APPROVAL SHEET

Title of Dissertation: Computational Methods in Finite Mixtures using Approximate Information and Regression Linked to the Mixture Mean

Name of Candidate: Andrew M. Raim
Doctor of Philosophy, 2014

Dissertation and Abstract Approved: ______________________________
Nagaraj K. Neerchal
Professor of Statistics
Department of Mathematics and Statistics

Date Approved: ____________________

CURRICULUM VITAE

Name: Andrew M. Raim

Education

University of Maryland, Baltimore County (UMBC)
Doctor of Philosophy, Statistics, Fall 2008 – Spring 2014
Master of Science, Statistics, Fall 2008 – Fall 2011
Master of Science, Computer Science, Spring 2003 – Fall 2008
Bachelor of Science, Computer Science, Fall 1998 – Spring 2002

Baltimore City College High School, Baltimore, MD, Class of 1998

Selected Publications

Andrew M. Raim, Nagaraj K. Neerchal, and Jorge G. Morel. An approximation to the information matrix of exponential family finite mixtures, (Submitted, 2014).

Andrew M. Raim and Nagaraj K. Neerchal. Modeling overdispersion in binomial data with regression linked to a finite mixture probability of success. In *JSM Proceedings, Statistical Computing Section. Alexandria, VA: American Statistical Association*, pages 2760–2774, 2013.

Andrew M. Raim, Minglei Liu, Nagaraj K. Neerchal, and Jorge G. Morel. On the method of approximate Fisher scoring for finite mixtures of multinomials. *Statistical Methodology*, 18:115–130, 2014.

Andrew M. Raim, Matthias K. Gobbert, Nagaraj K. Neerchal, and Jorge G. Morel. Maximum likelihood estimation of the random-clumped multinomial model as prototype problem for large-scale statistical computing. *Journal of Statistical Computation and Simulation*, 83(12):2178–2194, 2013.

Work Experience

Graduate Assistant, High Performance Computing Facility, UMBC, 2009–2014

Software Engineer, Advertising.com, Baltimore, MD, 2002–2008

Helpdesk Consultant, Office of Information Technology, UMBC, 1999–2001

ABSTRACT

Title of dissertation: COMPUTATIONAL METHODS IN FINITE MIXTURES USING APPROXIMATE INFORMATION AND REGRESSION LINKED TO THE MIXTURE MEAN

Andrew M. Raim
Doctor of Philosophy, 2014

Dissertation directed by: Nagaraj K. Neerchal
Professor of Statistics
Department of Mathematics and Statistics
University of Maryland, Baltimore County

Finite mixture distributions are used in applications because of their ability to support heterogeneity. They also present interesting analytical challenges, often requiring special consideration in the selection of an appropriate model, inference of unknown parameters, and identifiability. The main contributions of this thesis are providing an approximation to the information matrix of a finite mixture of an arbitrary member of the exponential family, and a novel extension of the generalized linear model (GLM) with an underlying finite mixture distribution.

Our approximation is equivalent to a complete data information matrix, which helps to explain previously noted connections between approximate scoring and the Expectation Maximization (EM) algorithm, and is further generalized to mixtures of an arbitrary member of the exponential family. To obtain convergence between exact and approximate information requires a "clustered sampling" assumption so that observations are sampled from the same (unknown) subpopulation of the mixture, providing an analogue to trials of a multinomial observation.

We also consider a logistic regression model using a binomial finite mixture, so that the regression model is linked to the mixture mean. This significant extension of GLM appears promising for many potential applications such as modeling overdispersion in the data. Because the mixture mean is a composite parameter which does not appear explicitly in the likelihood, model formulation and inference pose both theoretical and computational challenges. We propose a random effects model with effects drawn from a set representing enforcement of the link. Initial results show that the model is effective at capturing extra variability while supporting the regression of interest.

COMPUTATIONAL METHODS IN FINITE MIXTURES USING APPROXIMATE INFORMATION AND REGRESSION LINKED TO THE MIXTURE MEAN

by

Andrew M. Raim

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Advisory Committee:
Dr. Nagaraj K. Neerchal, Chair/Advisor
Dr. Thomas Mathew
Dr. Yaakov Malinovsky
Dr. Yi Huang
Dr. Jorge G. Morel

© Copyright by
Andrew M. Raim
2014

DEDICATION

To my parents, Marc and Ellen, and my brother, Brian.

ACKNOWLEDGMENTS

It has been a great privilege to study statistics at UMBC. Having previously completed a BS and MS degree in computer science and started a career in software engineering, I was exposed to many fascinating ideas in mathematics and statistics. I knew that I wanted to pursue a PhD, but was not sure if it was still possible for me. My opportunity came in late summer 2008, thanks to my cousin Martin Klein and his advisor Dr. Bimal Sinha. With their help, I was able to change course and begin the study of statistics.

I owe my deepest gratitude to my thesis advisor and mentor, Dr. Nagaraj Neerchal. He has been a constant source of enthusiasm and inspiration, and regularly went out of his way to make time for me. I owe my completion of this thesis to countless hours of discussion between us, along with his steady encouragement. I thank Drs. Thomas Mathew, Yaakov Malinovsky, Yi Huang, and Jorge Morel for graciously serving in my dissertation committee and for providing valuable feedback about the work. Much of this thesis is based on the joint work of Neerchal and Morel, and I am grateful that they have shared their research interests with me. Thanks to my classmate Dr. Michelle Danaher for sharing her R code to find the vertices of a general polyhedron.

A special acknowledgement goes to Dr. Matthias Gobbert, who has enthusiastically supported me as his research assistant through the UMBC High Performance Computing Facility since Fall 2009. This has given me the great opportunity to develop my interests in high performance computing while focusing on statistics.

The UMBC Department of Mathematics and Statistics has been a wonderful, supportive environment for graduate study. Thanks to Deneen Blair, Boris Alemi, Janet Burgee, Bobbie Butler, Rose Schutt, Angela McNulty, Tarsha Randolph, and all the staff who helped me during my time in the department. Thanks to all the faculty for their teaching and guidance, especially the statistics group: Drs. Bimal Sinha, Thomas Mathew, Nagaraj Neerchal, Anindya Roy, Yi Huang, Junyong Park, DoHwan Park, Kofi Adragni, and Yaakov Malinovsky. I also thank Drs. Muddappa Gowda, Jinglai Shen, Bikas Sinha, Martin Klein, and Elizabeth Stanwyck for their excellent courses. There are too many graduate students to name individually, and I do not want to leave anyone out, but it has been a pleasure having them as classmates and friends.

I would like to thank my family and friends for being patient with me since I started this endeavor, and for all of their support. Spring 2014 marks my 16th year as a student at UMBC, having started as a CMSC freshman in Fall 1998, continued as a part-time CMSC masters student in Spring 2003, and finally joined as a STAT graduate student in Fall 2008. UMBC has become like a second home to me, and I am proud to have been a part of it.

Table of Contents

List of Tables vi

List of Figures viii

1 Introduction 1
- 1.1 Mixture Formulation 5
- 1.2 Moments of Mixtures 10
- 1.3 Identifiability 11
- 1.4 Fisher Information 12
- 1.5 Estimation 14

2 Approximate Information and Scoring Under Multinomial Finite Mixtures 17
- 2.1 Introduction 17
- 2.2 Preliminaries and Notation 19
- 2.3 An Approximation to the Information Matrix 22
- 2.4 Approximate Scoring Algorithm 42
- 2.5 Simulation Studies 56
  - 2.5.1 Distance Between True and Approximate FIM 57
  - 2.5.2 Approximations to Wald and Score Test Statistics 59
  - 2.5.3 Effectiveness of AFSA method: Convergence Speed 64
  - 2.5.4 Effectiveness of AFSA method: Monte Carlo Study 66
- 2.6 Conclusions 70

3 Large Cluster Approximation to the Information Matrix under Exponential Family Finite Mixtures 74
- 3.1 Introduction 74
- 3.2 Preliminaries 76
- 3.3 Convergence of Approximate Information Matrix 80
- 3.4 Relationship to Classification Problem 98
- 3.5 Examples 100
- 3.6 Conclusions 116

4 Mixture Link Models for Binomial Data with Overdispersion 118
4.1 Introduction 118
4.2 Background 119
4.3 Model Formulation 123
4.4 Expectation and Variance under Mixture Link Model 131
4.5 Finding the vertices of A 135
4.6 Computing the Mixture Link Density 140
4.7 Density Comparison between Mixture Link with other Binomial Models with Extra Variation 147
4.8 Moment-Based Approximation to the Mixture Link Model 151
4.9 Computational Studies 165
4.9.1 Numerical Maximum Likelihood 165
4.9.2 Goodness-of-Fit Test 167
4.9.3 Chromosome Aberration Data 170
4.9.4 Mutagenic Data 179
4.10 Conclusions 184

5 Conclusion 186

A Additional Results 189

References 194

List of Tables

2.1 Estimates (and standard errors in parentheses) for the competing algorithms. FSA Hybrid produced similar results with ε_0 set to $0.001, 0.01, 0.1, 1$, and 10. . . . 66
2.2 Convergence of several competing algorithms. Hybrid FSA is shown with several choices of the warmup tolerance ε_0. Exact FSA corresponds to $\varepsilon_0 = \infty$. Note that a maximum of 100 iterations was allowed in each case. 67
2.3 Closeness between AFSA and EM estimates, over 1000 samples. Scenarios A–D represent binomial mixtures, E–G represent trinomial mixtures, and H-I represent multinomial mixtures with $k = 4$ categories. $\mathrm{V}(m_i)$ is shorthand for the variance of m_i. . . . 70
2.4 Convergence characteristics of AFSA and EM over 1000 samples. Here $p_2 = 0.5$ is fixed. The reported quantity is the average number of algorithm iterations per sample. Note that the tolerance for convergence was set to 10^{-8} and a maximum of 1000 iterations was allowed for each algorithm per sample. . . . 71

3.1 Results for Normal mixture. The diagonals $\widetilde{\mathcal{I}}_{ii}$ are given with corresponding $\mathcal{I}_{ii}$ in parentheses. The last column shows Frobenius norm of the matrix difference $\widetilde{\mathcal{I}} - \mathcal{I}$. . . . 106
3.2 Results for Weibull mixture. The diagonals $\widetilde{\mathcal{I}}_{ii}$ are given with corresponding $\mathcal{I}_{ii}$ in parentheses. The last column shows Frobenius norm of the matrix difference $\widetilde{\mathcal{I}} - \mathcal{I}$. All entries of $\mathcal{I}$ were approximated by Monte Carlo simulation using $L = 100{,}000$. . . . 115

4.1 Number of steps required for finding vertices of A in dimension J. The column Polyhedron denotes the algorithm discussed in Remark 4.5 for general polyhedra which requires $\binom{2J+2}{J}$ steps. The column FindVertices denotes Algorithm 1 which requires $J \cdot 2^{J-1}$ steps. . . . 141
4.2 Distance $D(f, g)$ between the density f of $\boldsymbol{c}^T \boldsymbol{\lambda}$ and density g of beta approximation. . . . 157
4.3 Distance $D(f, g)$ between exact Mixture Link density f and density g using beta approximation with $m = 5$ trials. . . . 161

4.4 Distance $D(f, g)$ between exact Mixture Link density f and density g using beta approximation with $m = 10$ trials. 162
4.5 Distance $D(f, g)$ between exact Mixture Link density f and density g using beta approximation with $m = 20$ trials. 163
4.6 Maximum likelihood estimates for candidate models, with standard errors in parentheses. 172
4.7 Standard errors of MixLinkJ2 MLE computed using 500 parametric bootstrap samples. 172
4.8 Model comparison statistics. 173
4.9 Details for GOF test computed from the six models. 175
4.9 (Continued). 176
4.10 Estimates for mutagenic datasets with standard errors in parentheses. . . . 180
4.11 GOF for Mutagenic Dataset I. 181
4.12 GOF for Mutagenic Dataset II. 182
4.13 GOF for Mutagenic Dataset III. 183

List of Figures

1.1 Comparison between two-component bivariate normal mixture vs. two-component trinomial mixture. 3

2.1 The next AFSA $\tilde{p}_{\ell j}^{(g+1)}$ iteration is a linear combination of $\hat{p}_{\ell j}^{(g+1)}$ and $p_{\ell j}^{(g)}$, which depends on the ratio $\hat{\pi}_{\ell}^{(g+1)}/\pi_{\ell}^{(g)}$. 55
2.2 Distance between exact and approximate FIM and its inverse, as m is varied. 60
2.3 Empirical CDF of approximated and exact Wald test statistics. 62
2.4 Empirical CDF of approximated and exact Score test statistics. 63
2.5 Expected counts, rounded to the nearest integer, for $n = 500$ observations sampled independently from each of the two subpopulations. Counts rounded to zero are not shown. 66
2.6 Counts of observations from the generated sample at each of the possible trinomial outcomes for $m = 20$. 67
2.7 Convergence of several competing algorithms for a small test problem . . 67
2.8 Boxplots for Scenarios D and E of the Monte Carlo study presented in Table 2.3. At this scale, the boxes appear as thin horizontal lines. 71

3.1 Densities for the bivariate normal finite mixture under the three scenarios. 105
3.2 Frobenius norm of $\widetilde{\mathcal{I}}_m - \mathcal{I}_m$, as m varies, for the three normal scenarios. . 107
3.3 Densities for the Weibull finite mixture under the two scenarios. 116

4.1 A sample drawn from A: (a) $n = 100$ with $J = 2$, $\boldsymbol{\pi} = (\frac{3}{4}, \frac{1}{4})$, $p = \frac{2}{3}$, and (b) $n = 300$ with $J = 3$, $\boldsymbol{\pi} = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$, $p = \frac{2}{3}$. 128
4.2 The set $\{\boldsymbol{\mu} \in [0, 1]^3 : \mu_1\pi_1 + \mu_2\pi_2 + \mu_3\pi_3 = p\}$ visualized with two different settings of $\boldsymbol{\pi}$. In each case, $p \in \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$ is shown (from front to back). 129
4.3 Dirichlet$_3(\boldsymbol{\lambda} \mid \kappa\mathbf{1})$ density for several settings of κ. Only λ_1 and λ_2 are plotted, as $\lambda_3 = 1 - \lambda_1 - \lambda_2$ is redundant. 130
4.4 Example set A with $\boldsymbol{\pi} = (\frac{11}{20}, \frac{9}{20})$ and $p = \frac{1}{2}$. The vertices are located at $\boldsymbol{v}_1 = (\frac{10}{11}, 0)$ and $\boldsymbol{v}_2 = (\frac{1}{11}, 1)$. 137
4.5 BB densities. 148
4.6 RCB densities. 148
4.7 Mixture Link densities. 149

4.8 Compare Mixture Link densities for $J = 2$ and $J = 3$ when $\kappa = 0.5$. 149
4.9 Compare Mixture Link densities for $J = 2$ and $J = 3$ when $\kappa = 1$. 150
4.10 Compare Mixture Link densities for $J = 2$ and $J = 3$ when $\kappa = 2$. 150
4.11 Comparison between density of $\boldsymbol{c}^T \boldsymbol{\lambda}$ and moment-matched beta taking $\boldsymbol{c} = (0, 0.1, 1)$. 158
4.12 Comparison between density of $\boldsymbol{c}^T \boldsymbol{\lambda}$ and moment-matched beta taking $\boldsymbol{c} = (0, 0.1, 1)$. 159
4.13 Comparison between density of $\boldsymbol{c}^T \boldsymbol{\lambda}$ and moment-matched beta taking $\boldsymbol{c} = (0, 0.05, 0.5, 0.95, 1)$. 160
4.14 Comparison of exact Mixture Link density f and density g using beta approximation with $m = 20$ trials and $\boldsymbol{\pi} = (\frac{1}{20}, \frac{2}{20}, \frac{3}{20}, \frac{4}{20}, \frac{5}{20})$. 164
4.15 ECDF of estimates $\hat{\kappa}^{(b)}$ from 500 parametric bootstrap samples. 173
4.16 GOF plots for observed vs. expected counts. The grey bars represent the observed counts for a given interval, and the black dots are the expected counts under the MLE. Note that the choice of intervals varies between models. 177
4.16 (Continued). 178
4.17 Empirical CDF of MixLinkJ2 GOF statistic computed from parametric bootstrap. 178
4.18 GOF for Mutagenic Dataset I. 181
4.19 GOF for Mutagenic Dataset II. 182
4.20 GOF for Mutagenic Dataset III. 183

Chapter 1

Introduction

Finite mixture models are widely used in practice and have long been studied in the statistical literature because of the analytical challenges they present. Finite mixtures are often used to model a population consisting of multiple subpopulations, whose characteristics vary for the question under study, but where the subpopulation membership of each observation is not observed. This is the setup of the classic clustering problem which is one natural application of finite mixtures. Mixtures can also be useful in handling extra variability when important covariates have not been recorded. In both cases, the mixture distribution helps to account for unknown sources of heterogeneity in the data. Ignoring heterogeneity or failing to capture other important nuances in the data generating process can lead to the situation of overdispersion (or extra variation), where more variation is present in the data than a given model is designed to handle. Beside their practical use in applications, the study of mixtures also leads to a variety of interesting theoretical issues in identifiability, inference on the mixing distribution, and model selection.

The book by Titterington et al. (1985) provides a comprehensive overview of finite mixtures from a classical perspective. McLachlan and Peel (2000) present a more modern overview of general issues in finite mixtures and demonstrate more recent Bayesian and Expectation-Maximization computational methods. Frühwirth-Schnatter (2006) provides another modern overview emphasizing Bayesian inference, mixtures of regressions, and models where the latent mixing process evolves over time. The monograph by Lindsay (1995) reviews theory on the geometry of mixtures and on the nonparametric maximum likelihood estimator. Morel and Neerchal (2012) present an overview of models for overdispersion, especially for count, binomial, and multinomial data.

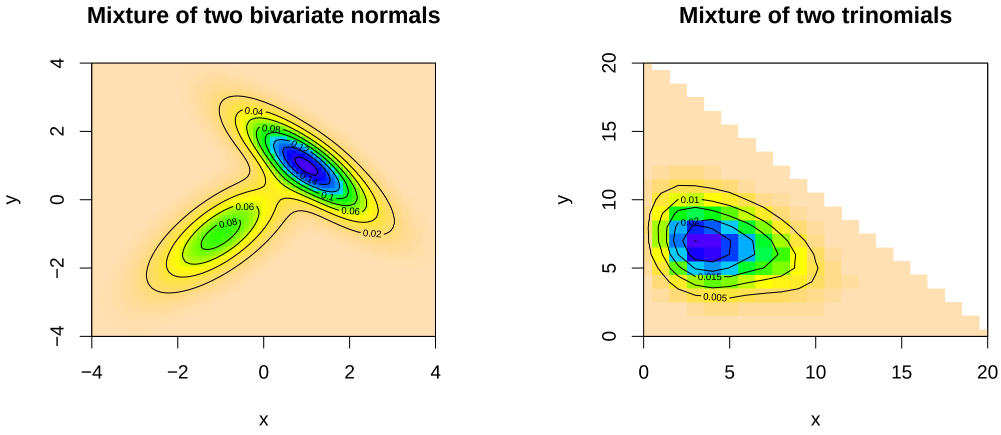
Much of the literature on finite mixtures focuses on continuous distributions such as normal, Student's t, and their multivariate extensions. Such models have a natural intuition, providing elliptical or near-elliptical shapes on, say, a k-dimensional Euclidean space. Populations can be viewed geometrically in this space, in terms of their modes, mutual distances, etc. This thesis focuses instead on finite mixtures of binomial and multinomial distributions, which are perhaps less intuitive, but often are more appropriate for modeling data observed under these circumstances. The finite mixture of multinomials has been applied to many areas including: clustering of internet traffic (Jorgensen, 2004), text/topic analysis (Hofmann, 1999), item response theory for analysis of educational or psychological tests (Bolt et al., 2001), and genetics (Toussile and Gassiat, 2009). Bayesian analysis of the finite mixture of multinomials is studied by Rufo et al. (2007). To emphasize the difference in graphical intuition, Figure 1.1a plots the density of a mixture of bivariate normals

$$\pi \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-\pi)\mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \quad \text{where} \quad \pi = 0.40,$$
$$\boldsymbol{\mu}_1 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.70 \\ 0.70 & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & -0.80 \\ -0.80 & 1 \end{pmatrix}.$$

Meanwhile, Figure 1.1b illustrates a mixture of two trinomials

$$\pi \mathrm{Mult}_3(m, \boldsymbol{p}_1) + (1-\pi)\mathrm{Mult}_3(m, \boldsymbol{p}_2), \quad \text{where}$$
$$\boldsymbol{p}_1 = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}, \quad \boldsymbol{p}_2 = \begin{pmatrix} 1/6 \\ 2/6 \\ 3/6 \end{pmatrix}, \quad \pi = 0.40, \quad m = 20.$$

Compared to Figure 1.1a, it is less evident in Figure 1.1b that there are two distinct sub-



(a) (b)

Figure 1.1: Comparison between two-component bivariate normal mixture vs. two-component trinomial mixture.

populations in the model. Of course, the two bivariate normal populations could be closer together, which would make them more difficult to distinguish.

This thesis investigates three problems which have arisen in the study of finite mixtures and overdispersion in binomial/multinomial data analysis.

Overview of Chapter 2. The first problem considers an approximate information matrix which was originally proposed for binomial and multinomial finite mixtures, and has been used in scoring iterations as a substitute for the Hessian or exact information matrix. This technique was developed for the random-clumped multinomial (RCM) model, which was proposed by Morel and Neerchal for overdispersed multinomial data. We show that the matrix approximation is equivalent to a complete data information matrix, which helps to explain previously noted connections between approximate scoring and the Expectation Maximization (EM) algorithm. This equivalence allows the technique of approximate scoring to be generalized beyond multinomial finite mixture analysis to any missing data problem where a complete data information matrix and the usual score vector can be formulated. This includes problems involving finite mixture and continuous mixtures.

The approximate scoring algorithm had previously been justified by showing that exact and approximate information matrices converge together as the number of multinomial trials $m \to \infty$; this justification does not immediately extend outside of the multinomial setting. Several simulation studies in this chapter show that approximate scoring and EM perform very similarly in the neighborhood of a solution, but that the approximate information may not act as a reliable substitute for the exact information to serve more general inference purposes such as obtaining standard errors and test statistics. It is also demonstrated that a hybrid algorithm, using approximate scoring to start an initial path toward a solution and then switching to Fisher scoring, combines the robustness of the former with the rapid convergence of the latter.

Overview of Chapter 3. The second problem extends the convergence of the approximate information discussed in Chapter 2 to the more general setting of exponential family finite mixtures. To do this, we use a "clustered sampling" scheme so that m observations are sampled from the same (unknown) subpopulation. The clustered sampling acts as an analogue to the m trials in the multinomial case. It is proved that the exact and approximate information matrices converge together as $m \to \infty$ and rates for the convergence are obtained. The proof requires a different approach than the multinomial case, which depends on properties of the multinomial distribution. It is also noted that the convergence does not occur when, instead, an independent and identically distributed sample of size m is taken from the finite mixture. Therefore, use of the information matrix approximation is justified in the very practical setting of exponential family finite mixtures, but only when the clustered sampling scheme applies. The results in this chapter help to shed light on the binomial/multinomial case, emphasizing that trials are samples taken within a common subpopulation.

Overview of Chapter 4. The third problem considers the setting of a binomial finite mixture, and investigates the objective of linking a regression to the mixture probability of success. This can be considered an extension of logistic regression model from the Generalized Linear Model framework, where a finite mixture is now assumed to support extra variation in the population. This approach can also be compared to the Generalized Estimating Equations (GEE) approach for ungrouped data, where the analyst is free to select a covariance structure for observations within-group to support extra variation. However, our approach supports grouped observations and is completely based on a likelihood. The finite mixture is not an exponential family, and the mixed probability of success is a composite parameter which does not appear directly in the likelihood. Therefore, a challenge is to formulate the model in such a way that the link is enforced yet the number of unknown parameters is kept manageable. In this work we formulate a random effects model called Mixture Link, where the random effects are drawn from a set representing enforcement of the link. Initial results show that the model is effective at capturing extra variability while supporting the regression; however, further study is required to more thoroughly address issues such as identifiability, computation of the likelihood, and estimation of parameters and standard errors. The approach itself appears to generalize beyond binomial data analysis, and may be considered for other problems where the analyst wishes to link a regression to a composite parameter which does not explicitly appear in the likelihood.

The remainder of Chapter 1 gives background on mixtures and various issues that arise in their use.

1.1 Mixture Formulation

Before focusing on finite mixtures, it is enlightening to formulate the general mixture model in a similar spirit to (Teicher, 1960) or (Teicher, 1961). Let $\mathcal{H}$ denote a class of densities; often $\mathcal{H} = \{h(\cdot \mid \boldsymbol{\phi}) : \boldsymbol{\phi} \in \Phi\}$ will be a family of densities indexed by a parameter $\boldsymbol{\phi} \in \Phi \subseteq \mathbb{R}^d$. Let $\mathcal{G}$ denote a family of mixing distributions on $\mathcal{H}$. $\mathcal{G}$ is often also indexed by a parameter, say $\boldsymbol{\theta} \in \Theta$. The class of mixture distributions, which will be denoted by $\mathcal{F}$, is then given by mixing all densities in $\mathcal{H}$ by a particular distribution $G \in \mathcal{G}$. A density in $\mathcal{F}$ may be written generally as $f(\boldsymbol{x} \mid G) = \int_{\mathcal{H}} h(\boldsymbol{x})\, dG(h)$. It is convenient to consider all unknown parameters of $f(\boldsymbol{x} \mid G)$ as belonging to the mixing distribution G. A few examples will help to illustrate this framework.

Example 1.1 (Student's t as a mixture of normals)**.** Let $\mathcal{H} = \{h(\cdot \mid \boldsymbol{\phi}) : \boldsymbol{\phi} \in \Phi\}$ be the family $N(0, \sigma^2)$, so that $\phi = \sigma^2$ and $\Phi = (0, \infty)$. Take $\mathcal{G}$ to be the Inverse Gamma family $\{\text{IG}(\alpha = v/2, \beta = v/2) : v > 0\}$, indexed by the parameter $\theta = v$ and

$$dG_\theta(\boldsymbol{\phi}) = \frac{\beta^\alpha}{\Gamma(\alpha)} \phi^{-\alpha-1} e^{-\beta/\phi} d\phi = \frac{(v/2)^{v/2}}{\Gamma(v/2)} \phi^{-v/2-1} e^{-v/2\phi} d\phi$$

Then we have

$$\begin{aligned} f(x \mid \theta) &= \int h(x \mid \phi) dG_\theta(\phi) \\ &= \int_0^\infty \frac{1}{\sqrt{\phi}\sqrt{2\pi}} e^{-x^2/2\phi} \cdot \frac{(v/2)^{v/2}}{\Gamma(v/2)} \phi^{-v/2-1} e^{-v/2\phi} d\phi \\ &= \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{x^2}{2}\right)^{-\frac{v+1}{2}}. \end{aligned}$$

This is the Student's t distribution with v degrees of freedom. Therefore, the t distribution can be considered a continuous mixture of normals. Liu and Rubin (1995) make use of the multivariate t distribution to handle extra variation beyond the standard multivariate normal $\text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ model. The mixture representation is used to provide a complete data model for use with Expectation-Maximization (see Section 1.5 for a brief overview of Expectation-Maximization). Finite mixtures of multivariate t's have been considered to handle even more variation (McLachlan and Peel, 2000, Chapter 7).

Example 1.2 (Beta-binomial)**.** Let $\mathcal{H} = \{\text{Bin}(x \mid n, \phi), \phi \in (0, 1)\}$, and take $\mathcal{G} = \{G_{\boldsymbol{\theta}}(\cdot) : \boldsymbol{\theta} \in \Theta\}$ to be the family $\text{Beta}(\phi \mid \alpha, \beta)$, so that $\boldsymbol{\theta} = (\alpha, \beta)$ and $\Theta = (0, \infty) \times$

$(0, \infty)$. Then we have

$$\begin{aligned} f(x \mid \theta) &= \int h(x \mid \phi) dG_{\boldsymbol{\theta}}(\phi) \\ &= \int_0^1 \binom{n}{x} \phi^x (1-\phi)^{n-x} \cdot \frac{\phi^{\alpha-1}(1-\phi)^{\beta-1}}{B(\alpha, \beta)} d\phi \\ &= \binom{n}{x} \frac{B(\alpha+x, \beta+n-x)}{B(\alpha, \beta)} \end{aligned}$$

This is the beta-binomial distribution parameterized by $\boldsymbol{\theta} = (\alpha, \beta)$. It has been used to model overdispersed binomial data when the standard binomial distribution cannot sufficiently capture the variation. See for example (Mosimann, 1962) and (Morel and Neerchal, 2012, Chapter 7).

Example 1.3 (Finite mixture)**.** Take $\mathcal{H} = \{h(\cdot \mid \boldsymbol{\phi}) : \boldsymbol{\phi} \in \Phi\}$ to be any parametric family, and let $\mathcal{G}$ be the family with densities

$$dG_{\boldsymbol{\theta}}(\boldsymbol{\phi}) = \sum_{j=1}^{s} \pi_j I(\boldsymbol{\phi} = \boldsymbol{\phi}_j) = \begin{cases} \boldsymbol{\phi}_1 & \text{w.p. } \pi_1 \\ & \vdots \\ \boldsymbol{\phi}_s & \text{w.p. } \pi_s \end{cases}$$

so that $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_s$ are points chosen from Φ, and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_s)$ forms a discrete probability distribution. The resulting mixture distribution is

$$f(\boldsymbol{x} \mid \boldsymbol{\theta}) = \int h(\boldsymbol{x} \mid \boldsymbol{\phi}) dG_{\boldsymbol{\theta}}(\boldsymbol{\phi}) = \sum_{\ell=1}^{s} \pi_\ell h(\boldsymbol{x} \mid \boldsymbol{\phi}_\ell) \tag{1.1}$$

This is the standard finite mixture model, with $\boldsymbol{\theta} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_s, \boldsymbol{\pi})$, which will be the focus for much of the thesis. In statistical applications, both the support points and masses of G will usually be unknown and hence need to be estimated. This thesis will often consider $\mathcal{H}$ to be the family of multinomial distributions $\{\text{Mult}_k(m, \boldsymbol{\phi}) : \boldsymbol{\phi} \in \Phi\}$, where Φ is the probability simplex in $\mathbb{R}^k$.

The finite mixture has an appealing intuition. Consider the two-stage sampling process

$$\begin{aligned} Z_i &\sim \text{Discrete}(1, \dots, s; \boldsymbol{\pi}), \\ \boldsymbol{X}_i \mid Z_i = j &\sim f(\boldsymbol{x} \mid \boldsymbol{\phi}_j), \end{aligned}$$

where $(\boldsymbol{X}_i, Z_i)$ are independent for $i = 1, \dots, n$, and $\text{Discrete}(a_1, \dots, a_s; \boldsymbol{\pi})$ denotes the distribution having support $a_1, \dots, a_s$ with respective probabilities $\pi_1, \dots, \pi_s$. If both $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ and $Z_1, \dots, Z_n$ are observed, we have classified data among s populations. A problem in this framework might be to predict the class Z for a new observation $\boldsymbol{X}$, given past observations. If only $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are observed, the data naturally belong to s unobserved clusters and are distributed according to (1.1). Estimating unknown parameters and determining cluster membership therefore can be seen as a more difficult problem than classification. Finite mixtures provide a natural likelihood-based way to approach this problem, and this use is called “model-based clustering”.

In practice, the number of support points s of G is not known and must be determined from the data or prior knowledge. Determination of s through the data has been considered a difficult theoretical issue. Approaches include information-theoretic model selection (e.g. AIC) and formal hypothesis testing. McLachlan and Peel (2000, Chapter 6) provide an overview.

Example 1.4 (Finite mixture with proportional odds model as mixing distribution)**.** Again take $\mathcal{H}$ to be any parametric family, and let $\mathcal{G}$ be the family with densities

$$dG_{\boldsymbol{\theta}}(\boldsymbol{\phi}) = \begin{cases} \boldsymbol{\phi}_1 & \text{w.p. } \mathrm{P}(\alpha_0 < Z - \boldsymbol{w}^T\boldsymbol{\beta} \le \alpha_1) \\ & \vdots \\ \boldsymbol{\phi}_s & \text{w.p. } \mathrm{P}(\alpha_{s-1} < Z - \boldsymbol{w}^T\boldsymbol{\beta} \le \alpha_s) \end{cases}$$

where $\alpha_0, \ldots, \alpha_s$ are fixed numbers such that $\alpha_0 < \ldots < \alpha_s$, $\alpha_0 = -\infty$, and $\alpha_s = \infty$. Let us assume that $Z - \boldsymbol{w}^T\boldsymbol{\beta} \sim \text{Logistic}(0, 1)$, and that $\boldsymbol{w}$ is a p-dimensional covariate with coefficients $\boldsymbol{\beta}$. This yields the finite mixture

$$f(\boldsymbol{x} \mid G_{\boldsymbol{\theta}}) = \int h(\boldsymbol{x} \mid \boldsymbol{\phi}) dG_{\boldsymbol{\theta}}(\boldsymbol{\phi}) = \sum_{\ell=1}^{s} \mathrm{P}(\alpha_{\ell-1} < Z - \boldsymbol{w}^T\boldsymbol{\beta} \leq \alpha_\ell) h(\boldsymbol{x} \mid \boldsymbol{\phi}_\ell),$$

where $\boldsymbol{\theta} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_s, \boldsymbol{\alpha}, \boldsymbol{\beta})$. Now the mixing probabilities themselves are parametric functions that depend on a regression. Such models have been referred to as latent regression models (Dayton and Macready, 1988). This particular model is being proposed by Huang (2012) to address covariate measurement error in the estimation of average causal effect.

Example 1.5 (Zero-Inflated Binomial)**.** Here is an example where the class $\mathcal{H}$ is not a parametric family. Let

$$\mathcal{H} = \{I(x = 0)\} \cup \{\text{Bin}(x \mid m, \phi) : \phi \in (0, 1)\},$$

and for elements $h \in \mathcal{H}$ suppose

$$dG_{\boldsymbol{\theta}}(h) = \begin{cases} I(x = 0) & \text{w.p. } \pi \\ \text{Bin}(x \mid m, p) & \text{w.p. } 1 - \pi \\ 0 & \text{o.w.} \end{cases}$$

so that $\boldsymbol{\theta} = (p, \pi)$. Then we obtain

$$f(x \mid G_{\boldsymbol{\theta}}) = \int_{\mathcal{H}} h(x) dG_{\boldsymbol{\theta}}(h) = \pi I(x = 0) + (1 - \pi)\text{Bin}(x \mid m, p)$$

which is the zero-inflated binomial (Hall, 2000).

One appealing aspect of mixture distributions is that it is usually easy to draw a sample from them. We have already mentioned the special case of finite mixtures. For a general mixture this can be done in two phases: first sample $h \sim G$ for a given $G \in \mathcal{G}$, then draw $\boldsymbol{X}$ from the density h. The resulting $\boldsymbol{X}$ will have the distribution $f(\cdot \mid G)$.

1.2 Moments of Mixtures

Let ν denote the dominating measure for densities in $\mathcal{H}$. Under a mixture distribution, the expected value and variance are given by

$$\mathrm{E}(\boldsymbol{X}) = \int\int \boldsymbol{x} h(\boldsymbol{x} \mid \boldsymbol{\phi}) dG(\boldsymbol{\phi}) d\nu(\boldsymbol{x}) = \int \mathrm{E}(\boldsymbol{X} \mid \boldsymbol{\phi}) dG(\boldsymbol{\phi}) = \mathrm{E}[\mathrm{E}(\boldsymbol{X} \mid \boldsymbol{\phi})]$$

and

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{X}) &= \int\int [\boldsymbol{x} - \mathrm{E}(\boldsymbol{X})] [\boldsymbol{x} - \mathrm{E}(\boldsymbol{X})]^T h(\boldsymbol{x} \mid \boldsymbol{\phi}) dG(\boldsymbol{\phi}) d\nu(\boldsymbol{x}) \\
&= \int \mathrm{Var}_{\boldsymbol{\phi}}(\boldsymbol{X}) dG(\boldsymbol{\phi}) + \int [\mathrm{E}(\boldsymbol{X} \mid \boldsymbol{\phi}) - \mathrm{E}(\boldsymbol{X})] [\mathrm{E}(\boldsymbol{X} \mid \boldsymbol{\phi}) - \mathrm{E}(\boldsymbol{X})]^T dG(\boldsymbol{\phi}) \\
&= \mathrm{E}[\mathrm{Var}(\boldsymbol{X} \mid \boldsymbol{\phi})] + \mathrm{Var}[\mathrm{E}(\boldsymbol{X} \mid \boldsymbol{\phi})],
\end{aligned}$$

provided that the integrals exist and the order of integration may be changed. Notice that $\mathrm{Var}[\mathrm{E}(\boldsymbol{X} \mid \boldsymbol{\phi})]$ is positive semidefinite, so we have the result

$$\mathrm{Var}(\boldsymbol{X}) - \mathrm{E}[\mathrm{Var}(\boldsymbol{X} \mid \boldsymbol{\phi})] \quad \text{is positive semidefinite.}$$

In other words the variance under the mixture is larger than the expected variance given a particular $\boldsymbol{\phi}$. This gives a simple justification for using mixtures to address overdispersion. For our main scenario of interest, the finite mixture, the mean and variance expressions become

$$\begin{aligned}\mathrm{E}(\boldsymbol{X}) &= \sum_{\ell=1}^{s} \pi_\ell \, \mathrm{E}(\boldsymbol{X} \mid \boldsymbol{\phi}_\ell) \quad \text{and} \\ \mathrm{Var}(\boldsymbol{X}) &= \sum_{\ell=1}^{s} \pi_\ell \, \mathrm{Var}(\boldsymbol{X} \mid \boldsymbol{\phi}_\ell) + \sum_{\ell=1}^{s} \pi_\ell \left[\mathrm{E}(\boldsymbol{X} \mid \boldsymbol{\phi}_\ell) - \mathrm{E}(\boldsymbol{X})\right] \left[\mathrm{E}(\boldsymbol{X} \mid \boldsymbol{\phi}_\ell) - \mathrm{E}(\boldsymbol{X})\right]^T .\end{aligned}$$

1.3 Identifiability

One major issue in the use of mixtures is identifiability. Consider the mapping $L : \mathcal{G} \to \mathcal{F}$ that takes a mixing distribution G, integrates the family $\mathcal{H}$ of component distributions with respect to G, and results in a specific mixture distribution.

Definition 1.6 (Identifiability)**.** A family of mixtures $\mathcal{F}$ is said to be identifiable if L is a one-to-one function. That is

$$f(\cdot \mid G) = f(\cdot \mid G^*) \implies G = G^*,$$

which can also be written as

$$f(\boldsymbol{x} \mid G) \overset{a.s.}{=} f(\boldsymbol{x} \mid G^*) \implies G = G^*.$$

If $\mathcal{G} = \{G_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ so that $\mathcal{F} = \{f(\boldsymbol{x} \mid G_{\boldsymbol{\theta}}) : \boldsymbol{\theta} \in \Theta\}$ is a parametric family, identifiability means that

$$f(\boldsymbol{x} \mid G_{\boldsymbol{\theta}}) \overset{a.s.}{=} f(\boldsymbol{x} \mid G_{\boldsymbol{\theta}^*}) \implies \boldsymbol{\theta} = \boldsymbol{\theta}^*.$$

Identifiability is often considered to be a minimum requirement for a statistical model to be useful. If it fails, two different models in $\mathcal{F}$ will yield the exact same likelihood no matter which data is observed, therefore one cannot hope to obtain evidence of one versus the other through the data. Proving or disproving identifiability in mixture models can be challenging. Rao (1992, Chapter 8) provides an overview of the theory of identifiability. Less strict criteria such as local identifiability can also be considered; see for example (Rothenberg, 1971) and (Paulino and de Bragança Pereira, 1994).

1.4 Fisher Information

The Fisher information matrix (FIM)

$$\begin{aligned}\mathcal{I}(\boldsymbol{\theta}) &= \operatorname{Var}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{x} \mid \boldsymbol{\theta})\right) \\ &= \operatorname{E}\left[\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{x} \mid \boldsymbol{\theta})\right\}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{x} \mid \boldsymbol{\theta})\right\}^T\right]\end{aligned}$$

plays an important role in statistics when it exists. We say that $\mathcal{I}(\boldsymbol{\theta})$ is the expected information about $\boldsymbol{\theta}$ contained in the random variable $\boldsymbol{X}$. When $\boldsymbol{\theta}$ is a scalar, the scalar $\mathcal{I}(\boldsymbol{\theta})$ is larger when there is increased ability for $\boldsymbol{X}$ to provide inference about $\boldsymbol{\theta}$. More generally, when $\boldsymbol{\theta} \in \mathbb{R}^q$, then $\mathcal{I}(\boldsymbol{\theta})$ is a $q \times q$ matrix. Under certain regularity conditions (Shao, 2008, Section 3.1) the matrix can be rewritten

$$\mathcal{I}(\boldsymbol{\theta}) = \operatorname{E}\left[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log f(\boldsymbol{x} \mid \boldsymbol{\theta})\right],$$

which emphasizes that the "information" is quantifying the degree of curvature in the log-likelihood at $\boldsymbol{\theta}$. A likelihood which is usually more flat (usually means with respect to the distribution of $\boldsymbol{X}$) around $\boldsymbol{\theta}$ will be reflected by a lower Fisher information. Fisher information is not appropriate when certain basic assumptions, such as differentiability of $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ and $\boldsymbol{\theta}$ being an interior point of Θ, are not met. The matrix $\mathcal{I}^{-1}(\boldsymbol{\theta})$ represents the asymptotically optimal variance (under appropriate regularity conditions) of an estimator $\hat{\boldsymbol{\theta}}$ under an independent and identically distributed sample $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ from $f(\boldsymbol{x} \mid \boldsymbol{\theta})$, as the sample size n tends to infinity. In this sense, a larger amount of information corresponds to more precise inference being possible. Furthermore, for an unbiased estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, the Cramer-Rao Lower Bound states that

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}) \geq \mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})$$

where "$\boldsymbol{A} \geq \boldsymbol{B}$" is taken to mean that $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite when $\boldsymbol{A}$ and $\boldsymbol{B}$ are $q \times q$ matrices. In this sense, $\mathcal{I}^{-1}(\boldsymbol{\theta})$ represents the most precise inference possible for any unbiased estimator, which may or may not be attainable by some estimator in that class. Under the reparameterization $\boldsymbol{\psi}(\boldsymbol{\theta})$, the information matrix with respect to $\boldsymbol{\psi}$ may be obtained using the Jacobian of the transformation $\boldsymbol{\psi} \mapsto \boldsymbol{\theta}$ as

$$\begin{aligned}
\mathcal{I}(\boldsymbol{\psi}) &= \mathrm{Var}\left(\frac{\partial}{\partial \boldsymbol{\psi}} \log f(\boldsymbol{x} \mid \boldsymbol{\psi})\right) \\
&= \mathrm{Var}\left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x} \mid \boldsymbol{\theta})\right) \\
&= \left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}}\right) \mathrm{Var}\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x} \mid \boldsymbol{\theta})\right) \left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}}\right)^T \\
&= \left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}}\right) \mathcal{I}(\boldsymbol{\theta}) \left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}}\right)^T.
\end{aligned}$$

Quantities such as

$$I(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log f(\boldsymbol{x} \mid \boldsymbol{\theta}) \quad \text{or} \quad J(\boldsymbol{\theta}) = \left\{\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x} \mid \boldsymbol{\theta})\right\} \left\{\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x} \mid \boldsymbol{\theta})\right\}^T$$

describing the "observed information" are often used to estimate the variance, and do not require computation of an expectation which may be analytically intractable. For example, Boldea and Magnus (2009) give expressions for $I(\boldsymbol{\theta})$ and $J(\boldsymbol{\theta})$ in the setting of the multivariate normal finite mixture. Properties of observed information such as near-singularity depend on the sample, which is one reason that $\mathcal{I}(\boldsymbol{\theta})$ would be preferred in variance estimation. In this thesis, we consider $\mathcal{I}(\boldsymbol{\theta})$ to be a quantity of interest in its own right, and generally do not make use of $I(\boldsymbol{\theta})$ and $J(\boldsymbol{\theta})$.

1.5 Estimation

Before electronic computers were widely available, estimation in finite mixtures was often carried out by the method of moments. The idea is to express the unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$ as functions of r moments $\mathrm{E}[\psi_1(\boldsymbol{X})], \ldots, \mathrm{E}[\psi_r(\boldsymbol{X})]$ so that

$$\begin{aligned} \theta_1 &= g_1\left(\mathrm{E}[\psi_1(\boldsymbol{X})], \ldots, \mathrm{E}[\psi_r(\boldsymbol{X})]\right) \\ &\vdots \\ \theta_q &= g_q\left(\mathrm{E}[\psi_1(\boldsymbol{X})], \ldots, \mathrm{E}[\psi_r(\boldsymbol{X})]\right). \end{aligned}$$

The moments $\mathrm{E}[\psi_j(\boldsymbol{X})]$ can then be substituted by sample moments, yielding an estimate for $\boldsymbol{\theta}$. Approaches such as the method of maximum likelihood (MLE) and Bayesian inference are often too computationally tedious to implement by hand for mixtures because they require iterative procedures to implement them. These approaches have become standard as computers have became available. MLEs are often computed using Newton-Raphson iterations

$$\begin{aligned} &\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)} - \boldsymbol{H}^{-1}(\boldsymbol{\theta}^{(g)})S(\boldsymbol{\theta}^{(g)}), \quad \text{where} \\ &S(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x} \mid \boldsymbol{\theta}), \quad \text{and} \quad \boldsymbol{H}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log f(\boldsymbol{x} \mid \boldsymbol{\theta}). \end{aligned}$$

Starting these iterations at an initial guess $\boldsymbol{\theta}^{(0)}$ and iterating until convergence yields an MLE $\hat{\boldsymbol{\theta}}$, or more precisely a critical point of the log-likelihood which may also be a maximizer. A useful byproduct is the matrix $-\boldsymbol{H}^{-1}(\hat{\boldsymbol{\theta}})$, which is an estimator of the asymptotic variance of the MLE. The square roots of the diagonal elements of this matrix are often taken as standard errors. A popular alternative is Fisher scoring, whose iterations are of the form

$$\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)} + \mathcal{I}^{-1}(\boldsymbol{\theta}^{(g)})S(\boldsymbol{\theta}^{(g)}).$$

Fisher scoring is often preferred over Newton-Raphson because the Hessian of the log-likelihood, and its properties such as near-singularity, may vary greatly based on the sample. On the other hand, the matrix $\mathcal{I}(\boldsymbol{\theta})$ only depends on the sample when $\boldsymbol{\theta}$ is estimated by $\hat{\boldsymbol{\theta}}$. Fisher scoring iterations yield the byproduct $\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})$, which is the asymptotic variance of the MLE. Dempster et al. (1977) helped to formalize the Expectation-Maximization (EM) algorithm, which has since been applied to an immense number of problems. The idea is to supplement the observed data $\boldsymbol{y}$ with missing data $\boldsymbol{z}$. The complete data $(\boldsymbol{y}, \boldsymbol{z})$ often has a likelihood with a much simpler form than the marginal likelihood of the observed data $\boldsymbol{y}$. EM considers the decomposition

$$\begin{aligned}
&\log f(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\theta}) = \log f(\boldsymbol{y} \mid \boldsymbol{\theta}) + \log f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\theta}) \\
&\implies \log f(\boldsymbol{y}, \boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}'}\left[\log f(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}) \mid \boldsymbol{y}\right] - \mathrm{E}_{\boldsymbol{\theta}'}\left[\log f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\theta}) \mid \boldsymbol{y}\right] \\
&\iff \log f(\boldsymbol{y}, \boldsymbol{\theta}) = Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}') - H(\boldsymbol{\theta} \mid \boldsymbol{\theta}'),
\end{aligned}$$

where the notation $\mathrm{E}_{\boldsymbol{\theta}'}(\cdot)$ means that the expectation is evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}'$, which is some given value of the parameters. The conditional expectation given $\boldsymbol{y}$ takes care of the unobservable $\boldsymbol{z}$ variables, transforming them into functions of the given $\boldsymbol{\theta}'$ we can be explicitly computed. EM focuses on maximizing only the $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')$ function, which often inherits a simple form from the complete data log-likelihood. The EM algorithm can then be written as follows, starting from an initial guess $\boldsymbol{\theta}^{(0)}$

E-step: Compute $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(0)})$

M-step: Maximize $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(0)})$ with respect to $\boldsymbol{\theta}$ to obtain $\boldsymbol{\theta}^{(1)}$

These steps are then repeated until some convergence criteria is met. It is straightforward to show that, as long as the $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(g)})$ function is increased, the $H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(g)})$ function is not decreased. Therefore, EM is also maximizing $\log f(\boldsymbol{y} \mid \boldsymbol{\theta})$, and hence computing an MLE. For the standard finite mixture from Example 1.3, the subpopulation of each observation can be considered missing data, leading to a complete data model

$$\begin{aligned} \boldsymbol{Y}_i \mid Z_i = \ell &\sim f(\boldsymbol{y} \mid \boldsymbol{\phi}_\ell), \\ Z_i &\sim \text{Discrete}(1, \ldots, s; \boldsymbol{\pi}), \end{aligned}$$

where $(\boldsymbol{X}_i, Z_i)$ are independent for $i = 1, \ldots, n$. The complete data likelihood then has a convenient product form

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{\ell=1}^{s} \left[\pi_\ell f(\boldsymbol{x}_i \mid \boldsymbol{\phi}_\ell) \right]^{I(Z_i = \ell)}.$$

Unlike Newton-Raphson and Fisher scoring, the EM algorithm does not yield an estimate for the asymptotic covariance of the MLE as a byproduct. Several extensions of EM such as (Louis, 1982) and (Meng and Rubin, 1991) have been proposed to yield estimates of the covariance.

Chapter 2

Approximate Information and Scoring Under Multinomial Finite Mixtures

2.1 Introduction

This chapter considers an approximate scoring technique proposed by Morel and Nagaraj (1993), and subsequently investigated in (Neerchal and Morel, 1998) and (Neerchal and Morel, 2005). These authors used the technique to compute maximum likelihood estimates (MLEs) in the study of a multinomial model with extra variation. The model, now known as the random-clumped multinomial (RCM) distribution, has made its way into mainstream use; for example, as an analytical tool in the SAS FMM procedure (SAS Institute Inc., 2011). The RCM distribution can be written as a finite mixture of multinomials, an extension of (Blischke, 1962, 1964), with specific constraints on parameters. Some details on RCM are given later in Example 2.11. Approximate scoring iterations were formulated in (Morel and Nagaraj, 1993) using the observed score vector along with a certain matrix which is an approximation to the Fisher information matrix (FIM). The approximation is motivated by the difficulty in formulating the exact FIM, as it does not have an analytically tractable form and may be expensive to compute accurately by simulation (e.g. Monte Carlo). The matrix approximation has been justified by convergence results showing that the approximate FIM and exact FIM become close for large numbers of multinomial trials.

The present work shows that the approximate scoring algorithm (AFSA) is closely connected to the extremely popular Expectation-Maximization (EM) algorithm (Dempster et al., 1977). In a neighborhood of a solution, the solution is seen to be obtained by both algorithms at the same convergence rate. An explanation for the connection between the two algorithms is provided, in that the FIM approximation is actually a "complete data" information matrix. Closed-form iterations for both EM and AFSA are also obtained, giving expressions with related terms. This work focuses on the finite mixture of multinomials model, motivated by the work on RCM and noting that RCM can be obtained as a special case by enforcing some additional constraints. However, once it is established that AFSA is scoring with a complete data information matrix, its use can be justified for other finite mixture models and missing data problems. For the cases presented in this chapter, an AFSA approach leads to practical procedures for computing MLEs.

A common complaint about EM in its basic form is the convergence rate, which can be slow depending on the proportion of missing data (Dempster et al., 1977). AFSA will be seen to have a similar convergence rate to EM. However, both algorithms possess a certain robustness to the initial value compared to faster methods such as Newton-Raphson or Fisher scoring, and are less likely to get stuck in neighborhoods of poor local maxima or to wander without any progress to a solution. We therefore recommend a hybrid algorithm, making use of both AFSA and exact scoring, where AFSA is used initially to progress to the neighborhood of a solution, and exact scoring is then used to give a fast convergence to that solution. We demonstrate that the proposed hybrid algorithm combines the best features of both AFSA and Fisher scoring.

The rest of the chapter is organized as follows. In Section 2.3, the approximation to the information matrix is presented, along with some of its properties. This approximate information matrix is easily computed and has an immediate application in scoring, which is presented in Section 2.4. Simulation studies are presented in Section 2.5 to illustrate convergence properties of the approximate information matrix and approximate scoring.

Concluding remarks are given in Section 2.6. The contents of this chapter are based on the paper (Raim et al., 2014), which in turn as an extension of the thesis (Liu, 2005). Some of the details from that thesis are reproduced here for completeness.

2.2 Preliminaries and Notation

Given an independent sample $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ with joint likelihood $L(\boldsymbol{\theta})$ and $\boldsymbol{\theta}$ having dimension $q \times 1$, the score vector is

$$S(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x}_i; \boldsymbol{\theta}).$$

For $\boldsymbol{X}_i \sim \text{Mult}_k(\boldsymbol{p}, m)$ the score vector for a single observation can be obtained from

$$\begin{aligned} \frac{\partial}{\partial p_a} \log f(\boldsymbol{x}; \boldsymbol{p}, m) &= \frac{\partial}{\partial p_a} \left[x_1 \log p_1 + \cdots + x_{k-1} \log p_{k-1} + x_k \log \left(1 - \sum_{j=1}^{k-1} p_j \right) \right] \\ &= x_a / p_a - x_k / p_k, \end{aligned} \tag{2.1}$$

so that

$$\frac{\partial}{\partial \boldsymbol{p}} \log f(\boldsymbol{x}; \boldsymbol{p}, m) = \begin{pmatrix} x_1/p_1 \\ \vdots \\ x_{k-1}/p_{k-1} \end{pmatrix} - \begin{pmatrix} x_k/p_k \\ \vdots \\ x_k/p_k \end{pmatrix} = \boldsymbol{D}^{-1} \boldsymbol{x}_{-k} - \frac{x_k}{p_k} \boldsymbol{1},$$

denoting $\boldsymbol{D} := \text{Diag}(p_1, \dots, p_{k-1})$ and $\boldsymbol{x}_{-k} := (x_1, \dots, x_{k-1})$.

The score vector for a single observation $\boldsymbol{X} \sim \text{MultMix}_k(m, \boldsymbol{\theta})$ can also be obtained,

$$\begin{aligned}\frac{\partial \log \mathrm{P}(\boldsymbol{x})}{\partial \boldsymbol{p}_a} &= \frac{\partial \log\{\sum_{\ell=1}^{s} \pi_\ell \,\mathrm{P}_\ell(\boldsymbol{x})\}}{\partial \boldsymbol{p}_a} \\ &= \frac{1}{\mathrm{P}(\boldsymbol{x})}\pi_a \frac{\partial \,\mathrm{P}_a(\boldsymbol{x})}{\partial \boldsymbol{p}_a} \\ &= \frac{\pi_a \,\mathrm{P}_a(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} \frac{\partial \log \mathrm{P}_a(\boldsymbol{x})}{\partial \boldsymbol{p}_a} \\ &= \frac{\pi_a \,\mathrm{P}_a(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} \left[\boldsymbol{D}_a^{-1}\boldsymbol{x}_{-k} - \frac{x_k}{p_{ak}}\mathbf{1}\right], \qquad a = 1, \ldots, s,\end{aligned}$$

where $\boldsymbol{D}_a := \mathrm{Diag}(p_{a1}, \ldots, p_{a,k-1})$, and

$$\begin{aligned}\frac{\partial \log \mathrm{P}(\boldsymbol{x})}{\partial \pi_a} &= \frac{\partial \log\{\sum_{\ell=1}^{s} \pi_\ell \,\mathrm{P}_\ell(\boldsymbol{x})\}}{\partial \pi_a} \\ &= \frac{\mathrm{P}_a(\boldsymbol{x}) - \mathrm{P}_s(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})}, \qquad a = 1, \ldots, s-1.\end{aligned}$$

Next, consider the $q \times q$ FIM for the independent sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$

$$\begin{aligned}\mathcal{I}(\boldsymbol{\theta}) = \mathrm{Var}(S(\boldsymbol{\theta})) &= \mathrm{E}\left[\left\{\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta})\right\}\left\{\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta})\right\}^T\right] \\ &= \mathrm{E}\left[-\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log L(\boldsymbol{\theta})\right].\end{aligned}$$

The last equality holds under appropriate regularity conditions. For the multinomial FIM, we may use (2.1) to obtain

$$\frac{\partial}{\partial p_a}\frac{\partial}{\partial p_b} \log f(\boldsymbol{x}; \boldsymbol{p}, m) = \begin{cases} x_k/p_k^2 & \text{if } a \neq b \\ -x_a/p_a^2 - x_k/p_k^2 & \text{otherwise} \end{cases}$$

and so

$$\frac{\partial}{\partial \boldsymbol{p} \partial \boldsymbol{p}^T} \log f(\boldsymbol{x}; \boldsymbol{p}, m) = \mathrm{Diag}\left(-\frac{x_1}{p_1^2}, \ldots, -\frac{x_{k-1}}{p_{k-1}^2}\right) - \frac{x_k}{p_k^2}\mathbf{1}\mathbf{1}^T.$$

Therefore, we have

$$\begin{aligned}
\mathcal{I}(\boldsymbol{p}) &= \mathrm{E}\left(-\frac{\partial}{\partial\boldsymbol{p}\partial\boldsymbol{p}^T}\log f(\boldsymbol{x};\boldsymbol{p},m)\right) \\
&= \mathrm{Diag}\left(\frac{mp_1}{p_1^2},\ldots,\frac{mp_{k-1}}{p_{k-1}^2}\right) + \frac{mp_k}{p_k^2}\mathbf{1}\mathbf{1}^T \\
&= m\left(\boldsymbol{D}^{-1} + p_k^{-1}\mathbf{1}\mathbf{1}^T\right).
\end{aligned}$$

The score vector and Hessian of the log-likelihood can be used to implement the Newton-Raphson algorithm, where the $(g+1)$th iteration is given by

$$\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)} - \left\{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log L(\boldsymbol{\theta}^{(g)})\right\}^{-1} S(\boldsymbol{\theta}^{(g)}).$$

The Hessian may be replaced with the FIM to implement Fisher Scoring

$$\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)} + \mathcal{I}^{-1}(\boldsymbol{\theta}^{(g)})\, S(\boldsymbol{\theta}^{(g)}).$$

In order for the estimation problem to be well-defined in the first place, the model must be identifiable. For finite mixtures, this is taken to mean that the equality

$$\sum_{\ell=1}^{s}\pi_\ell f(\boldsymbol{x};\boldsymbol{\theta}_\ell) \overset{a.s.}{=} \sum_{\ell=1}^{v}\lambda_\ell f(\boldsymbol{x};\boldsymbol{\xi}_\ell)$$

implies $s = v$, $\pi_\ell = \lambda_{\rho(\ell)}$, and $\boldsymbol{p}_\ell = \boldsymbol{\xi}_{\rho(\ell)}$ for all $\ell = 1,\ldots,s$, where $(\rho(1),\ldots,\rho(s))$ is some permutation of $(1,\ldots,s)$ (McLachlan and Peel, 2000, Section 1.14). Chandra (1977) provides some insight into the identifiability issue, and relates the identifiability of a family of multivariate mixtures to its corresponding marginal mixtures. In the present case, the multivariate mixtures consist of multinomial densities, and the univariate marginal densities are binomials. It is known that a finite mixture of s components from the family $\{\ \mathrm{Mult}_k(m,\boldsymbol{p}) : \boldsymbol{p}\in(0,1)^k, \sum_{j=1}^k p_j = 1\ \}$ is identifiable if and only if $m \geq 2s-1$; see, for example, Elmore and Wang (2003). Then a sufficient condition for model (2.3) to be identifiable is that $m_i \geq 2s - 1$ for at least one observation. This can be seen by the following lemma.

Lemma 2.1. *Suppose* $\boldsymbol{X}_i \overset{ind}{\sim} f_i(\boldsymbol{x}; \boldsymbol{\theta}), i = 1, \ldots, n$*, where* f_i *share a common parameter* $\boldsymbol{\theta}$*, and for at least one* $r \in \{1, \ldots, n\}$ *the family* $\{f_r(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ *is identifiable. Then the joint model is identifiable.*

Proof. WLOG assume that $r = 1$, and suppose we have

$$\prod_{i=1}^{n} f_i(\boldsymbol{x}_i; \boldsymbol{\theta}) \overset{a.s.}{=} \prod_{i=1}^{n} f_i(\boldsymbol{x}_i; \boldsymbol{\xi}).$$

Integrating both sides with respect to $\boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, using the appropriate dominating measure,

$$f_1(\boldsymbol{x}_1; \boldsymbol{\theta}) \overset{a.s.}{=} f_1(\boldsymbol{x}_1; \boldsymbol{\xi}).$$

Since the family $\{f_1(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ is identifiable, this implies $\boldsymbol{\theta} = \boldsymbol{\xi}$. Hence the joint family $\{\prod_{i=1}^{n} f_i(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ is identifiable. □

2.3 An Approximation to the Information Matrix

Consider the multinomial sample space with m trials placed into k categories at random,

$$\Omega = \Big\{(x_1, \ldots, x_k) : \; x_j \in \{0, 1, \ldots, m\}, \; \sum_{j=1}^{k} x_j = m\Big\}.$$

The standard multinomial density is

$$f(\boldsymbol{x}; \boldsymbol{p}, m) = \frac{m!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k} \cdot I(\boldsymbol{x} \in \Omega),$$

where $I(\cdot)$ is the indicator function, and the parameter space is

$$\Theta = \left\{(p_1, \dots, p_{k-1}) : \ 0 < p_j < 1, \ \sum_{j=1}^{k-1} p_j < 1\right\} \subseteq \mathbb{R}^{k-1}.$$

If a random variable $\boldsymbol{X}$ has distribution $f(\boldsymbol{x}; \boldsymbol{p}, m)$, we will write $\boldsymbol{X} \sim \text{Mult}_k(\boldsymbol{p}, m)$. Following the sampling and overdispersion literature, we will refer to the number of trials m as the "cluster size" of a multinomial observation.

Suppose there are s multinomial populations $\text{Mult}_k(\boldsymbol{p}_1, m), \dots, \text{Mult}_k(\boldsymbol{p}_s, m)$, where $\boldsymbol{p}_\ell = (p_{\ell 1}, \dots, p_{\ell,k-1})$ for $\ell = 1, \dots, s$, and the ℓth population occurs with proportion π_ℓ in the mixed population. If we draw $\boldsymbol{X}$ from the mixed population, its probability density is a finite mixture of multinomials

$$f(\boldsymbol{x}; \boldsymbol{\theta}, m) = \sum_{\ell=1}^{s} \pi_\ell f(\boldsymbol{x}; \boldsymbol{p}_\ell, m), \quad \text{with } \boldsymbol{\theta} = (\boldsymbol{p}_1, \dots, \boldsymbol{p}_s, \boldsymbol{\pi}) \tag{2.2}$$

and we will write $\boldsymbol{X} \sim \text{MultMix}_k(m, \boldsymbol{\theta})$. The dimension of $\boldsymbol{\theta}$ is $q = s(k-1) + (s-1) = sk - 1$, disregarding the redundant parameters $p_{1k}, \dots, p_{sk}, \pi_s$. We will also make use of the following slightly-less-cumbersome notation for densities: $\text{P}(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\theta}, m)$ for the mixture, and $\text{P}_\ell(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{p}_\ell, m)$ for the ℓth component of the mixture. The setting of this chapter will be an independent sample $\boldsymbol{X}_i \sim \text{MultMix}_k(m_i, \boldsymbol{\theta})$, for $i = 1, \dots, n$, with cluster sizes not necessarily equal; the resulting likelihood is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \sum_{\ell=1}^{s} \pi_\ell \left[\frac{m_i!}{x_{i1}! \dots x_{ik}!} p_{\ell 1}^{x_{i1}} \dots p_{\ell k}^{x_{ik}} \cdot I(\boldsymbol{x}_i \in \Omega) \right] \right\}. \tag{2.3}$$

The inner summation prevents closed-form likelihood maximization, hence our goal will be to compute the MLE $\hat{\boldsymbol{\theta}}$ numerically. Some additional preliminaries are given in the supplement.

In general, the information matrix for mixtures involves a complicated expectation which does not have a tractable form. Since the multinomial mixture has a finite sample space, it can be computed naively by using the definition of the expectation

$$\mathcal{I}(\boldsymbol{\theta}) = \sum_{\boldsymbol{x} \in \Omega} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x}; \boldsymbol{\theta}) \right\} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x}; \boldsymbol{\theta}) \right\}^T f(\boldsymbol{x}; \boldsymbol{\theta}), \tag{2.4}$$

given a particular value for $\boldsymbol{\theta}$. Although the number of terms $\binom{k+m-1}{m}$ in the summation is finite, it grows quickly with m and k, and this method becomes intractable as m and k increase. For example, when $m = 100$ and $k = 10$, the sample space Ω contains more than 4.2 trillion elements. To avoid these potentially expensive computations, we extend the approximate FIM approach of Morel and Nagaraj (1993) to the general finite mixture of multinomials. The following theorem presents the approximation and its justification. It was originally proved at this level of generality in (Liu, 2005) although some cases had been omitted; the proof is reproduced here for completeness. The reader may also refer to (Morel and Nagaraj, 1991) which addresses the $k = s$ case, as needed for the random-clumped multinomial distribution.

