An Approximation to the Information Matrix of Exponential Family Finite Mixtures

Andrew M. Raim*, Nagaraj K. Neerchal* & Jorge G. Morel†

*Department of Mathematics and Statistics, University of Maryland, Baltimore County
†Biometrics and Statistical Sciences Department, Procter & Gamble Company

Abstract

A simple closed form for the Fisher information matrix (FIM) usually cannot be obtained under finite mixtures. Under binomial and multinomial finite mixtures, several authors have considered a certain block-diagonal approximation to the FIM. The approximation has been used in scoring iterations and in demonstrating asymptotic relative efficiency of proposed estimators. Raim et al (2014, Statistical Methodology 18:115–130) have shown 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. Binomial and multinomial mixtures feature a number of trials, which, when taken to infinity, result in the FIM and the approximation becoming arbitrarily close. This work considers a certain clustered sampling scheme that 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 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., 2014), 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 is the subpopulation indicator. 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 was 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., 2014) 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 exploits 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) reviews 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 this work, we consider the FIM to be a quantity of interest in its own right, but mention that it may be preferred because properties of the observed information, such as invertibility, depend on 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, Section 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 $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ from the ℓth subpopulation, where $\boldsymbol{X}_j$ are d-dimensional random variables. We will suppose an exponential family density for $\boldsymbol{X}_i$ in the form

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

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

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

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

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

so that unconditionally,

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

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

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

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

$$f(\boldsymbol{t} \mid \boldsymbol{\theta}) = \sum_{\ell=1}^{s} \pi_\ell \exp\left\{\boldsymbol{\eta}_\ell^T \boldsymbol{t} + ma(\boldsymbol{\eta}_\ell)\right\} \tag{2.1}$$

with respect to the same dominating measure. We will use the notation Ω to refer to the abstract sample space with a typical element ω, and $\mathcal{T}$ to refer to the space of $\boldsymbol{T}(\omega)$. The score vectors can be obtained by noting that $\log f(\boldsymbol{t} \mid \boldsymbol{\eta}) = \boldsymbol{\eta}^T \boldsymbol{t} + ma(\boldsymbol{\eta})$ and $\frac{\partial}{\partial \boldsymbol{\eta}} \log f(\boldsymbol{t} \mid \boldsymbol{\eta}) = \boldsymbol{t} - \mathrm{E}(\boldsymbol{T})$, therefore

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\eta}_\ell} \log f(\boldsymbol{t} \mid \boldsymbol{\theta}) &= \frac{\pi_\ell f(\boldsymbol{t} \mid \boldsymbol{\eta}_\ell)}{f(\boldsymbol{t} \mid \boldsymbol{\theta})} \left[\boldsymbol{t} - \mathrm{E}(\boldsymbol{T} \mid Z = \ell)\right], \quad \text{for } \ell = 1, \ldots, s \\
\frac{\partial}{\partial \pi_\ell} \log f(\boldsymbol{t} \mid \boldsymbol{\theta}) &= \frac{f(\boldsymbol{t} \mid \boldsymbol{\eta}_\ell) - f(\boldsymbol{t} \mid \boldsymbol{\eta}_s)}{f(\boldsymbol{t} \mid \boldsymbol{\theta})}.
\end{aligned}$$

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

$$\mathrm{E}\left\{-\frac{\partial^2}{\partial \boldsymbol{\eta}_\ell \partial \boldsymbol{\eta}_\ell^T} \log f(\boldsymbol{t} \mid \boldsymbol{\eta}_\ell)\right\} = \mathrm{Var}(\boldsymbol{W}_\ell) = m\{\mathrm{Var}(\boldsymbol{U}_1 \mid Z = \ell)\}. \tag{2.2}$$

Denote $\mathcal{I}(\boldsymbol{\theta})$ as the FIM of $\boldsymbol{T}$ under the finite mixture and $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ as the FIM of the complete data $(\boldsymbol{T}, Z)$, both with respect to the parameter $\boldsymbol{\theta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_s, \boldsymbol{\pi})$. Let $q = sk + s - 1$ denote the dimension of $\boldsymbol{\theta}$ 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

$$\begin{aligned}
\widetilde{\mathcal{I}}(\boldsymbol{\theta}) &= \mathrm{Blockdiag}\left(\pi_1 \boldsymbol{F}_1, \ldots, \pi_s \boldsymbol{F}_s, \boldsymbol{F}_\pi\right), \quad \text{where} \\
&\boldsymbol{F}_\ell = m\{\mathrm{Var}(\boldsymbol{U}_1 \mid Z = \ell)\}, \quad \text{for } \ell = 1, \ldots, s, \\
&\boldsymbol{F}_\pi = \boldsymbol{D}_\pi^{-1} + \pi_s^{-1} \mathbf{1}\mathbf{1}^T.
\end{aligned} \tag{2.3}$$

