Ann Inst Stat Math (2017) 69:333–364
DOI 10.1007/s10463-015-0542-9

An approximation to the information matrix of exponential family finite mixtures

Andrew M. Raim[1,2] · **Nagaraj K. Neerchal**[1] · **Jorge G. Morel**[1]

Received: 10 April 2014 / Revised: 19 May 2015 / Published online: 1 October 2015

Abstract A simple closed form of the Fisher information matrix (FIM) usually cannot be obtained under a finite mixture. Several authors have considered a block-diagonal FIM approximation for binomial and multinomial finite mixtures, used in scoring and in demonstrating relative efficiency of proposed estimators. Raim et al. (Stat Methodol 18:115–130, 2014a) noted that this approximation coincides with the complete data FIM of the observed data and latent mixing process jointly. It can, therefore, be formulated for a wide variety of missing data problems. Multinomial mixtures feature a number of trials, which, when taken to infinity, result in the FIM and approximation becoming arbitrarily close. This work considers a clustered sampling scheme which allows the convergence result to be extended significantly to the class of exponential family finite mixtures. A series of examples demonstrate the convergence result and suggest that it can be further generalized.

Keywords Fisher information · Complete data · Clustered sampling · Misclassification rate

1 Introduction

We consider an approximation to the Fisher information matrix (FIM) for exponential family finite mixtures. Obtaining a simple closed form for this matrix is generally not possible. A computationally convenient approximation may be useful in frequentist

✉ Andrew M. Raim
andrew.raim@gmail.com

[1] Department of Mathematics and Statistics, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA

[2] Present Address: Center for Statistical Research and Methodology, U.S. Census Bureau, 4600 Silver Hill Road, Washington DC 20233, USA

estimation (e.g. the scoring algorithm), in inference (e.g. computing standard errors and confidence intervals), and numerous other applications in which the information matrix is used.

This paper follows on to Raim et al. (2014a), which considers a block-diagonal matrix originally proposed in Blischke (1962, 1964) to approximate the FIM for the finite mixture of binomials, and later extended to multinomial finite mixtures by Morel and Nagaraj (1993). The matrix is seen to be, in fact, a complete data information matrix, where the missing data are the subpopulation indicators. The approximation and true FIM are shown to become close as the number of multinomial trials are increased, which justifies the approximation. The approximation is shown to be useful in Fisher scoring iterations, resulting in an estimation method comparable to expectation–maximization. However, the FIM and the approximation are not necessarily close for small to moderate m. It is noted that the complete data FIM can be formulated for any finite mixture, or more generally, for likelihoods involving missing data. However, the convergence between approximation and true FIM could not immediately be extended beyond the scope of multinomial data analysis, as it was based on the number of trials becoming large.

This paper provides one such extension, to exponential family finite mixtures. We consider a special clustered sampling scheme; suppose that m observations are sampled from one of s subpopulations. It is unknown to which subpopulation the observations belong, as in the usual finite mixture, but it is known that they share a common subpopulation. This provides an analogue to the trials of a binomial or multinomial experiment and allows a convergence result to be formulated.

The proof in the multinomial setting (Morel and Nagaraj 1991; Raim et al. 2014a) had been based on bounds for tail probabilities of binomial random variables and used the fact that the sample space is bounded. The proof in the present paper utilizes the exponential family form and does not require restrictions on the sample space. It is shown that the FIM and the approximation converge together as $m \to \infty$, and the convergence is exponential in m. However, the exponent includes a term which depends on the distance between subpopulations so that the convergence is very slow when subpopulations are similar and very fast when dissimilar. Therefore, the approximation is most suitable when the mixed subpopulations are more distinct and m is larger.

Because of the intractability of deriving the expectations needed for the FIM of a finite mixture, "observed" information quantities such as the Hessian of the log-likelihood or outer product of the score vector are often used in inference applications. For example, McLachlan and Peel (2000, Chapter 2) review several methods based on observed information, such as one proposed by Louis (1982) to obtain standard errors from the expectation–maximization algorithm. More recently, Boldea and Magnus (2009) provide expressions for observed information under the multivariate normal finite mixture. In the present work, we consider the FIM to be a quantity of interest in its own right. One important distinction between expected and observed information is the properties of the latter, such as invertibility, vary with the sample.

The rest of the paper proceeds as follows: Section 2 gives the formulation of the problem. Section 3 proves that the complete data FIM and true FIM become arbitrarily close as m becomes large, and provides rates of convergence. Section 4 highlights a connection between the convergence rate and the probability of misclassification

among the s subpopulations using an optimal classification rule. Section 5 provides several examples of the convergence. Finally, Sect. 6 gives concluding remarks.

2 Problem formulation

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

$$f(\mathbf{x} \mid \boldsymbol{\eta}_\ell) = \exp\{h(\mathbf{x}) + \boldsymbol{\eta}_\ell^{\mathrm{T}} \mathbf{u}(\mathbf{x}) - \psi(\boldsymbol{\eta}_\ell)\}.$$

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

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_m \mid \boldsymbol{\eta}_\ell) = \exp\left\{\sum_{i=1}^{m} h(\mathbf{x}_i) + \boldsymbol{\eta}_\ell^{\mathrm{T}} \sum_{i=1}^{m} \mathbf{u}_i - m\psi(\boldsymbol{\eta}_\ell)\right\}. \tag{1}$$

By Lemma 2.7.2 of Lehmann and Romano (2005), the density of $\mathbf{T} = \sum_{i=1}^{m} \mathbf{U}_i$ conditional on the subpopulation $Z = \ell$ can be obtained from (1) as

$$f(\mathbf{t} \mid \boldsymbol{\eta}_\ell) = \exp\{\boldsymbol{\eta}_\ell^{\mathrm{T}} \mathbf{t} - m\psi(\boldsymbol{\eta}_\ell)\}, \tag{2}$$

with respect to a σ-finite measure ν obtained by transforming the original dominating measure. A distribution in the form of (2) is said to belong to a natural exponential family. Unconditionally, the distribution of $\mathbf{T}$ is given by

$$f(\mathbf{t} \mid \boldsymbol{\theta}) = \sum_{\ell=1}^{s} \pi_\ell \exp\{\boldsymbol{\eta}_\ell^{\mathrm{T}} \mathbf{t} - m\psi(\boldsymbol{\eta}_\ell)\}, \tag{3}$$

with respect to the measure ν, where $\boldsymbol{\theta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_s, \pi_1, \ldots, \pi_{s-1})$. We will assume $\boldsymbol{\eta}_\ell$ are distinct for $\ell = 1, \ldots, s$ to prevent an obviously degenerate finite mixture. The notation Ω will refer to the abstract sample space with a typical element ω. Let $\mathbf{W}_\ell$

be a random variable with the distribution of $\mathbf{T}$ when $Z = \ell$ is observed. Expectation, variance, and moment generating function (MGF) of $\mathbf{W}_\ell$ are given by

$$E(\mathbf{W}_\ell) = m\frac{\partial}{\partial \boldsymbol{\eta}_\ell}\psi(\boldsymbol{\eta}_\ell), \quad \mathrm{Var}(\mathbf{W}_\ell) = m\frac{\partial^2}{\partial \boldsymbol{\eta}_\ell \partial \boldsymbol{\eta}_\ell^{\mathrm{T}}}\psi(\boldsymbol{\eta}_\ell), \quad \text{and}$$
$$E(e^{\boldsymbol{\tau}^{\mathrm{T}}\mathbf{W}_\ell}) = e^{m[\psi(\boldsymbol{\eta}_\ell+\boldsymbol{\tau})-\psi(\boldsymbol{\eta}_\ell)]},$$

where the MGF exists for $\boldsymbol{\tau}$ in some ball $B(\mathbf{0}, \varepsilon)$ of radius $\varepsilon > 0$ about $\mathbf{0}$ (Shao 2008, Theorem 2.1). The score vector and Fisher information matrix of $\mathbf{W}_\ell$ are

$$\frac{\partial}{\partial \boldsymbol{\eta}_\ell}\log f(\mathbf{t} \mid \boldsymbol{\eta}_\ell) = \mathbf{t} - E(\mathbf{W}_\ell) \quad \text{and} \quad E\left\{-\frac{\partial^2}{\partial \boldsymbol{\eta}_\ell \partial \boldsymbol{\eta}_\ell^{\mathrm{T}}}\log f(\mathbf{T} \mid \boldsymbol{\eta}_\ell)\right\} = \mathrm{Var}(\mathbf{W}_\ell).$$

The score vector of $\mathbf{T}$ can be obtained as

$$\frac{\partial}{\partial \boldsymbol{\eta}_\ell}\log f(\mathbf{t} \mid \boldsymbol{\theta}) = \frac{\pi_\ell f(\mathbf{t} \mid \boldsymbol{\eta}_\ell)}{f(\mathbf{t} \mid \boldsymbol{\theta})}[\mathbf{t} - E(\mathbf{W}_\ell)], \quad \text{for } \ell = 1, \ldots, s$$
$$\frac{\partial}{\partial \pi_\ell}\log f(\mathbf{t} \mid \boldsymbol{\theta}) = \frac{f(\mathbf{t} \mid \boldsymbol{\eta}_\ell) - f(\mathbf{t} \mid \boldsymbol{\eta}_s)}{f(\mathbf{t} \mid \boldsymbol{\theta})}, \quad \text{for } \ell = 1, \ldots, s-1.$$

Denote $\mathcal{I}(\boldsymbol{\theta})$ as the FIM of $\mathbf{T}$ under the finite mixture and $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ as the FIM of the complete data $(\mathbf{T}, Z)$, both with respect to $\boldsymbol{\theta}$. Let $q = sk + s - 1$ denote the dimension of $\boldsymbol{\theta}$ so that $\mathcal{I}(\boldsymbol{\theta})$ and $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ are $q \times q$ matrices. We will sometimes use the subscript m to emphasize that the matrices depend on the number of observations m. The matrix $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ has a simple closed form

$$\begin{aligned}\widetilde{\mathcal{I}}(\boldsymbol{\theta}) &= \mathrm{Blockdiag}\left(\pi_1\mathbf{F}_1, \ldots, \pi_s\mathbf{F}_s, \mathbf{F}_\pi\right), \quad \text{where} \\ \mathbf{F}_\ell &= m\{\mathrm{Var}(\mathbf{U}_1 \mid Z = \ell)\}, \quad \text{for } \ell = 1, \ldots, s, \\ \mathbf{F}_\pi &= \mathbf{D}_\pi^{-1} + \pi_s^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}.\end{aligned} \tag{4}$$

Here, $\mathbf{D}_\pi = \mathrm{Diag}(\pi_1, \ldots, \pi_{s-1})$ and $\mathbf{1}$ denotes a vector of ones of the appropriate dimension. Notice that $\mathbf{F}_\ell$ is the $k \times k$ FIM with respect to $\mathbf{W}_\ell$, and $\mathbf{F}_\pi$ is the $(s-1) \times (s-1)$ FIM of $\mathrm{Mult}_s(\boldsymbol{\pi}, 1)$, the multinomial distribution on s categories with probabilities $\boldsymbol{\pi}$ and a single trial. To obtain expression (4), the complete data density for $(\mathbf{T}, Z)$ is

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

Let $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_s)$ with $\Delta_\ell = I(Z = \ell)$ so that $\boldsymbol{\Delta} \sim \mathrm{Mult}_s(1, \boldsymbol{\pi})$. Denote $\boldsymbol{\Delta}_{-s} = (\Delta_1, \ldots, \Delta_{s-1})$ and $\boldsymbol{\pi}_{-s} = (\pi_1, \ldots, \pi_{s-1})$. This complete data density yields a score vector with entries

$$\frac{\partial}{\partial \boldsymbol{\eta}_a} \log f(\mathbf{t}, z \mid \boldsymbol{\theta}) = \Delta_a \frac{\partial}{\partial \boldsymbol{\eta}_a} \log f(\mathbf{t} \mid \boldsymbol{\eta}_a), \quad \text{for } a = 1, \ldots, s,$$
$$\frac{\partial}{\partial \boldsymbol{\pi}_{-s}} \log f(\mathbf{t}, z \mid \boldsymbol{\theta}) = \mathbf{D}_{\pi}^{-1} \boldsymbol{\Delta}_{-s} - \frac{\Delta_s}{\pi_s} \mathbf{1}.$$

for $a = 1, \ldots s$. Taking second derivatives yields

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

for $a, b \in \{1, \ldots, s\}$. Taking the expected value of the negative of these terms, jointly with respect to $(\mathbf{T}, Z)$, obtains the blocks of (4).

In the specific case of multinomial finite mixtures, $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ is seen to serve the role of an approximate information matrix in Raim et al. (2014a). In Section 3 we show that $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta}) \to \mathbf{0}$ as $m \to \infty$.