Theorem 2.2. *Suppose* $\boldsymbol{X} \sim \text{MultMix}_k(m, \boldsymbol{\theta})$ *is a single observation from the mixed population. Denote the exact FIM with respect to* $\boldsymbol{X}$ *as* $\mathcal{I}(\boldsymbol{\theta})$*. Then an approximation to the FIM with respect to* $\boldsymbol{X}$ *is given by the* $(sk-1) \times (sk-1)$ *block-diagonal matrix*

$$\widetilde{\mathcal{I}}(\boldsymbol{\theta}) := \text{Blockdiag}\left(\pi_1 \boldsymbol{F}_1, \ldots, \pi_s \boldsymbol{F}_s, \boldsymbol{F}_\pi\right),$$

where for $\ell = 1, \ldots, s$

$$\boldsymbol{F}_\ell = m\left[\boldsymbol{D}_\ell^{-1} + p_{\ell k}^{-1} \mathbf{1}\mathbf{1}^T\right] \quad \textit{and} \quad \boldsymbol{D}_\ell = \text{Diag}(p_{\ell 1}, \ldots, p_{\ell, k-1})$$

are $(k-1) \times (k-1)$ *matrices,*

$$\boldsymbol{F}_\pi = \boldsymbol{D}_\pi^{-1} + \pi_s^{-1} \mathbf{1}\mathbf{1}^T \quad \textit{and} \quad \boldsymbol{D}_\pi = \text{Diag}(\pi_1, \ldots, \pi_{s-1})$$

are $(s-1)\times(s-1)$ *matrices, and* $\mathbf{1}$ *denotes a vector of ones of the appropriate dimension. To emphasize the dependence of the FIM and the approximation on* m*, we will also write* $\mathcal{I}_m(\boldsymbol{\theta})$ *and* $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta})$*. If the vectors* $\mathbf{p}_1, \ldots, \mathbf{p}_s$ *are distinct (i.e.* $\mathbf{p}_a \neq \mathbf{p}_b$ *for every pair of populations* $a \neq b$*), then* $\mathcal{I}_m(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) \to \mathbf{0}$ *as* $m \to \infty$*.*

Notice that the matrix $\boldsymbol{F}_\ell$ is exactly the FIM of $\text{Mult}_k(\boldsymbol{p}_\ell, m)$ for the ℓth population, and $\boldsymbol{F}_\pi$ is the FIM of $\text{Mult}_s(\boldsymbol{\pi}, 1)$ corresponding to the mixing probabilities $\boldsymbol{\pi}$. To prove Theorem 2.2, we will first establish a key inequality from Okamoto (1959) for the tail probability of the binomial distribution, which was also considered by Blischke (1962).

Lemma 2.3. *Suppose* $X \sim$ *Binomial*(m, p)*, then for* $c \geq 0$*,*

i. $\text{P}(X/m - p \geq c) \leq e^{-2mc^2}$,

ii. $\text{P}(X/m - p \leq -c) \leq e^{-2mc^2}$.

Note that the inequalities in Lemma 2.3 can be generalized to where X is a sum of independent bounded random variables (Hoeffding, 1963), and the bounds on the probabilities may be tightened as well. In Chapter 3, where we extend the convergence results from the present chapter, we will instead take a different approach.

Theorem 2.4. *For a given index* $b \in \{1, \ldots, s\}$ *we have*

$$\sum_{\boldsymbol{x}\in\Omega} \sum_{a\neq b}^{s} \frac{\pi_a \, \text{P}_a(x) \, \text{P}_b(x)}{\text{P}(x)} \leq \frac{2}{\pi_b} \sum_{a\neq b}^{s} e^{-\frac{m}{2}\delta_{ab}^2},$$

where $\delta_{ab} = \bigvee_{j=1}^{k} |p_{aj} - p_{bj}|$*.*

Proof. Denote as $\Omega(x_j)$ the multinomial sample space when the jth element of $\boldsymbol{x}$ is fixed at a number x_j. Then we have, for any $L \in \{1, \ldots, k\}$,

$$\begin{aligned}
\sum_{\boldsymbol{x}\in\Omega} \frac{\pi_a \,\mathrm{P}_a(\boldsymbol{x})\,\mathrm{P}_b(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} &= \sum_{x_L=0}^{m} \sum_{\boldsymbol{x}\in\Omega(x_L)} \frac{\pi_a \,\mathrm{P}_a(\boldsymbol{x})\,\mathrm{P}_b(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} \\
&= \sum_{x_L \le \frac{m}{2}(p_{aL}+p_{bL})} \sum_{\boldsymbol{x}\in\Omega(x_L)} \pi_a \,\mathrm{P}_a(\boldsymbol{x}) \frac{\mathrm{P}_b(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} + \sum_{x_L > \frac{m}{2}(p_{aL}+p_{bL})} \sum_{\boldsymbol{x}\in\Omega(x_L)} \frac{\pi_a \,\mathrm{P}_a(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} \mathrm{P}_b(\boldsymbol{x}) \\
&\le \sum_{x_L \le \frac{m}{2}(p_{aL}+p_{bL})} \sum_{\boldsymbol{x}\in\Omega(x_L)} \frac{\pi_a}{\pi_b} \mathrm{P}_a(\boldsymbol{x}) + \sum_{x_L > \frac{m}{2}(p_{aL}+p_{bL})} \sum_{\boldsymbol{x}\in\Omega(x_L)} \mathrm{P}_b(\boldsymbol{x}) \\
&= \frac{\pi_a}{\pi_b} \sum_{x_L \le \frac{m}{2}(p_{aL}+p_{bL})} \sum_{\boldsymbol{x}\in\Omega(x_L)} \mathrm{P}_a(\boldsymbol{x}) + \sum_{x_L > \frac{m}{2}(p_{aL}+p_{bL})} \sum_{\boldsymbol{x}\in\Omega(x_L)} \mathrm{P}_b(\boldsymbol{x}). \qquad (2.5)
\end{aligned}$$

Notice that the last statement above consists of marginal probabilities for the Lth coordinate of k-dimensional multinomials, which are binomial probabilities. Following Blischke (1962), suppose $A \sim \text{Binomial}(m, p_{aL})$ and $B \sim \text{Binomial}(m, p_{bL})$, then (2.5) is equal to

$$\frac{\pi_a}{\pi_b} \mathrm{P}\left\{A \le \frac{m}{2}(p_{aL}+p_{bL})\right\} + \mathrm{P}\left\{B > \frac{m}{2}(p_{aL}+p_{bL})\right\}. \qquad (2.6)$$

Taking $c = \frac{1}{2}(p_{aL} - p_{bL})$ yields

$$\begin{aligned}
m(p_{aL} - c) &= \frac{m}{2}(p_{aL}+p_{bL}), \\
m(p_{bL} + c) &= \frac{m}{2}(p_{aL}+p_{bL}),
\end{aligned}$$

and (2.6) is equivalent to

$$\begin{aligned}
&\frac{\pi_a}{\pi_b} \mathrm{P}\{A \le m(p_{aL} - c)\} + \mathrm{P}\{B > m(p_{bL} + c)\} \\
&\quad = \frac{\pi_a}{\pi_b} \mathrm{P}\{A/m - p_{aL} \le -c\} + \mathrm{P}\{B/m - p_{bL} > c\} \\
&\quad \le \frac{\pi_a}{\pi_b} e^{-2mc^2} + e^{-2mc^2}, \qquad \text{by Lemma 2.3} \\
&\quad = \left(\frac{\pi_a + \pi_b}{\pi_b}\right) e^{-\frac{1}{2}m(p_{aL}-p_{bL})^2}.
\end{aligned}$$

The upper bound can be made as small as possible by selecting $L \in \{1, \ldots, k\}$ to obtain the largest possible $(p_{aL} - p_{bL})^2$; i.e.

$$\frac{\pi_a}{\pi_b} \mathrm{P}\left\{A \le m(p_{aL} - c)\right\} + \mathrm{P}\left\{B > m(p_{bL} + c)\right\} = \left(\frac{\pi_a + \pi_b}{\pi_b}\right) e^{-\frac{1}{2} m \delta_{ab}^2}.$$

Now we have

$$\begin{aligned}
\sum_{\boldsymbol{x} \in \Omega} \sum_{a \neq b}^{s} \frac{\pi_a \mathrm{P}_a(\boldsymbol{x}) \mathrm{P}_b(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} &= \sum_{a \neq b}^{s} \sum_{\boldsymbol{x} \in \Omega} \frac{\pi_a \mathrm{P}_a(\boldsymbol{x}) \mathrm{P}_b(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} \\
&\le \sum_{a \neq b}^{s} \frac{\pi_a + \pi_b}{\pi_b} e^{-\frac{m}{2}(p_{aL} - p_{bL})^2} \\
&\le \frac{2}{\pi_b} \sum_{a \neq b}^{s} e^{-\frac{m}{2}(p_{aL} - p_{bL})^2}.
\end{aligned}$$

□

Corollary 2.5. *The following intermediate result was obtained in the proof of Theorem 2.4*

$$\sum_{\boldsymbol{x} \in \Omega} \frac{\pi_a \mathrm{P}_a(\boldsymbol{x}) \mathrm{P}_b(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} \le \left(\frac{\pi_a + \pi_b}{\pi_b}\right) e^{-\frac{1}{2} m \delta_{ab}^2} \le \frac{2}{\pi_b} e^{-\frac{1}{2} m \delta_{ab}^2}.$$

We are now prepared to prove Theorem 2.2. Following the strategy of Morel and Nagaraj (1991), we consider the difference between the $\mathcal{I}(\boldsymbol{\theta})$ and the limiting matrix $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ element by element for finite cluster sizes and obtain bounds which converge to zero as $m \to \infty$. The bound used by Morel and Nagaraj (1991) is slightly different than ours, since we do not require that $k = s$.

Proof of Theorem 2.2. Partition the exact FIM as

$$\mathcal{I}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{pmatrix}$$

where

$$\boldsymbol{C}_{11} = \begin{pmatrix} \boldsymbol{A}_{11} & \dots & \boldsymbol{A}_{1s} \\ \vdots & \ddots & \vdots \\ \boldsymbol{A}_{s1} & \dots & \boldsymbol{A}_{ss} \end{pmatrix}, \qquad \boldsymbol{C}_{12} = \begin{pmatrix} \boldsymbol{A}_{1\pi} \\ \vdots \\ \boldsymbol{A}_{s\pi} \end{pmatrix} = \boldsymbol{C}_{21}^T, \qquad \boldsymbol{C}_{22} = \boldsymbol{A}_{\pi\pi},$$

and

$$\begin{aligned}
\boldsymbol{A}_{ab} &= \mathrm{E}\left(\left\{ \frac{\partial \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{p}_a} \right\} \left\{ \frac{\partial \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{p}_b} \right\}^T \right), \quad \text{for } a = 1,\dots,s \text{ and } b = 1,\dots,s, \\
\boldsymbol{A}_{\pi b} &= \mathrm{E}\left(\left\{ \frac{\partial \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\pi}} \right\} \left\{ \frac{\partial \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{p}_b} \right\}^T \right), \quad \text{for } b = 1,\dots,s \\
&= \boldsymbol{A}_{b\pi}^T, \\
\boldsymbol{A}_{\pi\pi} &= \mathrm{E}\left(\left\{ \frac{\partial \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\pi}} \right\} \left\{ \frac{\partial \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\pi}} \right\}^T \right).
\end{aligned}$$

We must show that as $m \to \infty$,

$$\boldsymbol{C}_{11} - \mathrm{Blockdiag}(\pi_1 \boldsymbol{F}_1, \dots, \pi_s \boldsymbol{F}_s) \to \boldsymbol{0}, \tag{2.7}$$

$$\boldsymbol{C}_{21}^T = \boldsymbol{C}_{12} \to \boldsymbol{0}, \tag{2.8}$$

$$\boldsymbol{C}_{22} - \boldsymbol{F}_\pi \to \boldsymbol{0}. \tag{2.9}$$

Case (i) First consider the (i,i)th block of $\boldsymbol{C}_{11} - \mathrm{Blockdiag}(\pi_1\boldsymbol{F}_1, \ldots, \pi_s\boldsymbol{F}_s)$

$$\begin{aligned}
&\boldsymbol{D}_i\left(\boldsymbol{A}_{ii} - \pi_i\boldsymbol{F}_i\right)\boldsymbol{D}_i \\
&\quad= \boldsymbol{D}_i\left\{\mathrm{E}\left[\left\{\frac{\partial}{\partial\boldsymbol{p}_i}\log\mathrm{P}(\boldsymbol{x})\right\}\left\{\frac{\partial}{\partial\boldsymbol{p}_i}\log\mathrm{P}(\boldsymbol{x})\right\}^T\right] - \pi_i\boldsymbol{F}_i\right\}\boldsymbol{D}_i \\
&\quad= \pi_i^2\boldsymbol{D}_i\,\mathrm{E}\left[\frac{\mathrm{P}_i^2(\boldsymbol{x})}{\mathrm{P}^2(\boldsymbol{x})}\frac{\partial\log\mathrm{P}_i(\boldsymbol{x})}{\partial\boldsymbol{p}_i}\frac{\partial\log\mathrm{P}_i(\boldsymbol{x})}{\partial\boldsymbol{p}_i^T}\right]\boldsymbol{D}_i - \pi_i\boldsymbol{D}_i\boldsymbol{F}_i\boldsymbol{D}_i \\
&\quad= \pi_i^2\sum_{\boldsymbol{x}\in\Omega}\frac{\mathrm{P}_i(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})}\left(\boldsymbol{x}_{-k} - \frac{x_k}{p_{ik}}\boldsymbol{p}_i\right)\left(\boldsymbol{x}_{-k} - \frac{x_k}{p_{ik}}\boldsymbol{p}_i\right)^T\mathrm{P}_i(\boldsymbol{x}) \\
&\qquad\quad - \pi_i^2\sum_{\boldsymbol{x}\in\Omega}\frac{1}{\pi_i}\left(\boldsymbol{x}_{-k} - \frac{x_k}{p_{ik}}\boldsymbol{p}_i\right)\left(\boldsymbol{x}_{-k} - \frac{x_k}{p_{ik}}\boldsymbol{p}_i\right)^T\mathrm{P}_i(\boldsymbol{x}) \qquad (2.10) \\
&\quad= \pi_i^2\sum_{\boldsymbol{x}\in\Omega}\left(\boldsymbol{x}_{-k} - \frac{x_k}{p_{ik}}\boldsymbol{p}_i\right)\left(\boldsymbol{x}_{-k} - \frac{x_k}{p_{ik}}\boldsymbol{p}_i\right)^T\left(\frac{\mathrm{P}_i(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} - \frac{1}{\pi_i}\right)\mathrm{P}_i(\boldsymbol{x}) \\
&\quad= \frac{\pi_i}{p_{ik}^2}\sum_{\boldsymbol{x}\in\Omega}\left(p_{ik}\boldsymbol{x}_{-k} - x_k\boldsymbol{p}_i\right)\left(p_{ik}\boldsymbol{x}_{-k} - x_k\boldsymbol{p}_i\right)^T\left(\frac{\pi_i\,\mathrm{P}_i(\boldsymbol{x}) - \mathrm{P}(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})}\right)\mathrm{P}_i(\boldsymbol{x}). \qquad (2.11)
\end{aligned}$$

where x_k is the kth element of $\boldsymbol{x}$ and $\boldsymbol{x}_{-k} = (x_1, \ldots, x_{k-1})$. We have pre and post-multiplied by $\boldsymbol{D}_i$ so that Theorem 2.4 can be applied. But note that since $\boldsymbol{D}_i$ does not vary over m,

$$\boldsymbol{D}_i\left\{\boldsymbol{A}_{ii} - \pi_i\boldsymbol{F}_i\right\}\boldsymbol{D}_i \to \boldsymbol{0} \quad\Longrightarrow\quad \boldsymbol{A}_{ii} - \pi_i\boldsymbol{F}_i \to \boldsymbol{0}, \qquad \text{as } m \to \infty.$$

We have also used the fact in step (2.10) that

$$\boldsymbol{D}_i\frac{\partial\log\mathrm{P}_i(\boldsymbol{x})}{\partial\boldsymbol{p}_i} = \boldsymbol{D}_i\left\{\boldsymbol{D}_i^{-1}\boldsymbol{x}_{-k} - \frac{x_k}{p_{ik}}\boldsymbol{1}\right\} = \boldsymbol{x}_{-k} - \frac{x_k}{p_{ik}}\boldsymbol{p}_i.$$

We next have for $r, s \in \{1, \ldots, k-1\}$

$$[p_{ik}x_r - x_kp_{ir}]^2 \le [x_r + mp_{ir}]^2 \le 4m^2.$$

Also,

$$\begin{aligned}0 &\leq \Big[[p_{ik}x_r - x_k p_{ir}] + [p_{ik}x_s - x_k p_{is}]\Big]^2 \\ &= [p_{ik}x_r - x_k p_{ir}]^2 + [p_{ik}x_s - x_k p_{is}]^2 + 2[p_{ik}x_r - x_k p_{ir}][p_{ik}x_s - x_k p_{is}]\end{aligned}$$

and similarly

$$\begin{aligned}0 &\leq \Big[[p_{ik}x_r - x_k p_{ir}] - [p_{ik}x_s - x_k p_{is}]\Big]^2 \\ &= [p_{ik}x_r - x_k p_{ir}]^2 + [p_{ik}x_s - x_k p_{is}]^2 - 2[p_{ik}x_r - x_k p_{ir}][p_{ik}x_s - x_k p_{is}],\end{aligned}$$

which implies that

$$\begin{aligned}\Big|[p_{ik}x_r - x_k p_{ir}][p_{ik}x_s - x_k p_{is}]\Big| &\leq \frac{1}{2}\Big\{[x_r + mp_{ir}]^2 + [x_s + mp_{is}]^2\Big\} \\ &\leq 4m^2.\end{aligned}$$

Notice that this bound is free of r and s, so it holds uniformly over all $r, s \in \{1, \ldots, k{-}1\}$. If we denote the (r, s)th element of the matrix given in (2.11) by ε_{rs}, we have

$$\begin{aligned}|\varepsilon_{rs}| &\leq \frac{4\pi_i m^2}{p_{ik}^2} \sum_{\boldsymbol{x}\in\Omega} \frac{\mathrm{P}(x) - \pi_i \,\mathrm{P}_i(x)}{\mathrm{P}(x)} \,\mathrm{P}_i(x) = \frac{4\pi_i m^2}{p_{ik}^2} \sum_{\boldsymbol{x}\in\Omega} \sum_{j\neq i}^{s} \frac{\pi_j \,\mathrm{P}_i(x) \,\mathrm{P}_j(x)}{\mathrm{P}(x)} \\ &\leq \frac{8m^2}{p_{ik}^2} \sum_{j\neq i}^{s} e^{-\frac{m}{2}\delta_{ij}^2},\end{aligned}$$

by Theorem 2.4. By assumption, $\delta_{ij}^2 > 0$ for $i \neq j$, and therefore $\varepsilon_{rs} \to 0$ as $m \to \infty$.

Case (ii) Next, consider the (i,j)th block of $\boldsymbol{C}_{11} - \mathrm{Blockdiag}(\pi_1\boldsymbol{F}_1, \ldots, \pi_s\boldsymbol{F}_s)$ where $i \neq j$.

$$\begin{aligned}
&\boldsymbol{D}_i\boldsymbol{A}_{ij}\boldsymbol{D}_j \\
&\quad= \boldsymbol{D}_i\left\{\mathrm{E}\left[\left\{\frac{\partial}{\partial\boldsymbol{p}_i}\log\mathrm{P}(\boldsymbol{x})\right\}\left\{\frac{\partial}{\partial\boldsymbol{p}_j}\log\mathrm{P}(\boldsymbol{x})\right\}^T\right]\right\}\boldsymbol{D}_j \\
&\quad= \boldsymbol{D}_i\left[\mathrm{E}\left(\frac{\pi_i\pi_j}{\mathrm{P}^2(\boldsymbol{x})}\frac{\partial\,\mathrm{P}_i(\boldsymbol{x})}{\partial\boldsymbol{p}_i}\frac{\partial\,\mathrm{P}_j(\boldsymbol{x})}{\partial\boldsymbol{p}_j^T}\right)\right]\boldsymbol{D}_j \\
&\quad= \pi_i\pi_j\boldsymbol{D}_i\left[\mathrm{E}\left(\frac{\mathrm{P}_i(\boldsymbol{x})\,\mathrm{P}_j(\boldsymbol{x})}{\mathrm{P}^2(\boldsymbol{x})}\frac{\partial\log\mathrm{P}_i(\boldsymbol{x})}{\partial\boldsymbol{p}_i}\frac{\partial\log\mathrm{P}_j(\boldsymbol{x})}{\partial\boldsymbol{p}_j^T}\right)\right]\boldsymbol{D}_j \\
&\quad= \pi_i\pi_j\sum_{\boldsymbol{x}\in\Omega}\frac{\mathrm{P}_i(\boldsymbol{x})\,\mathrm{P}_j(\boldsymbol{x})}{\mathrm{P}^2(\boldsymbol{x})}\left(\boldsymbol{x}_{-k} - \frac{x_k}{p_{ik}}\boldsymbol{p}_i\right)\left(\boldsymbol{x}_{-k} - \frac{x_k}{p_{jk}}\boldsymbol{p}_j\right)^T\mathrm{P}(\boldsymbol{x}) \\
&\quad= \frac{\pi_i\pi_j}{p_{ik}p_{jk}}\sum_{\boldsymbol{x}\in\Omega}\frac{\mathrm{P}_i(\boldsymbol{x})\,\mathrm{P}_j(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})}\left(p_{ik}\boldsymbol{x}_{-k} - x_k\boldsymbol{p}_i\right)\left(p_{jk}\boldsymbol{x}_{-k} - x_k\boldsymbol{p}_j\right)^T. \qquad (2.12)
\end{aligned}$$

If we now denote the (r,s)th element of the matrix given in (2.12) by ε_{rs}, we have

$$|\varepsilon_{rs}| \leq \frac{4\pi_i\pi_j m^2}{p_{ik}p_{jk}}\sum_{\boldsymbol{x}\in\Omega}\frac{\mathrm{P}_i(x)\,\mathrm{P}_j(x)}{\mathrm{P}(x)} \leq \frac{8m^2}{p_{ik}p_{jk}}e^{-\frac{m}{2}\delta_{ij}^2}$$

for all (r,s), applying Theorem 2.5 and a similar argument to Case (i). Since $\delta_{ij}^2 > 0$ for $i \neq j$, $\varepsilon_{rs} \to 0$ as $m \to \infty$.

Case (iii) Now consider the matrix

$$\begin{aligned}
&\boldsymbol{A}_{\pi\pi} - \boldsymbol{F}_\pi \qquad (2.13)\\
&\quad = \mathrm{E}\left[\left\{\frac{\partial}{\partial\boldsymbol{\pi}}\log\mathrm{P}(\boldsymbol{x})\right\}\left\{\frac{\partial}{\partial\boldsymbol{\pi}}\log\mathrm{P}(\boldsymbol{x})\right\}^T\right] - \boldsymbol{F}_\pi\\
&\quad = \mathrm{E}\left[\frac{1}{\mathrm{P}^2(\boldsymbol{x})}\left\{\begin{pmatrix}\mathrm{P}_1(\boldsymbol{x})\\ \vdots \\ \mathrm{P}_{s-1}(\boldsymbol{x})\end{pmatrix} - \mathrm{P}_s(\boldsymbol{x})\cdot\mathbf{1}\right\}\left\{\begin{pmatrix}\mathrm{P}_1(\boldsymbol{x})\\ \vdots \\ \mathrm{P}_{s-1}(\boldsymbol{x})\end{pmatrix} - \mathrm{P}_s(\boldsymbol{x})\cdot\mathbf{1}\right\}^T\right]\\
&\qquad - \left(\boldsymbol{D}_\pi^{-1} + \pi_s^{-1}\mathbf{1}\mathbf{1}^T\right).
\end{aligned}$$

Pick out the (a, a)th entry which we will denote as ε_{aa}. We have

$$\begin{aligned}
\varepsilon_{aa} &= \mathrm{E}\left[\frac{[\mathrm{P}_a(\boldsymbol{x}) - \mathrm{P}_s(\boldsymbol{x})]^2}{\mathrm{P}^2(\boldsymbol{x})}\right] - (\pi_a^{-1} + \pi_s^{-1})\\
&= \sum_{\boldsymbol{x}\in\Omega}\frac{\mathrm{P}_a^2(\boldsymbol{x}) - 2\,\mathrm{P}_a(\boldsymbol{x})\,\mathrm{P}_s(\boldsymbol{x}) + \mathrm{P}_s^2(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} - (\pi_a^{-1} + \pi_s^{-1})\\
&= \sum_{\boldsymbol{x}\in\Omega}\left(\frac{\mathrm{P}_a^2(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} - \frac{\mathrm{P}_a(\boldsymbol{x})}{\pi_a}\right) + \sum_{\boldsymbol{x}\in\Omega}\left(\frac{\mathrm{P}_s^2(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} - \frac{\mathrm{P}_s(\boldsymbol{x})}{\pi_s}\right) - 2\sum_{\boldsymbol{x}\in\Omega}\frac{\mathrm{P}_a(\boldsymbol{x})\,\mathrm{P}_s(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})}\\
&= \frac{1}{\pi_a}\sum_{\boldsymbol{x}\in\Omega}\frac{\pi_a\,\mathrm{P}_a(\boldsymbol{x}) - \mathrm{P}(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})}\,\mathrm{P}_a(\boldsymbol{x}) + \frac{1}{\pi_s}\sum_{\boldsymbol{x}\in\Omega}\frac{\pi_s\,\mathrm{P}_s(\boldsymbol{x}) - \mathrm{P}(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})}\,\mathrm{P}_s(\boldsymbol{x}) - 2\sum_{\boldsymbol{x}\in\Omega}\frac{\mathrm{P}_a(\boldsymbol{x})\,\mathrm{P}_s(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})}\\
&= -\frac{1}{\pi_a}\sum_{\boldsymbol{x}\in\Omega}\sum_{\ell\neq a}^{s}\frac{\pi_\ell\,\mathrm{P}_\ell(\boldsymbol{x})\,\mathrm{P}_a(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} - \frac{1}{\pi_s}\sum_{\boldsymbol{x}\in\Omega}\sum_{\ell\neq s}^{s}\frac{\pi_\ell\,\mathrm{P}_\ell(\boldsymbol{x})\,\mathrm{P}_s(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} - \frac{2}{\pi_a}\sum_{\boldsymbol{x}\in\Omega}\frac{\pi_a\,\mathrm{P}_a(\boldsymbol{x})\,\mathrm{P}_s(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})}
\end{aligned}$$

Then by the triangle inequality,

$$|\varepsilon_{aa}| \le \frac{2}{\pi_a^2}\sum_{\ell\neq a}^{s} e^{-\frac{m}{2}\delta_{\ell a}^2} + \frac{2}{\pi_s^2}\sum_{\ell\neq s}^{s} e^{-\frac{m}{2}\delta_{\ell s}^2} + \frac{4}{\pi_a\pi_s}e^{-\frac{m}{2}\delta_{as}^2},$$

applying Theorem 2.4 to the first two terms, and Corollary 2.5 to the last term. Since $\delta_{ij}^2 > 0$ for $i \neq j$, we have $\varepsilon_{aa} \to 0$ for $a \in \{1, \ldots, s-1\}$ as $m \to \infty$.

Case (iv) Consider again the matrix $\boldsymbol{A}_{\pi\pi} - \boldsymbol{F}_\pi$ from (2.13), but now the case where $a \neq b$. We have

$$\begin{aligned}\varepsilon_{ab} &= \mathrm{E}\left[\frac{[\mathrm{P}_a(\boldsymbol{x}) - \mathrm{P}_s(\boldsymbol{x})][\mathrm{P}_b(\boldsymbol{x}) - \mathrm{P}_s(\boldsymbol{x})]}{\mathrm{P}^2(\boldsymbol{x})} - \pi_s^{-1}\right] \\ &= \sum_{\boldsymbol{x}\in\Omega}\frac{\mathrm{P}_a(\boldsymbol{x})\,\mathrm{P}_b(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} - \sum_{\boldsymbol{x}\in\Omega}\frac{\mathrm{P}_a(\boldsymbol{x})\,\mathrm{P}_s(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} - \sum_{\boldsymbol{x}\in\Omega}\frac{\mathrm{P}_b(\boldsymbol{x})\,\mathrm{P}_s(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} + \sum_{\boldsymbol{x}\in\Omega}\frac{\mathrm{P}_s^2(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} - \pi_s^{-1}.\end{aligned} \tag{2.14}$$

We can use Corollary 2.5 to handle the first three terms. For the last term, notice that

$$\sum_{\boldsymbol{x}\in\Omega}\frac{\mathrm{P}_s^2(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} - \frac{1}{\pi_s} = \sum_{\boldsymbol{x}\in\Omega}\left(\frac{\mathrm{P}_s(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} - \frac{1}{\pi_s}\right)\mathrm{P}_s(\boldsymbol{x}) = -\frac{1}{\pi_s}\sum_{\boldsymbol{x}\in\Omega}\sum_{\ell\neq s}\frac{\pi_\ell\,\mathrm{P}_\ell(\boldsymbol{x})\,\mathrm{P}_s(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})}.$$

Now, applying the triangle inequality to (2.14),

$$|\varepsilon_{ab}| \leq \frac{2}{\pi_a\pi_b}e^{-\frac{m}{2}\delta_{ab}^2} + \frac{2}{\pi_a\pi_s}e^{-\frac{m}{2}\delta_{as}^2} + \frac{2}{\pi_b\pi_s}e^{-\frac{m}{2}\delta_{bs}^2} + \frac{2}{\pi_s^2}\sum_{\ell\neq s}e^{-\frac{m}{2}\delta_{\ell s}^2}$$

Since $\delta_{ij}^2 > 0$ for $i \neq j$, we have $\varepsilon_{ab} \to 0$ for $a \neq b$ in $\{1, \dots, s-1\}$ as $m \to \infty$.

Case (v) Finally, consider the following matrix, for $j = 1, \dots, s$,

$$\begin{aligned}\boldsymbol{A}_{\pi j}\boldsymbol{D}_j &= \mathrm{E}\left[\left\{\frac{\partial}{\partial\boldsymbol{\pi}}\log P(\boldsymbol{x})\right\}\left\{\frac{\partial}{\partial\boldsymbol{p}_j}\log P(\boldsymbol{x})\right\}^T\right]\boldsymbol{D}_j \\ &= \mathrm{E}\left[\frac{1}{P(\boldsymbol{x})}\begin{pmatrix}P_1(\boldsymbol{x}) - P_s(\boldsymbol{x}) \\ \vdots \\ P_{s-1}(\boldsymbol{x}) - P_s(\boldsymbol{x})\end{pmatrix}\frac{\pi_j P_j(\boldsymbol{x})}{P(\boldsymbol{x})}\left(\boldsymbol{D}_j\boldsymbol{x}_{-k} - \frac{x_k}{p_k}\mathbf{1}\right)^T\right]\boldsymbol{D}_j \\ &= \mathrm{E}\left[\frac{\pi_j P_j(\boldsymbol{x})}{P^2(\boldsymbol{x})}\begin{pmatrix}P_1(\boldsymbol{x}) - P_s(\boldsymbol{x}) \\ \vdots \\ P_{s-1}(\boldsymbol{x}) - P_s(\boldsymbol{x})\end{pmatrix}\left(\boldsymbol{x}_{-k} - \frac{x_k}{p_k}\boldsymbol{p}_j\right)^T\right]\end{aligned}$$

whose (a,b)th element is

$$\begin{aligned}\varepsilon_{ab} &= \mathrm{E}\left[\frac{\pi_j P_j(\boldsymbol{x})}{P^2(\boldsymbol{x})}\left(P_a(\boldsymbol{x}) - P_s(\boldsymbol{x})\right)\left(x_b - \frac{x_k}{p_{jk}}p_{jb}\right)\right] \\ &= \sum_{\boldsymbol{x}\in\Omega}\frac{\pi_j P_j(\boldsymbol{x})}{P(\boldsymbol{x})}\left(P_a(\boldsymbol{x}) - P_s(\boldsymbol{x})\right)\left(x_b - \frac{x_k}{p_{jk}}p_{jb}\right). \end{aligned} \tag{2.15}$$

First suppose that $j \neq a$ and $j \neq s$. Since $|t_b p_{jk} - t_k p_{jb}| \leq t_b p_{jk} + t_k p_{jb} \leq 2m$ we have

$$\begin{aligned}|\varepsilon_{ab}| &\leq \frac{2m}{p_{jk}}\sum_{\boldsymbol{x}\in\Omega}\frac{\pi_j P_j(\boldsymbol{x})}{P(\boldsymbol{x})}\left|P_a(\boldsymbol{x}) - P_s(\boldsymbol{x})\right| \\ &\leq \frac{2m}{p_{jk}}\left\{\sum_{\boldsymbol{x}\in\Omega}\frac{\pi_j P_j(\boldsymbol{x})P_a(\boldsymbol{x})}{P(\boldsymbol{x})} + \sum_{\boldsymbol{x}\in\Omega}\frac{\pi_j P_j(\boldsymbol{x})P_s(\boldsymbol{x})}{P(\boldsymbol{x})}\right\} \\ &\leq \frac{2m}{p_{jk}}\left\{\frac{2}{\pi_a}e^{-\frac{m}{2}\delta_{ja}^2} + \frac{2}{\pi_s}e^{-\frac{m}{2}\delta_{js}^2}\right\},\end{aligned}$$

using Corollary 2.5. Since $\delta_{ja}^2 > 0$ and $\delta_{js}^2 > 0$, we have $\varepsilon_{ab} \to \infty$ as $m \to \infty$.

Now suppose $j = a$ or $j = s$, and notice that

$$\sum_{\boldsymbol{x}\in\Omega}\left(x_b - \frac{x_k}{p_{jk}}p_{jb}\right)P_j(\boldsymbol{x}) = \mathrm{E}\left(X_b - X_k\frac{p_{jb}}{p_{jk}}\,\middle|\, Z = j\right) = 0.$$

Therefore, the expression for ε_{ab} in (2.15) is equivalent to

$$\begin{aligned}\varepsilon_{ab} &= \sum_{\boldsymbol{x}\in\Omega}\left[\frac{\pi_j P_j(\boldsymbol{x})}{P(\boldsymbol{x})}\left(P_a(\boldsymbol{x}) - P_s(\boldsymbol{x})\right) + 2P_j(\boldsymbol{x})\right]\left(x_b - \frac{x_k}{p_{jk}}p_{jb}\right) \\ &= \sum_{\boldsymbol{x}\in\Omega}\pi_j P_j(\boldsymbol{x})\left[\frac{P_a(\boldsymbol{x}) - P_s(\boldsymbol{x})}{P(\boldsymbol{x})} + 2\pi_j^{-1}\right]\left(x_b - \frac{x_k}{p_{jk}}p_{jb}\right),\end{aligned}$$

and so

$$\begin{aligned}
\varepsilon_{ab} &\leq \frac{2m}{p_{jk}} \sum_{\boldsymbol{x}\in\Omega} \pi_j P_j(\boldsymbol{x}) \left[\frac{P_a(\boldsymbol{x}) - P_s(\boldsymbol{x})}{P(\boldsymbol{x})} + 2\pi_j^{-1}\right] \\
&= \frac{2m}{p_{jk}} \left\{ \sum_{\boldsymbol{x}\in\Omega} \frac{\pi_j P_j(\boldsymbol{x}) P_a(\boldsymbol{x})}{P(\boldsymbol{x})} - \sum_{\boldsymbol{x}\in\Omega} \frac{\pi_j P_j(\boldsymbol{x}) P_s(\boldsymbol{x})}{P(\boldsymbol{x})} + 2\pi_j^{-1}\right\} \\
&\leq \frac{2m}{p_{jk}} \left\{ \frac{2}{\pi_a} e^{-\frac{m}{2}\delta_{ja}^2} - \frac{2}{\pi_s} e^{-\frac{m}{2}\delta_{js}^2} + 2\pi_j^{-1}\right\} \\
&= \begin{cases} \frac{2m}{p_{jk}}\frac{2}{\pi_a}\exp\{-\frac{m}{2}\delta_{ja}^2\}, & \text{if } j = s \\ \frac{2m}{p_{jk}}\frac{2}{\pi_s}\exp\{-\frac{m}{2}\delta_{js}^2\}, & \text{if } j = a, \end{cases}
\end{aligned}$$

applying Corollary 2.5 on the second-to-last line. Similarly,

$$\varepsilon_{ab} \geq -\frac{2m}{p_{jk}} \left\{ \frac{2}{\pi_a} e^{-\frac{m}{2}\delta_{ja}^2} - \frac{2}{\pi_s} e^{-\frac{m}{2}\delta_{js}^2} + 2\pi_j^{-1}\right\} = \begin{cases} -\frac{2m}{p_{jk}}\frac{2}{\pi_a}\exp\{-\frac{m}{2}\delta_{ja}^2\}, & \text{if } j = s \\ -\frac{2m}{p_{jk}}\frac{2}{\pi_s}\exp\{-\frac{m}{2}\delta_{js}^2\}, & \text{if } j = a. \end{cases}$$

Therefore for both cases, $j = a$ and $j = s$, we have that $\varepsilon_{ab} \to 0$ as $m \to \infty$. $\square$

As an important corollary from the proof of Theorem 2.2, the convergence rate of the elements of $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})$ to zero is of exponential order, but depends on the closeness between subpopulations. The convergence can be slowed dramatically when $\boldsymbol{p}_a$ and $\boldsymbol{p}_b$ are close together for mixture components $a \neq b$.

Corollary 2.6 (Convergence rates)**.** *We have the following rates of convergence for the elements of* $\mathcal{I}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}(\boldsymbol{\theta})$ *as* $m \to \infty$. *Let* $\delta_* = \bigwedge_{a\neq b} \delta_{ab}^2 = \bigwedge_{a\neq b} \left[\bigvee_{j=1}^{k-1} p_{aj} - p_{bj}\right]$.

(i) For elements ε_{ab} *of the diagonal blocks* $\boldsymbol{A}_{ii} - \pi_i \boldsymbol{F}_i$, $i = 1, \ldots, s$, $\varepsilon_{ab} = O(m^2 e^{-\frac{m}{2}\delta_*^2})$

(ii) For elements ε_{ab} *of the off-diagonal blocks* $\boldsymbol{A}_{ij}$, $i, j \in \{1, \ldots, s\}$ *and* $i \neq j$, $\varepsilon_{ab} = O(m^2 e^{-\frac{m}{2}\delta_*^2})$

(iii) For diagonal elements ε_{aa} *of the last diagonal block* $\boldsymbol{A}_{\pi\pi} - \boldsymbol{F}_\pi$, $\varepsilon_{aa} = O(e^{-\frac{m}{2}\delta_*^2})$

(iv) For off-diagonal elements ε_{ab}, $a \neq b$, of the last diagonal block $\boldsymbol{A}_{\pi\pi} - \boldsymbol{F}_{\pi}$, $\varepsilon_{ab} = O(e^{-\frac{m}{2}\delta_*^2})$

(v) For elements ε_{ab} of the off-diagonal blocks $\boldsymbol{A}_{\pi j}$, $j = 1, \ldots, s$, $\varepsilon_{ab} = O(me^{-\frac{m}{2}\delta_*^2})$

where $i, j \in \{1, \ldots, s\}$.

The FIM approximation turns out to be equivalent to a complete data FIM, as shown below in Proposition 2.7, which gives an interesting connection to EM. The matrix $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ can therefore be formulated for any finite mixture whose components have a well-defined FIM, and is not limited to the case of multinomials. Denote $\text{Discrete}(a_1, \ldots, a_s; \boldsymbol{\pi})$ as the discrete distribution taking values $a_1, \ldots, a_s$ with corresponding probabilities $\pi_1, \ldots, \pi_s$.

Proposition 2.7. *The matrix $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ is equivalent to the FIM of $(\boldsymbol{X}, Z)$, where $(\boldsymbol{X} \mid Z = \ell) \sim$ Mult$_k(\boldsymbol{p}_\ell, m)$ and $Z \sim$ Discrete$(1, \ldots, s; \boldsymbol{\pi})$.*

Proof of Proposition 2.7. Here Z represents the population from which $\boldsymbol{X}$ was drawn. The complete data likelihood is then

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}, z) = \prod_{\ell=1}^{s} \left[\pi_\ell f(\boldsymbol{x} \mid \boldsymbol{p}_\ell, m)\right]^{I(z=\ell)}.$$

This likelihood leads to the score vectors

$$\frac{\partial}{\partial \boldsymbol{p}_a} \log L(\boldsymbol{\theta}) = \Delta_a \left[\boldsymbol{D}_a^{-1} \boldsymbol{x}_{-k} - \frac{x_k}{p_{ak}} \boldsymbol{1}\right],$$
$$\frac{\partial}{\partial \boldsymbol{\pi}} \log L(\boldsymbol{\theta}) = \boldsymbol{D}_\pi^{-1} \boldsymbol{\Delta}_{-s} - \frac{\Delta_s}{\pi_s} \boldsymbol{1},$$

where $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_s)$ so that $\Delta_\ell = I(Z = \ell)$ and $\boldsymbol{\Delta} \sim \text{Mult}_s(1, \boldsymbol{\pi})$, and $\boldsymbol{\Delta}_{-s}$ denotes the vector $(\Delta_1, \dots, \Delta_{s-1})$. Taking second derivatives yields

$$\begin{aligned}
\frac{\partial^2}{\partial \boldsymbol{p}_a \partial \boldsymbol{p}_a^T} \log L(\boldsymbol{\theta}) &= -\Delta_a \left[\boldsymbol{D}_a^{-2} \boldsymbol{x}_{-k} + \frac{x_k}{p_{ak}^2} \boldsymbol{1}\boldsymbol{1}^T \right], \\
\frac{\partial^2}{\partial \boldsymbol{p}_a \partial \boldsymbol{p}_b^T} \log L(\boldsymbol{\theta}) &= 0, \qquad \text{for } a \neq b, \\
\frac{\partial^2}{\partial \boldsymbol{p}_a \partial \boldsymbol{\pi}^T} \log L(\boldsymbol{\theta}) &= 0, \\
\frac{\partial^2}{\partial \boldsymbol{\pi} \partial \boldsymbol{\pi}^T} \log L(\boldsymbol{\theta}) &= -\left[\boldsymbol{D}_\pi^{-2} \boldsymbol{\Delta}_{-s} + \frac{\Delta_s}{\pi_s^2} \boldsymbol{1}\boldsymbol{1}^T \right].
\end{aligned}$$

Now take the expected value of the negative of each of these terms, jointly with respect to $(\boldsymbol{X}, Z)$, to obtain the blocks of $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$. □

Corollary 2.8. *Suppose $\boldsymbol{X}_i \sim$ MultMix$(m_i, \boldsymbol{\theta})$, $i = 1, \dots, n$, is an independent sample from the mixed population with varying cluster sizes, and $M = m_1 + \cdots + m_n$. Then the approximate FIM with respect to $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$ is given by*

$$\widetilde{\mathcal{I}}(\boldsymbol{\theta}) = \text{Blockdiag}\left(\pi_1 \boldsymbol{F}_1, \dots, \pi_s \boldsymbol{F}_s, \boldsymbol{F}_\pi\right),$$

where $\boldsymbol{F}_\ell = M\left[\boldsymbol{D}_\ell^{-1} + p_{\ell k}^{-1} \boldsymbol{1}\boldsymbol{1}^T\right]$ for $\ell = 1, \dots, s$, and $\boldsymbol{F}_\pi = n\left[\boldsymbol{D}_\pi^{-1} + \pi_s^{-1} \boldsymbol{1}\boldsymbol{1}^T\right]$.

Proof of Corollary 2.8. Let $\widetilde{\mathcal{I}}_{m_i}(\boldsymbol{\theta})$ represent the FIM approximation with respect to observation $\boldsymbol{X}_i$. The result is obtained using $\widetilde{\mathcal{I}}(\boldsymbol{\theta}) = \widetilde{\mathcal{I}}_{m_1}(\boldsymbol{\theta}) + \cdots + \widetilde{\mathcal{I}}_{m_n}(\boldsymbol{\theta})$, corresponding to the additive property of exact FIMs for independent samples. This additive property can be justified by noting that each $\widetilde{\mathcal{I}}_{m_i}(\boldsymbol{\theta})$ is a true (complete data) FIM, by Proposition 2.7. □

Since $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ is a block-diagonal matrix, some useful expressions can be obtained in closed-form.

Corollary 2.9. *Let $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ represent the FIM with respect to an independent sample $\boldsymbol{X}_i \sim$ MultMix$(m_i, \boldsymbol{\theta})$, $i = 1, \dots, n$. Then:*

(a) $\widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta}) = \text{Blockdiag}\left(\pi_1^{-1}\boldsymbol{F}_1^{-1}, \ldots, \pi_s^{-1}\boldsymbol{F}_s^{-1}, \boldsymbol{F}_\pi^{-1}\right)$*, where* $\boldsymbol{F}_\ell^{-1} = M^{-1}\{\boldsymbol{D}_\ell - \boldsymbol{p}_\ell\boldsymbol{p}_\ell^T\}$ *for* $\ell = 1, \ldots, s$ *and* $\boldsymbol{F}_\pi^{-1} = n^{-1}\{\boldsymbol{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}^T\}$.

(b) $\text{tr}\left(\widetilde{\mathcal{I}}(\boldsymbol{\theta})\right) = \sum_{\ell=1}^{s}\sum_{j=1}^{k-1} M\pi_\ell\left\{p_{\ell j}^{-1} + p_{\ell k}^{-1}\right\} + \sum_{\ell=1}^{s-1} n\left\{\pi_\ell^{-1} + \pi_s^{-1}\right\}$.

(c) $\det\left(\widetilde{\mathcal{I}}(\boldsymbol{\theta})\right) = \left(\prod_{\ell=1}^{s} p_{\ell k}^{-1}\prod_{j=1}^{k-1} M\pi_\ell p_{\ell j}^{-1}\right)\left(\pi_s^{-1}\prod_{\ell=1}^{s-1} n\pi_\ell^{-1}\right)$.

Proof of Corollary 2.9. **(a)** Since $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ is block diagonal, its inverse can be obtained by inverting the blocks. To find the expressions for the individual blocks, we can apply the Sherman-Morrison formula (see for example Rao (1965, chapter 1))

$$(\boldsymbol{C} + \boldsymbol{u}\boldsymbol{v}^T)^{-1} = \boldsymbol{C}^{-1} - \frac{\boldsymbol{C}^{-1}\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{C}^{-1}}{1 + \boldsymbol{v}^T\boldsymbol{C}^{-1}\boldsymbol{u}}.$$

For the case of $\boldsymbol{F}_\pi^{-1}$, for example, take $\boldsymbol{C} = \boldsymbol{D}_\pi^{-1}$, $\boldsymbol{u} = \pi_s^{-1/2}\mathbf{1}$, and $\boldsymbol{v} = \pi_s^{-1/2}\mathbf{1}^T$ and use the expressions in Corollary 2.8.

(b) Since the trace of a block diagonal matrix is the sum of the traces of its blocks, we have

$$\text{tr}\left(\widetilde{\mathcal{I}}(\boldsymbol{\theta})\right) = \pi_1\,\text{tr}\left(\boldsymbol{F}_1\right) + \cdots + \pi_s\,\text{tr}\left(\boldsymbol{F}_s\right) + \text{tr}\left(\boldsymbol{F}_\pi\right). \tag{2.16}$$

The individual traces can be obtained as

$$\text{tr}\left(\boldsymbol{F}_\ell\right) = \text{tr}\left[M(\boldsymbol{D}_\ell^{-1} + p_{\ell k}^{-1}\mathbf{1}\mathbf{1}^T)\right] = \sum_{j=1}^{k-1} M\left\{p_{\ell j}^{-1} + p_{\ell k}^{-1}\right\},$$

a summation over the diagonal elements. Similarly for the block corresponding to $\boldsymbol{\pi}$,

$$\text{tr}\left(\boldsymbol{F}_\pi\right) = \text{tr}\left[n\left(\boldsymbol{D}_\pi^{-1} + \pi_s^{-1}\mathbf{1}\mathbf{1}^T\right)\right] = \sum_{\ell=1}^{s-1} n\left\{\pi_\ell^{-1} + \pi_s^{-1}\right\}.$$

The result is obtained by replacing these expressions into (2.16).

(c) Since $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ has a block diagonal structure,

$$\begin{aligned}\det \widetilde{\mathcal{I}}(\boldsymbol{\theta}) &= \det\left\{\boldsymbol{F}_{\pi}\right\} \times \prod_{\ell=1}^{s} \det\left\{\pi_{\ell} \boldsymbol{F}_{\ell}\right\} \\ &= \left(n^{s-1} \det\left\{\boldsymbol{D}_{\pi}^{-1} + \pi_s^{-1} \mathbf{1}\mathbf{1}^T\right\}\right)\left(\prod_{\ell=1}^{s} \pi_{\ell}^{k-1} M^{k-1} \det\left\{\boldsymbol{D}_{\ell}^{-1} + \boldsymbol{p}_{\ell k}^{-1} \mathbf{1}\mathbf{1}^T\right\}\right)\end{aligned} \tag{2.17}$$

Recall the property (see for example Rao (1965, chapter 1)) that for $\boldsymbol{M}$ nonsingular, we have

$$\det(\boldsymbol{M} + \boldsymbol{u}\boldsymbol{u}^T) = \begin{vmatrix} \boldsymbol{M} & -\boldsymbol{u} \\ \boldsymbol{u}^T & 1 \end{vmatrix} = \det(\boldsymbol{M})\left(1 + \boldsymbol{u}^T \boldsymbol{M}^{-1} \boldsymbol{u}\right).$$

This yields, for instance

$$\begin{aligned}\det\left\{\boldsymbol{D}_{\pi}^{-1} + \pi_s^{-1} \mathbf{1}\mathbf{1}^T\right\} &= \det\left\{\boldsymbol{D}_{\pi}^{-1}\right\}\left(1 + \pi_s^{-1} \mathbf{1}^T \boldsymbol{D}_{\pi} \mathbf{1}\right) \\ &= \left[1 + \frac{1 - \pi_s}{\pi_s}\right] \prod_{\ell=1}^{s-1} \pi_{\ell}^{-1} = \pi_s^{-1} \prod_{\ell=1}^{s-1} \pi_{\ell}^{-1}.\end{aligned}$$

The result can be obtained by substituting the simplified determinants into (2.17). □

The determinant and trace of the FIM are not utilized in the computation of MLEs, but are used in the computation of many statistics in subsequent analysis. In such applications, it may be useful to have a closed-form approximation for these expressions. As one example, consider the Consistent Akaike Information Criterion with Fisher Information (CAICF) formulated in (Bozdogan, 1987). The CAICF is an information-theoretic criterion for model selection, and is a function of the log-determinant of the FIM.

It can also be shown that $\mathcal{I}_m^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta}) \to \mathbf{0}$ as $m \to \infty$, which we now state as a theorem. This result is perhaps more immediately relevant than Theorem 2.2 for the scoring technique presented in the following section.

Theorem 2.10. *Let $\mathcal{I}_m(\boldsymbol{\theta})$ and $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta})$ be defined as in Theorem 2.2 (namely the FIM* *and its approximation, with respect to a single observation with cluster size m). Then $\mathcal{I}_m^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta}) \to \mathbf{0}$ as $m \to \infty$.*

Proof of Theorem 2.10. This proof uses properties of matrix norms; refer to Lange (2010, Chapter 6) or Meyer (2001, Chapter 5) for background. Notice that for nonsingular $q \times q$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,

$$\boldsymbol{B}^{-1} - \boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{B}^{-1}.$$

Then for any matrix norm satisfying the sub-multiplicative property,

$$\|\boldsymbol{A}^{-1} - \boldsymbol{B}^{-1}\| \leq \|\boldsymbol{A}^{-1}\| \cdot \|\boldsymbol{A} - \boldsymbol{B}\| \cdot \|\boldsymbol{B}^{-1}\|. \tag{2.18}$$

Fix $\boldsymbol{\theta} \in \Theta$, take $\boldsymbol{A} = \widetilde{\mathcal{I}}_m(\boldsymbol{\theta})$ and $\boldsymbol{B} = \mathcal{I}_m(\boldsymbol{\theta})$, and take $\|\cdot\|$ to be the Frobenius matrix norm. Then (2.18) becomes

$$\|\mathcal{I}_m^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta})\|_F \leq \|\mathcal{I}_m^{-1}(\boldsymbol{\theta})\|_F \cdot \|\widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta})\|_F \cdot \|\mathcal{I}_m(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m(\boldsymbol{\theta})\|_F,$$

where $\|\boldsymbol{A}\|_F^2 = \sum_{i=1}^q \sum_{j=1}^q a_{ij}^2$, and a_{ij} denote the elements of $\boldsymbol{A}$. To show that the RHS converges to 0 as $m \to \infty$, we will handle the three terms separately. Since $\mathcal{I}_m(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) \to \mathbf{0}$ as $m \to \infty$ by Theorem 2.2, $\|\mathcal{I}_m(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m(\boldsymbol{\theta})\|_F \to 0$. Next, we address the $\|\widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta})\|_F$ term. Using the explicit form in Corollary 2.9, we have

$$\begin{aligned} 0 \leq \|\widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta})\|_F^2 &= \sum_{\ell=1}^s \|\pi_\ell^{-1}\boldsymbol{F}_\ell^{-1}\|_F^2 + \|\boldsymbol{F}_\pi^{-1}\|_F^2 \\ &= \sum_{\ell=1}^s m^{-2}\pi_\ell^{-2}\|\boldsymbol{D}_\ell - \boldsymbol{p}_\ell\boldsymbol{p}_\ell^T\|_F^2 + \|\boldsymbol{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}^T\|_F^2. \end{aligned}$$

All terms beside m^{-2} are free of m, therefore $\|\widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta})\|_F$ is seen to be decreasing in m, and hence is bounded in m.

We will now consider the term $\|\mathcal{I}_m^{-1}(\boldsymbol{\theta})\|$, with the 2-norm instead of the Frobenius norm. Let $\lambda_1(m) \geq \cdots \geq \lambda_q(m)$ be the eigenvalues of $\mathcal{I}_m(\boldsymbol{\theta})$ for a fixed m, all assumed to be positive. Since the 2-norm of a symmetric positive definite matrix is its largest eigenvalue, we have

$$\begin{aligned}0 \leq \|\mathcal{I}_m^{-1}(\boldsymbol{\theta})\|_2 = \frac{1}{\lambda_q(m)} &= \frac{1}{\min\limits_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T \mathcal{I}_m(\boldsymbol{\theta})\boldsymbol{x}} \\ &= \frac{1}{\min\limits_{\|\boldsymbol{x}\|=1} \left\{\boldsymbol{x}^T \left[\mathcal{I}_m(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m(\boldsymbol{\theta})\right] \boldsymbol{x} + \boldsymbol{x}^T \widetilde{\mathcal{I}}_m(\boldsymbol{\theta})\boldsymbol{x}\right\}}.\end{aligned}$$

Notice that

$$\begin{aligned}&\min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T \left[\mathcal{I}_m(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m(\boldsymbol{\theta})\right] \boldsymbol{x} + \min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T \widetilde{\mathcal{I}}_m(\boldsymbol{\theta})\boldsymbol{x} \\ &\qquad \leq \min_{\|\boldsymbol{x}\|=1} \left\{\boldsymbol{x}^T \left[\mathcal{I}_m(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m(\boldsymbol{\theta})\right] \boldsymbol{x} + \boldsymbol{x}^T \widetilde{\mathcal{I}}_m(\boldsymbol{\theta})\boldsymbol{x}\right\}\end{aligned}$$

since both LHS and RHS are lower bounds for $\boldsymbol{x}^T \left[\mathcal{I}_m(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m(\boldsymbol{\theta})\right] \boldsymbol{x} + \boldsymbol{x}^T \widetilde{\mathcal{I}}_m(\boldsymbol{\theta})\boldsymbol{x}$, and the RHS is the greatest such bound. Therefore

$$\begin{aligned}1/\lambda_q(m) &\leq \frac{1}{\min\limits_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T \left[\mathcal{I}_m(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m(\boldsymbol{\theta})\right] \boldsymbol{x} + \min\limits_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T \widetilde{\mathcal{I}}_m(\boldsymbol{\theta})\boldsymbol{x}} \\ &= \frac{1}{\beta_q(m) + \widetilde{\lambda}_q(m),}\end{aligned}$$

denoting the eigenvalues of $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta})$ as $\widetilde{\lambda}_1(m) \geq \cdots \geq \widetilde{\lambda}_q(m)$ (all positive), and the eigenvalues of $\mathcal{I}_m(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m(\boldsymbol{\theta})$ as $\beta_1(m) \geq \cdots \geq \beta_q(m)$. It is well known that the mapping from a matrix to its eigenvalues is a continuous function of its elements (Meyer, 2001, Chapter 7). Therefore

$$\mathcal{I}_m(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) \to \mathbf{0} \text{ as } m \to \infty \quad \Longrightarrow \quad \beta_q(m) \to 0 \text{ as } m \to \infty.$$

Now for any $\varepsilon > 0$, there exists a positive integer m_0 such that $|\beta_q(m)| < \varepsilon$ for all

$m \geq m_0$, and so we have

$$0 \leq \|\mathcal{I}_m^{-1}(\boldsymbol{\theta})\|_2 \leq \frac{1}{\beta_q(m) + \widetilde{\lambda}_q(m)} \leq \frac{1}{\widetilde{\lambda}_q(m) - \varepsilon} \tag{2.19}$$

for all $m \geq m_0$. Because $\|\boldsymbol{A}\|_2 \leq \|\boldsymbol{A}\|_F$, and $\|\widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta})\|$ was seen to be bounded, for all m there exists a $K > 0$ such that,

$$1/\widetilde{\lambda}_q(m) = \|\widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta})\|_2 \leq \|\widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta})\|_F \leq K \quad \iff \quad \widetilde{\lambda}_q(m) \geq 1/K.$$

WLOG assume that ε has been chosen so that $\widetilde{\lambda}_q(m) \geq 1/K > \varepsilon$, to avoid division by zero. The RHS of (2.19) is therefore bounded above by $(1/K - \varepsilon)^{-1}$ for all $m \geq m_0$, which implies $\|\mathcal{I}_m^{-1}(\boldsymbol{\theta})\|_2$ is bounded when $m \geq m_0$.

To conclude the proof, note that in general $q^{-1/2}\|\boldsymbol{A}\|_F \leq \|\boldsymbol{A}\|_2$, so that

$$\begin{aligned}
0 &\leq \|\mathcal{I}_m^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta})\|_F \\
&\leq \|\mathcal{I}_m^{-1}(\boldsymbol{\theta})\|_F \cdot \|\widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta})\|_F \cdot \|\mathcal{I}_m(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m(\boldsymbol{\theta})\|_F \\
&\leq \sqrt{q}\|\mathcal{I}_m^{-1}(\boldsymbol{\theta})\|_2 \cdot \|\widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta})\|_F \cdot \|\mathcal{I}_m(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m(\boldsymbol{\theta})\|_F.
\end{aligned}$$

It follows from the earlier steps that the RHS converges to zero as $m \to \infty$, and therefore $\|\mathcal{I}_m^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta})\|_F \to 0$, which implies $\mathcal{I}_m^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta}) \to \boldsymbol{0}$.

□

In the next section, we use the FIM approximation obtained in Theorem 2.2 to define an approximate scoring algorithm and investigate its properties.

2.4 Approximate Scoring Algorithm

Consider an independent sample with varying cluster sizes $\boldsymbol{X}_i \sim \text{MultMix}_k(m_i, \boldsymbol{\theta})$ for $i = 1, \ldots, n$. Let $\boldsymbol{\theta}^{(0)}$ be an initial guess for $\boldsymbol{\theta}$, and $S(\boldsymbol{\theta})$ be the score vector with respect to the sample. Then by independence, $S(\boldsymbol{\theta}) = \sum_{i=1}^{n} S(\boldsymbol{\theta}; \boldsymbol{x}_i)$, where $S(\boldsymbol{\theta}; \boldsymbol{x}_i)$ is the score vector with respect to the ith observation. The exact scoring algorithm is given by computing the iterations

$$\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)} + \mathcal{I}^{-1}(\boldsymbol{\theta}^{(g)})\, S(\boldsymbol{\theta}^{(g)}), \quad g = 1, 2, \ldots \tag{2.20}$$

until the convergence criteria

$$\left| \log L(\boldsymbol{\theta}^{(g+1)}) - \log L(\boldsymbol{\theta}^{(g)}) \right| < \varepsilon$$

is met, for some given tolerance $\varepsilon > 0$. In practice, a line search may be used for every iteration after determining a search direction, and other convergence criteria may be considered, but such modifications will not be considered here. Note that (2.20) uses the exact FIM which may not be easily computable. We propose to substitute the approximation $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ for $\mathcal{I}(\boldsymbol{\theta})$, and will refer to the resulting method as the approximate scoring algorithm (AFSA). The expressions for $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ and its inverse are available in closed-form, as seen in Corollaries 2.8 and 2.9.

AFSA can be applied to finite mixture of multinomial models which are not explicitly in the form of (2.3). We now give two such examples in which AFSA may be used to compute MLEs.

Example 2.11. The random-clumped multinomial (RCM) model (Morel and Nagaraj, 1993) is a special case of the finite mixture of multinomials where the mixing proportions $\boldsymbol{\pi}$ and the component probability vectors $\boldsymbol{p}_\ell$, for $\ell = 1, \ldots, s$, are functions of a smaller set of parameters $\boldsymbol{\eta}$. The Jacobian of this transformation can be used to write AFSA iterations in terms of $\boldsymbol{\eta}$. This example has also been discussed in the context of AFSA in (Liu, 2005). Before describing the iterations we will recall some details about the distribution.

The RCM distribution models overdispersion through "clumped" sampling in the multinomial framework. RCM represents an interesting model for exploring computational methods. Recently Zhou and Lange (2010) have used it as an illustrative example for the minorization-maximization principle. Raim et al. (2013) have explored parallel computing in maximum likelihood estimation using large RCM models as a test problem. It turns out that RCM conforms to the finite mixture of multinomials representation (2.2), and can therefore be fitted by AFSA. Once the mixture representation is established, the score vector and FIM approximation can be formulated by the use of transformations; see for example Lehmann and Casella (1998, Section 2.6). Hence, we can obtain the algorithm presented in Morel and Nagaraj (1993) and Neerchal and Morel (1998) as an AFSA-type algorithm.

Consider a cluster of m trials, where each trial results in one of k possible outcomes with probabilities $\pi_1, \dots, \pi_k$. Suppose a default category is also selected at random, so that each trial either results in this default outcome with probability ρ, or an independent choice with probability $1-\rho$. Intuitively, if $\rho \to 0$, this scheme approaches a standard multinomial distribution. An RCM random variable can be obtained from the following procedure. Let $\boldsymbol{Y}_0, \boldsymbol{Y}_1, \dots, \boldsymbol{Y}_m \overset{\text{iid}}{\sim} \text{Mult}_k(\boldsymbol{\pi}, 1)$ and $\boldsymbol{U}_1, \dots, \boldsymbol{U}_m \overset{\text{iid}}{\sim} \text{Uniform}(0,1)$ be independent samples, then

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{Y}_0 \sum_{i=1}^{m} I(\boldsymbol{U}_i \leq \rho) + \sum_{i=1}^{m} \boldsymbol{Y}_i I(\boldsymbol{U}_i > \rho) \\ &= \boldsymbol{Y}_0 N + (\boldsymbol{Z} \mid N) \end{aligned} \tag{2.21}$$

follows the distribution $\text{RCM}_k(\boldsymbol{\pi}, \rho)$. The representation (2.21) emphasizes that $N \sim \text{Binomial}(m, \rho)$, $(\boldsymbol{Z} \mid N) \sim \text{Mult}_k(\boldsymbol{\pi}, m - N)$, and $\boldsymbol{Y}_0 \sim \text{Mult}_k(\boldsymbol{\pi}, 1)$, where N and $\boldsymbol{Y}_0$ are independent.

RCM is also a special case of the finite mixture of multinomials, so that

$$\begin{aligned}
&\boldsymbol{X} \sim f(\boldsymbol{x}; \boldsymbol{\pi}, \rho) = \sum_{\ell=1}^{k} \pi_\ell f(\boldsymbol{x};\, \boldsymbol{p}_\ell, m), \\
&\boldsymbol{p}_\ell = (1-\rho)\boldsymbol{\pi} + \rho \boldsymbol{e}_\ell, \quad \text{for } \ell = 1, \ldots, k-1, \\
&\boldsymbol{p}_k = (1-\rho)\boldsymbol{\pi},
\end{aligned}$$

where $f(\boldsymbol{x}; \boldsymbol{p}, m)$ is our usual notation for the density of $\text{Mult}_k(\boldsymbol{p}, m)$. This mixture representation can be derived using moment generating functions, as shown in (Morel and Nagaraj, 1993). Notice that in this mixture $s = k$, so that the number of mixture components matches the number of categories. There are also only k distinct parameters rather than $sk - 1$ as in the general mixture.