Here, $\boldsymbol{D}_\pi = \mathrm{Diag}(\pi_1, \ldots, \pi_{s-1})$ and $\mathbf{1}$ denotes a vector of ones of the appropriate dimension. Notice that $\boldsymbol{F}_\ell$ is the $k \times k$ FIM with respect to $\boldsymbol{W}_\ell$, and $\boldsymbol{F}_\pi$ 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 (2.3), the complete data density for $(\boldsymbol{T}, Z)$ is

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

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

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

Taking second derivatives yields

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

Taking the expected value of the negative of these terms, jointly with respect to $(\boldsymbol{T}, Z)$, obtains the blocks of (2.3).

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

3 Convergence of Approximate Information Matrix

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

Lemma 3.1. *The matrix* $\widetilde{\mathcal{I}}(\boldsymbol{\theta}) - \mathcal{I}(\boldsymbol{\theta})$ *is equal to* $\mathrm{E}_{\boldsymbol{T}}\left[\mathcal{I}_{Z|\boldsymbol{T}}(\boldsymbol{\theta})\right]$.

Proof. Notice that

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

Therefore,

$$
\begin{aligned}
\widetilde{\mathcal{I}}(\boldsymbol{\theta}) &= \mathrm{E}_{\boldsymbol{T},Z}\left[\left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\boldsymbol{T}, Z) \right\} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\boldsymbol{T}, Z) \right\}^T \right] \\
&= \mathrm{E}_{\boldsymbol{T}}\left[\mathcal{I}_{Z|\boldsymbol{T}}(\boldsymbol{\theta})\right] + \boldsymbol{B} + \boldsymbol{B}^T + \mathcal{I}(\boldsymbol{\theta}). \qquad (3.1)
\end{aligned}
$$

Now we have

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

The result follows from rearranging terms in (3.1). □ □

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

Actual Information = Complete Information − Missing Information.

Before proceeding with the main result, we state several important consequences of Lemma 3.1. 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 $\boldsymbol{e}_j$ will be used to represent the jth column of the identity matrix of the appropriate dimension.

Corollary 3.2.

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

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

(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* $\boldsymbol{\theta}_0 \in \Theta$,

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

(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 the two inverse matrices respectively. Then* $\mathcal{I}^{jj}(\boldsymbol{\theta}) > \widetilde{\mathcal{I}}^{jj}(\boldsymbol{\theta})$ *for* $j = 1, \ldots, q$.

Proof.

(a) From Lemma 3.1, $\widetilde{\mathcal{I}}(\hat{\boldsymbol{\theta}}) - \mathcal{I}(\hat{\boldsymbol{\theta}}) = \mathrm{E}_{\boldsymbol{T}}\left[\mathcal{I}_{Z|\boldsymbol{T}}(\boldsymbol{\theta})\right]$, an expected value of a conditional information matrix which is positive semidefinite. Therefore the quantity

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

is nonnegative and the result follows.

(b) Lemma A.2 gives that $\mathcal{I}^{-1}(\boldsymbol{\theta}_0) - \widetilde{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}}_0)$ is positive definite, which implies that the quantity

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

is seen to be strictly positive, and the result follows.

(c) Lemma A.2 gives that $\mathcal{I}^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta})$ is positive definite, therefore the diagonal elements $\boldsymbol{e}_j^T \left[\mathcal{I}^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta})\right] \boldsymbol{e}_j$ are positive for $j = 1, \ldots, q$. □ □

A useful consequence of Lemma 3.1 is next given as Proposition 3.3, 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.

