An Approximate Fisher Scoring Algorithm for Finite Mixtures of Multinomials

Andrew M. Raim, Minglei Liu, Nagaraj K. Neerchal and Jorge G. Morel

Abstract

Finite mixture distributions arise naturally in many applications including clustering and classification. Since they usually do not yield closed forms for maximum likelihood estimates (MLEs), numerical methods using the well known Fisher Scoring or Expectation-Maximization algorithms are considered. In this work, an approximation to the Fisher Information Matrix of an arbitrary mixture of multinomial distributions is introduced. This leads to an Approximate Fisher Scoring algorithm (AFSA), which turns out to be closely related to Expectation-Maximization, and is more robust to the choice of initial value than Fisher Scoring iterations. A combination of AFSA and the classical Fisher Scoring iterations provides the best of both computational efficiency and stable convergence properties.

Key Words: Multinomial; Finite mixture, Maximum likelihood, Fisher Information Matrix, Fisher Scoring.

1 Introduction

A finite mixture model arises when observations are believed to originate from one of several populations, but it is unknown to which population each observation belongs. Model identifiability of mixtures is an important issue with a considerable literature; see, for example, Robbins (1948) and Teicher (1960). The book by McLachlan and Peel (2000) is a good entry point to the literature on mixtures. A majority of the mixture literature deals with mixtures of normal distributions; however, Blischke (1962; 1964), Krolikowska (1976), and Kabir (1968) are a few early works which address mixtures of discrete distributions. The present work focuses on mixtures of multinomials, which have wide applications in clustering and classification problems, as well as modeling overdispersion data (Morel and Nagaraj, 2012).

It is well known that computation of maximum likelihood estimates (MLE) under mixture distributions is often analytically intractable, and therefore iterative numerical methods are needed. Classical iterative techniques such as Newton-Raphson and Fisher Scoring are two widely used methods. The more recent Expectation-Maximization (EM) algorithm discussed in Dempster, Laird and Rubin (1977) has become another standard technique to compute MLEs. EM is a framework for performing estimation in missing data problems. The idea

Andrew Raim is graduate student at Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250, USA (Email: araim1@umbc.edu). Minglei Liu is Senior Principial Biostatistician at Medtronic, Santa Rosa, CA, USA. Nagaraj Neerchal is Professor at Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250, USA (Email: nagaraj@umbc.edu). Jorge Morel is Principal Statistician at Procter & Gamble, Cincinnati, OH, USA.

is to solve a difficult incomplete data problem by repeatedly solving tractable complete-data problems. If the unknown population labels are treated as missing information, then estimation under the mixture distribution can be considered a missing data problem, and an EM algorithm can be used. Unlike Fisher Scoring (FSA), EM does not require computation of an expected Hessian in each iteration, which is a great advantage if this matrix is difficult to compute. Slow speed of convergence has been cited as a disadvantage of EM. Variations and improved versions of the EM algorithm have been widely used for obtaining MLEs for mixtures (Mclachlan and Peel, 2000, chapter 2).

Fisher Scoring iterations require the inverse of the Fisher Information Matrix (FIM). In the mixture setting, computing the FIM involves a complicated expectation which does not have an analytically tractable form. The matrix can be approximated numerically by Monte Carlo simulation for example, but this is computationally expensive, especially when repeated over many iterations. Morel and Nagaraj (1991; 1993) proposed a variant of Fisher Scoring using an approximate FIM in their study of a multinomial model with extra variation. This model, now referred to as the Random Clumped Multinomial (see Example 3.1 for details), is a special case of the finite mixture of multinomials. The approximate FIM was justified asymptotically, and was used to obtain MLEs for the model and to demonstrate their efficiency. In the present paper, we extend the approximate FIM idea to general finite mixtures of multinomials and hence formulate the Approximate Fisher Scoring Algorithm (AFSA) for this family of distributions. By using the approximate FIM in place of the true FIM, we obtain an algorithm which is closely related to EM. Both AFSA and EM have a slower convergence rate than Fisher Scoring once they are in the proximity of a maximum, but both are also much more robust than Fisher Scoring in finding such regions from an arbitrary initial value.

The rest of the paper is organized as follows. In section 2, a large cluster approximation for the Fisher Information Matrix is derived and some of its properties are presented. This approximate information matrix is easily computed and has an immediate application in Fisher Scoring, which is presented in section 3. Simulation studies are presented in section 4, illustrating convergence properties of the approximate information matrix and approximate Fisher Scoring. Concluding remarks are given in section 5.

2 An Approximate Fisher Information Matrix

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

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

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

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

If a random variable $\boldsymbol{X}$ has distribution $f(\boldsymbol{x}; \boldsymbol{p}, m)$, we will write $\boldsymbol{X} \sim \text{Mult}_k(\boldsymbol{p}, m)$. Since $x_k = m - \sum_{j=1}^{k-1} x_j$ and $p_k = 1 - \sum_{j=1}^{k-1} p_j$, the kth category can be considered as redundant information. Following the sampling and overdispersion literature, we will refer to the number of trials m as the "cluster size" of a multinomial observation.

Now suppose there are s multinomial populations

$$\text{Mult}_k(\boldsymbol{p}_1, m), \ldots, \text{Mult}_k(\boldsymbol{p}_s, m), \qquad \boldsymbol{p}_\ell = (p_{\ell 1}, \ldots, p_{\ell,k-1})$$

where the ℓth population occurs with proportion π_ℓ for $\ell = 1, \ldots, s$. If we draw $\boldsymbol{X}$ from the mixed population, its probability density is a finite mixture of multinomials

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{\ell=1}^{s} \pi_\ell f(\boldsymbol{x}; \boldsymbol{p}_\ell, m), \qquad \boldsymbol{\theta} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_s, \boldsymbol{\pi}), \tag{2.1}$$

and we will write $\boldsymbol{X} \sim \text{MultMix}_k(\boldsymbol{\theta}, m)$. The dimension of $\boldsymbol{\theta}$ is

$$q := s(k-1) + (s-1) = sk - 1,$$

disregarding the redundant parameters $p_{1k}, \ldots, p_{sk}, \pi_s$. We will also make use of the following slightly-less-cumbersome notation for densities,

$$\begin{aligned} \text{P}(\boldsymbol{x}) &:= f(\boldsymbol{x}; \boldsymbol{\theta}, m): \quad \text{the mixture,} \\ \text{P}_\ell(\boldsymbol{x}) &:= f(\boldsymbol{x}; \boldsymbol{p}_\ell, m): \quad \text{the } \ell\text{th component of the mixture.} \end{aligned}$$

The setting of this paper will be an independent sample

$$\boldsymbol{X}_i \sim \text{MultMix}_k(\boldsymbol{\theta}, m_i), \qquad i = 1, \ldots, n$$

with cluster sizes not necessarily equal. The resulting likelihood is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{x}_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \sum_{\ell=1}^{s} \pi_\ell \left[\frac{m_i!}{x_{i1}! \ldots x_{ik}!} p_{\ell 1}^{x_{i1}} \ldots p_{\ell k}^{x_{ik}} \cdot I(\boldsymbol{x}_i \in \Omega) \right] \right\}. \tag{2.2}$$

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 Appendix A.

In general, as mentioned earlier, the Fisher Information Matrix (FIM) 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

$$\mathcal{I}(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}\in\Omega} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x};\boldsymbol{\theta}) \right\} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x};\boldsymbol{\theta}) \right\}^T f(\boldsymbol{x};\boldsymbol{\theta}), \tag{2.3}$$

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 intractible as m and k increase. For example, when $m = 100$ and $k = 4$, the sample space Ω contains more than 178,000 elements. To avoid these potentially expensive computations, we extend the approximate FIM approach of Morel and Nagaraj (1991; 1993) to the general finite mixture of multinomials. The following theorem states our main result.

Theorem 2.1. *Suppose $\boldsymbol{X} \sim MultMix_k(\boldsymbol{\theta}, m)$ 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,

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

is a $(s-1)\times(s-1)$ matrix, 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 $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_s$ are distinct (i.e. $\boldsymbol{p}_a \neq \boldsymbol{p}_b$ for every pair of populations $a \neq b$), then $\mathcal{I}_m(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m(\boldsymbol{\theta})$ as $m \to \infty$.

A proof is given in Appendix B. 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 $\boldsymbol{\pi}$; see Appendix A for details. The approximate FIM turns out to be equivalent to a complete data FIM, as shown in Proposition 2.2 below, which provides an interesting connection to EM. This matrix can be formulated for any finite mixture whose components have a well-defined FIM, and is not limited to the case of multinomials.