The FIM approximation for the RCM model can be obtained by transformation, starting with the expression for the general mixture. Consider transforming the k dimensional $\boldsymbol{\eta} = (\boldsymbol{\pi}, \rho)$ to the $q = sk - 1 = (k+1)(k-1)$ dimensional $\boldsymbol{\theta} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_s, \boldsymbol{\pi})$ so that

$$\boldsymbol{\theta}(\boldsymbol{\eta}) = \begin{pmatrix} (1-\rho)\boldsymbol{\pi} & + & \rho\boldsymbol{e}_1 \\ & \vdots & \\ (1-\rho)\boldsymbol{\pi} & + & \rho\boldsymbol{e}_{k-1} \\ (1-\rho)\boldsymbol{\pi} & & \\ \boldsymbol{\pi} & & \end{pmatrix}, \quad \text{yielding} \quad \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}} = \begin{pmatrix} (1-\rho)\boldsymbol{I}_{k-1} & -\boldsymbol{\pi} + \boldsymbol{e}_1 \\ \vdots & \vdots \\ (1-\rho)\boldsymbol{I}_{k-1} & -\boldsymbol{\pi} + \boldsymbol{e}_{k-1} \\ (1-\rho)\boldsymbol{I}_{k-1} & -\boldsymbol{\pi} \\ \boldsymbol{I}_{k-1} & \boldsymbol{0} \end{pmatrix}$$

as the $q \times k$ Jacobian of the transformation. Using the relations

$$\begin{aligned}
S(\boldsymbol{\eta}) &= \frac{\partial}{\partial \boldsymbol{\eta}} \log f(\boldsymbol{x}; \boldsymbol{\theta}) = \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}}\right)^T \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x}; \boldsymbol{\theta}), \\
\mathcal{I}(\boldsymbol{\eta}) &= \text{Var}\left(S(\boldsymbol{\eta})\right) = \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}}\right)^T \mathcal{I}(\boldsymbol{\theta}) \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}}\right),
\end{aligned}$$

it is possible to obtain an explicit form of the approximate FIM as stated in (Morel and

Nagaraj, 1993). The convergence $\widetilde{\mathcal{I}}(\boldsymbol{\eta}) - \mathcal{I}(\boldsymbol{\eta}) \to \mathbf{0}$ as $m \to \infty$ is proved explicitly in (Morel and Nagaraj, 1991). We then have AFSA iterations for RCM,

$$\boldsymbol{\eta}^{(g+1)} = \boldsymbol{\eta}^{(g)} + \widetilde{\mathcal{I}}^{-1}(\boldsymbol{\eta}^{(g)})\, S(\boldsymbol{\eta}^{(g)}), \quad g = 1, 2, \ldots$$

The following example involves a mixture of multinomials where the response probabilities are functions of covariates. The idea is analogous to the usual multinomial with logit link, but with links corresponding to each component of the mixture. Again, the Jacobian of a transformation can be used to formulate AFSA iterations.

Example 2.12. In practice a binomial or multinomial outcome is often studied as a response to a covariate. As an example showing how AFSA can be applied to such models, consider the finite mixture of binomial regressions model (Frühwirth-Schnatter, 2006, Chapter 9). Suppose Y follows the binomial mixture distribution $\text{MultMix}_2(m, \boldsymbol{\theta})$ so that

$$f(y \mid m, \boldsymbol{\theta}) = \sum_{\ell=1}^{s} \pi_\ell \binom{m}{y} p_\ell^y (1 - p_\ell)^{m-y}$$

with $\boldsymbol{\theta} = (p_1, \ldots, p_s, \pi_1, \ldots, \pi_{s-1})$, and a regression

$$p_\ell = G(\boldsymbol{x}^T \boldsymbol{\beta}_\ell), \quad \boldsymbol{x} \in \mathbb{R}^d,$$

is linked to each mixture component $\ell = 1, \ldots, s$ through the inverse link function $G(x) = 1/(1 + e^{-x})$. Denote $\boldsymbol{\vartheta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_s, \pi_1, \ldots, \pi_{s-1})$, the parameter of interest in studying the regression of Y on $\boldsymbol{x}$. AFSA requires the expressions

$$\frac{\partial}{\partial \boldsymbol{\vartheta}} \log f(y \mid m, \boldsymbol{\vartheta}) = \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\vartheta}} \right)^T \frac{\partial}{\partial \boldsymbol{\theta}} \log f(y \mid m, \boldsymbol{\theta}) \tag{2.22}$$

and

$$\widetilde{\mathcal{I}}(\boldsymbol{\vartheta}) = \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\vartheta}}\right)^T \widetilde{\mathcal{I}}(\boldsymbol{\theta}) \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\vartheta}}\right), \tag{2.23}$$

where the expressions for the score and FIM with respect to $\boldsymbol{\theta}$ are as given earlier. To obtain the Jacobian of the transformation $\boldsymbol{\vartheta} \mapsto \boldsymbol{\theta}$, we have

$$\frac{\partial p_a}{\partial \boldsymbol{\beta}_b} = \begin{cases} G'(\boldsymbol{x}^T\boldsymbol{\beta}_b)\boldsymbol{x}^T & \text{if } a = b \\ 0 & \text{o.w.} \end{cases}$$

This yields the $(2s-1) \times (sd+s-1)$ matrix

$$\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\vartheta}} = \left(\begin{array}{ccc|c} G'(\boldsymbol{x}^T\boldsymbol{\beta}_1)\boldsymbol{x}^T & & & \\ & \ddots & & \boldsymbol{0} \\ & & G'(\boldsymbol{x}^T\boldsymbol{\beta}_s)\boldsymbol{x}^T & \\ \hline & \boldsymbol{0} & & \boldsymbol{I}_{s-1} \end{array}\right),$$

where $\boldsymbol{I}_{s-1}$ is the identity matrix of dimension $s-1$. AFSA for an independent sample

$$\begin{aligned} &\boldsymbol{Y}_i \sim \text{MultMix}_2(m_i, \boldsymbol{\theta}_i), \quad \boldsymbol{\theta}_i = (p_{i1}, \ldots, p_{is}, \pi_1, \ldots, \pi_{s-1}) \\ &p_{i\ell} = G(\boldsymbol{x}_i^T\boldsymbol{\beta}_\ell), \quad \text{for } i = 1, \ldots, n \text{ and } \ell = 1, \ldots, s, \end{aligned}$$

can be written by summing expressions (2.22) and (2.23) for the score and FIM approximation over all observations.

Example 2.13. Example 2.12 can be extended to a more complicated regression model in the multinomial setting. The same transformation technique can be used to obtain the expressions needed for AFSA. Consider the response $\boldsymbol{Y} \sim \text{MultMix}_k(m, \boldsymbol{\theta})$ with covariates $\boldsymbol{x}$ and $\boldsymbol{w}$. For each $\boldsymbol{p}_\ell$ vector, $\ell = 1, \ldots, s$, assume a generalized logistic regression

$$\log \frac{p_{\ell j}}{p_{\ell k}} = \eta_{\ell j}, \quad \eta_{\ell j} = \boldsymbol{x}^T \boldsymbol{\beta}_{\ell j},$$

for $j = 1, \ldots, k-1$. For $\boldsymbol{\pi}$, assume a proportional odds model,

$$\log \frac{\pi_1 + \cdots + \pi_\ell}{\pi_{\ell+1} + \cdots + \pi_s} = \eta_\ell^\pi, \quad \eta_\ell^\pi = \nu_\ell + \boldsymbol{w}^T \boldsymbol{\alpha},$$

for $\ell = 1, \ldots, s-1$, taking $\eta_0^\pi := -\infty$ and $\eta_s^\pi := \infty$. The unknown parameters are the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_{\ell j}$, and the scalars $\nu_1 < \cdots < \nu_{s-1}$. Denote these parameters collectively as $\boldsymbol{\vartheta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_s, \boldsymbol{\nu}, \boldsymbol{\alpha})$ where $\boldsymbol{\beta}_\ell = (\boldsymbol{\beta}_{\ell 1}, \ldots, \boldsymbol{\beta}_{\ell, k-1})$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_s)$. Expressions for the $\boldsymbol{\theta}$ parameters can be obtained as

$$p_{\ell j} = \frac{e^{\eta_{\ell j}}}{1 + \sum_{b=1}^{k-1} e^{\eta_{\ell b}}} \quad \text{and} \quad \pi_\ell = \frac{1}{1 + e^{-\eta_\ell^\pi}} - \frac{1}{1 + e^{-\eta_{\ell-1}^\pi}},$$

for $\ell = 1, \ldots, s$ and $j = 1, \ldots, k-1$. To implement AFSA, a score vector and FIM approximation are needed. For the score vector we have

$$\frac{\partial}{\partial \boldsymbol{\vartheta}} \log f(\boldsymbol{y}; \boldsymbol{\theta}) = \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\vartheta}} \right)^T \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{y}; \boldsymbol{\theta}), \tag{2.24}$$

and the approximate FIM is given by

$$\widetilde{\mathcal{I}}(\boldsymbol{\vartheta}) = \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\vartheta}} \right)^T \widetilde{\mathcal{I}}(\boldsymbol{\theta}) \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\vartheta}} \right). \tag{2.25}$$

Finding and expression for the Jacobian is tedious but straightforward. AFSA for an independent sample

$$\boldsymbol{Y}_i \sim \text{MultMix}_k(m_i, \boldsymbol{\theta}_i), \quad \boldsymbol{\theta}_i = (\boldsymbol{p}_{i1}, \ldots, \boldsymbol{p}_{is}, \pi_1, \ldots, \pi_{s-1}),$$

with generalized logit and proportional odds models as above for $i = 1, \ldots, n$, can be written by summing the expressions (2.24) and (2.25) for the score and FIM approximation over all observations.

We have already seen that the FIM approximation is equivalent to a complete data FIM from EM. There is also an interesting connection between AFSA and EM, stated as Theorem 2.16, that the iterations are algebraically related. This was first observed by Liu (2005). Again, we reproduce the details on that issue in this thesis. To see the connection, explicit forms for AFSA and EM iterations are first presented in Propositions 2.14 and 2.15.

Proposition 2.14 (AFSA Iterations)**.** *The AFSA iterations*

$$\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)} + \widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta}^{(g)})\, S(\boldsymbol{\theta}^{(g)}), \quad g = 1, 2, \ldots \tag{2.26}$$

can be written explicitly as

$$\begin{aligned} \pi_\ell^{(g+1)} &= \pi_\ell^{(g)} \frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} \quad \textit{and} \\ p_{\ell j}^{(g+1)} &= \frac{1}{M} \sum_{i=1}^{n} \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} x_{ij} - p_{\ell j}^{(g)} \left[1 - \frac{1}{M} \sum_{i=1}^{n} m_i \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} \right], \end{aligned}$$

where $\ell = 1, \ldots, s$, $j = 1, \ldots, k$, and $M = m_1 + \cdots + m_n$.

Proof of Proposition 2.14. The general form for AFSA is given by

$$\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)} + \mathrm{Blockdiag}\left(\pi_1 \boldsymbol{F}_1^{-1} \ldots, \pi_s \boldsymbol{F}_s^{-1}, \boldsymbol{F}_\pi^{-1}\right) S(\boldsymbol{\theta}^{(g)})$$

so that the individual updates are

$$\begin{aligned} \boldsymbol{p}_\ell^{(g+1)} &= \boldsymbol{p}_\ell^{(g)} + \pi_\ell^{-1} \boldsymbol{F}_\ell^{-1} \frac{\partial}{\partial \boldsymbol{p}_\ell} \log L(\boldsymbol{\theta}^{(g)}), \qquad \ell = 1, \ldots, s \\ \boldsymbol{\pi}^{(g+1)} &= \boldsymbol{\pi}^{(g)} + \boldsymbol{F}_\pi^{-1} \frac{\partial}{\partial \boldsymbol{\pi}} \log L(\boldsymbol{\theta}^{(g)}). \end{aligned}$$

From Corollary 2.9 we have

$$\begin{aligned}
\boldsymbol{\pi}^{(g+1)} &= \boldsymbol{\pi}^{(g)} + (n\boldsymbol{F}_{\pi})^{-1} \sum_{i=1}^{n} \frac{\partial \log L(\boldsymbol{\theta}^{(g)} \mid \boldsymbol{x}_i)}{\partial \boldsymbol{\pi}} \\
&= \boldsymbol{\pi}^{(g)} + n^{-1} \left[\operatorname{Diag}\{\boldsymbol{\pi}^{(g)}\} - \boldsymbol{\pi}^{(g)} \boldsymbol{\pi}^{(g)T} \right] \sum_{i=1}^{n} \frac{\partial \log(\boldsymbol{\theta}^{(g)} \mid \boldsymbol{x}_i)}{\partial \boldsymbol{\pi}}.
\end{aligned}$$

Then for $\ell = 1, \ldots, s-1$,

$$\begin{aligned}
\pi_{\ell}^{(g+1)} &= \pi_{\ell}^{(g)} + n^{-1} \pi_{\ell}^{(g)} \sum_{i=1}^{n} \frac{\mathrm{P}_{\ell}(\boldsymbol{x}_i) - \mathrm{P}_{s}(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} - n^{-1} \sum_{i=1}^{n} \sum_{t=1}^{s-1} \pi_{\ell}^{(g)} \pi_{t}^{(g)} \frac{\mathrm{P}_{t}(\boldsymbol{x}_i) - \mathrm{P}_{s}(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} \\
&= \pi_{\ell}^{(g)} + n^{-1} \pi_{\ell}^{(g)} \sum_{i=1}^{n} \frac{\mathrm{P}_{\ell}(\boldsymbol{x}_i) - \mathrm{P}_{s}(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} \\
&\quad - n^{-1} \pi_{\ell}^{(g)} \sum_{i=1}^{n} \left\{ \frac{\mathrm{P}(\boldsymbol{x}_i) - \pi_{s}^{(g)} \mathrm{P}_{s}(\boldsymbol{x}_i) - (1 - \pi_{s}^{(g)}) \mathrm{P}_{s}(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} \right\} \\
&= \pi_{\ell}^{(g)} + n^{-1} \pi_{\ell}^{(g)} \sum_{i=1}^{n} \frac{\mathrm{P}_{\ell}(\boldsymbol{x}_i) - \mathrm{P}_{s}(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} - n^{-1} \pi_{\ell}^{(g)} \sum_{i=1}^{n} \left\{ 1 - \frac{\mathrm{P}_{s}(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} \right\} \\
&= \pi_{\ell}^{(g)} \frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{P}_{\ell}(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}.
\end{aligned}$$

Next, to obtain explicit iterations for $p_{\ell j}$'s, the blocks for $\ell = 1, \ldots, s$ are given by

$$\begin{aligned}
\boldsymbol{p}_{\ell}^{(g+1)} &= \boldsymbol{p}_{\ell}^{(g)} + \left(\pi_{\ell}^{(g)} \boldsymbol{F}_{\ell} \right)^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{p}_{\ell}} \log L(\boldsymbol{\theta}^{(g)} \mid \boldsymbol{x}_i) \\
&= \boldsymbol{p}_{\ell}^{(g)} + \frac{1}{M \pi_{\ell}^{(g)}} \left[\operatorname{Diag}\{\boldsymbol{p}_{\ell}^{(g)}\} - \boldsymbol{p}_{\ell}^{(g)} \boldsymbol{p}_{\ell}^{(g)T} \right] \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{p}_{\ell}} \log L(\boldsymbol{\theta}^{(g)} \mid \boldsymbol{x}_i).
\end{aligned}$$

For $j = 1, \ldots, k-1$,

$$\begin{aligned}
p_{\ell j}^{(g+1)} &= p_{\ell j}^{(g)} + \frac{1}{M} \sum_{i=1}^{n} p_{\ell j}^{(g)} \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} \left(\frac{x_{ij}}{p_{\ell j}^{(g)}} - \frac{x_{ik}}{p_{\ell k}^{(g)}} \right) \\
&\quad - \frac{1}{M} \sum_{i=1}^{n} \sum_{t=1}^{k-1} p_{\ell j}^{(g)} p_{\ell t}^{(g)} \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} \left(\frac{x_{it}}{p_{\ell t}^{(g)}} - \frac{x_{ik}}{p_{\ell k}^{(g)}} \right) \\
&= p_{\ell j}^{(g)} + \frac{1}{M} \sum_{i=1}^{n} \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} \left(x_{ij} - \frac{p_{\ell j}^{(g)}}{p_{\ell k}^{(g)}} x_{ik} \right) \\
&\quad - \frac{1}{M} \sum_{i=1}^{n} p_{\ell j}^{(g)} \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} \sum_{t=1}^{k-1} \left(x_{it} - \frac{p_{\ell t}^{(g)}}{p_{\ell k}^{(g)}} x_{ik} \right). \\
&= p_{\ell j}^{(g)} + \frac{1}{M} \sum_{i=1}^{n} \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} \left\{ \left(x_{ij} - \frac{p_{\ell j}^{(g)}}{p_{\ell k}^{(g)}} x_{ik} \right) - p_{\ell j}^{(g)} \sum_{t=1}^{k-1} \left(x_{it} - \frac{p_{\ell t}^{(g)}}{p_{\ell k}^{(g)}} x_{ik} \right) \right\} \quad (2.27)
\end{aligned}$$

Since $\sum_{t=1}^{k} x_{it} = m_i$ and $\sum_{t=1}^{k} p_{\ell t}^{(g)} = 1$,

$$\sum_{t=1}^{k-1} \left(x_{it} - \frac{p_{\ell t}^{(g)}}{p_{\ell k}^{(g)}} x_{ik} \right) = (m_i - x_{ik}) - x_{ik} \frac{1 - p_{\ell k}^{(g)}}{p_{\ell k}^{(g)}} = m_i - x_{ik}/p_{\ell k}^{(g)}.$$

Applying this result to (2.27) and simplifying we get

$$\begin{aligned}
p_{\ell j}^{(g+1)} &= p_{\ell j}^{(g)} + \frac{1}{M} \sum_{i=1}^{n} \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} \left(x_{ij} - m_i p_{\ell j}^{(g)} \right) \\
&= p_{\ell j}^{(g)} + \frac{1}{M} \sum_{i=1}^{n} \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} x_{ij} - \frac{p_{\ell j}^{(g)}}{M} \sum_{i=1}^{n} m_i \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}.
\end{aligned}$$

□

Proposition 2.15 (EM Iterations)**.** *Consider the complete data* $(\boldsymbol{X}_i, Z_i)$*, independent for* $i = 1, \ldots, n$*, where* $Z_i \sim$ *Discrete*$(1, \ldots, s; \boldsymbol{\pi})$ *and* $(\boldsymbol{X}_i \mid Z_i = \ell) \sim$ *Mult*$_k(\boldsymbol{p}_\ell, m_i)$*. Iterations for an EM algorithm are given by*

$$\pi_\ell^{(g+1)} = \frac{1}{n} \pi_\ell^{(g)} \sum_{i=1}^{n} \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} \quad \textit{and} \quad p_{\ell j}^{(g+1)} = \frac{\sum_{i=1}^{n} x_{ij} \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}}{\sum_{i=1}^{n} m_i \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}},$$

for $\ell = 1, \dots, s$ *and* $j = 1, \dots, k$.

Proof of Proposition 2.15. The complete data likelihood is

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{z}) = \prod_{i=1}^{n} \prod_{\ell=1}^{s} \Big[\pi_\ell f(\boldsymbol{x}_i \mid \boldsymbol{p}_\ell, m_i)\Big]^{\Delta_{i\ell}}.$$

where $\Delta_{i\ell} = I(z_i = \ell)$ and $(\Delta_{i1}, \dots, \Delta_{is}) \overset{\text{iid}}{\sim} \text{Mult}_s(1, \boldsymbol{\pi})$ for $i = 1, \dots, n$. Then the corresponding log-likelihood is

$$\log L(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{z}) = \sum_{i=1}^{n} \sum_{\ell=1}^{s} \Delta_{i\ell} \log \Big[\pi_\ell f(\boldsymbol{x}_i \mid \boldsymbol{p}_\ell, m_i)\Big]. \tag{2.28}$$

Since $z_1, \dots, z_n$ are not observed, we instead use the expected log-likelihood, conditional on $\boldsymbol{\theta} = \boldsymbol{\theta}^{(g)}$ and $\boldsymbol{x}$. First note that

$$\begin{aligned}
\gamma_{i\ell}^{(g)} &:= \text{E}\left(\Delta_{i\ell} \mid \boldsymbol{x}_1, \dots, \boldsymbol{x}_n, \boldsymbol{\theta}^{(g)}\right) = \text{P}(Z_i = \ell \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(g)}) \\
&= \frac{\text{P}(Z_i = \ell \mid \boldsymbol{\theta}^{(g)})\ \text{P}(\boldsymbol{x}_i \mid Z_i = \ell, \boldsymbol{\theta}^{(g)})}{f(\boldsymbol{x}_i \mid \boldsymbol{\theta}^{(g)}, m_i)} = \frac{\pi_\ell^{(g)} f(\boldsymbol{x}_i \mid \boldsymbol{p}_\ell^{(g)}, m_i)}{\sum_{a=1}^{s} \pi_a^{(g)} f(\boldsymbol{x}_i \mid \boldsymbol{p}_a^{(g)}, m_i)} = \frac{\pi_\ell^{(g)}\, \text{P}_\ell(\boldsymbol{x}_i)}{\text{P}(\boldsymbol{x}_i)}
\end{aligned}$$

is the posterior probability of population ℓ, given $\boldsymbol{x}_i$ and the previous iteration. Conditional on this information, the expectation of (2.28) becomes

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(g)}) := \sum_{i=1}^{n} \sum_{\ell=1}^{s} \gamma_{i\ell}^{(g)} \log \pi_\ell + \sum_{i=1}^{n} \sum_{\ell=1}^{s} \gamma_{i\ell}^{(g)} \log \Big[f(\boldsymbol{x}_i \mid \boldsymbol{p}_\ell, m_i)\Big].$$

Now to maximize this expression with respect to each parameter, equate partial derivatives to zero and solve for the parameter. For $\pi_1, \dots, \pi_{s-1}$ we have

$$\begin{aligned}
0 = \frac{\partial}{\partial \pi_a} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(g)}) &= \sum_{i=1}^{n} \frac{\gamma_{ia}^{(g)}}{\pi_a} - \sum_{i=1}^{n} \frac{\gamma_{is}^{(g)}}{\pi_s} \\
\iff \pi_s \sum_{i=1}^{n} \gamma_{ia}^{(g)} &= \pi_a \sum_{i=1}^{n} \gamma_{is}^{(g)}.
\end{aligned} \tag{2.29}$$

Summing both sides of (2.29) over $a = 1, \ldots, s$ we obtain

$$\begin{aligned} \pi_s \sum_{a=1}^{s} \sum_{i=1}^{n} \gamma_{ia}^{(g)} = \sum_{i=1}^{n} \gamma_{is}^{(g)} &\iff \pi_s n = \sum_{i=1}^{n} \gamma_{is}^{(g)} \\ &\iff \hat{\pi}_s^{(g+1)} = \frac{1}{n} \sum_{i=1}^{n} \gamma_{is}^{(g)} \end{aligned}$$

since the posterior probabilities $\gamma_{i1}^{(g)}, \ldots, \gamma_{is}^{(g)}$ sum to 1. Replacing this back into (2.29) yields

$$\hat{\pi}_a^{(g+1)} = \frac{\hat{\pi}_s^{(g+1)} \sum_{i=1}^{n} \gamma_{ia}^{(g)}}{\sum_{i=1}^{n} \gamma_{is}^{(g)}} = \frac{1}{n} \sum_{i=1}^{n} \gamma_{ia}^{(g)}.$$

Similar steps yield the EM iterations for the p_{ab}'s. For p_{ab} where $a = 1, \ldots, s$ and $b = 1, \ldots, k-1$,

$$\begin{aligned} 0 = \frac{\partial}{\partial p_{ab}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(g)}) &= \sum_{i=1}^{n} \gamma_{ia}^{(g)} \left(\frac{x_{ib}}{p_{ab}} - \frac{x_{ik}}{p_{ak}} \right) \\ \iff p_{ak} \sum_{i=1}^{n} \gamma_{ia}^{(g)} x_{ib} &= p_{ab} \sum_{i=1}^{n} \gamma_{ia}^{(g)} x_{ik}. \end{aligned} \tag{2.30}$$

Summing both sides of (2.30) over $b = 1, \ldots, k$ we obtain

$$p_{ak} \sum_{i=1}^{n} \gamma_{ia}^{(g)} m_i = \sum_{i=1}^{n} \gamma_{ia}^{(g)} x_{ik} \iff \hat{p}_{ak}^{(g+1)} = \frac{\sum_{i=1}^{n} x_{ik} \gamma_{ia}^{(g)}}{\sum_{i=1}^{n} m_i \gamma_{ia}^{(g)}}$$

since $x_{i1} + \cdots + x_{ik} = m_i$. Replacing this back into (2.30) yields

$$\hat{p}_{ab}^{(g+1)} = \hat{p}_{ak}^{(g+1)} \frac{\sum_{i=1}^{n} x_{ib} \gamma_{ia}^{(g)}}{\sum_{i=1}^{n} x_{ik} \gamma_{ia}^{(g)}} = \frac{\sum_{i=1}^{n} x_{ib} \gamma_{ia}^{(g)}}{\sum_{i=1}^{n} m_i \gamma_{ia}^{(g)}}.$$

□

Theorem 2.16. *Denote the estimator from EM by* $\hat{\boldsymbol{\theta}}$*, and the estimator from AFSA by* $\tilde{\boldsymbol{\theta}}$*. Suppose cluster sizes are equal, so that* $m_1 = \cdots = m_n = m$*. If the two algorithms start* *at the gth iteration with $\boldsymbol{\theta}^{(g)}$, then for the $(g+1)$th iteration,*

$$\tilde{\pi}_\ell^{(g+1)} = \hat{\pi}_\ell^{(g+1)} \quad \textit{and} \quad \tilde{p}_{\ell j}^{(g+1)} = \left(\frac{\hat{\pi}_\ell^{(g+1)}}{\pi_\ell^{(g)}}\right)\hat{p}_{\ell j}^{(g+1)} + \left(1 - \frac{\hat{\pi}_\ell^{(g+1)}}{\pi_\ell^{(g)}}\right)p_{\ell j}^{(g)}$$

for $\ell = 1, \ldots, s$ and $j = 1, \ldots, k$.

Proof of Theorem 2.16. It is immediate from Propositions 2.14 and 2.15 that $\tilde{\pi}_\ell^{(g+1)} = \hat{\pi}_\ell^{(g+1)}$, and that

$$\frac{\hat{\pi}_\ell^{(g+1)}}{\pi_\ell^{(g)}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}.$$

Now we have

$$\begin{aligned}
&\left(\frac{\hat{\pi}_\ell^{(g+1)}}{\pi_\ell^{(g)}}\right)\hat{p}_{\ell j}^{(g+1)} + \left(1 - \frac{\hat{\pi}_\ell^{(g+1)}}{\pi_\ell^{(g)}}\right)p_{\ell j}^{(g)} \\
&\quad = \frac{\sum_{i=1}^{n} x_{ij}\frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}}{mn\sum_{i=1}^{n}\frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}}\sum_{i=1}^{n}\frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} + p_{\ell j}^{(g)}\left[1 - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}\right] \\
&\quad = \frac{1}{mn}\sum_{i=1}^{n}\frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}x_{ij} + p_{\ell j}^{(g)}\left(1 - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}\right) = \tilde{p}_{\ell j}^{(g+1)}.
\end{aligned} \tag{2.31}$$

□

The AFSA iterate $\tilde{p}_{\ell j}^{(g+1)}$ can then be seen as a linear combination of the gth iterate and the $(g+1)$th step of EM. The coefficient $\hat{\pi}_\ell^{(g+1)}/\pi_\ell^{(g)}$ is nonnegative but may be larger than 1. Therefore $\tilde{p}_{\ell j}^{(g+1)}$ need not lie strictly between $\hat{p}_{\ell j}^{(g+1)}$ and $p_{\ell j}^{(g)}$. Figure 2.1 shows a plot of $\tilde{p}_{\ell j}^{(g+1)}$ as the ratio $\hat{\pi}_\ell^{(g+1)}/\pi_\ell^{(g)}$ varies. However, suppose that at the gth step the EM algorithm is close to convergence. Then

$$\hat{\pi}_\ell^{(g+1)} \approx \hat{\pi}_\ell^{(g)} \iff \frac{\hat{\pi}_\ell^{(g+1)}}{\hat{\pi}_\ell^{(g)}} \approx 1, \quad \text{for } \ell = 1, \ldots, s.$$

From (2.31) we will also have

$$\tilde{p}_{\ell j}^{(g+1)} \approx \hat{p}_{\ell j}^{(g+1)}, \quad \text{for } \ell = 1, \ldots, s, \text{ and } j = 1, \ldots, k.$$

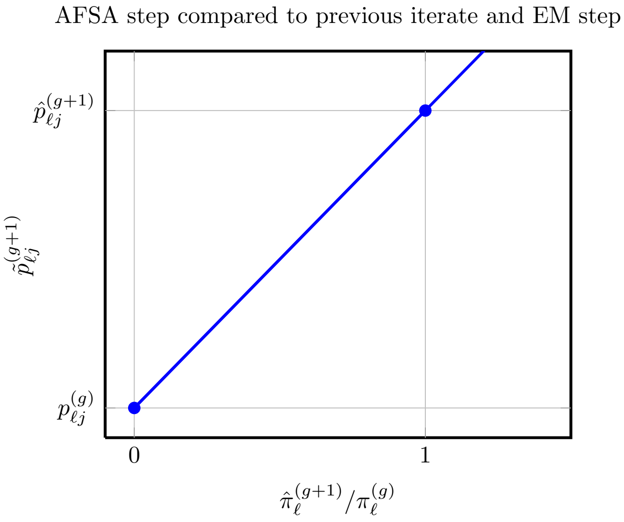


Figure 2.1: The next AFSA $\tilde{p}_{\ell j}^{(g+1)}$ iteration is a linear combination of $\hat{p}_{\ell j}^{(g+1)}$ and $p_{\ell j}^{(g)}$, which depends on the ratio $\hat{\pi}_{\ell}^{(g+1)}/\pi_{\ell}^{(g)}$.

From this point on, AFSA and EM iterations are approximately the same. Hence, in the vicinity of a solution, AFSA and EM will produce the same estimate. Note that this result holds for any m, and does not require a large cluster size justification. For the case of varying cluster sizes $m_1, \dots, m_n$,

$$\begin{aligned} &\frac{\hat{\pi}_{\ell}^{(g+1)}}{\pi_{\ell}^{(g)}} \hat{p}_{\ell j}^{(g+1)} + \left(1 - \frac{\hat{\pi}_{\ell}^{(g+1)}}{\pi_{\ell}^{(g)}}\right) p_{\ell j}^{(g)} \\ &\qquad = \frac{\sum_{i=1}^{n} x_{ij} \frac{\mathrm{P}_{\ell}(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}}{n \sum_{i=1}^{n} m_i \frac{\mathrm{P}_{\ell}(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}} \sum_{i=1}^{n} \frac{\mathrm{P}_{\ell}(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} + p_{\ell j}^{(g)} \left[1 - \frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{P}_{\ell}(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}\right], \end{aligned} \tag{2.32}$$

which does not simplify to $\tilde{p}_{\ell j}^{(g+1)}$ as in the proof of Theorem 2.16. However, this illustrates that EM and AFSA are still closely related. This also suggests an *ad hoc* revision to AFSA, letting $\tilde{p}_{\ell j}^{(g+1)}$ equal (2.32) so that the algebraic relationship to EM would be maintained as in (2.31) for the balanced case.

A more general connection is known between EM and iterations of the form

$$\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)} + \mathcal{I}_c^{-1}(\boldsymbol{\theta}^{(g)})\, S(\boldsymbol{\theta}^{(g)}), \quad g = 1, 2, \dots, \tag{2.33}$$

where $\mathcal{I}_c(\boldsymbol{\theta})$ is a complete data FIM. Titterington (1984) shows that the two iterations are approximately equivalent under appropriate regularity conditions. The equivalence is exact when the complete data likelihood is in an exponential family

$$L(\boldsymbol{\mu}) = \exp\left\{b(\boldsymbol{x}) + \boldsymbol{\eta}^T \boldsymbol{t} + a(\boldsymbol{\eta})\right\}, \quad \boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\mu}), \quad \boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x}),$$

and $\boldsymbol{\mu} := \mathrm{E}[\boldsymbol{t}(\boldsymbol{X})]$ is the parameter of interest. The complete data likelihood for our multinomial mixture is indeed an exponential family, but the parameter of interest $\boldsymbol{\theta}$ is a transformation of $\boldsymbol{\mu}$ rather than $\boldsymbol{\mu}$ itself. Therefore the equivalence is approximate, as we have seen in Theorem 2.16. The justification for AFSA leading to this chapter followed the historical approach of Blischke (1964), and not from the role of $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ as a complete data FIM. But the relationship between EM and the iterations (2.33) suggests that approximate scoring — that is, scoring with a complete data information matrix — is a reasonable approach for missing data problems beyond the finite mixture of multinomials setting.

2.5 Simulation Studies

The main result stated in Theorem 2.2 allows us to approximate the matrix $\mathcal{I}(\boldsymbol{\theta})$ by $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$, which is much more easily computed. Theorem 2.10 justifies $\widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta})$ as an approximation for the inverse FIM. In the present section, simulation studies investigate the quality of the two approximations as a function of m. We also present studies to demonstrate the convergence speed and solution quality of AFSA.

2.5.1 Distance Between True and Approximate FIM

Consider two concepts of distance to compare the closeness of the exact and approximate matrices. Based on the Frobenius norm $\|\boldsymbol{A}\|_F^2 = \sum_i \sum_j a_{ij}^2$, a distance metric

$$d_F(\boldsymbol{A}, \boldsymbol{B}) = \|\boldsymbol{A} - \boldsymbol{B}\|_F$$

can be constructed using the sum of squared differences of corresponding elements. This distance will be larger in general when the magnitudes of the elements are larger, so we will also consider a scaled version

$$d_S(\boldsymbol{A}, \boldsymbol{B}) = \frac{d_F(\boldsymbol{A}, \boldsymbol{B})}{\|\boldsymbol{B}\|_F} = \sqrt{\frac{\sum_i \sum_j (a_{ij} - b_{ij})^2}{\sum_i \sum_j b_{ij}^2}},$$

noting that this is not a true distance metric since it is not symmetric. Using these two metrics, we compare the distance between true and approximate FIMs, and also the distance between their inverses. Consider a mixture MultMix$_2(m, \boldsymbol{\theta})$ of three binomials, with parameters

$$\boldsymbol{p} = (1/7,\ 1/3,\ 2/3) \quad \text{and} \quad \boldsymbol{\pi} = (1/6,\ 2/6,\ 3/6).$$

Figure 2.2 plots the two distance types for both the FIM and inverse FIM as m varies. Note that distances are plotted on a log scale, so the vertical axis represents changes in orders of magnitude. To see more concretely what is being compared, for the moderate cluster size $m = 20$ we have

$$\begin{pmatrix} 27.222 & 0 & 0 & 0 & 0 \\ 0 & 30 & 0 & 0 & 0 \\ 0 & 0 & 45 & 0 & 0 \\ 0 & 0 & 0 & 8 & 2 \\ 0 & 0 & 0 & 2 & 5 \end{pmatrix} \text{ vs. } \begin{pmatrix} 14.346 & -2.453 & -0.184 & -3.341 & 1.625 \\ -2.453 & 12.605 & -6.749 & -4.440 & -0.944 \\ -0.184 & -6.749 & 34.175 & -1.205 & -2.914 \\ -3.341 & -4.440 & -1.205 & 6.022 & 2.536 \\ 1.625 & -0.944 & -2.914 & 2.536 & 3.621 \end{pmatrix}$$

for the approximate and exact FIMs respectively, and

$$\begin{pmatrix} 0.037 & 0 & 0 & 0 & 0 \\ 0 & 0.033 & 0 & 0 & 0 \\ 0 & 0 & 0.022 & 0 & 0 \\ 0 & 0 & 0 & 0.139 & -0.056 \\ 0 & 0 & 0 & -0.056 & 0.222 \end{pmatrix} \text{ vs.}$$

$$\begin{pmatrix} 0.216 & 0.160 & 0.020 & 0.366 & -0.295 \\ 0.160 & 0.251 & 0.043 & 0.383 & -0.240 \\ 0.020 & 0.043 & 0.040 & 0.053 & -0.003 \\ 0.366 & 0.383 & 0.053 & 0.953 & -0.690 \\ -0.295 & -0.240 & -0.003 & -0.690 & 0.827 \end{pmatrix}$$

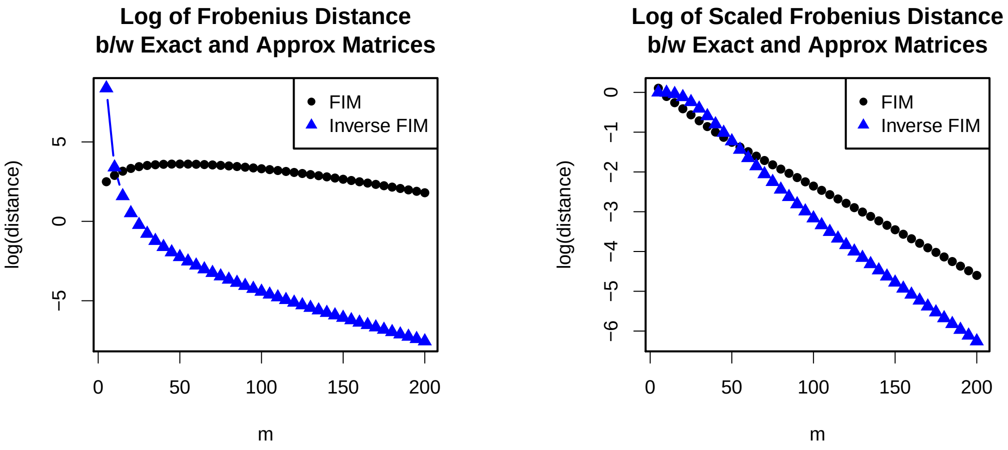
for the approximate and exact inverse FIMs. Since the approximations are block-diagonal matrices they have no way of capturing the off-diagonal blocks, which are present in the exact matrices but are eventually dominated by the block-diagonal elements as $m \to \infty$. This emphasizes one obvious disadvantage of the FIM approximation, which is that it cannot be used to estimate all asymptotic covariances for the MLEs for a fixed cluster size. For this $m = 20$ case, the block-diagonal elements for both pairs of matrices are not very close, although they are at least the same order of magnitude with the same signs. The magnitudes of elements in the inverse FIMs are in general much smaller than those in the FIMs, so the unscaled distance will naturally be smaller between the inverses.

Now in Figure 2.2 consider the distance $d_F(\widetilde{\mathcal{I}}(\boldsymbol{\theta}), \mathcal{I}(\boldsymbol{\theta}))$ as m is varied. For the FIM, the distance appears to be moderate at first, then increasing with m, and finally beginning to vanish as m becomes large. What is not reflected here is that the magnitudes of the elements themselves are increasing; this is inflating the distance until the convergence of Theorem 2.2 begins to kick in. Considering the scaled distance $d_S(\widetilde{\mathcal{I}}(\boldsymbol{\theta}), \mathcal{I}(\boldsymbol{\theta}))$ helps to suppress the effect of the element magnitudes and gives a clearer picture of the convergence.

Focusing next on the inverse FIM, consider the distance $d_F(\widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta}), \mathcal{I}^{-1}(\boldsymbol{\theta}))$. For $m < 5$ the exact FIM is computationally singular, so its inverse cannot be computed. Note that in this case the conditions for identifiability are not satisfied (see the supplement). This is not just a coincidence; there is a known relationship between model non-identifiability and singularity of the FIM (Rothenberg, 1971). For m between 5 and about 23, the distance is very large at first because of near-singularity of the FIM, but quickly returns to a reasonable magnitude. As m increases further, the distance quickly vanishes toward zero. We also consider the scaled distance $d_S(\widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta}), \mathcal{I}^{-1}(\boldsymbol{\theta}))$. Again, this helps to remove the effects of the element magnitudes, which are becoming very small as m increases. Even after taking into account the scale of the elements, the distance between the inverse matrices appears in Figure 2.2 to be converging more quickly in comparison to the distance between the FIM and its approximation. This may be interesting from an inference perspective since the inverse of the FIM corresponds to the asymptotic covariance. For small to medium cluster sizes, neither the approximate FIM nor its inverse appear to be very close to the exact matrices.

2.5.2 Approximations to Wald and Score Test Statistics

In the previous section, we saw that the inverse FIM and the inverse approximation appeared to be converging together more quickly than the FIM and the approximation.



(a) Using unscaled distance

(b) Using scaled distance

Figure 2.2: Distance between exact and approximate FIM and its inverse, as m is varied.

The following illustrates the use of the approximation and inverse approximation in inference. The Wald statistic for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ is

$$W_n(\hat{\boldsymbol{\theta}}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathcal{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

and the score statistic is

$$R_n(\boldsymbol{\theta}_0) = [S(\boldsymbol{\theta}_0)]^T [\mathcal{I}(\boldsymbol{\theta}_0)]^{-1} [S(\boldsymbol{\theta}_0)],$$

The usual large sample result gives that $W_n(\hat{\boldsymbol{\theta}}) \xrightarrow{\mathcal{L}} \chi^2_q$ and $R_n(\boldsymbol{\theta}_0) \xrightarrow{\mathcal{L}} \chi^2_q$ as $n \to \infty$, where $q = sk - 1$. In addition to carrying out the hypothesis test, the Wald statistic can be used to construct a large sample $1 - \alpha$ level confidence region

$$\left\{ \boldsymbol{\theta}_0 : (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathcal{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \leq \chi^2_{q,\alpha} \right\}, \tag{2.34}$$

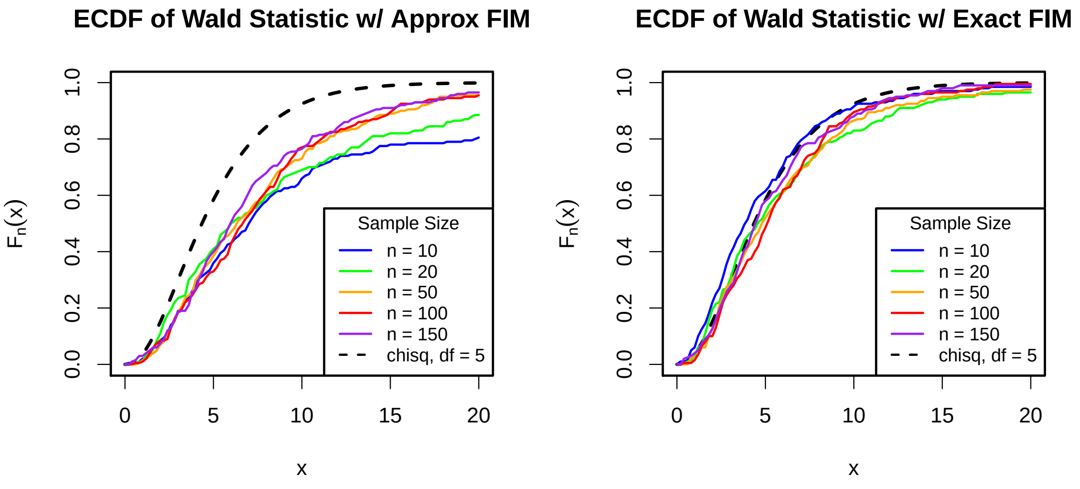
an ellipsoid in $\mathbb{R}^q$ centered at the MLE $\hat{\boldsymbol{\theta}}$, with shape determined by the FIM. Similarly, the score test statistic can be used to construct the large sample $1 - \alpha$ level confidence region

$$\left\{ \boldsymbol{\theta}_0 : [S(\boldsymbol{\theta}_0)]^T [\widetilde{\mathcal{I}}(\boldsymbol{\theta}_0)]^{-1} [S(\boldsymbol{\theta}_0)] \leq \chi^2_{q,\alpha} \right\}. \tag{2.35}$$

Consider replacing $\mathcal{I}(\hat{\boldsymbol{\theta}})$ with the approximate FIM, yielding approximated test statistics $\widetilde{W}_n(\hat{\boldsymbol{\theta}}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \widetilde{\mathcal{I}}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. and $\widetilde{R}_n(\hat{\boldsymbol{\theta}}) = [S(\boldsymbol{\theta}_0)]^T [\widetilde{\mathcal{I}}(\boldsymbol{\theta}_0)]^{-1} [S(\boldsymbol{\theta}_0)]$. Both of the approximated statistics are more easily computed than their exact counterparts based on the exact information matrix. We now compare the distributions of the exact and approximated statistics by simulation. Observations are drawn from the three-component binomial mixture with

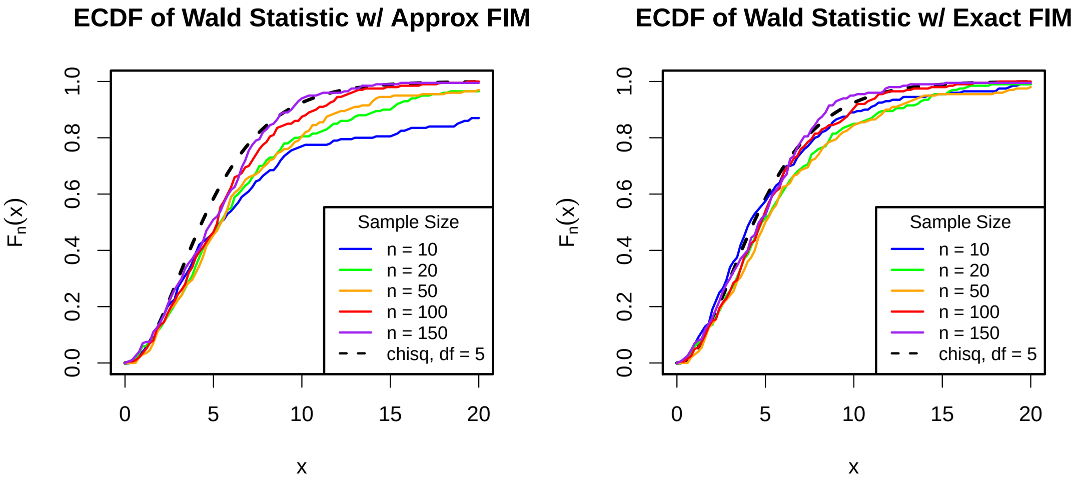
$$(p_1, p_2, p_3) = \left(\frac{1}{7}, \frac{1}{3}, \frac{2}{3} \right) \quad \text{and} \quad \boldsymbol{\pi} = \left(\frac{1}{6}, \frac{2}{6}, \frac{3}{6} \right).$$

Samples were drawn from this mixture 200 times for several m and n. For each sample we compute the four statistics, and hence obtained their empirical distributions under H_0 using the 200 samples. Figure 2.3 compares the exact and approximated Wald statistics using $m \in \{50, 100\}$ and $n \in \{10, 20, 50, 100, 120\}$. Figure 2.4 compares the exact and approximated Score statistics at the same sample sizes. In each plot, the limiting χ^2_5 cumulative distribution function (CDF) is shown for reference. We can see that even for a fairly large number of trials $m = 50$, the CDF of $\widetilde{W}_n$ is much further away from the target χ^2_q distribution than that of W_n even for the largest sample size $n = 150$. The situation under $m = 100$ is improved, and $\widetilde{W}_n$ under $n = 150$ is close to χ^2_q. On the other hand, the statistic $\widetilde{R}_n$ is close to the target χ^2_q even for $m = 50$ under the smaller sample sizes. This may be expected based on the study in Section 2.5.1, where the inverse FIM and its approximation seemed to be converging together faster than $\mathcal{I}_m(\boldsymbol{\theta})$ and $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta})$.



(a) Approximated Wald statistic, $m = 50$.

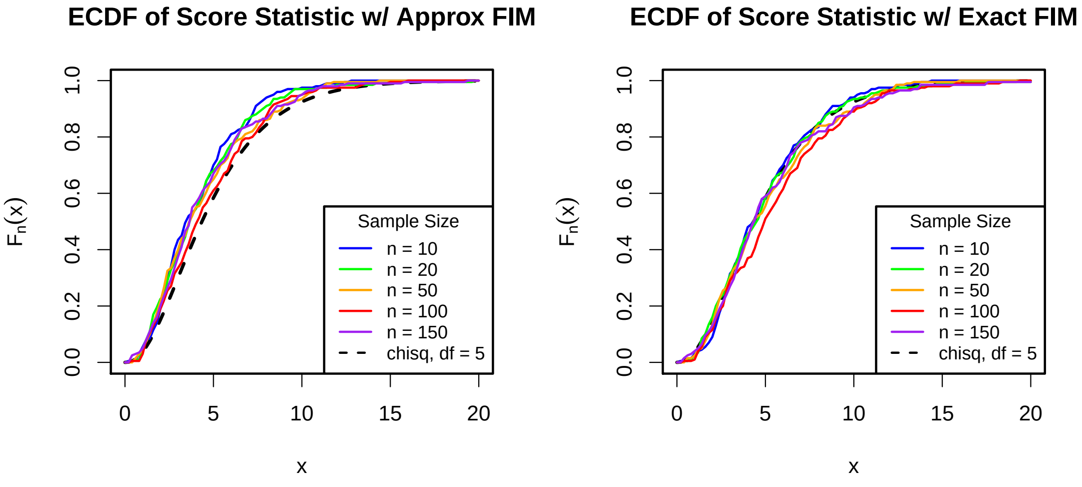
(b) Exact Wald statistic, $m = 50$.



(c) Approximated Wald statistic, $m = 100$.

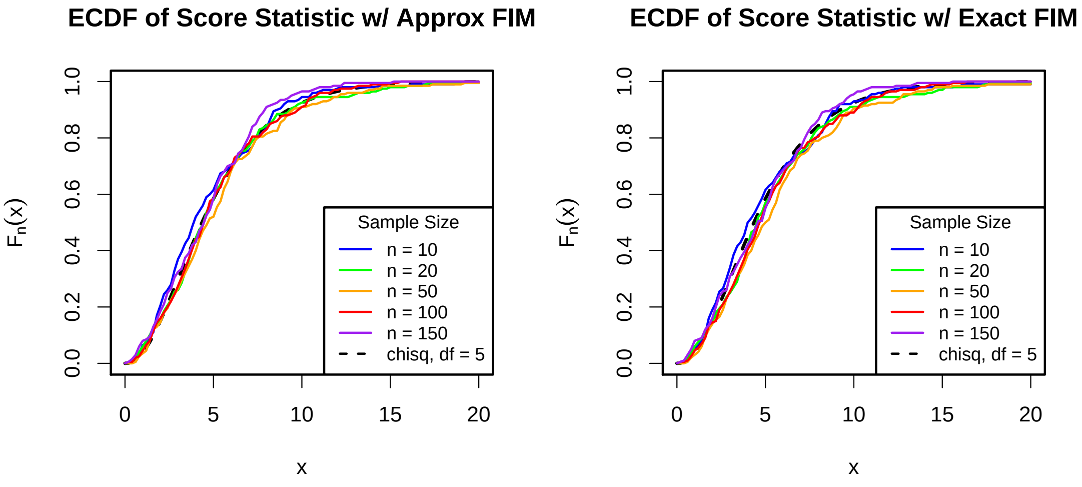
(d) Exact Wald statistic, $m = 100$.

Figure 2.3: Empirical CDF of approximated and exact Wald test statistics.



(a) Approximated Score statistic, $m = 50$.

(b) Exact Score statistic, $m = 50$.



(c) Approximated Score statistic, $m = 100$.

(d) Exact Score statistic, $m = 100$.

Figure 2.4: Empirical CDF of approximated and exact Score test statistics.

2.5.3 Effectiveness of AFSA method: Convergence Speed

We first observe the convergence speed of AFSA and several of its competitors. Consider the mixture of two trinomials

$$\boldsymbol{Y}_i \overset{\text{iid}}{\sim} \text{MultMix}_3(m = 20, \boldsymbol{\theta}), \quad i = 1, \ldots, n = 500$$

$$\boldsymbol{p}_1 = (1/3,\ 1/3,\ 1/3), \quad \boldsymbol{p}_2 = (0.1,\ 0.3,\ 0.6), \quad \pi = 0.75.$$

We now apply AFSA, FSA, and EM to a single randomly generated dataset using the same initial value $\boldsymbol{\theta}^{(0)}$. This allows for a simple comparison between the algorithms. Of course, the exact behavior of the algorithms will vary depending on the sample; the behavior over many samples is studied in Section 2.5.4. Figure 2.5 shows the expected counts for $n = 500$ observations in each of the two subpopulations while Figure 2.6 shows the particular sample we have drawn from the mixture. The sample displays evidence of two visually distinguishable modes which correspond to the two subpopulations plotted in Figures 2.5a and 2.5b. A larger proportion of observations belong to the first mode, as expected, since $\pi = 0.75$. After the gth iteration of any of the algorithms, the quantity

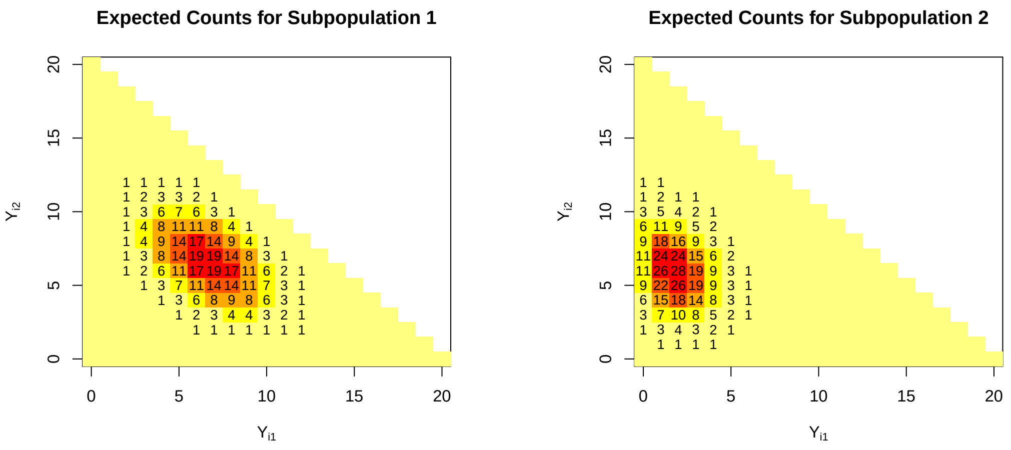
$$\delta^{(g)} = \log L(\boldsymbol{\theta}^{(g)}) - \log L(\boldsymbol{\theta}^{(g-1)})$$

is measured. The sequence $\log|\delta^{(g)}|$ is plotted for each algorithm in Figure 2.7. Note that $\delta^{(g)}$ may be negative, except for example in EM which guarantees an improvement to the log-likelihood in every step. A negative $\delta^{(g)}$ can be interpreted as negative progress, at least from a local maximum. The absolute value is taken to make plotting possible on the log scale, but some steps with negative progress have been obscured. The resulting estimates and standard errors for all algorithms are shown in Table 2.1, and additional summary information is shown in Table 2.2.

We see that AFSA and EM have almost exactly the same rate of convergence toward the same solution, as suggested by Theorem 2.16. FSA had severe problems, and was not able to converge within 100 iterations; i.e. $\delta^{(g)} < 10^{-8}$ was not attained. The situation for FSA is worse than it appears in the plot; although $\log|\delta^{(g)}|$ is becoming small, FSA's steps result in both positive and negative $\delta^{(g)}$'s until the iteration limit is reached. This indicates a failure to approach any maximum of the log-likelihood.

We also consider an FSA hybrid with a "warmup period", where for a given $\varepsilon_0 > 0$ the FIM approximation is used until the first time $\delta^{(g)} < \varepsilon_0$ is crossed. Notice that $\varepsilon_0 = \infty$ corresponds to "no warmup period". After the warmup period, exact scoring iterations are used until the final convergence criterion $\delta^{(g)} < \varepsilon$ is reached. A similar idea has been considered by Neerchal and Morel (2005), who proposed a two-stage procedure for AFSA in the RCM setting of Example 2.11. The first stage consisted of running AFSA iterations until convergence, and in the second stage one additional iteration of exact scoring was performed. The purpose of the FSA iteration was to improve standard error estimates, which were previously found to be inaccurate when computed directly from the FIM approximation (Neerchal and Morel, 1998). Here we note that FSA also offers a faster convergence rate than AFSA, given an initial path to a solution. Therefore, AFSA can be used in early iterations to move to the vicinity of a solution, then a switch to FSA will give an accelerated convergence to the solution. This approach depends on the exact FIM being feasible to compute, so the sample space cannot be too large to make use of the naive summation (2.4). Hence, there is a trade-off in the choice of ε_0 between energy spent on computing the exact FIM for FSA, and a larger number of iterations required for AFSA. Figure 2.7 shows that the hybrid strategy is effective, addressing the erratic behavior of FSA from an arbitrary starting value and the slower convergence rates of EM and AFSA. Table 2.2 shows that even a very limited warmup period such as that allowed by $\varepsilon_0 = 10$ can give a good result.

The Newton-Raphson algorithm, which has not been discussed, performed similarly to exact scoring but has issues with singularity of the Hessian in some samples.



(a) $\text{Mult}_3(m = 20, \boldsymbol{p}_1 = (1/3, 1/3, 1/3))$ (b) $\text{Mult}_3(m = 20, \boldsymbol{p}_2 = (0.1, 0.3, 0.6))$

Figure 2.5: Expected counts, rounded to the nearest integer, for $n = 500$ observations sampled independently from each of the two subpopulations. Counts rounded to zero are not shown.

Table 2.1: Estimates (and standard errors in parentheses) for the competing algorithms. FSA Hybrid produced similar results with ε_0 set to $0.001, 0.01, 0.1, 1$, and 10.

	FSA	AFSA	EM	FSA Hybrid
$\hat{p}_{11}$	0.2744 (0.0045)	0.3282 (0.0054)	0.3282 (NA)	0.3282 (0.0062)
$\hat{p}_{12}$	0.3189 (0.0047)	0.3325 (0.0054)	0.3325 (NA)	0.3325 (0.0056)
$\hat{p}_{21}$	0.0804 (0.0882)	0.1006 (0.0062)	0.1006 (NA)	0.1006 (0.0087)
$\hat{p}_{22}$	0.9193 (0.0886)	0.2749 (0.0092)	0.2749 (NA)	0.2749 (0.0106)
$\hat{\pi}$	0.9990 (0.0014)	0.7637 (0.0190)	0.7381 (NA)	0.7381 (0.0247)

Standard errors for AFSA were obtained as $\sqrt{a^{11}}, \ldots, \sqrt{a^{qq}}$, denoting $\widetilde{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}}) = ((a^{ij}))$. For FSA and FSA-Hybrid, the inverse of the exact FIM was used instead. The basic EM algorithm does not yield standard error estimates. Several extensions have been proposed to address this, such as by Louis (1982) and Meng and Rubin (1991). In light of Theorem 2.16, standard errors from $\widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta})$ evaluated at EM estimates could also be used to obtain similar results to AFSA.

2.5.4 Effectiveness of AFSA method: Monte Carlo Study

We next consider a Monte Carlo study of the difference between AFSA and EM estimators to assess the behavior of AFSA over a large number of samples. EM is con-

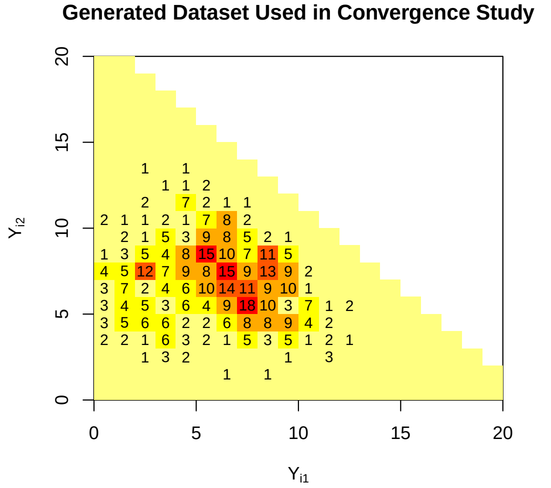


Figure 2.6: Counts of observations from the generated sample at each of the possible trinomial outcomes for $m = 20$.

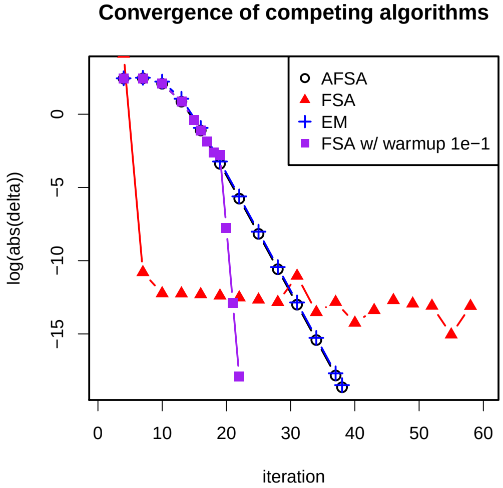


Figure 2.7: Convergence of several competing algorithms for a small test problem

Table 2.2: Convergence of several competing algorithms. Hybrid FSA is shown with several choices of the warmup tolerance ε_0. Exact FSA corresponds to $\varepsilon_0 = \infty$. Note that a maximum of 100 iterations was allowed in each case.

Method	ε_0	LogLik	Tol	Iter
AFSA	—	-2247.834	7.99×10^{-09}	38
EM	—	-2247.834	9.26×10^{-09}	38
FSA	∞	-2424.330	-4.04×10^{-07}	100
FSA	10	-2247.834	3.46×10^{-09}	15
FSA	1	-2247.834	1.44×10^{-09}	20
FSA	0.1	-2247.834	1.08×10^{-10}	23
FSA	0.01	-2247.834	1.43×10^{-09}	25
FSA	0.001	-2247.834	1.28×10^{-10}	28

sidered to produce reliable estimates, hence it is desired to achieve solutions close to EM with high probability. Observations were generated from

$$\boldsymbol{Y}_i \overset{\text{ind}}{\sim} \text{MultMix}_k(m_i, \boldsymbol{\theta}), \quad i = 1, \ldots, n = 500,$$

given varying cluster sizes $m_1, \ldots, m_n$ which themselves were generated as

$$Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta), \quad m_i = \lceil Z_i \rceil.$$

Several different settings of $\boldsymbol{\theta}$ are considered, with $s = 2$ mixing components and proportion $\pi = 0.75$ for the first component. The parameters α and β were chosen such that $\text{E}(Z_i) = \alpha\beta = 20$. This gives $\beta = 20/\alpha$ so that only α is free, and $\text{Var}(Z_i) = \alpha\beta^2 = 400/\alpha$ can be chosen as desired. The expectation and variance of m_i are intuitively similar to Z_i, and their exact values may be computed numerically.

Once the n observations are generated, an AFSA estimator $\tilde{\boldsymbol{\theta}}$ and an EM estimator $\hat{\boldsymbol{\theta}}$ are fit. This process is repeated 1000 times yielding $\tilde{\boldsymbol{\theta}}^{(r)}$ and $\hat{\boldsymbol{\theta}}^{(r)}$ for $r = 1, \ldots, 1000$. A default initial value was selected for each setting of $\boldsymbol{\theta}$ and is used for both algorithms in every repetition. To measure the closeness of the two estimators,

$$\bar{D} = \frac{1}{1000} \sum_{r=1}^{1000} D_r, \quad \text{where } D_r = \bigvee_{j=1}^{q} \left| \frac{\tilde{\theta}_j^{(r)} - \hat{\theta}_j^{(r)}}{\tilde{\theta}_j^{(r)}} \right|$$

is the maximum relative difference taken over all components of $\boldsymbol{\theta}$, averaged over all repetitions. Here $\bigvee$ represents the "maximum" operator. Notice that obtaining a good result for $\bar{D}$ depends on the vectors $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ being ordered in the same way. To help ensure this, we add the constraint $\pi_1 > \cdots > \pi_s$, which is enforced in both algorithms by reordering the estimates for $\pi_1, \ldots, \pi_s$ and $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_s$ accordingly after every iteration. Table 2.3 shows the results of the simulation. Nine different scenarios for $\boldsymbol{\theta}$ are considered. The cluster sizes $m_1, \ldots, m_n$ are selected in three different ways: a balanced case where $m_i = 20$ for $i = 1, \ldots, n$, cluster sizes selected at random with small variability (using $\alpha = 100$), and cluster sizes selected at random with moderate variability (using $\alpha = 25$). As seen in Section 2.5.1, clusters sizes on the order of $m = 20$ may not provide a high accuracy of the FIM approximation to the exact FIM, but are adequate here for AFSA.

Both AFSA and EM are susceptible to finding local maxima of the likelihood, as are all iterative optimization procedures, but in this experiment AFSA encountered the problem much more frequently. These cases stood out because the local maxima occurred with one of the mixing proportions or category probabilities close to zero, i.e. a convergence to the boundary of the parameter space. This is especially apparent in our Monte Carlo statistic $\bar{D}$, which can become very large if this occurs even once for a given scenario. The problem occurred most frequently for the case $\boldsymbol{p}_1 = (0.1, 0.3)$ and $\boldsymbol{p}_2 = (1/3, 1/3)$. To counter this, we restarted AFSA with a random starting value whenever a solution with any estimate less than 0.01 was obtained. For this experiment, no more than 15 out of 1000 samples required a restart, and no more than two restarts were needed for the same sample. In practice, we recommend starting AFSA with several initial values to ensure that any solutions on the boundary are not missteps taken by the algorithm.