Proposition 3.3. *Denote the* (i,j)*th element of* $\mathcal{I}_{Z|\boldsymbol{T}}(\boldsymbol{\theta})$ *as* $C_{ij}^{(m)}$ *when the sample size is* m*. Then*

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

Proof. Recall that $\mathrm{E}(C_{ij}^{(m)})$ is the (i,j)th element of $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})$ by Lemma 3.1. Because $\mathcal{I}_{Z|\boldsymbol{T}}(\boldsymbol{\theta})$ is the covariance matrix of a score function, we may apply the Cauchy-Schwarz inequality to obtain

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

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

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

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

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

which gives the result. □ □

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

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

so that $\widetilde{\mathcal{I}}_m(\boldsymbol{\psi}) - \mathcal{I}_m(\boldsymbol{\psi}) \to \mathbf{0}$ as $m \to \infty$ if and only if $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta}) \to \mathbf{0}$. It is also clear that the rate of convergence of the elements of $\widetilde{\mathcal{I}}_m(\boldsymbol{\psi}) - \mathcal{I}_m(\boldsymbol{\psi})$ to zero will be equivalent to the rate of the elements of $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})$.

Now consider the block decomposition of the true information matrix

$$\mathcal{I}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{A}_{11} & \ldots & \boldsymbol{A}_{1s} & \boldsymbol{A}_{1\pi} \\ \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{A}_{s1} & \ldots & \boldsymbol{A}_{ss} & \boldsymbol{A}_{s\pi} \\ \boldsymbol{A}_{\pi 1} & \ldots & \boldsymbol{A}_{\pi s} & \boldsymbol{A}_{\pi\pi} \end{pmatrix}, \tag{3.2}$$

with blocks

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

By Proposition 3.3, 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

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

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

$$\mathrm{E}_{\boldsymbol{T}}[Q_i^{(m)}(\boldsymbol{T})] = \mathrm{E}_{\boldsymbol{T}}\left\{\mathrm{E}_{Z|\boldsymbol{T}}[I(Z = \ell) \mid \boldsymbol{T}]\right\} = \mathrm{P}(Z = \ell) = \pi_\ell. \tag{3.3}$$

Later we will encounter the same expectation but under the density $f(\boldsymbol{t} \mid \boldsymbol{\eta}_\ell)$, in which case the simplification (3.3) does not happen. Block (i, i) of the decomposition (3.2) can be written as

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

whose jth diagonal element is

$$\boldsymbol{e}_j^T \left[\pi_i \boldsymbol{F}_i - \boldsymbol{A}_{ii}\right] \boldsymbol{e}_j = \pi_i^2 \, \mathrm{E}\left\{\left[1 - Q_i^{(m)}(\boldsymbol{W}_i)\right]\left[W_{ij} - \mathrm{E}(W_{ij})\right]^2\right\}. \tag{3.4}$$

The lower right block of the decomposition (3.2) is

$$\boldsymbol{F}_\pi - \boldsymbol{A}_{\pi\pi} = \left(\boldsymbol{D}_\pi^{-1} + \pi_s^{-1}\mathbf{1}\mathbf{1}^T\right) \tag{3.5}$$
$$- \operatorname{E}\left[\frac{1}{f^2(\boldsymbol{t} \mid \boldsymbol{\theta})} \begin{pmatrix} f(\boldsymbol{t} \mid \boldsymbol{\eta}_1) - f(\boldsymbol{t} \mid \boldsymbol{\eta}_s) \\ \vdots \\ f(\boldsymbol{t} \mid \boldsymbol{\eta}_{s-1}) - f(\boldsymbol{t} \mid \boldsymbol{\eta}_s) \end{pmatrix} \begin{pmatrix} f(\boldsymbol{t} \mid \boldsymbol{\eta}_1) - f(\boldsymbol{t} \mid \boldsymbol{\eta}_s) \\ \vdots \\ f(\boldsymbol{t} \mid \boldsymbol{\eta}_{s-1}) - f(\boldsymbol{t} \mid \boldsymbol{\eta}_s) \end{pmatrix}^T\right].$$

whose ath diagonal element can be expressed as

$$\boldsymbol{e}_a^T \left[\boldsymbol{F}_\pi - \boldsymbol{A}_{\pi\pi}\right] \boldsymbol{e}_a = (\pi_a^{-1} + \pi_s^{-1}) - \pi_a^{-1} \operatorname{E}\left[Q_a^{(m)}(\boldsymbol{W}_a)\right]$$
$$- \pi_s^{-1} \operatorname{E}\left[Q_s^{(m)}(\boldsymbol{W}_s)\right] + 2\pi_a^{-1} \operatorname{E}\left[Q_a^{(m)}(\boldsymbol{W}_s)\right] \tag{3.6}$$

