Mixture Models for Failure-Time Data 268
10.1 Introduction 268
10.2 Competing Risks 269
  10.2.1 Mixtures of Survival Functions 269
  10.2.2 Latent Failure-Time Approach 270
  10.2.3 ML Estimation for Mixtures of Survival Functions 271
10.3 Example 10.1: Heart-Valve Data 272
  10.3.1 Description of Problem 272
  10.3.2 Mixture Models with Unconstrained Components 273
  10.3.3 Constrained Mixture Models 274
  10.3.4 Conditional Probability of a Reoperation 276
  10.3.5 Advantages of Mixture Model-Based Approach 276
10.4 Long-Term Survivor Model 277
  10.4.1 Definition 277
  10.4.2 Modified Long-Term Survivor Model 278
  10.4.3 Partial ML Approach for Modified Long-Term Survival Model 279
10.4.4 Interpretation of Cure Rate in Presence of Competing Risks 280
10.4.5 Example 9.2: Breast Cancer Data 280

10.5 Analysis of Masked System-Life Data 283
10.5.1 Masked Cause of Failure 283
10.5.2 Application of EM Algorithm 283
10.5.3 Exponential Components 284
10.5.4 Weibull Components 285

11 Mixture Analysis of Directional Data 287
11.1 Introduction 287
11.2 Joint Sets 287
11.3 Directional Data 291
11.4 Initial Work on Clustering of Directional Data 292
11.5 Mixture of Kent Distributions 292
11.6 Moment Estimation of Kent Distribution 293
11.7 Uniform Component for Background Noise 295
11.8 Application of EM Algorithm 296
11.9 Example 11.1: Two Mining Samples 297
11.10 Determining the Number of Joint Sets 298
11.11 Discussion 301