The entries in Table 2.3 show that small to moderate variation of the cluster sizes does not have a significant impact on the equivalence of AFSA and EM. On the other hand, as $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ are moved closer together, the quantity $\bar{D}$ tends to become larger. Theorem 2.2 depends on the distinctness of the category probability vectors, so the quality of the FIM approximation at moderate cluster sizes may begin to suffer in this case. The estimation problem itself also intuitively becomes more difficult as $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ become closer. Although the $\bar{D}$ value in the three columns for Scenario E are on the order 10^{-3}, they are reduced to the order 10^{-6} upon removal of one one large outlier in each case; see Figure 2.8. Recall that the dimension of $\boldsymbol{p}_i$ is $k - 1$; it can be seen from Table 2.3 that increasing k from 2 to 4 does not necessarily have a negative effect on the results.

Table 2.3: Closeness between AFSA and EM estimates, over 1000 samples. Scenarios A–D represent binomial mixtures, E–G represent trinomial mixtures, and H-I represent multinomial mixtures with $k = 4$ categories. $\mathrm{V}(m_i)$ is shorthand for the variance of m_i.

	(kth probability not shown)		All m_i equal	$\alpha = 100$	$\alpha = 25$
	$\boldsymbol{p}_1$	$\boldsymbol{p}_2$	$m_i = 20$	$\mathrm{V}(m_i) \approx 4.083$	$\mathrm{V}(m_i) \approx 16.083$
A.	(0.1)	(0.5)	2.178×10^{-6}	2.019×10^{-6}	2.080×10^{-6}
B.	(0.3)	(0.5)	4.073×10^{-5}	3.501×10^{-5}	3.890×10^{-5}
C.	(0.35)	(0.5)	8.683×10^{-4}	2.625×10^{-4}	2.738×10^{-4}
D.	(0.4)	(0.5)	9.954×10^{-3}	6.206×10^{-2}	6.563×10^{-2}
E.	$(0.1, 0.3)$	$(1/3, 1/3)$	1.342×10^{-3}	1.009×10^{-3}	1.878×10^{-3}
F.	$(0.1, 0.5)$	$(1/3, 1/3)$	1.408×10^{-6}	1.338×10^{-6}	1.334×10^{-6}
G.	$(0.3, 0.5)$	$(1/3, 1/3)$	3.884×10^{-6}	3.943×10^{-6}	3.885×10^{-6}
H.	$(0.1, 0.1, 0.3)$	$(0.25, 0.25, 0.25)$	8.389×10^{-7}	8.251×10^{-7}	8.440×10^{-7}
I.	$(0.1, 0.2, 0.3)$	$(0.25, 0.25, 0.25)$	1.523×10^{-6}	1.472×10^{-6}	1.408×10^{-6}

Table 2.4 shows the results of a follow-up study to compare the convergence behavior of AFSA and EM over a large number of samples, as cluster size and separation between mixture components are varied. Here we consider the mixture of two binomials, where $p_2 = 0.5$ is fixed and p_1 varies in scenarios A–D which match to Table 2.3, and a common m is used for all observations. For each setting of m and p_1, 1000 samples were generated, and AFSA and EM were applied in turn to each sample. As expected, the number of iterations required for convergence is similar for both algorithms, and more iterations are required to find a suitable solution when $|p_2 - p_1|$ is small or when m is small.

2.6 Conclusions

A large cluster approximation was presented for the FIM of the multinomial finite mixture in Theorem 2.2, which has been proposed in (Morel and Nagaraj, 1991) and further studied in (Liu, 2005). This matrix has a convenient block-diagonal form where each non-zero block is the FIM of a standard multinomial observation. We observed that the approximation is equivalent to a complete data FIM, had the subpopulation label been recorded for each observation; this was stated as Proposition 2.7. Using this approxima-

Table 2.4: Convergence characteristics of AFSA and EM over 1000 samples. Here $p_2 = 0.5$ is fixed. The reported quantity is the average number of algorithm iterations per sample. Note that the tolerance for convergence was set to 10^{-8} and a maximum of 1000 iterations was allowed for each algorithm per sample.

		$m = 20$		$m = 50$		$m = 100$	
	p_1	AFSA	EM	AFSA	EM	AFSA	EM
A.	0.1	12.60	12.64	6.13	5.58	4.41	3.32
B.	0.3	142.79	142.67	30.37	30.51	12.20	12.24
C.	0.35	*435.90	*435.62	77.85	77.55	25.98	25.72
D.	0.4	*795.36	*796.15	*348.55	*345.50	84.67	82.93

*Results for some samples failed to converge within the allowed number of iterations. For the case (D, $m = 20$), this occurred with AFSA in 576 samples and with EM in 579 samples. For (C, $m = 20$), both algorithms failed to converge in 74 samples, while (D, $m = 50$) resulted in both algorithms failing to converge 31 samples.

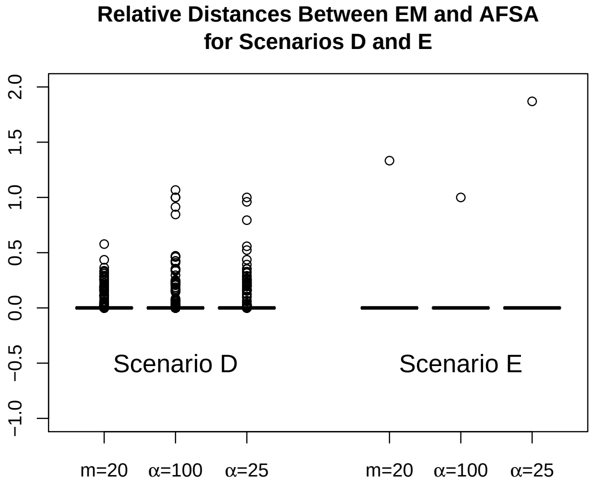


Figure 2.8: Boxplots for Scenarios D and E of the Monte Carlo study presented in Table 2.3. At this scale, the boxes appear as thin horizontal lines.

tion to the FIM, one can formulate an approximate scoring algorithm (AFSA). As first seen in (Liu, 2005), AFSA iterations are closely related to the well-known Expectation-Maximization (EM) algorithm for finite mixtures (Theorem 2.16). Simulations show that, although large cluster sizes are needed before the exact and approximate FIM are close, the approximation is quite effective in obtaining estimates through AFSA. However, for standard error computations and ensuing inference, it is advisable to use the exact FIM, especially for small to moderate cluster sizes.

We have seen that AFSA (and also EM) has an advantage, in terms of robustness to initial values, over the more standard Fisher scoring and Newton-Raphson algorithms. This comes at the cost of a slower convergence rate. For Newton-Raphson iterations, the invertibility of the Hessian depends on the sample, in addition to the current iterate $\boldsymbol{\theta}^{(g)}$ and the model. Fisher scoring iterations can be computed when the cluster size is not too small (ensuring that the FIM is nonsingular), but may converge to a poor solution or be unable to make progress at all using an arbitrarily chosen starting point. On the other hand, Fisher scoring converges very quickly given a sufficiently good starting point. Therefore, we recommend a hybrid approach: use AFSA iterations for an initial warmup period, then switch to exact scoring once a path toward a solution has been established.

Although AFSA and EM are closely related and often tend toward the same solution, AFSA is not necessarily restricted to the parameter space of the problem. AFSA also tended to converge to the boundary of the space more often than EM. These issues are not specific to AFSA; Newton-type iterations in general are prone to them without additional precautions. For the simulations in this work, we have simply restarted AFSA with a different initial value if it left the space or converged to the boundary. It is recommended to try several initial values in practice and check the solutions; this not only avoids selecting poor solutions on the boundary, but also improves the chance of finding a global maximum. Other measures could be considered as well, such as manipulating the step size at each iteration or reparameterizing the problem so that the parameter space is $\mathbb{R}^q$. These heuristics may be preferred to more complicated algorithms for constrained optimization.

AFSA may be preferable to EM in situations where it is more natural to formulate. Derivation of the E-step conditional log-likelihood may involve evaluating a complicated expectation, but this is not required for AFSA. On the other hand, AFSA requires the score vector for the observed data; this may involve a messy differentiation but is arguably easier to address numerically than the E-step. AFSA can be formulated for special finite mixtures of multinomials, such as the random-clumped multinomial from Example 2.11 and the mixture with linked regressions from Example 2.13, using Jacobians of appropriate transformations.

It is interesting to note the relationship between FSA, AFSA, and EM as Newton-type algorithms. Fisher scoring is a classic algorithm where the Hessian is replaced by its expectation. In AFSA, the Hessian is replaced instead by a complete data FIM. EM can be considered a Newton-type algorithm also, where the entire likelihood is replaced by a complete data likelihood with missing data integrated out. It is in this light that EM and AFSA iterations are seen to be approximately equivalent. Because the AFSA approach is equivalent to scoring with a complete data FIM, the technique can be applied to other finite mixture models and other missing data problems, just as EM.

In this chapter, convergence between the complete data FIM and exact FIM has only been established for binomial and multinomial mixtures and is obtained by letting the number of trials m tend to infinity. Chapter 3 extends this result to exponential family mixtures, but this must be done in such a way that there are still m "trials" within each observation. It is also desirable to find a small cluster correction that could be applied to improve the approximation. Although not addressed in this thesis, this might allow standard errors and confidence regions, such as those discussed in Section 2.5.2, to be reliably computed from the FIM approximation.

Chapter 3

Large Cluster Approximation to the Information Matrix under Exponential Family Finite Mixtures

3.1 Introduction

In this chapter, we consider approximation to the Fisher information matrix (FIM) for exponential family finite mixtures. Obtaining a simple closed form for this information matrix is generally not possible. A computationally convenient approximation may be useful in frequentist estimation (e.g. the scoring algorithm), in inference (e.g. computing standard errors and confidence intervals), as well as numerous other applications in which the information matrix is used. In Chapter 2 it was seen empirically that scoring may not require a highly accurate approximation to work well. On the other hand, inference procedures such as Wald and Score confidence intervals may require a high accuracy to produce results close to those based on the exact information matrix. Therefore, it is important in general to study the closeness between the proposed approximation and the exact matrix.

In Chapter 2, we considered an approximate information matrix which was originally proposed in (Blischke, 1962, 1964) for the finite mixture of binomials, and later extended to finite mixture of multinomials by Morel and Nagaraj (1993) and Liu (2005). We saw that it was, in fact, a complete data matrix with respect to the latent subpopulation indicator. The approximation and the true FIM are seen to converge together as the number of multinomial trials are increased. Furthermore, the approximation is useful in estimation by scoring. This chapter extends the matrix approximation to finite mixtures of exponential family densities. Such a result allows the idea of approximate information to be extended outside the scope of multinomial data analysis. We consider a special clustered sampling scheme; suppose that m observations are sampled from one of s subpopulations. It is unknown to which subpopulation the observations belong, as in the usual finite mixture, but it is known that they share a common subpopulation. This provides an analogue to the trials of a binomial or multinomial experiment, and allows us to formulate convergence of the approximate information matrix in a similar way.

The proof of convergence in the present setting is very different than the one used in the multinomial setting (Morel and Nagaraj, 1991; Liu, 2005). The multinomial proof is based on bounds for tail probabilities of binomial random variables and that the sample space is bounded. The proof in the present chapter exploits the exponential family form and does not require restrictions on the sample space. It is shown that the approximate and exact information matrices converge together as $m \to \infty$, and the convergence is exponential in m. However, the exponent includes a term which depends on the distance between subpopulations so that the convergence is very slow when the subpopulations are similar and very fast when they are not. Therefore, the approximation is most suitable when the mixed subpopulations are more distinct and m is large.

The rest of this chapter proceeds as follows. Section 3.2 provides the setup and notation needed for the rest of the chapter. Section 3.3 proves the convergence of the approximate information matrix, as well as providing rates of convergence. Section 3.4 shows an interesting connection between the convergence rate and the theoretical probability of misclassification among the s subpopulations using an optimal classification rule. Section 3.5 provides several examples of the convergence. Finally, Section 3.6 concludes the chapter.

3.2 Preliminaries

Suppose a population consists of s subpopulations, and that the ℓth subpopulation occurs with proportion π_ℓ, for $\ell = 1, \dots, s$. Let $Z \sim \text{Discrete}(1, \dots, s; \pi_1, \dots, \pi_s)$ be the result of drawing one of the populations at random; that is, $Z = \ell$ with probability π_ℓ for $\ell = 1, \dots, s$. Consider drawing an independent and identically distributed sample $\boldsymbol{X}_1, \dots, \boldsymbol{X}_m$ from the ℓth subpopulation, where $\boldsymbol{X}_j$ are d-dimensional random variables. We will suppose an exponential family density for $\boldsymbol{X}_i$ in the form

$$f(\boldsymbol{x} \mid \boldsymbol{\phi}_\ell) = \exp\left\{b(\boldsymbol{x}) + \boldsymbol{\eta}(\boldsymbol{\phi}_\ell)^T \boldsymbol{u}(\boldsymbol{x}) + a(\boldsymbol{\eta}(\boldsymbol{\phi}_\ell))\right\},$$

with respect to a dominating measure (say) λ common to $\ell = 1, \dots, s$, which can be written in terms of the natural parameter $\boldsymbol{\eta}_\ell$ as

$$f(\boldsymbol{x} \mid \boldsymbol{\eta}_\ell) = \exp\left\{b(\boldsymbol{x}) + \boldsymbol{\eta}_\ell^T \boldsymbol{u}(\boldsymbol{x}) + a(\boldsymbol{\eta}_\ell)\right\}.$$

The quantity $\boldsymbol{U}(\boldsymbol{X})$ is the sufficient statistic in this formulation, assumed to be a vector of dimension k. The subpopulation densities $f(\boldsymbol{x} \mid \boldsymbol{\eta}_\ell)$, for $\ell = 1, \dots, s$, are members of the exponential family $\mathcal{F} = \{f(\cdot \mid \boldsymbol{\eta}) : \boldsymbol{\eta} \in \Xi\}$. We will assume Ξ is an open convex set in $\mathbb{R}^k$ so that the $\mathcal{F}$ is an exponential family of full rank, and derivatives of the density may be taken at any $\boldsymbol{\eta} \in \Xi$. These assumptions ensure important regularity conditions in the theory of Fisher information which are discussed in (Shao, 2008, Section 3.1) and (Lehmann and Casella, 1998, Section 2.5), yet also cover a wide range of practically used densities. The joint density of $\boldsymbol{X}_1, \dots, \boldsymbol{X}_m$ conditional on selecting subpopulation $Z = \ell$ can be written as

$$f(\boldsymbol{x}_1, \dots, \boldsymbol{x}_m \mid \boldsymbol{\eta}_\ell) = \exp\left\{\sum_{i=1}^m b(\boldsymbol{x}_i) + \boldsymbol{\eta}_\ell^T \sum_{i=1}^m \boldsymbol{u}_i + m a(\boldsymbol{\eta}_\ell)\right\},$$

so that unconditionally,

$$f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \mid \boldsymbol{\theta}) = \sum_{\ell=1}^{s} \pi_\ell \exp\left\{\sum_{i=1}^{m} b(\boldsymbol{x_i}) + \boldsymbol{\eta}_\ell^T \sum_{i=1}^{m} \boldsymbol{u}_i + ma(\boldsymbol{\eta}_\ell)\right\},$$

where $\boldsymbol{\theta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_s, \pi_1, \ldots, \pi_{s-1})$. By Lemma 2.7.2 of Lehmann and Romano (2005), the density of $\boldsymbol{T} = \sum_{i=1}^{m} \boldsymbol{U}_i$ conditional on the subpopulation $Z = \ell$ can be written as

$$f(\boldsymbol{t} \mid \boldsymbol{\eta}_\ell) = \exp\left\{\boldsymbol{\eta}_\ell^T \boldsymbol{t} + ma(\boldsymbol{\eta}_\ell)\right\}$$

with respect to some dominating σ-finite measure ν. Therefore, unconditionally,

$$f(\boldsymbol{t} \mid \boldsymbol{\theta}) = \sum_{\ell=1}^{s} \pi_\ell \exp\left\{\boldsymbol{\eta}_\ell^T \boldsymbol{t} + ma(\boldsymbol{\eta}_\ell)\right\} \tag{3.1}$$

with respect to the same dominating measure. We will use the notation Ω to refer to the abstract sample space with a typical element ω, and $\mathcal{T}$ to refer to the space of $\boldsymbol{T}(\omega)$. The score vectors can be obtained by noting that

$$\log f(\boldsymbol{t} \mid \boldsymbol{\eta}) = \boldsymbol{\eta}^T \boldsymbol{t} + ma(\boldsymbol{\eta}),$$

and therefore

$$\frac{\partial}{\partial \boldsymbol{\eta}} \log f(\boldsymbol{t} \mid \boldsymbol{\eta}) = \boldsymbol{t} + m\frac{\partial}{\partial \boldsymbol{\eta}} a(\boldsymbol{\eta}), = \boldsymbol{t} - \mathrm{E}(\boldsymbol{T}),$$

and

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\eta}_\ell} \log f(\boldsymbol{t} \mid \boldsymbol{\theta}) &= \frac{\frac{\partial}{\partial \boldsymbol{\eta}_\ell} f(\boldsymbol{t} \mid \boldsymbol{\theta})}{f(\boldsymbol{t} \mid \boldsymbol{\theta})} \\
&= \frac{\pi_\ell \frac{\partial}{\partial \boldsymbol{\eta}_\ell} f(\boldsymbol{t} \mid \boldsymbol{\eta}_\ell)}{f(\boldsymbol{t} \mid \boldsymbol{\theta})} \\
&= \frac{\pi_\ell f(\boldsymbol{t} \mid \boldsymbol{\eta}_\ell)}{f(\boldsymbol{t} \mid \boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\eta}_\ell} \log f(\boldsymbol{t} \mid \boldsymbol{\eta}_\ell) \\
&= \frac{\pi_\ell f(\boldsymbol{t} \mid \boldsymbol{\eta}_\ell)}{f(\boldsymbol{t} \mid \boldsymbol{\theta})} \left[\boldsymbol{t} - \mathrm{E}(\boldsymbol{T} \mid Z = \ell)\right], \\
\frac{\partial}{\partial \pi_\ell} \log f(\boldsymbol{t} \mid \boldsymbol{\theta}) &= \frac{f(\boldsymbol{t} \mid \boldsymbol{\eta}_\ell) - f(\boldsymbol{t} \mid \boldsymbol{\eta}_s)}{f(\boldsymbol{t} \mid \boldsymbol{\theta})},
\end{aligned}$$

for $\ell = 1, \ldots, s$. Let $\boldsymbol{W}_\ell$ be a random variable with the distribution of $\boldsymbol{T}$ when $Z = \ell$ is observed. The Fisher information matrix in $\boldsymbol{W}_\ell$ for $\boldsymbol{\eta}_\ell$ can be obtained as

$$\begin{aligned}
\mathrm{E}\left\{-\frac{\partial^2}{\partial \boldsymbol{\eta}_\ell \partial \boldsymbol{\eta}_\ell^T} \log f(\boldsymbol{t} \mid \boldsymbol{\eta}_\ell)\right\} &= \mathrm{E}\left\{-m \frac{\partial^2}{\partial \boldsymbol{\eta}_\ell \partial \boldsymbol{\eta}_\ell^T} a(\boldsymbol{\eta}_\ell)\right\} \\
&= \mathrm{Var}(\boldsymbol{W}_\ell) \\
&= m\{\mathrm{Var}(\boldsymbol{U}_1 \mid Z = \ell)\}. \qquad (3.2)
\end{aligned}$$

Denote $\mathcal{I}(\boldsymbol{\theta})$ as the FIM of $\boldsymbol{T}$ under the finite mixture unconditional on Z, and $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ as the FIM of the complete data $(\boldsymbol{T}, Z)$, both with respect to the parameter $\boldsymbol{\theta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_s, \boldsymbol{\pi})$. Let $q = sk + s - 1$ denote the dimension of $\boldsymbol{\theta}$. We will sometimes use the subscript m to emphasize that the matrices depend on the number of observations m. The following proposition gives a closed form for $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$.

Proposition 3.1. *$\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ can be written as the $q \times q$ block diagonal matrix*

$$\widetilde{\mathcal{I}}(\boldsymbol{\theta}) = \mathrm{Blockdiag}\left(\pi_1 \boldsymbol{F}_1, \ldots, \pi_s \boldsymbol{F}_s, \boldsymbol{F}_\pi\right),$$

where for $\ell = 1, \ldots, s$,

$$\boldsymbol{F}_\ell = m\{\mathrm{Var}(\boldsymbol{U}_1 \mid Z = \ell)\}$$

is the $k \times k$ *FIM with respect to* $\boldsymbol{T} \mid Z = \ell$*, and*

$$\boldsymbol{F}_\pi = \boldsymbol{D}_\pi^{-1} + \pi_s^{-1}\mathbf{1}\mathbf{1}^T \quad \textit{and} \quad \boldsymbol{D}_\pi = \mathrm{Diag}(\pi_1, \ldots, \pi_{s-1})$$

is the FIM of Mult$_s(\boldsymbol{\pi}, 1)$ *of dimension* $(s-1) \times (s-1)$*. Here* $\mathbf{1}$ *denotes a vector of ones of the appropriate dimension.*

Proof. The complete data likelihood for $(\boldsymbol{T}, Z)$ is

$$f(\boldsymbol{t}, z \mid \boldsymbol{\theta}) = \prod_{\ell=1}^{s} \Big[\pi_\ell f(\boldsymbol{t} \mid \boldsymbol{\eta}_\ell)\Big]^{I(z=\ell)}.$$

Let $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_s)$ so that $\Delta_\ell = I(Z = \ell) \sim \text{Bernoulli}(\pi_\ell)$, and let $\boldsymbol{\Delta}_{-s}$ denote the vector $(\Delta_1, \ldots, \Delta_{s-1})$. This complete data likelihood leads to the score vector with

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\eta}_a} \log f(\boldsymbol{t}, z \mid \boldsymbol{\theta}) &= \Delta_a \frac{\partial}{\partial \boldsymbol{\eta}_a} \log f(\boldsymbol{t} \mid \boldsymbol{\eta}_a), \\
\frac{\partial}{\partial \boldsymbol{\pi}} \log f(\boldsymbol{t}, z \mid \boldsymbol{\theta}) &= \boldsymbol{D}_\pi^{-1}\boldsymbol{\Delta}_{-s} - \frac{\Delta_s}{\pi_s}\mathbf{1}.
\end{aligned}$$

Taking second derivatives yields

$$\begin{aligned}
\frac{\partial^2}{\partial \boldsymbol{\eta}_a \partial \boldsymbol{\eta}_a^T} \log f(\boldsymbol{t}, z \mid \boldsymbol{\theta}) &= \Delta_a \frac{\partial^2}{\partial \boldsymbol{\eta}_a \partial \boldsymbol{\eta}_a^T} \log f(\boldsymbol{t}, \mid \boldsymbol{\eta}_a) \\
\frac{\partial^2}{\partial \boldsymbol{\eta}_a \partial \boldsymbol{\eta}_b^T} \log f(\boldsymbol{t}, z \mid \boldsymbol{\theta}) &= 0, \qquad \text{for } a \neq b, \\
\frac{\partial^2}{\partial \boldsymbol{\eta}_a \partial \boldsymbol{\pi}^T} \log f(\boldsymbol{t}, z \mid \boldsymbol{\theta}) &= 0, \\
\frac{\partial^2}{\partial \boldsymbol{\pi} \partial \boldsymbol{\pi}^T} \log f(\boldsymbol{t}, z \mid \boldsymbol{\theta}) &= -\left[\boldsymbol{D}_\pi^{-2}\boldsymbol{\Delta}_{-s} + \frac{\Delta_s}{\pi_s^2}\mathbf{1}\mathbf{1}^T\right].
\end{aligned}$$

Now take the expected value of the negative of each of these terms, jointly with respect to $(\boldsymbol{T}, Z)$, to obtain the blocks of $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$. $\square$

The matrix $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ is seen to serve the role of the approximate information matrix, which was introduced in Chapter 2 in the more restricted context of the finite mixture of multinomials. In Section 3.3 we show that $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta}) \to \mathbf{0}$ as $m \to \infty$, just as in the setting of multinomial finite mixtures.

3.3 Convergence of Approximate Information Matrix

The proof of the convergence of $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})$ to $\mathbf{0}$ will proceed in several steps. We will first show that this difference is the expected value an the information matrix. One simple consequence of this is that the difference must be positive semidefinite. Denote $\mathcal{I}_{Z|\boldsymbol{T}}(\boldsymbol{\theta})$ as the FIM of Z conditional on $\boldsymbol{T}$.

Lemma 3.2. *The matrix* $\widetilde{\mathcal{I}}(\boldsymbol{\theta}) - \mathcal{I}(\boldsymbol{\theta})$ *is equal to* $\mathrm{E}_{\boldsymbol{T}}\left[\mathcal{I}_{Z|\boldsymbol{T}}(\boldsymbol{\theta})\right]$.

Proof. Notice that

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\boldsymbol{T}, Z) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(Z \mid \boldsymbol{T}) + \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\boldsymbol{T}).$$

Therefore,

$$\begin{aligned}
\widetilde{\mathcal{I}}(\boldsymbol{\theta}) &= \mathrm{E}_{\boldsymbol{T},Z}\left[\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(\boldsymbol{T},Z)\right\}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(\boldsymbol{T},Z)\right\}^T\right] \\
&= \mathrm{E}_{\boldsymbol{T}}\ \mathrm{E}_{Z|\boldsymbol{T}}\left[\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(Z\mid\boldsymbol{T})\right\}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(Z\mid\boldsymbol{T})\right\}^T\right] \\
&\qquad + \mathrm{E}_{\boldsymbol{T},Z}\left[\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(Z\mid\boldsymbol{T})\right\}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(\boldsymbol{T})\right\}^T\right] \\
&\qquad + \mathrm{E}_{\boldsymbol{T},Z}\left[\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(\boldsymbol{T})\right\}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(Z\mid\boldsymbol{T})\right\}^T\right] \\
&\qquad + \mathrm{E}_{\boldsymbol{T}}\left[\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(\boldsymbol{T})\right\}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(\boldsymbol{T})\right\}^T\right] \\
&= \mathrm{E}_{\boldsymbol{T}}\left[\mathcal{I}_{Z|\boldsymbol{T}}(\boldsymbol{\theta})\right] + \boldsymbol{B}_* + \boldsymbol{B}_*^T + \mathcal{I}(\boldsymbol{\theta}) \qquad (3.3)
\end{aligned}$$

where the last term is equal to $\mathcal{I}(\boldsymbol{\theta})$. Now we have

$$\begin{aligned}
\boldsymbol{B}_* &= \mathrm{E}_{\boldsymbol{T},Z}\left[\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(Z\mid\boldsymbol{T})\right\}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(\boldsymbol{T})\right\}^T\right] \\
&= \mathrm{E}_{\boldsymbol{T}}\ \mathrm{E}_{Z|\boldsymbol{T}}\left[\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(Z\mid\boldsymbol{T})\right\}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(\boldsymbol{T})\right\}^T\right] \\
&= \mathrm{E}_{\boldsymbol{T}}\ \mathrm{E}_{Z|\boldsymbol{T}}\left[\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(Z\mid\boldsymbol{T})\right\}\right]\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(\boldsymbol{T})\right\}^T \\
&= \boldsymbol{0}
\end{aligned}$$

The result follows from rearranging terms in (3.3). □

The quantity $\mathrm{E}_{\boldsymbol{T}}\left[\mathcal{I}_{Z|\boldsymbol{T}}(\boldsymbol{\theta})\right]$ has been referred to as the "missing information" (Orchard and Woodbury, 1972), so that we have

$$\text{Observed Information} = \text{Complete Information} - \text{Missing Information}.$$

Before proceeding with the main result, we can state a few important consequences of the previous lemma. The first says that standard errors obtained from the approximate information matrix are systematically too optimistic (small) compared to those obtained from the exact information matrix.

Corollary 3.3. *Suppose* $\mathcal{I}(\boldsymbol{\theta})$ *and* $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ *are nonsingular, and that* $\widetilde{\mathcal{I}}(\boldsymbol{\theta}) - \mathcal{I}(\boldsymbol{\theta})$ *is positive definite. Denote by* $\mathcal{I}^{ij}(\boldsymbol{\theta})$ *and* $\widetilde{\mathcal{I}}^{ij}(\boldsymbol{\theta})$ *the elements of the two inverse matrices respectively. Then* $\mathcal{I}^{jj}(\boldsymbol{\theta}) > \widetilde{\mathcal{I}}^{jj}(\boldsymbol{\theta})$ *for* $j = 1, \ldots, q$.

Proof. Proposition A.3 gives that $\mathcal{I}^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta})$ is positive definite. Therefore the diagonal elements

$$\boldsymbol{e}_j^T \left[\mathcal{I}^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta})\right] \boldsymbol{e}_j, \qquad j = 1, \ldots, q.$$

are positive. □

Similarly, we can show that a Wald-like test statistic based on the approximation will be systematically too large, and a Score-like test statistic will be too small.

Corollary 3.4. *For any* $\boldsymbol{\theta}_0 \in \Theta$*,*

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \widetilde{\mathcal{I}}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \geq (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathcal{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

Proof. Consider the quantity

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \left(\widetilde{\mathcal{I}}(\hat{\boldsymbol{\theta}}) - \mathcal{I}(\hat{\boldsymbol{\theta}})\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0). \tag{3.4}$$

From Lemma 3.2, $\widetilde{\mathcal{I}}(\hat{\boldsymbol{\theta}}) - \mathcal{I}(\hat{\boldsymbol{\theta}}) = \mathrm{E}_{\boldsymbol{T}}\left[\mathcal{I}_{Z|\boldsymbol{T}}(\boldsymbol{\theta})\right]$, an expected value of a conditional information matrix which is positive semidefinite. Therefore (3.4) is seen to be nonnegative and the result follows. □

Corollary 3.5. *Under the same conditions as Corollary 3.3,*

$$[S(\boldsymbol{\theta}_0)]^T \mathcal{I}^{-1}(\boldsymbol{\theta}_0)[S(\boldsymbol{\theta}_0)] > [S(\boldsymbol{\theta}_0)]^T \widetilde{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}}_0)[S(\boldsymbol{\theta}_0)]$$

for any $\boldsymbol{\theta}_0 \in \Theta$.

Proof. This follows from applying Proposition A.3 to

$$[S(\boldsymbol{\theta}_0)]^T \left(\mathcal{I}^{-1}(\boldsymbol{\theta}_0) - \widetilde{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}}_0)\right)[S(\boldsymbol{\theta}_0)]. \tag{3.5}$$

Because $\mathcal{I}^{-1}(\boldsymbol{\theta}_0) - \widetilde{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}}_0)$ is positive definite, the quantity (3.5) is seen to be strictly positive, and the result follows. □

An important consequence of Lemma 3.2 is given as Proposition 3.6, which concludes that the diagonal elements of $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})$ have the largest magnitudes in the matrix. Therefore, the convergence of all elements to zero will follow if we can show that the diagonal elements convergence to zero.

Proposition 3.6. *Denote the* (i, j)*th element of* $\mathcal{I}_{Z|\boldsymbol{T}}(\boldsymbol{\theta})$ *as* $C_{ij}^{(m)}$ *when the sample size is* m*. Then*

$$\mathrm{E}\,|C_{ij}^{(m)}| \leq \left\{\mathrm{E}(C_{ii}^{(m)})\right\}^{1/2} \left\{\mathrm{E}(C_{jj}^{(m)})\right\}^{1/2}.$$

Proof. Recall that $\mathrm{E}(C_{ij}^{(m)})$ is the (i, j)th element of $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})$ by Lemma 3.2. Because $\mathcal{I}_{Z|\boldsymbol{T}}(\boldsymbol{\theta})$ is the covariance matrix of a score function, we may apply the Cauchy-Schwarz inequality to obtain

$$|C_{ij}^{(m)}| \leq [C_{ii}^{(m)}]^{1/2} \cdot [C_{jj}^{(m)}]^{1/2}.$$

for any pair (i, j), and therefore

$$\mathrm{E}\,|C_{ij}^{(m)}| \leq \mathrm{E}\left\{[C_{ii}^{(m)}]^{1/2} \cdot [C_{jj}^{(m)}]^{1/2}\right\}.$$

Now apply the Cauchy-Schwarz inequality to the right hand side to obtain

$$\begin{aligned}\mathrm{E}\left\{[C_{ii}^{(m)}]^{1/2} \cdot [C_{jj}^{(m)}]^{1/2}\right\} &\leq \left\{\mathrm{E}[(C_{ii}^{(m)})^{\frac{1}{2}\cdot 2}]\right\}^{1/2} \cdot \left\{\mathrm{E}[(C_{jj}^{(m)})^{\frac{1}{2}\cdot 2}]\right\}^{1/2} \\ &\leq \left\{\mathrm{E}[C_{ii}^{(m)}]\right\}^{1/2} \cdot \left\{\mathrm{E}[C_{jj}^{(m)}]\right\}^{1/2},\end{aligned}$$

which gives the result. □

We are focusing on the parameterization $\boldsymbol{\theta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_s, \boldsymbol{\pi})$ for its convenience. The following remark discusses the convergence behavior under a transformation $\boldsymbol{\psi}(\boldsymbol{\theta})$ of the parameters.

Remark 3.7. Suppose $\boldsymbol{\psi}(\boldsymbol{\theta})$ is a transformation of $\boldsymbol{\theta}$ which does not depend on m. We have that

$$\widetilde{\mathcal{I}}_m(\boldsymbol{\psi}) - \mathcal{I}_m(\boldsymbol{\psi}) = \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\psi}}\right) \left[\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})\right] \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\psi}}\right)^T,$$

so that $\widetilde{\mathcal{I}}_m(\boldsymbol{\psi}) - \mathcal{I}_m(\boldsymbol{\psi})$ as $m \to \infty$ if and only if $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})$. Also note that the rate of convergence of the elements of $\widetilde{\mathcal{I}}_m(\boldsymbol{\psi}) - \mathcal{I}_m(\boldsymbol{\psi})$ is determined by the rate of the elements of $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})$. Therefore, the development focuses on the parameterization $\boldsymbol{\theta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_s, \boldsymbol{\pi})$ even though our eventual interest may be in $\boldsymbol{\psi}(\boldsymbol{\theta})$.

Now consider the block decomposition of the exact information matrix

$$\mathcal{I}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{A}_{11} & \dots & \boldsymbol{A}_{1s} & \boldsymbol{A}_{1\pi} \\ \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{A}_{s1} & \dots & \boldsymbol{A}_{ss} & \boldsymbol{A}_{s\pi} \\ \boldsymbol{A}_{\pi 1} & \dots & \boldsymbol{A}_{\pi s} & \boldsymbol{A}_{\pi\pi} \end{pmatrix}, \tag{3.6}$$

with blocks

$$\begin{aligned} \boldsymbol{A}_{ab} &= \mathrm{E}\left[\left\{\frac{\partial}{\partial \boldsymbol{\eta}_a} \log f(\boldsymbol{t} \mid \boldsymbol{\theta}\right\}\left\{\frac{\partial}{\partial \boldsymbol{\eta}_b} \log f(\boldsymbol{t} \mid \boldsymbol{\theta})\right\}^T\right], \quad a, b \in \{1, \dots, s\}, \\ \boldsymbol{A}_{b\pi}^T = \boldsymbol{A}_{\pi b} &= \mathrm{E}\left[\left\{\frac{\partial}{\partial \boldsymbol{\pi}} \log f(\boldsymbol{t} \mid \boldsymbol{\theta})\right\}\left\{\frac{\partial}{\partial \boldsymbol{\eta}_b} \log f(\boldsymbol{t} \mid \boldsymbol{\theta})\right\}^T\right], \quad b \in \{1, \dots, s\}, \\ \boldsymbol{A}_{\pi\pi} &= \mathrm{E}\left[\left\{\frac{\partial}{\partial \boldsymbol{\pi}} \log f(\boldsymbol{x} \mid \boldsymbol{\theta})\right\}\left\{\frac{\partial}{\partial \boldsymbol{\pi}} \log f(\boldsymbol{t} \mid \boldsymbol{\theta})\right\}^T\right]. \end{aligned}$$

By Proposition 3.6, we only need to show convergence for the diagonal elements of $\widetilde{\mathcal{I}}(\boldsymbol{\theta}) - \mathcal{I}(\boldsymbol{\theta})$; to do this, we will obtain expressions for the diagonal blocks. It will be helpful to define

$$\begin{aligned} &R_i^{(m)}(\boldsymbol{t}) = \sum_{\ell \neq i}^{s} \pi_\ell \exp\{(\boldsymbol{\eta}_\ell - \boldsymbol{\eta}_i)^T \boldsymbol{t} + m[a(\boldsymbol{\eta}_\ell) - a(\boldsymbol{\eta}_i)]\} \equiv \frac{f(\boldsymbol{t} \mid \boldsymbol{\theta})}{f(\boldsymbol{t} \mid \boldsymbol{\eta}_i)} - \pi_i, \quad \text{and} \\ &Q_i^{(m)}(\boldsymbol{t}) = \frac{\pi_i f(\boldsymbol{t} \mid \boldsymbol{\eta}_i)}{f(\boldsymbol{t} \mid \boldsymbol{\theta})}. \end{aligned}$$

Notice that $Q_i^{(m)}(\boldsymbol{T}) = \mathrm{P}(Z = \ell \mid \boldsymbol{T})$ is the posterior probability of observing the ℓth subpopulation given an observed $\boldsymbol{T}$, hence taking expectation with respect to the mixture density of $f(\boldsymbol{t} \mid \boldsymbol{\theta})$ yields

$$\mathrm{E}_{\boldsymbol{T}}[Q_i^{(m)}(\boldsymbol{T})] = \mathrm{E}_{\boldsymbol{T}}\left\{\mathrm{E}_{Z|\boldsymbol{T}}[I(Z = \ell) \mid \boldsymbol{T}]\right\} = \mathrm{P}(Z = \ell) = \pi_\ell. \tag{3.7}$$

Later we will encounter the same expectation but under the density $f(\boldsymbol{t} \mid \boldsymbol{\eta}_\ell)$, in which case the simplification (3.7) does not happen. Now consider block (i,i) of the decomposition (3.6). We have

$$\begin{aligned}
\pi_i \boldsymbol{F}_i - \boldsymbol{A}_{ii} &= \pi_i \boldsymbol{F}_i - \mathrm{E}\left[\left\{\frac{\partial}{\partial \boldsymbol{\eta}_i} \log f(\boldsymbol{t} \mid \boldsymbol{\theta})\right\}\left\{\frac{\partial}{\partial \boldsymbol{\eta}_i} \log f(\boldsymbol{t} \mid \boldsymbol{\theta})\right\}^T\right] \\
&= \pi_i \int \Big(\boldsymbol{t} + ma'(\eta_i)\Big)\Big(\boldsymbol{t} + ma'(\eta_i)\Big)^T f(\boldsymbol{t} \mid \boldsymbol{\theta}) d\nu(\boldsymbol{t}) \\
&\quad - \int \left(\frac{\pi_i f(\boldsymbol{t} \mid \boldsymbol{\eta}_i)}{f(\boldsymbol{t} \mid \boldsymbol{\theta})}\right)^2 \Big(\boldsymbol{t} + ma'(\eta_i)\Big)\Big(\boldsymbol{t} + ma'(\eta_i)\Big)^T f(\boldsymbol{t} \mid \boldsymbol{\theta}) d\nu(\boldsymbol{t}) \\
&= \pi_i^2 \int \left(\frac{1}{\pi_i} - \frac{\pi_i f(\boldsymbol{t} \mid \boldsymbol{\eta}_i)}{f(\boldsymbol{t} \mid \boldsymbol{\theta})}\right) \Big(\boldsymbol{t} - \mathrm{E}(\boldsymbol{W}_i)\Big)\Big(\boldsymbol{t} - \mathrm{E}(\boldsymbol{W}_i)\Big)^T f(\boldsymbol{t} \mid \boldsymbol{\eta}_i) d\nu(\boldsymbol{t}) \\
&= \pi_i^2 \int \left(\frac{f(\boldsymbol{t} \mid \boldsymbol{\theta}) - \pi_i f(\boldsymbol{t} \mid \boldsymbol{\eta}_i)}{f(\boldsymbol{t} \mid \boldsymbol{\theta})}\right) \Big(\boldsymbol{t} - \mathrm{E}(\boldsymbol{W}_i)\Big)\Big(\boldsymbol{t} - \mathrm{E}(\boldsymbol{W}_i)\Big)^T f(\boldsymbol{t} \mid \boldsymbol{\eta}_i) d\nu(\boldsymbol{t}) \\
&= \pi_i^2 \int \Big[1 - Q_i^{(m)}(\boldsymbol{t})\Big] \Big(\boldsymbol{t} - \mathrm{E}(\boldsymbol{W}_i)\Big)\Big(\boldsymbol{t} - \mathrm{E}(\boldsymbol{W}_i)\Big)^T f(\boldsymbol{t} \mid \boldsymbol{\eta}_i) d\nu(\boldsymbol{t})
\end{aligned}$$

The jth diagonal element of this block is therefore

$$\begin{aligned}
&\pi_i^2 \int \Big[1 - Q_i^{(m)}(\boldsymbol{t})\Big] \Big[t_j - \mathrm{E}(W_{ij})\Big]^2 f(\boldsymbol{t} \mid \boldsymbol{\eta}_i) d\nu(\boldsymbol{t}) \\
&\quad = \pi_i^2 \, \mathrm{E}\left\{\Big[1 - Q_i^{(m)}(\boldsymbol{W}_i)\Big] \Big[W_{ij} - \mathrm{E}(W_{ij})\Big]^2\right\}
\end{aligned} \tag{3.8}$$

Now consider the lower right block of the decomposition (3.6),

$$\boldsymbol{F}_\pi - \boldsymbol{A}_{\pi\pi} = \left(\boldsymbol{D}_\pi^{-1} + \pi_s^{-1} \mathbf{1}\mathbf{1}^T\right) - \mathrm{E}\left[\left\{\frac{\partial}{\partial \boldsymbol{\pi}} \log f(\boldsymbol{t} \mid \boldsymbol{\theta})\right\}\left\{\frac{\partial}{\partial \boldsymbol{\pi}} \log f(\boldsymbol{t} \mid \boldsymbol{\theta})\right\}^T\right] \tag{3.9}$$

$$\begin{aligned}
&= \left(\boldsymbol{D}_\pi^{-1} + \pi_s^{-1} \mathbf{1}\mathbf{1}^T\right) \\
&\quad - \mathrm{E}\left[\frac{1}{f^2(\boldsymbol{t} \mid \boldsymbol{\theta})} \begin{pmatrix} f(\boldsymbol{t} \mid \boldsymbol{\eta}_1) - f(\boldsymbol{t} \mid \boldsymbol{\eta}_s) \\ \vdots \\ f(\boldsymbol{t} \mid \boldsymbol{\eta}_{s-1}) - f(\boldsymbol{t} \mid \boldsymbol{\eta}_s) \end{pmatrix} \begin{pmatrix} f(\boldsymbol{t} \mid \boldsymbol{\eta}_1) - f(\boldsymbol{t} \mid \boldsymbol{\eta}_s) \\ \vdots \\ f(\boldsymbol{t} \mid \boldsymbol{\eta}_{s-1}) - f(\boldsymbol{t} \mid \boldsymbol{\eta}_s) \end{pmatrix}^T\right].
\end{aligned}$$

Denote the (a,b)th entry of $\boldsymbol{F}_\pi - \boldsymbol{A}_{\pi\pi}$ as ξ_{ab}. Expressions for the diagonal elements are given by

$$
\begin{aligned}
\xi_{aa} &= (\pi_a^{-1} + \pi_s^{-1}) - \mathrm{E}\left[\left(\frac{f(\boldsymbol{t} \mid \boldsymbol{\eta}_a) - f(\boldsymbol{t} \mid \boldsymbol{\eta}_s)}{f(\boldsymbol{t} \mid \boldsymbol{\theta})}\right)^2\right] \\
&= (\pi_a^{-1} + \pi_s^{-1}) - \mathrm{E}\left[\frac{f^2(\boldsymbol{t} \mid \boldsymbol{\eta}_a) - 2f(\boldsymbol{t} \mid \boldsymbol{\eta}_a)f(\boldsymbol{t} \mid \boldsymbol{\eta}_s) + f^2(\boldsymbol{t} \mid \boldsymbol{\eta}_s)}{f^2(\boldsymbol{t} \mid \boldsymbol{\theta})}\right] \\
&= (\pi_a^{-1} + \pi_s^{-1}) - \int \frac{f^2(\boldsymbol{t} \mid \boldsymbol{\eta}_a)}{f(\boldsymbol{t} \mid \boldsymbol{\theta})} d\nu(\boldsymbol{t}) - \int \frac{f^2(\boldsymbol{t} \mid \boldsymbol{\eta}_s)}{f(\boldsymbol{t} \mid \boldsymbol{\theta})} d\nu(\boldsymbol{t}) \\
&\quad + 2\int \frac{f(\boldsymbol{t} \mid \boldsymbol{\eta}_a)f(\boldsymbol{t} \mid \boldsymbol{\eta}_s)}{f(\boldsymbol{t} \mid \boldsymbol{\theta})} d\nu(\boldsymbol{t}) + 2\int \frac{f(\boldsymbol{t} \mid \boldsymbol{\eta}_a)f(\boldsymbol{t} \mid \boldsymbol{\eta}_s)}{f(\boldsymbol{t} \mid \boldsymbol{\theta})} d\nu(\boldsymbol{t}) \\
&= (\pi_a^{-1} + \pi_s^{-1}) - \pi_a^{-1} \int Q_a^{(m)}(\boldsymbol{t}) f(\boldsymbol{t} \mid \boldsymbol{\eta}_a) d\nu(\boldsymbol{t}) - \pi_s^{-1} \int Q_s^{(m)}(\boldsymbol{t}) f(\boldsymbol{t} \mid \boldsymbol{\eta}_s) d\nu(\boldsymbol{t}) \\
&\quad + 2\pi_a^{-1} \int Q_a^{(m)}(\boldsymbol{t}) f(\boldsymbol{t} \mid \boldsymbol{\eta}_s) d\nu(\boldsymbol{t}) \\
&= (\pi_a^{-1} + \pi_s^{-1}) - \pi_a^{-1}\,\mathrm{E}\left[Q_a^{(m)}(\boldsymbol{W}_a)\right] - \pi_s^{-1}\,\mathrm{E}\left[Q_s^{(m)}(\boldsymbol{W}_s)\right] + 2\pi_a^{-1}\,\mathrm{E}\left[Q_a^{(m)}(\boldsymbol{W}_s)\right]
\end{aligned}
\tag{3.10}
$$

The following Lemma gives a convexity result for exponential family densities which will determine the behavior of $R_i^{(m)}(\boldsymbol{W}_j)$ and $Q_i^{(m)}(\boldsymbol{W}_j)$ as $m \to \infty$.

Lemma 3.8. *Suppose the density* $f(\boldsymbol{t} \mid \boldsymbol{\eta}) = \exp\{\boldsymbol{\eta}^T \boldsymbol{t} + ma(\boldsymbol{\eta})\}$, *has natural parameter space* Ξ *which is an open convex set, and FIM* $\mathcal{I}_m(\boldsymbol{\eta})$ *is positive definite on* Ξ. *Then for any* $\boldsymbol{\eta}^* \in \Xi$

$$
a(\boldsymbol{\eta}) - a(\boldsymbol{\eta}^*) < a'(\boldsymbol{\eta}^*)^T(\boldsymbol{\eta} - \boldsymbol{\eta}^*), \quad \forall \boldsymbol{\eta} \in \Xi. \tag{3.11}
$$

where $a'(\boldsymbol{\eta})$ *denotes the derivative of* a *at* $\boldsymbol{\eta}$.

Proof. This proof uses a convexity argument (Boyd and Vandenberghe, 2004, c.f.). Notice that the Hessian $H_a(\boldsymbol{\eta})$ of $a(\boldsymbol{\eta})$ can be obtained from

$$
\frac{\partial^2}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} \log f(\boldsymbol{t} \mid \boldsymbol{\eta}) = \frac{\partial}{\partial \boldsymbol{\eta}} \Big[\boldsymbol{t} + ma'(\boldsymbol{\eta})\Big] = mH_a(\boldsymbol{\eta}).
$$

Now we have that $\mathcal{I}_m(\boldsymbol{\eta}) = -mH_a(\boldsymbol{\eta})$ is positive definite on Ξ, which implies that $-a$ is a strictly convex function. Since a is differentiable on the convex set Ξ we have, for

$g := -a$,

$$g(\boldsymbol{\eta}) - g(\boldsymbol{\eta}^*) > g'(\boldsymbol{\eta}^*)^T(\boldsymbol{\eta} - \boldsymbol{\eta}^*), \quad \forall \boldsymbol{\eta} \in \Xi,$$

which is equivalent to the result (3.11). $\square$

Next, the behavior of $R_i^{(m)}(\boldsymbol{W}_j)$ and $Q_i^{(m)}(\boldsymbol{W}_j)$ can be determined for large m; note that the behavior depends on which the distribution, $j = 1, \ldots, s$, is assumed for $\boldsymbol{W}_j$. Let us define the expressions

$$\begin{aligned} -\gamma_{IJK} &= -a'(\boldsymbol{\eta}_J)^T(\boldsymbol{\eta}_I - \boldsymbol{\eta}_K) + [a(\boldsymbol{\eta}_I) - a(\boldsymbol{\eta}_K)], \\ c_i^* &= \bigwedge_{\ell \neq i}^{s} \gamma_{\ell i i}, \quad d_{ij}^* = \bigvee_{\ell \neq i}^{s} \{-\gamma_{\ell j i}\}, \quad \text{and} \quad c^{**} = \bigwedge_{\ell=1}^{s} c_\ell^*, \end{aligned} \tag{3.12}$$

which will be used throughout the rest of the chapter.

Proposition 3.9. *Suppose* $\boldsymbol{\eta}_a \neq \boldsymbol{\eta}_b$ *for all* $a \neq b$. *Then*

(a) $R_i^{(m)}(\boldsymbol{W}_i) \overset{a.s.}{=} o(e^{-mc_i^*})$ *for* $c_i^* > 0$, *so that* $R_i^{(m)}(\boldsymbol{W}_i) \overset{a.s.}{\to} 0$ *as* $m \to \infty$.

(b) If $j \neq i$ *then for* $d_{ij}^* > 0$ *and* $\gamma_{ijj} > 0$,

$$O(e^{m\gamma_{ijj}}) \leq R_i^{(m)}(\boldsymbol{W}_j) \leq O(e^{md_{ij}^*}), \qquad \textit{almost surely, for all large } m.$$

As a consequence, $R_i^{(m)}(\boldsymbol{W}_j) \overset{a.s.}{\to} \infty$ *as* $m \to \infty$.

Proof. By the strong law of large numbers and continuity, we have that for almost any $\omega \in \Omega$ and any $\varepsilon > 0$, there exists an M_ω such that, for all $m \geq M_\omega$,

$$\begin{aligned} &\left| (\boldsymbol{\eta}_\ell - \boldsymbol{\eta}_i)^T[-a'(\boldsymbol{\eta}_j)] - (\boldsymbol{\eta}_\ell - \boldsymbol{\eta}_i)^T \boldsymbol{W}_j(\omega)/m \right| < \varepsilon \\ &\quad \iff -a'(\boldsymbol{\eta}_j)^T(\boldsymbol{\eta}_\ell - \boldsymbol{\eta}_i) - \varepsilon < (\boldsymbol{\eta}_\ell - \boldsymbol{\eta}_i)^T \boldsymbol{W}_j(\omega)/m < -a'(\boldsymbol{\eta}_j)^T(\boldsymbol{\eta}_\ell - \boldsymbol{\eta}_i) + \varepsilon. \end{aligned}$$

This implies that $\forall m \geq M_\omega$

$$\begin{aligned} R_i^{(m)}(\boldsymbol{W}_j(\omega)) &\leq \sum_{\ell \neq i}^{s} \pi_\ell \exp\Big\{ m\Big[-a'(\boldsymbol{\eta}_j)^T(\boldsymbol{\eta}_\ell - \boldsymbol{\eta}_i) + [a(\boldsymbol{\eta}_\ell) - a(\boldsymbol{\eta}_i)] + \varepsilon \Big]\Big\} \\ &= \sum_{\ell \neq i}^{s} \pi_\ell \exp\{m\,(-\gamma_{\ell j i} + \varepsilon)\}, \end{aligned}$$

and

$$\begin{aligned} R_i^{(m)}(\boldsymbol{W}_j(\omega)) &\geq \sum_{\ell \neq i}^{s} \pi_\ell \exp\Big\{ m\Big[-a'(\boldsymbol{\eta}_j)^T(\boldsymbol{\eta}_\ell - \boldsymbol{\eta}_i) + [a(\boldsymbol{\eta}_\ell) - a(\boldsymbol{\eta}_i)] - \varepsilon \Big]\Big\} \\ &= \sum_{\ell \neq i}^{s} \pi_\ell \exp\{m\,(-\gamma_{\ell j i} - \varepsilon)\}. \end{aligned}$$

Case (a). Suppose $j = i$. From Lemma 3.8 we have

$$\gamma_{\ell i i} = a'(\boldsymbol{\eta}_i)^T(\boldsymbol{\eta}_\ell - \boldsymbol{\eta}_i) - [a(\boldsymbol{\eta}_\ell) - a(\boldsymbol{\eta}_i)] > 0$$

for all $\ell \neq i$, so that for $m \geq M_\omega$,

$$\begin{aligned} 0 &\leq R_i^{(m)}(\boldsymbol{W}_i(\omega)) \\ &\leq \sum_{\ell \neq i}^{s} \pi_\ell e^{m(-\gamma_{\ell i i} + \varepsilon)} \\ &= \sum_{\ell \neq i}^{s} \pi_\ell e^{-m(\gamma_{\ell i i} - \varepsilon)} \\ &\leq e^{-m(c_i^* - \varepsilon)} \sum_{\ell \neq i}^{s} \pi_\ell \\ &\leq e^{-m(c_i^* - \varepsilon)} \\ &\to 0, \qquad \text{as } m \to \infty. \end{aligned}$$

Note that $c_i^* > \varepsilon$ when $\varepsilon > 0$ is taken arbitrarily small. Since this holds for almost every $\omega \in \Omega$, we have $R_i^{(m)}(\boldsymbol{W}_i) \overset{a.s.}{\to} 0$ and $R_i^{(m)}(\boldsymbol{W}_i) \overset{a.s.}{=} o(e^{-mc_i^*})$.

Case (b). Now suppose $j \neq i$. Consider for $\ell = 1, \ldots, s$,

$$-\gamma_{\ell ji} = -a'(\boldsymbol{\eta}_j)^T(\boldsymbol{\eta}_\ell - \boldsymbol{\eta}_i) + [a(\boldsymbol{\eta}_\ell) - a(\boldsymbol{\eta}_i)].$$

Notice that

$$\begin{aligned}-\gamma_{jji} &= -a'(\boldsymbol{\eta}_j)^T(\boldsymbol{\eta}_j - \boldsymbol{\eta}_i) + [a(\boldsymbol{\eta}_j) - a(\boldsymbol{\eta}_i)] \\ &= a'(\boldsymbol{\eta}_j)^T(\boldsymbol{\eta}_i - \boldsymbol{\eta}_j) - [a(\boldsymbol{\eta}_i) - a(\boldsymbol{\eta}_j)] \\ &= \gamma_{ijj},\end{aligned}$$

where $\gamma_{ijj} > 0$ by Lemma 3.8. Then for $m \geq M_\omega$,

$$R_i^{(m)}(\boldsymbol{W}_j(\omega)) \geq \sum_{\ell \neq i}^{s} \pi_\ell e^{m(-\gamma_{\ell ji} - \varepsilon)} \geq \pi_j e^{m(-\gamma_{jji} - \varepsilon)} = \pi_j e^{m(\gamma_{ijj} - \varepsilon)} \to \infty, \tag{3.13}$$

as $m \to \infty$, since $\gamma_{ijj} - \varepsilon > 0$ for arbitrarily small $\varepsilon > 0$. Therefore $R_i^{(m)}(\boldsymbol{W}_j) \overset{a.s.}{\to} \infty$. We can also obtain an upper bound using

$$R_i^{(m)}(\boldsymbol{W}_j(\omega)) \leq \sum_{\ell \neq i}^{s} \pi_\ell e^{m(-\gamma_{\ell ji} + \varepsilon)} \leq \sum_{\ell \neq i}^{s} \pi_\ell e^{m(d_{ij}^* + \varepsilon)} \leq e^{m(d_{ij}^* + \varepsilon)}, \tag{3.14}$$

noting that

$$d_{ij}^* = \bigvee_{\ell \neq i}^{s} \{-\gamma_{\ell ji}\} \geq -\gamma_{jji} = \gamma_{ijj} > 0.$$

We have therefore found the upper and lower bounds

$$\pi_j e^{m(\gamma_{ijj} - \varepsilon)} \leq R_i^{(m)}(\boldsymbol{W}_j(\omega)) \leq e^{m(d_{ij}^* + \varepsilon)}, \qquad \forall m \geq M_\omega,$$

and hence the desired almost sure bounds

$$\pi_j e^{m\gamma_{ijj}} \leq R_i^{(m)}(\boldsymbol{W}_j) \leq e^{md_{ij}^*}, \qquad \text{for all large } m.$$

are obtained. $\square$

Proposition 3.10. *Suppose $\boldsymbol{\eta}_a \neq \boldsymbol{\eta}_b$ for all $a \neq b$. Then*

(a) $1 - Q_i^{(m)}(\boldsymbol{W}_i) \overset{a.s.}{=} O(e^{-mc_i^*})$*, so that* $Q_i^{(m)}(\boldsymbol{W}_i) \overset{a.s.}{\to} 1$ *as* $m \to \infty$*,*

(b) If $j \neq i$ *then* $Q_i^{(m)}(\boldsymbol{W}_j) \overset{a.s.}{=} O(e^{-m\gamma_{ijj}})$*, so that* $Q_i^{(m)}(\boldsymbol{W}_j) \overset{a.s.}{\to} 0$ *as* $m \to \infty$*,*

with c_i^* *and* γ_{ijj} *as defined in* (3.12).

Proof. Notice that

$$\begin{aligned} Q_i^{(m)}(\boldsymbol{t}) &= \frac{\pi_i \exp\{\boldsymbol{\eta}_i^T \boldsymbol{t} + ma(\boldsymbol{\eta}_i)\}}{\sum_{\ell=1}^{s} \pi_\ell \exp\{\boldsymbol{\eta}_\ell^T \boldsymbol{t} + ma(\boldsymbol{\eta}_\ell)\}} \\ &= \frac{\pi_i}{\pi_i + \sum_{\ell \neq i}^{s} \pi_\ell \exp\{(\boldsymbol{\eta}_\ell - \boldsymbol{\eta}_i)^T \boldsymbol{t} + m[a(\boldsymbol{\eta}_\ell) - (\boldsymbol{\eta}_\ell)]\}} \\ &= \frac{\pi_i}{\pi_i + R_i^{(m)}(\boldsymbol{t})} \end{aligned} \tag{3.15}$$

Now apply Proposition 3.9 to obtain the limits. To obtain the rates, first take $\boldsymbol{T} = \boldsymbol{W}_i$, and notice that

$$1 - Q_i^{(m)}(\boldsymbol{W}_i) = 1 - \frac{\pi_i}{\pi_i + R_i^{(m)}(\boldsymbol{W}_i)} = \frac{1}{\pi_i \left[R_i^{(m)}(\boldsymbol{W}_i)\right]^{-1} + 1}$$

Since $R_i^{(m)}(\boldsymbol{W}_i) \overset{a.s.}{=} O(e^{-mc_i^*})$ by Proposition 3.9, there exists a constant K such that

$$\begin{aligned} &\left| \frac{R_i^{(m)}(\boldsymbol{t})}{e^{-mc_i^*}} \right| < K, \\ \iff &R_i^{(m)}(\boldsymbol{W}_i) < Ke^{-mc_i^*} \\ \iff &\left[R_i^{(m)}(\boldsymbol{W}_i)\right]^{-1} > K^{-1} e^{mc_i^*}, \end{aligned}$$

almost surely, for all m large. Then we have

$$\begin{aligned} e^{mc_i^*}\left[1-Q_i^{(m)}(\boldsymbol{W}_i)\right] &\leq \frac{e^{mc_i^*}}{\pi_i K^{-1} e^{mc_i^*}+1}, \quad \text{almost surely, for all } m \text{ large} \\ &\to \frac{K}{\pi_i}, \quad \text{as } m\to\infty, \end{aligned}$$

and so we have the result $1-Q_i^{(m)}(\boldsymbol{t}) = O(e^{-mc_i^*})$.

Now take $\boldsymbol{T}=\boldsymbol{W}_j$ for $j\neq i$. Notice that

$$Q_i^{(m)}(\boldsymbol{W}_j) = \frac{\pi_i}{\pi_i + R_i^{(m)}(\boldsymbol{W}_j)},$$

and Proposition 3.9 gives that

$$R_i^{(m)}(\boldsymbol{W}_j) \geq e^{m\gamma_{ijj}}, \quad \text{almost surely for all large } m.$$

Then we have

$$\begin{aligned} e^{m\gamma_{ijj}} Q_i^{(m)}(\boldsymbol{W}_j) &= \frac{\pi_i e^{m\gamma_{ijj}}}{\pi_i + R_i^{(m)}(\boldsymbol{W}_j)} \\ &\leq \frac{\pi_i e^{m\gamma_{ijj}}}{\pi_i + O(e^{m\gamma_{ijj}})}, \quad \text{almost surely for all large } m, \end{aligned}$$

which converges to a constant as $m\to\infty$. Then we have the result $Q_i^{(m)}(\boldsymbol{W}_j) \overset{a.s.}{=} O(e^{-m\gamma_{ijj}})$. $\square$

Proposition 3.10 suggests that the convergence between the exact and approximate information will be fast when both of the following happen quickly as m is increased: (1) the posterior probability of being in the ℓth subpopulation goes to 1 when the true subpopulation $Z=\ell$, and (2) the posterior probability of being in the ℓth subpopulation goes to 0 when the true subpopulation $Z\neq\ell$. It is clear from Proposition 3.10 and dominated convergence that the expectation (3.10) converges to zero. We also note that

$W_{ij} - \mathrm{E}(W_{ij})$ is a sum of independent and identically distributed random variables, so that $[W_{ij} - \mathrm{E}(W_{ij})]^2 = O(m^2)$, and therefore

$$\pi_i^2 \left[1 - Q_i^{(m)}(\boldsymbol{W}_i)\right] \left[W_{ij} - \mathrm{E}(W_{ij})\right]^2 \overset{a.s.}{=} O(m^2 e^{-mc_i^*}).$$

Therefore, its expectation (3.8) converges to zero if and only if the sequence

$$[1 - Q_i^{(m)}(\boldsymbol{W}_i)][W_{ij} - \mathrm{E}(W_{ij})]^2, \quad m = 1, 2, \dots \tag{3.16}$$

is uniformly integrable (Resnick, 1999, chapter 6). The convergence of $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})$ can therefore be characterized in the following theorem.

Theorem 3.11. $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta}) \to 0$ *as* $m \to \infty$ *if and only if the sequence* (3.16) *is uniformly integrable for each* $i = 1, \dots, s$.

Some additional work will allow us to conclude convergence of the expectations without needing to verify uniform integrability, and also to obtain rates.

Lemma 3.12. $\mathrm{E}\left[Q_i^{(m)}(\boldsymbol{W}_i)\right] = 1 - O(e^{-mc_i^*})$ *with* c_i^* *defined as in* (3.12).