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

Lemma 3.4. *Suppose the density $f(\boldsymbol{t} \mid \boldsymbol{\eta}) = \exp\{\boldsymbol{\eta}^T\boldsymbol{t} + ma(\boldsymbol{\eta})\}$, has natural parameter space Ξ which is an open convex set, and FIM $\mathcal{I}_m(\boldsymbol{\eta})$ is positive definite on Ξ. Then for any $\boldsymbol{\eta}^* \in \Xi$*

$$a(\boldsymbol{\eta}) - a(\boldsymbol{\eta}^*) < a'(\boldsymbol{\eta}^*)^T(\boldsymbol{\eta} - \boldsymbol{\eta}^*), \quad \forall \boldsymbol{\eta} \in \Xi. \tag{3.7}$$

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

Proof. Notice that

$$\frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T} \log f(\boldsymbol{t} \mid \boldsymbol{\eta}) = m\frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T} a(\boldsymbol{\eta}).$$

Because $\mathcal{I}_m(\boldsymbol{\eta}) = -m\frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T} a(\boldsymbol{\eta})$ is positive definite on Ξ, this implies $-a$ is a strictly convex function. Since a is differentiable on the convex set Ξ we have, for $g := -a$,

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

which is equivalent to the result (3.7). □ □

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

$$-\gamma_{IJK} = -a'(\boldsymbol{\eta}_J)^T(\boldsymbol{\eta}_I - \boldsymbol{\eta}_K) + [a(\boldsymbol{\eta}_I) - a(\boldsymbol{\eta}_K)],$$
$$c_i^* = \bigwedge_{\ell \neq i}^{s} \gamma_{\ell ii}, \quad d_{ij}^* = \bigvee_{\ell \neq i}^{s} \left\{-\gamma_{\ell ji}\right\}, \quad \text{and} \quad c^{**} = \bigwedge_{\ell=1}^{s} c_\ell^*, \tag{3.8}$$

which will be used for the remainder of the paper.

Proposition 3.5. *Suppose $\boldsymbol{\eta}_a \neq \boldsymbol{\eta}_b$ for all $a \neq b$. Then*

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

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

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

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

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

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

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

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

and

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

Case (a). Suppose $j = i$. From Lemma 3.4 we have

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

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

$$0 \leq R_i^{(m)}(\boldsymbol{W}_i(\omega)) \leq \sum_{\ell \neq i}^{s} \pi_\ell e^{m(-\gamma_{\ell i i} + \varepsilon)} = \sum_{\ell \neq i}^{s} \pi_\ell e^{-m(\gamma_{\ell i i} - \varepsilon)} \leq e^{-m(c_i^* - \varepsilon)} \sum_{\ell \neq i}^{s} \pi_\ell \leq e^{-m(c_i^* - \varepsilon)} \to 0$$

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

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

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

Notice that

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

where $\gamma_{ijj} > 0$ by Lemma 3.4. Then for $m \geq M_\omega$,

$$R_i^{(m)}(\boldsymbol{W}_j(\omega)) \geq \sum_{\ell \neq i}^{s} \pi_\ell e^{m(-\gamma_{\ell j i} - \varepsilon)} \geq \pi_j e^{m(-\gamma_{jji} - \varepsilon)} = \pi_j e^{m(\gamma_{ijj} - \varepsilon)} \to \infty, \tag{3.9}$$

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

$$R_i^{(m)}(\boldsymbol{W}_j(\omega)) \le \sum_{\ell\neq i}^{s} \pi_\ell e^{m(-\gamma_{\ell j i}+\varepsilon)} \le \sum_{\ell\neq i}^{s} \pi_\ell e^{m(d^*_{ij}+\varepsilon)} \le e^{m(d^*_{ij}+\varepsilon)}, \tag{3.10}$$

noting that $d^*_{ij} = \bigvee_{\ell\neq i}^{s}\{-\gamma_{\ell j i}\} \ge -\gamma_{jji} = \gamma_{ijj} > 0$. We have therefore found the upper and lower bounds

$$\pi_j e^{m(\gamma_{ijj}-\varepsilon)} \le R_i^{(m)}(\boldsymbol{W}_j(\omega)) \le e^{m(d^*_{ij}+\varepsilon)}, \qquad \forall m \ge M_\omega,$$

and hence the desired almost sure bounds

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

are obtained. □ □

Proposition 3.6. *Suppose $\boldsymbol{\eta}_a \neq \boldsymbol{\eta}_b$ for all $a \neq b$. Then*

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

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

with c_i^* and γ_{ijj} as defined in (3.8).