12 Variants of the EM Algorithm for Large Databases 302
12.1 Introduction 302
12.2 Incremental EM Algorithm 303
12.2.1 Introduction 303
12.2.2 Definition of Partial E-Step 303
12.2.3 Block Updating of Sufficient Statistics 303
12.2.4 Justification of IEM Algorithm 305
12.2.5 Gain in Convergence Time 305
12.2.6 IEM Algorithm for Singleton Blocks 306
12.2.7 Efficient Updating Formulas 306
12.3 Simulations for IEM Algorithm 307
12.3.1 Simulation 1 307
12.3.2 Simulation 2 309
12.4 Lazy EM Algorithm 310
12.5 Sparse EM Algorithm 311
12.6 Sparse IEM Algorithm 312
12.6.1 Some Simulation Results 312
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.6.2</td>
<td>Summary of Results for the IEM and SPIEM Algorithms</td>
<td>315</td>
</tr>
<tr>
<td>12.7</td>
<td>A Scalable EM Algorithm</td>
<td>316</td>
</tr>
<tr>
<td>12.7.1</td>
<td>Introduction</td>
<td>316</td>
</tr>
<tr>
<td>12.7.2</td>
<td>Primary Compression of the Data</td>
<td>316</td>
</tr>
<tr>
<td>12.7.3</td>
<td>Updating of Parameter Estimates</td>
<td>318</td>
</tr>
<tr>
<td>12.7.4</td>
<td>Merging of Sufficient Statistics</td>
<td>319</td>
</tr>
<tr>
<td>12.7.5</td>
<td>Secondary Data Compression</td>
<td>319</td>
</tr>
<tr>
<td>12.7.6</td>
<td>Tuning Constants</td>
<td>320</td>
</tr>
<tr>
<td>12.7.7</td>
<td>Simulation Results</td>
<td>321</td>
</tr>
<tr>
<td>12.8</td>
<td>Multiresolution KD-Trees</td>
<td>323</td>
</tr>
<tr>
<td>12.8.1</td>
<td>Introduction</td>
<td>323</td>
</tr>
<tr>
<td>12.8.2</td>
<td>EM Algorithm Based on Multiresolution KD-Trees</td>
<td>323</td>
</tr>
<tr>
<td>13</td>
<td>Hidden Markov Models</td>
<td>326</td>
</tr>
<tr>
<td>13.1</td>
<td>Introduction</td>
<td>326</td>
</tr>
<tr>
<td>13.2</td>
<td>Hidden Markov Chain</td>
<td>328</td>
</tr>
<tr>
<td>13.2.1</td>
<td>Definition</td>
<td>328</td>
</tr>
<tr>
<td>13.2.2</td>
<td>Some Examples</td>
<td>329</td>
</tr>
<tr>
<td>13.3</td>
<td>Applying EM Algorithm to Hidden Markov Chain Model</td>
<td>329</td>
</tr>
<tr>
<td>13.3.1</td>
<td>EM Framework</td>
<td>329</td>
</tr>
<tr>
<td>13.3.2</td>
<td>E-Step</td>
<td>330</td>
</tr>
<tr>
<td>13.3.3</td>
<td>Forward–Backward Recursions on E-Step</td>
<td>330</td>
</tr>
<tr>
<td>13.3.4</td>
<td>M-Step</td>
<td>332</td>
</tr>
<tr>
<td>13.3.5</td>
<td>Numerical Instabilities</td>
<td>332</td>
</tr>
<tr>
<td>13.4</td>
<td>Hidden Markov Random Field</td>
<td>332</td>
</tr>
<tr>
<td>13.4.1</td>
<td>Specification of Markov Random Field</td>
<td>333</td>
</tr>
<tr>
<td>13.4.2</td>
<td>Application of EM Algorithm</td>
<td>333</td>
</tr>
<tr>
<td>13.4.3</td>
<td>Restoration Step</td>
<td>334</td>
</tr>
<tr>
<td>13.4.4</td>
<td>An Improved Approximation to EM Solution</td>
<td>335</td>
</tr>
<tr>
<td>13.4.5</td>
<td>Approximate M-Step for Normal Components</td>
<td>336</td>
</tr>
<tr>
<td>13.5</td>
<td>Example 13.1: Segmentation of MR Images</td>
<td>336</td>
</tr>
<tr>
<td>13.6</td>
<td>Bayesian Approach</td>
<td>338</td>
</tr>
<tr>
<td>13.7</td>
<td>Examples of Gibbs Sampling with Hidden Markov Chains</td>
<td>339</td>
</tr>
</tbody>
</table>

Appendix Mixture Software 343
CONTENTS

A.1 EMMIX 343
A.2 Some Other Mixture Software 345

References 349

Author Index 395

Subject Index 407
Preface

The importance of finite mixture models in the statistical analysis of data is underscored by the ever-increasing rate at which articles on mixture applications appear in the statistical and general scientific literature. The aim of this monograph is to provide an up-to-date account of the theory and applications of modeling via finite mixture distributions. Since the appearance of the monograph of McLachlan and Basford (1988) on finite mixtures, the literature has expanded enormously to the extent that another monograph on the topic is apt. In the past decade the extent and the potential of the applications of finite mixture models have widened considerably. Because of their flexibility, mixture models are being increasingly exploited as a convenient, semiparametric way in which to model unknown distributional shapes. This is in addition to their obvious applications where there is group-structure in the data or where the aim is to explore the data for such structure, as in a cluster analysis.

In this book the more recent work is surveyed against the background of the existing literature. The widespread use of mixture models in recent times is demonstrated by the fact that of the 800 or so references in this book, almost 40% of them have been published since 1995. A comprehensive account of the major issues involved with modeling via finite mixture distributions is provided. They include identifiability problems, the actual fitting of finite mixtures through use of the EM algorithm, the properties of the maximum likelihood estimators so obtained, the assessment of the number of components to be used in the mixture, and the applicability of asymptotic theory in providing a basis for the solutions to some of these problems. The intent is to provide guidelines to users of mixture models on these various issues. The emphasis is on the applications of mixture models, not only in mainstream statistical analyses, but also in other areas such as unsupervised pattern recognition, speech recognition, and medical imaging.