In a practical data analysis situation under clustered sampling, a random sample $\mathbf{T}_1, \ldots, \mathbf{T}_n$ would be observed. Here, each $\mathbf{T}_i$ represents the sufficient statistic based on m_i individual observations $\mathbf{X}_{i1}, \ldots, \mathbf{X}_{im_i}$ drawn from the subpopulation labeled by a common unobserved Z_i. If we assume that the distribution of Z_i is Discrete$(1, \ldots, J; \boldsymbol{\pi})$, we obtain the finite mixture model (3). The score vector, FIM, and approximate FIM for the sample are formed by summing the corresponding n quantities for each clustered observation. In this case, closeness of $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ and $\mathcal{I}(\boldsymbol{\theta})$ will be ensured when all m_i are sufficiently large.

3 Convergence of approximate information matrix

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

Lemma 1 *The matrix $\widetilde{\mathcal{I}}(\boldsymbol{\theta}) - \mathcal{I}(\boldsymbol{\theta})$ is equal to $E_{\mathbf{T}}[\mathcal{I}_{Z|\mathbf{T}}(\boldsymbol{\theta})]$.*

Proof Notice that

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

Therefore,

$$\begin{aligned}\widetilde{\mathcal{I}}(\boldsymbol{\theta}) &= E_{\mathbf{T},Z}\left[\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(\mathbf{T},Z)\right\}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(\mathbf{T},Z)\right\}^{\mathrm{T}}\right] \\ &= E_{\mathbf{T}}[\mathcal{I}_{Z|\mathbf{T}}(\boldsymbol{\theta})] + \mathbf{B} + \mathbf{B}^{\mathrm{T}} + \mathcal{I}(\boldsymbol{\theta}), \end{aligned} \tag{5}$$

where

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

The result follows from rearranging terms in (5). □

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

$$\text{Actual information} = \text{complete information} - \text{missing information}.$$

Before proceeding with the main result, we state several important consequences of Lemma 1. A Wald-like test statistic based on the approximation will be systematically too large, and a Score-like test statistic will be too small. Also, standard errors obtained from the approximate information matrix will be systematically too optimistic (small). The notation $\mathbf{e}_j$ will be used to represent the jth column of the identity matrix of the appropriate dimension.

Corollary 1 (a) (Wald Statistic) *For any* $\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0 \in \Theta$,

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

(b) (Score Statistic) *Suppose* $\mathcal{I}(\boldsymbol{\theta})$ *and* $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ *are nonsingular and that* $\widetilde{\mathcal{I}}(\boldsymbol{\theta}) - \mathcal{I}(\boldsymbol{\theta})$ *is positive definite. Then for any* $\boldsymbol{\theta}_0 \in \Theta$,

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

(c) (Standard Errors) *Suppose* $\mathcal{I}(\boldsymbol{\theta})$ *and* $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ *are nonsingular and that* $\widetilde{\mathcal{I}}(\boldsymbol{\theta}) - \mathcal{I}(\boldsymbol{\theta})$ *is positive definite. Denote by* $\mathcal{I}^{ij}(\boldsymbol{\theta})$ *and* $\widetilde{\mathcal{I}}^{ij}(\boldsymbol{\theta})$ *the elements of* $\mathcal{I}^{-1}(\boldsymbol{\theta})$ *and* $\widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta})$, *respectively. Then* $\{\mathcal{I}^{jj}(\boldsymbol{\theta})\}^{1/2} > \{\widetilde{\mathcal{I}}^{jj}(\boldsymbol{\theta})\}^{1/2}$ *for* $j = 1, \ldots, q$.

Proof (a) From Lemma 1, $\widetilde{\mathcal{I}}(\hat{\boldsymbol{\theta}}) - \mathcal{I}(\hat{\boldsymbol{\theta}}) = E_{\mathbf{T}}[\mathcal{I}_{Z|\mathbf{T}}(\boldsymbol{\theta})]$, an expected value of a conditional information matrix which is positive semidefinite. Therefore, the quantity $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}}(\widetilde{\mathcal{I}}(\hat{\boldsymbol{\theta}}) - \mathcal{I}(\hat{\boldsymbol{\theta}}))(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is nonnegative and the result follows.

(b) Lemma 6 in the Appendix gives that $\mathcal{I}^{-1}(\boldsymbol{\theta}_0) - \widetilde{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}}_0)$ is positive definite, which implies that the quantity $[S(\boldsymbol{\theta}_0)]^{\mathrm{T}}(\mathcal{I}^{-1}(\boldsymbol{\theta}_0) - \widetilde{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}}_0))[S(\boldsymbol{\theta}_0)]$ is strictly positive, and the result follows.
(c) Lemma 6 gives that $\mathcal{I}^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta})$ is positive definite; therefore, the diagonal elements $\mathbf{e}_j^{\mathrm{T}}[\mathcal{I}^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta})]\mathbf{e}_j$ are positive for $j = 1, \ldots, q$. □

A useful consequence of Lemma 1 is next given as Proposition 1, which states that the off-diagonal elements of the matrix $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})$ have magnitudes which are bounded by the diagonal elements. This will allow our convergence proof to focus only on the diagonal elements.

Proposition 1 *Denote the* (i, j)*th element of* $\mathcal{I}_{Z|\mathbf{T}}(\boldsymbol{\theta})$ *as* C_{ij}*. Then*

$$E|C_{ij}| \leq \{E(C_{ii})\}^{1/2}\{E(C_{jj})\}^{1/2}.$$

Proof Recall that $E(C_{ij})$ is the (i, j)th element of $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})$ by Lemma 1. Because $\mathcal{I}_{Z|\mathbf{T}}(\boldsymbol{\theta})$ is the covariance matrix of a score vector, we may apply the Cauchy-Schwarz inequality to obtain

$$|C_{ij}| \leq C_{ii}^{1/2} \cdot C_{jj}^{1/2},$$

for any pair (i, j), which implies that

$$E|C_{ij}| \leq E\left\{C_{ii}^{1/2} \cdot C_{jj}^{1/2}\right\}.$$

Apply Cauchy–Schwarz again to the right-hand side to obtain

$$E\left\{C_{ii}^{1/2} \cdot C_{jj}^{1/2}\right\} \leq \{E[C_{ii}]\}^{1/2} \cdot \{E[C_{jj}]\}^{1/2},$$

which gives the result. □

We focus on the parameterization $\boldsymbol{\theta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_s, \boldsymbol{\pi}_{-s})$ for convenience, but note that the convergence behavior is preserved under transformations. Suppose $\boldsymbol{\theta}(\boldsymbol{\vartheta})$ is a differentiable transformation of $\boldsymbol{\vartheta}$ which does not depend on m. We have that

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

so that $\widetilde{\mathcal{I}}_m(\boldsymbol{\vartheta}) - \mathcal{I}_m(\boldsymbol{\vartheta}) \to \mathbf{0}$ as $m \to \infty$ if and only if $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta}) \to \mathbf{0}$, with equivalent rates of convergence.

Now consider the block decomposition of the true information matrix

$$\mathcal{I}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{A}_{11} & \ldots & \mathbf{A}_{1s} & \mathbf{A}_{1\pi} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{A}_{s1} & \ldots & \mathbf{A}_{ss} & \mathbf{A}_{s\pi} \\ \mathbf{A}_{\pi 1} & \ldots & \mathbf{A}_{\pi s} & \mathbf{A}_{\pi\pi} \end{pmatrix}, \tag{6}$$

with blocks

$$\mathbf{A}_{ab} = E\left[\left\{\frac{\partial}{\partial \boldsymbol{\eta}_a}\log f(\mathbf{t}\mid\boldsymbol{\theta})\right\}\left\{\frac{\partial}{\partial \boldsymbol{\eta}_b}\log f(\mathbf{t}\mid\boldsymbol{\theta})\right\}^{\mathrm{T}}\right], \quad \mathrm{a},\mathrm{b}\in\{1,\ldots,s\},$$

$$\mathbf{A}_{b\pi}^{\mathrm{T}} = \mathbf{A}_{\pi b} = E\left[\left\{\frac{\partial}{\partial \boldsymbol{\pi}_{-s}}\log f(\mathbf{t}\mid\boldsymbol{\theta})\right\}\left\{\frac{\partial}{\partial \boldsymbol{\eta}_b}\log f(\mathbf{t}\mid\boldsymbol{\theta})\right\}^{\mathrm{T}}\right], \quad \mathrm{b}\in\{1,\ldots,s\},$$

$$\mathbf{A}_{\pi\pi} = E\left[\left\{\frac{\partial}{\partial \boldsymbol{\pi}_{-s}}\log f(\mathbf{t}\mid\boldsymbol{\theta})\right\}\left\{\frac{\partial}{\partial \boldsymbol{\pi}_{-s}}\log f(\mathbf{t}\mid\boldsymbol{\theta})\right\}^{\mathrm{T}}\right].$$

By Proposition 1, it is only necessary to show convergence of the diagonal elements of $\widetilde{\mathcal{I}}(\boldsymbol{\theta}) - \mathcal{I}(\boldsymbol{\theta})$ to zero. To do this, we will obtain expressions for the diagonal blocks. It will be helpful to define

$$R_i^{(m)}(\mathbf{t}) = \sum_{\ell\neq i}^{s}\pi_\ell \exp\{(\boldsymbol{\eta}_\ell - \boldsymbol{\eta}_i)^{\mathrm{T}}\mathbf{t} - m[\psi(\boldsymbol{\eta}_\ell) - \psi(\boldsymbol{\eta}_i)]\} = \frac{f(\mathbf{t}\mid\boldsymbol{\theta})}{f(\mathbf{t}\mid\boldsymbol{\eta}_i)} - \pi_i, \quad \text{and}$$

$$Q_i^{(m)}(\mathbf{t}) = \frac{\pi_i f(\mathbf{t}\mid\boldsymbol{\eta}_i)}{f(\mathbf{t}\mid\boldsymbol{\theta})} = \frac{\pi_i}{\pi_i + R_i^{(m)}(\mathbf{t})}.$$

Notice that $Q_i^{(m)}(\mathbf{T}) = P(Z = \ell \mid \mathbf{T})$ is the posterior probability of observing the ℓth subpopulation given an observed $\mathbf{T}$; hence, taking expectation with respect to the mixture density of $f(\mathbf{t}\mid\boldsymbol{\theta})$ yields $E_{\mathbf{T}}[Q_\ell^{(m)}(\mathbf{T})] = P(Z=\ell) = \pi_\ell$. Later we will encounter $E[Q_\ell^{(m)}(\mathbf{W}_\ell)]$, the expectation taken with respect to $f(\mathbf{t}\mid\boldsymbol{\eta}_\ell)$, which does not simplify trivially.

Consider the decomposition of $\widetilde{\mathcal{I}}(\boldsymbol{\theta}) - \mathcal{I}(\boldsymbol{\theta})$ into blocks according to (6); block (i,i) can be written as

$$\pi_i\mathbf{F}_i - \mathbf{A}_{ii} = \pi_i^2\int[1 - Q_i^{(m)}(\mathbf{t})](\mathbf{t} - E(\mathbf{W}_i))(\mathbf{t} - E(\mathbf{W}_i))^{\mathrm{T}} f(\mathbf{t}\mid\boldsymbol{\eta}_i)d\nu(\mathbf{t}),$$

whose jth diagonal element is

$$\mathbf{e}_j^{\mathrm{T}}[\pi_i\mathbf{F}_i - \mathbf{A}_{ii}]\mathbf{e}_j = \pi_i^2 E\left\{[1 - Q_i^{(m)}(\mathbf{W}_i)][W_{ij} - E(W_{ij})]^2\right\}. \tag{7}$$

Here, W_{ij} represents the jth element of $\mathbf{W}_i$. The lower right diagonal block of $\widetilde{\mathcal{I}}(\boldsymbol{\theta}) - \mathcal{I}(\boldsymbol{\theta})$ is

$$\mathbf{F}_\pi - \mathbf{A}_{\pi\pi} = (\mathbf{D}_\pi^{-1} + \pi_s^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}) - E\left[\frac{1}{f^2(\mathbf{T}\mid\boldsymbol{\theta})}\begin{pmatrix} f(\mathbf{T}\mid\boldsymbol{\eta}_1) - f(\mathbf{T}\mid\boldsymbol{\eta}_s)\\ \vdots \\ f(\mathbf{T}\mid\boldsymbol{\eta}_{s-1}) - f(\mathbf{T}\mid\boldsymbol{\eta}_s)\end{pmatrix}\begin{pmatrix} f(\mathbf{T}\mid\boldsymbol{\eta}_1) - f(\mathbf{T}\mid\boldsymbol{\eta}_s)\\ \vdots \\ f(\mathbf{T}\mid\boldsymbol{\eta}_{s-1}) - f(\mathbf{T}\mid\boldsymbol{\eta}_s)\end{pmatrix}^{\mathrm{T}}\right]. \tag{8}$$

whose ath diagonal element can be expressed as

$$\mathbf{e}_a^{\mathrm{T}}[\mathbf{F}_\pi - \mathbf{A}_{\pi\pi}]\mathbf{e}_a = (\pi_a^{-1} + \pi_s^{-1}) - \pi_a^{-1} E[Q_a^{(m)}(\mathbf{W}_a)] - \pi_s^{-1} E[Q_s^{(m)}(\mathbf{W}_s)] + 2\pi_a^{-1} E[Q_a^{(m)}(\mathbf{W}_s)]. \tag{9}$$

The following lemma gives a simple convexity result for exponential family densities which will determine the behavior of $R_i^{(m)}(\mathbf{W}_j)$ and $Q_i^{(m)}(\mathbf{W}_j)$ as $m \to \infty$. See Boyd and Vandenberghe (2004) for background on convex analysis.