Proposition 2.2. *The matrix $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ is equivalent to the FIM of $(\boldsymbol{X}, Z)$, where*

$$Z = \begin{cases} 1 & \textit{with probability } \pi_1 \\ \vdots & \\ s & \textit{with probability } \pi_s, \end{cases} \quad \textit{and} \quad (\boldsymbol{X} \mid Z = \ell) \sim Mult_k(\boldsymbol{p}_\ell, m). \tag{2.4}$$

Proof of Proposition 2.2. 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

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

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

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

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

Corollary 2.3. *Suppose $\boldsymbol{X}_i \sim$ MultMix$(\boldsymbol{\theta}, m_i)$, $i = 1, \ldots, n$, 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 $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is given by*

$$\begin{aligned}
\widetilde{\mathcal{I}}(\boldsymbol{\theta}) &= \text{Blockdiag}\left(\pi_1\boldsymbol{F}_1, \ldots, \pi_s\boldsymbol{F}_s, \boldsymbol{F}_\pi\right), \\
\boldsymbol{F}_\ell &= M\left[\boldsymbol{D}_\ell^{-1} + p_{\ell k}^{-1}\mathbf{1}\mathbf{1}^T\right], \qquad \ell = 1, \ldots, s \\
\boldsymbol{F}_\pi &= n\left[\boldsymbol{D}_\pi^{-1} + \pi_s^{-1}\mathbf{1}\mathbf{1}^T\right].
\end{aligned}$$

Proof of Corollary 2.3. Let $\widetilde{\mathcal{I}}_i(\boldsymbol{\theta})$ represent the approximate FIM with respect to observation $\boldsymbol{X}_i$. The result is obtained using

$$\widetilde{\mathcal{I}}(\boldsymbol{\theta}) = \widetilde{\mathcal{I}}_1(\boldsymbol{\theta}) + \cdots + \widetilde{\mathcal{I}}_n(\boldsymbol{\theta}),$$

corresponding to the additive property of exact FIMs for independent samples. The additive property can be justified by noting that each $\widetilde{\mathcal{I}}_i(\boldsymbol{\theta})$ is a true (complete data) FIM, by Proposition 2.2. □

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

Corollary 2.4. *Let $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ represent the FIM with respect to an independent sample $\boldsymbol{X}_i \sim$ MultMix$(\boldsymbol{\theta}, m_i)$, $i = 1, \ldots, n$. Then:*

(a) The inverse of $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ is given by

$$\begin{aligned} \widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta}) &= \text{Blockdiag}\left(\pi_1^{-1}\boldsymbol{F}_1^{-1}, \ldots, \pi_s^{-1}\boldsymbol{F}_s^{-1}, \boldsymbol{F}_\pi^{-1}\right), \\ \boldsymbol{F}_\ell^{-1} &= M^{-1}\{\boldsymbol{D}_\ell - \boldsymbol{p}_\ell \boldsymbol{p}_\ell^T\}, \quad \ell = 1, \ldots, s \\ \boldsymbol{F}_\pi^{-1} &= n^{-1}\{\boldsymbol{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}^T\}. \end{aligned} \tag{2.5}$$

(b) The trace of $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ is given by

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

(c) The determinant of $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ is given by

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

Proof of Corollary 2.4 (a). Since $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ is block diagonal, its inverse can be obtained by inverting the blocks, which can immediately be seen to be (2.5). 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 2.3. □

Proof of Corollary 2.4 (b). Since the trace of a block diagonal matrix is the sum of the traces of its blocks, we have

$$\text{tr}\left(\widetilde{\mathcal{I}}(\boldsymbol{\theta})\right) = \pi_1 \,\text{tr}\left(\boldsymbol{F}_1\right) + \cdots + \pi_s \,\text{tr}\left(\boldsymbol{F}_s\right) + \text{tr}\left(\boldsymbol{F}_\pi\right). \tag{2.6}$$

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 $\boldsymbol{\pi}$,

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

The result is obtained by replacing these expressions into (2.6). □

Proof of Corollary 2.4 (c). Since $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ has a block diagonal structure,

$$\begin{aligned}\det \widetilde{\mathcal{I}}(\boldsymbol{\theta}) &= \det\{\boldsymbol{F}_\pi\} \times \prod_{\ell=1}^{s} \det\{\pi_\ell \boldsymbol{F}_\ell\} \\ &= \left(n^{s-1} \det\left\{\boldsymbol{D}_\pi^{-1} + \pi_s^{-1}\mathbf{1}\mathbf{1}^T\right\}\right)\left(\prod_{\ell=1}^{s} \pi_\ell^{k-1} M^{k-1} \det\left\{\boldsymbol{D}_\ell^{-1} + \boldsymbol{p}_{\ell k}^{-1}\mathbf{1}\mathbf{1}^T\right\}\right) \qquad (2.7)\end{aligned}$$

Recall the property (see for example Rao 1965, chapter 1) that for $\boldsymbol{M}$ non-singular, 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 (2.7). □

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 preferable to have a closed form 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. A proof is given in Appendix B. This result is perhaps more immediately relevant than Theorem 2.1 for our Fisher Scoring application presented in the following section.

Theorem 2.5. *Let $\mathcal{I}_m(\boldsymbol{\theta})$ and $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta})$ be defined as in Theorem 2.1 (namely the FIM and approximate FIM 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 $m \to \infty$.*

In the next section, we use the approximate FIM obtained in Theorem 2.1 to define an approximate Fisher Scoring algorithm and investigate its properties.

3 Approximate Fisher Scoring Algorithm

Consider an independent sample with varying cluster sizes

$$\boldsymbol{X}_i \sim \text{MultMix}_k(\boldsymbol{\theta}, m_i), \qquad i = 1, \ldots, n.$$

Let $\boldsymbol{\theta}^{(0)}$ be an initial guess for $\boldsymbol{\theta}$, and $S(\boldsymbol{\theta})$ be the score vector with respect to the sample (see Appendix A). 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 Fisher Scoring Algorithm is given by computing the iterations

$$\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)} + \mathcal{I}^{-1}(\boldsymbol{\theta}^{(g)})\, S(\boldsymbol{\theta}^{(g)}), \qquad g = 1, 2, \ldots \tag{3.1}$$

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, but such modifications will not be considered here. Note that (3.1) 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 Fisher Scoring Algorithm (AFSA). The expressions for $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ and its inverse are available in closed form, as seen in Corollaries 2.3 and 2.4.

AFSA can be applied to finite mixture of multinomial models which are not explicitly in the form of (2.2). We now give two examples which use AFSA to compute MLEs for such models. The first is the Random Clumped model for overdispersed multinomial data. The second is an arbirtrary mixture of multinomials with links from parameters to covariates.

Example 3.1. In section 1 we have mentioned the Random Clumped Multinomial (RCM), a distribution that addresses overdispersion due to "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 (2012) 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 (2.1), and can therefore be fitted by the AFSA algorithm. Once the mixture representation is established, the score vector and approximate FIM can be formulated by the use of transformations; see for example section 2.6 of Lehmann and Casella (1998). 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 $(\pi_1, \ldots, \pi_k)$. 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 $\rho \to 0$, RCM approaches a standard multinomial distribution. Using this idea, an RCM random variable can be obtained from the following procedure. Let $\boldsymbol{Y}_0, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m \overset{\text{iid}}{\sim} \text{Mult}_k(\boldsymbol{\pi}, 1)$ and $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_m \overset{\text{iid}}{\sim} U(0,1)$ be independent samples, then

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{Y}_0 \sum_{i=1}^{m} I(\boldsymbol{U}_i \le \rho) + \sum_{i=1}^{m} \boldsymbol{Y}_i I(\boldsymbol{U}_i > \rho) \\ &= \boldsymbol{Y}_0 N + (\boldsymbol{Z} \mid N) \end{aligned} \tag{3.2}$$

follows the distribution $\text{RCM}_k(\boldsymbol{\pi}, \rho)$. The representation (3.2) 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 approximate FIM 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

$$\boldsymbol{\theta}(\boldsymbol{\eta}) = \begin{pmatrix} (1-\rho)\boldsymbol{\pi} & + & \rho \boldsymbol{e}_1 \\ & \vdots & \\ (1-\rho)\boldsymbol{\pi} & + & \rho \boldsymbol{e}_{k-1} \\ (1-\rho)\boldsymbol{\pi} & & \\ \boldsymbol{\pi} & & \end{pmatrix}.$$

The $q \times k$ Jacobian of this transformation is

$$\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}} = \left(\left(\frac{\partial \theta_i}{\partial \eta_j} \right) \right) = \begin{pmatrix} (1-\rho)\boldsymbol{I}_{k-1} & -\boldsymbol{\pi} + \boldsymbol{e}_1 \\ \vdots & \vdots \\ (1-\rho)\boldsymbol{I}_{k-1} & -\boldsymbol{\pi} + \boldsymbol{e}_{k-1} \\ (1-\rho)\boldsymbol{I}_{k-1} & -\boldsymbol{\pi} \\ \boldsymbol{I}_{k-1} & \boldsymbol{0} \end{pmatrix}.$$

Using the relations

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

it is possible to obtain an explicit form of the approximate FIM as $\widetilde{\mathcal{I}}(\boldsymbol{\eta}) = ((a_{ij}))$, where

$$a_{ij} = \begin{cases} m(1-\rho)^2(\beta_i + \beta_k) - (\pi_i^{-1} + \pi_k^{-1}), & i = j, \quad i, j \in \{1, \ldots, k-1\} \\ m(1-\rho)^2 \beta_k - \pi_k^{-1} & i \neq j, \quad i, j \in \{1, \ldots, k-1\} \\ m(1-\rho)(\gamma_i - \gamma_k), & j = k, \quad i \in \{1, \ldots, k-1\} \\ \frac{m}{(1-\rho)} \sum_{i=1}^{k} \pi_i(1-\pi_i) \left[(1-\rho)\pi_i + \rho\right]^{-1}, & i = k, j = k \end{cases}$$

and

$$\beta_i = \frac{\pi_i}{(1-\rho)\pi_i + \rho} + \frac{1-\pi_i}{(1-\rho)\pi_i}, \qquad \gamma_i = \frac{\pi_i(1-\pi_i)}{(1-\rho)\pi_i + \rho} + \frac{\pi_i}{(1-\rho)}, \qquad i = 1, \ldots, k.$$

It can be shown rigorously that $\widetilde{\mathcal{I}}(\boldsymbol{\eta}) - \mathcal{I}(\boldsymbol{\eta}) \to \mathbf{0}$ as $m \to \infty$, as stated in (Morel and Nagaraj, 1993), and proved in detail in (Morel and Nagaraj, 1991). The proof is similar in spirit to the proof of Theorem 2.1. We then have AFSA iterations for RCM,

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

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.