With the advent of inexpensive, high-speed computers and the simultaneous rapid development in posterior simulation techniques such as Markov chain Monte Carlo
(MCMC) methods for enabling Bayesian estimation to be undertaken, practitioners are increasingly turning to Bayesian methods for the analysis of complicated statistical models. In this book, we consider the latest developments in Bayesian estimation of mixture models.

New topics that are covered in this book include the scaling of the EM algorithm to allow mixture models to be used in data mining applications involving massively huge databases. In the same spirit, there is also an account of the use of the sparse/incremental EM algorithm and of multiresolution \( kd \)-trees for speeding up the implementation of the standard EM algorithm for the fitting of mixture models. Another topic concerns the use of hierarchical mixtures-of-experts models as a powerful new approach to nonlinear regression that is a serious competitor to well-known statistical procedures such as MARS and CART. Other recent developments covered include the use of mixture models for handling overdispersion in generalized linear models and proposals for dealing with mixed continuous and categorical variables. In other recent work, there is the proposal to use \( t \) components in the mixture model to provide a robust approach to mixture modeling. A further topic is the use of mixtures of factor analyzers which provide a way of fitting mixture models to high-dimensional data. As mixture models provide a convenient basis for the modeling of dependent data by allowing the component-indicator variables to have a Markovian structure, there is also coverage of the latest developments in hidden Markov models, including the Bayesian approach to this problem. Another problem considered is the fitting of mixture models to multivariate data in binned form, which arises in some important medical applications in practice.

The book also covers the latest developments on existing issues with mixture modeling, such as assessing the number of components to be used in a mixture model and the associated problem of determining how many clusters there are in clustering applications with mixture models.

In presenting these latest results, the authors have attempted to draw together the statistical literature with the machine learning and pattern recognition literature.

It is intended that the book should appeal to both applied and theoretical statisticians, as well as to investigators working in the many diverse areas in which relevant use can be made of finite mixture models. It will be assumed that the reader has a fair mathematical or statistical background. The main parts of the book describing the formulation of the finite mixture approach, 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. The emphasis is on the practical applications of mixture models; and to this end, numerous examples are given.

Chapter 1 begins with a discussion of mixture models and their applications and gives a brief overview of the current state of the area. It also includes a brief history of mixture models.

Chapter 2 focuses on the maximum likelihood fitting of mixture models via the EM algorithm. It covers issues such as the choice of starting values, stopping criteria, the calculation of the observed information matrix, and the provision of standard errors either by information-based methods or by the bootstrap.
Chapter 3 specializes the results in Chapter 2 to mixtures of normal components. Given the tractability of the multivariate normal distribution, it is not surprising that mixture modeling of continuous data is invariably undertaken by normal mixtures. This chapter also discusses the occurrence and identification of so-called spurious local maximizers of the likelihood function, which is an issue with the fitting of normal mixture models with no restrictions on the component-covariance matrices. A number of illustrative examples are presented.

In Chapter 4, we consider the Bayesian approach to the fitting of mixture models. Estimation in a Bayesian framework is now feasible using posterior simulation via recently developed MCMC methods. Bayes estimators for mixture models are well-defined so long as the prior distributions are proper. One main hindrance is that improper priors yield improper posterior distributions. We discuss the use of "partially proper priors," which do not require subjective input for the component parameters, yet the posterior is proper. We also discuss ways of handling other hindrances, including the effect of label switching, which arises when there is no real prior information that allows one to discriminate between the components of a mixture model belonging to the same parametric family.

In Chapter 5, we consider the fitting of mixture models with nonnormal component densities, including components suitable for mixed feature variables, where some are continuous and some are categorical. The maximum likelihood fitting of commonly used discrete components such as the binomial and Poisson are undertaken within the wider framework of a mixture of generalized linear models (GLMs). The latter also has the capacity to handle the regression case, where the response is allowed to depend on the value of a vector of covariates. The use of mixtures of GLMs for handling overdispersion in a single GLM component is discussed. In work related to mixtures of GLMs, the mixtures-of-experts model is considered, along with its extension, the hierarchical mixtures-of-experts model. This approach which combines aspects of finite mixture models and GLMs provides a comparatively fast learning and good generalization for nonlinear regression problems, including classification.