Lemma 2 *Consider the density* $f(\mathbf{t} \mid \boldsymbol{\eta}) = \exp\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{t} - m\psi(\boldsymbol{\eta})\}$ *with natural parameter space* Ξ *an open convex set. Suppose its FIM* $\mathcal{I}_m(\boldsymbol{\eta})$ *is positive definite on* Ξ. *For any* $\boldsymbol{\eta}, \boldsymbol{\eta}^* \in \Xi$,

$$\psi(\boldsymbol{\eta}) - \psi(\boldsymbol{\eta}^*) > \psi'(\boldsymbol{\eta}^*)^{\mathrm{T}}(\boldsymbol{\eta} - \boldsymbol{\eta}^*), \tag{10}$$

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

Proof Notice that

$$\frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^{\mathrm{T}}} \log f(\mathbf{t} \mid \boldsymbol{\eta}) = m \frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^{\mathrm{T}}} \psi(\boldsymbol{\eta}) = \mathcal{I}_m(\boldsymbol{\eta}),$$

implying ψ is a strictly convex function. Since ψ is differentiable on the convex set Ξ we have the result (10). □

Next, the behavior of $R_i^{(m)}(\mathbf{W}_j)$ and $Q_i^{(m)}(\mathbf{W}_j)$ will be determined for large m. Note that the behavior depends on the subpopulation, $j = 1, \ldots, s$, which characterizes the distribution of $\mathbf{W}_j$. The expressions

$$\gamma_{IJK} = -\psi'(\boldsymbol{\eta}_J)^{\mathrm{T}}(\boldsymbol{\eta}_I - \boldsymbol{\eta}_K) + [\psi(\boldsymbol{\eta}_I) - \psi(\boldsymbol{\eta}_K)],$$
$$c_i^* = \bigwedge_{\ell \neq i}^{s} \gamma_{\ell i i}, \quad d_{ij}^* = \bigvee_{\ell \neq i}^{s} \{-\gamma_{\ell j i}\}, \quad \text{and} \quad c^{**} = \bigwedge_{\ell=1}^{s} c_\ell^*, \tag{11}$$

will be used for the remainder of the paper. To describe the almost sure behavior of a sequence $\{X_m\}$ of random variables, let $\{a_m\}$ and $\{b_m\}$ be sequences of real numbers. We will write:

1. $X_m \overset{a.s.}{=} O(a_m)$ if there exists a set A having probability 1 where, for each $\omega \in A$, there exists a $K(\omega)$ such that $|X_m(\omega)/a_m| < K(\omega)$ for all $m \in \{1, 2, \ldots\}$. The quantity $K(\omega)$ may depend on ω but is free of m.
2. $O(a_m) \leq X_m \leq O(b_m)$ almost surely if both $a_m/X_m \overset{a.s.}{=} O(1)$ and $X_m/b_m \overset{a.s.}{=} O(1)$.

Proposition 2 *The sequence* $R_i^{(m)}(\mathbf{W}_j)$ *behaves as follows for large* m:

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

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

$$O(e^{m\gamma_{ijj}}) \leq R_i^{(m)}(\mathbf{W}_j) \leq O(e^{md_{ij}^*}), \quad \textit{almost surely}.$$

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

Proof Let g be the continuous function $g(\mathbf{x}) = (\boldsymbol{\eta}_\ell - \boldsymbol{\eta}_i)^{\mathrm{T}}\mathbf{x}$. By the strong law of large numbers, there exists an $A \subseteq \Omega$ having probability 1, where $g(\mathbf{W}_j(\omega)/m) \to g(\psi'(\boldsymbol{\eta}_j))$ as $m \to \infty$ for all $\omega \in A$. Select any $\varepsilon \in (0, c_i^*)$. For any $\omega \in A$, there exists an $M(\omega)$ such that for all $m \geq M(\omega)$,

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

We have that

$$\begin{aligned}R_i^{(m)}(\mathbf{W}_j(\omega)) &< \sum_{\ell \neq i}^{s} \pi_\ell \exp\{m[\psi'(\boldsymbol{\eta}_j)^{\mathrm{T}}(\boldsymbol{\eta}_\ell - \boldsymbol{\eta}_i) - [\psi(\boldsymbol{\eta}_\ell) - \psi(\boldsymbol{\eta}_i)] + \varepsilon]\} \\ &= \sum_{\ell \neq i}^{s} \pi_\ell \exp\{m(-\gamma_{\ell ji} + \varepsilon)\}\end{aligned}$$

and

$$\begin{aligned}R_i^{(m)}(\mathbf{W}_j(\omega)) &> \sum_{\ell \neq i}^{s} \pi_\ell \exp\{m[\psi'(\boldsymbol{\eta}_j)^{\mathrm{T}}(\boldsymbol{\eta}_\ell - \boldsymbol{\eta}_i) - [\psi(\boldsymbol{\eta}_\ell) - \psi(\boldsymbol{\eta}_i)] - \varepsilon]\} \\ &= \sum_{\ell \neq i}^{s} \pi_\ell \exp\{m(-\gamma_{\ell ji} - \varepsilon)\}\end{aligned}$$

for all $m \geq M(\omega)$.

Case (*a*) Suppose $j = i$. From Lemma 2 we have

$$\gamma_{\ell ii} = -\psi'(\boldsymbol{\eta}_i)^{\mathrm{T}}(\boldsymbol{\eta}_\ell - \boldsymbol{\eta}_i) + [\psi(\boldsymbol{\eta}_\ell) - \psi(\boldsymbol{\eta}_i)] > 0$$

for all $\ell \neq i$, so that for $m \geq M(\omega)$,

$$\begin{aligned}0 &\leq R_i^{(m)}(\mathbf{W}_i(\omega)) \\ &< \sum_{\ell \neq i}^{s} \pi_\ell e^{m(-\gamma_{\ell ii} + \varepsilon)} = \sum_{\ell \neq i}^{s} \pi_\ell e^{-m(\gamma_{\ell ii} - \varepsilon)} < e^{-m(c_i^* - \varepsilon)} \sum_{\ell \neq i}^{s} \pi_\ell \\ &= (1 - \pi_i)e^{-m(c_i^* - \varepsilon)}. \qquad (12)\end{aligned}$$

Recall that $c_i^* > \varepsilon$, so the exponent on the RHS of (12) is negative. Since (12) can be obtained for each $\omega \in A$, and ε can be selected arbitrarily close to 0, we have $R_i^{(m)}(\mathbf{W}_i) \overset{a.s.}{=} O(e^{-mc_i^*})$.

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

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

Notice that

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

where $\gamma_{ijj} > 0$ by Lemma 2. Then for $m \geq M(\omega)$,

$$R_i^{(m)}(\mathbf{W}_j(\omega)) > \sum_{\ell \neq i}^{s} \pi_\ell e^{m(-\gamma_{\ell ji} - \varepsilon)} > \pi_j e^{m(-\gamma_{jji} - \varepsilon)} = \pi_j e^{m(\gamma_{ijj} - \varepsilon)}. \tag{13}$$

Note that $\gamma_{ijj} - \varepsilon > 0$ through the choice of a sufficiently small ε. Since the expression (13) can be obtained for each $\omega \in A$, and ε could have been taken arbitrarily small, we have that $\pi_j e^{m\gamma_{ijj}} / R_i^{(m)}(\mathbf{W}_j) \overset{a.s.}{=} O(1)$. We can also obtain an upper bound using

$$R_i^{(m)}(\mathbf{W}_j(\omega)) < \sum_{\ell \neq i}^{s} \pi_\ell e^{m(-\gamma_{\ell ji} + \varepsilon)} < (1 - \pi_i) e^{m(d_{ij}^* + \varepsilon)}, \tag{14}$$

noting that $d_{ij}^* = \bigvee_{\ell \neq i}^{s} \{-\gamma_{\ell ji}\} \geq -\gamma_{jji} = \gamma_{ijj} > 0$. The expression (14) can be obtained for each $\omega \in A$ for arbitrarily small ε; therefore, $R_i^{(m)}(\mathbf{W}_j) / e^{m(d_{ij}^*)} \overset{a.s.}{=} O(1)$. We have obtained the desired almost sure bounds

$$O(e^{m\gamma_{ijj}}) \leq R_i^{(m)}(\mathbf{W}_j) \leq O(e^{md_{ij}^*}).$$

□

Note that multipliers involving $\boldsymbol{\pi}$ are constant in m so we have dropped them from the rates. However, the quality of the approximation of $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ to $\mathcal{I}(\boldsymbol{\theta})$ is not uniform in $\boldsymbol{\pi}$; this is demonstrated in Raim et al. (2014b).

Proposition 3 *The sequence $Q_i^{(m)}(\mathbf{W}_j)$ behaves as follows for large m:*

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

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

Proof Case (*a*) We have

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

Since $R_i^{(m)}(\mathbf{W}_i) \overset{a.s.}{=} O(e^{-mc_i^*})$ by Proposition 2, there exists a constant $K(\omega)$ for each ω in a set A with probability 1 such that

$$\left| \frac{R_i^{(m)}(\mathbf{W}_i(\omega))}{e^{-mc_i^*}} \right| < K(\omega) \iff [R_i^{(m)}(\mathbf{W}_i(\omega))]^{-1} > K(\omega)^{-1} e^{mc_i^*},$$

so that

$$e^{mc_i^*}[1 - Q_i^{(m)}(\mathbf{W}_i(\omega))] < \frac{e^{mc_i^*}}{K(\omega)^{-1} e^{mc_i^*} + 1} < K(\omega).$$

This gives the desired rate $1 - Q_i^{(m)}(\mathbf{W}_i) \overset{a.s.}{=} O(e^{-mc_i^*})$.

Case (*b*) Proposition 2 gives a set $A \subseteq \Omega$ of probability 1 where, for each $\omega \in A$, there exists a $K(\omega)$ such that $R_i^{(m)}(\mathbf{W}_j(\omega)) > K(\omega)^{-1} e^{m\gamma_{ijj}}$. Then, we have

$$e^{m\gamma_{ijj}} Q_i^{(m)}(\mathbf{W}_j(\omega)) = \frac{\pi_i e^{m\gamma_{ijj}}}{\pi_i + R_i^{(m)}(\mathbf{W}_j(\omega))} < \frac{\pi_i e^{m\gamma_{ijj}}}{\pi_i + K(\omega)^{-1} e^{m\gamma_{ijj}}} < \pi_i K(\omega).$$

This gives the desired rate $Q_i^{(m)}(\mathbf{W}_j) \overset{a.s.}{=} O(e^{-m\gamma_{ijj}})$. □

Proposition 3 suggests that the convergence between the FIM and approximate information will be fast when both of the following happen as m is increased. First, the posterior probability of membership in the ℓth subpopulation should quickly approach 1 when the true subpopulation $Z = \ell$. Second, the posterior probability of membership in the ℓth subpopulation should quickly approach 0 when the true subpopulation $Z \neq \ell$. It is clear from Proposition 3 and dominated convergence that the expectation (9) converges to zero. Also note that $W_{ij} - E(W_{ij})$ is a sum of independent and identically distributed random variables, so that $[W_{ij} - E(W_{ij})]^2 \overset{a.s.}{=} O(m^2)$, and, therefore,

$$\pi_i^2[1 - Q_i^{(m)}(\mathbf{W}_i)][W_{ij} - E(W_{ij})]^2 \overset{a.s.}{=} O(m^2 e^{-mc_i^*}). \tag{15}$$

Then the expectation (7) converges to zero if and only if the LHS of (15) is uniformly integrable (Shao 2008, Theorem 1.8). The convergence of $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})$ can, therefore, be characterized in the following theorem:

Theorem 1 *$\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta}) \to 0$ as $m \to \infty$ if and only if the sequence (15) is uniformly integrable for each $i = 1, \ldots, s$ and $j = 1, \ldots, k$.*

Some additional work will allow us to prove $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta}) \to 0$ directly without checking uniform integrability, and also to obtain rates of convergence. The following corollary is an immediate consequence of Proposition 4 in the Appendix, noting that $\frac{1}{2}[\gamma_{jii} + \gamma_{ijj}] \geq c^{**}$.

Corollary 2 *There exist* $\kappa > 0$ *and* $\zeta > 0$ *such that*

$$E[1 - Q_i^{(m)}(\mathbf{W}_i)] = O(e^{-m(\kappa c^{**} - \zeta)}),$$

where $\kappa c^{**} > \zeta$.