Example 3.2. In practice there are often covariates to be linked into the model. As an example for how AFSA can be applied, consider the following fixed effect model for response $\boldsymbol{Y} \sim \mathrm{MultMix}_k(\boldsymbol{\theta}(\boldsymbol{x}), m)$ with $d \times 1$ covariates $\boldsymbol{x}$ and $\boldsymbol{z}$. To each $\boldsymbol{p}_\ell$ vector, a generalized logit link will be added

$$\log \frac{p_{\ell j}(\boldsymbol{x})}{p_{\ell k}(\boldsymbol{x})} = \eta_{\ell j}, \quad \eta_{\ell j} = \boldsymbol{x}^T \boldsymbol{\beta}_{\ell j}, \qquad \text{for } \ell = 1, \ldots, s \text{ and } j = 1, \ldots, k-1.$$

A proportional odds model will be assumed for $\boldsymbol{\pi}$,

$$\log \frac{\pi_1(\boldsymbol{z}) + \cdots + \pi_\ell(\boldsymbol{z})}{\pi_{\ell+1}(\boldsymbol{z}) + \cdots + \pi_s(\boldsymbol{z})} = \eta_\ell^\pi, \quad \eta_\ell^\pi = \nu_\ell + \boldsymbol{z}^T \boldsymbol{\alpha}, \qquad \text{for } \ell = 1, \ldots, s-1,$$

taking $\eta_0^\pi := -\infty$ and $\eta_s^\pi := \infty$. The unknown parameters are the $d \times 1$ vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_{\ell j}$, and the scalars ν_ℓ. Denote these parameters collectively as

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_s \\ \boldsymbol{\nu} \\ \boldsymbol{\alpha} \end{pmatrix}, \qquad \text{where } \boldsymbol{\beta}_\ell = \begin{pmatrix} \boldsymbol{\beta}_{\ell 1} \\ \vdots \\ \boldsymbol{\beta}_{\ell, k-1} \end{pmatrix} \text{ and } \boldsymbol{\nu} = \begin{pmatrix} \boldsymbol{\nu}_1 \\ \vdots \\ \boldsymbol{\nu}_s \end{pmatrix}.$$

Expressions for the $\boldsymbol{\theta}$ parameters can be obtained as

$$p_{\ell j}(\boldsymbol{x}) = \frac{e^{\eta_{\ell j}}}{1 + \sum_{b=1}^{k-1} e^{\eta_{\ell b}}} \qquad \text{for } \ell = 1, \ldots, s \text{ and } j = 1, \ldots, k-1,$$
$$\pi_\ell(\boldsymbol{z}) = \frac{e^{\eta_\ell^\pi}}{1 + e^{\eta_\ell^\pi}} - \frac{e^{\eta_{\ell-1}^\pi}}{1 + e^{\eta_{\ell-1}^\pi}} \qquad \text{for } \ell = 1, \ldots, s.$$

To implement AFSA, a score vector and approximate FIM are needed. For the score vector we have

$$S(\boldsymbol{B}) = \frac{\partial}{\partial \boldsymbol{B}} \log f(\boldsymbol{y}; \boldsymbol{\theta}) = \left(\frac{\partial \boldsymbol{N}}{\partial \boldsymbol{B}}\right)^T \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{N}}\right)^T \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{y}; \boldsymbol{\theta})$$

where $\boldsymbol{N} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_s, \boldsymbol{\eta}_\pi)$, $\boldsymbol{\eta}_\ell = (\eta_{\ell 1}, \ldots, \eta_{\ell, k-1})$, and $\boldsymbol{\eta}_\pi = (\eta_1^\pi, \ldots, \eta_{s-1}^\pi)$. For the FIM we have

$$\mathcal{I}(\boldsymbol{B}) = \operatorname{Var}\left(S(\boldsymbol{B})\right) = \left(\frac{\partial \boldsymbol{N}}{\partial \boldsymbol{B}}\right)^T \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{N}}\right)^T \mathcal{I}(\boldsymbol{\theta}) \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{N}}\right) \left(\frac{\partial \boldsymbol{N}}{\partial \boldsymbol{B}}\right).$$

Finding expressions for the two Jacobians is tedious but straightforward.

Propositions 3.3 and 3.4 and Theorem 3.5 state consequences of the main approximation result, which have significant implications on the computation of MLEs. We have already seen that the approximate FIM is equivalent to a complete data FIM from EM. There is also an interesting connection between AFSA and EM, in that the iterations are algebraically related. To see this connection, explicit forms for AFSA and EM iterations are first presented, with proofs given in Appendix B.

Proposition 3.3 (AFSA Iterations)**.** *The AFSA iterations*

$$\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)} + \widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta}^{(g)})\, S(\boldsymbol{\theta}^{(g)}), \qquad g = 1, 2, \ldots \tag{3.3}$$

can be written explicitly as

$$\pi_\ell^{(g+1)} = \pi_\ell^{(g)} \frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}, \quad \ell = 1, \ldots, s$$
$$p_{\ell j}^{(g+1)} = \frac{1}{M} \sum_{i=1}^{n} \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} x_{ij} - p_{\ell j}^{(g)} \left[1 - \frac{1}{M} \sum_{i=1}^{n} m_i \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}\right], \quad \ell = 1, \ldots, s, \;\; j = 1, \ldots, k.$$

where $M = m_1 + \cdots + m_n$.

Proposition 3.4 (EM Iterations)**.** *Consider the complete data*

$$Z_i = \begin{cases} 1 & \textit{with probability } \pi_1 \\ \vdots & \\ s & \textit{with probability } \pi_s, \end{cases} \qquad \textit{and} \qquad (\boldsymbol{X}_i \mid Z_i = \ell) \;\sim \textit{Mult}_k(\boldsymbol{p}_\ell, m_i),$$

where $(\boldsymbol{X}_i, Z_i)$ *are independent for* $i = 1, \ldots, n$. *Denote* $\gamma_{i\ell}^{(g)} := \mathrm{P}(Z_i = \ell \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(g)})$ *as the posterior probability that the* i*th observation belongs to the* ℓ*th group. Iterations for an EM algorithm are given by*

$$\pi_\ell^{(g+1)} = \frac{1}{n}\sum_{i=1}^{n}\gamma_{i\ell}^{(g)} = \frac{1}{n}\pi_\ell^{(g)}\sum_{i=1}^{n}\frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}, \quad \ell = 1, \ldots, s,$$

$$p_{\ell j}^{(g+1)} = \frac{\sum_{i=1}^{n} x_{ij}\gamma_{i\ell}^{(g)}}{\sum_{i=1}^{n} m_i\gamma_{i\ell}^{(g)}} = \frac{\sum_{i=1}^{n} x_{ij}\frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}}{\sum_{i=1}^{n} m_i\frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}}, \quad \ell = 1, \ldots, s, \;\; j = 1, \ldots, k.$$

The iterations for AFSA or EM are repeated for $g = 1, 2, \ldots$, with a given initial guess $\boldsymbol{\theta}^{(0)}$, until

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

where $\varepsilon > 0$ is a given tolerance, which is taken to be the stopping criteria for the remainder of this paper.