Proof. From the Markov inequality we have,

$$\mathrm{P}\left(Q_i^{(m)}(\boldsymbol{W}_i) \geq \varepsilon\right) \leq \frac{\mathrm{E}\left[Q_i^{(m)}(\boldsymbol{W}_i)\right]}{\varepsilon} \leq \frac{1}{\varepsilon}, \quad \text{for any } \varepsilon > 0,$$

recalling that $0 \leq Q_i^{(m)}(\boldsymbol{W}_i) \leq 1$. Equivalently,

$$\varepsilon \, \mathrm{P}\left(Q_i^{(m)}(\boldsymbol{W}_i) \geq \varepsilon\right) \leq \mathrm{E}\left[Q_i^{(m)}(\boldsymbol{W}_i)\right] \leq 1.$$

Proposition 3.10 gives that $Q_i^{(m)}(\boldsymbol{W}_i) \overset{a.s.}{=} 1 - O(e^{-mc_i^*})$, which implies

$$\mathrm{P}\left(Q_i^{(m)}(\boldsymbol{W}_i) \geq \varepsilon\right) = 1 - O(e^{-mc_i^*}),$$

assuming that $0 < \varepsilon < 1$. Therefore

$$\varepsilon\left[1 - O(e^{-mc_i^*})\right] \le \mathrm{E}\left[Q_i^{(m)}(\boldsymbol{W}_i)\right] \le 1.$$

Taking $\varepsilon < 1$ arbitrarily close to 1 gives the result. □

Lemma 3.13. *Let $S_n = X_1 + \cdots + X_n$ where $\{X_i\}$ are independent and identically distributed and $\mathrm{E}(|X_1|^k) < \infty$ for a given positive integer $k \ge 0$. Then $\mathrm{E}(S_n^k) = O(n^k)$.*

Proof. Notice that

$$\begin{aligned}
\mathrm{E}(S_n^k) &= \mathrm{E}[(X_1 + \cdots + X_n)^k] \\
&= \mathrm{E}\left[\sum_{\boldsymbol{z}\in\Omega_{n,k}} \frac{k!}{z_1!\cdots z_n!} X_1^{z_1}\cdots X_n^{z_n}\right] \\
&= \sum_{\boldsymbol{z}\in\Omega_{n,k}} \frac{k!}{z_1!\cdots z_n!}\,\mathrm{E}[X_1^{z_1}]\cdots\mathrm{E}[X_n^{z_n}] \\
&= \sum_{\boldsymbol{z}\in\Omega_{n,k}} \frac{k!}{z_1!\cdots z_n!}\,\mathrm{E}[X_1^{z_1}]\cdots\mathrm{E}[X_1^{z_n}]
\end{aligned}$$

where $\Omega_{n,k}$ is the multinomial sample space with n categories and k trials. Let

$$\xi = \max_{\boldsymbol{z}\in\Omega_{n,k}} \left|\mathrm{E}[X_1^{z_1}]\cdots\mathrm{E}[X_1^{z_n}]\right|$$

and note that $\xi \ge 0$ is finite since the expression involves only moments of X_1 up to order k, which are all assumed to be finite. Now we have

$$\left|\mathrm{E}(S_n^k)\right| \le \xi \sum_{\boldsymbol{z}\in\Omega_{n,k}} \frac{k!}{z_1!\cdots z_n!} = \xi n^k$$

which gives the result. □

The following theorem gives rates for the diagonal elements of the matrix $\widetilde{\mathcal{I}}(\boldsymbol{\theta}) - \mathcal{I}(\boldsymbol{\theta})$, which dominate the other elements of the matrix. We require that fourth moments are finite for all components of the original $\boldsymbol{X}_i$ given $Z = \ell$ for $\ell = 1, \ldots, s$. But this does not represent any additional restriction; an Exponential family of full rank has a moment generating function which is finite in a neighborhood of zero (Shao, 2008, Theorem 2.1), therefore all moments exist.

Theorem 3.14. *Consider the matrix* $\widetilde{\mathcal{I}}(\boldsymbol{\theta}) - \mathcal{I}(\boldsymbol{\theta})$*;*

(a) For the j*th diagonal element of the* i*th diagonal block,*

$$\boldsymbol{e}_j^T \left(\pi_i \boldsymbol{F}_i - \boldsymbol{A}_{ii}\right) \boldsymbol{e}_j = O(m^2 e^{-\frac{m}{2} c_i^*}),$$

provided that $\mathrm{E}[|X_{1j}|^4 \mid Z = i] < \infty$.

(b) For the j*th diagonal element of the* $\boldsymbol{\pi}$ *diagonal block,*

$$\boldsymbol{e}_j^T \left(\boldsymbol{F}_\pi - \boldsymbol{A}_{\pi\pi}\right) \boldsymbol{e}_j = O(e^{-mc_j^*}) + O(e^{-mc_s^*}) + O(e^{-m\gamma_{jss}}), \quad j = 1, \ldots, s-1$$

Proof. For (a) we have

$$\begin{aligned}
&\pi_i^2 \, \mathrm{E}\left\{\left[1 - Q_i^{(m)}(\boldsymbol{W}_i)\right]\left[W_{ij} - \mathrm{E}(W_{ij})\right]^2\right\} \\
&\leq \pi_i^2 \sqrt{\mathrm{E}\left[\left(1 - Q_i^{(m)}(\boldsymbol{W}_i)\right)^2\right]} \sqrt{\mathrm{E}\left[\left(W_{ij} - \mathrm{E}(W_{ij})\right)^4\right]} \\
&\quad \text{(by Cauchy-Schwarz inequality)} \\
&\leq \pi_i^2 \sqrt{\mathrm{E}\left[1 - Q_i^{(m)}(\boldsymbol{W}_i)\right]} \sqrt{\mathrm{E}\left[\left(W_{ij} - \mathrm{E}(W_{ij})\right)^4\right]} \\
&\quad \text{(since } 0 \leq X \leq 1 \implies X^2 \leq X \implies \mathrm{E}(X^2) \leq \mathrm{E}(X)\text{)} \\
&= \pi_i^2 \left\{O(e^{-mc_i^*}) O(m^4)\right\}^{1/2} \\
&\quad \text{(by Corollary 3.12 and Lemma 3.13)} \\
&= O(m^2 e^{-\frac{m}{2} c_i^*})
\end{aligned}$$

For (b), use Proposition 3.10 with the expectation (3.10) to obtain

$$\begin{aligned}(\pi_j^{-1}+\pi_s^{-1}) - \pi_j^{-1}\,\mathrm{E}\left[Q_j^{(m)}(\boldsymbol{W}_j)\right] - \pi_s^{-1}\,\mathrm{E}\left[Q_s^{(m)}(\boldsymbol{W}_s)\right] + 2\pi_j^{-1}\,\mathrm{E}\left[Q_j^{(m)}(\boldsymbol{W}_s)\right] \\ = \pi_j^{-1}O(e^{-mc_j^*}) + \pi_s^{-1}O(e^{-mc_s^*}) + 2\pi_j^{-1}O(e^{-m\gamma_{jss}}).\end{aligned}$$

$\square$

Note that the rates obtained in Theorem 3.14 match Corollary 2.6 for the multinomial case.

Remark 3.15. Hölder's Inequality can be used to weaken the assumption of a finite fourth moment in Theorem 3.14, at the cost of a slower exponential rate in the bound. Suppose $u, v > 1$ such that $1/u + 1/v = 1$; then $v = u/(u-1)$ and Hölder's Inequality gives

$$\begin{aligned}&\pi_i^2\,\mathrm{E}\left\{\left[1-Q_i^{(m)}(\boldsymbol{W}_i)\right]\left[W_{ij}-\mathrm{E}(W_{ij})\right]^2\right\} \\ &\quad\le \pi_i^2\left\{\mathrm{E}\left[\left(1-Q_i^{(m)}(\boldsymbol{W}_i)\right)^u\right]\right\}^{1/u}\left\{\mathrm{E}\left[\left(W_{ij}-\mathrm{E}(W_{ij})\right)^{\frac{2u}{u-1}}\right]\right\}^{\frac{u-1}{u}}.\end{aligned}$$

Now u can be taken arbitrarily large so that $u/(u-1) < 1+\varepsilon$ for any $\varepsilon > 0$. Then we have

$$\begin{aligned}&\pi_i^2\,\mathrm{E}\left\{\left[1-Q_i^{(m)}(\boldsymbol{W}_i)\right]\left[W_{ij}-\mathrm{E}(W_{ij})\right]^2\right\} \\ &\quad\le \pi_i^2\left\{\mathrm{E}\left[\left(1-Q_i^{(m)}(\boldsymbol{W}_i)\right)^u\right]\right\}^{1/u}\left\{\mathrm{E}\left[\left(W_{ij}-\mathrm{E}(W_{ij})\right)^{2(1+\varepsilon)}\right]\right\}^{\frac{u-1}{u}} \\ &\quad\le \pi_i^2\left\{\mathrm{E}\left[1-Q_i^{(m)}(\boldsymbol{W}_i)\right]\right\}^{1/u}\left\{\mathrm{E}\left[\left(W_{ij}-\mathrm{E}(W_{ij})\right)^{2(1+\varepsilon)}\right]\right\}^{\frac{u-1}{u}} \\ &\quad= O(e^{-\frac{m}{u}c^*})\left\{\mathrm{E}\left[\left(W_{ij}-\mathrm{E}(W_{ij})\right)^{2(1+\varepsilon)}\right]\right\}^{\frac{u-1}{u}}.\end{aligned}$$

Therefore, the moment assumption can be weakened significantly, although the upper bound $O(e^{-\frac{m}{u}c_i^*})$ will become increasingly less useful as u is taken to be larger. In the case of the full rank Exponential family we have all moments, and there is no need to weaken the assumptions.

Because of the convenient block-diagonal form of the information matrix approximation, its inverse $\widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta}) = \text{Blockdiag}(\pi_1^{-1}\boldsymbol{F}_1^{-1}, \ldots, \pi_s^{-1}\boldsymbol{F}_s^{-1}, \boldsymbol{F}_\pi^{-1})$ is also block-diagonal. As in Chapter 2, the convergence result for $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})$ can be used to show convergence between the inverses.

Lemma 3.16. *Suppose $\mathcal{I}_m(\boldsymbol{\theta})$ and $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta})$ are nonsingular. Then $\mathcal{I}_m^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta}) \to \mathbf{0}$ as $m \to \infty$.*

Proof. The proof is adapted from the proof of Theorem 2.10. Notice here that

$$\begin{aligned}
\|\widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta})\|_F^2 &= \sum_{\ell=1}^{s} \|\pi_\ell^{-1}\boldsymbol{F}_\ell^{-1}\|_F^2 + \|\boldsymbol{F}_\pi^{-1}\|_F^2 \\
&= \sum_{\ell=1}^{s} m^{-2}\pi_\ell^{-2}\|\widetilde{\mathcal{I}}_1^{-1}(\boldsymbol{\eta}_\ell)\|_F^2 + \|\boldsymbol{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}^T\|_F^2 \\
&= \|\boldsymbol{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}^T\|_F^2 + O(m^{-2})
\end{aligned}$$

where $\widetilde{\mathcal{I}}_1(\boldsymbol{\eta}_\ell) = \text{Var}(\boldsymbol{U}_1 \mid Z = \ell)$, as obtained in (3.2), is free of m. Let $c^{**} = \bigwedge_{i=1}^{s} c_i^*$, and use Proposition 3.6 and Theorem 3.14 to obtain the simple bound

$$\|\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})\|_F^2 = q^2 O(m^2 e^{-\frac{m}{2}c^{**}}).$$

By the same argument as in Theorem 2.10, $\|\mathcal{I}_m^{-1}(\boldsymbol{\theta})\|_F = O(1)$. We then have

$$\begin{aligned}
\|\mathcal{I}_m^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta})\|_F &\leq \|\mathcal{I}_m^{-1}(\boldsymbol{\theta})\|_F \cdot \|\widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta})\|_F \cdot \|\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})\|_F \\
&= O(1) \cdot \left\{\|\boldsymbol{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}^T\|_F^2 + O(m^{-2})\right\} \cdot \left\{q^2 O(m^2 e^{-\frac{m}{2}c^{**}})\right\}^{1/2},
\end{aligned}$$

which gives the result. □

3.4 Relationship to Classification Problem

There is a fundamental connection between the convergence behavior of $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})$ and the probability of misclassification using an optimal rule. Namely, both depend on the separation between subpopulations in a similar way. Suppose that there are s subpopulations with densities $f(\boldsymbol{x} \mid \boldsymbol{\phi}_1), \ldots, f(\boldsymbol{x} \mid \boldsymbol{\phi}_s)$ from an exponential family, which occur in the overall population in respective proportions $\pi_1, \ldots, \pi_s$. Now let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ be independently and identically distributed from subpopulation $Z = j$, but Z is not observed. We consider classification rules on $\boldsymbol{T} = \sum_{i=1}^m \boldsymbol{U}(\boldsymbol{X}_i)$ which is sufficient given Z. The classification problem is to specify a rule, described by regions $\boldsymbol{\mathcal{D}} = \{\mathcal{D}_1, \ldots, \mathcal{D}_s\}$ which partition the space $\mathcal{T}$ of $\boldsymbol{T}$ so that

$$\boldsymbol{T} \in \mathcal{D}_\ell \iff \boldsymbol{T} \text{ belongs to } \ell\text{th subpopulation.}$$

One objective is to specify a rule $\boldsymbol{\mathcal{D}}$ which minimizes the probability of misclassification $p(\boldsymbol{\mathcal{D}})$. (Another may be to minimize the cost of misclassification, if the possible misclassifications are assigned different costs). It is well-known (Anderson, 2003) that the rule $\boldsymbol{\mathcal{D}}^* = \{\mathcal{D}_1^*, \ldots, \mathcal{D}_s^*\}$ using

$$\mathcal{D}_\ell^* = \left\{ \boldsymbol{t} \in \mathcal{T} : \ell = \underset{a}{\operatorname{argmax}}\, \pi_a f(\boldsymbol{t} \mid \boldsymbol{\phi}_a) \right\}$$

minimizes $p(\boldsymbol{\mathcal{D}})$. Using this optimal rule, we may compute

$$\begin{aligned}
p(\boldsymbol{\mathcal{D}}^*) &= \sum_{\ell=1}^{s} \mathrm{P}(\boldsymbol{T} \notin \mathcal{D}_\ell^* \mid Z = \ell)\, \mathrm{P}(Z = \ell) \\
&= \sum_{\ell=1}^{s} \pi_\ell\, \mathrm{P}\left(\bigcup_{j \neq \ell} [\boldsymbol{T} \in \mathcal{D}_j^*] \,\middle|\, Z = \ell \right) \\
&= \sum_{\ell=1}^{s} \pi_\ell\, \mathrm{P}\left(\bigcup_{j \neq \ell} [\pi_j f(\boldsymbol{T} \mid \boldsymbol{\phi}_j) \geq \pi_\ell f(\boldsymbol{T} \mid \boldsymbol{\phi}_\ell)] \,\middle|\, Z = \ell \right) \\
&\leq \sum_{\ell=1}^{s} \pi_\ell\, \mathrm{P}\left(\sum_{j \neq \ell} \pi_j f(\boldsymbol{T} \mid \boldsymbol{\phi}_j) \geq \pi_\ell f(\boldsymbol{T} \mid \boldsymbol{\phi}_\ell) \,\middle|\, Z = \ell \right) \\
&= \sum_{\ell=1}^{s} \pi_\ell\, \mathrm{P}\left(R_\ell^{(m)}(\boldsymbol{W}_\ell) \geq \pi_\ell \right).
\end{aligned}$$

The optimal probability of misclassification $p(\boldsymbol{\mathcal{D}}^*)$ provides an objective measurement on the degree of separation between the s subpopulations; a higher probability indicates that it is more difficult to distinguish among them. However, the rule $\mathcal{D}^*$ can only be applied when all $\boldsymbol{\phi}_\ell$ and $\boldsymbol{\pi}$ are known. Recall that $R_\ell^{(m)}(\boldsymbol{W}_\ell) \overset{a.s.}{=} o(e^{-mc_i^*})$, so that we obtain $p(\boldsymbol{\mathcal{D}}^*) = o(e^{-mc_i^*})$ where c_i^* was defined in (3.12). Therefore, collection of additional observations for $\boldsymbol{T} = \sum_{i=1}^{m} \boldsymbol{U}(\boldsymbol{X}_i)$ may drastically improve $p(\boldsymbol{\mathcal{D}}^*)$ if c_i^* is large, and has almost no effect when c_i^* is very small.

To see the relationship between $p(\boldsymbol{\mathcal{D}}^*)$ and the convergence rate of the approximate information matrix, notice that

$$\begin{aligned}
&\mathrm{P}\left(R_\ell^{(m)}(\boldsymbol{W}_\ell) \leq \pi_\ell \right) \\
&\qquad = \lim_{\varepsilon \uparrow 1} \varepsilon\, \mathrm{P}\left[R_\ell^{(m)}(\boldsymbol{W}_\ell) \leq \pi_\ell \left(\frac{1}{\varepsilon} - 1 \right) \right] \\
&\qquad = \lim_{\varepsilon \uparrow 1} \varepsilon\, \mathrm{P}\left(Q_\ell^{(m)}(\boldsymbol{W}_\ell) \geq \varepsilon \right) \\
&\qquad \leq \mathrm{E}\left[Q_\ell^{(m)}(\boldsymbol{W}_\ell) \right] \\
\iff &\ \mathrm{E}\left[1 - Q_\ell^{(m)}(\boldsymbol{W}_\ell) \right] \leq \mathrm{P}\left(R_\ell^{(m)}(\boldsymbol{W}_\ell) \geq \pi_\ell \right)
\end{aligned}$$

so that $\mathrm{P}\left(R_\ell^{(m)}(\boldsymbol{W}_\ell) \geq \pi_\ell\right)$ gives an upper bound on the probability of misclassifying $\boldsymbol{T}$ when $Z = \ell$. Recall that the convergence rate of the ℓth block of $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})$ depends on $\mathrm{E}\left[1 - Q_\ell^{(m)}(\boldsymbol{W}_\ell)\right]$, as in the proof of Theorem 3.14. Proposition 3.9 gives

$$\mathrm{P}\left(R_\ell^{(m)}(\boldsymbol{W}_\ell) \geq \pi_\ell\right) \leq \mathrm{P}\left(O(e^{-mc_\ell^*}) \geq \pi_\ell\right) = O(e^{-mc_\ell^*})$$

so that

$$p(\boldsymbol{\mathcal{D}}^*) \leq \sum_{\ell=1}^{s} \pi_\ell O(e^{-mc_\ell^*}).$$

3.5 Examples

Example 3.17 (Multinomial Populations)**.** The multinomial case was the focus of Chapter 2. To apply the more general results from this chapter, let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m \mid Z = j$ be independent and identically distributed as $\text{Mult}_{k+1}(1, \boldsymbol{p}_j)$, with $Z \sim \text{Discrete}(1, \ldots, s; \boldsymbol{\pi})$. Let $\boldsymbol{T} = \sum_{i=1}^m \boldsymbol{X}_i$. Recall that the multinomial subpopulations are exponential families with

$$\begin{aligned} f(\boldsymbol{t} \mid m, \boldsymbol{p}_\ell) &= \exp\left\{\log \frac{m!}{t_1! \cdots t_{k+1}!} + \sum_{a=1}^{k+1} t_a \log p_{\ell a}\right\} \\ &= \exp\left\{\log \frac{m!}{t_1! \cdots t_{k+1}!} + \sum_{a=1}^{k} t_a \log p_{\ell a} + \left(m - \sum_{a=1}^{k} t_a\right) \log p_{\ell,k+1}\right\} \\ &= \exp\left\{\log \frac{m!}{t_1! \cdots t_{k+1}!} + \sum_{a=1}^{k} t_a \log \frac{p_{\ell a}}{p_{\ell,k+1}} + m \log p_{\ell,k+1}\right\}, \end{aligned}$$

where $p_{\ell,k+1} = 1 - \sum_{a=1}^k p_{\ell a}$. The natural parameter is then

$$\boldsymbol{\eta}_j = \left(\log \frac{p_{\ell 1}}{p_{\ell,k+1}}, \cdots, \log \frac{p_{\ell k}}{p_{\ell,k+1}}\right).$$

The approximate information matrix with respect to $\boldsymbol{\theta} = (\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_s, \boldsymbol{\pi})$ is then

$$\widetilde{\mathcal{I}}(\boldsymbol{\theta}) = \text{Blockdiag}(\pi_1 \boldsymbol{F}_1, \dots, \pi_s \boldsymbol{F}_s, \boldsymbol{F}_\pi)$$

where $\boldsymbol{F}_\ell = m \operatorname{Var}(\boldsymbol{U}_1) = m\{\text{Diag}(\boldsymbol{p}_\ell) - \boldsymbol{p}_\ell \boldsymbol{p}_\ell^T\}$. It can also be shown that $\partial \boldsymbol{\eta}_\ell / \partial \boldsymbol{p}_\ell = \text{Diag}(\boldsymbol{p}_\ell)^{-1} + p_{\ell,k+1}^{-1} \mathbf{1}\mathbf{1}^T$, so that

$$\begin{aligned}
\widetilde{\mathcal{I}}(\boldsymbol{p}_\ell) &= \left(\frac{\partial \boldsymbol{\eta}_\ell}{\partial \boldsymbol{p}_\ell} \right) \widetilde{\mathcal{I}}(\boldsymbol{\eta}_\ell) \left(\frac{\partial \boldsymbol{\eta}_\ell}{\partial \boldsymbol{p}_\ell} \right)^T \\
&= \left\{ \text{Diag}(\boldsymbol{p}_\ell)^{-1} + p_{\ell,k+1}^{-1} \mathbf{1}\mathbf{1}^T \right\} m\{\text{Diag}(\boldsymbol{p}_\ell) - \boldsymbol{p}_\ell \boldsymbol{p}_\ell^T\} \left\{ \text{Diag}(\boldsymbol{p}_\ell)^{-1} + p_{\ell,k+1}^{-1} \mathbf{1}\mathbf{1}^T \right\} \\
&= m \left\{ \text{Diag}(\boldsymbol{p}_\ell)^{-1} + p_{\ell,k+1}^{-1} \mathbf{1}\mathbf{1}^T \right\}.
\end{aligned}$$

Therefore we obtain the form of $\widetilde{\mathcal{I}}(\boldsymbol{\psi})$ with respect to the parameter $\boldsymbol{\psi} = (\boldsymbol{p}_1, \dots, \boldsymbol{p}_s, \boldsymbol{\pi})$, which was studied in Chapter 2.

Example 3.18 (Normal Populations)**.** Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_m \mid Z = j$ be independent and identically distributed in $\mathbb{R}^k$ as $\text{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, with $Z \sim \text{Discrete}(1, \dots, s; \boldsymbol{\pi})$, so that $\boldsymbol{T} = \sum_{i=1}^m \boldsymbol{X}_i \mid Z = j \sim \text{N}(m\boldsymbol{\mu}_j, m\boldsymbol{\Sigma})$. Let us compare the approximate and exact information matrices with respect to $\boldsymbol{\psi} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_s, \boldsymbol{\pi})$ where $\boldsymbol{\Sigma}$ is taken to be known for the sake of demonstration. Recall that the Normal subpopulations are exponential families with

$$\begin{aligned}
f(\boldsymbol{t} \mid m\boldsymbol{\mu}_j, m\boldsymbol{\Sigma}) &= \exp\left\{ -\frac{1}{2}(\boldsymbol{t} - m\boldsymbol{\mu}_j)^T (m\boldsymbol{\Sigma})^{-1} (\boldsymbol{t} - m\boldsymbol{\mu}_j) - \frac{k}{2}\log(2\pi) - \frac{1}{2}\log|m\boldsymbol{\Sigma}| \right\} \\
&= \exp\Big\{ -\frac{1}{2}\left[\boldsymbol{t}^T (m\boldsymbol{\Sigma})^{-1} \boldsymbol{t} - 2(m\boldsymbol{\mu}_j)^T (m\boldsymbol{\Sigma})^{-1} \boldsymbol{t} + (m\boldsymbol{\mu}_j)^T (m\boldsymbol{\Sigma})^{-1} (m\boldsymbol{\mu}_j) \right] \\
&\qquad - \frac{k}{2}\log(2\pi) - \frac{1}{2}\log|m\boldsymbol{\Sigma}| \Big\} \\
&= \exp\left\{ \boldsymbol{\eta}_j^T \boldsymbol{t} + m a(\boldsymbol{\eta}_j) + h(\boldsymbol{t}) \right\}.
\end{aligned}$$

Here, $\boldsymbol{\eta}_j^T \boldsymbol{t} = (m\boldsymbol{\mu}_j)^T (m\boldsymbol{\Sigma})^{-1} \boldsymbol{t}$ and hence $\boldsymbol{\eta}_j = \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j$ is the natural parameter. We also have

$$ma(\boldsymbol{\eta_j}) = -\frac{1}{2}(m\boldsymbol{\mu}_j)^T(m\boldsymbol{\Sigma})^{-1}(m\boldsymbol{\mu}_j) = -m\frac{1}{2}\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j = -m\frac{1}{2}\boldsymbol{\eta}_j^T\boldsymbol{\Sigma}\boldsymbol{\eta}_j.$$

The first and second derivatives of the log-density under $Z = j$, with respect to $\boldsymbol{\eta}_j$, are

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\eta}_j} \log f(\boldsymbol{t} \mid \boldsymbol{\eta}_j) &= \frac{\partial}{\partial \boldsymbol{\eta}_j}\left\{\boldsymbol{\eta}_j^T\boldsymbol{t} + ma(\boldsymbol{\eta}_j)\right\} = \boldsymbol{t} - m\boldsymbol{\Sigma}\boldsymbol{\eta}_j, \\ -\frac{\partial^2}{\partial \boldsymbol{\eta}_j \partial \boldsymbol{\eta}_j^T} \log f(\boldsymbol{t} \mid \boldsymbol{\eta}_j) &= m\boldsymbol{\Sigma}, \end{aligned}$$

therefore the information contained in $\boldsymbol{\mu}_j$ in $\boldsymbol{T}$ under the jth subpopulation is given by

$$\mathcal{I}(\boldsymbol{\mu}_j) = \left(\frac{\partial \boldsymbol{\eta}_j}{\partial \boldsymbol{\mu}_j}\right)\mathcal{I}(\boldsymbol{\eta}_j)\left(\frac{\partial \boldsymbol{\eta}_j}{\partial \boldsymbol{\mu}_j}\right)^T = \boldsymbol{\Sigma}^{-1}(m\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1} = m\boldsymbol{\Sigma}^{-1}.$$

The approximate information matrix for the mixed population with respect to $\boldsymbol{\psi}$ is then

$$\widetilde{\mathcal{I}}(\boldsymbol{\psi}) = \text{Blockdiag}(\pi_1\boldsymbol{F}_1, \ldots, \pi_s\boldsymbol{F}_s, \boldsymbol{F}_\pi), \quad \text{with } \boldsymbol{F}_j = m\boldsymbol{\Sigma}^{-1} \text{ for } j = 1, \ldots, s,$$

and $\boldsymbol{F}_\pi = \boldsymbol{D}_\pi^{-1} + \pi_s^{-1}\mathbf{1}\mathbf{1}^T$. The exact information matrix will be computed numerically, using the `cubature` package[1] in `R` for multivariate integration. Let us concretely take the dimension $k = 2$ and the number of populations $s = 2$, with

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\pi} = \begin{pmatrix} 1/4 \\ 3/4 \end{pmatrix}.$$

Notice that for a mixture with $s = 2$ components, we have

$$\gamma_{111} = a'(\boldsymbol{\eta}_1)^T(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_1) - [a(\boldsymbol{\eta}_1) - a(\boldsymbol{\eta}_1)] = 0,$$

---

[1] http://cran.r-project.org/web/packages/cubature

and likewise $\gamma_{121} = \gamma_{212} = \gamma_{222} = 0$. We also have

$$\begin{aligned}\gamma_{112} &= a'(\boldsymbol{\eta}_1)^T(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2) - [a(\boldsymbol{\eta}_1) - a(\boldsymbol{\eta}_2)] \\ &= -a'(\boldsymbol{\eta}_1)^T(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1) + [a(\boldsymbol{\eta}_2) - a(\boldsymbol{\eta}_1)] \\ &= -\gamma_{211}\end{aligned}$$

and

$$\begin{aligned}\gamma_{221} &= a'(\boldsymbol{\eta}_2)^T(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1) - [a(\boldsymbol{\eta}_2) - a(\boldsymbol{\eta}_1)] \\ &= -a'(\boldsymbol{\eta}_2)^T(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2) + [a(\boldsymbol{\eta}_1) - a(\boldsymbol{\eta}_2)] \\ &= -\gamma_{122},\end{aligned}$$

where γ_{211} and γ_{122} are nonnegative by Lemma 3.8. Therefore, the numbers γ_{211} and γ_{122} together are sufficient to compute the orders for the convergence rates. We will consider three scenarios for the subpopulation means,

- Scenario 1: $\boldsymbol{\mu}_1 = (-1, 1)$, $\boldsymbol{\mu}_2 = (1, -1)$, so that $\gamma_{221} = \gamma_{122} = 8$.
- Scenario 2: $\boldsymbol{\mu}_1 = (-1/2, 1/2)$, $\boldsymbol{\mu}_2 = (1/2, -1/2)$, so that $\gamma_{221} = \gamma_{122} = 2$.
- Scenario 3: $\boldsymbol{\mu}_1 = (-1/8, 1/8)$, $\boldsymbol{\mu}_2 = (1/8, -1/8)$, so that $\gamma_{221} = \gamma_{122} = 1/8$.

Figure 3.1 plots the mixed populations for the three scenarios. The subpopulations are well-separated in Scenario 1, while in Scenario 2 there is only a small hint of separation, and in Scenario 3 the two groups are visually indistinguishable.

Table 3.1 compares the diagonal elements of $\widetilde{\mathcal{I}}_m(\boldsymbol{\psi})$ with those of $\mathcal{I}_m(\boldsymbol{\psi})$, where the latter have been computed numerically. Also shown is the Frobenius norm of the matrix $\widetilde{\mathcal{I}}_m(\boldsymbol{\psi}) - \mathcal{I}_m(\boldsymbol{\psi})$. Recall from the proof of Lemma 3.16, and by the transformation Remark 3.7, that

$$\|\widetilde{\mathcal{I}}_m(\boldsymbol{\psi}) - \mathcal{I}_m(\boldsymbol{\psi})\|_F^2 = q^2 O(m^2 e^{-\frac{m}{2}c^{**}}).$$

As expected, the approximate and exact matrix elements converge together quickly for Scenario 1, and more slowly for Scenario 2. For Scenario 3, the Frobenius norm at first increases with m because of the slow the convergence rate, and eventually begins decreasing when m is large. Figure 3.2 plots the norms from all three scenarios. The numerical integration sometimes produced inaccurate results which appear to be caused by the very large limits of integration we provided. For example, in Scenario 1 when $m = 8$ and in Scenario 2 when $m = 26$, $\mathcal{I}_{55} = 4.6667$ instead of the expected 5.3333. These results have been omitted from the tables and plots.

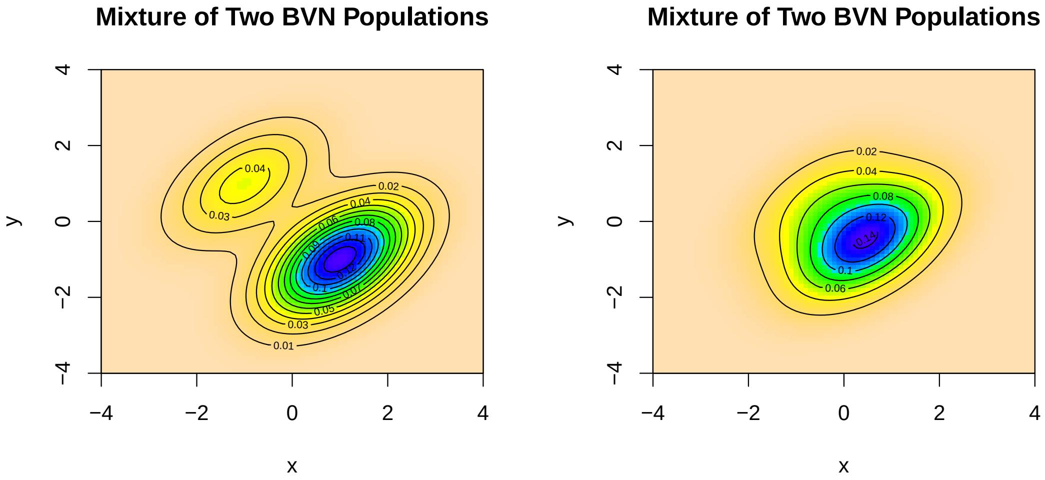
Example 3.19 (Common-Subpopulation Sampling vs. iid Sampling)**.** It is natural to ask if there is relationship between the information matrix of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ independently and identically distributed from $f(\boldsymbol{x} \mid \boldsymbol{\phi}_Z)$, but where Z is not observed, and the information matrix of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ independently and identically distributed from the finite mixture $f(\boldsymbol{x} \mid \boldsymbol{\theta})$. The convergence theory in this chapter was developed strictly for the former case. As a concrete example, suppose $X_1, \ldots, X_m$ are Normal random variables. Let $\mathcal{I}_m(\boldsymbol{\theta})$ denote the information matrix of $T = \sum_{i=1}^m X_i$, where $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_s, \pi_1, \ldots, \pi_{s-1})$ and

$$T \sim \sum_{\ell=1}^{s} \pi_\ell \frac{1}{\sqrt{2\pi m}} \exp\left\{ -\frac{1}{2m}(t - m\mu_\ell)^2 \right\}.$$

On the other hand, if X_i are iid from density

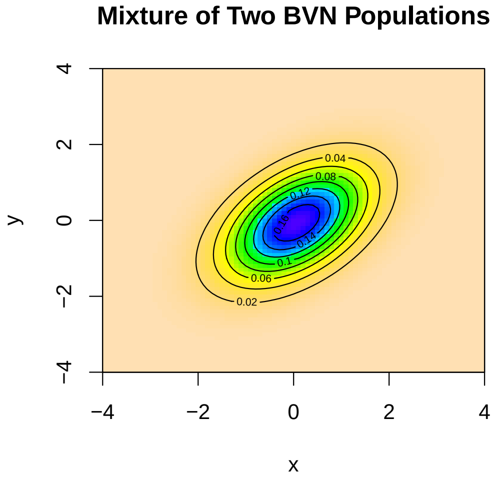
$$f(x \mid \boldsymbol{\theta}) = \sum_{\ell=1}^{s} \pi_\ell \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}(x - \mu_\ell)^2 \right\},$$

then the information matrix is $m\mathcal{I}_1(\boldsymbol{\theta})$. Suppose we take $s = 2$ mixing components with $\mu_1 = -1$, $\mu_2 = 1$, and $\pi = 1/4$. Computing the two information matrices, we have:



(a) Normal Scenario 1.

(b) Normal Scenario 2.



(c) Normal Scenario 3.

Figure 3.1: Densities for the bivariate normal finite mixture under the three scenarios.

Table 3.1: Results for Normal mixture. The diagonals $\widetilde{\mathcal{I}}_{ii}$ are given with corresponding $\mathcal{I}_{ii}$ in parentheses. The last column shows Frobenius norm of the matrix difference $\widetilde{\mathcal{I}}-\mathcal{I}$.

(a) Scenario 1

m	$\widetilde{\mathcal{I}}_{11}$	$\widetilde{\mathcal{I}}_{22}$	$\widetilde{\mathcal{I}}_{33}$	$\widetilde{\mathcal{I}}_{44}$	$\widetilde{\mathcal{I}}_{55}$	$\lVert\widetilde{\mathcal{I}}-\mathcal{I}\rVert_F$
1	0.333 (0.276)	0.333 (0.273)	1.0 (0.910)	1.0 (0.920)	5.333 (4.914)	0.6486
2	0.667 (0.643)	0.667 (0.643)	2.0 (1.971)	2.0 (1.971)	5.333 (5.290)	0.1419
3	1.000 (0.994)	1.000 (0.994)	3.0 (2.993)	3.0 (2.993)	5.333 (5.328)	0.0304
4	1.333 (1.332)	1.333 (1.332)	4.0 (3.999)	4.0 (3.999)	5.333 (5.333)	0.0060
5	1.667 (1.666)	1.667 (1.666)	5.0 (5.000)	5.0 (5.000)	5.333 (5.333)	0.0011
6	2.000 (2.000)	2.000 (1.999)	6.0 (6.000)	6.0 (6.000)	5.333 (5.333)	0.0002

(b) Scenario 2

m	$\widetilde{\mathcal{I}}_{11}$	$\widetilde{\mathcal{I}}_{22}$	$\widetilde{\mathcal{I}}_{33}$	$\widetilde{\mathcal{I}}_{44}$	$\widetilde{\mathcal{I}}_{55}$	$\lVert\widetilde{\mathcal{I}}-\mathcal{I}\rVert_F$
1	0.333 (0.192)	0.333 (0.192)	1 (0.777)	1 (0.777)	5.333 (2.729)	3.0006
2	0.667 (0.452)	0.667 (0.452)	2 (1.670)	2 (1.670)	5.333 (3.968)	2.1626
3	1.000 (0.761)	1.000 (0.761)	3 (2.653)	3 (2.653)	5.333 (4.592)	1.7011
...	...	...	...	...	...	...
23	7.667 (7.666)	7.667 (7.666)	23 (23.000)	23 (23.000)	5.333 (5.333)	0.0013
24	8.000 (8.000)	8.000 (8.000)	24 (24.000)	24 (24.000)	5.333 (5.333)	0.0008
25	8.333 (8.333)	8.333 (8.333)	25 (25.000)	25 (25.000)	5.333 (5.333)	0.0005

(c) Scenario 3

m	$\widetilde{\mathcal{I}}_{11}$	$\widetilde{\mathcal{I}}_{22}$	$\widetilde{\mathcal{I}}_{33}$	$\widetilde{\mathcal{I}}_{44}$	$\widetilde{\mathcal{I}}_{55}$	$\lVert\widetilde{\mathcal{I}}-\mathcal{I}\rVert_F$
1	0.333 (0.100)	0.333 (0.100)	1 (0.746)	1 (0.746)	5.333 (0.245)	5.1939
2	0.667 (0.227)	0.667 (0.227)	2 (1.488)	2 (1.488)	5.333 (0.480)	5.2334
3	1.000 (0.375)	1.000 (0.375)	3 (2.231)	3 (2.231)	5.333 (0.703)	5.3942
...	...	...	...	...	...	...
28	9.333 (6.112)	9.333 (6.117)	28 (22.989)	28 (22.989)	5.333 (3.736)	13.4873
29	9.667 (6.387)	9.667 (6.369)	29 (23.907)	29 (23.913)	5.333 (3.798)	13.7428
30	10.000 (6.598)	10.000 (6.648)	30 (24.839)	30 (24.843)	5.333 (3.857)	13.9949
...	...	...	...	...	...	...
78	26.000 (21.254)	26.000 (22.472)	78 (73.687)	78 (73.685)	5.333 (5.085)	14.0573
79	26.333 (21.627)	26.333 (22.845)	79 (74.743)	79 (74.737)	5.333 (5.093)	14.0086
80	26.667 (22.001)	26.667 (23.218)	80 (75.800)	80 (75.798)	5.333 (5.102)	13.8565

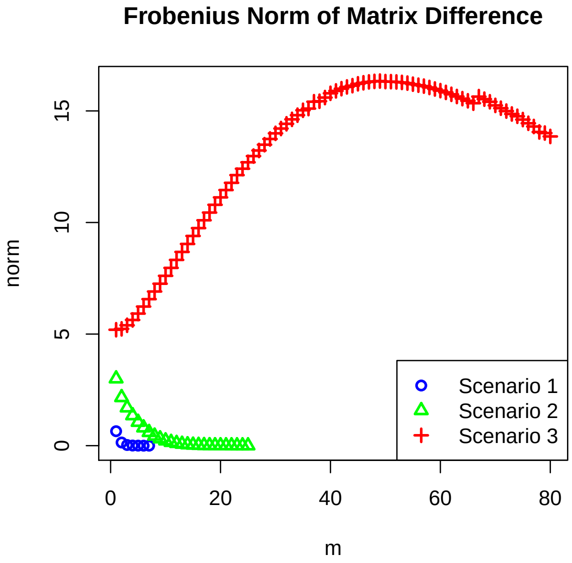


Figure 3.2: Frobenius norm of $\widetilde{\mathcal{I}}_m - \mathcal{I}_m$, as m varies, for the three normal scenarios.

- For $m = 3$,

$$\mathcal{I}_m(\boldsymbol{\theta}) = \begin{pmatrix} 0.5370 & -0.2023 & -0.3692 \\ -0.2023 & 1.9289 & -0.4653 \\ -0.3692 & -0.4653 & 4.5916 \end{pmatrix} \quad \text{vs.}$$

$$m\mathcal{I}_1(\boldsymbol{\theta}) = \begin{pmatrix} 0.4177 & -0.0951 & -1.1399 \\ -0.0951 & 1.6739 & -1.7900 \\ -1.1399 & -1.7900 & 8.1871 \end{pmatrix}.$$

- For $m = 20$,

$$\mathcal{I}_m(\boldsymbol{\theta}) = \begin{pmatrix} 4.9990 & -0.0010 & -0.0003 \\ -0.0010 & 14.9989 & -0.0003 \\ -0.0003 & -0.0003 & 5.3333 \end{pmatrix} \quad \text{vs.}$$

$$m\mathcal{I}_1(\boldsymbol{\theta}) = \begin{pmatrix} 2.7845 & -0.6341 & -7.5991 \\ -0.6341 & 11.1592 & -11.9331 \\ -7.5991 & -11.9331 & 54.5809 \end{pmatrix}.$$

- For $m = 50$,

$$\mathcal{I}_m(\boldsymbol{\theta}) = \begin{pmatrix} 12.5 & 0.0 & 0.0000 \\ 0.0 & 37.5 & 0.0000 \\ 0.0 & 0.0 & 5.3333 \end{pmatrix} \quad \text{vs.}$$

$$m\mathcal{I}_1(\boldsymbol{\theta}) = \begin{pmatrix} 6.9612 & -1.5853 & -18.9977 \\ -1.5853 & 27.8981 & -29.8327 \\ -18.9977 & -29.8327 & 136.4524 \end{pmatrix}.$$

It is evident that the convergence discussed in this chapter does not manifest itself when the sample is unclustered.

Example 3.20 (Dirichlet-Multinomial)**.** Recall from Example 1.2 that Beta-Binomial is a continuous mixture of Binomial. More generally, Dirichlet-Multinomial is a continuous mixture of multinomial, and therefore we may write

$$T \mid \mu \sim \text{Mult}_J(m, \boldsymbol{\mu}), \quad \boldsymbol{\mu} \sim \text{Dirichlet}_J(\boldsymbol{\alpha}),$$

so that the complete data distribution of $(\boldsymbol{T}, \boldsymbol{\mu})$ is

$$f(\boldsymbol{t}, \boldsymbol{\mu} \mid \boldsymbol{\alpha}) = f(\boldsymbol{t} \mid \boldsymbol{\mu}) f(\boldsymbol{\mu} \mid \boldsymbol{\alpha}), \quad \text{where}$$

$$f(\boldsymbol{t} \mid \boldsymbol{\mu}) = \frac{m!}{t_1! \cdots t_J!} \mu_1^{t_1} \cdots \mu_J^{t_J} \quad \text{and} \quad f(\boldsymbol{\mu} \mid \boldsymbol{\alpha}) = \frac{\mu_1^{\alpha_1 - 1} \cdots \mu_J^{\alpha_J - 1}}{B(\alpha_1, \ldots, \alpha_J)}.$$

Take $J = k + 1$ to ensure the parameter space of the multinomial family contains an open set in $\mathbb{R}^k$, which was assumed at the beginning of the chapter. The Dirichlet-Multinomial distribution is obtained by finding the marginal distribution of $\boldsymbol{T}$,

$$f(\boldsymbol{t} \mid \boldsymbol{\alpha}) = \int f(\boldsymbol{t} \mid \boldsymbol{\mu}) f(\boldsymbol{\mu} \mid \boldsymbol{\alpha}) d\boldsymbol{\mu}, \tag{3.17}$$

where the integral may be computed in closed form. Although the theory in this chapter has been developed specifically for finite mixtures of exponential families, we can construct an approximate information matrix using the complete data. Note that the distribution of $\boldsymbol{T} \mid \boldsymbol{\mu}$ is free of $\boldsymbol{\alpha}$ so that

$$\frac{\partial}{\partial \boldsymbol{\alpha}} \log f(\boldsymbol{t}, \boldsymbol{\mu} \mid \boldsymbol{\alpha}) = \frac{\partial}{\partial \boldsymbol{\alpha}} \log f(\boldsymbol{\mu} \mid \boldsymbol{\alpha});$$

therefore, the complete data information matrix is just the FIM with respect to $\text{Dirichlet}_J(\boldsymbol{\alpha})$. This is analogous to the finite mixture case, where the first s diagonal blocks correspond to the support points of the mixing distribution

$$\text{Discrete}(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_s; \boldsymbol{\pi}),$$

and the lower-right block of the matrix corresponds to $\boldsymbol{\pi}$. Now the mixing process follows a Dirichlet distribution whose support is the probability simplex in $\mathbb{R}^J$ (i.e. which does not have corresponding entries in the information matrix). Theorem 1 in Neerchal and Morel (1998) shows that the exact information matrix under the marginal distribution (3.17) of

$\boldsymbol{T}$ converges to the FIM of Dirichlet$_k(\boldsymbol{\alpha})$ as $m \to \infty$. Therefore, the theory of this chapter may extend to more general settings than when the latent mixing process follows a finite mixture distribution.

Example 3.21 (Normal-Normal)**.** Let us consider a second continuous mixture along the lines of Example 3.20. The normal-normal hierarchical model is popular in Bayesian analysis (Gelman et al., 2003, Section 5.4), with one application (for example) in the Fay-Herriot model for small area estimation (Rao, 2003). The results from this chapter can be applied in the following sense. Suppose

$$\begin{aligned} \bar{X} \mid \mu &\sim \mathrm{N}(\mu, \sigma^2/m), \\ \mu &\sim \mathrm{N}(\theta, \tau^2). \end{aligned}$$

and take σ^2 and τ^2 to be known for the sake of demonstration. Recall that if $T = \sum_{i=1}^m X_m \sim \mathrm{N}(m\mu, m\sigma^2)$, then $\bar{X} = T/m$ and we may obtain the density of $\bar{X}$ by transformation from

$$\begin{aligned} f_{\bar{X}}(x \mid \theta) &= \int f_{\bar{X}}(x \mid \mu) f_\mu(\mu \mid \theta) d\mu \\ &= \left| \frac{\partial T}{\partial \bar{X}} \right| \int f_T(t \mid \mu) f_\mu(\mu \mid \theta) d\mu \\ &= \left| \frac{\partial T}{\partial \bar{X}} \right| f_T(x \mid \theta). \end{aligned}$$

Therefore,

$$\frac{\partial}{\partial \theta} \log f_{\bar{X}}(x \mid \theta) = \frac{\partial}{\partial \theta} \log f_T(t \mid \theta),$$

and the information is the same whether we work with $\bar{X}$ or T. It can be shown that marginally,

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{m} + \tau^2\right),$$

therefore the information about θ in $\bar{X}$ is $\mathcal{I}_m(\theta) = (\sigma^2/m + \tau^2)^{-1}$. As in Example 3.20, the complete data information about θ in $(\bar{X}, \mu)$ is $\widetilde{\mathcal{I}}(\theta) = \tau^{-2}$. Now we have convenient forms for both the exact and approximate information, and it is clear that $\mathcal{I}_m(\theta) \to \widetilde{\mathcal{I}}(\theta)$ as $m \to \infty$.

Example 3.22 (Mixture of Finite Mixtures)**.** Let us consider another example where the results in this chapter apply to a distribution which does not immediately appear to be an exponential family finite mixture. Consider the finite mixture of RCB densities

$$f(t \mid m, \boldsymbol{\theta}) = \sum_{\ell=1}^{s} w_\ell \mathrm{RCB}(t \mid m, \rho_\ell, \pi_\ell).$$

where $\boldsymbol{\theta} = (\rho_1, \pi_1, \ldots, \rho_s, \pi_s, w_1, \ldots, w_{s-1})$. The density may be rewritten as a binomial finite mixture

$$\begin{aligned} f(t \mid m, \boldsymbol{\theta}) &= \sum_{\ell=1}^{s} w_\ell \sum_{j=1}^{2} \pi_{\ell j} \mathrm{Bin}(t \mid m, \xi_\ell) \\ &= \sum_{\ell=1}^{2s} \lambda_\ell \mathrm{Bin}(t \mid m, \xi_\ell) \end{aligned}$$

where

$$\xi_\ell = \begin{cases} (1 - \rho_{\frac{\ell+1}{2}})\pi_{\frac{\ell+1}{2}} + \rho_{\frac{\ell+1}{2}} & \text{if } \ell \text{ is odd} \\ (1 - \rho_{\ell/2})\pi_{\ell/2} & \text{o.w.} \end{cases}$$

and

$$\lambda_\ell = \begin{cases} w_{\frac{\ell+1}{2}} \pi_{\frac{\ell+1}{2}} & \text{if } \ell \text{ is odd} \\ w_{\ell/2}(1 - \pi_{\ell/2}) & \text{o.w.} \end{cases}$$

for $\ell = 1, \ldots, 2s$. It is now clear that the approximate information matrix $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta})$ may be formulated by first forming the approximate information matrix with respect to $\boldsymbol{\vartheta} = (\xi_1, \ldots, \xi_{2s}, \lambda_1, \ldots, \lambda_{2s-1})$,

$$\widetilde{\mathcal{I}}_m(\boldsymbol{\vartheta}) = \text{Blockdiag}\left(\frac{m}{\xi_1(1-\xi_1)}, \ldots, \frac{m}{\xi_{2s}(1-\xi_{2s})}, \boldsymbol{D}_\lambda^{-1} + \lambda_{2s}^{-1} \mathbf{1}\mathbf{1}^T \right)$$

and then using the Jacobian of the transformation $\boldsymbol{\theta} \mapsto \boldsymbol{\vartheta}$ to obtain

$$\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) = \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\vartheta}} \right) \widetilde{\mathcal{I}}_m(\boldsymbol{\vartheta}) \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\vartheta}} \right)^T .$$

The convergence of $\widetilde{\mathcal{I}}_m(\boldsymbol{\vartheta}) - \mathcal{I}_m(\boldsymbol{\vartheta})$ to zero follows from Theorem 3.11, and the convergence of $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})$ to zero follows from Remark 3.7.

Example 3.23 (Weibull Finite Mixture)**.** Consider the Weibull density

$$f(x \mid \beta, \lambda) = \frac{\beta}{\lambda} \left(\frac{x}{\lambda} \right)^{\beta - 1} e^{-(x/\lambda)^\beta} I(x > 0),$$

where $\beta > 0$ and $\lambda > 0$. For a random variable X with this distribution we will write $X \sim \text{Weibull}(\beta, \lambda)$. Consider the case when λ is known but β is unknown so that $\{f(\cdot \mid \beta, \lambda) : \beta > 0\}$ is not an exponential family. In this case, the score can be written as

$$\frac{\partial}{\partial \beta} \log f(\boldsymbol{x} \mid \beta, \lambda) = \frac{1}{\beta} - \left[1 - \left(\frac{x}{\lambda} \right)^\beta \right] \log \left(\frac{x}{\lambda} \right),$$

and the Fisher information is therefore found by computing

$$\mathcal{I}(\beta) = \int_0^\infty \left\{ \frac{1}{\beta} - \left[1 - \left(\frac{x}{\lambda}\right)^\beta \right] \log\left(\frac{x}{\lambda}\right) \right\}^2 f(x \mid \beta, \lambda) dx. \tag{3.18}$$

Although the theory developed in this chapter does not apply because of the departure from exponential family, let us investigate the convergence of the approximate information as we did in the other examples. Suppose $\boldsymbol{X} = (X_1, \ldots, X_m)$ given $Z = \ell$ are a random sample from Weibull$(\beta_\ell, \lambda_\ell)$. Therefore, the marginal density of $\boldsymbol{X}$ is given by

$$\begin{aligned} f(\boldsymbol{x} \mid \boldsymbol{\theta}) &= \sum_{\ell=1}^{s} \pi_\ell f(\boldsymbol{x} \mid \beta_\ell, \lambda_\ell) \\ &= \sum_{\ell=1}^{s} \pi_\ell \left[\left(\frac{\beta_\ell}{\lambda_\ell}\right)^m \left(\prod_{i=1}^{m} \frac{x_i}{\lambda_\ell} \right)^{\beta_\ell - 1} \exp\left\{ - \sum_{i=1}^{m} (x_i/\lambda_\ell)^{\beta_\ell} \right] \right\} \end{aligned} \tag{3.19}$$

where $\boldsymbol{\theta} = (\beta_1 \ldots, \beta_s, \pi_1, \ldots, \pi_{s-1})$. The corresponding score vector contains entries

$$\begin{aligned} &\frac{\partial}{\partial \beta_a} \log f(\boldsymbol{x} \mid \boldsymbol{\theta}) \\ &\quad = \frac{\pi_a f(\boldsymbol{x} \mid \beta_a, \lambda_a)}{f(\boldsymbol{x} \mid \boldsymbol{\theta})} \left[\frac{m}{\beta_a} + \sum_{i=1}^{m} \log x_i - m \log \lambda_a - \sum_{i=1}^{m} \left(\frac{x_i}{\lambda_a}\right)^{\beta_a} \log\left(\frac{x_i}{\lambda_a}\right) \right], \end{aligned}$$

for $a = 1, \ldots, s$ and

$$\frac{\partial}{\partial \pi_a} \log f(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{f(\boldsymbol{x} \mid \beta_a, \lambda_a) - f(\boldsymbol{x} \mid \beta_s, \lambda_s)}{f(\boldsymbol{x} \mid \boldsymbol{\theta})}$$

for $a = 1, \ldots, s-1$. The exact information matrix $\mathcal{I}(\boldsymbol{\theta})$ can be computed approximately by Monte Carlo simulation, with the (a, b)th entry for the upper-triangular portion of matrix computed as

$$\mathcal{I}_{ab}(\boldsymbol{\theta}) \approx \begin{cases} \frac{1}{L}\sum_{r=1}^{L}\left\{\frac{\partial}{\partial \beta_a}\log f(\boldsymbol{X}^{(r)} \mid \boldsymbol{\theta})\right\}\left\{\frac{\partial}{\partial \beta_b}\log f(\boldsymbol{X}^{(r)} \mid \boldsymbol{\theta})\right\}, \\ \quad \text{if } a \leq s \text{ and } a \leq b \leq s \\ \frac{1}{L}\sum_{r=1}^{L}\left\{\frac{\partial}{\partial \beta_a}\log f(\boldsymbol{X}^{(r)} \mid \boldsymbol{\theta})\right\}\left\{\frac{\partial}{\partial \pi_{b-s}}\log f(\boldsymbol{X}^{(r)} \mid \boldsymbol{\theta})\right\}, \\ \quad \text{if } a \leq s \text{ and } b > s \\ \frac{1}{L}\sum_{r=1}^{L}\left\{\frac{\partial}{\partial \pi_{a-s}}\log f(\boldsymbol{X}^{(r)} \mid \boldsymbol{\theta})\right\}\left\{\frac{\partial}{\partial \pi_{b-s}}\log f(\boldsymbol{X}^{(r)} \mid \boldsymbol{\theta})\right\}, \\ \quad \text{if } a > s,\, b > s, \text{ and } a \leq b \end{cases}$$

where $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(L)}$ are random samples from (3.19), and the number of repetitions L is taken to be larger for more accuracy. By symmetry, the lower-triangular entries $\mathcal{I}_{ab}(\boldsymbol{\theta})$ can be taken as $\mathcal{I}_{ba}(\boldsymbol{\theta})$ for $a \in \{1, \ldots, 2s-1\}$ and $b < a$. The approximation information matrix is given by

$$\widetilde{\mathcal{I}}(\boldsymbol{\theta}) = \text{Blockdiag}(\pi_1 F_1, \ldots, \pi_s F_s, \boldsymbol{F}_\pi)$$

where F_ℓ is given by multiplying the Weibull$(\beta_\ell, \lambda_\ell)$ information (3.18) by m, and $\boldsymbol{F}_\pi = \boldsymbol{D}_\pi^{-1} + \pi_s^{-1}\mathbf{1}\mathbf{1}^T$ as usual for finite mixtures.

Consider two scenarios of the form

$$\pi\text{Weibull}(\beta_1, \lambda_1) + (1-\pi)\text{Weibull}(\beta_2, \lambda_2)$$

with

- Scenario 1: $(\beta_1 = 1, \lambda_1 = 1)$, $(\beta_2 = 4, \lambda_2 = 4)$, and $\pi = 1/3$,
- Scenario 2: $(\beta_1 = 1, \lambda_1 = 1)$, $(\beta_2 = 2, \lambda_2 = 2)$, and $\pi = 1/3$.

Figure 3.3 plots the subpopulations and mixed population for each scenario. Table 3.2

Table 3.2: Results for Weibull mixture. The diagonals $\widetilde{\mathcal{I}}_{ii}$ are given with corresponding $\mathcal{I}_{ii}$ in parentheses. The last column shows Frobenius norm of the matrix difference $\widetilde{\mathcal{I}} - \mathcal{I}$. All entries of $\mathcal{I}$ were approximated by Monte Carlo simulation using $L = 100{,}000$.

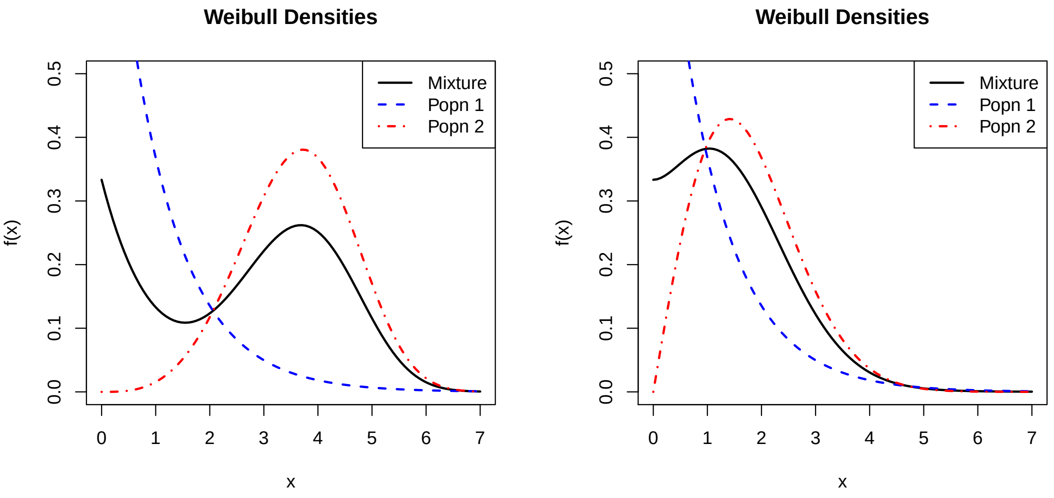
(a) Scenario 1

m	$\widetilde{\mathcal{I}}_{11}$	$\widetilde{\mathcal{I}}_{22}$	$\widetilde{\mathcal{I}}_{33}$	$\Vert\widetilde{\mathcal{I}} - \mathcal{I}\Vert_F$
1	0.6079 (0.3787)	0.0760 (0.0535)	4.5000 (3.2304)	1.3201
2	1.2158 (1.0521)	0.1520 (0.1279)	4.5000 (4.0346)	0.5472
3	1.8237 (1.7571)	0.2280 (0.2112)	4.5000 (4.3218)	0.2397
4	2.4316 (2.3626)	0.3039 (0.2926)	4.5000 (4.4237)	0.1256
5	3.0395 (2.9479)	0.3799 (0.3772)	4.5000 (4.4805)	0.1122
6	3.6474 (3.5409)	0.4559 (0.4494)	4.5000 (4.4914)	0.1097
7	4.2553 (4.3264)	0.5319 (0.5281)	4.5000 (4.5106)	0.0729
8	4.8632 (4.9649)	0.6079 (0.6077)	4.5000 (4.4984)	0.1082
9	5.4711 (5.4920)	0.6839 (0.6854)	4.5000 (4.5032)	0.0257
10	6.0790 (6.0419)	0.7599 (0.7637)	4.5000 (4.5010)	0.0404

(b) Scenario 2

m	$\widetilde{\mathcal{I}}_{11}$	$\widetilde{\mathcal{I}}_{22}$	$\widetilde{\mathcal{I}}_{33}$	$\Vert\widetilde{\mathcal{I}} - \mathcal{I}\Vert_F$
1	0.6079 (0.3919)	0.3039 (0.1696)	4.5000 (1.0642)	3.4731
2	1.2158 (0.8718)	0.6079 (0.3840)	4.5000 (1.7997)	2.8164
3	1.8237 (1.3980)	0.9118 (0.6135)	4.5000 (2.3182)	2.3894
4	2.4316 (1.9380)	1.2158 (0.8703)	4.5000 (2.7546)	2.0388
5	3.0395 (2.5468)	1.5197 (1.1423)	4.5000 (3.0743)	1.7982
…	…	…	…	…
23	13.9816 (13.7489)	6.9908 (6.8029)	4.5000 (4.4462)	0.3482
24	14.5895 (14.5347)	7.2947 (7.1399)	4.5000 (4.4513)	0.2575
25	15.1974 (15.0696)	7.5987 (7.5052)	4.5000 (4.4704)	0.2163
26	15.8053 (15.9109)	7.9026 (7.8191)	4.5000 (4.4645)	0.1920
27	16.4132 (16.3579)	8.2066 (8.1740)	4.5000 (4.4682)	0.1320

compares the approximate and exact information matrices for these scenarios, respectively. Evaluation of the approximate information matrix requires evaluation of (3.18), which we compute by numerical integration. The exact information matrix is computed by Monte Carlo simulation, as mentioned, with $L = 100{,}000$. The accuracy is less than ideal, but enough for our purpose of establishing whether the convergence is taking place. It is clear that the convergence is taking place, and, as expected, is faster for Scenario 1 where the subpopulations are further apart.



(a) Scenario 1. (b) Scenario 2.

Figure 3.3: Densities for the Weibull finite mixture under the two scenarios.

3.6 Conclusions

In this chapter, we have extended the approximate information matrix discussed in Chapter 2 from multinomial finite mixtures to exponential family finite mixtures. The extension became possible upon noticing that the approximate information matrix is the complete data information matrix of the response and the latent mixing process. This makes the approximation applicable to statistical analysis beyond binomial and multinomial data.

In the multinomial case, the approximation had originally been justified by its convergence to the exact information matrix as the number of multinomial trials m becomes large. In the exponential family case, we instead consider sampling m observations from a common, but unknown, subpopulation. Under this construction, the exact and approximate information matrices are again seen to converge together as m becomes large. The proof in the exponential family case is quite different than the one specific to multinomial, which is given in full detail in Chapter 2, and does not depend on the properties of multinomial. Rates of convergence were obtained showing that the convergence is exponential, but the exponent depends on both m and the similarity between subpopulations. Convergence is very fast when subpopulations are distinct, but becomes slow as they are moved closer together. Example 3.19 suggests that the approximation does not converge to the information matrix of an independent and identically distributed sample of size m taken from the finite mixture.

There are several interesting questions to consider at this point. The setting of exponential family finite mixtures covers many cases that may be useful in application, but our convergence proof requires this assumption (e.g. the $R_i(\cdot)$ and $Q_i(\cdot)$ functions are critical to the proof). Examples 3.20 and 3.21 provide evidence of the convergence even when the latent mixing process is a continuous distribution rather than the discrete distribution assumed in the finite mixture. Example 3.23 shows the convergence in a Weibull finite mixture which does not meet the exponential family assumption. This suggests that the convergence result can be generalized beyond what has been proved in this chapter. Finally, it would be of interest to have a reliable method of correcting accuracy in the approximate information when m is not large or the subpopulations are not well-separated.

Chapter 4

Mixture Link Models for Binomial Data with Overdispersion

4.1 Introduction

A common problem in the analysis of binomial data using logistic regression occurs when more variation is present in the data than can be expressed by the model. This can happen when basic modeling assumptions are not met, and when it does overdispersion is said to occur. This chapter considers a novel way of handling overdispersion in the binomial regression setting by linking predictors, through a regression, to the probability of success in a finite mixture of binomials. Such a mixture presumes J latent binomial subpopulations with success probabilities μ_j, for $1, \ldots, J$, and who compose the overall population with proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)$. The quantity $\boldsymbol{\mu}^T \boldsymbol{\pi} = \pi_1 \mu_1 + \cdots + \pi_J \mu_J$ can be interpreted as the overall probability of success for a single trial. The finite mixture has a natural appeal for its ability to model extra variation in a robust way, and the proposed model promises to use this ability to reflect extra variability in estimates for a single regression over the entire overall population. For example, this would be desirable in the case of a “primary” subpopulation and a “contamination” subpopulation. Finite mixtures are also used to accommodate more general departures from simple assumptions, such as multiple modes. In such cases, it may be desired to study the mean behavior of the overall population, but to ensure that inference is done with decreased precision.

There are technical challenges that must be overcome in establishing the link to

$\boldsymbol{\mu}^T \boldsymbol{\pi}$ and carrying out even basic computation of the likelihood. This chapter develops one possible implementation of the model. Initial results show that it provides a good fit for a real dataset with known overdispersion issues, faring well in comparison to several other binomial models with extra variation.

The rest of the chapter proceeds as follows. Section 4.2 introduces the binomial regression problem and discusses some existing approaches to handling extra variation. Section 4.3 develops the new model, which is termed the Mixture Link distribution. Section 4.4 obtains the first few moments of the distribution. In order to work with the proposed model, it is necessary to compute the vertices of the set which represents the link to the regression; this is discussed in Section 4.5. Section 4.6 discusses evaluation of the density, which appears to require numerical approximation except in some simple cases. Section 4.7 presents plots of the density to give an idea of its ability to capture extra-binomial variation. Section 4.8 proposes a moment-matched beta approximation to the random effects density, to reduce the amount of computation needed to evaluate the density. Illustrative data analyses are presented in Section 4.9, where several binomial models are compared using a goodness-of-fit test as well as AIC and BIC. Finally, Section 4.10 concludes the chapter.