Proof. Recall that $Q_i^{(m)}(\boldsymbol{t}) = \pi_i \cdot \{\pi_i + R_i^{(m)}(\boldsymbol{t})\}^{-1}$ and apply Proposition 3.5 to obtain the limit. To obtain the rates, first take $\boldsymbol{T} = \boldsymbol{W}_i$, and notice that

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

Since $R_i^{(m)}(\boldsymbol{W}_i) \overset{a.s.}{=} O(e^{-mc_i^*})$ by Proposition 3.5, there exists a constant K such that

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

almost surely for all m large, so that

$$e^{mc_i^*}\left[1 - Q_i^{(m)}(\boldsymbol{W}_i)\right] \le \frac{e^{mc_i^*}}{\pi_i K^{-1} e^{mc_i^*} + 1} \to \frac{K}{\pi_i}, \quad \text{as } m \to \infty,$$

This gives the result $1 - Q_i^{(m)}(\boldsymbol{t}) \overset{a.s.}{=} O(e^{-mc_i^*})$.

Now take $\boldsymbol{T} = \boldsymbol{W}_j$ for $j \neq i$. We have $Q_i^{(m)}(\boldsymbol{W}_j) = \pi_i \cdot \{\pi_i + R_i^{(m)}(\boldsymbol{W}_j)\}^{-1}$, and Proposition 3.5 gives that $R_i^{(m)}(\boldsymbol{W}_j) \ge e^{m\gamma_{ijj}}$ almost surely for all large m. Then we have

$$e^{m\gamma_{ijj}} Q_i^{(m)}(\boldsymbol{W}_j) = \frac{\pi_i e^{m\gamma_{ijj}}}{\pi_i + R_i^{(m)}(\boldsymbol{W}_j)} \le \frac{\pi_i e^{m\gamma_{ijj}}}{\pi_i + O(e^{m\gamma_{ijj}})}$$

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

Proposition 3.6 suggests that the convergence between the FIM and approximate information will be fast when both of the following happen quickly as m is increased: (1) the posterior probability of being in the ℓth subpopulation goes to 1 when the true subpopulation $Z = \ell$, and (2) the posterior probability of being in the ℓth subpopulation goes to 0 when the true subpopulation $Z \neq \ell$. It is clear from Proposition 3.6 and dominated convergence that the expectation (3.6) converges to zero. We also note that $W_{ij} - \mathrm{E}(W_{ij})$ is a sum of independent and identically distributed random variables, so that $[W_{ij} - \mathrm{E}(W_{ij})]^2 \overset{a.s.}{=} O(m^2)$, and therefore

$$\pi_i^2 \left[1 - Q_i^{(m)}(\boldsymbol{W}_i)\right] \left[W_{ij} - \mathrm{E}(W_{ij})\right]^2 \overset{a.s.}{=} O(m^2 e^{-mc_i^*}). \tag{3.11}$$

Then the expectation (3.4) converges to zero if and only if the LHS of (3.11) is uniformly integrable (Resnick, 1999, chapter 6). The convergence of $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})$ can therefore be characterized in the following theorem.

Theorem 3.7. *$\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta}) \to 0$ as $m \to \infty$ if and only if the sequence* (3.11) *is uniformly integrable for each $i = 1, \dots, s$.*

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.

Lemma 3.8. $\mathrm{E}\left[Q_i^{(m)}(\boldsymbol{W}_i)\right] = 1 - O(e^{-mc_i^*})$ *with c_i^* defined as in Proposition* (3.8).

Proof. From the Markov inequality we have,

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

recalling that $0 \leq Q_i^{(m)}(\boldsymbol{W}_i) \leq 1$. Proposition 3.6 gives that $Q_i^{(m)}(\boldsymbol{W}_i) \overset{a.s.}{=} 1 - O(e^{-mc_i^*})$, which implies $\mathrm{P}\left(Q_i^{(m)}(\boldsymbol{W}_i) \geq \varepsilon\right) = 1 - O(e^{-mc_i^*})$, assuming that $0 < \varepsilon < 1$. Therefore