Theorem 3.5. *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* g*th iteration with* $\boldsymbol{\theta}^{(g)}$*, then for the* $(g+1)$*th iteration,*

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

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

Proof of Theorem 3.5. It is immediate from Propositions 3.3 and 3.4 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,

$$\begin{aligned}
&\left(\frac{\hat{\pi}_\ell^{(g+1)}}{\pi_\ell^{(g)}}\right)\hat{p}_{\ell j}^{(g+1)} + \left(1 - \frac{\hat{\pi}_\ell^{(g+1)}}{\pi_\ell^{(g)}}\right)p_{\ell j}^{(g)} \\
&\quad= \frac{\sum_{i=1}^{n} x_{ij}\frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}}{m\sum_{i=1}^{n}\frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}}\frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} + p_{\ell j}^{(g)}\left(1 - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}\right) \\
&\quad= \frac{1}{mn}\sum_{i=1}^{n}\frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}x_{ij} + p_{\ell j}^{(g)}\left(1 - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}\right) \\
&\quad= \tilde{p}_{\ell j}^{(g+1)}. \qquad (3.4)
\end{aligned}$$

□

The $(g+1)$th AFSA iterate 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 non-negative 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 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 gth step the EM algorithm is close to convergence. Then

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

From (3.4) we will also have

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

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 $m_1, \ldots, m_n$,

$$\begin{aligned} &\frac{\hat{\pi}_\ell^{(g+1)}}{\pi_\ell^{(g)}} \hat{p}_{\ell j}^{(g+1)} + \left(1 - \frac{\hat{\pi}_\ell^{(g+1)}}{\pi_\ell^{(g)}}\right) p_{\ell j}^{(g)} \\ &\qquad = \frac{\sum_{i=1}^n x_{ij} \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}}{\sum_{i=1}^n m_i \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}} \frac{1}{n} \sum_{i=1}^n \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} + p_{\ell j}^{(g)} \left(1 - \frac{1}{n} \sum_{i=1}^n \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)}\right), \end{aligned} \tag{3.5}$$

which does not simplify to $\tilde{p}_{\ell j}^{(g+1)}$ as in the proof of Theorem 3.5. 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 (3.5) so that the algebraic relationship to EM would be maintained as in (3.4) for the balanced case.

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

$$\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)} + \mathcal{I}_c^{-1}(\boldsymbol{\theta}^{(g)})\, S(\boldsymbol{\theta}^{(g)}), \qquad g = 1, 2, \ldots, \tag{3.6}$$

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 a regular exponential family

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

and $\boldsymbol{\mu} := \mathrm{E}(\boldsymbol{t}(\boldsymbol{X}))$ is the parameter of interest. The complete data likelihood for our multinomial mixture is indeed a regular exponential family, but the parameter of interest $\boldsymbol{\theta}$ is a transformation of $\boldsymbol{\mu}$ rather than $\boldsymbol{\mu}$ itself. Therefore the equivalance is approximate, as we have seen in Theorem 3.5. The justification for AFSA leading to this paper 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 (3.6) suggests that AFSA is a reasonable approach for finite mixtures beyond the multinomial setting,

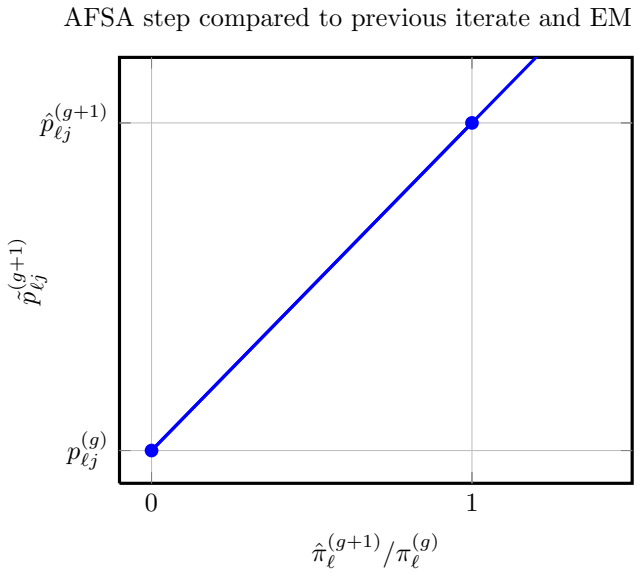


Figure 1: The next AFSA $\tilde{p}_{\ell j}^{(g+1)}$ iteration is a linear combination of $\hat{p}_{\ell j}^{(g+1)}$ and $p_{\ell j}^{(g)}$, which depends on the ratio $\hat{\pi}_\ell^{(g+1)}/\pi_\ell^{(g)}$.

4 Simulation Studies

The main result stated in Theorem 2.1 allows us to approximate the matrix $\mathcal{I}(\boldsymbol{\theta})$ by $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$, which is much more easily computed. Theorem 2.5 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.

4.1 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 $\|\boldsymbol{A}\|_F = \sqrt{\sum_i \sum_j a_{ij}^2}$, 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 dis-

tance between their inverses. Consider a mixture $\text{MultMix}_2(\boldsymbol{\theta}, m)$ of three binomials, with parameters

$$\boldsymbol{p} = \begin{pmatrix} 1/7 & 1/3 & 2/3 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\pi} = \begin{pmatrix} 1/6 & 2/6 & 3/6 \end{pmatrix}.$$

Figure 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 orders of magnitude. To see more concretely what is being compared, for the moderate cluster size $m = 20$ we have, respectively for the approximate and exact FIMs,

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

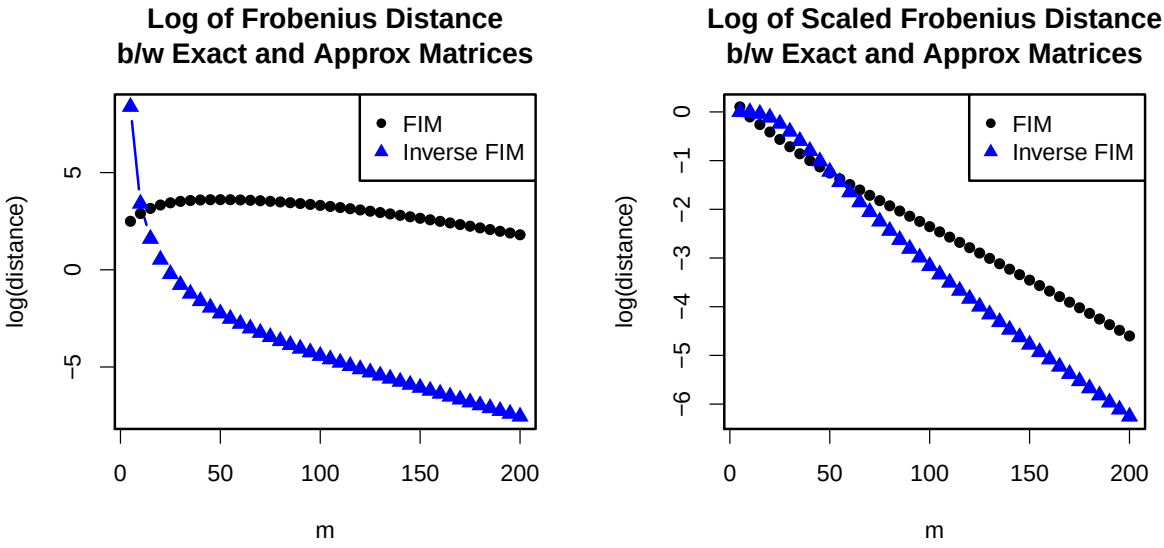
and for the approximate and exact inverse FIMs,

$$\begin{pmatrix} 0.037 & 0 & 0 & 0 & 0 \\ 0 & 0.033 & 0 & 0 & 0 \\ 0 & 0 & 0.022 & 0 & 0 \\ 0 & 0 & 0 & 0.139 & -0.056 \\ 0 & 0 & 0 & -0.056 & 0.222 \end{pmatrix} \text{ vs. } \begin{pmatrix} 0.216 & 0.160 & 0.020 & 0.366 & -0.295 \\ 0.160 & 0.251 & 0.043 & 0.383 & -0.240 \\ 0.020 & 0.043 & 0.040 & 0.053 & -0.003 \\ 0.366 & 0.383 & 0.053 & 0.953 & -0.690 \\ -0.295 & -0.240 & -0.003 & -0.690 & 0.827 \end{pmatrix}.$$

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 approximate FIM, which is that it cannot be used to estimate all the 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 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 Thereom 2.1 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 Appendix A). 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



(a) Using unscaled distance

(b) Using scaled distance

Figure 2: Distance between exact and approximate FIM and its inverse, as m is varied.

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 to be converging more quickly than the distance between the FIM and approximate FIM.

4.2 Effectiveness of AFSA method

4.2.1 Convergence Speed

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

$$\begin{aligned} &\boldsymbol{Y}_i \overset{\text{iid}}{\sim} \text{MultMix}_3(\boldsymbol{\theta}, m = 20), \qquad i = 1, \ldots, n = 500 \\ &\boldsymbol{p}_1 = \begin{pmatrix} 1/3 & 1/3 & 1/3 \end{pmatrix}, \qquad \boldsymbol{p}_2 = \begin{pmatrix} 0.1 & 0.3 & 0.6 \end{pmatrix}, \qquad \pi = 0.75. \end{aligned}$$

We fit the MLE using AFSA, FSA, and EM. After the gth iteration, 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 3. 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 1, and additional summary information is shown in Table 2.

We see that AFSA and EM have almost exactly the same rate of convergence toward the same solution, as suggested by Thereom 3.5. 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 considered an FSA hybrid with a "warmup period", where for a given $\varepsilon_0 > 0$ the approximate FIM is used until the first time $\delta^{(g)} < \varepsilon_0$ is crossed. Notice that $\varepsilon_0 = \infty$ corresponds to "no warmup period". 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 3.1. The first stage consisted of running AFSA iterations until convergence, and in the second stage one additional iteration of exact Fisher 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 approximate FIM (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 converge to the solution. This approach depends on the exact FIM being feasible to compute, so the sample space cannot be too large. For the present simulations, we make use of the naive summation (2.3). Hence, there is a trade-off in the choice of ε_0 between energy spent on computing the exact FIM and a larger number of iterations required for AFSA. Figure 3 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 shows that even a very limited warmup period such as $\varepsilon_0 = 10$ can be sufficient.