Chapter 6 is devoted to the estimation of the order of a mixture model. It covers the two main approaches. One way is based on a penalized form of the log likelihood whereby the likelihood is penalized by the subtraction of a term that "penalizes" the model for the number of parameters in it. The other main way for deciding on the order of a mixture model is to carry out a hypothesis test, using the likelihood ratio as the test statistic.

In Chapter 7, we consider the fitting of mixtures of (multivariate) $t$ distributions, as proposed in McLachlan and Peel (1998a) and Peel and McLachlan (2000). The $t$ distribution provides a longer-tailed alternative to the normal distribution. Hence it provides a more robust approach to the fitting of normal mixture models, as observations that are atypical of a component are given reduced weight in the calculation of its parameters. Also, the use of $t$ components gives less extreme estimates of the posterior probabilities of component membership of the mixture model.

In Chapter 8, we consider mixtures of factor analyzers from the perspective of both (a) a method for model-based density estimation from high-dimensional data, and hence for the clustering of such data, and (b) a method for local dimensionality
reduction. We also discuss the close link of mixtures of factor analyzers with mixtures of probabilistic principal component analyzers. The mixtures of factor analyzers model enables a normal mixture model to be fitted to high-dimensional data. The number of free parameters is controlled through the dimension $q$ of the latent factor space. It allows thus an interpolation in model complexities from isotropic to full covariance structures without any restrictions.

In Chapter 9, we consider the fitting of finite mixture models to binned and truncated multivariate data by maximum likelihood via the EM algorithm. The solution for an arbitrary number of dimensions of the feature vector is specialized to the case of bivariate normal mixtures.

Chapter 10 is on the use of mixture distributions to model failure-time data in a variety of situations, which occur in reliability and survival analyses. The focus is on the use of mixture distributions to model time to failure in the case of competing risks or failures.

In Chapter 11, a case study is provided to illustrate the use of mixture models in the analysis of multivariate directional data. Mixtures of Kent distributions are used as an aid in joint set identification.

In Chapter 12, we consider methods for improving the speed of the EM algorithm for the maximum likelihood fitting of mixture models to large databases that preserve the simplicity of implementation of the EM in its standard form. They include the incremental version of the EM algorithm, where only a partial E-step is performed before each M-step, and a sparse version, where not all the posterior probabilities of component membership are updated on each iteration. The use of multiresolution $kd$-trees to speed up the implementation of the E-step is also described. In addition, we consider how the EM algorithm can be scaled to handle very large databases with a limited memory buffer.

In Chapter 13, recent advances on hidden Markov models are covered. Hidden Markov models are increasingly being adopted in applications, since they provide a convenient way of formulating an extension of a mixture model to allow for dependent data. We discuss hidden Markov chain models in the one-dimensional case and hidden Markov random fields in two or higher dimensions.

A brief account of some of the available software for the fitting of mixture models is provided in the Appendix. This account includes a description of the program EMMIX (McLachlan et al., 1999).

The authors wish to thank Dr. Angus Ng for many helpful and insightful discussions on mixture models, and Katrina Monico for her constructive comments and suggestions on drafts of the manuscript. They would also like to acknowledge gratefully financial support from the Australian Research Council. Thanks are due too to the authors and owners of copyrighted material for permission to reproduce tables and figures.