Lemma 3 *Let* $S_n = X_1 + \cdots + X_n$ *where* $\{X_i\}$ *are independent and identically distributed and* $E(|X_1|^k) < \infty$ *for a given positive integer* $k \geq 0$. *Then* $E(S_n^k) = O(n^k)$.

Proof Notice that

$$E(S_n^k) = E[(X_1 + \cdots + X_n)^k] = \sum_{\mathbf{z} \in \Omega_{n,k}} \frac{k!}{z_1! \cdots z_n!} E[X_1^{z_1}] \cdots E[X_1^{z_n}]$$

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

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

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

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

which gives the result. □

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

Theorem 2 *There exist* $\kappa > 0$ *and* $\zeta > 0$ *such that the diagonal elements of* $\widetilde{\mathcal{I}}(\boldsymbol{\theta}) - \mathcal{I}(\boldsymbol{\theta})$ *satisfy the following*:

(a) *For the* j*th diagonal element of the* i*th diagonal block,* $j = 1, \ldots, k$ *and* $i = 1, \ldots, s$,

$$\mathbf{e}_j^{\mathrm{T}}(\pi_i \mathbf{F}_i - \mathbf{A}_{ii})\mathbf{e}_j = O(m^2 e^{-\frac{m}{2}(\kappa c^{**} - \zeta)}).$$

(b) *For the* j*th diagonal element of the* $\boldsymbol{\pi}$ *diagonal block,* $j = 1, \ldots, s - 1$,

$$\mathbf{e}_j^{\mathrm{T}} \left(\mathbf{F}_\pi - \mathbf{A}_{\pi\pi} \right) \mathbf{e}_j = O(e^{-m(\kappa c^{**} - \zeta)}).$$

Proof Corollary 2 provides a pair $(\kappa_\ell, \zeta_\ell)$ for the order of each $E[1 - Q_\ell^{(m)}(\mathbf{W}_\ell)]$, $\ell = 1, \ldots, s$. Define (κ, ζ) to be the pair which minimizes $\kappa_\ell c^{**} - \zeta_\ell$ over $\ell = 1, \ldots, s$. For (a) we have

$$
\begin{aligned}
&\pi_i^2 E\{[1 - Q_i^{(m)}(\mathbf{W}_i)][W_{ij} - E(W_{ij})]^2\} \\
&\quad \le \pi_i^2 \{E[(1 - Q_i^{(m)}(\mathbf{W}_i))^2]\}^{1/2} \{E[(W_{ij} - E(W_{ij}))^4]\}^{1/2} \qquad (16) \\
&\quad \le \pi_i^2 \{E[1 - Q_i^{(m)}(\mathbf{W}_i)]\}^{1/2} \{E[(W_{ij} - E(W_{ij}))^4]\}^{1/2} \qquad (17) \\
&\quad = \pi_i^2 \{O(e^{-m(\kappa c^{**} - \zeta)}) O(m^4)\}^{1/2} \qquad (18) \\
&\quad = O(m^2 e^{-\frac{m}{2}(\kappa c^{**} - \zeta)}).
\end{aligned}
$$

Notice that (16) follows from the Cauchy–Schwarz inequality, (17) because $0 \le X \le 1$ implies $E(X^2) \le E(X)$, and (18) by Corollary 2 and Lemma 3.

For (b), use Corollary 2 with the expectation (9) to obtain

$$
\begin{aligned}
&\mathbf{e}_j^{\mathrm{T}} (\mathbf{F}_\pi - \mathbf{A}_{\pi\pi}) \mathbf{e}_j \\
&\quad = \pi_j^{-1} E[1 - Q_j^{(m)}(\mathbf{W}_j)] + \pi_s^{-1} E[1 - Q_s^{(m)}(\mathbf{W}_s)] + 2\pi_j^{-1} E[Q_j^{(m)}(\mathbf{W}_s)] \\
&\quad = \pi_j^{-1} O(e^{-m(\kappa c^{**} - \zeta)}) + \pi_s^{-1} O(e^{-m(\kappa c^{**} - \zeta)}) + 2\pi_j^{-1} O(e^{-m(\kappa c^{**} - \zeta)}).
\end{aligned}
$$

The fact that $E[Q_j^{(m)}(\mathbf{W}_s)] = O(e^{-m(\kappa c^{**} - \zeta)})$ follows from

$$
\begin{aligned}
E[Q_j^{(m)}(\mathbf{W}_s)] &= \int \frac{\pi_j f(\mathbf{w} \mid \boldsymbol{\eta}_j)}{f(\mathbf{w})} f(\mathbf{w} \mid \boldsymbol{\eta}_s) \mathrm{d}\nu(\mathbf{w}) \\
&\le \int \sum_{\ell \ne s} \frac{\pi_\ell f(\mathbf{w} \mid \boldsymbol{\eta}_\ell)}{f(\mathbf{w})} f(\mathbf{w} \mid \boldsymbol{\eta}_s) \mathrm{d}\nu(\mathbf{w}) = E[1 - Q_s^{(m)}(\mathbf{W}_s)].
\end{aligned}
$$

□

Because of the convenient block-diagonal form of the complete data FIM, its inverse $\widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta}) = \mathrm{Blockdiag}(\pi_1^{-1}\mathbf{F}_1^{-1}, \ldots, \pi_s^{-1}\mathbf{F}_s^{-1}, \mathbf{F}_\pi^{-1})$ is also block-diagonal. As in Raim et al. (2014a, Theorem 2.5), the convergence result for $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})$ can be used to show convergence between the inverses. This is stated as a theorem, and the proof is left to the Appendix.

Theorem 3 *Suppose $\mathcal{I}_m(\boldsymbol{\theta})$ and $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta})$ are nonsingular. Then $\mathcal{I}_m^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta}) \to \mathbf{0}$ as $m \to \infty$.*

Remark 1 Until this point we have assumed a given $\boldsymbol{\theta}$ to evaluate the information matrix. An anonymous referee notes that, in practice, an unknown $\boldsymbol{\theta}$ will be estimated by some $\hat{\boldsymbol{\theta}}$ based on the data. We can justify $\widetilde{\mathcal{I}}_m^{-1}(\hat{\boldsymbol{\theta}})$ as an estimator for the large sample variance $\mathcal{I}_m^{-1}(\boldsymbol{\theta})$ as follows.

Note that $\mathcal{I}_m^{-1}(\cdot)$ and $\widetilde{\mathcal{I}}_m^{-1}(\cdot)$ represent sequences of functions that vary with m and $\hat{\boldsymbol{\theta}}$ represents a sequence of estimators based on m clustered observations. By the triangle inequality,

$$
\|\widetilde{\mathcal{I}}_m^{-1}(\hat{\boldsymbol{\theta}}) - \mathcal{I}_m^{-1}(\boldsymbol{\theta})\| \le \|\widetilde{\mathcal{I}}_m^{-1}(\hat{\boldsymbol{\theta}}) - \widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta})\| + \|\widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta}) - \mathcal{I}_m^{-1}(\boldsymbol{\theta})\|.
$$

Theorem 3 gives that $\|\widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta}) - \mathcal{I}_m^{-1}(\boldsymbol{\theta})\| \to 0$ as $m \to \infty$. Taking $\|\cdot\|$ to be the Frobenius norm and using a similar decomposition as in the proof of Theorem 3,

$$\|\widetilde{\mathcal{I}}_m^{-1}(\hat{\boldsymbol{\theta}}) - \widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta})\|^2 = \sum_{\ell=1}^{s} m^{-2} \|\hat{\pi}_\ell^{-1} \mathcal{I}_1^{-1}(\hat{\boldsymbol{\eta}}_\ell) - \pi_\ell^{-1} \mathcal{I}_1^{-1}(\boldsymbol{\eta}_\ell)\|^2 + \|[\mathbf{D}_{\hat{\pi}} - \hat{\boldsymbol{\pi}}_{-s} \hat{\boldsymbol{\pi}}_{-s}^{\mathrm{T}}] - [\mathbf{D}_{\pi} - \boldsymbol{\pi}_{-s} \boldsymbol{\pi}_{-s}^{\mathrm{T}}]\|^2,$$

where $\mathcal{I}_1(\boldsymbol{\eta}_\ell)$ represents the exact FIM under the ℓth subpopulation for a single observation. The expressions

$$\|\hat{\pi}_\ell^{-1} \mathcal{I}_1^{-1}(\hat{\boldsymbol{\eta}}_\ell) - \pi_\ell^{-1} \mathcal{I}_1^{-1}(\boldsymbol{\eta}_\ell)\|^2 \quad \text{and} \tag{19}$$

$$\|[\mathbf{D}_{\hat{\pi}} - \hat{\boldsymbol{\pi}}_{-s} \hat{\boldsymbol{\pi}}_{-s}^{\mathrm{T}}] - [\mathbf{D}_{\pi} - \boldsymbol{\pi}_{-s} \boldsymbol{\pi}_{-s}^{\mathrm{T}}]\|^2 \tag{20}$$

do not depend on m except through the estimator $\hat{\boldsymbol{\theta}}$. It is apparent that (20) is a continuous function of $\hat{\boldsymbol{\theta}}$. Continuity of (19) in $\hat{\boldsymbol{\theta}}$ can be verified by obtaining $\mathcal{I}_1^{-1}(\boldsymbol{\eta}_\ell)$ for the distribution under consideration; this is a prerequisite in formulating $\widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta})$. Under continuity and given a consistent estimator $\hat{\boldsymbol{\theta}}$, we have that (19) and (20) are of order $o_p(1)$; therefore, $\|\widetilde{\mathcal{I}}_m^{-1}(\hat{\boldsymbol{\theta}}) - \mathcal{I}_m^{-1}(\boldsymbol{\theta})\| = o_p(1)$.

Neerchal and Morel (2005) and Raim et al. (2014a) have considered the setting of multinomial finite mixtures where the FIM can be computed exactly (e.g. by brute force, summing over the sample space), though it may be time-consuming. When $\widetilde{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}})$ itself does not provide an accurate estimate of the large sample covariance, $\widetilde{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}})$ is suggested as an aid to find the MLE, but $\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})$ is used in the final steps of estimation and to obtain standard errors.

4 Relationship to classification problem

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

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

The objective is to specify a rule $\boldsymbol{\mathcal{D}}$ which minimizes the probability of misclassification $p(\boldsymbol{\mathcal{D}})$. (Alternatively, the objective may be to minimize the cost of

misclassification if the possible misclassifications are assigned different costs.) It is well-known that the rule $\mathcal{D}^* = \{\mathcal{D}_1^*, \ldots, \mathcal{D}_s^*\}$ given by

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

minimizes $p(\mathcal{D})$ (Anderson 2003). Of course the rule $\mathcal{D}^*$ assumes full knowledge of all $f(\mathbf{x} \mid \boldsymbol{\phi}_\ell)$ and π_ℓ, and, therefore, is not directly usable in practice. Under $\mathcal{D}^*$, we may consider the optimal probability of misclassification $p(\mathcal{D}^*)$, which provides a measurement for the degree of mutual separation between the s subpopulations. A larger $p(\mathcal{D}^*)$ indicates that it is more difficult to distinguish among them. We can relate $p(\mathcal{D}^*)$ to the convergence rates obtained in Sect. 3 by

$$\begin{aligned}
p(\mathcal{D}^*) &= \sum_{\ell=1}^{s} P(\mathbf{T} \notin \mathcal{D}_\ell^* \mid Z = \ell)\, P(Z = \ell) \\
&= \sum_{\ell=1}^{s} \pi_\ell\, P\left(\bigcup_{j \neq \ell} [\mathbf{T} \in \mathcal{D}_j^*] \,\middle|\, Z = \ell \right) \\
&= \sum_{\ell=1}^{s} \pi_\ell\, P\left(\bigcup_{j \neq \ell} \left[\pi_j f(\mathbf{T} \mid \boldsymbol{\phi}_j) \geq \pi_\ell f(\mathbf{T} \mid \boldsymbol{\phi}_\ell) \right] \,\middle|\, Z = \ell \right) \\
&\leq \sum_{\ell=1}^{s} \pi_\ell\, P\left(\sum_{j \neq \ell} \pi_j f(\mathbf{T} \mid \boldsymbol{\phi}_j) \geq \pi_\ell f(\mathbf{T} \mid \boldsymbol{\phi}_\ell) \,\middle|\, Z = \ell \right) \\
&= \sum_{\ell=1}^{s} \pi_\ell\, P(Q_\ell^{(m)}(\mathbf{W}_\ell) \leq 1/2). \qquad (21)
\end{aligned}$$

Corollary 2, along with the Markov inequality, can be used to obtain a rate for the RHS of (21),

$$\begin{aligned}
\sum_{\ell=1}^{s} \pi_\ell\, P(Q_\ell^{(m)}(\mathbf{W}_\ell) \leq 1/2) &\leq \sum_{\ell=1}^{s} \pi_\ell 2E[1 - Q_\ell^{(m)}(\mathbf{W}_\ell)] \\
&= O(e^{-m(\kappa c^{**} - \zeta)}).
\end{aligned}$$

The optimal probability of misclassification will, therefore, decrease rapidly to 0 as m increases if the s subpopulations are well-separated.