The Newton-Raphson algorithm, which has not been shown here, performed similarly to Fisher Scoring but has issues with singularity in some samples. 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 3.5, standard errors from $\widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta})$ evaluated at EM estimates could also be used to obtain similar results to AFSA.

4.2.2 Monte Carlo Study

We next consider a Monte Carlo study of the difference between AFSA and EM estimators. Observations were generated from

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

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

$$Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta), \qquad m_i = \lceil Z_i \rceil. \tag{4.1}$$

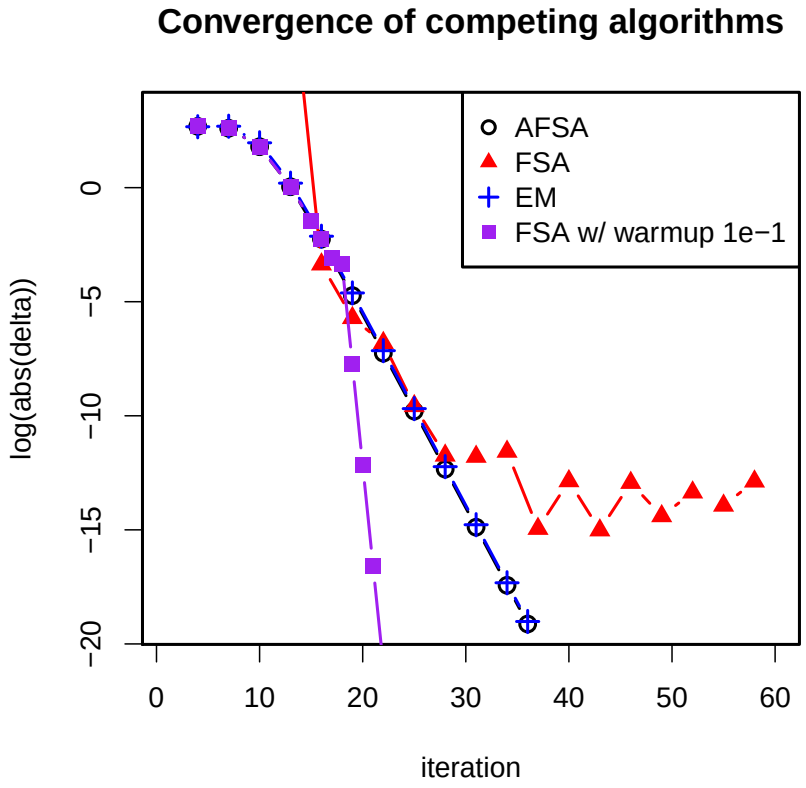


Figure 3: Convergence of several competing algorithms for a small test problem

Table 1: Estimates and standard errors for the competing algorithms. FSA Hybrid produced the same results with ε_0 set to $0.001, 0.01, 0.1, 1$, and 10.

	FSA	AFSA	EM	FSA Hybrid
$\hat{p}_{11}$	0.2690	0.3315	0.3315	0.3315
SE	0.0044	0.0055	—	0.0065
$\hat{p}_{12}$	0.3341	0.3403	0.3403	0.3403
SE	0.0047	0.0055	—	0.0057
$\hat{p}_{21}$	0.3341	0.0928	0.0928	0.0928
SE	0.0908	0.0057	—	0.0081
$\hat{p}_{22}$	0.0002	0.3167	0.3167	0.3167
SE	0.0055	0.0091	—	0.0102
$\hat{\pi}$	0.9990	0.7381	0.7381	0.7381
SE	0.0014	0.0197	—	0.0253

Table 2: Convergence of several competing algorithms. Hybrid FSA is shown with several choices of the warmup tolerance ε_0. Exact FSA uses $\varepsilon_0 = \infty$.

method	ε_0	logLik	tol	iter
AFSA	—	-2234.655	4.94×10^{-09}	36
EM	—	-2234.655	5.50×10^{-09}	36
FSA	∞	-2423.864	-1.26×10^{-07}	100
FSA	10	-2234.655	4.46×10^{-10}	16
FSA	1	-2234.655	1.34×10^{-10}	20
FSA	0.1	-2234.655	7.84×10^{-10}	22
FSA	0.01	-2234.655	1.42×10^{-10}	24
FSA	0.001	-2234.655	9.05×10^{-10}	26

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 $\mathrm{E}(Z_i) = \alpha\beta = 20$. This gives $\beta = 20/\alpha$ so only α is free, and $\mathrm{Var}(Z_i) = \alpha\beta^2 = 400/\alpha$ 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 used for both algorithms in every repetition. To measure the closeness of the two estimators, a maximum relative difference is taken over all components of $\boldsymbol{\theta}$, then averaged over all repetitions:

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

Here $\bigvee$ represents the "maximum" operator. Notice that obtaining a good result for $\overline{D}$ 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 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 $\alpha = 25$).

Both algorithms are susceptible to finding local maxima of the likelihood, 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 an especially bad situation for our Monte Carlo statistic $\overline{D}$, which can become very large if

Table 3: Closeness between AFSA and EM estimates, over 1000 trials

	$\boldsymbol{p}_1$	$\boldsymbol{p}_2$	Cluster sizes equal $m_i = 20$	$\alpha = 100$ $\text{Var}(m_i) \approx 4.083$	$\alpha = 25$ $\text{Var}(m_i) \approx 16.083$
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}

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 trials required a restart, and no more than two restarts were needed for the same trial. 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 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 $\overline{D}$ tends to become larger. Theorem 2.1 depends on the distinctness of the category probability vectors, so the quality of the FIM approximation at moderate cluster size 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. Recall that the dimension of $\boldsymbol{p}_i$ is $k - 1$; it can be seen from Table 3 that increasing k from 2 to 4 does not necessarily have a negative effect on the results.

In Scenario E, $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ are not too close together, yet $\overline{D}$ has a similar magnitude to Scenario D where the two vectors are closer. Figure 4 shows a plot of the individual D_r for Scenarios D and E. Notice that in Scenario E, one particular simulation in each case is responsible for the large magnitude of $\overline{D}$. Upon removal of these simulations, the order of $\overline{D}$ is reduced from 10^{-3} to 10^{-6}. However, many large D_r were present in the Scenario D results.

5 Conclusions

A large cluster approximation was presented for the FIM of the finite mixture of multinonials model (Theorem 2.1). This matrix has a convenient block diagonal form, where each non-zero block is the FIM of a standard multinomial observation. Furthermore, the approximation is equivalent to the "complete data" FIM, had population labels been recorded for each

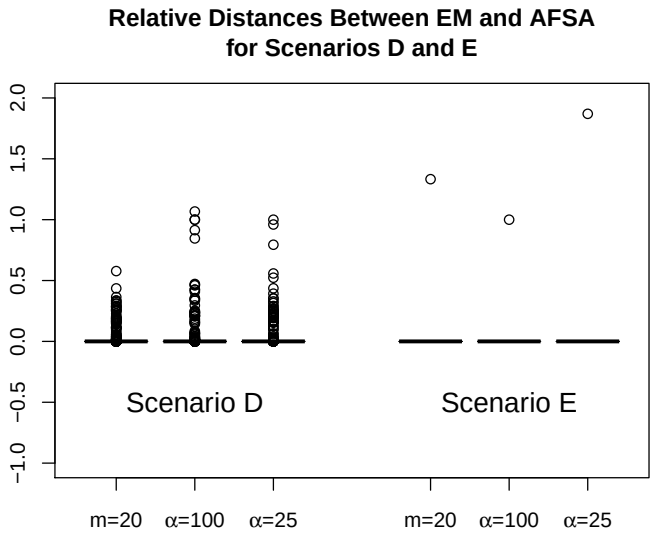


Figure 4: Boxplots for Scenarios D and E of Monte Carlo study. At this scale, the boxes appear as thin horizontal lines.

observation (Proposition 2.2). Using this approximation to the FIM, we formulated the Approximate Fisher Scoring Algorithm (AFSA), and showed that its iterations are closely related to the well known Expectation-Maximization (EM) algorithm for finite mixtures (Theorem 3.5). Simulations show that a rather large cluster size is needed before the exact and approximate FIM are close; this is not surprising given that a block diagonal matrix is being used to approximate a dense matrix. A large cluster size is also needed for a close approximation of the inverse, although the inverses are seen to converge together more quickly. Therefore, the approximate FIM and its inverse are not well-suited to replace the exact matrices for general usage. This means, for example, that one should be cautious about computing standard errors for the MLE from the approximate inverse FIM.