4.2 Background

Under the usual logistic regression model, T_i successes are observed in m_i trials for $i = 1, \ldots, n$. The probability of success p_i for each observation is modeled on a covariate $\boldsymbol{x}_i \in \mathbb{R}^d$, which is taken to be fixed. It is assumed that $p_i = G(\boldsymbol{x}_i^T \boldsymbol{\beta})$ for $\boldsymbol{\beta} \in \mathbb{R}^d$ and $G : \mathbb{R} \to (0, 1)$ is a prespecified inverse link function. For this chapter, G will be taken to be the cumulative distribution function (CDF) for the logistic distribution $G(x) = 1/(1 + e^{-x})$, but at no point does the development require this. The model just described may be written briefly as

$$T_i \overset{\text{ind}}{\sim} \text{Bin}(m_i, p_i), \quad p_i = G(\boldsymbol{x}_i^T \boldsymbol{\beta}).$$

In practice, T_i, $\boldsymbol{x}_i$, and m_i are observed, and statistical inference on the parameter $\boldsymbol{\beta}$ is a primary objective of analysis. Logistic regression is a special case of the generalized linear model (GLM) framework (McCullagh and Nelder, 1989), which allows non-normal, non-continuous outcomes to be modeled as responses to a regression. However, a frequent problem with GLM is that the data exhibit more variation than the underlying exponential family distribution is capable of expressing (Morel and Neerchal, 2012). Overdispersion may be caused, for example, when important covariates have not been included in the regression, or when the implicit assumption of independence within the m_i trials has been violated. The limitation in the amount of modeled variability in the binomial GLM can be seen by noting the relationship between the mean and variance

$$\text{E}(T_i) = m_i p_i \quad \text{and} \quad \text{Var}(T_i) = m_i p_i (1 - p_i);$$

therefore, the same regression used to model the probabilities of success of the T_i also must explain the mean and variance.

A simple workaround is to extend the model with a dispersion parameter ϕ so that $\text{Var}(T_i) = \phi m_i p_i (1 - p_i)$ (Agresti, 2002, Section 4.7). The resulting model is referred to as quasi-likelihood because it no longer corresponds to a true distribution. For longitudinal data, a popular quasi-likelihood method is the generalized estimating equations (GEE) developed by Liang and Zeger (1986). GEE proposes inference on $\boldsymbol{\beta}$ to be based on a score-like equation and allows the analyst to assume a working correlation structure to induce dependence for observations within a subject. This idea may be used when ungrouped Bernoulli trials of a binomial experiment are observed. GEE has some desirable properties, such as consistency even under misspecification of the working correlation;

however it may not be based on a real likelihood.

There are also a variety of likelihood-based models that can be used to induce extra variation; we will mention several here. The zero-inflated binomial (ZIB) distribution discussed by Hall (2000),

$$\mathrm{P}(T = t \mid m, p, \phi) = \phi I(t = 0) + (1 - \phi)\text{Bin}(t \mid m, p),$$

assumes a latent process that generates a zero with probability ϕ and a binomial random variable with probability $1 - \phi$. Similarly, any of the support values $0, 1, \ldots, m$ may be selected by the analyst to be inflated. The random-clumped binomial (RCB) distribution (Morel and Nagaraj, 1993) may be used when the inflated value is not known ahead of time and is considered to be drawn randomly. An RCB distributed random variable $T = NY + (X \mid N)$ is obtained using

$$Y \sim \text{Ber}(p), \quad N \sim \text{Bin}(m, \phi), \quad (X \mid N) \sim \text{Bin}(m - N, p),$$

where Y represents success/failure of a leader, N is the number of trials that follows the leader, and $(X \mid N)$ are remaining trials that are selected independently. Here, $p \in (0, 1)$ is interpreted as the success probability for the trials, and $\phi \in (0, 1)$ is the probability of following the leader. Perhaps the most popular binomial distribution supporting extra variation is beta-binomial (BB), which assumes a hierarchy,

$$T \mid \mu \sim \text{Bin}(m, \mu), \quad \mu \sim \text{Beta}(\alpha, \beta),$$

where the probability of success is drawn from a beta distribution. BB may be reparameterized, as noted in (Morel and Neerchal, 2012, Section 4.2) and (Prentice, 1986) for example, using

$$\begin{aligned} &\alpha = p\phi^{-1}(1-\phi) \quad \text{and} \quad \beta = (1-p)\phi^{-1}(1-\phi) \\ \iff\ &p = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \phi = \frac{1}{\alpha+\beta+1}, \end{aligned}$$

so that $p \equiv \mathrm{E}(\mu) \in (0,1)$ can be interpreted as a probability of success and

$$\mathrm{Var}(T) = mp(1-p)\{1+\phi(m-1)\}.$$

For the ZIB, RCB, and BB distributions as stated here, $\phi \in (0,1)$ is seen as an overdispersion parameter with respect to the binomial distribution where the limiting case of $\phi = 0$ corresponds to "no overdispersion". Although ZIB, RCB, and BB are not exponential families, and therefore do not fall into the classical GLM framework, regressions may be linked to p and/or ϕ, and inference for $\boldsymbol{\beta}$ may be carried out through the linked likelihood.

Adding random effects to the regression model of a GLM is a flexible way to model extra variation between observations or to group observations that naturally belong to the same cluster (c.f. Agresti, 2002; Morel and Neerchal, 2012). However, because random effects are unobserved and manifest themselves as integrals in the likelihood, computation quickly becomes difficult as random effect structures are allowed to become more elaborate. A compromise between flexibility and computation is found in the random intercept model, where only a random intercept is assumed. Logistic regression with a random intercept has been considered by Follmann and Lambert (1989) and Aitkin (1996), among others, who use nonparametric maximum likelihood (NPMLE) to avoid making assumptions about the distribution of the random intercept.

Finite mixture distributions are often used to model the situation of multiple latent subpopulations. In the basic finite mixture of binomials,

$$f(t \mid m, \boldsymbol{\theta}) = \sum_{j=1}^{J} \pi_j \text{Bin}(t \mid m, \mu_j), \tag{4.1}$$

it is assumed that there are J subpopulations, and a latent process Z is selecting from the labels $(1, \dots, J)$ with corresponding probabilities $(\pi_1, \dots, \pi_J)$. The finite mixture (4.1) can be extended to a finite mixture of regressions by linking regressions

$$\mu_j = G(\boldsymbol{x}^T \boldsymbol{\beta}_j), \quad \text{for } j = 1, \dots, J.$$

This idea is discussed in Frühwirth-Schnatter (2006), which also considers distributions other than simple discrete for the mixing process Z.

The remainder of this chapter presents Mixture Link: a completely likelihood-based binomial model for extra variation. Mixture Link models a binomial outcome with a finite mixture, and links a regression to the mixture probability of success. The finite mixture is used to handle heterogeneity in a robust way, but unlike the finite mixture of regressions model the interest is in a single regression for the overall population. Therefore, the finite mixture of regressions can be thought of as "conditional modeling" with respect to latent subpopulations, while Mixture Link is "marginal modeling" on the entire population, with built-in tolerance for heterogeneity across subpopulations.

4.3 Model Formulation

Consider a random variable T following the finite mixture of binomials distribution (4.1), which we will denote as $T \sim \text{BinMix}(m, \boldsymbol{\mu}, \boldsymbol{\pi})$. Without further restriction, the component probabilities of success $\boldsymbol{\mu} = (\mu_1, \dots, \mu_J)$ naturally lie within the rectangle $[0, 1]^J$, and the subpopulation proportions $\boldsymbol{\pi} = (\pi_1, \dots, \pi_J)$ are within the J-dimensional probability simplex $\mathcal{S}^J = \{\boldsymbol{\mu} \in [0,1]^J : \sum_{j=1}^J \mu_j = 1\}$. Notice that

$$\mathrm{E}(T) = \sum_{j=1}^{J} \pi_j m \mu_j = m\boldsymbol{\mu}^T\boldsymbol{\pi} \tag{4.2}$$

where $\boldsymbol{\mu}^T\boldsymbol{\pi}$ is the mixture probability of success. Analogously to logistic regression under the GLM framework, our goal is to link the regression $\boldsymbol{x}^T\boldsymbol{\beta}$ to the finite mixture by enforcing the constraint

$$\boldsymbol{\mu}^T\boldsymbol{\pi} = p, \quad \text{where } p = G(\boldsymbol{x}^T\boldsymbol{\beta}).$$

The space of all $\boldsymbol{\mu}$ that honors the link is then

$$A(p, \boldsymbol{\pi}) = \{\boldsymbol{\mu} \in [0,1]^J : \boldsymbol{\mu}^T\boldsymbol{\pi} = p\}; \tag{4.3}$$

when there is no confusion, we will write A as shorthand. We will often write the Mixture Link model in terms of p rather than $\boldsymbol{\beta}$, with the understanding that the regression $p = G(\boldsymbol{x}^T\boldsymbol{\beta})$ can be linked when desired. Although Mixture Link was developed with regression in mind, the distribution is well-defined without the link.

Consider an independent sample

$$T_i \overset{\text{ind}}{\sim} \text{BinMix}(m_i, \boldsymbol{\mu}_i, \boldsymbol{\pi}), \quad \boldsymbol{\mu}_i \in A_i, \quad i = 1, \ldots, n,$$

where $A_i = A(p_i, \boldsymbol{\pi})$ and $p_i = G(\boldsymbol{x}_i^T\boldsymbol{\beta})$. Here A_i and $\boldsymbol{\mu}_i$ vary with i to reflect that observations may have distinct covariates $\boldsymbol{x}_i$. We assume that $\boldsymbol{\pi}$ is common to all observations. When $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n$ are treated as fixed and unknown quantities, taking a maximum likelihood approach would mean maximizing

$$\prod_{i=1}^{n}\left\{\sum_{j=1}^{J}\pi_j \text{Bin}(t_i \mid m_i, \mu_{ij})\right\}, \quad \text{subject to} \quad \boldsymbol{\gamma}(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_n,\boldsymbol{\pi}) = \boldsymbol{X}\boldsymbol{\beta}, \tag{4.4}$$

$$\boldsymbol{\gamma}(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_n,\boldsymbol{\pi}) = \begin{pmatrix} g(\boldsymbol{\mu}_1^T\boldsymbol{\pi}) \\ \vdots \\ g(\boldsymbol{\mu}_n^T\boldsymbol{\pi}) \end{pmatrix} : n \times 1, \quad \text{and} \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} : n \times d,$$

where $g = G^{-1}$. The maximum likelihood estimator subject to constraints has been studied, for example, in (Aitchison and Silvey, 1958). In (4.4), the parameter $\boldsymbol{\beta}$ only enters the optimization problem through the constraint, which suggests a that a profile likelihood approach such as

$$Q(\boldsymbol{\beta}) = \sup_{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_n,\boldsymbol{\pi}} \left\{\log L(\boldsymbol{\theta}) : \boldsymbol{\gamma}(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_n,\boldsymbol{\pi}) = \boldsymbol{X}\boldsymbol{\beta}\right\}, \quad \text{or} \tag{4.5}$$

$$Q(\boldsymbol{\beta},\boldsymbol{\pi}) = \sup_{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_n} \left\{\log L(\boldsymbol{\theta}) : \boldsymbol{\gamma}(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_n,\boldsymbol{\pi}) = \boldsymbol{X}\boldsymbol{\beta}\right\}, \tag{4.6}$$

may be more natural to consider. This removes the nuisance $\boldsymbol{\mu}_i$ variables from consideration by optimizing over them. However, the overall optimization problem is still on the space

$$\underbrace{[0,1]^J \times \cdots \times [0,1]^J}_{n} \times \mathcal{S}^J \times \mathbb{R}^d,$$

whose dimension is increasing with the sample size n due to the nuisance parameters $\boldsymbol{\mu}_i$ for $i = 1,\ldots,n$, This is generally not a desirable quality for a model.

Instead of taking on the optimization problem (4.4), we consider a hierarchical model where the $\boldsymbol{\mu}_i$ are unobservable random effects. Hence, the $\boldsymbol{\mu}_i$ will be probabilities of success for a finite mixture of J binomials, permitted to vary among observations by the assumption of being drawn from a distribution on $A_i \subseteq [0,1]^J$, where there A_i support the common objective of linking the desired regression to the finite mixture likelihood. The effects must be integrated out (rather than optimized over, as in profile likelihood) to obtain the likelihood of the observed data. This can be contrasted to the profile optimization which removes the $\boldsymbol{\mu}_i$ from consideration by an inner optimization. The tradeoff between having too many fixed nuisance parameters vs. unobservable random effects is traditionally seen in linear mixed models (McCulloch et al., 2008). A first question for the present case is to determine a distribution for the random effects. A simple result will give one possible answer to this question.

Lemma 4.1. *For a fixed p and $\boldsymbol{\pi}$, the set $A(p, \boldsymbol{\pi})$ as defined in (4.3) is bounded and convex.*

Proof. Notice that

$$A(p, \boldsymbol{\pi}) = [0,1]^J \cap \{\boldsymbol{\mu} \in \mathbb{R}^J : \boldsymbol{\mu}^T \boldsymbol{\pi} = G(\boldsymbol{x}^T \boldsymbol{\beta})\},$$

which is an intersection of two convex sets, a rectangle and a hyperplane, in $\mathbb{R}^J$. Therefore $A(p, \boldsymbol{\pi})$ itself is convex. It is also bounded because $A(p, \boldsymbol{\pi}) \subseteq [0,1]^J$. $\square$

Because any given A_i is bounded and convex, we can find vertices $\boldsymbol{v}_1^{(i)}, \ldots, \boldsymbol{v}_{k_i}^{(i)} \in \mathbb{R}^J$ such that A_i is equivalent to their convex hull, defined as

$$\operatorname{conv}(\boldsymbol{v}_1^{(i)}, \ldots, \boldsymbol{v}_{k_i}^{(i)}) = \Big\{ \sum_{\ell=1}^{k_i} \lambda_\ell \boldsymbol{v}_\ell^{(i)} : \boldsymbol{\lambda} \in \mathcal{S}^{k_i} \Big\} = \Big\{ \boldsymbol{V}^{(i)} \boldsymbol{\lambda} : \boldsymbol{\lambda} \in \mathcal{S}^{k_i} \Big\}, \tag{4.7}$$

where $\mathcal{S}^{k_i}$ is the k_i-dimensional probability simplex and $\boldsymbol{V}^{(i)} = (\boldsymbol{v}_1^{(i)} \cdots \boldsymbol{v}_{k_i}^{(i)}) \in \mathbb{R}^{J \times k_i}$. Therefore, any point in A_i can be expressed as a convex combination of the vertices. A proof that this decomposition is possible is given, for example, in (Bazaraa et al., 2009, Theorem 2.1). Vertices of A_i are also extreme points of A_i, whose definition is as follows.

Definition 4.2 (Extreme Point of a Convex Set)**.** A point x in a convex set S is called an extreme point of S if it cannot be written as a convex combination of other points in S.

That is,

$$\boldsymbol{x} = \lambda \boldsymbol{y} + (1-\lambda)\boldsymbol{z} \text{ for some } \lambda \in [0,1] \quad \Longrightarrow \quad \boldsymbol{x} = \boldsymbol{y} = \boldsymbol{z}.$$

Note that $\boldsymbol{V}^{(i)}$ may be different for each observation when the set A_i depends on a covariate $\boldsymbol{x}_i$. The number of vertices k_i may also vary with each observation. It is assumed that k_i is chosen to be the minimum number of vertices so that (4.7) holds; i.e. all $\boldsymbol{v}_\ell^{(i)}$ are extreme points of A_i and all $\boldsymbol{v}_\ell^{(i)}$ are distinct points.

Now a natural way to place a distribution on the set A is to let $\boldsymbol{\lambda} \sim \text{Dirichlet}_k(\boldsymbol{\alpha})$, whose density is

$$f(\boldsymbol{\lambda} \mid \boldsymbol{\alpha}) = \frac{\lambda_1^{\alpha_1 - 1} \cdots \lambda_k^{\alpha_k - 1}}{\mathrm{B}(\boldsymbol{\alpha})} \cdot I(\boldsymbol{\lambda} \in \mathcal{S}^k), \quad \text{where } \mathrm{B}(\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)}{\Gamma(\alpha_1 + \cdots + \alpha_k)}.$$

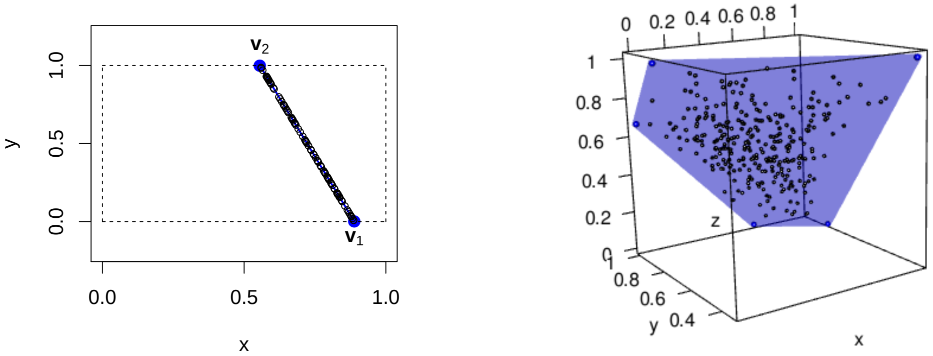
Danaher et al. (2012) recently proposed priors based on the Minkowski-Weyl decomposition to enforce (biologically motivated) polyhedral constraints for parameters in Bayesian analysis. Recall that a direction of a polyhedron P is a vector $\boldsymbol{\xi}$ such that $\boldsymbol{\mu} + \delta\boldsymbol{\xi} \in P$ for all $\delta > 0$, for any $\boldsymbol{\mu} \in P$. The Minkowski-Weyl decomposition says that

$$P = \operatorname{conv}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k) + \operatorname{cone}(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_h),$$

$$\text{where } \operatorname{cone}(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_h) = \left\{ \sum_{\ell=1}^{h} \lambda_\ell \boldsymbol{\xi}_\ell : \boldsymbol{\lambda} \geq 0, \right\},$$

for extreme points $\boldsymbol{v}_\ell$ and extreme directions $\boldsymbol{\xi}_\ell$ of P. Danaher et al. (2012) propose a Dirichlet prior distribution for the simplex between the extreme points, while gamma priors are proposed for the positive coefficients on the extreme directions. Because the sets A_i are bounded polyhedra for the present problem, no directions are contained within the set and we need only consider the extreme points.

Figure 4.1 shows an example of the set $A(p, \boldsymbol{\pi})$ for $J = 2$ and $J = 3$, along with a random sample taken from the set assuming a $\text{Dirichlet}_k(1, \ldots, 1)$ distribution. When



(a) (b)

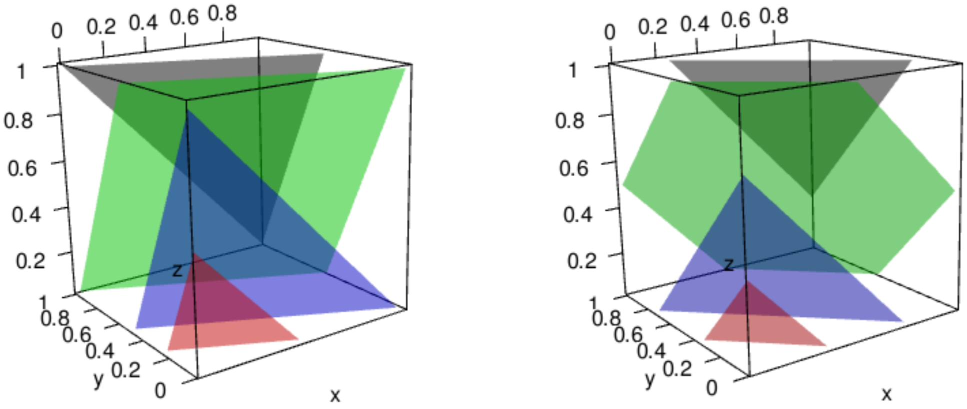
Figure 4.1: A sample drawn from A: (a) $n = 100$ with $J = 2$, $\boldsymbol{\pi} = (\frac{3}{4}, \frac{1}{4})$, $p = \frac{2}{3}$, and (b) $n = 300$ with $J = 3$, $\boldsymbol{\pi} = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$, $p = \frac{2}{3}$.

$J = 3$, Figure 4.2 shows how the set $A(p, \boldsymbol{\pi})$ changes as p is varied. Note that the number of vertices k may change, and so may the placement of the hyperplane segment. It is clear that for $J = 3$ it is possible for k to take on values at least in $\{3, 4, 5, 6\}$, and certainly $k = J$ need not hold.

We can now write the Mixture Link model as the hierarchy

$$\begin{aligned}
&T_i \mid \boldsymbol{\mu}_i, \boldsymbol{\pi} \overset{\text{ind}}{\sim} \text{BinMix}(m_i, \boldsymbol{\mu}_i, \boldsymbol{\pi}), \\
&\boldsymbol{\mu}_i = \boldsymbol{V}^{(i)} \boldsymbol{\lambda}^{(i)}, \quad \text{where } \boldsymbol{V}^{(i)} = (\boldsymbol{v}_1^{(i)} \cdots \boldsymbol{v}_{k_i}^{(i)}) \text{ are vertices of } A(p_i, \boldsymbol{\pi}), \\
&\boldsymbol{\lambda}^{(i)} \overset{\text{ind}}{\sim} \text{Dirichlet}_{k_i}(\boldsymbol{\alpha}^{(i)}). \qquad (4.8)
\end{aligned}$$

Notice that the dimension of $\boldsymbol{\alpha}^{(i)} = (\alpha_1^{(i)}, \ldots, \alpha_{k_i}^{(i)})$ may vary between observations, depending on $\boldsymbol{\pi}$ and p_i. Because our main interest is the regression case where $p_1, \ldots, p_n$ are not equal, we make the further assumption that $\boldsymbol{\alpha}^{(i)} = \kappa \mathbf{1}$ where $\mathbf{1} = (1, \ldots, 1)$ and $\kappa > 0$. The Dirichlet($\kappa \mathbf{1}$) distribution is sometimes referred to as Symmetric Dirichlet. There are also identifiability issues in letting the components of $\boldsymbol{\alpha}^{(i)}$ vary because the vertices in $\boldsymbol{V}^{(i)}$ are not strictly ordered, therefore it is difficult to maintain a correspondence between $\boldsymbol{v}_\ell^{(i)}$ and $\alpha_\ell^{(i)}$. Figure 4.3 shows Dirichlet distributions plotted for several settings of κ when $J = 3$. Notice that $\kappa = 1$ corresponds to the uniform distribution



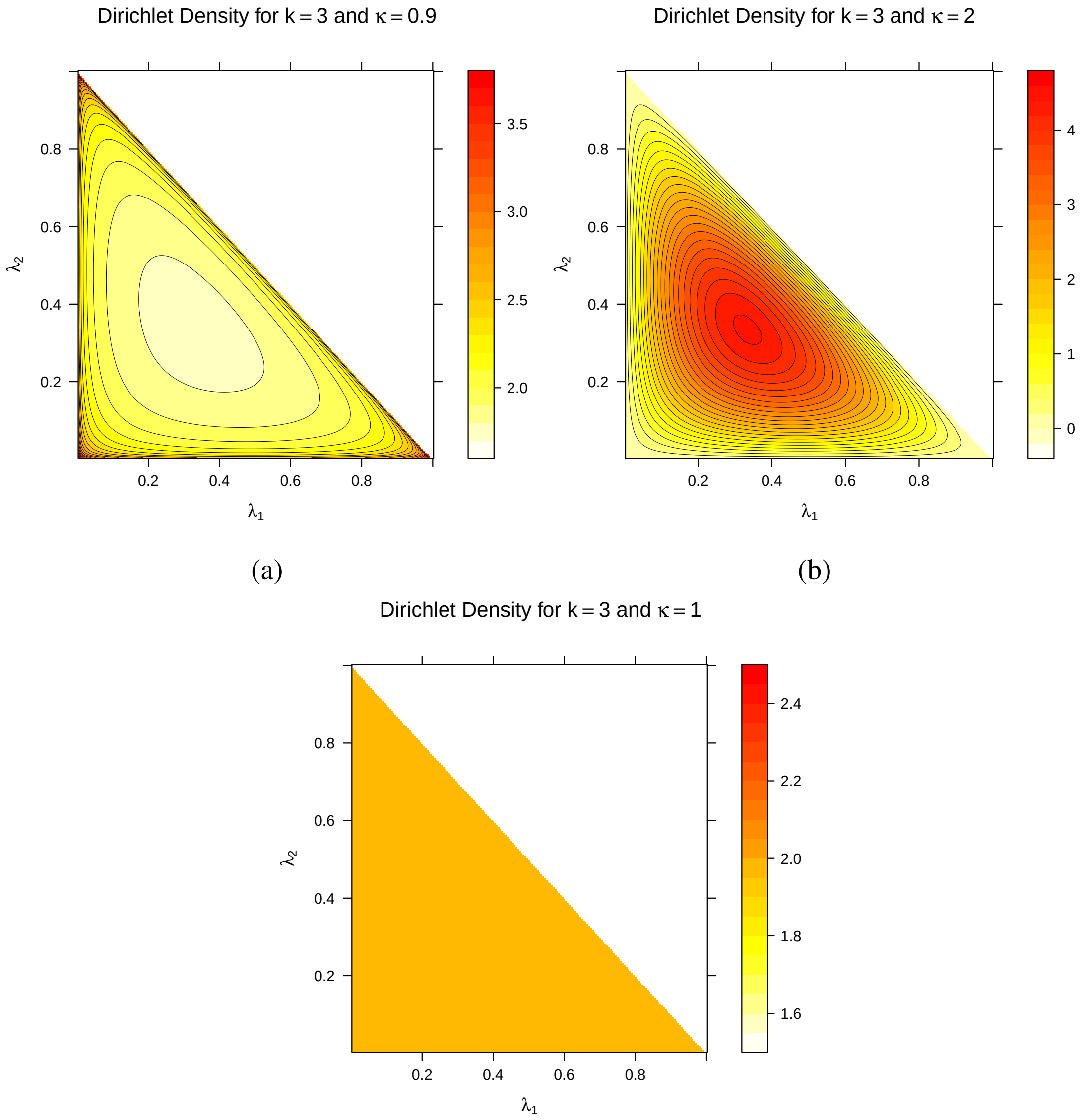
(a) $\boldsymbol{\pi} = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. (b) $\boldsymbol{\pi} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

Figure 4.2: The set $\{\boldsymbol{\mu} \in [0,1]^3 : \mu_1\pi_1 + \mu_2\pi_2 + \mu_3\pi_3 = p\}$ visualized with two different settings of $\boldsymbol{\pi}$. In each case, $p \in \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$ is shown (from front to back).

of $\boldsymbol{\lambda}^{(i)}$ on the simplex (and furthermore to a uniform distribution of $\boldsymbol{\mu}_i$ on A_i), while $0 < \kappa < 1$ results in more density focused toward the vertices than the interior, and $\kappa > 1$ yields more density in the interior of the simplex. The hierarchy (4.8) is parameterized by $\boldsymbol{\theta} = (p, \boldsymbol{\pi}, \kappa) \in \mathbb{R}^{1+(J-1)+1}$ if T_i are taken to be independent and identically distributed, or $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\pi}, \kappa) \in \mathbb{R}^{d+(J-1)+1}$ in the case of a regression. In a frequentist analysis, $\boldsymbol{\theta}$ will be a fixed but unknown parameter. A Bayesian analysis would put a prior on $\boldsymbol{\beta}$ or p, the main parameters of interest, and perhaps on $\boldsymbol{\pi}$ and κ as well. Denote the elements of $\boldsymbol{V}$ as $v_{j\ell}$, $\boldsymbol{v}_{j\cdot}^T$ as the jth row, and $\boldsymbol{v}_{\cdot\ell}$ as the ℓth column. The Mixture Link density is given by

$$\begin{aligned} f(t \mid m, p, \boldsymbol{\pi}, \kappa) &= \int \sum_{j=1}^{J} \pi_j \left\{ \binom{m}{t} \mu_j^t (1-\mu_j)^{m-t} \right\} \cdot f_A(\boldsymbol{\mu}) d\boldsymbol{\mu} \\ &= \binom{m}{t} \sum_{j=1}^{J} \pi_j \int_{\ell_j}^{u_j} w^t (1-w)^{m-t} \cdot f_{A^{(j)}}(w) dw \qquad (4.9) \end{aligned}$$

where $f_A(\boldsymbol{\mu})$ is the joint density of $\boldsymbol{\mu} = \boldsymbol{V}\boldsymbol{\lambda}$ on the set $A(p, \boldsymbol{\pi})$, $f_{A^{(j)}}(w)$ is the marginal



(a)

(b)

(c)

Figure 4.3: Dirichlet$_3(\boldsymbol{\lambda} \mid \kappa\mathbf{1})$ density for several settings of κ. Only λ_1 and λ_2 are plotted, as $\lambda_3 = 1 - \lambda_1 - \lambda_2$ is redundant.

density of $\mu_j = \boldsymbol{v}_{j\cdot}^T \boldsymbol{\lambda}$. The limits of integration are

$$\ell_j = \min\{v_{j1}, \ldots, v_{jk}\} \quad \text{and} \quad u_j = \max\{v_{j1}, \ldots, v_{jk}\} \quad \text{for } j = 1, \ldots, J.$$

The notation $T \sim \text{MixLink}_J(m_i, p, \boldsymbol{\pi}, \kappa)$ will be used to say that T is drawn from this distribution. The joint likelihood of the sample $T_i \overset{\text{ind}}{\sim} \text{MixLink}_J(m_i, p, \boldsymbol{\pi}, \kappa)$ for $i = 1, \ldots, n$ is then

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \binom{m_i}{t_i} \sum_{j=1}^{J} \pi_j \int w^{t_i}(1-w)^{m_i - t_i} \cdot f_{A_i^{(j)}}(w) dw \right\}. \tag{4.10}$$

Note that in Chapter 3, an important issue in the finite mixture was the distinguishability between subpopulations. In the binomial mixture case, this amounts to similarity between the elements of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_J)$. However, Mixture Link integrates $\boldsymbol{\mu}$ out according to the density f_A; hence, the similarity of the elements does not affect the distribution except in an aggregated sense.

4.4 Expectation and Variance under Mixture Link Model

It is customary to compute moments such as the expectation, variance, and moment-generating function when introducing a probability distribution. In the case of Mixture Link, the calculations will become useful in approximating the density, as discussed in Section 4.8. Suppose $T \sim \text{MixLink}_J(m, p, \boldsymbol{\pi}, \kappa)$. To compute expectations of T, it is helpful to consider the complete data distribution

$$\begin{aligned}
&T \mid \boldsymbol{\mu}, \boldsymbol{\pi}, (Z = j) \sim \text{Binomial}(m, \mu_j), \\
&Z \sim \text{Discrete}(1, \ldots, J; \boldsymbol{\pi}), \\
&\boldsymbol{\mu} = \boldsymbol{V}\boldsymbol{\lambda}, \quad \text{where } \boldsymbol{V} = (\boldsymbol{v}_1 \cdots \boldsymbol{v}_k) \text{ are vertices of } A, \\
&\boldsymbol{\lambda} \sim \text{Dirichlet}_k(\kappa \mathbf{1}),
\end{aligned}$$

where $\boldsymbol{\lambda}$ and Z are independently distributed. It can be verified that this is a complete data model for Mixture Link by integrating out Z and $\boldsymbol{\lambda}$. Now, since $T \sim$ Binomial(m, μ_j) given Z and $\boldsymbol{\lambda}$, we have

$$\begin{aligned}
\mathrm{E}(T) &= \mathrm{E}_{Z,\boldsymbol{\lambda}}[\mathrm{E}(T \mid Z, \boldsymbol{\lambda})] \\
&= \mathrm{E}_{Z,\boldsymbol{\lambda}}\left[\sum_{j=1}^{J} I(Z = j) m \mu_j\right] \\
&= m \sum_{j=1}^{J} \pi_j \, \mathrm{E}(\mu_j) \\
&= m \sum_{j=1}^{J} \pi_j \boldsymbol{e}_j^T \boldsymbol{V} \, \mathrm{E}(\boldsymbol{\lambda}) \\
&= m \sum_{j=1}^{J} \pi_j \boldsymbol{e}_j^T \boldsymbol{V} (k^{-1} \mathbf{1}) \\
&= m k^{-1} \sum_{j=1}^{J} \pi_j \boldsymbol{v}_{j.}^T \mathbf{1} \\
&= m \sum_{j=1}^{J} \pi_j \bar{v}_{j.}
\end{aligned}$$

where $\bar{v}_{j.}$ denotes the mean of the elements of $\boldsymbol{v}_{j.}$, and recalling that for $\boldsymbol{D} \sim$ Dirichlet$_k(\boldsymbol{\alpha})$, $\mathrm{E}(\boldsymbol{D}) = \boldsymbol{\alpha}/\alpha_0$ where $\alpha_0 = \sum_{\ell=1}^{k} \alpha_\ell$. Notice that $\mathrm{E}(T)$ is free of κ, and is a function of $\boldsymbol{\pi}$ and p as linear combination of the vertices. We may also write $\mathrm{E}(T) = mk^{-1}\boldsymbol{\pi}^T \boldsymbol{V} \mathbf{1}$.

Now, computing the second moment of T,

$$\begin{aligned}
\mathrm{E}(T^2) &= \mathrm{E}_{Z,\boldsymbol{\lambda}}[\mathrm{E}(T^2 \mid Z, \boldsymbol{\lambda})] \\
&= \mathrm{E}_{Z,\boldsymbol{\lambda}} \left\{ \sum_{j=1}^{J} I(Z=j) \left[m\mu_j(1-\mu_j) + m^2\mu_j^2 \right] \right\} \\
&= \sum_{j=1}^{J} \pi_j \left[m\,\mathrm{E}(\mu_j) - m\,\mathrm{E}(\mu_j^2) + m^2\,\mathrm{E}(\mu_j^2) \right] \\
&= m \sum_{j=1}^{J} \pi_j\,\mathrm{E}(\mu_j) + m(m-1) \sum_{j=1}^{J} \pi_j\,\mathrm{E}(\mu_j^2) \\
&= \mathrm{E}(T) + m(m-1) \sum_{j=1}^{J} \pi_j\,\mathrm{E}(\mu_j^2).
\end{aligned}$$

Now we have

$$\begin{aligned}
\mathrm{E}(\mu_j^2) &= \boldsymbol{e}_j^T \boldsymbol{V}\,\mathrm{E}(\boldsymbol{\lambda}\boldsymbol{\lambda}^T) \boldsymbol{V}^T \boldsymbol{e}_j \\
&= \boldsymbol{e}_j^T \boldsymbol{V} \left[\mathrm{Var}(\boldsymbol{\lambda}) + \mathrm{E}(\boldsymbol{\lambda})\,\mathrm{E}(\boldsymbol{\lambda}^T) \right] \boldsymbol{V}^T \boldsymbol{e}_j \\
&= \boldsymbol{v}_{j\cdot}^T \left[\frac{\boldsymbol{I} + \kappa \boldsymbol{1}\boldsymbol{1}^T}{k(1+\kappa k)} \right] \boldsymbol{v}_{j\cdot} \\
&= \frac{\boldsymbol{v}_{j\cdot}^T \boldsymbol{v}_{j\cdot} + \kappa k^2 \bar{v}_{j\cdot}^2}{k(1+\kappa k)}.
\end{aligned}$$

We have used the fact that

$$\begin{aligned}
&\mathrm{Var}(\boldsymbol{\lambda}) = \frac{\alpha_0\,\mathrm{Diag}(\boldsymbol{\alpha}) - \boldsymbol{\alpha}\boldsymbol{\alpha}^T}{\alpha_0^2(\alpha_0+1)} = \frac{k\kappa^2 \boldsymbol{I} - \kappa^2 \boldsymbol{1}\boldsymbol{1}^T}{k^2\kappa^2(k\kappa+1)} = \frac{k\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T}{k^2(k\kappa+1)}, \quad \text{and} \\
&\mathrm{E}(\boldsymbol{\lambda}\boldsymbol{\lambda}^T) = \mathrm{Var}(\boldsymbol{\lambda}) + \mathrm{E}(\boldsymbol{\lambda})\,\mathrm{E}(\boldsymbol{\lambda}^T) = \frac{\alpha_0\,\mathrm{Diag}(\boldsymbol{\alpha}) - \boldsymbol{\alpha}\boldsymbol{\alpha}^T}{\alpha_0^2(\alpha_0+1)} + \frac{\boldsymbol{\alpha}\boldsymbol{\alpha}^T}{\alpha_0^2} = \frac{\mathrm{Diag}(\boldsymbol{\alpha}) + \boldsymbol{\alpha}\boldsymbol{\alpha}^T}{\alpha_0(\alpha_0+1)}.
\end{aligned}$$

Therefore we obtain the second moment and second factorial moment

$$\begin{aligned} \mathrm{E}(T^2) &= \mathrm{E}(T) + m(m-1)\sum_{j=1}^{J}\pi_j \frac{\boldsymbol{v}_{j.}^T\boldsymbol{v}_{j.} + \kappa k^2 \bar{v}_{j.}^2}{k(1+\kappa k)}, \\ \mathrm{E}\left[\frac{T(T-1)}{m(m-1)}\right] &= \sum_{j=1}^{J}\pi_j \frac{\boldsymbol{v}_{j.}^T\boldsymbol{v}_{j.} + \kappa k^2 \bar{v}_{j.}^2}{k(1+\kappa k)}, \end{aligned}$$

and the variance

$$\begin{aligned} \mathrm{Var}(T) &= \mathrm{E}(T^2) - \mathrm{E}^2(T) \\ &= m\sum_{j=1}^{J}\pi_j\bar{v}_{j.} + m(m-1)\sum_{j=1}^{J}\pi_j\frac{\boldsymbol{v}_{j.}^T\boldsymbol{v}_{j.} + \kappa(k\bar{v}_{j.})^2}{k(1+\kappa k)} - \left(m\sum_{j=1}^{J}\pi_j\bar{v}_{j.}\right)^2 \\ &= m\sum_{j=1}^{J}\pi_j\bar{v}_{j.}\left(1 - m\sum_{j=1}^{J}\pi_j\bar{v}_{j.}\right) + m(m-1)\sum_{j=1}^{J}\pi_j\frac{\boldsymbol{v}_{j.}^T\boldsymbol{v}_{j.} + \kappa(k\bar{v}_{j.})^2}{k(1+\kappa k)}. \end{aligned} \tag{4.11}$$

Notice that as $\kappa \to \infty$, $\mathrm{Var}(\boldsymbol{v}_{j.}^T\boldsymbol{\lambda}) \to 0$, therefore the distribution of $\boldsymbol{v}_{j.}^T\boldsymbol{\lambda}$ approaches a point mass at $\mathrm{E}(\boldsymbol{v}_{j.}^T\boldsymbol{\lambda}) = \bar{v}_{j.}$ and the Mixture Link distribution becomes

$$f(t \mid m, p, \boldsymbol{\pi}) = \binom{m}{t}\sum_{j=1}^{J}\pi_j\bar{v}_{j.}^t(1-\bar{v}_{j.})^{m-t}, \tag{4.12}$$

Therefore, as $\kappa \to \infty$, the Mixture Link density becomes a finite mixture of binomials with m trials whose jth probability of success $\bar{v}_{j.}$ is the mean of the jth coordinate of the k vertices of $A(p, \boldsymbol{\pi})$. In this case the variance becomes

$$\mathrm{Var}(T) = m\sum_{j=1}^{J}\pi_j\bar{v}_{j.}\left(1 - m\sum_{j=1}^{J}\pi_j\bar{v}_{j.}\right) + m(m-1)\sum_{j=1}^{J}\pi_j\bar{v}_{j.}^2.$$

Also notice that as $\kappa \to 0$,

$$\mathrm{Var}(T) \to m \sum_{j=1}^{J} \pi_j \bar{v}_{j\cdot} \left(1 - m \sum_{j=1}^{J} \pi_j \bar{v}_{j\cdot}\right) + m(m-1) \sum_{j=1}^{J} \pi_j \frac{\boldsymbol{v}_{j\cdot}^T \boldsymbol{v}_{j\cdot}}{k}.$$

The moment generating function (MGF) of T can be obtained as an integral by noticing that

$$\mathrm{E}(e^{\gamma T} \mid \boldsymbol{\mu}, Z = j) = [\mu_j e^{\gamma} + 1 - \mu_j]^m$$

is the MGF of binomial. Hence

$$\begin{aligned} \phi_T(\gamma) &= \mathrm{E}_{\boldsymbol{\mu}}\{\mathrm{E}_Z[\mathrm{E}(e^{\gamma T} \mid \boldsymbol{\mu}, Z)\} \\ &= \mathrm{E}_{\boldsymbol{\mu}} \left\{ \sum_{j=1}^{J} \pi_j [\mu_j e^{\gamma} + 1 - \mu_j]^m \right\} \\ &= \sum_{j=1}^{J} \pi_j \int_{\ell_j}^{u_j} [w e^{\gamma} + 1 - w]^m f_{A_j}(w) dw \end{aligned}$$

is the MGF of T.

4.5 Finding the vertices of A

Computation of the Mixture Link density and its moments depends on the vertices of the set A. In this section we will see how the vertices can be determined; first for the simple case when $J = 2$, then extending to $J > 2$. For the case $J = 2$, it is easy to identify the vertices of A graphically by following the line to the points at which it intersects the unit rectangle. An illustration is given in Figure 4.4, and the result is stated now as a lemma.

Lemma 4.3. *Suppose $J = 2$ and A has two distinct vertices $\boldsymbol{v}_1, \boldsymbol{v}_2$. Then the vertices are* *given by*

$$\boldsymbol{v}_1 = \begin{cases} \left(\frac{1}{\pi_1}p, 0\right), & \textit{if } \frac{1}{\pi_1}p \leq 1 \\ \left(1, \frac{1}{\pi_2}(p - \pi_1)\right), & \textit{otherwise}, \end{cases}$$

$$\boldsymbol{v}_2 = \begin{cases} \left(\frac{1}{\pi_1}(p - \pi_2), 1\right), & \textit{if } \frac{1}{\pi_1}(p - \pi_2) \geq 0 \\ \left(0, \frac{1}{\pi_2}p\right), & \textit{otherwise}, \end{cases}$$

where $\pi_2 = 1 - \pi_1$.

Proof. Using $\mu_1\pi_1 + \mu_2\pi_2 = p$ we have

$$\mu_1 = \frac{1}{\pi_1}(p - \mu_2\pi_2) \quad \text{and} \quad \mu_2 = \frac{1}{\pi_2}(p - \mu_1\pi_1), \tag{4.13}$$

where $\mu_1 \in [0, 1]$ and $\mu_2 \in [0, 1]$ must hold. To obtain $\boldsymbol{v}_1$, take μ_1 as large as possible noting expressions (4.13). If $\mu_1 = 1$ is a valid solution (i.e. a point in A), then $\mu_2 = \frac{1}{\pi_1}(p - \pi_2)$. Otherwise take μ_2 as small as possible to maximize μ_1; this yields $\mu_1 = \frac{1}{\pi_1}p$ and $\mu_2 = 0$. A similar argument taking μ_1 as small as possible yields $\boldsymbol{v}_2$. $\square$

We may also locate the vertices $(\boldsymbol{v}_1, \boldsymbol{v}_2)$ systematically in the following way. Fix μ_2 at both 0 and 1, and solve for μ_1 so that (μ_1, μ_2) is on the hyperplane. Now fix μ_1 at both 0 and 1, and solve for μ_2 so that (μ_1, μ_2) is on the hyperplane. For all four points $\boldsymbol{\mu} = (\mu_1, \mu_2)$, if $\boldsymbol{\mu} \in A$, then it is a vertex of A. We will see later that this idea generalizes to $J > 2$. Note that it is also possible to have $k = 1$ vertices when $J = 2$. For example, if $\boldsymbol{\pi} = (1/2, 1/2)$ and $p = 1$, then $\mu_1 = 1, \mu_2 = 1$ is the only solution to $\mu_1\pi_1 + \mu_2\pi_2 = p$ in $[0, 1]^2$, and therefore A is a singleton set.

For the general ($J > 2$) case, the following lemma characterizes points in A which need to be considered when searching for the extreme points.

Lemma 4.4 (Characterization of Extreme Points of A)**.** *Suppose* $\boldsymbol{v} = (v_1, \ldots, v_J)$ *is a point in* A *with two or more components* $v_j \notin \{0, 1\}$. *Then* $\boldsymbol{v}$ *is not an extreme point of*

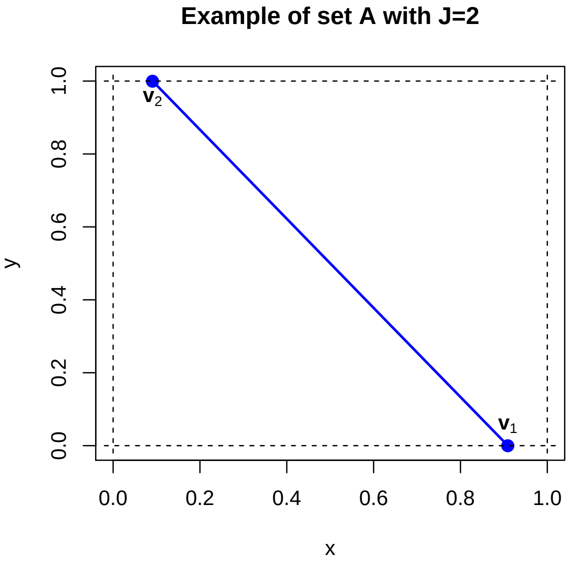


Figure 4.4: Example set A with $\boldsymbol{\pi} = (\frac{11}{20}, \frac{9}{20})$ and $p = \frac{1}{2}$. The vertices are located at $\boldsymbol{v}_1 = (\frac{10}{11}, 0)$ and $\boldsymbol{v}_2 = (\frac{1}{11}, 1)$.

A.

Proof. Suppose WLOG that $\boldsymbol{v} \in A$ with $v_1 \in (0,1)$ and $v_2 \in (0,1)$. We have that

$$\begin{aligned} \boldsymbol{v}^T\boldsymbol{\pi} = p \quad & \iff \quad v_1\pi_1 + v_2\pi_2 + (v_3\pi_3 + \cdots + v_J\pi_J) = p \\ & \iff \quad v_1\pi_1 + v_2\pi_2 = p^*, \end{aligned}$$

where $p^* = p - (v_3\pi_3 + \cdots + v_J\pi_J)$. We can now use Lemma 4.3 to obtain vertices, say $\boldsymbol{a}$ and $\boldsymbol{b}$, of the line segment

$$L = \left\{(\mu_1, \mu_2, v_3, \ldots, v_J) \in [0,1]^J : \mu_1\pi_1 + \mu_2\pi_2 = p^*\right\},$$

where $(v_3, \dots, v_J)$ are held fixed and only (μ_1, μ_2) may vary. Explicitly, we have

$$\boldsymbol{a} = \begin{cases} \left(\frac{1}{\pi_1}p^*, 0, v_3, \dots, v_J\right), & \text{if } \frac{1}{\pi_1}p^* \leq 1 \\ \left(1, \frac{1}{\pi_2}(p^* - \pi_1), v_3, \dots, v_J\right), & \text{otherwise,} \end{cases}$$

$$\boldsymbol{b} = \begin{cases} \left(\frac{1}{\pi_1}(p^* - \pi_2), 1, v_3, \dots, v_J\right), & \text{if } \frac{1}{\pi_1}(p^* - \pi_2) \geq 0 \\ \left(0, \frac{1}{\pi_2}p^*, v_3, \dots, v_J\right), & \text{otherwise.} \end{cases}$$

By construction, we have that $\boldsymbol{v}$ is in the line segment between $\boldsymbol{a}$ and $\boldsymbol{b}$, with $\boldsymbol{a} \neq \boldsymbol{b}$. Furthermore since $L \subseteq A$, we have that $\boldsymbol{a}, \boldsymbol{b} \in A$. Therefore, $\boldsymbol{v}$ can not be an extreme point of A. □

Lemma 4.4 suggests that in searching for extreme points, we must only consider those with at most one component not equal to 0 or 1. This can be used to formulate a simple algorithm. The idea is as follows:

- Consider the jth dimension, which is the one we will permit to take values in $[0, 1]$ other than $\{0, 1\}$.
- For $(\mu_1, \dots, \mu_{j-1}, \mu_{j+1}, \dots, \mu_J) \in \{0, 1\}^{J-1}$, find μ_j^* that solves $\boldsymbol{\mu}^T \boldsymbol{\pi} = p$.
- The point $(\mu_1, \dots, \mu_{j-1}, \mu_j^*, \mu_{j+1}, \dots, \mu_J) \in \{0, 1\}^J$ is a vertex of A if and only if it is in A.

The procedure is given formally as Algorithm 1. Notice that it checks $J \cdot 2^{J-1}$ points, and therefore becomes impractical for large J. It would be useful if some parts of the search could be excluded from consideration.

Remark 4.5 (Vertex finding in general polyhedra)**.** Bazaraa et al. (2009, Chapter 2) discusses vertex identification for the general polyhedron

$$P = \{x \in \mathbb{R}^n : Cx \leq b, x \geq 0\},$$

where C is a $m \times n$ matrix and b is a $m \times 1$ vector. Denote c_i^T as the ith row of C. There are $m + n$ total restrictions in P: $c_1^T x \leq b_1, \ldots, c_m^T x \leq b_m$ and $x_1 \geq 0, \ldots, x_n \geq 0$. The vertices can be characterized as follows: construct an $n \times n$ matrix $\widetilde{C}$ from n of the $m + n$ rows of C, and a vector $\tilde{b}$ out of the corresponding entries of b. The vertices v of P occur when $\widetilde{C}$ is nonsingular, so that the restrictions are linearly independent, and when $v = \widetilde{C}^{-1}\tilde{b}$ is a point in P. A simple algorithm is therefore to inspect all $\binom{m+n}{n}$ combinations of restrictions, and for each combination, to form $\widetilde{C}$ and $\tilde{b}$ and save all points where $\widetilde{C}^{-1}\tilde{b} \in P$.

For the Mixture Link problem, we may rewrite A as

$$\begin{aligned} A &= \{\boldsymbol{\mu} \in [0,1]^J : \boldsymbol{\mu}^T \boldsymbol{\pi} = p\} \\ &= \{\boldsymbol{\mu} \in \mathbb{R}^J : \boldsymbol{\mu} \geq \mathbf{0}, \boldsymbol{\mu} \leq \mathbf{1}, \boldsymbol{\mu}^T \boldsymbol{\pi} \leq p, \boldsymbol{\mu}^T \boldsymbol{\pi} \geq p\} \\ &= \{\boldsymbol{\mu} \in \mathbb{R}^J : \boldsymbol{C}\boldsymbol{\mu} \leq \boldsymbol{b}, \boldsymbol{\mu} \geq \mathbf{0}\}, \end{aligned}$$

where

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{I} \\ \boldsymbol{\pi}^T \\ -\boldsymbol{\pi}^T \end{pmatrix} \in \mathbb{R}^{(J+2) \times J}, \quad \text{and} \quad \boldsymbol{b} = \begin{pmatrix} \mathbf{1} \\ p \\ -p \end{pmatrix} \in \mathbb{R}^{(J+2) \times 1}.$$

To carry out the general vertex finding algorithm, we must inspect $\binom{2J+2}{J}$ combinations of restrictions. Table 4.1 compares this to the number of candidate vertices considered by Algorithm 1. The general vertex finding method requires significantly more steps, and the steps are more computationally involved: each requiring a check for singularity and possibly a linear solve. Therefore, Algorithm 1 provides a huge computational improvement when J is not very small. We expect that $J = 12$ mixture components (the last entry in the table) is too many to use in practice. However, even for $J = 3$ or $J = 4$, the savings in computation would give a noticeable improvement when the density must be computed many times, such as in an iterative algorithm for estimation. More involved methods for vertex finding in polyhedra have been investigated in the mathematics literature; a survey is given in (Matheiss and Rubin, 1980).

Remark 4.6 (Differentiability of the vertices)**.** It is clear that the vertices of $A(p, \boldsymbol{\pi})$ are not differentiable over all p and $\boldsymbol{\pi}$. The number of vertices k can change so that a particular vertex may suddenly appear or disappear as p and $\boldsymbol{\pi}$ vary. Even in the simple case where $k = 2$ is fixed, the vertices are not differentiable at $p, \boldsymbol{\pi}$ where $p = \pi_1$ or $p = \pi_2$; this can be verified using the explicit expressions in Lemma 4.3.

Algorithm 1 Find vertices of the set $A(p, \boldsymbol{\pi})$.

```
function FINDVERTICES($p, \boldsymbol{\pi}$)
    $\mathcal{V} \leftarrow \varnothing$
    for $j = 1, \ldots, J$ do
        if $\pi_j > 0$ then
            for all $\boldsymbol{\mu}_{-j} \in \{0, 1\}^{J-1}$ do
                $\mu_j^* \leftarrow \frac{1}{\pi_j} \left[ p - \boldsymbol{\mu}_{-j}^T \boldsymbol{\pi}_{-j} \right]$
                $\boldsymbol{v}^* \leftarrow (\mu_1, \ldots, \mu_{j-1}, \mu_j^*, \mu_{j+1}, \ldots, \mu_J)$
                if $\boldsymbol{v}^* \in A$ then
                    $\mathcal{V} \leftarrow \mathcal{V} \cup \boldsymbol{v}^*$
    return $\mathcal{V}$
```

4.6 Computing the Mixture Link Density

A first step in making use of the Mixture Link model is being able to compute the density efficiently and in turn to compute the likelihood. One required step, given p and $\boldsymbol{\pi}$, is to compute the vertices $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ of $A(p, \boldsymbol{\pi})$ as discussed previously. Next, the density (4.9) involves an integration over a random variable which is a linear combination of a Dirichlet distributed random vector.

Provost and Cheong (2000) note that the distribution of a linear combination of a Dirichlet random vector is closely related to a well studied distribution — the linear combination of chi-square random variables. To see this, recall that if $X_j \overset{\text{ind}}{\sim} \chi^2_{v_j}$ for

Table 4.1: Number of steps required for finding vertices of A in dimension J. The column Polyhedron denotes the algorithm discussed in Remark 4.5 for general polyhedra which requires $\binom{2J+2}{J}$ steps. The column FindVertices denotes Algorithm 1 which requires $J \cdot 2^{J-1}$ steps.

J	Polyhedron	FindVertices
1	4	1
2	15	4
3	56	12
4	210	32
5	792	80
6	3,003	192
7	11,440	448
8	43,758	1,024
9	167,960	2,304
10	646,646	5,120
11	2,496,144	11,264
12	9,657,700	24,576

$j = 1, \ldots, k$, then

$$\left(\frac{X_1}{\sum_{j=1}^{k} X_j}, \ldots, \frac{X_k}{\sum_{j=1}^{k} X_j} \right) \sim \text{Dirichlet}_k(\boldsymbol{\alpha})$$

where $\alpha_j = v_j/2$. Details are given in (Kotz et al., 2000), and are also reproduced in Appendix A. Now if $\boldsymbol{\lambda} \sim \text{Dirichlet}_k(\boldsymbol{\alpha})$, we may write the distribution of a linear combination $\boldsymbol{c}^T\boldsymbol{\lambda}$ as

$$F_{\boldsymbol{c}^T\boldsymbol{\lambda}}(x) = \mathrm{P}\left(\sum_{j=1}^{k} c_j \lambda_j \leq x \right) = \mathrm{P}\left(\sum_{j=1}^{k} c_j \frac{X_j}{\sum_{\ell=1}^{k} X_\ell} \leq x \right) = \mathrm{P}\left(\sum_{j=1}^{k} (c_j - x) X_j \leq 0 \right). \tag{4.14}$$

Provost and Cheong show how this probability can be computed through an expression given by Imhof (1961). Imhof obtains the CDF of a linear combination of chi-squares $\boldsymbol{b}^T\boldsymbol{X}$ using the inversion formula for the CDF

$$F(x) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty t^{-1} \mathfrak{Im}\{e^{-itx}\phi(t)\} dt,$$

and the characteristic function

$$\phi_{\boldsymbol{b}^T \boldsymbol{X}}(t) = \prod_{j=1}^{k} (1 - 2b_j it)^{-v_j/2},$$

where $\mathfrak{Im}(z)$ denotes the imaginary part of z. The exact expression

$$\mathrm{P}\left(\sum_{j=1}^{k} b_j X_j \le x \right) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{\sin(\frac{1}{2}\sum_{j=1}^{k} v_j \arctan(b_j u) - \frac{1}{2}xu)}{u \prod_{j=1}^{k} (1 + b_j^2 u^2)^{v_j/4}} du \tag{4.15}$$

is obtained, along with error bounds for numerical computation of the integral. Expression (4.15) may be combined with (4.14) to give

$$F_{\boldsymbol{c}^T \boldsymbol{\lambda}}(x) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{\sin(\sum_{j=1}^{k} \alpha_j \arctan((c_j - x)u))}{u \prod_{j=1}^{k} (1 + (c_j - x)^2 u^2)^{\alpha_j/2}} du. \tag{4.16}$$

as the CDF for $\boldsymbol{c}^T \boldsymbol{\lambda}$. For this work, we compute (4.16) using the `imhof` function from the `CompQuadForm` package (Duchesne and Micheaux, 2010) for `R`, and then evaluate the density of $\boldsymbol{c}^T \boldsymbol{\lambda}$ by numerical differentiation

$$f_{\boldsymbol{c}^T \boldsymbol{\lambda}}(x) = \frac{F_{\boldsymbol{c}^T \boldsymbol{\lambda}}(x + \varepsilon) - F_{\boldsymbol{c}^T \boldsymbol{\lambda}}(x)}{\varepsilon}$$

for a small $\varepsilon > 0$. In this way, the density of $\mu_j = \boldsymbol{e}_j^T \boldsymbol{V} \boldsymbol{\lambda}$ is obtained, where $\boldsymbol{e}_j$ represents the jth column of an identity matrix of the appropriate dimension. We use `integrate` to compute the Mixture Link density (4.9), where the integral is taken with respect to the numerically computed density $f_A(\mu_j)$. Hence, two one-dimensional numerical integrations and a numerical differentiation are used for a single evaluation of the Mixture Link density. Although this straightforward method provides a general way to compute the density, we have found it to be unacceptably slow for applications involving data analysis or simulation

As briefly mentioned by Provost and Cheong (2000), in the case of $k = 2$ vertices there is again a closed form for the density. Suppose $X_j \overset{\text{ind}}{\sim} \chi^2_{v_j}$ for $j = 1, 2$ and notice that

$$
\begin{aligned}
Z = \frac{a_1 X_1 + a_2 X_2}{X_1 + X_2} = \frac{a_1 - a_2 + a_2 + a_2 X_2/X_1}{1 + X_2/X_1} = \frac{a_1 - a_2}{1 + X_2/X_1} + a_2 = \frac{a_1 - a_2}{1 + \frac{v_2}{v_1} F} + a_2 \\
\iff F = \frac{v_1}{v_2} \left(\frac{a_1 - a_2}{Z - a_2} - 1 \right)
\end{aligned}
$$

where $F = \frac{X_1/v_1}{X_2/v_2}$ follows an F-distribution with degrees of freedom (v_1, v_2). Notice that $F \in (0, \infty)$ implies that $Z \in (a_2, a_1)$, assuming that $a_2 < a_1$. The case $a_1 = a_2$ need not be considered since Z becomes a point mass at $a_1 = a_2$. The Jacobian of this transformation is given by

$$
\frac{\partial F}{\partial Z} = -\frac{v_1}{v_2} \frac{a_1 - a_2}{(Z - a_2)^2},
$$

and therefore the density of Z can be written as

$$
\begin{aligned}
f_Z(z) = f_F(F) \cdot |\partial F / \partial Z| &= \frac{(v_2/v_1)^{v_2/2}}{B\left(\frac{v_2}{2}, \frac{v_1}{2}\right)} F^{v_2/2 - 1} \left(1 + \frac{v_2}{v_1} F\right)^{-v_1/2 - v_2/2} \left[\frac{v_1}{v_2} \frac{a_2 - a_1}{(z - a_2)^2} \right] \\
&= \frac{(a_1 - z)^{v_2/2 - 1} (z - a_2)^{v_1/2 - 1} (a_1 - a_2)^{1 - v_1/2 - v_2/2}}{B\left(\frac{v_2}{2}, \frac{v_1}{2}\right)}.
\end{aligned}
$$

Therefore, when $\boldsymbol{\lambda} \sim \text{Dirichlet}_2(\boldsymbol{\alpha})$, the density for $c_1 \lambda_1 + c_2 \lambda_2$ is

$$
f_{\boldsymbol{c}^T \boldsymbol{\lambda}}(z) = \frac{(c_1 - z)^{\alpha_2 - 1} (z - c_2)^{\alpha_1 - 1} (c_1 - c_2)^{1 - \alpha_1 - \alpha_2}}{B(\alpha_2, \alpha_1)}. \tag{4.17}
$$

Closed form expressions for the density when $k = 3$ and $k = 4$ are also discussed by Provost and Cheong (2000), but they are progressively less convenient to compute.

We have focused on obtaining the linear combination of Dirichlet density, but our real objective is the density of the Mixture Link distribution itself. Obtaining a closed form the Mixture Link density is possible in some simple situations.

Trivial case: $m = 1$. It should first be noted that the Bernoulli Mixture Link model, obtained when $m = 1$, simplifies in a trivial way since,

$$
\begin{aligned}
f(t \mid m, p, \boldsymbol{\pi}, \kappa) &= \int \left\{ \sum_{j=1}^{J} \pi_j \mu_j^t (1-\mu_j)^{1-t} \right\} f_A(\boldsymbol{\mu}) d\boldsymbol{\mu} \\
&= \begin{cases} \displaystyle\int \boldsymbol{\mu}^T \boldsymbol{\pi} f_A(\mu_j) d\boldsymbol{\mu} = p, & \text{if } t = 1, \\ \displaystyle\int (1 - \boldsymbol{\mu}^T \boldsymbol{\pi}) f_A(\mu_j) d\boldsymbol{\mu} = 1 - p & \text{if } t = 0. \end{cases} \\
&= p^t (1-p)^{1-t}.
\end{aligned}
$$

Therefore, in this case, Mixture Link is equivalent to a usual Bernoulli distribution. However, if $m \geq 2$ so that there is more than one Bernoulli trial, the two models no longer coincide. This result suggests that extra variation modeled by Mixture Link comes from the interdependence among the m trials.

Simple case: $J = 2$ and $\kappa = 1$. In the case $J = 2$, the set A has either one or two vertices. When there is only one vertex, $A = \{\boldsymbol{\mu}\}$ is a singleton set, and we immediately obtain the finite mixture density

$$
f(t \mid m, \boldsymbol{\theta}) = \sum_{j=1}^{J} \pi_j \binom{m}{t} \mu_j^t (1-\mu_j)^{m-t}
$$

evaluated at $\boldsymbol{\mu}$. Suppose now that there are two vertices, say $\boldsymbol{v}_1 = \begin{pmatrix} v_{11} \\ v_{12} \end{pmatrix}$ and $\boldsymbol{v}_2 = \begin{pmatrix} v_{21} \\ v_{22} \end{pmatrix}$. Then we have

$$
A = \{\lambda \boldsymbol{v}_1 + (1-\lambda)\boldsymbol{v}_2 : \lambda \in [0,1]\}
$$

so that any $\boldsymbol{\mu} \in A$ can be written as

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \lambda \boldsymbol{v}_1 + (1-\lambda)\boldsymbol{v}_2 = \begin{pmatrix} \lambda v_{11} + (1-\lambda)v_{21} \\ \lambda v_{12} + (1-\lambda)v_{22} \end{pmatrix}.$$

We may then write the density as

$$\begin{aligned} f(t \mid m, \boldsymbol{\theta}) &= \int \left\{ \sum_{j=1}^{2} \binom{m}{t} \pi_j \mu_j^t (1-\mu_j)^{m-t} \right\} \cdot f_A(\mu_j) d\mu_j \\ &= \binom{m}{t} \sum_{j=1}^{2} \pi_j \int_0^1 [\lambda v_{1j} + (1-\lambda)v_{2j}]^t [1 - (\lambda v_{1j} + (1-\lambda)v_{2j})]^{m-t} d\lambda, \end{aligned} \tag{4.18}$$

where $\lambda \sim \mathrm{U}(0,1)$. Now consider the transformation of the integrals in (4.18) using $w = \lambda v_{1j} + (1-\lambda)v_{2j}$ which gives

$$\begin{aligned} f(t \mid m, \boldsymbol{\theta}) &= \binom{m}{t} \sum_{j=1}^{2} \pi_j \frac{1}{v_{1j} - v_{2j}} \int_{v_{2j}}^{v_{1j}} w^t (1-w)^{m-t} dw \\ &= \binom{m}{t} \sum_{j=1}^{2} \pi_j \frac{\mathrm{B}_{v_{1j}}(t+1, m-t+1) - \mathrm{B}_{v_{2j}}(t+1, m-t+1)}{v_{1j} - v_{2j}}, \end{aligned} \tag{4.19}$$

where $\mathrm{B}_x(\alpha, \beta) = \int_0^x w^{\alpha-1}(1-w)^{\beta-1} dw$ is the incomplete beta function. The expression (4.19) is routine to compute with statistical software once the two vertices have been determined.