Brisbane, Australia

Geoffrey J. McLachlan
David Peel
1

General Introduction

1.1 INTRODUCTION

1.1.1 Flexible Method of Modeling

Finite mixtures of distributions have provided a mathematical-based approach to the statistical modeling of a wide variety of random phenomena. Because of their usefulness as an extremely flexible method of modeling, finite mixture models have continued to receive increasing attention over the years, from both a practical and theoretical point of view. Indeed, in the past decade the extent and the potential of the applications of finite mixture models have widened considerably. Fields in which mixture models have been successfully applied include astronomy, biology, genetics, medicine, psychiatry, economics, engineering, and marketing, among many other fields in the biological, physical, and social sciences. In these applications, finite mixture models underpin a variety of techniques in major areas of statistics, including cluster and latent class analyses, discriminant analysis, image analysis, and survival analysis, in addition to their more direct role in data analysis and inference of providing descriptive models for distributions.
model unknown distributional shapes, whatever the objective, whether it be, say, density estimation or the flexible construction of Bayesian priors. For example, Priebe (1994) showed that with \( n = 10,000 \) observations, a log normal density can be well approximated by a mixture of about 30 normals. In contrast, a kernel density estimator uses a mixture of 10,000 normals. A mixture model is able to model quite complex distributions through an appropriate choice of its components to represent accurately the local areas of support of the true distribution. It can thus handle situations where a single parametric family is unable to provide a satisfactory model for local variations in the observed data. Inferences about the modeled phenomenon can be made without difficulties from the mixture components, since the latter are chosen for their tractability. This flexibility allows mixture models to play a useful role in neural networks (Bishop, 1995, Section 5.9). For with neural networks formed using radial basis functions, the input data can be modeled by a mixture model (for example, a normal mixture). That is, the basis functions can be taken to be the components of this mixture model after estimation by maximum likelihood from the input data. The second-layer weights in the neural network can then be estimated from the input data and their known outputs.

1.1.2 Initial Approach to Mixture Analysis

One of the first major analyses involving the use of mixture models was undertaken just over 100 years ago by the famous biometrician Karl Pearson. In his now classic paper, Pearson (1894) fitted a mixture of two normal probability density functions with different means \( \mu_1 \) and \( \mu_2 \) and variances \( \sigma_1^2 \) and \( \sigma_2^2 \) in proportions \( \pi_1 \) and \( \pi_2 \) to some data provided by Weldon (1892, 1893). The latter paper may have been the first ever to advocate statistical analysis as a primary method for studying biological problems (Pearson, 1906); see Stigler (1986, Chapter 8) and Tarter and Lock (1993, Chapter 1) for a more detailed account. The data set analyzed by Pearson (1894) consisted of measurements on the ratio of forehead to body length of \( n = 1000 \) crabs sampled from the Bay of Naples. These measurements, which were recorded in the form of \( v = 29 \) intervals, are displayed in Figure 1.1, along with the plot of the density of a single normal distribution fitted to them. Weldon (1893) had speculated that the asymmetry in the histogram of these data might be a signal that this population was evolving toward two new subspecies. Sensing that his own mathematical training was inadequate, Weldon turned to his colleague Karl Pearson for assistance.

Pearson's (1894) mixture model-based approach suggested that there were two subspecies present. This paper was the first of two monstrous memoirs in a series of "Contributions to the Mathematical Theory of Evolution" (Stigler, 1986, Chapter 10). In Figure 1.1 we have plotted the density of the two-component normal mixture, as obtained by using maximum likelihood to fit this model to the data in their original interval form. Pearson (1894) had used the method of moments to fit this mixture model to the mid-points of the intervals, which, for this data set, gives a fit very similar to that obtained with the more efficient method of maximum likelihood. It can be seen
that this mixture of two normal heteroscedastic components (components with unequal variances) does the job that Pearson (1894) had intended it to do, and that was to accommodate the apparent skewness in the data, which cannot be modeled adequately by the symmetric normal distribution. Pearson (1894) obtained his moments-based estimates of the five parameters of his normal heteroscedastic mixture model as a solution of a ninth degree polynomial (nonic). The computational effort in fitting this model (that is, finding the roots of a nonic) must have been at the time a daunting prospect to potential users of this mixture methodology. Indeed, as Everitt (1996) noted in his historical review of the development of finite mixture models, Charlier (1906) was led to comment "The solution of an equation of the ninth degree, where almost all powers, to the ninth, of the unknown quantity are existing, is, however, a very laborious task. Mr. Pearson has indeed possessed the energy to perform this heroic task in some instances in his first memoir on these topics from the year 1894. But I fear that he will have few successors, if the dissection of the frequency curve into two components is not very urgent." Not surprisingly, various attempts were made over the ensuing years to simplify Pearson’s (1894) moments-based approach to the fitting of a normal mixture model.