As another example of a general use for the approximate FIM, consider approximate $(1-\alpha)$ level Wald-type and Score-type confidence regions,

$$\left\{\boldsymbol{\theta}_0 : (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \, \widetilde{\mathcal{I}}(\hat{\boldsymbol{\theta}}) \, (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \leq \chi^2_{q,\alpha}\right\} \quad \text{and} \quad \left\{\boldsymbol{\theta}_0 : S(\boldsymbol{\theta}_0)^T \, \widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta}_0) \, S(\boldsymbol{\theta}_0) \leq \chi^2_{q,\alpha}\right\}, \quad (5.1)$$

respectively, using the approximate FIM in place of the exact FIM. Such regions are very practical to compute, but will likely not have the desired coverage for $\boldsymbol{\theta}$. However, we might expect the Score region to perform better for moderate cluster sizes because it involves the inverse matrix. On the other hand, the approximate FIM works well as a tool for estimation in the AFSA algorithm. This is interesting because the more standard Fisher Scoring and Newton-Raphson algorithms do not work well on their own. For Newton-Raphson, the invertibility of the Hessian depends on the sample as well as the current iterate $\boldsymbol{\theta}^{(g)}$ and the model. Fisher Scoring can be computed when the cluster size is not too small (so that

the FIM is non-singular), but it is often unable to make progress at all from an arbitrarily chosen starting point. In this case, AFSA (or EM) is useful for giving FSA some initial help. If FSA has a sufficiently good starting point, it can converge very quickly. Therefore we recommend a hybrid approach: use AFSA iterations for an initial warmup period, then switch to FSA once a path toward a solution has been established. This approach may also help to reduce the number of exact FIM computations needed, which may be expensive. Although AFSA and EM are closely related and often tend toward the same solution, AFSA is not restricted to the parameter space. Additional precautions may therefore be needed to prevent AFSA iterations from drifting outside of the space. AFSA also tended to converge to the boundary of the space more often than EM; hence, we reiterate the usual advice of trying several initial values as a good practice. 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 is not required for AFSA. A trade-off for AFSA is that the score vector for the observed data must be computed; this may involve a messy differentiation, but is arguably easier to address numerically than the E-step. AFSA iterations were obtained for the Random-Clumped Multinomial in Example 3.1, starting from a general multinomial mixture and using an appropriate transformation of the parameters.

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. In this light, EM and AFSA iterations are seen to be approximately equivalent.

Several interesting questions can be raised at this point. There is a relationship between AFSA and EM which extends beyond the multinomial mixture; we wonder if the relationship between the exact and complete data information matrix generalizes as well. Also, for the present multinomial mixture, perhaps there is a small cluster bias correction that could be applied to improve the approximation. This might allow standard errors and confidence regions such as (5.1) to safely be derived from the approximate FIM.

6 Acknowledgements

The hardware used in the computational studies is part of the UMBC High Performance Computing Facility (HPCF). The facility is supported by the U.S. National Science Foundation through the MRI program (grant no. CNS–0821258) and the SCREMS program (grant no. DMS–0821311), with additional substantial support from the University of Maryland, Baltimore County (UMBC). See www.umbc.edu/hpcf for more information on HPCF and the projects using its resources. The first author additionally acknowledges financial support as HPCF RA.

References

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.

H. Bozdogan. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

S. Chandra. On the mixtures of probability distributions. *Scandinavian Journal of Statistics*, 4:105–112, 1977.

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.

A. B. Kabir. Estimation of parameters of a finite mixture of distributions. *Journal of the Royal Statistical Society. Series B*, 30:472–482, 1968.

K. Krolikowska. Estimation of the parameters of any finite mixture of geometric distributions. *Demonstratio Mathematica*, 9:573–582, 1976.

K. Lange. *Numerical Analysis for Statisticians*. Springer, 2nd edition, 2010.

E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, 2nd edition, 1998.

T. A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 44:226–233, 1982.

G. Mclachlan and D. Peel. *Finite Mixture Models*. Wiley-Interscience, 2000.

X. L. Meng and D. B. Rubin. Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association*, 86(416):899–909, 1991.

C. D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, 2001.

J. G. Morel and N. K. Nagaraj. A finite mixture distribution for modeling multinomial extra variation. Technical Report Research report 91–03, Department of Mathematics and Statistics, University of Maryland, Baltimore County, 1991.

J. G. Morel and N. K. Nagaraj. A finite mixture distribution for modelling multinomial extra variation. *Biometrika*, 80(2):363–371, 1993.

J. G. Morel and N. K. Nagaraj. *Overdispersion Models in SAS*. SAS Institute, 2012.

N. K. Neerchal and J. G. Morel. Large cluster results for two parametric multinomial extra variation models. *Journal of the American Statistical Association*, 93(443):1078–1087, 1998.

N. K. Neerchal and J. 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.

M. Okamoto. Some inequalities relating to the partial sum of binomial probabilities. *Annals of the Institute of Statistical Mathematics*, 10:29–35, 1959.

A. M. Raim, M. K. Gobbert, N. K. Neerchal, and J. G. Morel. Maximum likelihood estimation of the random-clumped multinomial model as prototype problem for large-scale statistical computing. Accepted, 2012.

C. R. Rao. *Linear statistical inference and its applications*. John Wiley and Sons Inc, 1965.

H. Robbins. Mixture of distributions. *The Annals of Mathematical Statistics*, 19(3):360–369, 1948.

T. J. Rothenberg. Identification in parametric models. *Econometrica*, 39:577–591, 1971.

H. Teicher. On the mixture of distributions. *The Annals of Mathematical Statistics*, 31(1): 55–73, 1960.

D. M. Titterington. Recursive parameter estimation using incomplete data. *Journal of the Royal Statistical Society. Series B*, 46:257–267, 1984.

H. Zhou and K. Lange. MM algorithms for some discrete multivariate distributions. *Journal of Computational and Graphical Statistics*, 19(3):645–665, 2010.

A Appendix: Preliminaries and Notation

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

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

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

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

so that

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

denoting $\boldsymbol{D} := \operatorname{diag}(p_1, \ldots, p_{k-1})$ and $\boldsymbol{x}_{-k} := (x_1, \ldots, x_{k-1})$.

The score vector for a single observation $\boldsymbol{X} \sim \operatorname{MultMix}_k(\boldsymbol{\theta}, m)$ 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 $\boldsymbol{D}_a := \operatorname{diag}(p_{a1}, \ldots, p_{a,k-1})$, 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 $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$

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

The last equality holds under appropriate regularity conditions. For the multinomial FIM, we may use (A.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

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

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

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

implies $s = v$ and terms within the sums are equal, except the indicies may be permuted (Mclachlan and Peel, 2000, section 1.14). Chandra (1977) provides some insight into the identifiability issue, and shows that a family of multivariate mixtures is identifiable if any of the corresponding marginal mixtures are identifiable. In the present case, the multivariate mixtures consist of multinomial densities, and the marginal densities are binomials. It is well known that a finite mixture of s components from

$$\{\ \mathrm{Binomial}(m,\theta) : \theta \in (0,1)\ \}$$

is identifiable if and only if $m \geq 2s-1$; see, for example, Blischke (1964). Then a sufficient condition for model (2.2) to be identifiable is that $m_i \geq 2s-1$ for at least one observation. This can be seen by the following lemma.

Lemma A.1. *Suppose $\boldsymbol{X}_i \overset{ind}{\sim} f_i(\boldsymbol{x};\boldsymbol{\theta}), i = 1,\ldots,n$, where f_i's are densities, 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 of Lemma A.1. WLOG assume that $r = 1$, and suppose we have

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

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

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

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. □

B Appendix: Additional Proofs

To prove Theorem 2.1, we will first establish a key inequality. A similar strategy was used by Morel and Nagaraj (1991), but they considered the special case $k = s$, so that the number of mixture components is equal to the number of categories within each component. Here we generalize their argument to the general case where $k = s$ need not hold. The original proof was inspired by the following inequality from Okamoto (1959) for the tail probability of the binomial distribution, which was also considered by Blischke (1962).

Lemma B.1. *Suppose $X \sim$ Binomial(m, p) and let $f(x; m, p)$ be its density. Then for $c \geq 0$,*

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

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

Theorem B.2. *For a given index $b \in \{1, \ldots, s\}$ we have*

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

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

Proof of Theorem B.2. For $a, b \in \{1, \ldots, s\}$, assume WLOG that

$$\delta_{ab} := \bigvee_{j=1}^{k-1}(p_{aj} - p_{bj}) = (p_{aL} - p_{bL}), \qquad \text{for some } L \in \{1, \ldots k-1\}$$

is positive. 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

$$\begin{aligned}
\sum_{\boldsymbol{x}\in\Omega} \frac{\pi_a \, \mathrm{P}_a(\boldsymbol{x}) \, \mathrm{P}_b(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} &= \sum_{x_L=0}^{m} \sum_{\boldsymbol{x}\in\Omega(x_L)} \frac{\pi_a \, \mathrm{P}_a(\boldsymbol{x}) \, \mathrm{P}_b(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} \\
&= \sum_{x_L \leq \frac{m}{2}(p_{aL}+p_{bL})} \sum_{\boldsymbol{x}\in\Omega(x_L)} \pi_a \, \mathrm{P}_a(\boldsymbol{x}) \frac{\mathrm{P}_b(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} + \sum_{x_L > \frac{m}{2}(p_{aL}+p_{bL})} \sum_{\boldsymbol{x}\in\Omega(x_L)} \frac{\pi_a \, \mathrm{P}_a(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} \, \mathrm{P}_b(\boldsymbol{x}) \\
&\leq \sum_{x_L \leq \frac{m}{2}(p_{aL}+p_{bL})} \sum_{\boldsymbol{x}\in\Omega(x_L)} \frac{\pi_a}{\pi_b} \, \mathrm{P}_a(\boldsymbol{x}) + \sum_{x_L > \frac{m}{2}(p_{aL}+p_{bL})} \sum_{\boldsymbol{x}\in\Omega(x_L)} \mathrm{P}_b(\boldsymbol{x}) \\
&= \frac{\pi_a}{\pi_b} \sum_{x_L \leq \frac{m}{2}(p_{aL}+p_{bL})} \sum_{\boldsymbol{x}\in\Omega(x_L)} \mathrm{P}_a(\boldsymbol{x}) + \sum_{x_L > \frac{m}{2}(p_{aL}+p_{bL})} \sum_{\boldsymbol{x}\in\Omega(x_L)} \mathrm{P}_b(\boldsymbol{x}). \qquad \text{(B.1)}
\end{aligned}$$