5 Examples

We now present several examples demonstrating the closeness of $\mathcal{I}(\boldsymbol{\theta})$ and $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ in the mixture setting.

Example 1 (Multinomial Finite Mixture) Let $\mathbf{X}_1, \ldots, \mathbf{X}_m$ be independent and identically distributed as $\text{Mult}_{k+1}(1, \mathbf{p}_Z)$, with $Z \sim \text{Discrete}(1, \ldots, s; \boldsymbol{\pi})$. Take $\mathbf{T} = \sum_{i=1}^{m} \mathbf{X}_i$. The multinomial subpopulations are exponential families with $f(\mathbf{t} \mid m, \mathbf{p}_\ell) = \exp\left\{\boldsymbol{\eta}_\ell^{\mathrm{T}}\mathbf{t} - m\psi(\boldsymbol{\eta}_\ell) + h(\mathbf{t})\right\}$, where

$$\boldsymbol{\eta}_\ell = \left(\log \frac{p_{\ell 1}}{p_{\ell,k+1}}, \cdots, \log \frac{p_{\ell k}}{p_{\ell,k+1}}\right) \quad \text{and} \quad \psi(\boldsymbol{\eta}_\ell) = -\log p_{\ell,k+1},$$

with $p_{\ell,k+1} = 1 - \sum_{a=1}^{k} p_{\ell a}$. The approximate information matrix with respect to $\boldsymbol{\theta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_s, \boldsymbol{\pi}_{-s})$ is then $\widetilde{\mathcal{I}}(\boldsymbol{\theta}) = \text{Blockdiag}(\pi_1\mathbf{F}_1, \ldots, \pi_s\mathbf{F}_s, \mathbf{F}_\pi)$, where $\mathbf{F}_\ell = m\{\text{Diag}(\mathbf{p}_\ell) - \mathbf{p}_\ell\mathbf{p}_\ell^{\mathrm{T}}\}$ and $\mathbf{F}_\pi = \mathbf{D}_\pi^{-1} + \pi_s^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}$. Transforming to $\boldsymbol{\vartheta}(\boldsymbol{\theta}) = (\mathbf{p}_1, \ldots, \mathbf{p}_s, \boldsymbol{\pi}_{-s})$ gives $\partial\boldsymbol{\eta}_\ell/\partial\mathbf{p}_\ell = \text{Diag}(\mathbf{p}_\ell)^{-1} + p_{\ell,k+1}^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}$ so that

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

Therefore, we obtain the form of $\widetilde{\mathcal{I}}(\boldsymbol{\vartheta})$ which was studied in Raim et al. (2014a).

Example 2 (Multivariate Normal Finite Mixture) Let $\mathbf{X}_1, \ldots, \mathbf{X}_m$ be independent and identically distributed in $\mathbb{R}^k$ as $\text{N}(\boldsymbol{\mu}_Z, \boldsymbol{\Sigma})$, with $Z \sim \text{Discrete}(1, \ldots, s; \boldsymbol{\pi})$. Then $\mathbf{T} = \sum_{i=1}^{m} \mathbf{X}_i \sim \text{N}(m\boldsymbol{\mu}_Z, m\boldsymbol{\Sigma})$ given Z. Let us compare the FIM and approximation with respect to $\boldsymbol{\vartheta} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_s, \boldsymbol{\pi}_{-s})$, where $\boldsymbol{\Sigma}$ is taken to be known. The normal subpopulations are exponential families with $f(\mathbf{t} \mid m\boldsymbol{\mu}_j, m\boldsymbol{\Sigma}) = \exp\{\boldsymbol{\eta}_j^{\mathrm{T}}\mathbf{t} - m\psi(\boldsymbol{\eta}_j) + h(\mathbf{t})\}$, where $\boldsymbol{\eta}_j = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j$ and $m\psi(\boldsymbol{\eta}_j) = m\frac{1}{2}\boldsymbol{\eta}_j^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\eta}_j$. Under $Z = j$, the first and second derivative of the log-density with respect to $\boldsymbol{\eta}_j$ are given by

$$\frac{\partial}{\partial\boldsymbol{\eta}_j} \log f(\mathbf{t} \mid \boldsymbol{\eta}_j) = \mathbf{t} - m\boldsymbol{\Sigma}\boldsymbol{\eta}_j \quad \text{and} \quad -\frac{\partial^2}{\partial\boldsymbol{\eta}_j\partial\boldsymbol{\eta}_j^{\mathrm{T}}} \log f(\mathbf{t} \mid \boldsymbol{\eta}_j) = m\boldsymbol{\Sigma}.$$

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

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

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

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

and $\mathbf{F}_\pi = \mathbf{D}_\pi^{-1} + \pi_s^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}$. We now study the closeness between $\mathcal{I}(\boldsymbol{\vartheta})$ and $\widetilde{\mathcal{I}}(\boldsymbol{\vartheta})$ by numerical experiment. The true information matrix is computed using the `cubature`

package[1] in R for numerical multivariate integration. Let us concretely take dimension $k = 2$ and number of populations $s = 2$, with

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\pi} = \begin{pmatrix} 0.25 \\ 0.75 \end{pmatrix}.$$

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

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

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

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

and likewise $\gamma_{221} = -\gamma_{122}$, where γ_{211} and γ_{122} are positive by Lemma 2. We have listed all eight possible γ_{IJK} constants, and it is apparent that γ_{211} and γ_{122} together characterize the convergence rates. We will consider three scenarios for the subpopulation means:

- Scenario 1: $\boldsymbol{\mu}_1 = (-1, 1)$, $\boldsymbol{\mu}_2 = (1, -1)$, so that $\gamma_{221} = \gamma_{122} = 8$.
- Scenario 2: $\boldsymbol{\mu}_1 = (-0.5, 0.5)$, $\boldsymbol{\mu}_2 = (0.5, -0.5)$, so that $\gamma_{221} = \gamma_{122} = 2$.
- Scenario 3: $\boldsymbol{\mu}_1 = (-0.125, 0.125)$, $\boldsymbol{\mu}_2 = (0.125, -0.125)$, so that $\gamma_{221} = \gamma_{122} = 0.125$.

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

Table 1 shows the diagonal elements of $\widetilde{\mathcal{I}}_m(\boldsymbol{\vartheta})$ compared with those of $\mathcal{I}_m(\boldsymbol{\vartheta})$, where the latter have been computed numerically. Also shown is the Frobenius norm of the matrix $\widetilde{\mathcal{I}}_m(\boldsymbol{\vartheta}) - \mathcal{I}_m(\boldsymbol{\vartheta})$. Note from the proof of Theorem 3 in the Appendix that

$$\|\widetilde{\mathcal{I}}_m(\boldsymbol{\vartheta}) - \mathcal{I}_m(\boldsymbol{\vartheta})\|_{\mathrm{F}}^2 = q^2 O(m^4 e^{-m(\kappa c^{**} - \zeta)}).$$

Figure 2 plots the norms for the three scenarios. As expected, the elements of the FIM and the approximation converge together quickly for Scenario 1, and more slowly for Scenario 2. For Scenario 3, the Frobenius norm initially increases with m because of the extremely slow convergence rate, and eventually begins decreasing when m is large.

Example 3 (Sampling iid from Normal Finite Mixture)

It is natural to ask if there is relationship between the information matrix of $\mathbf{X}_1, \ldots, \mathbf{X}_m$ independently and identically distributed from $f(\mathbf{x} \mid \boldsymbol{\phi}_Z)$, but where Z is not observed, and the information matrix of $\mathbf{X}_1, \ldots, \mathbf{X}_m$ independently and identically

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

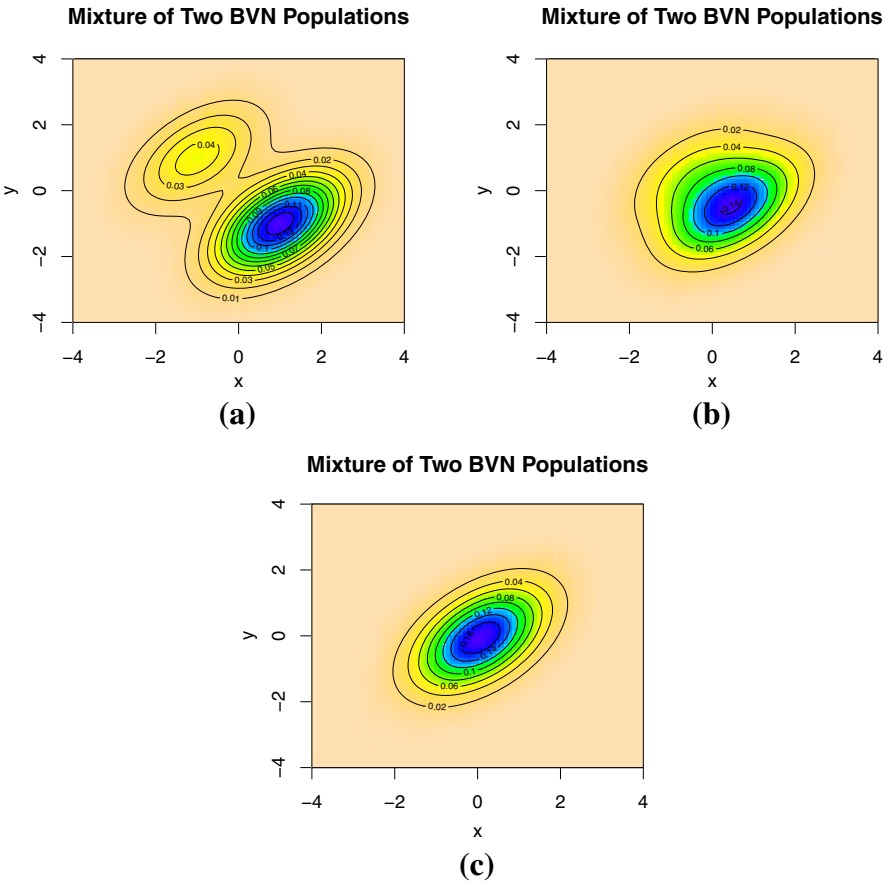

Fig. 1 Densities for the bivariate normal finite mixture under the three scenarios. **a** Scenario 1, **b** Scenario 2, **c** Scenario 3

distributed from the finite mixture $f(\mathbf{x} \mid \boldsymbol{\theta})$. The results in this paper were developed strictly for the former case. As a concrete example, suppose $X_1, \ldots, X_m$ are drawn independently from $\mathrm{N}(\mu_Z, 1)$, where $Z \sim \mathrm{Discrete}(1, \ldots, s; \boldsymbol{\pi})$. Let $\mathcal{I}_m(\boldsymbol{\theta})$ denote the information matrix of $T = \sum_{i=1}^{m} X_i$, where $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_s, \pi_1, \ldots, \pi_{s-1})$ and the density of T is

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

On the other hand, if X_i are drawn iid from the finite mixture

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

Table 1 Results for bivariate normal mixture

m	Entry (1, 1)	Entry (2, 2)	Entry (3, 3)	Entry (4, 4)	Entry (5, 5)	$\|\widetilde{\mathcal{I}} - \mathcal{I}\|_{\mathrm{F}}$
(a) Scenario 1						
1	0.333 (0.276)	0.333 (0.276)	1 (0.921)	1 (0.921)	5.333 (4.921)	0.6459
2	0.667 (0.643)	0.667 (0.643)	2 (1.971)	2 (1.971)	5.333 (5.290)	0.1419
3	1.000 (0.994)	1.000 (0.994)	3 (2.993)	3 (2.993)	5.333 (5.328)	0.0304
4	1.333 (1.332)	1.333 (1.332)	4 (3.999)	4 (3.999)	5.333 (5.333)	0.0060
5	1.667 (1.666)	1.667 (1.666)	5 (5.000)	5 (5.000)	5.333 (5.333)	0.0011
6	2.000 (2.000)	2.000 (1.999)	6 (6.000)	6 (6.000)	5.333 (5.333)	0.0002
(b) Scenario 2						
1	0.333 (0.192)	0.333 (0.192)	1 (0.777)	1 (0.777)	5.333 (2.729)	3.0005
2	0.667 (0.452)	0.667 (0.452)	2 (1.670)	2 (1.670)	5.333 (3.968)	2.1626
3	1.000 (0.761)	1.000 (0.761)	3 (2.653)	3 (2.653)	5.333 (4.592)	1.7011
–	–	–	–	–	–	–
23	7.667 (7.666)	7.667 (7.666)	23 (23.000)	23 (23.000)	5.333 (5.333)	0.0013
24	8.000 (8.000)	8.000 (8.000)	24 (24.000)	24 (24.000)	5.333 (5.333)	0.0009
25	8.333 (8.333)	8.333 (8.333)	25 (25.000)	25 (25.000)	5.333 (5.333)	0.0006
(c) Scenario 3						
1	0.333 (0.100)	0.333 (0.100)	1 (0.746)	1 (0.746)	5.333 (0.245)	5.1939
2	0.667 (0.227)	0.667 (0.227)	2 (1.488)	2 (1.488)	5.333 (0.480)	5.2334
3	1.000 (0.375)	1.000 (0.375)	3 (2.231)	3 (2.231)	5.333 (0.703)	5.3942
–	–	–	–	–	–	–
28	9.333 (6.118)	9.333 (6.118)	28 (22.989)	28 (22.989)	5.333 (3.736)	13.4860
29	9.667 (6.393)	9.667 (6.393)	29 (23.913)	29 (23.913)	5.333 (3.798)	13.7348
30	10.000 (6.670)	10.000 (6.670)	30 (24.844)	30 (24.844)	5.333 (3.857)	13.9718
–	–	–	–	–	–	–
78	26.000 (21.770)	26.000 (22.770)	78 (73.692)	78 (73.692)	5.333 (5.084)	13.7007
79	26.333 (23.138)	26.333 (23.138)	79 (74.748)	79 (74.748)	5.333 (5.093)	13.5528
80	26.667 (23.506)	26.667 (23.506)	80 (75.804)	80 (75.804)	5.333 (5.102)	13.4030