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

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

Lemma 3.9. *Let $S_n = X_1 + \cdots + X_n$ where $\{X_i\}$ are independent and identically distributed and* $\mathrm{E}(|X_1|^k) < \infty$ *for a given positive integer $k \geq 0$. Then $\mathrm{E}(S_n^k) = O(n^k)$.*

Proof. Notice that

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

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

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

and note that $\xi \geq 0$ 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

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

which gives the result. □ □

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

Theorem 3.10. *Consider the matrix* $\widetilde{\mathcal{I}}(\boldsymbol{\theta}) - \mathcal{I}(\boldsymbol{\theta})$*;*

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

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

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

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

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

Proof. For (a) we have

$$\begin{aligned}
&\pi_i^2 \,\mathrm{E}\left\{\left[1 - Q_i^{(m)}(\boldsymbol{W}_i)\right]\left[W_{ij} - \mathrm{E}(W_{ij})\right]^2\right\} \\
&\leq \pi_i^2 \sqrt{\mathrm{E}\left[\left(1 - Q_i^{(m)}(\boldsymbol{W}_i)\right)^2\right]} \sqrt{\mathrm{E}\left[\left(W_{ij} - \mathrm{E}(W_{ij})\right)^4\right]} && (3.12) \\
&\leq \pi_i^2 \sqrt{\mathrm{E}\left[1 - Q_i^{(m)}(\boldsymbol{W}_i)\right]} \sqrt{\mathrm{E}\left[\left(W_{ij} - \mathrm{E}(W_{ij})\right)^4\right]} && (3.13) \\
&= \pi_i^2 \left\{O(e^{-mc_i^*}) O(m^4)\right\}^{1/2} && (3.14) \\
&= O(m^2 e^{-\frac{m}{2} c_i^*}).
\end{aligned}$$

Notice that (3.12) follows from the Cauchy-Schwarz inequality, (3.13) because $0 \leq X \leq 1$ implies $\mathrm{E}(X^2) \leq \mathrm{E}(X)$, and (3.14) by Lemmas 3.8 and 3.9.

For (b), use Proposition 3.6 with the expectation (3.6) to obtain

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

□ □

Note that Theorem 3.10 (b) implies the simpler bound $\boldsymbol{e}_j^T \left(\boldsymbol{F}_\pi - \boldsymbol{A}_{\pi\pi}\right) \boldsymbol{e}_j = O(e^{-mc_j^*}) + O(e^{-mc_s^*}) = O(e^{-mc^{**}})$ since $c_\ell^* = \bigwedge_{a=1}^s \gamma_{\ell aa}$.

Because of the convenient block-diagonal form of the information matrix approximation, its inverse $\widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta}) = \text{Blockdiag}(\pi_1^{-1}\boldsymbol{F}_1^{-1}, \ldots, \pi_s^{-1}\boldsymbol{F}_s^{-1}, \boldsymbol{F}_\pi^{-1})$ is also block-diagonal. As in (Raim et al., 2014, 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 3.11. *Suppose* $\mathcal{I}_m(\boldsymbol{\theta})$ *and* $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta})$ *are nonsingular. Then* $\mathcal{I}_m^{-1}(\boldsymbol{\theta}) - \widetilde{\mathcal{I}}_m^{-1}(\boldsymbol{\theta}) \to \boldsymbol{0}$ *as* $m \to \infty$*.*

4 Relationship to Classification Problem

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

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

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

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

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

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

The optimal probability of misclassification $p(\boldsymbol{\mathcal{D}}^*)$ provides a measurement on the degree of mutual separation between the s subpopulations; a higher probability indicates that it is more difficult to distinguish among them. Of course the rule $\boldsymbol{\mathcal{D}}^*$ can only be applied when all $\boldsymbol{\phi}_\ell$ and $\boldsymbol{\pi}$ are known. Recall that $R_\ell^{(m)}(\boldsymbol{W}_\ell) \overset{a.s.}{=} o(e^{-mc_\ell^*})$, where c_ℓ^* was defined in (3.8), so that we obtain $p(\boldsymbol{\mathcal{D}}^*) = o(e^{-mc^{**}})$. Therefore, collection of additional observations for $\boldsymbol{T} = \sum_{i=1}^m \boldsymbol{U}(\boldsymbol{X}_i)$ may drastically improve $p(\boldsymbol{\mathcal{D}}^*)$ if all c_ℓ^* are large, and has almost no effect when some c_ℓ^* are very small.