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 (B.1) is equal to

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

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

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

and (B.2) is equivalent to

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

Now we have

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

□

Corollary B.3. *The following intermediate result was obtained in the proof of Theorem B.2*

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

We are now prepared to prove Theorem 2.1. 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 2.1. 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

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

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

$$\boldsymbol{C}_{11} - \mathrm{Blockdiag}(\pi_1\boldsymbol{F}_1,\ldots,\pi_s\boldsymbol{F}_s) \to \boldsymbol{0}, \tag{B.3}$$

$$\boldsymbol{C}_{21}^T = \boldsymbol{C}_{12} \to \boldsymbol{0}, \tag{B.4}$$

$$\boldsymbol{C}_{22} - \boldsymbol{F}_\pi \to \boldsymbol{0}. \tag{B.5}$$

We will show (B.3) and (B.5) only; (B.4) is similar. The reader may refer to Morel and Nagaraj (1991) for more details, shown for the $k = s$ case.

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

$$\begin{aligned}
&\boldsymbol{D}_i\left(\boldsymbol{A}_{ii} - \pi_i\boldsymbol{F}_i\right)\boldsymbol{D}_i\\
&\quad= \boldsymbol{D}_i\left\{\mathrm{E}\left[\left\{\frac{\partial}{\partial \boldsymbol{p}_i}\log \mathrm{P}(\boldsymbol{x})\right\}\left\{\frac{\partial}{\partial \boldsymbol{p}_i}\log \mathrm{P}(\boldsymbol{x})\right\}^T\right] - \pi_i\boldsymbol{F}_i\right\}\boldsymbol{D}_i\\
&\quad= \pi_i^2\boldsymbol{D}_i\,\mathrm{E}\left[\frac{\mathrm{P}_i^2(\boldsymbol{x})}{\mathrm{P}^2(\boldsymbol{x})}\frac{\partial \log \mathrm{P}_i(\boldsymbol{x})}{\partial \boldsymbol{p}_i}\frac{\partial \log \mathrm{P}_i(\boldsymbol{x})}{\partial \boldsymbol{p}_i^T}\right]\boldsymbol{D}_i - \pi_i\boldsymbol{D}_i\boldsymbol{F}_i\boldsymbol{D}_i\\
&\quad= \pi_i^2\sum_{\boldsymbol{x}\in\Omega}\frac{\mathrm{P}_i(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})}\left(\boldsymbol{x}_{-k} - \frac{x_k}{p_{ik}}\boldsymbol{p}_i\right)\left(\boldsymbol{x}_{-k} - \frac{x_k}{p_{ik}}\boldsymbol{p}_i\right)^T\mathrm{P}_i(\boldsymbol{x})\\
&\qquad - \pi_i^2\sum_{\boldsymbol{x}\in\Omega}\frac{1}{\pi_i}\left(\boldsymbol{x}_{-k} - \frac{x_k}{p_{ik}}\boldsymbol{p}_i\right)\left(\boldsymbol{x}_{-k} - \frac{x_k}{p_{ik}}\boldsymbol{p}_i\right)^T\mathrm{P}_i(\boldsymbol{x}) \qquad \text{(B.6)}\\
&\quad= \pi_i^2\sum_{\boldsymbol{x}\in\Omega}\left(\boldsymbol{x}_{-k} - \frac{x_k}{p_{ik}}\boldsymbol{p}_i\right)\left(\boldsymbol{x}_{-k} - \frac{x_k}{p_{ik}}\boldsymbol{p}_i\right)^T\left(\frac{\mathrm{P}_i(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} - \frac{1}{\pi_i}\right)\mathrm{P}_i(\boldsymbol{x})\\
&\quad= \frac{\pi_i}{p_{ik}^2}\sum_{\boldsymbol{x}\in\Omega}\left(p_{ik}\boldsymbol{x}_{-k} - x_k\boldsymbol{p}_i\right)\left(p_{ik}\boldsymbol{x}_{-k} - x_k\boldsymbol{p}_i\right)^T\left(\frac{\pi_i\,\mathrm{P}_i(\boldsymbol{x}) - \mathrm{P}(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})}\right)\mathrm{P}_i(\boldsymbol{x}). \qquad \text{(B.7)}
\end{aligned}$$

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 B.2 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 (B.6) that

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

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

$$[p_{ik} x_r - x_k p_{ir}]^2 \leq [x_r + m p_{ir}]^2 \leq 4m^2.$$

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

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

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 (B.7) 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 B.2. 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 $\boldsymbol{C}_{11} - \text{Blockdiag}(\pi_1\boldsymbol{F}_1, \ldots, \pi_s\boldsymbol{F}_s)$ where $i \neq j$.

$$
\begin{aligned}
&\boldsymbol{D}_i\boldsymbol{A}_{ij}\boldsymbol{D}_j \\
&\quad= \boldsymbol{D}_i \left\{ \mathrm{E}\left[\left\{ \frac{\partial}{\partial \boldsymbol{p}_i} \log \mathrm{P}(\boldsymbol{x}) \right\} \left\{ \frac{\partial}{\partial \boldsymbol{p}_j} \log \mathrm{P}(\boldsymbol{x}) \right\}^T \right] \right\} \boldsymbol{D}_j \\
&\quad= \boldsymbol{D}_i \left[\mathrm{E}\left(\frac{\pi_i\pi_j}{\mathrm{P}^2(\boldsymbol{x})} \frac{\partial \mathrm{P}_i(\boldsymbol{x})}{\partial \boldsymbol{p}_i} \frac{\partial \mathrm{P}_j(\boldsymbol{x})}{\partial \boldsymbol{p}_j^T} \right) \right] \boldsymbol{D}_j \\
&\quad= \pi_i\pi_j \boldsymbol{D}_i \left[\mathrm{E}\left(\frac{\mathrm{P}_i(\boldsymbol{x})\,\mathrm{P}_j(\boldsymbol{x})}{\mathrm{P}^2(\boldsymbol{x})} \frac{\partial \log \mathrm{P}_i(\boldsymbol{x})}{\partial \boldsymbol{p}_i} \frac{\partial \log \mathrm{P}_j(\boldsymbol{x})}{\partial \boldsymbol{p}_j^T} \right) \right] \boldsymbol{D}_j \\
&\quad= \pi_i\pi_j \sum_{\boldsymbol{x}\in\Omega} \frac{\mathrm{P}_i(\boldsymbol{x})\,\mathrm{P}_j(\boldsymbol{x})}{\mathrm{P}^2(\boldsymbol{x})} \left(\boldsymbol{x}_{-k} - \frac{x_k}{p_{ik}}\boldsymbol{p}_i \right) \left(\boldsymbol{x}_{-k} - \frac{x_k}{p_{jk}}\boldsymbol{p}_j \right)^T \mathrm{P}(\boldsymbol{x}) \\
&\quad= \frac{\pi_i\pi_j}{p_{ik}p_{jk}} \sum_{\boldsymbol{x}\in\Omega} \frac{\mathrm{P}_i(\boldsymbol{x})\,\mathrm{P}_j(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} \left(p_{ik}\boldsymbol{x}_{-k} - x_k\boldsymbol{p}_i \right) \left(p_{jk}\boldsymbol{x}_{-k} - x_k\boldsymbol{p}_j \right)^T . \qquad \text{(B.8)}
\end{aligned}
$$

If we now denote the (r,s)th element of the matrix given in (B.8) 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 B.3 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

$$
\begin{aligned}
&\boldsymbol{A}_{\pi\pi} - \boldsymbol{F}_\pi \qquad \text{(B.9)} \\
&\quad= \mathrm{E}\left[\left\{ \frac{\partial}{\partial \boldsymbol{\pi}} \log \mathrm{P}(\boldsymbol{x}) \right\} \left\{ \frac{\partial}{\partial \boldsymbol{\pi}} \log \mathrm{P}(\boldsymbol{x}) \right\}^T \right] - \boldsymbol{F}_\pi \\
&\quad= \mathrm{E}\left[\frac{1}{\mathrm{P}^2(\boldsymbol{x})} \left\{ \begin{pmatrix} \mathrm{P}_1(\boldsymbol{x}) \\ \vdots \\ \mathrm{P}_{s-1}(\boldsymbol{x}) \end{pmatrix} - \mathrm{P}_s(\boldsymbol{x}) \cdot \boldsymbol{1} \right\} \left\{ \begin{pmatrix} \mathrm{P}_1(\boldsymbol{x}) \\ \vdots \\ \mathrm{P}_{s-1}(\boldsymbol{x}) \end{pmatrix} - \mathrm{P}_s(\boldsymbol{x}) \cdot \boldsymbol{1} \right\}^T \right] - \left(\boldsymbol{D}_\pi^{-1} + \pi_s^{-1}\boldsymbol{1}\boldsymbol{1}^T \right) .
\end{aligned}
$$