![Plot of forehead to body length data on 1000 crabs and of the fitted one-component (dashed line) and two-component (solid line) normal mixture models.](image)

**Fig. 1.1** Plot of forehead to body length data on 1000 crabs and of the fitted one-component (dashed line) and two-component (solid line) normal mixture models.

### 1.1.3 Impact of EM Algorithm

As to be elaborated on in Section 1.1.8, where we give a brief history of finite mixture models, it has only been in the last 20 or so years that considerable advances have been made in the fitting of finite mixture models, in particular by the method of maximum likelihood. Even with the advent of high-speed computers, there had been some reluctance in the past to fit mixture models to data of more than one dimension, possibly because of a lack of understanding of issues that arise with their fitting. They include the presence of multiple maxima in the mixture likelihood function and the unboundedness of the likelihood function in the case of normal
components with unequal covariance matrices. But as the difficulties concerning these computational issues came to be properly understood and successfully addressed, it led to the increasing use of mixture models in practice.

In the 1960s, the fitting of finite mixture models by maximum likelihood had been studied in a number of papers, including the seminal papers by Day (1969) and Wolfe (1965, 1967, 1970). However, it was the publication of the seminal paper of Dempster, Laird, and Rubin (1977) on the EM algorithm that greatly stimulated interest in the use of finite mixture distributions to model heterogeneous data. This is because the fitting of mixture models by maximum likelihood is a classic example of a problem that is simplified considerably by the EM’s conceptual unification of maximum likelihood (ML) estimation from data that can be viewed as being incomplete. As Aitkin and Aitkin (1994) noted, almost all the post-1978 applications of mixture modeling reported in the books on mixtures by Titterington, Smith, and Makov (1985) and McLachlan and Basford (1988) use the EM algorithm; see McLachlan and Krishnan (1997, Section 1.8). This also applies to the applications in this book.

1.2 OVERVIEW OF BOOK

The use of the EM algorithm for the fitting of finite mixture models, especially normal mixture models, has been demonstrated by McLachlan and Basford (1988) for the analysis of data arising from a wide variety of fields. For most commonly used parametric formulations of finite mixture models, the use of the EM algorithm to find a local maximizer of the likelihood function is straightforward. However, a number of issues remain. For example, as the likelihood function for mixture models usually has multiple local maxima, there is the question of which root of the likelihood equation corresponding to a local maximum of the likelihood function (that is, which local maximizer) to choose as the estimate of the vector of unknown parameters. Typically, the desired root corresponds to the global maximizer of the likelihood function in those situations where the likelihood function is bounded over the parameter space. But with mixtures of normal components with unequal variances in the univariate case or unequal covariance matrices in the multivariate case, the likelihood function is unbounded. In this case, the choice of root of the likelihood equation is not as obvious as in the bounded case and so requires careful consideration in practice. There is also the associated problem of how to select suitable starting values for the EM algorithm in the search of appropriate roots of the likelihood equation in the first instance. Another important consideration with the fitting of finite mixture models concerns the choice of the number of components \( g \) in the mixture model in those applications where \( g \) has to be inferred from the data.

In this book we give an extensive coverage of these problems, focusing on the latest developments. Indeed, almost 40% of the references in the book have been published since 1995.

Practitioners are increasingly turning to Bayesian methods for the analysis of complicated statistical models. This move is due in large part to the advent of inexpensive high speed computers and the simultaneous rapid development in posterior simula-
tion techniques such as Markov chain Monte Carlo (MCMC) methods for enabling Bayesian estimation to be undertaken. In this book, we consider the latest developments in Bayesian estimation of mixture models.

One of the major problems of interest in mixture models concerns the choice of the number of components. In recent times, a number of new criteria have been suggested, some of which, like the integrated classification criterion, have given encouraging results in empirical studies designed to test their performance. In this book we discuss these various criteria, along with standard criteria such as the Bayesian information criterion (BIC) and the bootstrap likelihood ratio test.