The five "entry" columns show the diagonal elements $\widetilde{\mathcal{I}}_{ii}$ along with corresponding $\mathcal{I}_{ii}$ in parentheses. The last column shows the Frobenius norm of $\widetilde{\mathcal{I}} - \mathcal{I}$

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

- for $m = 3$, $\mathcal{I}_m(\boldsymbol{\theta})$ vs. $m\mathcal{I}_1(\boldsymbol{\theta})$ is

$$\begin{pmatrix} 0.5370 & -0.2023 & -0.3692 \\ -0.2023 & 1.9289 & -0.4653 \\ -0.3692 & -0.4653 & 4.5916 \end{pmatrix} \quad \text{vs.} \quad \begin{pmatrix} 0.4177 & -0.0951 & -1.1399 \\ -0.0951 & 1.6739 & -1.7900 \\ -1.1399 & -1.7900 & 8.1871 \end{pmatrix}.$$

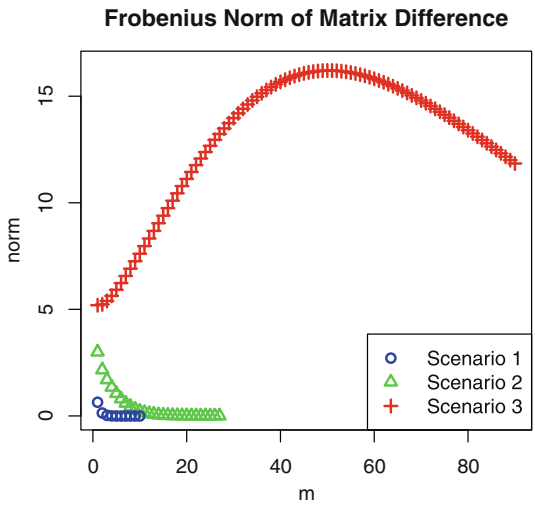

Fig. 2 Frobenius norm of $\widetilde{\mathcal{I}}_m - \mathcal{I}_m$, as m varies, for the three normal scenarios

- for $m = 50$, $\mathcal{I}_m(\boldsymbol{\theta})$ vs. $m\mathcal{I}_1(\boldsymbol{\theta})$ is

$$\begin{pmatrix} 12.5 & 0.0 & 0.0000 \\ 0.0 & 37.5 & 0.0000 \\ 0.0 & 0.0 & 5.3333 \end{pmatrix} \quad \text{vs.} \quad \begin{pmatrix} 6.9612 & -1.5853 & -18.9977 \\ -1.5853 & 27.8981 & -29.8327 \\ -18.9977 & -29.8327 & 136.4524 \end{pmatrix}.$$

It is evident that $m\mathcal{I}_1(\boldsymbol{\theta})$ does not become close to $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta})$, and, therefore, convergence of the complete data FIM may not occur when the sample is not drawn in a clustered manner.

Example 4 (Dirichlet-Multinomial) A Dirichlet-multinomial random variable may be obtained from the marginal distribution of $\mathbf{T}$ when

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

This marginal distribution is a continuous mixture of multinomials. In the special case of $J = 2$ categories, a beta-binomial random variable is obtained. The complete data density of $(\mathbf{T}, \boldsymbol{\mu})$ is

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

Let $k = J - 1$ to ensure the parameter space of the multinomial family contains an open set in $\mathbb{R}^k$. The Dirichlet-multinomial density is

$$f(\mathbf{t} \mid \boldsymbol{\alpha}) = \frac{m!}{t_1! \cdots t_J!} \frac{\prod_{j=1}^{J} \Gamma(\alpha_j + t_j)}{\Gamma(\sum_{j=1}^{J} \alpha_j)} \frac{\Gamma(\sum_{j=1}^{J} \alpha_j + m)}{\prod_{j=1}^{J} \Gamma(\alpha_j)}. \tag{22}$$

Although the results in this paper have been developed for finite mixtures of exponential families and not continuous mixtures, we may consider the complete data information matrix and ask whether it approximates the true information matrix. Note that the distribution of $\mathbf{T} \mid \boldsymbol{\mu}$ is free of $\boldsymbol{\alpha}$ so that $\frac{\partial}{\partial \boldsymbol{\alpha}} \log f(\mathbf{t}, \boldsymbol{\mu} \mid \boldsymbol{\alpha}) = \frac{\partial}{\partial \boldsymbol{\alpha}} \log f(\boldsymbol{\mu} \mid \boldsymbol{\alpha})$; therefore, the complete data information matrix is just the FIM with respect to $\text{Dirichlet}_J(\boldsymbol{\alpha})$. This is an analog to the finite mixture case, where the first s diagonal blocks correspond to the support points of the mixing distribution, $\text{Discrete}(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_s; \boldsymbol{\pi}_{-s})$, and the lower-right block corresponds to $\boldsymbol{\pi}$. Here the mixing process follows a Dirichlet distribution whose support is the probability simplex in $\mathbb{R}^J$, which is known and does not require corresponding entries in the information matrix. Neerchal and Morel (1998, Theorem 1) show that the FIM of $\mathbf{T}$ converges to the FIM of $\text{Dirichlet}_k(\boldsymbol{\alpha})$ as $m \to \infty$. Therefore, the results in this paper may extend beyond the assumption of the latent mixing process following a finite mixture distribution.

Example 5 (Normal–Normal) Let us consider a second continuous mixture along the lines of Example 4. The normal–normal hierarchical model is popular in Bayesian analysis (Gelman et al. 2003, Section 5.4), with one application, for example, in the Fay-Herriot model for small area estimation (Rao 2003). The results from this paper can be applied in the following sense. Suppose

$$\bar{X} \mid \mu \sim \text{N}(\mu, \sigma^2/m), \quad \mu \sim \text{N}(\theta, \tau^2).$$

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

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

Therefore, $\frac{\partial}{\partial \theta} \log f_{\bar{X}}(x \mid \theta) = \frac{\partial}{\partial \theta} \log f_T(t \mid \theta)$, and the information is the same whether we work with $\bar{X}$ or T. It can be shown that marginally, $\bar{X} \sim \text{N}\left(\mu, \sigma^2/m + \tau^2\right)$; therefore, the true information about θ in $\bar{X}$ is $\mathcal{I}_m(\theta) = (\sigma^2/m + \tau^2)^{-1}$. The complete data information about θ in $(\bar{X}, \mu)$ is $\widetilde{\mathcal{I}}(\theta) = \tau^{-2}$. Now we have convenient forms for both the true information and complete data information, and it is clear that $\mathcal{I}_m(\theta) \to \widetilde{\mathcal{I}}(\theta)$ as $m \to \infty$. Note that the right-hand side of the limit is fixed, as in Example 4.

Example 6 (Mixture of Finite Mixtures) Consider the random-clumped binomial (RCB) distribution introduced in Morel and Nagaraj (1993) to model binomial data with extra variation. An RCB random variable T can be written as $T = NY + (X \mid N)$, where

$$Y \sim \text{Ber}(\pi), \quad N \sim \text{Bin}(m, \rho), \quad (X \mid N) \sim \text{Bin}(m - N, \pi),$$

so that N of the m trials mimic the outcome in Y, and the remaining trials are drawn independently for X. The RCB density can be expressed as the finite mixture of two binomial densities, $\text{RCB}(t \mid m, \rho, \pi) = \pi \text{Bin}(t \mid m, \xi_1) + (1 - \pi)\text{Bin}(t \mid m, \xi_2)$, where $\xi_1 = (1 - \rho)\pi + \rho$ and $\xi_2 = (1 - \rho)\pi$. Consider now a finite mixture of RCB densities

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

where $\boldsymbol{\vartheta} = (\rho_1, \pi_1, \ldots, \rho_s, \pi_s, w_1, \ldots, w_{s-1})$. This does not immediately appear to be an exponential family finite mixture; however, the density may be rewritten as a binomial finite mixture

$$f(t \mid m, \boldsymbol{\vartheta}) = \sum_{\ell=1}^{s} w_\ell \sum_{j=1}^{2} \pi_{\ell j} \text{Bin}(t \mid m, \xi_\ell) = \sum_{\ell=1}^{2s} \lambda_\ell \text{Bin}(t \mid m, \xi_\ell),$$

where

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

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

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

with respect to $\boldsymbol{\theta} = (\xi_1, \ldots, \xi_{2s}, \lambda_1, \ldots, \lambda_{2s-1})$, and then using the Jacobian of the transformation $\boldsymbol{\vartheta} \mapsto \boldsymbol{\theta}$ to obtain

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

The convergence of $\widetilde{\mathcal{I}}_m(\boldsymbol{\vartheta}) - \mathcal{I}_m(\boldsymbol{\vartheta})$ to zero follows from Theorem 2.

Example 7 (Weibull Finite Mixture) Consider the Weibull density

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

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

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

and the Fisher information is, therefore, found by computing

$$\mathcal{I}(\beta) = \int_0^{\infty} \left\{\frac{1}{\beta} - \left[1 - \left(\frac{x}{\lambda}\right)^{\beta}\right] \log\left(\frac{x}{\lambda}\right)\right\}^2 f(x \mid \beta, \lambda) \mathrm{d}x. \tag{23}$$

Although the results developed in this paper do not apply because of the departure from exponential family, we will proceed to investigate the convergence of the approximate information. Suppose $\mathbf{X} = (X_1, \ldots, X_m)$ given Z are a random sample from Weibull(β_Z, λ_Z) and $Z \sim \text{Discrete}(1, \ldots, s; \boldsymbol{\pi})$. The marginal density of $\mathbf{X}$ is then given by

$$f(\mathbf{x} \mid \boldsymbol{\theta}) = \sum_{\ell=1}^{s} \pi_\ell \left[\left(\frac{\beta_\ell}{\lambda_\ell}\right)^m \left(\prod_{i=1}^{m} \frac{x_i}{\lambda_\ell}\right)^{\beta_\ell - 1} \exp\left\{-\sum_{i=1}^{m} (x_i/\lambda_\ell)^{\beta_\ell}\right\}\right], \tag{24}$$

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

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

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

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

for $a = 1, \ldots, s - 1$. The approximate information matrix is given by

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

where F_ℓ is given by multiplying the Weibull$(\beta_\ell, \lambda_\ell)$ information (23) by m, and $\mathbf{F}_\pi = \mathbf{D}_\pi^{-1} + \pi_s^{-1} \mathbf{1}\mathbf{1}^{\mathrm{T}}$ as usual for finite mixtures.