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,

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

applying Theorem B.2 to the first two terms, and Corollary B.3 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 (B.9), but now the case where $a \neq b$. We have

$$\begin{aligned}
\varepsilon_{ab} &= \mathrm{E}\left[\frac{[\mathrm{P}_a(\boldsymbol{x}) - \mathrm{P}_s(\boldsymbol{x})][\mathrm{P}_b(\boldsymbol{x}) - \mathrm{P}_s(\boldsymbol{x})]}{\mathrm{P}^2(\boldsymbol{x})} - \pi_s^{-1}\right] \\
&= \sum_{\boldsymbol{x}\in\Omega}\frac{\mathrm{P}_a(\boldsymbol{x})\,\mathrm{P}_b(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} - \sum_{\boldsymbol{x}\in\Omega}\frac{\mathrm{P}_a(\boldsymbol{x})\,\mathrm{P}_s(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} - \sum_{\boldsymbol{x}\in\Omega}\frac{\mathrm{P}_b(\boldsymbol{x})\,\mathrm{P}_s(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} + \sum_{\boldsymbol{x}\in\Omega}\frac{\mathrm{P}_s^2(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} - \pi_s^{-1}. \qquad \text{(B.10)}
\end{aligned}$$

We can use Corollary B.3 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 (B.10),

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

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

Proof of Theorem 2.5. This proof uses properties of matrix norms; refer to Lange (2010, Chapter 6) or Meyer (2001, Chapter 5) for background. Notice that for non-singular $q \times q$ 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,

$$\|\boldsymbol{A}^{-1} - \boldsymbol{B}^{-1}\| \leq \|\boldsymbol{A}^{-1}\| \cdot \|\boldsymbol{A} - \boldsymbol{B}\| \cdot \|\boldsymbol{B}^{-1}\|. \tag{B.11}$$

Fix $\boldsymbol{\theta} \in \Theta$, take $\boldsymbol{A} = \widetilde{\mathcal{I}}(\boldsymbol{\theta})$ and $\boldsymbol{B} = \mathcal{I}(\boldsymbol{\theta})$, and take $\|\cdot\|$ to the Frobenius matrix norm for convenience. Then (B.11) becomes

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

where $\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^q \sum_{j=1}^q a_{ij}^2}$, 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 $\mathcal{I}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}(\boldsymbol{\theta}) \to \boldsymbol{0}$ as $m \to \infty$ by Theorem 2.1, $\|\mathcal{I}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}(\boldsymbol{\theta})\|_F \to 0$. Next, we address the $\|\widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta})\|_F$ term. Using the explicit form in Corollary 2.4, we have

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

Then $\|\widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta})\|_F$ is decreasing in m, and hence is bounded in m.

We will now consider the term $\|\mathcal{I}^{-1}(\boldsymbol{\theta})\|$, with the 2-norm in place of the Frobenius norm. Let $\lambda_1(m) \geq \cdots \geq \lambda_q(m)$ be the eigenvalues of $\mathcal{I}(\boldsymbol{\theta})$ 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

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

Notice that

$$\min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T \left[\mathcal{I}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}(\boldsymbol{\theta}) \right] \boldsymbol{x} + \min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T \widetilde{\mathcal{I}}(\boldsymbol{\theta}) \boldsymbol{x} \quad \leq \quad \min_{\|\boldsymbol{x}\|=1} \left\{ \boldsymbol{x}^T \left[\mathcal{I}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}(\boldsymbol{\theta}) \right] \boldsymbol{x} + \boldsymbol{x}^T \widetilde{\mathcal{I}}(\boldsymbol{\theta}) \boldsymbol{x} \right\}$$

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

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

denoting the eigenvalues of $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ as $\widetilde{\lambda}_1(m) \geq \cdots \geq \widetilde{\lambda}_q(m)$ (all positive), and the eigenvalues of $\mathcal{I}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}(\boldsymbol{\theta})$ 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

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

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

$$0 \leq \|\mathcal{I}^{-1}(\boldsymbol{\theta})\|_2 \leq \frac{1}{\beta_q(m) + \widetilde{\lambda}_q(m)} \leq \frac{1}{\widetilde{\lambda}_q(m) - \varepsilon} \tag{B.12}$$

for all $m \geq m_0$. Because $\|\boldsymbol{A}\|_2 \leq \|\boldsymbol{A}\|_F$, the result in Part (ii) gives that for all m there exists a $K > 0$ such that,

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

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 (B.12) is bounded above by $(1/K - \varepsilon)^{-1}$ for all $m \geq m_0$, which implies $\|\mathcal{I}^{-1}(\boldsymbol{\theta})\|_2$ 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

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

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

□

Proof of Proposition 3.3. The general form for AFSA is given by

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

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 2.4 we have

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

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

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

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

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

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

$$\begin{aligned}
p_{\ell j}^{(g+1)} &= p_{\ell j}^{(g)} + \frac{1}{M} \sum_{i=1}^{n} p_{\ell j}^{(g)} \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} \left(\frac{x_{ij}}{p_{\ell j}^{(g)}} - \frac{x_{ik}}{p_{\ell k}^{(g)}} \right) - \frac{1}{M} \sum_{i=1}^{n} \sum_{t=1}^{k-1} p_{\ell j}^{(g)} p_{\ell t}^{(g)} \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} \left(\frac{x_{it}}{p_{\ell t}^{(g)}} - \frac{x_{ik}}{p_{\ell k}^{(g)}} \right) \\
&= p_{\ell j}^{(g)} + \frac{1}{M} \sum_{i=1}^{n} \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} \left(x_{ij} - \frac{p_{\ell j}^{(g)}}{p_{\ell k}^{(g)}} x_{ik} \right) - \frac{1}{M} \sum_{i=1}^{n} p_{\ell j}^{(g)} \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} \sum_{t=1}^{k-1} \left(x_{it} - \frac{p_{\ell t}^{(g)}}{p_{\ell k}^{(g)}} x_{ik} \right). \\
&= p_{\ell j}^{(g)} + \frac{1}{M} \sum_{i=1}^{n} \frac{\mathrm{P}_\ell(\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{x}_i)} \left\{ \left(x_{ij} - \frac{p_{\ell j}^{(g)}}{p_{\ell k}^{(g)}} x_{ik} \right) - p_{\ell j}^{(g)} \sum_{t=1}^{k-1} \left(x_{it} - \frac{p_{\ell t}^{(g)}}{p_{\ell k}^{(g)}} x_{ik} \right) \right\} \qquad \text{(B.13)}
\end{aligned}$$

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 (B.13) and simplifying we get

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

□

Proof of Proposition 3.4. 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 $\Delta_{i\ell} = I(z_i = \ell) \sim \text{Bernoulli}(\pi_\ell)$. Then the corresponding log-likelihood is

$$\log L(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{z}) = \sum_{i=1}^{n} \sum_{\ell=1}^{s} \Delta_{i\ell} \log \Big[\pi_\ell f(\boldsymbol{x}_i \mid \boldsymbol{p}_\ell, m_i)\Big]. \tag{B.14}$$

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

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

is the posterior probability of population ℓ, given $\boldsymbol{x}_i$ and the previous iteration. Conditional on this information, the expectation of (B.14) 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 $\pi_1, \ldots, \pi_{s-1}$ we have

$$\begin{aligned}0 = \frac{\partial}{\partial \pi_a} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(g)}) &= \sum_{i=1}^{n} \frac{\gamma_{ia}^{(g)}}{\pi_a} - \sum_{i=1}^{n} \frac{\gamma_{is}^{(g)}}{\pi_s} \\ \iff \pi_s \sum_{i=1}^{n} \gamma_{ia}^{(g)} &= \pi_a \sum_{i=1}^{n} \gamma_{is}^{(g)}. \end{aligned} \tag{B.15}$$

Summing both sides of (B.15) 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 (B.15) 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 $b = 1, \ldots, k-1$,

$$0 = \frac{\partial}{\partial p_{ab}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(g)}) = \sum_{i=1}^{n} \gamma_{ia}^{(g)} \left(\frac{x_{ib}}{p_{ab}} - \frac{x_{ik}}{p_{ik}} \right)$$

$$\iff p_{ak} \sum_{i=1}^{n} \gamma_{ia}^{(g)} x_{ib} = p_{ab} \sum_{i=1}^{n} \gamma_{ia}^{(g)} x_{ik}. \tag{B.16}$$

Summing both sides of (B.16) 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 (B.16) yields

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

□