As many applications of nonnormal mixtures are with components belonging to the exponential family, consideration is given to the family of mixtures of generalized linear models (GLMs). In this framework, the mixing proportions, as well as the component distributions, are allowed to depend on some associated covariates. A common way in which mixtures of GLMs arises in practice is in the handling of overdispersion in a single GLM. A convenient way to proceed in this case is to introduce a random effect into the linear predictor and to consider a mixture of such models with different intercepts in different proportions.

In work related to mixtures of GLMs, the mixtures-of-experts model is considered, along with its extension, the hierarchical mixtures-of-experts (HMEs) model. This approach, which combines aspects of finite mixture models and GLMs, provides a comparatively fast learning and good generalization for nonlinear regression problems, including classification. It is thus a serious competitor to well-known statistical procedures such as MARS and CART.

On other material concerning the use of nonnormal components, there is treatment of the case where the feature variables are mixed, with some being categorical and some continuous. There is also a separate chapter devoted to mixture distributions in modeling failure-time data with competing risks. A case study is presented on the use of mixtures of Kent distributions for the analysis of multivariate directional data.

The problem of fitting finite mixture models to binned and truncated data is also covered. The methodology is illustrated with a case study on the diagnosis of iron-deficient anemia by the mixture modeling of binned and truncated data on a patient in the form of volume and hemoglobin concentration of red blood cells, as measured by a cytometric blood cell counter.

This book contains a number of recent results on mixture models by the authors, including the use of mixtures of $t$ distributions to provide a robust extension of normal mixture models, and the use of mixtures of factor analyzers to enable the fitting of normal mixture models to high-dimensional data. With the considerable attention being given to the analysis of large data sets, as in typical data mining applications, recent work on speeding up the implementation of the EM algorithm is discussed, including (a) the use of the sparse/incremental EM and of multiresolution $kd$-trees and (b) the scaling of the EM algorithm to massively large databases where there is a limited memory buffer.

Hidden Markov models are increasingly being used, as they provide a way of formulating an extension of mixture models to allow for dependent data. This book reviews the latest results on maximum likelihood and Bayesian methods of estimation.
for such models. It concludes with a concise summary of software available for the fitting of mixture models, including the EMMIX program (McLachlan et al., 1999).

Numerous examples of applications of mixture models are given throughout the book to demonstrate the methodology. Where available, the data sets considered in this book may be found on the World Wide Web at http://www.maths.uq.edu.au/~gjm.

1.3 BASIC DEFINITION

We let \( Y_1, \ldots, Y_n \) denote a random sample of size \( n \), where \( Y_j \) is a \( p \)-dimensional random vector with probability density function \( f(y_j) \) on \( \mathbb{R}^p \). In practice, \( Y_j \) contains the random variables corresponding to \( p \) measurements made on the \( j \)th recording of some features on the phenomenon under study. We let \( Y = (Y_1^T, \ldots, Y_n^T)^T \), where the superscript \( T \) denotes vector transpose. Note that we are using \( Y \) to represent the entire sample; that is, \( Y \) is an \( n \)-tuple of points in \( \mathbb{R}^p \). Where possible, a realization of a random vector is denoted by the corresponding lower-case letter. For example, \( y = (y_1^T, \ldots, y_n^T)^T \) denotes an observed random sample where \( y_j \) is the observed value of the random vector \( Y_j \).

Although we are taking the feature vector \( Y_j \) to be a continuous random vector here, we can still view \( f(y_j) \) as a density in the case where \( Y_j \) is discrete by the adoption of counting measure. We suppose that the density \( f(y_j) \) of \( Y_j \) can be written in the form

\[
\sum_{i=1}^{g} \pi_i f_i(y_j),
\]

where the \( f_i(y_j) \) are densities and the \( \pi_i \) are nonnegative quantities that sum to one; that is,

\[
0 \leq \pi_i \leq 1 \quad (i = 1, \ldots, g)
\]

and

\[
\sum_{i=1}^{g} \pi_i = 1.
\]