Consider a numerical study comparing $\mathcal{I}(\boldsymbol{\theta})$ and $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ under the density

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

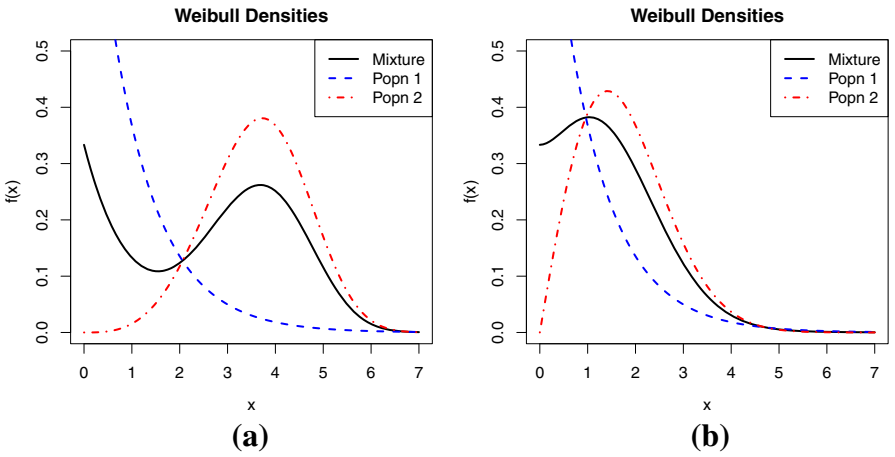

Fig. 3 Densities for the Weibull finite mixture under the two scenarios. **a** Scenario 1, **b** Scenario 2

Table 2 Results for Weibull mixture

m	Entry (1, 1)	Entry (2, 2)	Entry (3, 3)	$\|\widetilde{\mathcal{I}} - \mathcal{I}\|_{\mathrm{F}}$
(a) Scenario 1				
1	0.6079 (0.3787)	0.0760 (0.0535)	4.5 (3.2304)	1.3201
2	1.2158 (1.0521)	0.1520 (0.1279)	4.5 (4.0346)	0.5472
3	1.8237 (1.7571)	0.2280 (0.2112)	4.5 (4.3218)	0.2397
4	2.4316 (2.3626)	0.3039 (0.2926)	4.5 (4.4237)	0.1256
5	3.0395 (2.9479)	0.3799 (0.3772)	4.5 (4.4805)	0.1122
6	3.6474 (3.5409)	0.4559 (0.4494)	4.5 (4.4914)	0.1097
7	4.2553 (4.3264)	0.5319 (0.5281)	4.5 (4.5106)	0.0729
8	4.8632 (4.9649)	0.6079 (0.6077)	4.5 (4.4984)	0.1082
9	5.4711 (5.4920)	0.6839 (0.6854)	4.5 (4.5032)	0.0257
10	6.0790 (6.0419)	0.7599 (0.7637)	4.5 (4.5010)	0.0404
(b) Scenario 2				
1	0.6079 (0.3919)	0.3039 (0.1696)	4.5 (1.0642)	3.4731
2	1.2158 (0.8718)	0.6079 (0.3840)	4.5 (1.7997)	2.8164
3	1.8237 (1.3980)	0.9118 (0.6135)	4.5 (2.3182)	2.3894
4	2.4316 (1.9380)	1.2158 (0.8703)	4.5 (2.7546)	2.0388
5	3.0395 (2.5468)	1.5197 (1.1423)	4.5 (3.0743)	1.7982
–	–	–	–	–
23	13.9816 (13.7489)	6.9908 (6.8029)	4.5 (4.4462)	0.3482
24	14.5895 (14.5347)	7.2947 (7.1399)	4.5 (4.4513)	0.2575
25	15.1974 (15.0696)	7.5987 (7.5052)	4.5 (4.4704)	0.2163
26	15.8053 (15.9109)	7.9026 (7.8191)	4.5 (4.4645)	0.1920
27	16.4132 (16.3579)	8.2066 (8.1740)	4.5 (4.4682)	0.1320

The three "entry" columns show the diagonal elements $\widetilde{\mathcal{I}}_{ii}$ along with corresponding $\mathcal{I}_{ii}$ in parentheses. The last column shows the Frobenius norm of $\widetilde{\mathcal{I}} - \mathcal{I}$

for the following two scenarios:

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

Figure 3 plots the subpopulations and mixed population for each scenario. Table 2 compares the approximate and true information matrices in these scenarios. Evaluation of (23) for the approximate FIM is carried out by numerical integration. The true FIM is computed by basic Monte Carlo simulation using 100,000 draws. While a more sophisticated method could be used to improve the accuracy, there is clear evidence of the convergence in Table 2. As expected, the rate is faster in Scenario 1 where the subpopulations are further apart.

6 Conclusions

This paper extended Raim et al. (2014a) from multinomial finite mixtures to the more general class of exponential family finite mixtures, making the work relevant to statistical analysis beyond binomial and multinomial data. The main convergence result showed that the true FIM and complete data FIM become close as the number of observations m becomes large, provided that the observations are drawn according to the clustered sampling scheme. This justifies the use of the complete data FIM as an approximation to the true FIM. Rates of convergence were seen to be exponential, but the exponent depends on both m and the similarity between subpopulations. Example 3 suggests that the complete data FIM does not become close to the information matrix of an independent and identically distributed sample of size m drawn from the finite mixture.

There are several interesting questions to consider at this point. The setting of exponential family finite mixtures covers many cases that may be useful in application. Our convergence proof relies on this setting; for example, the $R_i(\cdot)$ and $Q_i(\cdot)$ functions are critical to the proof. However, Examples 4 and 5 provide evidence of the convergence even when the latent mixing process has a continuous distribution. Example 7 shows the convergence in a Weibull finite mixture which does not meet the exponential family assumption. These examples suggest that the convergence result can be generalized further. It would also be of interest to establish a reliable and easily computable method to improve accuracy of the approximation when m is not large or the subpopulations are not well-separated.

7 Appendix: additional results

Lemma 4 and Proposition 4 below are needed for Corollary 2. These results generalize an argument used in Raim et al. (2014a). Lemma 4 gives bounds for tail probabilities involving a linear transformation of a random variable in a natural exponential family. The result is similar in spirit to Okamoto (1959), which specifically considers the binomial distribution. Subsequently, Proposition 4 obtains an upper bound for $E[1 - Q_i^{(m)}(\mathbf{W}_i)]$.

Lemma 4 *Suppose* $\mathbf{U}_1, \ldots, \mathbf{U}_m$ *are independent and identically distributed copies of* $\mathbf{U}$ *with natural exponential family density* $f(\mathbf{u} \mid \boldsymbol{\eta}) = \exp\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u} - \psi(\boldsymbol{\eta})\}$*, and let* $\mathbf{T} = \mathbf{U}_1 + \cdots + \mathbf{U}_m$ *and* $\boldsymbol{\alpha} \in \mathbb{R}^k$*. For any* $c > 0$*,*

(a) *there exist* $\lambda > 0$ *and* $\delta > 0$ *such that*

$$P[\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{T}/m - E(\mathbf{U})) \geq c] \leq e^{-m[\lambda c - \delta]}, \quad \textit{with } \lambda c > \delta.$$

(b) *there exist* $\lambda > 0$ *and* $\delta > 0$ *such that*

$$P[\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{T}/m - E(\mathbf{U})) \leq -c] \leq e^{-m[\lambda c - \delta]}, \quad \textit{with } \lambda c > \delta.$$

Proof Recall that $E(\mathbf{U}) = \psi'(\boldsymbol{\eta})$, $\mathrm{Var}(\mathbf{U}) = \psi''(\boldsymbol{\eta})$, and the MGF $E(e^{\boldsymbol{\tau}^{\mathrm{T}}\mathbf{U}}) = \exp\{\psi(\boldsymbol{\eta} + \boldsymbol{\tau}) - \psi(\boldsymbol{\eta})\}$ exists for all $\boldsymbol{\tau}$ in some ball $B(\mathbf{0}, \varepsilon)$ of radius $\varepsilon > 0$ centered around $\mathbf{0}$. If $\mathrm{Var}(\mathbf{U})$ is strictly positive definite for all $\boldsymbol{\eta} \in \Xi$, Lemma 2 gives

$$\psi(\boldsymbol{\eta} + \boldsymbol{\tau}) - \psi(\boldsymbol{\eta}) > \boldsymbol{\tau}^{\mathrm{T}}\psi'(\boldsymbol{\eta}). \quad \text{when } \boldsymbol{\eta} + \boldsymbol{\tau}, \boldsymbol{\eta} \in \Xi. \tag{25}$$

Let $\lambda > 0$ so that $\lambda\boldsymbol{\alpha} \in B(\mathbf{0}, \varepsilon)$; we have

$$\begin{aligned} P[\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{T}/m - E(\mathbf{U})) \geq c] &= P[e^{\lambda\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{T} - mE(\mathbf{U}))} \geq e^{\lambda mc}] \\ &\leq e^{-\lambda mc} E\{e^{\lambda\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{T} - mE(\mathbf{U}))}\} \\ &= e^{-\lambda mc} \exp\{m[\psi(\boldsymbol{\eta} + \lambda\boldsymbol{\alpha}) - \psi(\boldsymbol{\eta}) - \lambda\boldsymbol{\alpha}^{\mathrm{T}}\psi'(\boldsymbol{\eta})]\}. \end{aligned} \tag{26}$$

Define $p(\boldsymbol{\eta}, \boldsymbol{\alpha}, \lambda) = \psi(\boldsymbol{\eta} + \lambda\boldsymbol{\alpha}) - \psi(\boldsymbol{\eta}) - \lambda\boldsymbol{\alpha}^{\mathrm{T}}\psi'(\boldsymbol{\eta})$, which is positive for all λ from (25) when $\boldsymbol{\eta} + \lambda\boldsymbol{\alpha} \in \Xi$ and $\boldsymbol{\eta} \in \Xi$. Also, $p(\boldsymbol{\eta}, \boldsymbol{\alpha}, \lambda) \to 0$ as $\lambda \downarrow 0$. Then (26) becomes

$$P[\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{T}/m - E(\mathbf{U})) \geq c] \leq e^{-m[\lambda c - p(\boldsymbol{\eta}, \boldsymbol{\alpha}, \lambda)]}.$$

To obtain a useful upper bound, define $g(\lambda) = \lambda c - p(\boldsymbol{\eta}, \boldsymbol{\alpha}, \lambda)$; our goal is to find λ with $g(\lambda) > 0$. We have $g'(\lambda) = c - \boldsymbol{\alpha}^{\mathrm{T}}\psi'(\boldsymbol{\eta} + \lambda\boldsymbol{\alpha}) + \boldsymbol{\alpha}^{\mathrm{T}}\psi'(\boldsymbol{\eta})$ and $g'(0) = c > 0$. Notice that $g'(\lambda) \to c$ as $\lambda \to 0$; then for any $g^* > 0$ there exists a $\lambda^* > 0$ so that

$$\lambda \in B(0, \lambda^*) \implies g'(\lambda) \in B(c, g^*).$$

Let $g^* > 0$ be chosen so that $B(c, g^*)$ does not contain 0 and therefore contains only positive numbers. Then $g'(\lambda) > 0$ for all $\lambda \in (0, \lambda^*)$, which implies that $g(\lambda)$ is increasing on this interval and, therefore, $g(\lambda) > 0$. A satisfactory λ can now be found in the set $(0, \lambda^*) \cap \{\lambda : \lambda\boldsymbol{\alpha} \in B(\mathbf{0}, \varepsilon)\}$. With this choice of λ, let $\delta = p(\boldsymbol{\eta}, \boldsymbol{\alpha}, \lambda)$ to obtain part (a) of the result.

To obtain a bound for the probability of the lower tail, take $\lambda < 0$ such that $\lambda\boldsymbol{\alpha} \in B(\mathbf{0}, \varepsilon)$, and note that

$$\begin{aligned} P[\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{T}/m - E(\mathbf{U})) \leq -c] &= P[e^{\lambda\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{T}-mE(\mathbf{U}))} \geq e^{-\lambda mc}] \\ &\leq e^{\lambda mc} E\{e^{\lambda\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{T}-mE(\mathbf{U}))}\} \\ &= e^{-m[-\lambda c - p(\boldsymbol{\eta},\boldsymbol{\alpha},\lambda)]}. \end{aligned}$$

A similar argument as above gives a $\delta > 0$ to satisfy part (b) of the result. □

Proposition 4 *There exists a* $\kappa > 0$ *and* $\zeta > 0$ *such that*

$$E[1 - Q_i^{(m)}(\mathbf{W}_i)] \leq \frac{1}{\pi_i} \sum_{j \neq i}^{s} e^{-m(\frac{\kappa}{2}[\gamma_{jii} + \gamma_{ijj}] - \zeta)}.$$

Proof We have

$$E[1 - Q_i^{(m)}(\mathbf{W}_i)] = \int \sum_{j \neq i}^{s} \frac{\pi_j f(\mathbf{w} \mid \boldsymbol{\eta}_j)}{f(\mathbf{w})} f(\mathbf{w} \mid \boldsymbol{\eta}_i) \mathrm{d}\nu(\mathbf{w}).$$

Let $\boldsymbol{\alpha} \in \mathbb{R}^k$ and $\beta \in \mathbb{R}$, to be determined explicitly. For a particular $j \in \{1, \ldots, s\} \backslash \{i\}$, we obtain

$$\begin{aligned} &\int \frac{\pi_j f(\mathbf{w} \mid \boldsymbol{\eta}_j)}{f(\mathbf{w})} f(\mathbf{w} \mid \boldsymbol{\eta}_i) \mathrm{d}\nu(\mathbf{w}) \\ &= \int_{[\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{w} \leq \beta]} \frac{\pi_j f(\mathbf{w} \mid \boldsymbol{\eta}_j)}{f(\mathbf{w})} f(\mathbf{w} \mid \boldsymbol{\eta}_i) \mathrm{d}\nu(\mathbf{w}) + \int_{[\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{w} > \beta]} \frac{\pi_j f(\mathbf{w} \mid \boldsymbol{\eta}_j)}{f(\mathbf{w})} f(\mathbf{w} \mid \boldsymbol{\eta}_i) \mathrm{d}\nu(\mathbf{w}) \\ &\leq \int_{[\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{w} \leq \beta]} f(\mathbf{w} \mid \boldsymbol{\eta}_i) \mathrm{d}\nu(\mathbf{w}) + \int_{[\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{w} > \beta]} \frac{\pi_j}{\pi_i} f(\mathbf{w} \mid \boldsymbol{\eta}_j) \mathrm{d}\nu(\mathbf{w}) \\ &= P[\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{W}_i \leq \beta] + \frac{\pi_j}{\pi_i} P[\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{W}_j > \beta]. \end{aligned}$$