Remark 4.7 (Identifiability)**.** There is a natural invariance in the Mixture Link density to the order of the elements of $\boldsymbol{\pi}$. Consider evaluating (4.19) with a fixed t and m using $\boldsymbol{\pi} = (\pi, 1-\pi)$ and $\boldsymbol{\pi} = (1-\pi, \pi)$; the vertices $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ switch roles in the two representations, but the density (4.19) is invariant to the change. For the general Mixture

Link density,

$$f(t \mid m, p, \boldsymbol{\pi}, \kappa) = \binom{m}{t} \sum_{j=1}^{J} \pi_j \int_{\ell_j}^{u_j} w^t (1-w)^{m-t} \cdot f_{A^{(j)}}(w) dw,$$

suppose that π_j and $\pi_{j'}$ are swapped for some $j, j' \in \{1, \ldots, J\}$. Recall that $f_{A^{(j)}}$ is the density of $\boldsymbol{v}_{j\cdot}^T \boldsymbol{\lambda}$, and that $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ are solutions to

$$\mu_1 \pi_1 + \cdots + \mu_J \pi_J = p, \quad 0 \leq \mu_j \leq 1.$$

Therefore when π_j swaps with $\pi_{j'}$, the jth and j'th element of each $\boldsymbol{v}_\ell$ are swapped and therefore the distributions of $\boldsymbol{v}_{j\cdot}^T \boldsymbol{\lambda}$ and $\boldsymbol{v}_{j'\cdot}^T \boldsymbol{\lambda}$ swap. It also follows that (ℓ_j, u_j) swaps with $(\ell_{j'}, u_{j'})$ because ℓ_j and u_j are the smallest and largest element of $\boldsymbol{v}_{j\cdot}$. Therefore, the general Mixture Link density is invariant to permutation of the elements of $\boldsymbol{\pi}$. This is similar to the label switching problem in the usual finite mixture (McLachlan and Peel, 2000), where the concept of identifiability is relaxed slightly to allow components of the mixture to be permuted. We can add the constraint

$$\pi_1 < \cdots < \pi_J$$

to avoid ambiguity in the Mixture Link likelihood due to permutation invariance.

The general question of identifiability for the Mixture Link model still remains to be addressed. It is clear that it does not always hold. Consider the trivial case $m = 1$ and recall that the density simplifies to

$$f(t \mid m, p, \boldsymbol{\pi}, \kappa) = p^t (1-p)^{1-t}$$

when no regression is linked. This expression is free of $\boldsymbol{\pi}$ and κ, and therefore there is no hope of using the data to identify those parameters. Therefore, we must have at least

$m > 1$ for identifiability to possibly hold. Recall that in the simple binomial/multinomial finite mixture with m trials and J subpopulations, a necessary and sufficient condition for identifiability is that $m \geq 2J - 1$; refer to Section 2.2.

4.7 Density Comparison between Mixture Link with other Binomial Models with Extra Variation

To understand the utility of Mixture Link for modeling overdispersion in practice, we now examine some plots of the density. Here we consider the distribution $\text{MixLink}_J(m, p, \boldsymbol{\pi}, \kappa)$, i.e. without regression. Plotted in Figures 4.5 and 4.6 are the densities for the RCB and BB distributions, respectively, which were introduced in Section 4.2. For each of $p \in \{0.25, 0.50\}$, the density is plotted for $m = 20$ trials and several settings of ϕ. Figure 4.7 shows corresponding plots for the Mixture Link density letting $J = 2$ and $\kappa = 1$. Each shows the binomial density for reference. For beta-binomial, as the overdispersion parameter ϕ increases, the density moves from the standard binomial to one where most mass is at the extreme support values 0 and 20. Under RCB, increasing ϕ leads to the formation of a second mode. For the Mixture Link density, increasing π has the effect of fattening the tails compared to the standard binomial.

Figures 4.8, 4.9, and 4.10 show several more cases of the Mixture Link density, focusing only on the case $p = 0.5$ but varying $\kappa \in \{0.5, 1, 2\}$ and $J \in \{2, 3\}$. A variety of shapes can be seen for the limited settings of $\boldsymbol{\pi}$ that are shown. Expressing two modes is possible, as is inflating mass at the extreme support values 0 and 20.

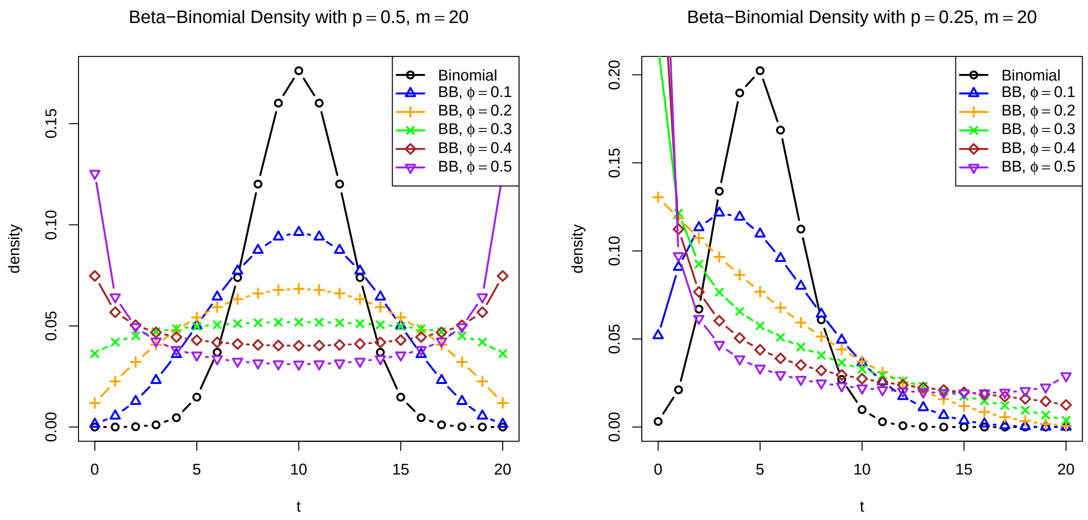


Figure 4.5: BB densities.

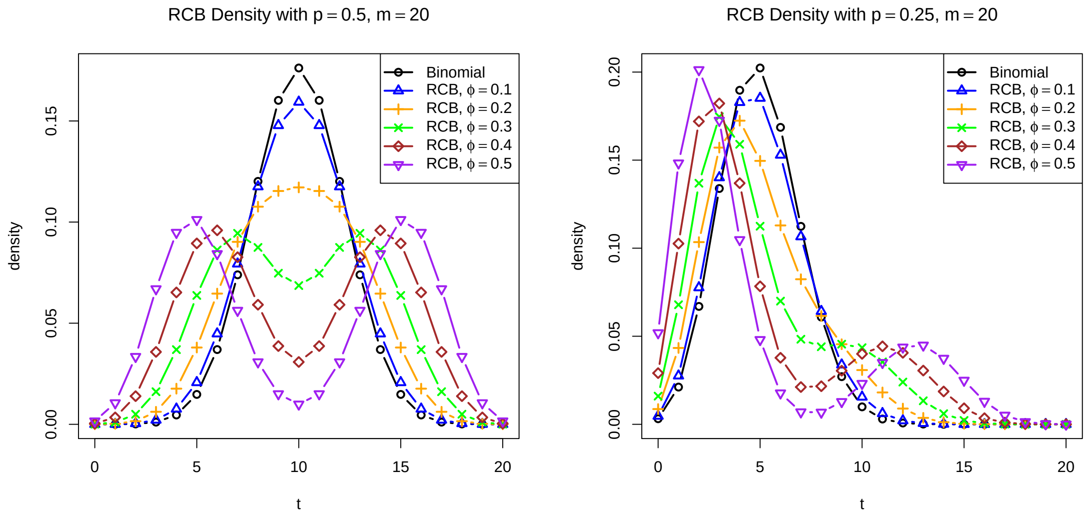


Figure 4.6: RCB densities.

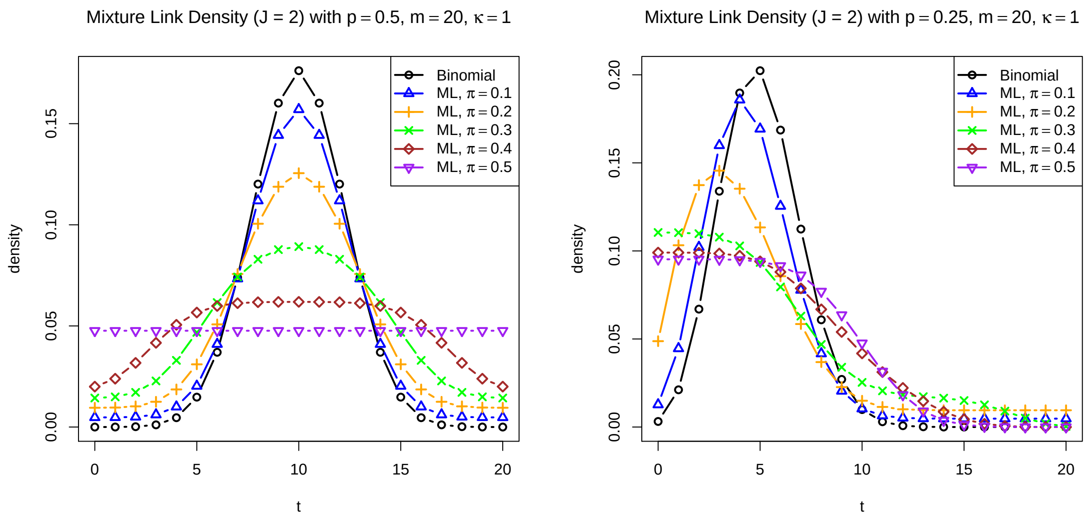


Figure 4.7: Mixture Link densities.

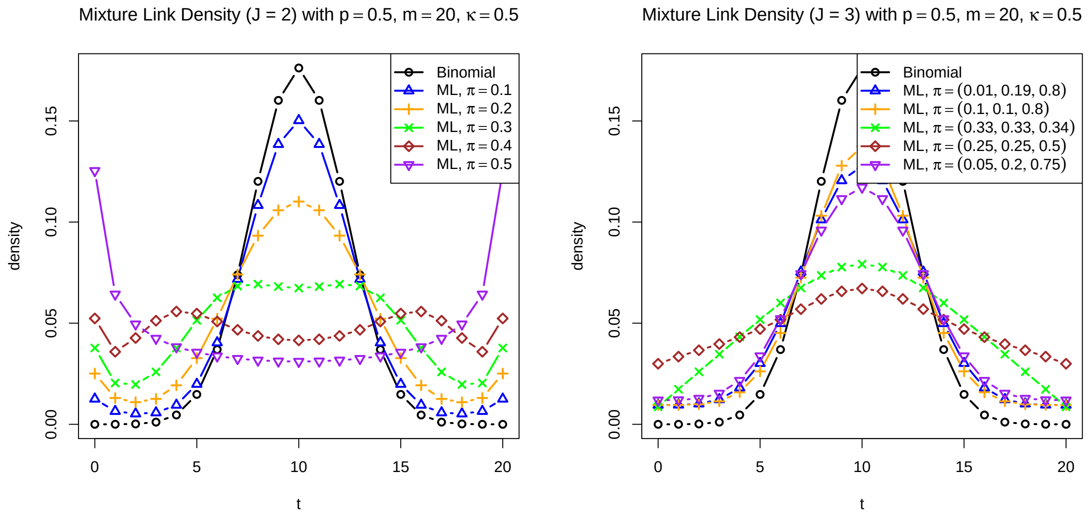


Figure 4.8: Compare Mixture Link densities for $J = 2$ and $J = 3$ when $\kappa = 0.5$.

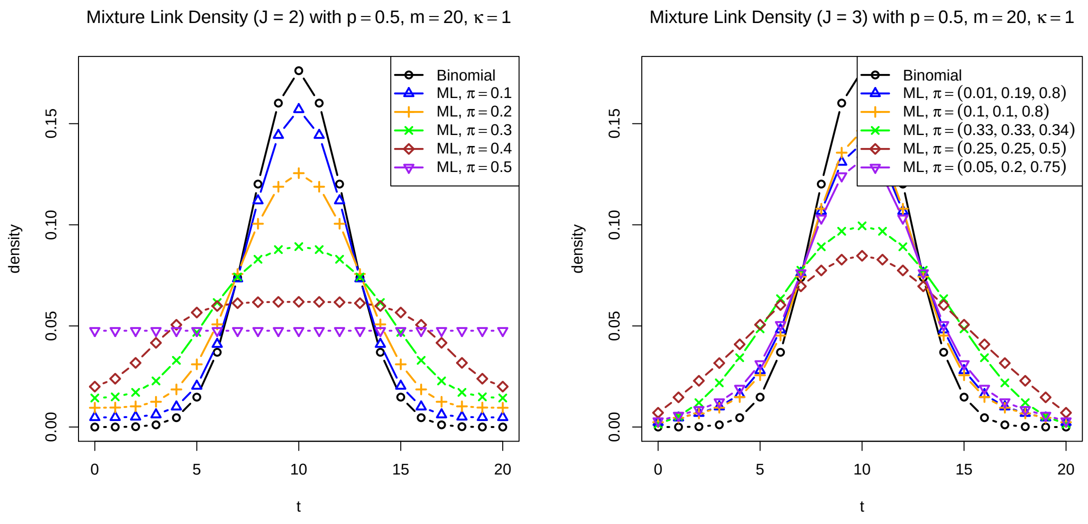


Figure 4.9: Compare Mixture Link densities for $J = 2$ and $J = 3$ when $\kappa = 1$.

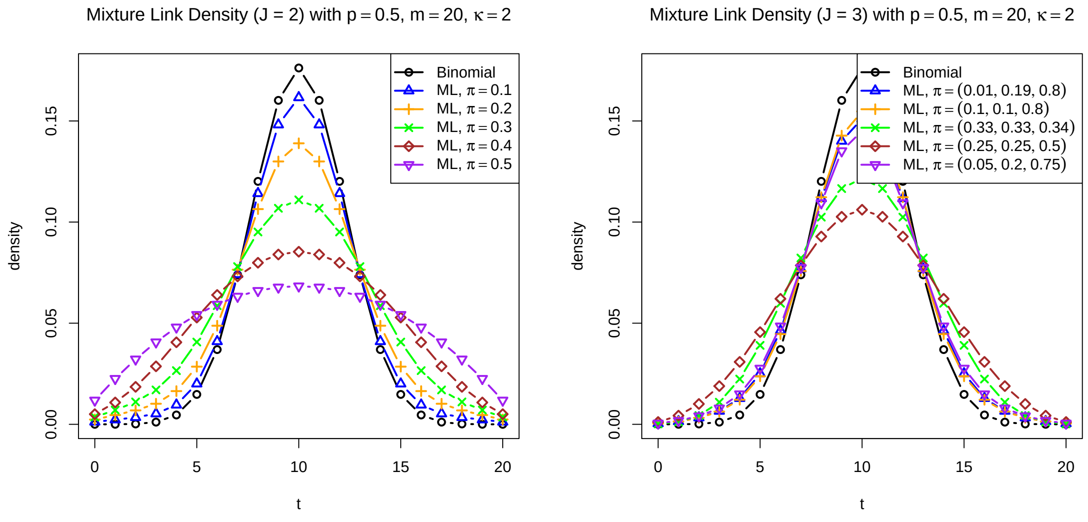


Figure 4.10: Compare Mixture Link densities for $J = 2$ and $J = 3$ when $\kappa = 2$.

4.8 Moment-Based Approximation to the Mixture Link Model

The density of a linear combination of Dirichlet random vector does not have a simple closed form in general, and numerical integration of the Imhof expression (4.16) is time consuming when repeated many times as needed when evaluating the density or likelihood for Mixture Link. The issue is more severe when the likelihood must be evaluated many times, as required in many MCMC sampling and numerical optimization approaches. In this section we consider approximating the linear combination of Dirichlet density by a simpler beta density, where the parameters are selected by moment-matching. The resulting approximation to the Mixture Link density is easily evaluated numerically, for example by quadrature. This moment matching evokes the classical approximation by Satterthwaite (1946), which is used in approximating the distribution of the two sample t-test when the population variances are unequal. Satterthwaite (1946) uses a single χ^2_v random variable to approximate a linear combination of chi-squares $\sum_{i=1}^{k} a_i X_i$, where the degrees of freedom v is selected by equating first and second moments.

Suppose $B \sim \text{Beta}(a, b)$ and $B^* = (u - \ell)B + \ell$ for given numbers $\ell < u$. Then B^* has a shifted/scaled beta distribution on the interval (ℓ, u). We have for B^* that

$$\begin{aligned} \mathrm{E}(B^*) &= (u-\ell)\frac{a}{a+b} + \ell, \quad \text{and} \\ \mathrm{Var}(B^*) &= (u-\ell)^2 \frac{ab}{(a+b)^2(a+b+1)}. \end{aligned}$$

Recall that for a given $\boldsymbol{c} \in \mathbb{R}^k$ and $\boldsymbol{\lambda} \sim \text{Dirichlet}_k(\kappa \mathbf{1})$, the expected value and variance of $\boldsymbol{c}^T\boldsymbol{\lambda}$ are given by

$$\begin{aligned} \xi &= \mathrm{E}(\boldsymbol{c}^T\boldsymbol{\lambda}) = \bar{c}, \quad \text{and} \\ \tau^2 &= \mathrm{Var}(\boldsymbol{c}^T\boldsymbol{\lambda}) = \frac{k\boldsymbol{c}^T\boldsymbol{c} - (k\bar{c})^2}{k^2(1+k\kappa)} \end{aligned}$$

respectively. Equating $\mathrm{E}(B^*) = \xi$ and $\mathrm{Var}(B^*) = \tau^2$ and solving for a and b, we obtain that

$$a = \left(\frac{\xi - \ell}{\tau}\right)^2 \frac{u - \xi}{u - \ell} - \frac{\xi - \ell}{u - \ell} \quad \text{and} \quad b = a\left(\frac{u - \xi}{\xi - \ell}\right)$$

For the Mixture Link model, let ℓ_j and u_j be the smallest and largest elements of $\boldsymbol{v}_{j\cdot}$, respectively, so that the interval (ℓ_j, u_j) represents the range of $\boldsymbol{v}_{j\cdot}^T\boldsymbol{\lambda}$ for $j = 1, \ldots, J$. To obtain the beta approximation to Mixture Link, let

$$\begin{aligned} \xi_j &= \mathrm{E}(\boldsymbol{v}_{j\cdot}^T\boldsymbol{\lambda}) = \bar{v}_{j\cdot} \\ \tau_j^2 &= \mathrm{Var}(\boldsymbol{v}_{j\cdot}^T\boldsymbol{\lambda}) = \frac{k\boldsymbol{v}_{j\cdot}^T\boldsymbol{v}_{j\cdot} - (k\bar{v}_{j\cdot})^2}{k^2(1+k\kappa)}, \end{aligned}$$

so that $B_j^* \sim (u_j - \ell_j)\text{Beta}(a_j, b_j) + \ell_j$ is a moment-matched shifted/scaled beta with

$$a_j = \left(\frac{\xi_j - \ell_j}{\tau_j}\right)^2 \frac{u_j - \xi_j}{u_j - \ell_j} - \frac{\xi_j - \ell_j}{u_j - \ell_j} \quad \text{and} \quad b_j = a_j\left(\frac{u_j - \xi_j}{\xi_j - \ell_j}\right).$$

Now an approximation to the Mixture Link density may be computed as

$$f(t \mid m, p, \boldsymbol{\pi}, \kappa) = \binom{m}{t} \sum_{j=1}^{J} \pi_j \int_{\ell_j}^{u_j} w^t (1-w)^{m-t} \cdot \frac{1}{u_j - \ell_j} h\left(\frac{w - \ell_j}{u_j - \ell_j} \,\middle|\, a_j, b_j\right) dw \tag{4.20}$$

where $h(\cdot \mid a, b)$ represents the standard $\text{Beta}(a, b)$ density. Letting $z = (w - \ell_j)/(u_j - \ell_j)$

we may transform to

$$
\begin{aligned}
&f(t \mid m, p, \boldsymbol{\pi}, \kappa) \\
&\quad = \binom{m}{t} \sum_{j=1}^{J} \pi_j \int_0^1 \Big((u_j - \ell_j) z + \ell_j \Big)^t \Big(1 - [(u_j - \ell_j) z + \ell_j] \Big)^{m-t} g(z \mid a_j, b_j) dz,
\end{aligned}
\tag{4.21}
$$

which emphasizes that the expression is not quite in a conjugate beta form, unless $\ell_j = 0$ and $u_j = 1$, in which case

$$
f(t \mid m, p, \boldsymbol{\pi}, \kappa) = \sum_{j=1}^{J} \pi_j \binom{m}{t} \frac{B(a_j + t, b_j + m - t)}{B(a_j, b_j)}
$$

can be recognized as a beta-binomial finite mixture (recall Example 1.2). In general, (4.21) may be evaluated numerically by any method which can evaluate one-dimensional integrals over a bounded range. For example, taking N quadrature points $0 < \tilde{z}_1 < \cdots < \tilde{z}_N < 1$ spread (say) uniformly over $(0, 1)$, and corresponding weights

$$
\tilde{w}_d^{(j)} = \frac{w_d^{(j)}}{\sum_{\ell=1}^{N} w_\ell^{(j)}}, \quad \text{where} \quad w_d^{(j)} = \frac{1}{u_j - \ell_j} h\left(\frac{\tilde{z}_d - \ell_j}{u_j - \ell_j} \,\middle|\, a_j, b_j \right),
$$

for $d = 1, \ldots, N$ and $j = 1, \ldots, J$. We may then compute the density as

$$
f(t \mid m, p, \boldsymbol{\pi}, \kappa) \approx \binom{m}{t} \sum_{j=1}^{J} \pi_j \sum_{d=1}^{N} \Big[(u_j - \ell_j) \tilde{z}_d + \ell_j \Big]^t \Big[1 - ((u_j - \ell_j) \tilde{z}_d + \ell_j) \Big]^{m-t} \tilde{w}_d^{(j)}.
$$

Note that it is possible for $u_j = \ell_j$; in this case, the approximation

$$
\int_{\ell_j}^{u_j} w^t (1 - w)^{m-t} \cdot \frac{1}{u_j - \ell_j} h\left(\frac{w - \ell_j}{u_j - \ell_j} \,\middle|\, a_j, b_j \right) dw
$$

cannot be computed, but the original integral simplifies to

$$\int_{\ell_j}^{u_j} w^t(1-w)^{m-t} \cdot f_{A^{(j)}}(w)dw = \ell_j^t(1-\ell_j)^{m-t},$$

regardless of the distribution assumed for $\boldsymbol{\mu}$. Notice that when $\kappa \to \infty$,

$$\frac{a_j}{a_j+b_j} = \frac{a_j}{a_j(\frac{u_j-\xi_j}{\xi_j-\ell_j}+1)} = \frac{\xi_j-\ell_j}{u_j-\ell_j}$$

is free of κ, and

$$(a_j+b_j+1) = 1 + a_j\left(\frac{u_j-\xi_j}{\xi_j-\ell_j}+1\right) \to \infty$$

since $a_j \to \infty$. Therefore

$$\begin{aligned} \mathrm{E}(B_j^*) &= (u_j-\ell_j)\frac{a_j}{a_j+b_j}+\ell \\ &\to (u_j-\ell_j)\frac{\xi_j-\ell_j}{u_j-\ell_j}+\ell_j = \xi_j \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}(B_j^*) &= (u_j-\ell_j)^2\frac{a_jb_j}{(a_j+b_j)^2(a_j+b_j+1)} \\ &= (u_j-\ell_j)^2\frac{a_j}{a_j+b_j}\left[1-\frac{a_j}{a_j+b_j}\right]\frac{1}{(a_j+b_j+1)} \to 0, \end{aligned}$$

and hence the distribution of B_j^* converges to a point mass at ξ_j. In this case, the approximate Mixture Link density becomes

$$f(t \mid m, p, \boldsymbol{\pi}) = \binom{m}{t}\sum_{j=1}^{J}\pi_j\xi_j^t(1-\xi_j)^{m-t} = \binom{m}{t}\sum_{j=1}^{J}\pi_j\bar{v}_{j.}^t(1-\bar{v}_{j.})^{m-t},$$

just as in (4.12) under the exact linear combination of Dirichlet distribution.

To evaluate the accuracy of the beta approximation, let us first compare the density of $\boldsymbol{c}^T \boldsymbol{\lambda}$ with that of $B^* \sim (u - \ell)\text{Beta}(a, b) + \ell$, where $\boldsymbol{\lambda} \sim \text{Dirichlet}_k(\kappa \mathbf{1})$, $\boldsymbol{c} = (c_1, \ldots, c_k)$, ℓ is the smallest element of $\boldsymbol{c}$, u is the largest element of $\boldsymbol{c}$, and a, b are given by the moment matching discussed earlier in this section. Consider the distance based on the supremum norm, defined as

$$D(f, g) = \sup_{x \in \Omega} |f(x) - g(x)| \tag{4.22}$$

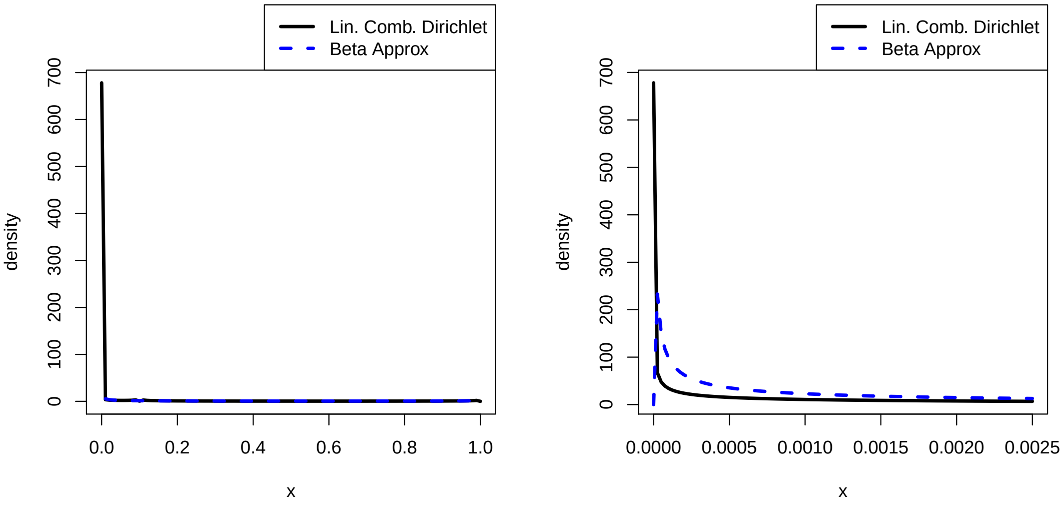
for densities f and g, where Ω is the sample space. We have that $D(f, g) \geq 0$, with $D(f, g) = 0$ attained only when $f(x) = g(x)$ for all $x \in \Omega$. Table 4.2 shows results of computing the sup norm distance numerically for several settings of $\boldsymbol{c}$ and κ. For all results, f is taken to be the density of $\boldsymbol{c}^T \boldsymbol{\lambda}$ computed by the Imhof procedure, and g is the moment-matched beta density with the R `integrate` function used to compute integrals. Figures 4.11 and 4.12 plot both densities with $\boldsymbol{c} = (0, 0.1, 1)$ using several of the κ settings from the table. Similarly, Figure 4.13 plots entries corresponding to $\boldsymbol{c} = (0, 0.05, 0.5, 0.95, 1)$. It can be seen that, for a given $\boldsymbol{c}$, the beta approximation may have little resemblance to f for small κ, but becomes more accurate as κ is increased from zero. The approximation also appears to work better when the entries of $\boldsymbol{c}$ are more uniformly spaced and more symmetric about the middle of the interval $[\ell, u]$. For $k = 2$, the approximation is very accurate for all κ (and matches exactly when $\kappa = 1$), but this is the least useful case because the density of $\boldsymbol{c}^T \boldsymbol{\lambda}$ has the convenient closed form (4.17).

Now let us compare the exact Mixture Link density f integrated by the Imhof method to the Mixture Link density g with random effects approximated by moment-matched beta random effects. Tables 4.3, 4.4, and 4.5 show the sup norm distance for $m = 5, 10, 20$ respectively. Figure 4.14 shows the two densities plotted for $\boldsymbol{\pi} = (\frac{1}{20}, \frac{2}{20}, \frac{3}{20}, \frac{4}{20}, \frac{5}{20})$, $p \in \{0.1, 0.5\}$, and $\kappa \in \{0.5, 1, 2\}$. As expected, when all other settings are kept fixed, the magnitude of the distance decreases as κ increases. It also appears to increase as m increases. It is not immediately clear how changing $\boldsymbol{\pi}$ and p affects the magnitude of the distance. Figure 4.14 shows that, even in a case from Table 4.5 with larger distances, the difference between the two distributions is small, indicating that the approximation is very good overall.

Empirical evidence given in this section suggests that the moment-matched beta distribution gives a very close result to the exact random effects distribution in computing integrals required to evaluate the Mixture Link density. It would be desirable to give a theoretical justification as well. The moment-matched beta density has great practical advantage over the linear combination of Dirichlet density, in that it is routine to evaluate using standard statistical software.

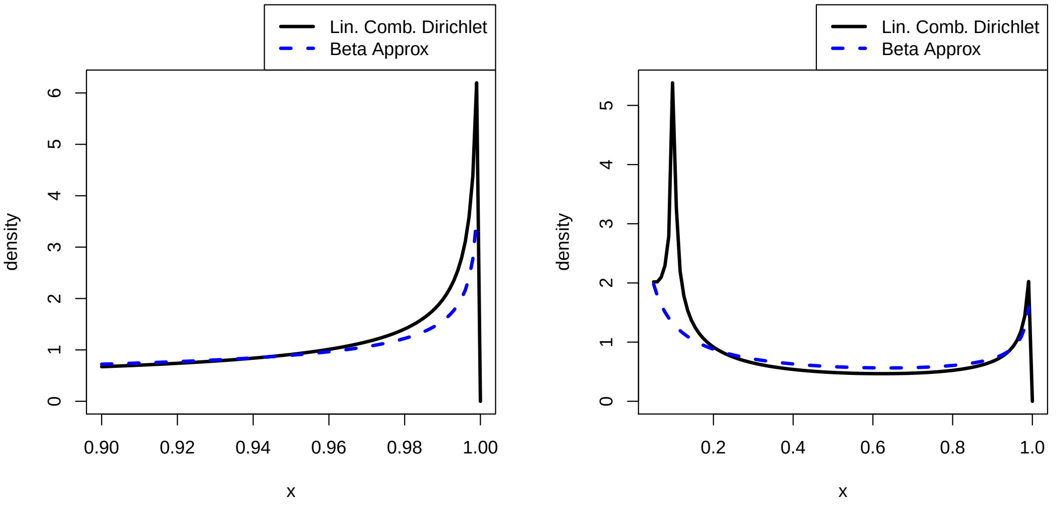
Table 4.2: Distance $D(f, g)$ between the density f of $\boldsymbol{c}^T\boldsymbol{\lambda}$ and density g of beta approximation.

	κ						
$\boldsymbol{c}$	0.25	0.5	0.75	1	2	3	4
$(0, 1)$	2.220E−16	3.331E−16	2.220E−16	0	2.220E−16	8.882E−16	1.776E−15
$(0.5, 1)$	4.441E−16	1.776E−15	2.220E−16	0	8.882E−16	1.554E−15	6.217E−15
$(0, 0.5, 1)$	2.883E−01	1.856E−01	1.377E−01	1.094E−01	1.045E−01	6.738E−02	5.171E−02
$(0, 0.2, 1)$	1.766E−01	1.501E−01	1.362E−01	1.277E−01	1.124E−01	3.085E−02	7.737E−02
$(0, 0.1, 1)$	9.891E−02	8.731E−02	8.104E−02	4.920E−01	1.546E−01	1.174E−01	1.032E−01
$(0, 0.3, 0.7, 1)$	2.778E−01	1.404E−01	8.464E−02	5.690E−02	7.806E−03	2.108E−02	1.930E−02
$(0, 0.1, 0.9, 1)$	1.008E−01	7.192E−02	8.948E−02	5.712E−02	2.725E−02	2.657E−02	2.274E−02
$(0, 0.1, 0.2, 1)$	2.308E−01	1.054E+00	8.257E−01	6.862E−01	4.016E−01	3.222E−01	2.888E−01
$(0, 0.05, 0.5, 0.95, 1)$	1.056E−01	4.710E−02	4.414E−02	2.379E−02	5.369E−03	1.915E−03	6.801E−04
$(0, 0.1, 0.2, 0.3, 1)$	4.228E+00	1.090E+00	8.757E−01	7.756E−01	5.776E−01	4.980E−01	4.584E−01



(a) $\kappa = 0.25$

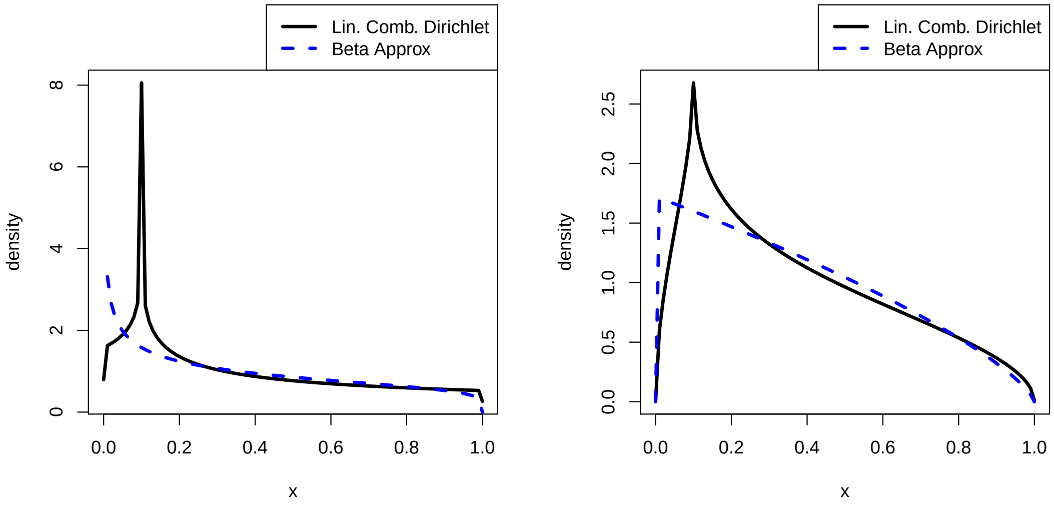
(b) $\kappa = 0.25$, focusing on $x \in (0, 0.1)$.



(c) $\kappa = 0.25$, focusing on $x \in (0.9, 1)$.

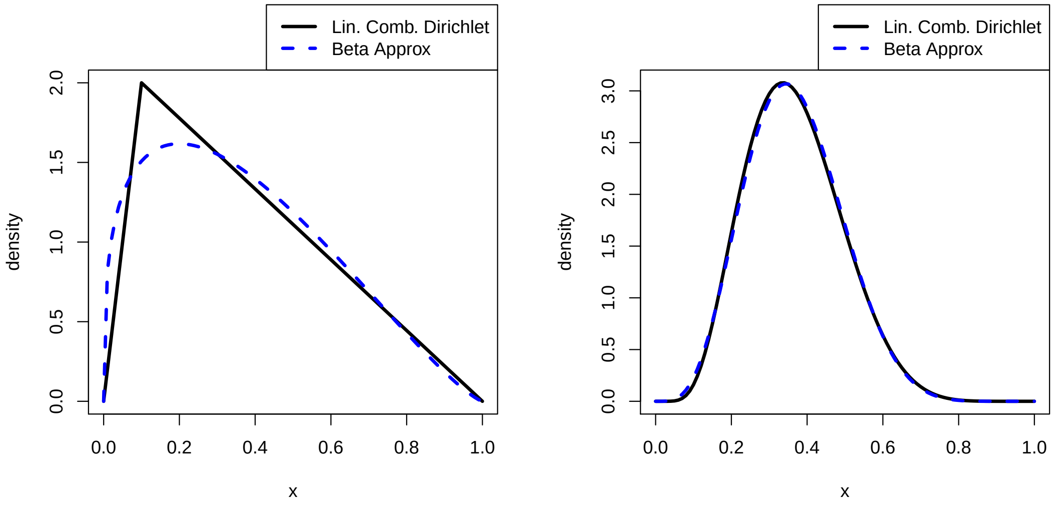
(d) $\kappa = 0.25$, focusing on $x \in (0.05, 1)$

Figure 4.11: Comparison between density of $\boldsymbol{c}^T\boldsymbol{\lambda}$ and moment-matched beta taking $\boldsymbol{c} = (0, 0.1, 1)$.



(a) $\kappa = 0.5$

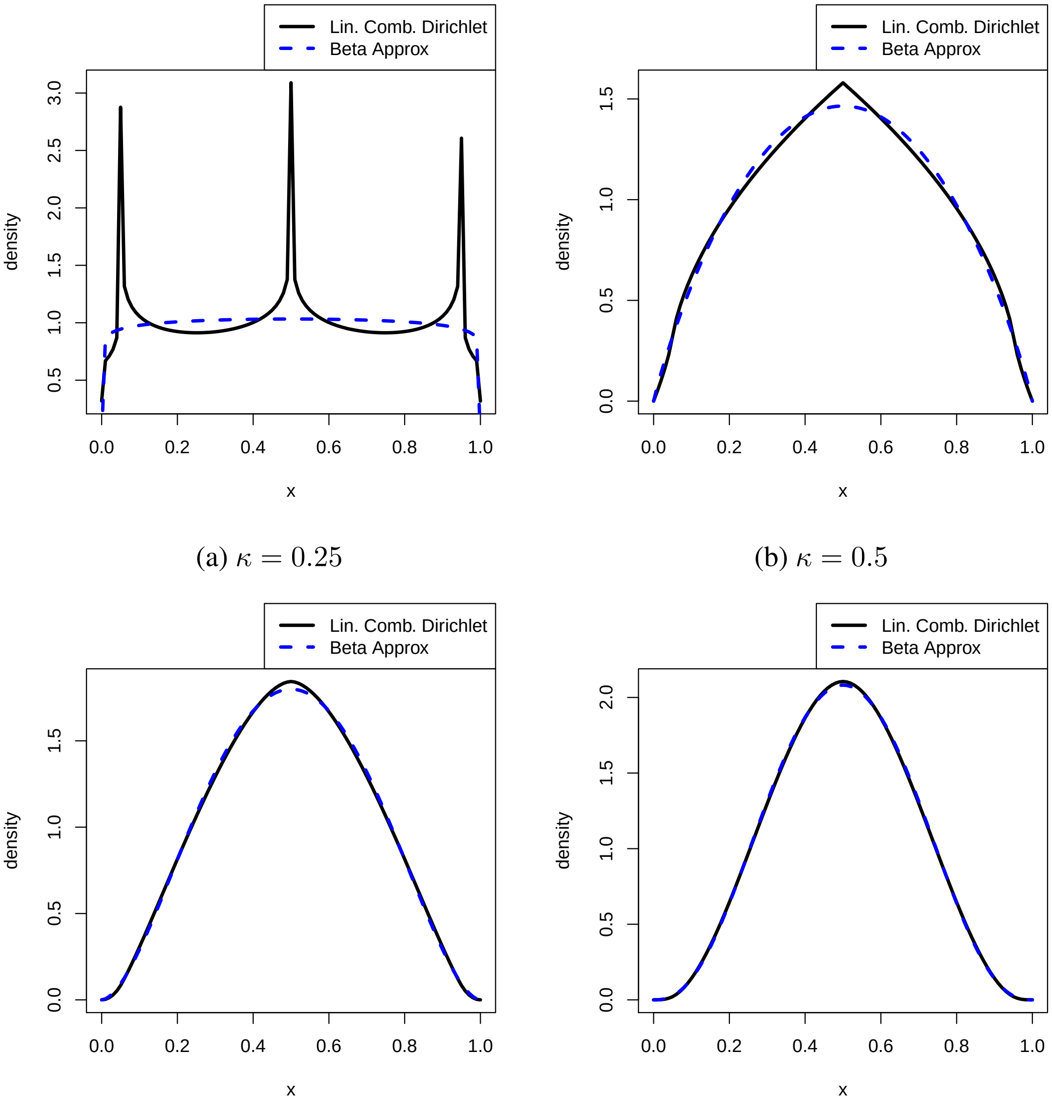
(b) $\kappa = 0.75$



(c) $\kappa = 1$

(d) $\kappa = 4$

Figure 4.12: Comparison between density of $\boldsymbol{c}^T\boldsymbol{\lambda}$ and moment-matched beta taking $\boldsymbol{c} = (0, 0.1, 1)$.



(a) $\kappa = 0.25$

(b) $\kappa = 0.5$

(c) $\kappa = 0.75$

(d) $\kappa = 1$

Figure 4.13: Comparison between density of $\boldsymbol{c}^T\boldsymbol{\lambda}$ and moment-matched beta taking $\boldsymbol{c} = (0, 0.05, 0.5, 0.95, 1)$.

Table 4.3: Distance $D(f,g)$ between exact Mixture Link density f and density g using beta approximation with $m = 5$ trials.

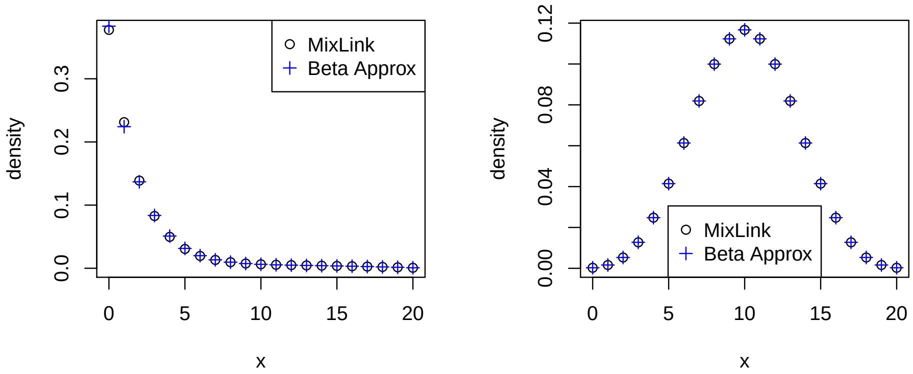
$\boldsymbol{\pi}$	p	$\kappa = 0.5$	$\kappa = 1$	$\kappa = 2$
$(\frac{1}{2}, \frac{1}{2})$	0.05	2.960E−05	1.110E−16	8.327E−17
	0.1	1.654E−05	1.665E−16	1.110E−16
	0.5	2.948E−07	5.551E−17	5.551E−17
$(\frac{1}{4}, \frac{3}{4})$	0.05	1.415E−05	5.551E−17	5.551E−16
	0.1	2.440E−05	1.110E−16	3.331E−16
	0.5	7.371E−08	1.665E−16	2.776E−17
$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	0.05	1.719E−03	4.614E−06	2.067E−06
	0.1	1.215E−03	2.167E−06	1.697E−06
	0.5	1.953E−03	4.468E−04	8.028E−05
$(\frac{1}{6}, \frac{2}{6}, \frac{3}{6})$	0.05	1.828E−03	5.872E−06	2.039E−06
	0.1	1.126E−03	2.373E−06	1.688E−06
	0.5	1.502E−03	4.218E−04	8.825E−05
$(\frac{1}{10}, \frac{2}{10}, \frac{7}{10})$	0.05	2.094E−03	7.767E−06	1.726E−06
	0.1	1.390E−03	2.329E−06	1.256E−06
	0.5	7.094E−05	1.949E−05	4.573E−06
$(0.05, 0.1, 0.85)$	0.05	2.407E−03	1.256E−05	2.125E−06
	0.1	3.086E−04	9.323E−05	2.609E−05
	0.5	3.385E−05	2.500E−06	5.931E−07
$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	0.05	1.639E−03	5.445E−06	2.047E−06
	0.1	1.159E−03	2.040E−06	1.747E−06
	0.5	5.389E−06	3.134E−07	3.846E−07
$(\frac{1}{10}, \frac{2}{10}, \frac{3}{10}, \frac{4}{10})$	0.05	1.843E−03	6.928E−06	2.102E−06
	0.1	6.676E−04	2.749E−06	2.033E−06
	0.5	2.732E−04	4.892E−05	7.742E−06
$(0.05, 0.1, 0.15, 0.7)$	0.05	2.255E−03	1.363E−05	2.238E−06
	0.1	2.949E−04	2.085E−04	9.095E−05
	0.5	1.804E−05	4.227E−06	1.112E−06
$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$	0.05	1.998E−03	5.942E−06	1.726E−06
	0.1	1.154E−03	2.283E−06	1.358E−06
	0.5	2.131E−05	3.395E−06	8.483E−07
$(\frac{1}{15}, \frac{2}{15}, \frac{3}{15}, \frac{4}{15}, \frac{5}{15})$	0.05	1.959E−03	7.703E−06	1.999E−06
	0.1	9.682E−04	3.919E−04	1.310E−04
	0.5	2.353E−06	9.916E−07	5.511E−07
$(\frac{1}{20}, \frac{2}{20}, \frac{3}{20}, \frac{4}{20}, \frac{10}{20})$	0.05	2.267E−03	1.091E−05	1.974E−06
	0.1	4.655E−04	2.410E−04	9.071E−05
	0.5	6.229E−05	1.127E−05	2.077E−06

Table 4.4: Distance $D(f, g)$ between exact Mixture Link density f and density g using beta approximation with $m = 10$ trials.

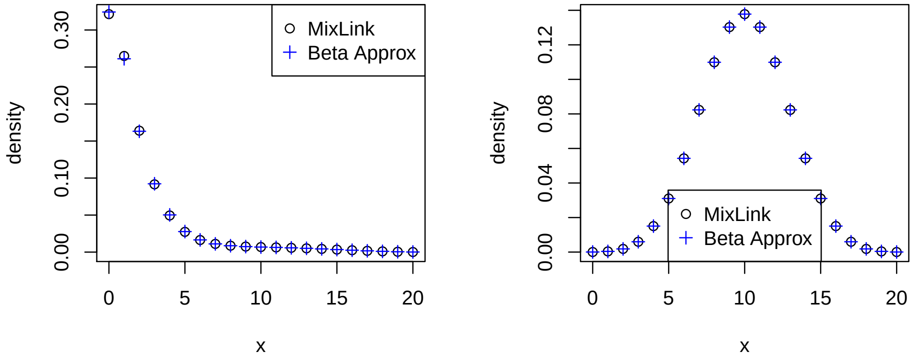
$\boldsymbol{\pi}$	p	$\kappa = 0.5$	$\kappa = 1$	$\kappa = 2$
$(\frac{1}{2}, \frac{1}{2})$	0.05	4.196E−04	1.221E−15	1.388E−15
	0.1	6.439E−04	1.499E−15	1.665E−15
	0.5	1.540E−07	4.163E−16	5.412E−16
$(\frac{1}{4}, \frac{3}{4})$	0.05	2.122E−04	1.110E−15	1.332E−15
	0.1	3.846E−04	1.277E−15	1.665E−15
	0.5	7.486E−04	6.384E−16	8.604E−16
$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	0.05	1.718E−03	3.364E−06	3.261E−06
	0.1	1.215E−03	1.404E−06	2.197E−06
	0.5	3.336E−03	1.008E−03	2.392E−04
$(\frac{1}{6}, \frac{2}{6}, \frac{3}{6})$	0.05	1.827E−03	4.462E−06	3.292E−06
	0.1	1.126E−03	1.454E−06	2.305E−06
	0.5	2.023E−03	8.228E−04	2.333E−04
$(\frac{1}{10}, \frac{2}{10}, \frac{7}{10})$	0.05	2.093E−03	6.615E−06	2.776E−06
	0.1	1.393E−03	2.040E−06	1.934E−06
	0.5	2.969E−04	7.880E−05	1.920E−05
$(0.05, 0.1, 0.85)$	0.05	2.407E−03	1.091E−05	3.732E−06
	0.1	4.386E−04	1.536E−04	4.998E−05
	0.5	9.326E−05	8.998E−06	9.050E−07
$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	0.05	1.638E−03	4.173E−06	3.252E−06
	0.1	1.158E−03	1.350E−06	2.273E−06
	0.5	5.296E−06	2.024E−07	2.751E−07
$(\frac{1}{10}, \frac{2}{10}, \frac{3}{10}, \frac{4}{10})$	0.05	1.841E−03	5.584E−06	3.380E−06
	0.1	1.024E−03	1.942E−06	2.696E−06
	0.5	5.283E−04	1.292E−04	2.540E−05
$(0.05, 0.1, 0.15, 0.7)$	0.05	2.432E−03	1.208E−05	3.730E−06
	0.1	3.312E−04	2.801E−04	1.483E−04
	0.5	7.777E−05	1.818E−05	3.262E−06
$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$	0.05	1.997E−03	4.657E−06	2.938E−06
	0.1	1.153E−03	1.566E−06	1.901E−06
	0.5	6.269E−05	1.115E−05	1.719E−06
$(\frac{1}{15}, \frac{2}{15}, \frac{3}{15}, \frac{4}{15}, \frac{5}{15})$	0.05	1.957E−03	6.334E−06	3.296E−06
	0.1	3.754E−03	1.785E−03	6.688E−04
	0.5	5.044E−06	1.951E−06	6.966E−07
$(\frac{1}{20}, \frac{2}{20}, \frac{3}{20}, \frac{4}{20}, \frac{10}{20})$	0.05	2.265E−03	9.459E−06	3.355E−06
	0.1	9.932E−04	6.646E−04	2.905E−04
	0.5	2.091E−04	4.430E−05	7.484E−06

Table 4.5: Distance $D(f, g)$ between exact Mixture Link density f and density g using beta approximation with $m = 20$ trials.

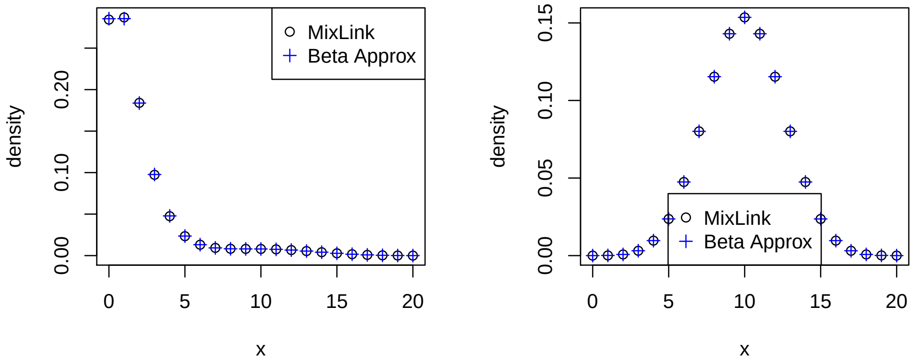
$\boldsymbol{\pi}$	p	$\kappa = 0.5$	$\kappa = 1$	$\kappa = 2$
$(\frac{1}{2}, \frac{1}{2})$	0.05	1.390E−03	2.220E−16	1.665E−16
	0.1	1.595E−03	5.135E−16	5.829E−16
	0.5	4.767E−07	4.718E−16	6.800E−16
$(\frac{1}{4}, \frac{3}{4})$	0.05	9.483E−04	1.665E−16	3.331E−16
	0.1	1.332E−03	4.163E−16	3.608E−16
	0.5	1.014E−03	9.576E−16	1.193E−15
$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	0.05	1.716E−03	2.047E−06	4.290E−06
	0.1	1.214E−03	8.808E−07	2.207E−06
	0.5	3.488E−03	1.268E−03	3.766E−04
$(\frac{1}{6}, \frac{2}{6}, \frac{3}{6})$	0.05	1.825E−03	2.906E−06	4.546E−06
	0.1	1.125E−03	7.904E−07	2.529E−06
	0.5	1.935E−03	8.333E−04	3.257E−04
$(\frac{1}{10}, \frac{2}{10}, \frac{7}{10})$	0.05	2.092E−03	5.015E−06	4.519E−06
	0.1	1.396E−03	1.486E−06	2.663E−06
	0.5	5.658E−04	2.556E−04	5.106E−05
$(0.05, 0.1, 0.85)$	0.05	2.408E−03	8.446E−06	6.074E−06
	0.1	4.902E−04	1.828E−04	6.382E−05
	0.5	3.740E−04	7.331E−05	2.303E−06
$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	0.05	1.637E−03	2.766E−06	4.332E−06
	0.1	1.157E−03	7.914E−07	2.343E−06
	0.5	2.531E−07	1.200E−07	1.882E−07
$(\frac{1}{10}, \frac{2}{10}, \frac{3}{10}, \frac{4}{10})$	0.05	1.925E−03	3.953E−06	4.724E−06
	0.1	1.142E−03	1.199E−06	3.008E−06
	0.5	5.745E−04	1.818E−04	4.667E−05
$(0.05, 0.1, 0.15, 0.7)$	0.05	2.490E−03	9.792E−06	5.821E−06
	0.1	3.946E−04	4.358E−04	2.040E−04
	0.5	2.026E−04	5.511E−05	9.627E−06
$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$	0.05	1.995E−03	3.198E−06	4.050E−06
	0.1	1.152E−03	9.531E−07	2.010E−06
	0.5	9.851E−05	2.251E−05	4.207E−06
$(\frac{1}{15}, \frac{2}{15}, \frac{3}{15}, \frac{4}{15}, \frac{5}{15})$	0.05	1.902E−03	4.634E−06	4.698E−06
	0.1	7.119E−03	3.804E−03	1.515E−03
	0.5	6.814E−06	3.633E−06	9.605E−07
$(\frac{1}{20}, \frac{2}{20}, \frac{3}{20}, \frac{4}{20}, \frac{10}{20})$	0.05	2.230E−03	7.487E−06	5.065E−06
	0.1	1.473E−03	9.554E−04	4.061E−04
	0.5	3.739E−04	9.918E−05	2.944E−05



(a) $p = 0.1, \kappa = 0.5$ (b) $p = 0.5, \kappa = 0.5$



(c) $p = 0.1, \kappa = 1$ (d) $p = 0.5, \kappa = 1$



(e) $p = 0.1, \kappa = 2$ (f) $p = 0.5, \kappa = 2$

Figure 4.14: Comparison of exact Mixture Link density f and density g using beta approximation with $m = 20$ trials and $\boldsymbol{\pi} = (\frac{1}{20}, \frac{2}{20}, \frac{3}{20}, \frac{4}{20}, \frac{5}{20})$.

4.9 Computational Studies

We now present an application of the Mixture Link model to an example dataset studying the effect of radiation dose on probability of chromosome aberration. We first describe how numerical optimization is used to obtain the maximum likelihood estimator and standard errors. A generalized goodness-of-fit (GOF) test is then recalled from the literature, which can be carried out in the "independent but not identically distributed binomial" case using the estimates obtained by numerical MLE. Finally, equipped with the GOF test, a study is carried out on the chromosome aberration data to compare Mixture Link to several binomial models for overdispersion.

4.9.1 Numerical Maximum Likelihood

In need of a practical way to carry out inference under Mixture Link, we make use of general numerical optimization via the `optim` function in R. As we have noted in Remark 4.6, the likelihood is not differentiable at all points in Θ, but as we will see later, pressing ahead with the numerical MLE gives reasonable results.

Consider the regression case $T_i \overset{\text{ind}}{\sim} \text{MixLink}(m_i, p_i, \boldsymbol{\pi}, \kappa)$, $p_i = G(\boldsymbol{x}_i^T \boldsymbol{\beta})$, where the objective for inference is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\pi}, \kappa)$. Denote $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \phi_3)$ as a transformed version of $\boldsymbol{\theta}$ to $\mathbb{R}^q$, where $\boldsymbol{\phi}_1 \in \mathbb{R}^d$, $\boldsymbol{\phi}_2 \in \mathbb{R}^J$, and $\phi_3 \in \mathbb{R}$. As the first step in computing the likelihood, the point $\boldsymbol{\phi}$ proposed by `optim` is transformed as $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\phi})$ where

$$\boldsymbol{\theta} = \begin{pmatrix} \beta \\ \boldsymbol{\pi} \\ \kappa \end{pmatrix} = \begin{pmatrix} \boldsymbol{\theta}_1(\boldsymbol{\phi}_1) \\ \boldsymbol{\theta}_2(\boldsymbol{\phi}_2) \\ \theta_3(\phi_3) \end{pmatrix}.$$

Recall that G is taken to be the logistic CDF, so that its derivative G' represents the logistic density. Furthermore, let us denote $\boldsymbol{G}(\boldsymbol{x}) = (G(x_1), \ldots, G(x_J))$, $\boldsymbol{G}'(\boldsymbol{x}) =$

$(G'(x_1), \ldots, G'(x_J))$. We take the transformation $\boldsymbol{\theta}$ to be

$$\boldsymbol{\theta}_1(\boldsymbol{x}) = \boldsymbol{x}, \quad \boldsymbol{\theta}_2(\boldsymbol{x}) = \frac{\boldsymbol{G}(\boldsymbol{x})}{\mathbf{1}^T \boldsymbol{G}(\boldsymbol{x})}, \quad \text{and} \quad \theta_3(x) = \exp(x).$$

The Jacobian of this transformation is

$$\frac{\partial \boldsymbol{\theta}(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \begin{pmatrix} \boldsymbol{I}_d & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\partial \boldsymbol{\theta}_2(\boldsymbol{\phi}_2)}{\partial \boldsymbol{\phi}_2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \exp(\phi_3) \end{pmatrix}$$

with

$$\frac{\partial \boldsymbol{\theta}_2(\boldsymbol{x})}{\partial \boldsymbol{x}} = \left[\mathbf{1}^T \{\boldsymbol{G}(\boldsymbol{x})\}\right]^{-2} \left\{ \{\mathbf{1}^T \boldsymbol{G}(\boldsymbol{x})\} \cdot \text{Diag}\{\boldsymbol{G}'(\boldsymbol{x})\} - \{\boldsymbol{G}(\boldsymbol{x})\}\{\boldsymbol{G}'(\boldsymbol{x})\}^T \right\}.$$

Then, given an estimate $\hat{\boldsymbol{\phi}}$ and the corresponding Hessian $\boldsymbol{H}(\hat{\boldsymbol{\phi}})$ from `optim`, an estimate for the covariance of $\hat{\boldsymbol{\theta}}$ is

$$\widehat{\text{Var}}(\hat{\boldsymbol{\theta}}) = \left[\frac{\partial \boldsymbol{\theta}(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right] [-\boldsymbol{H}(\boldsymbol{\phi})]^{-1} \left[\frac{\partial \boldsymbol{\theta}(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right]^T \Bigg|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}}, \tag{4.23}$$

and standard errors for $\hat{\boldsymbol{\theta}}$ can be computed by taking the square roots of the diagonal entries. In computing (4.23), entries in $\boldsymbol{H}$ and the Jacobian corresponding to the redundant Jth element of $\boldsymbol{\pi}$ are dropped from the calculation to avoid singularity. We specifically make use of the quasi-Newton `L-BFGS-B` method in `optim`, whereby the gradient and Hessian are computed by numerical differentiation.

Whether (4.23) is a valid estimator of the covariance of the MLE has not been established. An alternate way of computing standard errors is by bootstrapping; in this work we will consider the parametric bootstrap (as opposed to the more common nonparametric variant) to emphasize that we are interested in properties of Mixture Link but not necessarily the "true" distribution of the data. In the parametric bootstrap approach, we consider B samples

$$T_i^{(b)} \overset{\text{ind}}{\sim} \text{MixLink}_J(m_i, \hat{p}_i, \hat{\boldsymbol{\pi}}, \hat{\kappa}), \quad \hat{p}_i = G(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}),$$

for $i = 1, \ldots, n$ and $b = 1, \ldots, B$, where $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}, \hat{\kappa})$ is the MLE fitted to the original data. We then obtain $\hat{\boldsymbol{\theta}}^{(b)}$ by computing the MLE for the bth sample for $b = 1, \ldots, B$. Taking $\bar{\boldsymbol{\theta}}_{\text{Boot}} = \frac{1}{B} \sum_{b=1}^{B} \hat{\boldsymbol{\theta}}^{(b)}$, the covariance of $\hat{\boldsymbol{\theta}}$ may then be estimated by

$$\widehat{\text{Var}}_{\text{Boot}}(\hat{\boldsymbol{\theta}}) = \frac{1}{B-1} \sum_{b=1}^{B} \left[\hat{\boldsymbol{\theta}}^{(b)} - \bar{\boldsymbol{\theta}}_{\text{Boot}}\right] \left[\hat{\boldsymbol{\theta}}^{(b)} - \bar{\boldsymbol{\theta}}_{\text{Boot}}\right]^T, \tag{4.24}$$

and standard errors for each of the estimated quantities may be computed by the square roots of the diagonal elements.

4.9.2 Goodness-of-Fit Test

To compare several binomial models with extra variation on the same dataset, we consider the goodness-of-fit (GOF) test

$$H_0 : T_i \overset{\text{ind}}{\sim} f(t_i \mid m_i, \boldsymbol{\theta}, \boldsymbol{x}_i) \text{ for some } \boldsymbol{\theta} \in \Theta \quad \text{vs.} \quad H_1 : \text{Not},$$

where f is fully specified up to a possibly unknown parameter $\boldsymbol{\theta}$ in the space $\Theta \subseteq \mathbb{R}^q$. For binomial data with m_i varying with observations, Neerchal and Morel (1998) proposed the following variation to the usual Pearson chi-square test statistic. Suppose $\mathcal{A}_1, \ldots, \mathcal{A}_r$ are disjoint intervals that cover $[0, 1]$, and define the GOF test statistic

$$X(\boldsymbol{\theta}) = \sum_{\ell=1}^{r} \frac{[O_\ell - E_\ell(\boldsymbol{\theta})]^2}{E_\ell(\boldsymbol{\theta})}, \tag{4.25}$$

where

$$E_\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{t=0}^{m_i} \mathrm{P}(t \mid m_i, \boldsymbol{\theta}) I\left(\frac{t}{m_i} \in \mathcal{A}_\ell\right), \quad \text{and}$$

$$O_\ell = \sum_{i=1}^{n} I\left(\frac{t_i}{m_i} \in \mathcal{A}_\ell\right).$$

Sutradhar et al. (2008) shows that, when the null distribution f is RCB, $X(\boldsymbol{\theta}) \sim \chi^2_{r-1}$ when all parameters are known and $X(\hat{\boldsymbol{\theta}}) \sim \chi^2_{r-1-q}$ when $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^q$ is estimated by maximizing the *grouped* likelihood

$$L_g(\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{\ell=1}^{r} \left[\mathrm{P}\left(\frac{t_i}{m_i} \in \mathcal{A}_\ell \,\middle|\, m_i, \boldsymbol{\theta}\right)^{I\left(\frac{t_i}{m_i} \in \mathcal{A}_\ell\right)} \right],$$

In practice, it is more natural to work with the *ungrouped* likelihood

$$L_u(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i \mid m_i, \boldsymbol{\theta})$$

of the observed T_i. There is a noted “recovery” of degrees of freedom in the GOF statistic when the ungrouped MLE is used, so that $X(\hat{\boldsymbol{\theta}})$ follows a χ^2_ν distribution with ν between $r-1-q$ and $r-1$. Although the theory in (Sutradhar et al., 2008) is stated specifically for the RCB distribution, proofs are given for general binomial models with varying m_i. A number of regularity conditions are assumed; for example, to ensure first-order efficiency of the MLE.

Our GOF studies use the ungrouped MLE and consider p-values based on $\nu = r - 1 - q$. Recall that a smaller degrees of freedom ν will result in a more right-skewed χ^2_ν distribution. Consequently, when ν is reduced and $X(\hat{\boldsymbol{\theta}})$ is held fixed, there will appear to be stronger evidence against the hypothesis of adequate fit H_0. Therefore, taking ν to be the smallest value in the range $[r-1-q, r-1]$ is conservative to test a model for adequate fit. The selection of intervals $\mathcal{A}_\ell$ is left up to the analyst, but it is suggested to follow the rule of thumb that all $E_\ell(\boldsymbol{\theta}) \geq 5$ to ensure that the distribution theory holds. Some discussion on interval selection is given in (Kendall and Stuart, 1979, Section 30.2); common choices include equal width intervals or intervals having equal probability.

The parametric bootstrap is useful in applying the GOF test under Mixture Link. It is not immediately clear that the required regularity conditions hold for Mixture Link. Also, it may be desired to obtain a single p-value for the GOF test, especially when the range of p-values computed by χ^2_{r-1-q} and χ^2_{r-1} is large, or the range contains values which indicate both an acceptable and unacceptable fit. The parametric bootstrap may be used to verify the distribution of the GOF test statistic and to compute a more accurate p-value. Similarly to Section 4.9.1, the bootstrap requires B samples

$$T_i^{(b)} \overset{\text{ind}}{\sim} \text{MixLink}_J(m_i, \hat{p}_i, \hat{\boldsymbol{\pi}}, \hat{\kappa}), \quad \hat{p}_i = G(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}),$$

for $i = 1, \ldots, n$ and $b = 1, \ldots, B$, drawn using the (ungrouped) MLE $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}, \hat{\kappa})$ from the observed data . Using the bootstrap sample $T_1^{(b)}, \ldots, T_n^{(b)}$ in place of the observed data, the (ungrouped) bootstrap MLE $\hat{\boldsymbol{\theta}}^{(b)}$ is computed. The bootstrap GOF test statistic $X^{(b)}$ is then computed using the bootstrap sample and bootstrap MLE. The theoretical distribution of $X(\hat{\boldsymbol{\theta}})$ can then be studied through the empirical distribution of $X^{(1)}, \ldots, X^{(B)}$. A bootstrapped p-value may be computed as

$$\text{p-value}_{\text{Boot}} = \frac{1}{B} \sum_{b=1}^{B} I(X^{(b)} \geq X(\hat{\boldsymbol{\theta}})).$$

Note that the above bootstrap procedure reflects the use of the ungrouped likelihood and the necessity to estimate the parameters. If instead the MLE $\hat{\boldsymbol{\theta}}$ of the observed data were used to compute each $X^{(b)}$, the empirical distribution of $X^{(1)}, \ldots, X^{(B)}$ would be comparable to the distribution of $X(\boldsymbol{\theta})$ when all parameters are known.