The quantities \( \pi_1, \ldots, \pi_g \) are called the mixing proportions or weights. As the functions, \( f_1(y_j), \ldots, f_g(y_j), \) are densities, it is obvious that (1.1) defines a density. The \( f_i(y_j) \) are called the component densities of the mixture. We shall refer to the density (1.1) as a \( g \)-component finite mixture density and refer to its corresponding distribution function \( F(y_j) \) as a \( g \)-component finite mixture distribution. Since we shall be focusing almost exclusively on finite mixtures of distributions, we shall usually refer to finite mixture models as just mixture models in the sequel.

In this formulation of the mixture model, the number of components \( g \) is considered fixed. But of course in many applications, the value of \( g \) is unknown and has to be inferred from the available data, along with the mixing proportions and the parameters in the specified forms for the component densities.

When the number of components is allowed to increase with the sample size \( n \), the model is called a Gaussian mixture sieve; see Geman and Hwang (1982), Roeder
1.4 INTERPRETATION OF MIXTURE MODELS

An obvious way of generating a random vector $Y_j$ with the $g$-component mixture density $f(y_j)$, given by (1.1), is as follows. Let $Z_j$ be a categorical random variable taking on the values $1, \ldots, g$ with probabilities $\pi_1, \ldots, \pi_g$, respectively, and suppose that the conditional density of $Y_j$ given $Z_j = i$ is $f_i(y_j)$ ($i = 1, \ldots, g$). Then the unconditional density of $Y_j$ (that is, its marginal density) is given by $f(y_j)$. In this context, the variable $Z_j$ can be thought of as the component label of the feature vector $Y_j$. In later work, it is convenient to work with a $g$-dimensional component-label vector $Z_j$ in place of the single categorical variable $Z_j$, where the $i$th element of $Z_j$, $Z_{ij} = (Z_j)_i$, is defined to be one or zero, according to whether the component of origin of $Y_j$ in the mixture is equal to $i$ or not ($i = 1, \ldots, g$). Thus $Z_j$ is distributed according to a multinomial distribution consisting of one draw on $g$ categories with probabilities $\pi_1, \ldots, \pi_g$; that is,

$$\text{pr}\{Z_j = z_j\} = \pi_1^{z_{1j}} \pi_2^{z_{2j}} \ldots \pi_g^{z_{gj}}. \quad (1.4)$$

We write

$$Z_j \sim \text{Mult}_g(1, \pi), \quad (1.5)$$

where $\pi = (\pi_1, \ldots, \pi_g)^T$.

In the interpretation above of a mixture model, an obvious situation where the $g$-component mixture model (1.1) is directly applicable is where $Y_j$ is drawn from a population $G$ which consists of $g$ groups, $G_1, \ldots, G_g$, in proportions $\pi_1, \ldots, \pi_g$. If the density of $Y_j$ in group $G_i$ is given by $f_i(y_j)$ for $i = 1, \ldots, g$, then the density of $Y_j$ has the $g$-component mixture form (1.1). In this situation, the $g$ components of the mixture can be physically identified with the $g$ externally existing groups, $G_1, \ldots, G_g$.

In biometric applications for instance, a source of the heterogeneity is often age, sex, species, geographical origin, and cohort status. For example, a population $G$ may consist of two groups $G_1$ and $G_2$, corresponding to those members with or without a particular disease that is under study. The problem may be to estimate the disease prevalence (that is, the mixing proportion $\pi_1$ here) on the basis of some feature vector measured on a randomly selected sample of members of the population. In the case study of Do and McLachlan (1984), in which $p = 4$ variables were measured on the skulls of Malaysian rats collected from owl pellets, the components of the fitted mixture corresponded to $g = 7$ different species of rats. The aim of their study was to assess the rat diet of owls in terms of the proportion of each species of rat represented in the fitted mixture model.

We shall see in this book that there are many other examples in practice where the population is a mixture of $g$ distinct groups that are known a priori to exist in some physical sense. However, there are also many examples involving the use of mixture