Let us select β so that

$$m[\boldsymbol{\alpha}^{\mathrm{T}} E(\mathbf{W}_i) - c] = \beta \quad \text{and} \quad m[\boldsymbol{\alpha}^{\mathrm{T}} E(\mathbf{W}_j) + c] = \beta,$$

which implies that $\beta = \frac{m}{2}\{\boldsymbol{\alpha}^{\mathrm{T}} E(\mathbf{W}_i) + \boldsymbol{\alpha}^{\mathrm{T}} E(\mathbf{W}_j)\}$ and $2c = \boldsymbol{\alpha}^{\mathrm{T}} E(\mathbf{W}_i) - \boldsymbol{\alpha}^{\mathrm{T}} E(\mathbf{W}_j)$. Let us make the selection $\boldsymbol{\alpha} = -(\boldsymbol{\eta}_j - \boldsymbol{\eta}_i)$ so that

$$\begin{aligned} 2c &= -(\boldsymbol{\eta}_j - \boldsymbol{\eta}_i)^{\mathrm{T}} E(\mathbf{W}_i) + -(\boldsymbol{\eta}_i - \boldsymbol{\eta}_j)^{\mathrm{T}} E(\mathbf{W}_j) \\ &= -(\boldsymbol{\eta}_j - \boldsymbol{\eta}_i)^{\mathrm{T}} E(\mathbf{W}_i) + [\psi(\boldsymbol{\eta}_j) - \psi(\boldsymbol{\eta}_i)] + -(\boldsymbol{\eta}_i - \boldsymbol{\eta}_j)^{\mathrm{T}} E(\mathbf{W}_j) \\ &\quad + [\psi(\boldsymbol{\eta}_i) - \psi(\boldsymbol{\eta}_j)] \\ &= \gamma_{jii} + \gamma_{ijj}, \end{aligned}$$

which is positive by Lemma 2. Now by Lemma 4, there is a $\tau > 0$ and $\delta > 0$ such that

$$\begin{aligned} \mathrm{P}[\boldsymbol{\alpha}^T \boldsymbol{W}_i \leq \beta] &= \mathrm{P}\left[\boldsymbol{\alpha}^T \boldsymbol{W}_i \leq m[\boldsymbol{\alpha}^T \mathrm{E}(\boldsymbol{W}_i) - c]\right] \\ &= P[\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{W}_i/m - E(\mathbf{W}_i)) \leq -[\gamma_{jii} + \gamma_{ijj}]/2] \\ &\leq e^{-m(\tau[\gamma_{jii}+\gamma_{ijj}]/2-\delta)} \end{aligned}$$

and similarly a $\lambda > 0$ and $\rho > 0$ such that

$$P[\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{W}_j > \beta] \leq e^{-m(\lambda[\gamma_{jii}+\gamma_{ijj}]/2-\rho)},$$

where $\tau[\gamma_{jii} + \gamma_{ijj}]/2 - \delta > 0$ and $\lambda[\gamma_{jii} + \gamma_{ijj}]/2 - \rho > 0$. Define κ and ζ so that

$$\kappa[\gamma_{jii} + \gamma_{ijj}]/2 - \zeta = \min\{\tau[\gamma_{jii} + \gamma_{ijj}]/2 - \delta, \lambda[\gamma_{jii} + \gamma_{ijj}]/2 - \rho\},$$

and, therefore,

$$P[\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{W}_i \leq \beta] + \frac{\pi_j}{\pi_i} P[\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{W}_j > \beta] \leq \left(\frac{\pi_j}{\pi_i} + 1\right) e^{-m(\kappa[\gamma_{jii}+\gamma_{ijj}]/2-\zeta)}.$$

Now we have

$$\begin{aligned} E[1 - Q_i^{(m)}(\mathbf{W}_i)] &= \int \sum_{j\neq i}^{s} \frac{\pi_j f(\mathbf{w} \mid \boldsymbol{\eta}_j)}{f(\mathbf{w})} f(\mathbf{w} \mid \boldsymbol{\eta}_i) \mathrm{d}\nu(\mathbf{w}) \\ &\leq \sum_{j\neq i}^{s} \left(\frac{\pi_j + \pi_i}{\pi_i}\right) e^{-m(\kappa[\gamma_{jii}+\gamma_{ijj}]/2-\zeta)} \\ &\leq \frac{1}{\pi_i} \sum_{j\neq i}^{s} e^{-m(\frac{\kappa}{2}[\gamma_{jii}+\gamma_{ijj}]-\zeta)}. \end{aligned}$$

□

Lemma 5 *Suppose* $\mathbf{A}$ *and* $\mathbf{B}$ *are* $q \times q$ *nonsingular matrices. Then* $\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{B}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{A}^{-1}$.

Lemma 6 *Suppose* $\mathbf{A}, \mathbf{B}$ *are* $q \times q$ *symmetric positive definite matrices, and* $\mathbf{B} - \mathbf{A}$ *is positive definite. Then* $\mathbf{A}^{-1} - \mathbf{B}^{-1}$ *is positive definite.*

Proof By Lemma 5, $\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{B}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{A}^{-1}$. Suppose λ is an eigenvalue of $\mathbf{A}^{-1} - \mathbf{B}^{-1}$, then

$$\begin{aligned} &\det(\mathbf{B}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{A}^{-1} - \lambda I) = 0 \\ &\iff \det(\mathbf{B}^{-1/2}(\mathbf{B} - \mathbf{A})^{1/2}\mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})^{1/2}\mathbf{B}^{-1/2} - \lambda I) = 0. \end{aligned}$$

Therefore, $\mathbf{A}^{-1} - \mathbf{B}^{-1}$ and $\mathbf{B}^{-1/2}(\mathbf{B} - \mathbf{A})^{1/2}\mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})^{1/2}\mathbf{B}^{-1/2}$ have the same eigenvalues. Since the latter is symmetric positive definite, all eigenvalues are positive and the result follows. □

The following proof of Theorem 3 follows a similar argument to that of Raim et al. (2014a, Theorem 2.5) but is included in its entirety for completeness.

Proof (Theorem 3) Lemma 5 gives $\mathcal{I}^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta}) = \mathcal{I}^{-1}(\boldsymbol{\theta})[\mathcal{I}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}(\boldsymbol{\theta})]\widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta})$ for any $\boldsymbol{\theta} \in \Theta$. For any matrix norm,

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

therefore, it is sufficient to show that the RHS converges to 0 as $m \to \infty$. To do this, we will consider the three terms separately. Note that for a $q \times q$ matrix $\mathbf{A}$, the matrix 2-norm $\|\cdot\|_2$ and the Frobenius norm $\|\cdot\|_{\mathrm{F}}$ are related by $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_{\mathrm{F}} \leq \sqrt{q}\|\mathbf{A}\|_2$. Therefore, in showing the convergence of $\|\mathbf{A}_m\|$ to zero, we may consider whichever norm is more convenient.

Recalling that $\|\mathbf{A}\|_{\mathrm{F}}^2 = \sum_i \sum_j a_{ij}^2$, Proposition 1 and Theorem 2 give the simple bound

$$\|\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})\|_{\mathrm{F}}^2 = q^2 O(m^4 e^{-m(\kappa c^{**} - \zeta)}).$$

Next, we have

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

where $\widetilde{\mathcal{I}}_1(\boldsymbol{\eta}_\ell) = \mathrm{Var}(\mathbf{U}_1 \mid Z = \ell)$ is free of m.

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

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

Notice that

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

since both LHS and RHS are lower bounds for $\mathbf{x}^{\mathrm{T}}[\mathcal{I}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}(\boldsymbol{\theta})]\mathbf{x} + \mathbf{x}^{\mathrm{T}}\widetilde{\mathcal{I}}(\boldsymbol{\theta})\mathbf{x}$, and the RHS is the greatest such bound. Therefore, denoting the eigenvalues of $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ as $\widetilde{\lambda}_1(m) \geq \cdots \geq \widetilde{\lambda}_q(m) > 0$ and the eigenvalues of $\mathcal{I}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}(\boldsymbol{\theta})$ as $0 \geq \beta_1(m) \geq \cdots \geq \beta_q(m)$,

$$1/\lambda_q(m) \leq \frac{1}{\min\limits_{\|\mathbf{x}\|=1} \mathbf{x}^{\mathrm{T}}[\mathcal{I}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}(\boldsymbol{\theta})]\mathbf{x} + \min\limits_{\|\mathbf{x}\|=1} \mathbf{x}^{\mathrm{T}}\widetilde{\mathcal{I}}(\boldsymbol{\theta})\mathbf{x}} = \frac{1}{\beta_q(m) + \widetilde{\lambda}_q(m)}.$$

The mapping from a matrix to its eigenvalues is a continuous function of its elements (Meyer 2001, Chapter 7); therefore, $\mathcal{I}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}(\boldsymbol{\theta}) \to \mathbf{0}$ as $m \to \infty$ implies that $\beta_q(m) \to 0$.

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

$$\frac{1}{\beta_q(m) + \widetilde{\lambda}_q(m)} \leq \frac{1}{\widetilde{\lambda}_q(m) - \varepsilon} \tag{27}$$

for all $m \geq m_0$. Because $1/\widetilde{\lambda}_q(m) = \|\widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta})\| = O(1)$, there exists a $K > 0$ such that $1/\widetilde{\lambda}_q(m) \leq K$. 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 (27) 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 $m \geq m_0$.

We now have

$$\begin{aligned}
&\|\mathcal{I}_m^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta})\|_{\mathrm{F}} \\
&\quad \leq O(1) \cdot \{\|\mathbf{D}_\pi - \boldsymbol{\pi}_{-s}\boldsymbol{\pi}_{-s}^{\mathrm{T}}\|_{\mathrm{F}}^2 + O(m^{-2})\} \cdot \{q^2 O(m^4 e^{-m(\kappa c^{**} - \zeta)})\}^{1/2},
\end{aligned}$$

which gives the result. □

Acknowledgements We thank Professors Thomas Mathew, Yi Huang, and Yaakov Malinovsky at the University of Maryland, Baltimore County for helpful discussion in preparing the manuscript. Computational resources were provided by the High Performance Computing Facility (http://www.umbc.edu/hpcf) at the university. The first author thanks the facility for financial support as an RA. We additionally thank the editor and two anonymous referees for comments which helped us to significantly improve the paper.

References

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). Hoboken: Wiley.

Blischke, W. R. (1962). Moment estimators for the parameters of a mixture of two binomial distributions. *The Annals of Mathematical Statistics*, *33*(2), 444–454.

Blischke, W. R. (1964). Estimating the parameters of mixtures of binomial distributions. *Journal of the American Statistical Association*, *59*(306), 510–528.

Boldea, O., Magnus, J. R. (2009). Maximum likelihood estimation of the multivariate normal mixture model. *Journal of the American Statistical Association*, *104*(488), 1539–1549.

Boyd, S., Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). Boca Raton: Chapman and Hall/CRC.

Lehmann, E. L., Casella, G. (1998). *Theory of point estimation* (2nd ed.). New York: Springer.

Lehmann, E. L., Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). New York: Springer.

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

McLachlan, G., Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Meyer, C. D. (2001). *Matrix analysis and applied linear algebra*. Philadelphia: Society for Industrial and Applied Mathematics.

Morel, J. G., Nagaraj, N. K. (1991). *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.

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

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

Neerchal, N. K., Morel, J. G. (2005). An improved method for the computation of maximum likelihood estimates for multinomial overdispersion models. *Computational Statistics & Data Analysis*, *49*(1), 33–43.

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

Orchard, T., Woodbury, M. A. (1972). A missing information principle: Theory and applications. In: L. M. Le Cam, J. Neyman, E. L. Scott (Eds.), *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Theory of Statistics* (Vol. 1, pp. 697–715). Berkeley: University of California Press.

Raim, A. M., Liu, M., Neerchal, N. K., Morel, J. G. (2014a). On the method of approximate Fisher scoring for finite mixtures of multinomials. *Statistical Methodology*, *18*, 115–130.

Raim, A. M., Neerchal, N. K., Morel, J. G. (2014b) Large cluster approximation to the finite mixture information matrix with an application to meta-analysis. In *JSM Proceedings, Statistical Computing Section*. Alexandria: American Statistical Association, pp. 4025–4037.

Rao, J. N. K. (2003). *Small area estimation*. Hoboken, NJ: Wiley.

Shao, J. (2008). *Mathematical statistics* (2nd ed.). New York: Springer.