4.9.3 Chromosome Aberration Data

Awa et al. (1971) and Sofuni et al. (1978) study the effects of radiation exposure on chromosome aberrations in survivors of the atomic bombs that were used in Hiroshima and Nagasaki. Subjects in the study consist of 649 residents in Hiroshima and 403 residents in Nagasaki for whom radiation dose estimates were available. Subjects were placed into exposed and control groups. Individuals in the control group were either not present in the cities at the time of the bombings, or received an estimated dose of less than one rad. A chromosome analysis is carried out on m_i circulating lymphocytes for the ith subject, and of those, the number of chromosome aberrations t_i is recorded[1]. Two types of radiation exposure are measured, neutron and gamma, where higher doses of neutron exposure in Hiroshima are suspected of leading to increased incidence of aberration. Otake and Prentice (1984) analyze this data using beta-binomial, acknowledging the need for overdispersion modeling.

A subset of this data is featured in Morel and Neerchal (2012) as an illustrative example for goodness-of-fit in binomial models for extra variation. It is natural to suspect that overdispersion will be an issue in this data under standard logistic regression, as the presence or absence of aberrations within the m_i circulating lymphocytes of a particular subject may not be independent. Here, $n = 648$ observations from the Hiroshima portion of the original data are considered, and the covariate d_i represents the sum of neutron and gamma exposure for the ith subject. The total exposure is then normalized to

$$z_i = \frac{d_i - \bar{d}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - \bar{d})^2}}, \quad i = 1, \ldots n.$$

Using the methodology described in Sections 4.9.1 and 4.9.2, the following models will

---

[1] A lymphocyte is a type of white blood cell that plays a fundamental role in the immune system. An aberration is an abnormality involving the structure or number of chromosomes. See `www.britannica.com/EBchecked/topic/352799/lymphocyte` and `www.britannica.com/EBchecked/topic/116040/chromosomal-mutation`.

now be compared for goodness-of-fit for the chromosome aberration dataset:

- Logistic: $T_i \overset{\text{ind}}{\sim} \text{Bin}(m_i, p_i)$,
- RCB: $T_i \overset{\text{ind}}{\sim} \text{RCB}(m_i, p_i, \phi)$,
- BB: $T_i \overset{\text{ind}}{\sim} \text{BB}(m_i, p_i, \phi)$,
- RCB-Reg: $T_i \overset{\text{ind}}{\sim} \text{RCB}(m_i, p_i, \phi_i)$,
- BB-Reg: $T_i \overset{\text{ind}}{\sim} \text{BB}(m_i, p_i, \phi_i)$,
- MixLinkJ2: $T_i \overset{\text{ind}}{\sim} \text{MixLink}_2(m_i, p_i, \boldsymbol{\pi}, \kappa)$.

Taking $g = G^{-1}$ as the logistic link function, the regression $g(p_i) = \beta_0 + \beta_1 z_i + \beta_2 z_i^2$ is used for all models and $g(\phi_i) = \gamma_0 + \gamma_1 z_i + \gamma_2 z_i^2$ for the two "-Reg" models. The models RCB-Reg and BB-Reg have been considered in (Morel and Neerchal, 2012). The quadratic term in the regression model was previously suggested in (Sofuni et al., 1978). Morel and Neerchal (2012) consider linking the regression to the overdispersion parameter in RCB and BB, in addition to the probability of aberration, indicating that the amount of overdispersion also varies with radiation dose.

The MLEs and corresponding standard errors for the candidate models are given in Table 4.6. All models give roughly similar estimates of $\boldsymbol{\beta}$, having the same sign and similar magnitude. The standard errors for Logistic are noticeably smaller than the other models, indicating that the extra variation in the data is not being reflected as uncertainty in the estimates. The other models allow the standard error to be inflated. The standard errors for MixLinkJ2 have been computed by numerical Hessian as in (4.23); for comparison, the standard errors via (4.24) using $B = 500$ bootstrap samples are given in Table 4.7. The two methods give similar standard errors, with κ having the most notable difference. Figure 4.15 displays the empirical CDF of the $\hat{\kappa}^{(b)}$, showing several large values which seem to break away from the natural curve of the remaining values. It is possible that these larger values are artifacts of the numerical optimization and the Hessian-based estimate $\widehat{\text{Var}}(\hat{\boldsymbol{\theta}})$ is a reasonable variance estimate of the MLE.

Table 4.8 shows the GOF test result for each model along with other standard model

Table 4.6: Maximum likelihood estimates for candidate models, with standard errors in parentheses.

	Logistic		RCB		BB
β_0	-3.0306 (0.0246)	β_0	-2.9901 (0.0352)	β_0	-2.9487 (0.0445)
β_1	1.3017 (0.0343)	β_1	1.2040 (0.0415)	β_1	1.1144 (0.0550)
β_2	-0.3071 (0.0158)	β_2	-0.3429 (0.0242)	β_2	-0.2676 (0.0276)
		ϕ	0.1511 (0.0080)	ϕ	0.1661 (0.0076)

	RCB-Reg		BB-Reg		MixLinkJ2
β_0	-3.0699 (0.0338)	β_0	-3.0145 (0.0445)	β_0	-3.0061 (0.0441)
β_1	1.3010 (0.0444)	β_1	1.3594 (0.0564)	β_1	1.3656 (0.0562)
β_2	-0.3705 (0.0244)	β_2	-0.3449 (0.0332)	β_2	-0.3383 (0.0314)
γ_0	-2.3526 (0.0965)	γ_0	-1.8611 (0.0737)	π_1	0.3297 (0.0175)
γ_1	0.9331 (0.1569)	γ_1	0.7993 (0.1109)	κ	1.6293 (0.2472)
γ_2	-0.2365 (0.0565)	γ_2	-0.1610 (0.0525)		

Table 4.7: Standard errors of MixLinkJ2 MLE computed using 500 parametric bootstrap samples.

	Bootstrap SE
β_0	0.0458
β_1	0.0520
β_2	0.0306
π_1	0.0170
κ	0.2858

selection criteria: $-2\,\text{LogLik}$, Akaike information criteria (AIC) and Bayesian information criterion (BIC). Here, LogLik is the maximized value of the log-likelihood so that AIC $= -2\,\text{LogLik} + 2q$ and BIC $= -2\,\text{LogLik} + q\log(n)$. First consider the information theoretic (AIC and BIC) criteria, where a smaller value indicates a preferable model. As expected, Logistic results in the largest AIC/BIC because it does not account for the suspected overdispersion. The two RCB models have smaller AIC/BIC than Logistic, but not as small as in the two BB models. The BB-Reg model fares appears to fit significantly better than BB, indicating that the overdispersion parameter varies with radiation dose. The MixLinkJ2 model fits almost as well as BB-Reg, even without modeling $\boldsymbol{\pi}$ or κ as a function of radiation dose.

The GOF results give additional insight into the quality of the fits. For each model,

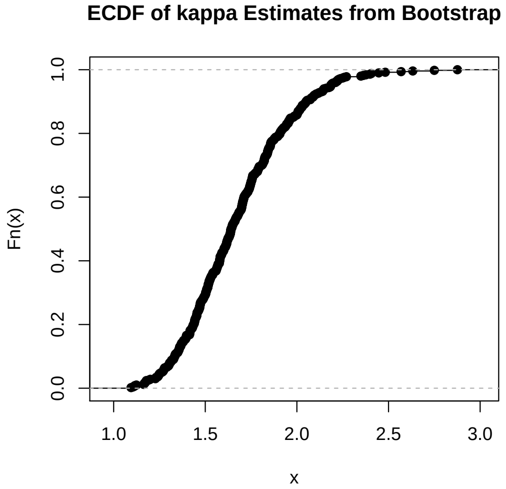


Figure 4.15: ECDF of estimates $\hat{\kappa}^{(b)}$ from 500 parametric bootstrap samples.

Table 4.8: Model comparison statistics.

					GOF		
Model	LogLik	q	AIC	BIC	statistic	df range	p-value
Logistic	-1814.189	3	3634.400	3647.799	110.38	[17,20]	$< 10^{-13}$
RCB	-1567.499	4	3142.997	3160.893	68.25	[15,19]	$< 10^{-6}$
BB	-1487.923	4	2983.847	3001.742	93.79	[12,18]	$< 10^{-11}$
RCB-Reg	-1546.612	6	3105.224	3132.067	63.96	[18,22]	$< 10^{-5}$
BB-Reg	-1429.605	6	2871.211	2898.054	19.40	[17,23]	> 0.3063
MixLinkJ2	-1433.331	5	2876.662	2905.506	19.50	[18,23]	> 0.3615

the intervals $\mathcal{A}_\ell$ were chosen by first considering

$$\mathcal{A}_1 = [0, 0.0099], \mathcal{A}_2 = (0.0099, 0.0198], \ldots, \mathcal{A}_{r-1} = (0.2970, 0.3069]$$

of the same length, and $\mathcal{A}_r = (0.3069, 1]$. This partitioning was selected so the results can be compared to (Morel and Neerchal, 2012). Using the (ungrouped) MLE for the model, expected counts for each $\mathcal{A}_\ell$ were computed, and $\mathcal{A}_\ell$ having expected counts less than 5 were merged with a neighboring interval. Table 4.9 shows detailed computations for the GOF statistic (4.25) for the models under consideration, and Figure 4.16 shows corresponding plots. The grey bars represent the observed counts for a given interval, and the black dots plot the expected counts using the MLE. The GOF comparison gives a similar ranking of models as the AIC/BIC comparison. The BB-Reg and MixLinkJ2 models give both give a statistically adequate fit, while the others do not. MixLinkJ2 attains a slightly better fit (a larger p-value) than does BB-Reg, as it features one less unknown parameter. One feature which seems to be a challenge to model is the large number of observations with a very low proportion of aberrations; these are counted in the first interval.

To validate the assumptions needed for the GOF test, Figure 4.17 shows the empirical CDF of the GOF test statistic. It has been computed using $B = 200$ parametric bootstrap samples in the manner discussed in Section 4.9.2. As predicted by the theory, the distribution appears to be a χ^2, between χ^2_{r-1} where no degrees of freedom are spent, and χ^2_{r-1-q} where q degrees of freedom are used to estimated the q unknown parameters. The bootstrap procedure also yields p-value$_{\text{Boot}} = 0.46$, which can be compared to the range $[0.3615, 0.6717]$ computed from χ^2_{r-1-q} and χ^2_{r-1}.

Table 4.9: Details for GOF test computed from the six models.

(a) Logistic GOF

Interval	Obs	Exp
$[0.0000, 0.0099]$	140	94.23
$(0.0099, 0.0198]$	107	108.24
$(0.0198, 0.0297]$	90	90.38
$(0.0297, 0.0396]$	52	62.30
$(0.0396, 0.0495]$	30	39.14
$(0.0495, 0.0594]$	32	30.35
$(0.0594, 0.0693]$	27	29.69
$(0.0693, 0.0792]$	33	22.06
$(0.0792, 0.0891]$	14	20.15
$(0.0891, 0.0990]$	13	17.63
$(0.0990, 0.1089]$	11	18.18
$(0.1089, 0.1188]$	11	15.49
$(0.1188, 0.1287]$	7	15.27
$(0.1287, 0.1386]$	8	15.04
$(0.1386, 0.1485]$	9	13.05
$(0.1485, 0.1584]$	8	12.63
$(0.1584, 0.1683]$	8	11.35
$(0.1683, 0.1782]$	6	8.94
$(0.1782, 0.1881]$	6	7.10
$(0.1881, 0.2079]$	8	9.44
$(0.2079, 1.0000]$	28	7.33
GOF statistic	110.38	
df	[17,20]	
p-value	$< 10^{-13}$	

(b) RCB GOF

Interval	Obs	Exp
$[0.0000, 0.0099]$	140	102.20
$(0.0099, 0.0198]$	107	112.13
$(0.0198, 0.0297]$	90	91.09
$(0.0297, 0.0396]$	52	63.54
$(0.0396, 0.0495]$	30	41.92
$(0.0495, 0.0594]$	32	34.40
$(0.0594, 0.0693]$	27	34.36
$(0.0693, 0.0792]$	33	26.91
$(0.0792, 0.0891]$	14	24.92
$(0.0891, 0.0990]$	13	20.54
$(0.0990, 0.1089]$	11	19.32
$(0.1089, 0.1188]$	11	14.26
$(0.1188, 0.1287]$	7	11.14
$(0.1287, 0.1386]$	8	8.86
$(0.1386, 0.1485]$	9	5.88
$(0.1485, 0.1683]$	16	8.44
$(0.1683, 0.1980]$	18	7.46
$(0.1980, 0.2376]$	10	9.08
$(0.2376, 0.2673]$	7	5.53
$(0.2673, 1.0000]$	13	6.02
GOF statistic	68.25	
df	[15,19]	
p-value	$< 10^{-6}$	

(c) RCB-Reg GOF

Interval	Obs	Exp
$[0.0000, 0.0099]$	140	107.72
$(0.0099, 0.0198]$	107	113.90
$(0.0198, 0.0297]$	90	89.83
$(0.0297, 0.0396]$	52	62.57
$(0.0396, 0.0495]$	30	42.26
$(0.0495, 0.0594]$	32	35.64
$(0.0594, 0.0693]$	27	35.96
$(0.0693, 0.0792]$	33	28.37
$(0.0792, 0.0891]$	14	26.01
$(0.0891, 0.0990]$	13	21.14
$(0.0990, 0.1089]$	11	19.45
$(0.1089, 0.1188]$	11	13.98
$(0.1188, 0.1287]$	7	10.58
$(0.1287, 0.1386]$	8	8.14
$(0.1386, 0.1485]$	9	5.07
$(0.1485, 0.1881]$	28	9.74
$(0.1881, 0.2376]$	16	6.01
$(0.2376, 0.2871]$	12	6.03
$(0.2871, 1.0000]$	8	5.59
GOF statistic	93.79	
df	[12,18]	
p-value	$< 10^{-11}$	

Table 4.9: (Continued).

(d) BB GOF

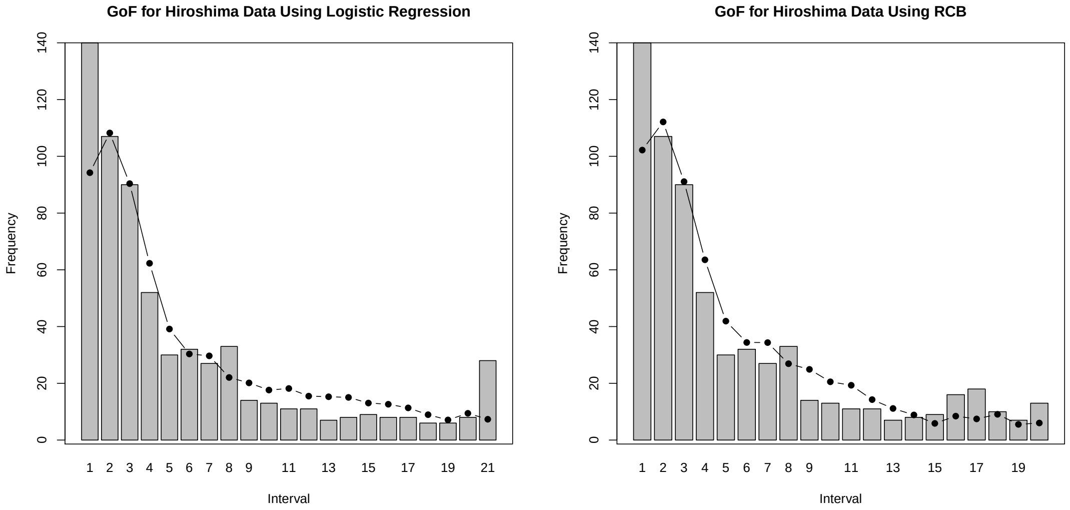
Interval	Obs	Exp
$[0.0000, 0.0099]$	140	159.97
$(0.0099, 0.0198]$	107	77.62
$(0.0198, 0.0297]$	90	60.51
$(0.0297, 0.0396]$	52	50.54
$(0.0396, 0.0495]$	30	38.39
$(0.0495, 0.0594]$	32	32.78
$(0.0594, 0.0693]$	27	31.91
$(0.0693, 0.0792]$	33	24.67
$(0.0792, 0.0891]$	14	22.63
$(0.0891, 0.0990]$	13	19.28
$(0.0990, 0.1089]$	11	19.01
$(0.1089, 0.1188]$	11	15.22
$(0.1188, 0.1287]$	7	13.98
$(0.1287, 0.1386]$	8	12.62
$(0.1386, 0.1485]$	9	10.20
$(0.1485, 0.1584]$	8	9.39
$(0.1584, 0.1683]$	8	8.52
$(0.1683, 0.1782]$	6	6.78
$(0.1782, 0.1881]$	6	5.77
$(0.1881, 0.2079]$	8	9.37
$(0.2079, 0.2277]$	6	6.39
$(0.2277, 0.2574]$	6	6.11
$(0.2574, 1.0000]$	16	6.33
GOF statistic	63.96	
df	[18,22]	
p-value	$< 10^{-5}$	

(e) BB-Reg GOF

Interval	Obs	Exp
$[0.0000, 0.0099]$	140	135.19
$(0.0099, 0.0198]$	107	109.96
$(0.0198, 0.0297]$	90	83.69
$(0.0297, 0.0396]$	52	59.89
$(0.0396, 0.0495]$	30	38.96
$(0.0495, 0.0594]$	32	29.39
$(0.0594, 0.0693]$	27	26.79
$(0.0693, 0.0792]$	33	19.35
$(0.0792, 0.0891]$	14	17.06
$(0.0891, 0.0990]$	13	14.30
$(0.0990, 0.1089]$	11	13.89
$(0.1089, 0.1188]$	11	10.87
$(0.1188, 0.1287]$	7	9.92
$(0.1287, 0.1386]$	8	9.13
$(0.1386, 0.1485]$	9	7.29
$(0.1485, 0.1584]$	8	6.90
$(0.1584, 0.1683]$	8	6.44
$(0.1683, 0.1782]$	6	5.30
$(0.1782, 0.1980]$	12	8.79
$(0.1980, 0.2178]$	5	7.48
$(0.2178, 0.2376]$	5	5.75
$(0.2376, 0.2673]$	7	6.60
$(0.2673, 0.3069]$	5	5.92
$(0.3069, 1.0000]$	8	9.12
GOF statistic	19.40	
df	[17,23]	
p-value	> 0.3063	

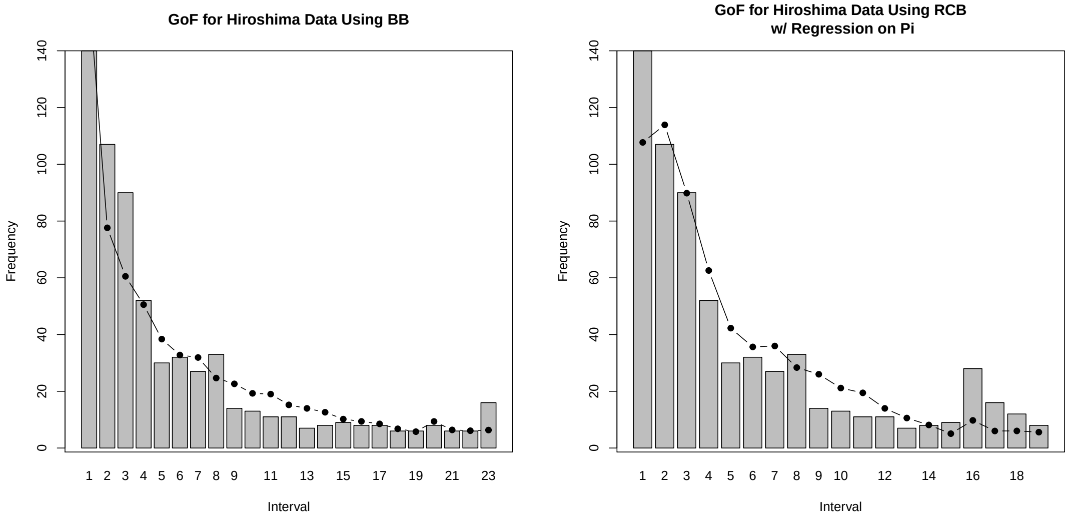
(f) MixLinkJ2 GOF

Interval	Obs	Exp
$[0.0000, 0.0099]$	140	138.76
$(0.0099, 0.0198]$	107	106.83
$(0.0198, 0.0297]$	90	80.94
$(0.0297, 0.0396]$	52	58.27
$(0.0396, 0.0495]$	30	38.16
$(0.0495, 0.0594]$	32	28.91
$(0.0594, 0.0693]$	27	26.42
$(0.0693, 0.0792]$	33	19.14
$(0.0792, 0.0891]$	14	17.03
$(0.0891, 0.0990]$	13	14.44
$(0.0990, 0.1089]$	11	14.21
$(0.1089, 0.1188]$	11	11.30
$(0.1188, 0.1287]$	7	10.50
$(0.1287, 0.1386]$	8	9.71
$(0.1386, 0.1485]$	9	7.90
$(0.1485, 0.1584]$	8	7.52
$(0.1584, 0.1683]$	8	7.00
$(0.1683, 0.1782]$	6	5.78
$(0.1782, 0.1980]$	12	9.48
$(0.1980, 0.2178]$	5	7.78
$(0.2178, 0.2376]$	5	5.71
$(0.2376, 0.2673]$	7	6.24
$(0.2673, 0.3069]$	5	5.59
$(0.3069, 1.0000]$	8	10.39
GOF statistic	19.50	
df	[18,23]	
p-value	> 0.3615	



(a) Logistic.

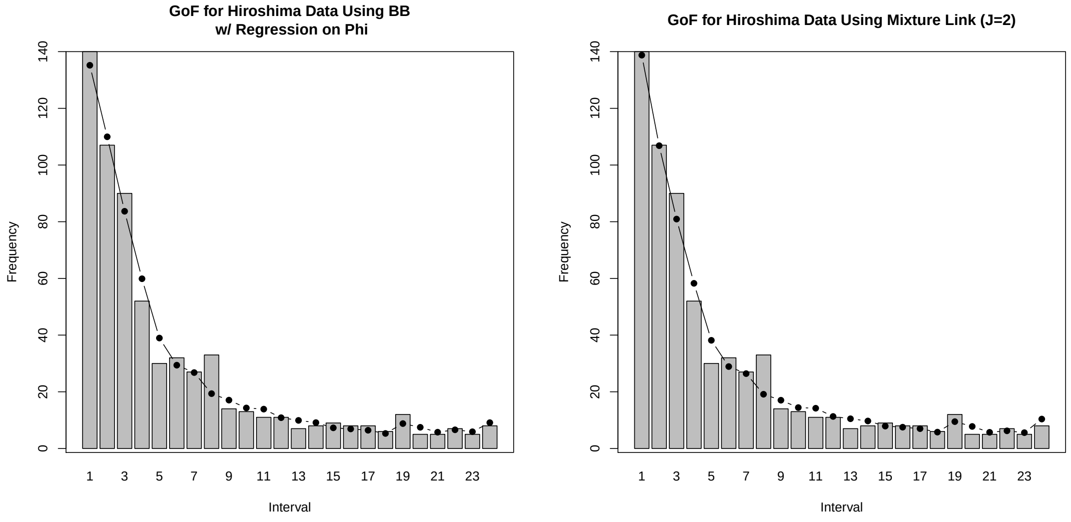
(b) RCB.



(c) BB.

(d) RCB-Reg.

Figure 4.16: GOF plots for observed vs. expected counts. The grey bars represent the observed counts for a given interval, and the black dots are the expected counts under the MLE. Note that the choice of intervals varies between models.



(e) BB-Reg. (f) MixLinkJ2.

Figure 4.16: (Continued).

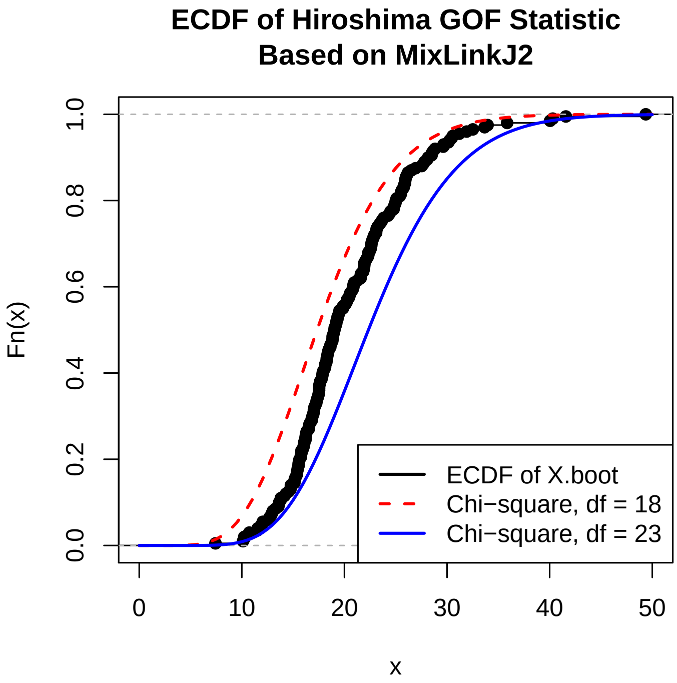


Figure 4.17: Empirical CDF of MixLinkJ2 GOF statistic computed from parametric bootstrap.

4.9.4 Mutagenic Data

Neerchal and Morel (1998) present a series of illustrative datasets where male mice are treated with a suspected mutagen and paired with one or more female mice. The n female mice are exposed to the male partner for a length of time and then sacrificed. The uterus of the ith female is then examined. The total number of implanted fetuses is denoted as m_i, and the number which are inviable (i.e. not alive) is t_i. Hence, t_i can be considered binomial data from m_i trials. The m_i have previously been treated as fixed for simplicity, as will be done in the present study. The data do not indicate which male was paired to each female, and the pairing of one male with multiple females could violate a naive assumption of independence among t_i, leading to overdispersion. Neerchal and Morel (1998) compare the viability of RCB and BB for three different datasets, denoted Dataset I, II, and III which correspond to three mutagenic experiments, using the GOF test statistic discussed in Section 4.9.2. In this section, we extend the comparison to include Mixture Link.

The following models are now under consideration:

- Binomial: $T_i \overset{\text{ind}}{\sim} \text{Bin}(m_i, p)$,
- RCB: $T_i \overset{\text{ind}}{\sim} \text{RCB}(m_i, p, \phi)$,
- BB: $T_i \overset{\text{ind}}{\sim} \text{BB}(m_i, p, \phi)$,
- MixLinkJ2: $T_i \overset{\text{ind}}{\sim} \text{MixLink}_2(m_i, p, \boldsymbol{\pi}, \kappa)$.

Notice that there is no covariate, and that all four models are applicable. The three models with extra variation suggest different explanations for the departure from binomial. RCB suspects a certain dependence among the m_i fetuses for each female mouse. BB supposes the probability of inviability for each litter is drawn randomly. MixLinkJ2 makes use of a mixture of two latent binomial subpopulations, each having different probability of inviability, where p represents the marginal inviability probability of the overall population.

Table 4.10 gives estimates and standard errors for the four models on Datasets I,

Table 4.10: Estimates for mutagenic datasets with standard errors in parentheses.

		Dataset I	Dataset II	Dataset III
Binomial	p	0.0893 (0.0034)	0.1079 (0.0030)	0.0718 (0.0032)
RCB	p	0.0890 (0.0045)	0.1088 (0.0035)	0.0760 (0.0049)
	ϕ	0.2550 (0.0223)	0.2031 (0.0195)	0.3235 (0.0302)
BB	p	0.0901 (0.0047)	0.1086 (0.0035)	0.0739 (0.0045)
	ϕ	0.2611 (0.0188)	0.2070 (0.0172)	0.2741 (0.0207)
MixLinkJ2	p	0.0921 (0.0050)	0.1091 (0.0035)	0.0756 (0.0053)
	π_1	0.2083 (0.0255)	0.2935 (0.0309)	0.0438 (0.0114)
	κ	2.2469 (1.3163)	3.4942 (1.8510)	5.6057 (5.5856)

II, and III. The estimates were obtained using numerical optimization (`optim` in `R`), and standard errors are computed from the Hessian. The estimates for RCB and BB obtained in this manner match exactly to (Neerchal and Morel, 1998), but some of the standard errors differ. As expected, the standard errors of $\hat{p}$ from the extra variation models are larger than those computed under Binomial. Tables 4.11, 4.12, and 4.13 show GOF statistics, along with AIC and BIC. Note that a range is given for the degrees of freedom (df) and p-value because of the recovery of df phenomenon discussed in Section 4.9.2. The accompanying Figures 4.18, 4.19, and 4.20 display the fits graphically. The grey bars represent observed counts and the black dots represent expected counts under the fitted model. The intervals chosen for GOF computation have been selected to match (Neerchal and Morel, 1998).

For all three datasets, the fit for MixLinkJ2 is on par with the better fit between BB and RCB, if not providing the best fit itself. For Dataset I, MixLinkJ2 is best in terms of the GOF statistic, while BB has the smallest AIC and BIC. For Dataset II, none of the models have a p-value indicating a particularly good fit, but the plots show that BB, RCB, and MixLinkJ2 are all capturing the general shape of the data. Finally, for Dataset III, MixLinkJ2 appears to give a slightly better fit than RCB in terms of GOF, AIC, and BIC, while BB is rejected by the GOF test.

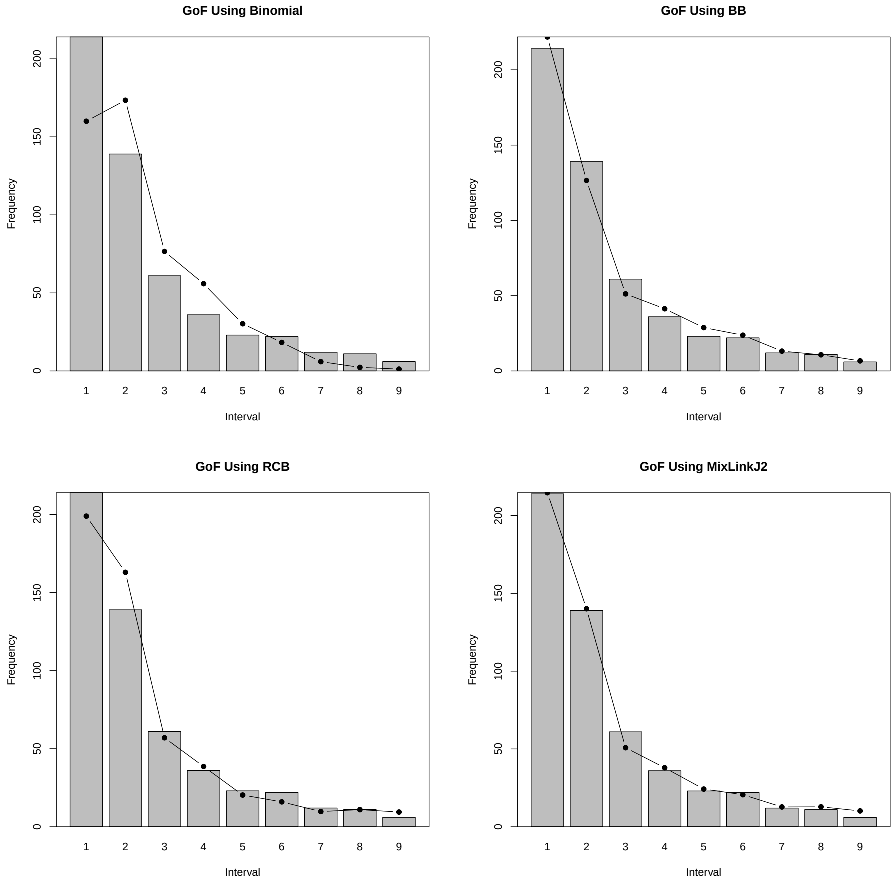


Figure 4.18: GOF for Mutagenic Dataset I.

Table 4.11: GOF for Mutagenic Dataset I.

		Expected			
Interval	Obs	Binomial	RCB	BB	MixLinkJ2
$[0, 1/21]$	214	160.01	199.01	221.82	214.69
$(1/21, 2/21]$	139	173.46	163.03	126.52	140.11
$(2/21, 3/21]$	61	76.59	56.99	51.19	50.78
$(3/21, 4/21]$	36	55.92	38.61	41.31	37.92
$(4/21, 5/21]$	23	30.25	20.31	28.79	24.21
$(5/21, 6/21]$	22	18.29	15.94	23.76	20.61
$(6/21, 7/21]$	12	5.97	9.74	13.18	12.71
$(7/21, 9/21]$	11	2.31	10.95	10.74	12.80
$(9/21, 1]$	6	1.20	9.42	6.69	10.17
X		95.724	9.556	5.545	4.320
df		$[7, 8]$	$[6, 8]$	$[6, 8]$	$[5, 8]$
p-value (lower)		0	0.1446	0.4760	0.5043
p-value (upper)		0	0.2977	0.6980	0.8272
AIC		1687.779	1570.628	1559.833	1562.879
BIC		1692.041	1579.151	1568.356	1581.925

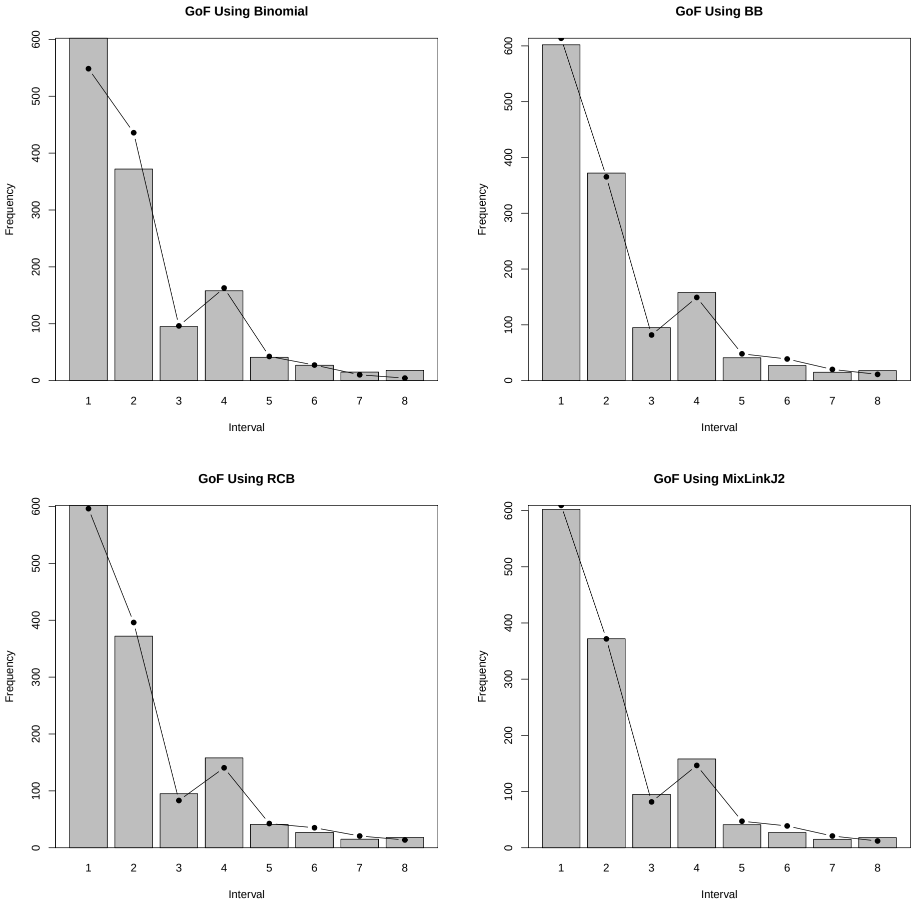


Figure 4.19: GOF for Mutagenic Dataset II.

Table 4.12: GOF for Mutagenic Dataset II.

		Expected			
Interval	Obs	Binomial	RCB	BB	MixLinkJ2
$[0, 1/14]$	602	548.48	596.26	613.71	609.30
$(1/14, 2/14]$	372	435.88	396.04	365.28	371.72
$(2/14, 3/14]$	95	96.22	83.07	81.81	81.59
$(3/14, 4/14]$	158	162.90	140.56	149.11	146.46
$(4/14, 5/14]$	41	42.49	42.56	48.01	47.19
$(5/14, 6/14]$	27	27.35	35.16	38.79	38.81
$(6/14, 7/14]$	15	10.25	20.64	20.07	20.90
$(7/14, 1]$	18	4.42	13.71	11.22	12.03
X		58.789	10.227	12.990	12.238
df		$[6, 7]$	$[5, 7]$	$[5, 7]$	$[4, 7]$
p-value (lower)		7.9280E-11	0.0690	0.0235	0.0157
p-value (upper)		2.6329E-10	0.1761	0.0723	0.0930
AIC		3375.531	3317.198	3318.598	3323.018
BIC		3380.723	3327.581	3328.981	3345.783

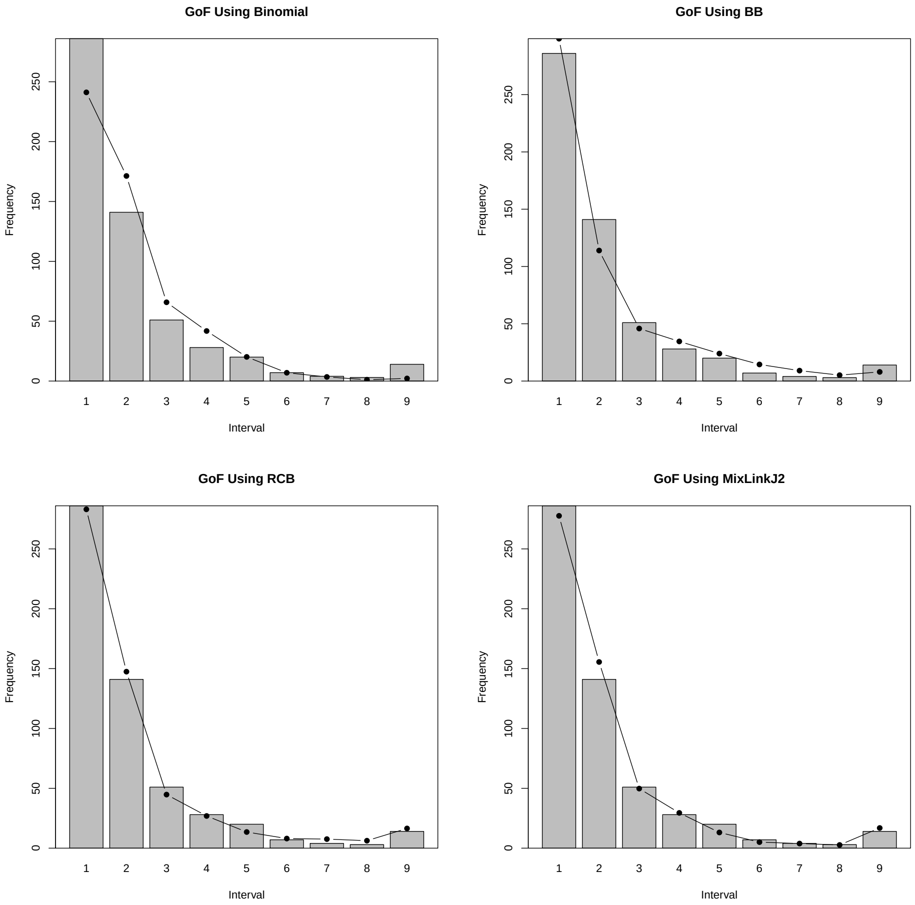


Figure 4.20: GOF for Mutagenic Dataset III.

Table 4.13: GOF for Mutagenic Dataset III.

		Expected			
Interval	Obs	Binomial	RCB	BB	MixLinkJ2
$[0, 1/19]$	286	241.16	283.05	298.85	277.62
$(1/19, 2/19]$	141	171.33	147.44	113.95	155.54
$(2/19, 3/19]$	51	65.81	44.76	45.85	49.77
$(3/19, 4/19]$	28	41.82	26.90	34.59	29.47
$(4/19, 5/19]$	20	20.19	13.50	24.00	13.13
$(5/19, 6/19]$	7	6.94	8.04	14.50	5.07
$(6/19, 7/19]$	4	3.44	7.58	9.10	3.85
$(7/19, 8/19]$	3	1.15	6.25	5.16	2.68
$(8/19, 1]$	14	2.18	16.47	7.99	16.86
X		88.786	8.248	21.638	6.570
df		$[7, 8]$	$[6, 8]$	$[6, 8]$	$[5, 8]$
p-value (lower)		2.2204E-16	0.2205	0.0014	0.2546
p-value (upper)		7.7716E-16	0.4096	0.0056	0.5837
AIC		1532.123	1382.750	1406.663	1376.636
BIC		1536.441	1391.384	1415.297	1395.904

4.10 Conclusions

In this chapter, we have presented a new binomial model with extra variation called Mixture Link, starting from the finite mixture of binomials and linking a regression to the mixture probability of success. This lead us to consider a random effects model on the set representing the link from the likelihood to the regression. The random effects are modeled by a Dirichlet distribution placed on the simplex between extreme points of the set. An algorithm was given to find all extreme points, which is needed to use the distribution in any practical way. The expectation and variance of Mixture Link were obtained through the moments of Dirichlet. Evaluation of the Mixture Link density is seen to involve an integral over the distribution of linear combination of Dirichlet, which must be computed numerically except in some special cases. One general method of exact numerical evaluation was discussed, but we have found it to be too slow for use in applications. A moment-matched beta approximation to the linear combination of Dirichlet distribution was proposed to facilitate computation; empirical results show that integrated density matches closely with the exact density, but theoretical justification is needed. Plots of the Mixture Link density show that it takes on a variety of expressive shapes. As promising first applications, Mixture Link is shown to fit the chromosome aberration and mutagenic data well in terms of AIC/BIC and goodness-of-fit. Initial results for Mixture Link are encouraging, and the model appears worthy of further study as a tool for the analysis of binomial data.

Future work is needed so that Mixture Link can be used in application. A more appropriate estimation method than numerical MLE is desired; ideally it would avoid derivatives because of differentiability issues seen in the likelihood. Theoretical properties, such as the consistency and asymptotic distribution of estimators, must be investigated in light of the fact that usual regularity conditions may not be satisfied. The effect of increasing J remains to be studied: whether more variation will effectively be modeled or there will be diminishing returns.

Although we have focused exclusively on binomial data, the Mixture Link approach can be extended to other kinds of finite mixtures such as normal and poisson. In the normal case, it may be desired to link a regression $\phi = \boldsymbol{x}^T\boldsymbol{\beta}$ to the mixture mean $\sum_{j=1}^J \pi_j\mu_j = \boldsymbol{\mu}^T\boldsymbol{\pi}$. The set $A(\phi, \boldsymbol{\pi}) = \{\boldsymbol{\mu} \in \mathbb{R}^J : \boldsymbol{\mu}^T\boldsymbol{\pi} = \phi\}$ is now an unbounded hyperplane, and a finite set of vertices is no longer appropriate to characterize it. In the case of poisson count data where $\mu_j > 0$ is a rate, $A(\phi, \boldsymbol{\pi}) = \{\boldsymbol{\mu} \in [0, \infty)^J : \boldsymbol{\mu}^T\boldsymbol{\pi} = \phi\}$ is a hyperplane constrained within the nonnegative orthant. In a broader sense, it may be desired to link a regression to a composite parameter which does not explicitly appear in the likelihood of interest. Our approach can be considered for these cases, provided an appropriate random effects distribution on the set A can be determined.

Chapter 5

Conclusion

In this dissertation, we have focused on three problems in the area of finite mixtures and overdispersion modeling of binomial and multinomial data. We first considered a matrix which had previously been proposed to approximate the information matrix in multinomial finite mixtures. We showed that this matrix is actually the information matrix of the joint complete data, where the latent subpopulation label is observed in addition to the multinomial outcome. This allows a similar information matrix approximation to be formulated for any missing data problem, including those involving mixtures. It also allows the technique of approximate scoring to be applied in these general settings, and brings to light the close relationship between approximate scoring and EM which was first noted in previous work. Simulation studies demonstrated the closeness of the two algorithms. While it was noted that the exact information matrix and the approximation themselves may not be close when the number of multinomial trials is not large, the approximation was seen to be quite effective when used to compute estimates by scoring. A hybrid method using approximate and exact scoring was seen to take advantage of both the robustness of approximate scoring and the fast convergence of exact scoring.

An extension of the approximate information matrix to exponential family finite mixtures was considered next. The extension supposes a clustered sampling scheme so that m observations are sampled within the same (but unknown) subpopulation. This provides an analogue to the m trials in the multinomial setting and allows the main convergence result, that the exact and approximate information matrices converge together as m is taken to infinity, to be extended to exponential family finite mixtures. A bound on the rate of convergence is obtained, which is exponential in m but where the exponent depends on the similarity between pairs of subpopulations. When two subpopulations are more identical, the convergence rate is slowed considerably. The rate of convergence between approximate and exact information is seen to be related to the optimal probably of misclassification. Simulations showed that the convergence result does not hold for the usual independent and identically distributed sampling scheme (as opposed to clustered sampling). It was seen in several examples that the convergence result holds for continuous mixing distributions and finite mixtures of non-exponential family densities. It is future work to consider extending the proof to these cases. It would also be of interest if the accuracy of the information matrix approximation could be improved when m is not large, or populations are close together.

The last part of the thesis considered linking a regression model to the mixed probability of success in a binomial finite mixture. The Mixture Link model was formulated as a random effects model, and some results were obtained to find vertices of the set where the link is enforced, to compute the likelihood, and to compute moments. Computing the likelihood exactly requires an expectation over the linear combination of Dirichlet distribution, which is known to be difficult. We saw empirically that moment-matching a simple beta distribution serves as a very good approximation, and simplifies computation. Initial results show that Mixture Link is useful in data analysis using the numerical MLE, despite differentiability issues in the likelihood. Further study is required to more thoroughly address issues such as identifiability, computation of the likelihood, and estimation of parameters and standard errors. We believe our random effects approach will generalize beyond the binomial setting, and may be useful for linking regressions to composite parameters in other models.

This thesis has taken a frequentist perspective, but the study of finite mixtures and overdispersion is also of interest in Bayesian statistical analysis. One particularly interesting topic in Bayesian mixture analysis is the Dirichlet Process Mixture (Ferguson, 1973; Neal, 2000), which has been heavily studied in recent years. An appealing aspect of the Dirichlet Process Mixture is that it avoids the need to for the analyst to select a number of mixture components and instead infers this from the data. It may therefore be worthwhile to revisit the approximate information and Mixture Link model from a Bayesian perspective.

Appendix A

Additional Results

Remark A.1 (Transformation from chi-square to Dirichlet)**.** Here we show the transformation from chi-square to Dirichlet. Suppose $X_j \overset{\text{ind}}{\sim} \chi^2_{v_j/2}$ for $j = 0, \ldots, k$ and consider the transformation

$$W_0 = \sum_{j=0}^{k} X_j, \quad W_1 = X_1 / \sum_{j=0}^{k} X_j, \quad \ldots, \quad W_k = X_k / \sum_{j=0}^{k} X_j$$

Notice that $\sum_{j=1}^{k} W_j = 1 - X_0 / \sum_{j=0}^{k} X_j$ so the inverse transformation is

$$X_0 = W_0 \left(1 - \sum_{j=1}^{k} W_j \right), \quad X_1 = W_1 W_0, \quad \ldots, \quad X_k = W_k W_0,$$

and the ranges of the new variables are

$$W_0 \in (0, \infty), \quad \text{and} \quad \begin{pmatrix} (1 - \sum_{j=1}^{k} W_j) \\ W_1 \\ \vdots \\ W_k \end{pmatrix} \in \mathcal{S}^{k+1},$$

where $\mathcal{S}^{k+1}$ is the $(k+1)$-dimensional probability simplex. The Jacobian of the transformation is

$$\begin{aligned}
\boldsymbol{J} &= \begin{pmatrix} \frac{\partial X_0}{\partial W_0} & \cdots & \frac{\partial X_0}{\partial W_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial X_k}{\partial W_0} & \cdots & \frac{\partial X_k}{\partial W_k} \end{pmatrix} \\
&= \begin{pmatrix} (1-\sum_{j=1}^{k} W_j) & -W_0 & -W_0 & \cdots & -W_0 & -W_0 \\ W_1 & W_0 & 0 & \cdots & 0 & 0 \\ W_2 & 0 & W_0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ W_{k-1} & 0 & 0 & \cdots & W_0 & 0 \\ W_k & 0 & 0 & \cdots & 0 & W_0 \end{pmatrix} \\
&= \begin{pmatrix} 1-\sum_{j=1}^{k} W_j & -W_0 \mathbf{1}_k^T \\ \boldsymbol{W}_{-0} & W_0 \boldsymbol{I}_k \end{pmatrix}
\end{aligned}$$

where $\mathbf{1}_k$ is a k-dimensional vector of ones, $\boldsymbol{I}_k$ is the $k \times k$ identity matrix, and $\boldsymbol{W}_{-0} = (W_1, \ldots, W_k)$. We may now find the determinant using the well-known property for block matrices

$$\det \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix} = \det(\boldsymbol{D}) \det(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}).$$

This gives

$$\begin{aligned}
\det \boldsymbol{J} &= \det(W_0^k) \det \left(1 - \sum_{j=1}^{k} W_j - -W_0 \mathbf{1}_k^T (W_0 \boldsymbol{I}_k)^{-1} \boldsymbol{W}_{-0} \right) \\
&= W_0^k \det \left(1 - \sum_{j=1}^{k} W_j + \mathbf{1}_k^T \boldsymbol{W}_{-0} \right) \\
&= W_0^k
\end{aligned}$$

Therefore,

$$\begin{aligned}
f(w_0, w_1, \ldots, w_k) &= f(x_0, x_1, \ldots, x_k) |\det \boldsymbol{J}| \\
&= \prod_{j=0}^{k} \frac{x_j^{v_j/2-1} e^{-x_j/2}}{\Gamma(v_j/2) 2^{v_j/2}} w_0^k \\
&= \frac{[w_0(1 - \sum_{j=1}^k w_j)]^{v_0/2-1} e^{-\frac{1}{2} w_0 (1 - \sum_{j=1}^k)}}{\Gamma(v_0/2) 2^{v_0/2}} \prod_{j=1}^{k} \frac{(w_0 w_j)^{v_j/2-1} e^{-w_0 w_j/2}}{\Gamma(v_j/2) 2^{v_j/2}} w_0^k \\
&= \frac{w_0^{\frac{1}{2}\sum_{j=0}^k v_j/2-1} e^{-\frac{1}{2} w_0}}{\Gamma(v_0/2) 2^{v_0/2}} \left(1 - \sum_{j=1}^{k} w_j\right)^{v_0/2-1} \prod_{j=1}^{k} \frac{w_j^{v_j/2-1}}{\Gamma(v_j/2) 2^{v_j/2}}.
\end{aligned}$$

Now, to find the marginal distribution of $(W_1, \ldots, W_k)$,

$$\begin{aligned}
f(w_1, \ldots, w_k) &= \int f(w_0, w_1, \ldots, w_k) dw_0 \\
&= \left(1 - \sum_{j=1}^{k} w_j\right)^{v_0/2-1} \prod_{j=1}^{k} \frac{w_j^{v_j/2-1}}{\Gamma(v_j/2) 2^{v_j/2}} \frac{1}{\Gamma(v_0/2) 2^{v_0/2}} \int_0^\infty w_0^{\frac{1}{2}\sum_{j=0}^k v_j/2-1} e^{-\frac{1}{2} w_0} dw_0 \\
&= \left(1 - \sum_{j=1}^{k} w_j\right)^{v_0/2-1} \prod_{j=1}^{k} \frac{w_j^{v_j/2-1}}{\Gamma(v_j/2) 2^{v_j/2}} \frac{\Gamma(\sum_{j=0}^k v_j/2) 2^{\sum_{j=0}^k v_j/2}}{\Gamma(v_0/2) 2^{v_0/2}} \\
&= \frac{\Gamma(\sum_{j=0}^k v_j/2)}{\Gamma(v_0/2)\Gamma(v_1/2) \cdots \Gamma(v_k/2)} w_1^{v_1/2-1} \cdots w_k^{v_k/2-1} \left(1 - \sum_{j=1}^{k} w_j\right)^{v_0/2-1}.
\end{aligned}$$

Let $D_0 = 1 - \sum_{j=1}^k W_j$ and $D_j = W_j$ for $j = 1, \ldots, k$; then we have that $(D_0, D_1, \ldots, D_k)$ is distributed as Dirichlet$_{k+1}(v_0/2, v_1/2, \ldots, v_k/2)$.

Let $\boldsymbol{a} = (a_1, \ldots, a_k)$ and $\boldsymbol{\lambda} \sim$ Dirichlet$_k(\boldsymbol{\alpha})$. Notice that to obtain the joint distribution of $(a_1\lambda_1, \ldots, a_k\lambda_k)$, we may write

$$\begin{pmatrix} a_1\lambda_1 \\ \vdots \\ a_k\lambda_k \end{pmatrix} = \boldsymbol{D}_a \boldsymbol{\lambda}, \quad \text{where } \boldsymbol{D}_a = \mathrm{Diag}(\boldsymbol{a})$$

and therefore

$$\mathrm{P}(\boldsymbol{D}_a\boldsymbol{\lambda} \leq \boldsymbol{x}) = \mathrm{P}(\boldsymbol{\lambda} \leq \boldsymbol{D}_a^{-1}\boldsymbol{x}).$$

Finding the marginal distribution of $\boldsymbol{a}^T\boldsymbol{\lambda}$ from the joint of $\boldsymbol{D}_a\boldsymbol{\lambda}$ is of great interest to Chapter 4. But this involves a complicated multidimensional integral for general k which apparently does not have a simple closed form. Therefore, numerical procedures must be considered when working with this distribution.

Lemma A.2. *Suppose $\boldsymbol{A}$ and $\boldsymbol{B}$ are non-singular $q \times q$ matrices. Then*

$$\boldsymbol{A}^{-1} - \boldsymbol{B}^{-1} = \boldsymbol{B}^{-1}(\boldsymbol{B} - \boldsymbol{A})\boldsymbol{A}^{-1}$$

Proof. We have

$$\begin{aligned} \boldsymbol{B} - \boldsymbol{A} &= \boldsymbol{B} - \boldsymbol{A} \\ \iff I - \boldsymbol{B}^{-1}\boldsymbol{A} &= \boldsymbol{B}^{-1}(\boldsymbol{B} - \boldsymbol{A}) \\ \iff \boldsymbol{A}^{-1} - \boldsymbol{B}^{-1} &= \boldsymbol{B}^{-1}(\boldsymbol{B} - \boldsymbol{A})\boldsymbol{A}^{-1}. \end{aligned}$$

□

Proposition A.3. *Suppose $\boldsymbol{A}, \boldsymbol{B}$ are $q \times q$ symmetric positive definite matrices, and $\boldsymbol{B} - \boldsymbol{A}$ is positive definite. Then $\boldsymbol{A}^{-1} - \boldsymbol{B}^{-1}$ is positive definite.*

Proof. Notice that

$$\boldsymbol{A}^{-1} - \boldsymbol{B}^{-1} = \boldsymbol{B}^{-1}(\boldsymbol{B} - \boldsymbol{A})\boldsymbol{A}^{-1}. \tag{A.1}$$

Suppose λ is an eigenvalue of (A.1), then we have

$$\begin{aligned} & \det(\boldsymbol{B}^{-1}(\boldsymbol{B}-\boldsymbol{A})\boldsymbol{A}^{-1}-\lambda I)=0 \\ \iff \quad & \det(\boldsymbol{B}^{-1/2}(\boldsymbol{B}-\boldsymbol{A})^{1/2}\boldsymbol{A}^{-1}(\boldsymbol{B}-\boldsymbol{A})^{1/2}\boldsymbol{B}^{-1/2}-\lambda I)=0, \end{aligned}$$

and therefore (A.1) and

$$\boldsymbol{B}^{-1/2}(\boldsymbol{B}-\boldsymbol{A})^{1/2}\boldsymbol{A}^{-1}(\boldsymbol{B}-\boldsymbol{A})^{1/2}\boldsymbol{B}^{-1/2} \tag{A.2}$$

have the same eigenvalues. Since (A.2) is symmetric positive definite, all eigenvalues are positive, and hence (A.1) is positive definite. □

Bibliography

Alan Agresti. *Categorical Data Analysis*. Wiley-Interscience, 2nd edition, 2002.

J. Aitchison and S. D. Silvey. Maximum-likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics*, 29:813–828, 1958.

Murray Aitkin. A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, 6:251–262, 1996.

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience, 3rd edition, 2003.

AkioA. Awa, Takeo Honda, Toshio Sofuni, Shotaro Neriishi, MichihiroC. Yoshida, and Takashi Matsui. Chromosome-aberration frequency in cultured blood-cells in relation to radiation dose of A-bomb survivor. *The Lancet*, 298(7730):903–905, 1971.

Mokhtar S. Bazaraa, John J. Jarvis, and Hanif D. Sherali. *Linear Programming and Network Flows*. Wiley, 4th edition, 2009.

W. R. Blischke. Moment estimators for the parameters of a mixture of two binomial distributions. *The Annals of Mathematical Statistics*, 33(2):444–454, 1962.

W. R. Blischke. Estimating the parameters of mixtures of binomial distributions. *Journal of the American Statistical Association*, 59(306):510–528, 1964.

Otilia Boldea and Jan R. Magnus. Maximum likelihood estimation of the multivariate normal mixture model. *Journal of the American Statistical Association*, 104(488): 1539–1549, 2009.

Daniel M. Bolt, Allan S. Cohen, and James A. Wollack. A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics*, 26(4):381–409, 2001.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Hamparsum Bozdogan. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

Satish Chandra. On the mixtures of probability distributions. *Scandinavian Journal of Statistics*, 4:105–112, 1977.

Michelle R. Danaher, Anindya Roy, Zhen Chen, Sunni L. Mumford, and Enrique F. Schisterman. Minkowski-Weyl priors for models with parameter constraints: An analysis of the biocycle study. *Journal of the American Statistical Association*, 107(500):1395–1409, 2012.

C. Mitchell Dayton and George B. Macready. Concomitant-variable latent-class models. *Journal of the American Statistical Association*, 83(401):173–178, 1988.

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

Pierre Duchesne and Pierre Lafaye De Micheaux. Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. *Computational Statistics & Data Analysis*, 54(4):858–862, 2010.

Ryan Elmore and Shaoli Wang. Identifiability and estimation in finite mixture models with multinomial components. Technical Report 03-04, Penn State University, Department of Statistics, 2003.

Thomas S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

Dean A. Follmann and Diane Lambert. Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association*, 84(405):295–300, 1989.

Sylvia Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer, 2006.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd edition, 2003.

Daniel B. Hall. Zero-inflated poisson and binomial regression with random effects: A case study. *Biometrics*, 56(4):1030–1039, 2000.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR ’99, pages 50–57. ACM, 1999.

Yi Huang. Average causal effect (ACE) estimation allowing covariate measurement error: A finite mixture modeling framework, 2012. Unpublished manuscript.

J. P. Imhof. Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48(3):419–426, 1961.

Murray Jorgensen. Using multinomial mixture models to cluster internet traffic. *Australian & New Zealand Journal of Statistics*, 46(2):205–218, 2004.

Maurice Kendall and Alan Stuart. *Kendalls Advanced Theory of Statistics*, volume 2. Charles Griffin & Company Limited, 4th edition, 1979.

Samuel Kotz, N. Balakrishnan, and Norman L. Johnson. *Continuous Multivariate Distributions, Volume 1, Models and Applications*. Wiley-Interscience, 2nd edition, 2000.

Kenneth Lange. *Numerical Analysis for Statisticians*. Springer, 2nd edition, 2010.

E. L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, 2nd edition, 1998.

E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, 3rd edition, 2005.

Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.

Bruce G. Lindsay. *Mixture Models: Theory, Geometry and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics. Institute of Mathematical Statistics, 1995.

Chuanhai Liu and Donald B. Rubin. ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, 5:19–39, 1995.

Minglei Liu. *Estimation for Finite Mixture Multinomial Models*. Phd thesis, University of Maryland, Baltimore County, Department of Mathematics and Statistics, 2005.

Thomas A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 44:226–233, 1982.

T. H. Matheiss and David S. Rubin. A survey and comparison of methods for finding all vertices of convex polyhedral sets. *Mathematics of Operations Research*, 5(2):167–185, 1980.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall/CRC, 2nd edition, 1989.

Charles E. McCulloch, Shayle R. Searle, and John M. Neuhaus. *Generalized, Linear, and Mixed Models*, volume 2. Wiley-Interscience, 2nd edition, 2008.

Geoffrey McLachlan and David Peel. *Finite Mixture Models*. Wiley-Interscience, 2000.

Xiao-Li Meng and Donald B. Rubin. Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association*, 86(416): 899–909, 1991.

Carl D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, 2001.

Jorge G. Morel and Nagaraj K. Nagaraj. A finite mixture distribution for modeling multinomial extra variation. Research Report 91–03, Department of Mathematics and Statistics, University of Maryland, Baltimore County, 1991.

Jorge G. Morel and Neerchal K. Nagaraj. A finite mixture distribution for modelling multinomial extra variation. *Biometrika*, 80(2):363–371, 1993.

Jorge G. Morel and Nagaraj K. Neerchal. *Overdispersion Models in SAS*. SAS Institute, 2012.

James E. Mosimann. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, 49(1):65–82, 1962.

Radford M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

Nagaraj K. Neerchal and Jorge G. Morel. Large cluster results for two parametric multinomial extra variation models. *Journal of the American Statistical Association*, 93 (443):1078–1087, 1998.

Nagaraj K. Neerchal and Jorge G. Morel. An improved method for the computation of maximum likelihood estimates for multinomial overdispersion models. *Computational Statistics & Data Analysis*, 49(1):33–43, 2005.

Masashi Okamoto. Some inequalities relating to the partial sum of binomial probabilities. *Annals of the Institute of Statistical Mathematics*, 10:29–35, 1959.

Terence Orchard and Max A. Woodbury. A missing information principle: theory and applications. In L. M. Le Cam, J. Neyman, and E. L. Scott, editors, *Proceedings of the Sixth Berkely Symposium on Mathematics, Statistics and Probability, Volume 1.*, pages 697–715. Berkeley: University of California Press, 1972.

Masanori Otake and Ross L. Prentice. The analysis of chromosomally aberrant cells based on beta-binomial distribution. *Radiation research*, 98(3):456–470, 1984.

Carlos Paulino and Carlos de Bragança Pereira. On identifiability of parametric statistical models. *Statistical Methods & Applications*, 3:125–151, 1994.

R. L. Prentice. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association*, 81(394):321–327, 1986.

Serge B. Provost and Young-Ho Cheong. On the distribution of linear combinations of the components of a dirichlet random vector. *Canadian Journal of Statistics*, 28(2): 417–425, 2000.

Andrew M. Raim, Matthias K. Gobbert, Nagaraj K. Neerchal, and Jorge G. Morel. Maximum likelihood estimation of the random-clumped multinomial model as prototype problem for large-scale statistical computing. *Journal of Statistical Computation and Simulation*, 83(12):2178–2194, 2013.

Andrew M. Raim, Minglei Liu, Nagaraj K. Neerchal, and Jorge G. Morel. On the method of approximate Fisher scoring for finite mixtures of multinomials. *Statistical Methodology*, 18:115–130, 2014.

B. L. S. Prakasa Rao. *Identifiability in stochastic models*. Academic Press, 1992.

C. R. Rao. *Linear statistical inference and its applications*. John Wiley and Sons Inc, 1965.

J. N. K. Rao. *Small Area Estimation*. Wiley-Interscience, 2003.

Sidney Resnick. *A Probability Path*. Birkhäuser, 1999.

Thomas. J. Rothenberg. Identification in parametric models. *Econometrica*, 39:577–591, 1971.

M.J. Rufo, C.J. Pérez, and J. Martín. Bayesian analysis of finite mixtures of multinomial and negative-multinomial distributions. *Computational Statistics & Data Analysis*, 51 (11):5452–5466, 2007.

SAS Institute Inc. *The FMM Procedure*. SAS Publishing, 2011.

Franklin E. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6):110–114, 1946.

Jun Shao. *Mathematical Statistics*. Springer, 2nd edition, 2008.

T. Sofuni, T. Honda, M. Itoh, S. Neriishi, and M. Otake. Relationship between the radiation dose and chromosome aberrations in atomic bomb survivors of Hiroshima and Nagasaki. *Journal of Radiation Research*, 19(2):126–140, 1978.

Santosh C. Sutradhar, Nagaraj K. Neerchal, and Jorge G. Morel. A goodness-of-fit test for overdispersed binomial (or multinomial) models. *Journal of Statistical Planning and Inference*, 138(5):1459–1471, 2008.

Henry Teicher. On the mixture of distributions. *The Annals of Mathematical Statistics*, 31(1):55–73, 1960.

Henry Teicher. Identifiability of mixtures. *The Annals of Mathematical Statistics*, 32(1): 244–248, 1961.

D. M. Titterington. Recursive parameter estimation using incomplete data. *Journal of the Royal Statistical Society. Series B*, 46:257–267, 1984.

D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, 1985.

Wilson Toussile and Elisabeth Gassiat. Variable selection in model-based clustering using multilocus genotype data. *Advances in Data Analysis and Classification*, 3:109–134, 2009.

H. Zhou and K. Lange. MM algorithms for some discrete multivariate distributions. *Journal of Computational and Graphical Statistics*, 19(3):645–665, 2010.