A connection between $p(\boldsymbol{\mathcal{D}}^*)$ and the convergence rate of the approximate information matrix can be seen from

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

so that $\mathrm{P}\left(R_\ell^{(m)}(\boldsymbol{W}_\ell) \geq \pi_\ell \right)$ gives an upper bound on the probability of misclassifying $\boldsymbol{T}$ when $Z = \ell$. Recall that the convergence rate of the ℓth block of $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})$ depends on $\mathrm{E}\left[1 - Q_\ell^{(m)}(\boldsymbol{W}_\ell) \right]$, as in the proof of Theorem 3.10. Proposition 3.5 gives

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

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

5 Examples

Remark 5.1 (Multinomial Finite Mixture)**.** Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ be independent and identically distributed as $\text{Mult}_{k+1}(1, \boldsymbol{p}_Z)$, with $Z \sim \text{Discrete}(1, \ldots, s; \boldsymbol{\pi})$. Take $\boldsymbol{T} = \sum_{i=1}^m \boldsymbol{X}_i$. The multinomial subpopulations are exponential families with $f(\boldsymbol{t} \mid m, \boldsymbol{p}_\ell) = \exp\left\{\boldsymbol{\eta}_\ell^T \boldsymbol{t} + ma(\boldsymbol{\eta_\ell}) + h(\boldsymbol{t})\right\}$, where

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

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

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

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

Remark 5.2 (Multiivariate Normal Finite Mixture)**.** Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ 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 $\boldsymbol{T} = \sum_{i=1}^m \boldsymbol{X}_i \sim \text{N}(m\boldsymbol{\mu}_Z, m\boldsymbol{\Sigma})$ given Z. Let us compare the FIM and approximation with respect to $\boldsymbol{\psi} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_s, \boldsymbol{\pi})$, where $\boldsymbol{\Sigma}$ is taken to be known. The Normal subpopulations are exponential families with $f(\boldsymbol{t} \mid m\boldsymbol{\mu}_j, m\boldsymbol{\Sigma}) = \exp\left\{\boldsymbol{\eta}_j^T \boldsymbol{t} + ma(\boldsymbol{\eta_j}) + h(\boldsymbol{t})\right\}$ where $\boldsymbol{\eta}_j = \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j$ and $ma(\boldsymbol{\eta_j}) = -m\frac{1}{2}\boldsymbol{\eta}_j^T \boldsymbol{\Sigma} \boldsymbol{\eta}_j$. Under $Z = j$, the first and second derivative of the log-density with respect to $\boldsymbol{\eta}_j$ are given by

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

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

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

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

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

and $\boldsymbol{F}_\pi = \boldsymbol{D}_\pi^{-1} + \pi_s^{-1} \boldsymbol{1}\boldsymbol{1}^T$. We will study the closeness between the FIM and the approximation by numerical experiment. The true information matrix will be 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

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

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

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

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

and

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

where γ_{211} and γ_{122} are nonnegative by Lemma 3.4. Therefore, the numbers γ_{211} and γ_{122} together are sufficient to describe the orders for the convergence rates. We will consider three scenarios for the subpopulation means,

- Scenario 1: $\boldsymbol{\mu}_1 = (-1, 1)$, $\boldsymbol{\mu}_2 = (1, -1)$, so that $\gamma_{221} = \gamma_{122} = 8$.
- Scenario 2: $\boldsymbol{\mu}_1 = (-0.5, 0.5)$, $\boldsymbol{\mu}_2 = (0.5, -0.5)$, 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 $\widetilde{\mathcal{I}}_m(\boldsymbol{\psi})$ compared with those of $\mathcal{I}_m(\boldsymbol{\psi})$, where the latter have been computed numerically. Also shown is the Frobenius norm of the matrix $\widetilde{\mathcal{I}}_m(\boldsymbol{\psi}) - \mathcal{I}_m(\boldsymbol{\psi})$. Note from the proof of Theorem 3.11 that

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

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. Figure 2 plots the norms for the three scenarios.[2]

Remark 5.3 (Sampling iid from Normal Finite Mixture)**.** It is natural to ask if there is relationship between the information matrix of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ independently and identically distributed from $f(\boldsymbol{x} \mid \boldsymbol{\phi}_Z)$, but where Z is not observed, and the information matrix of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ independently and identically distributed from the finite mixture $f(\boldsymbol{x} \mid \boldsymbol{\theta})$. The convergence theory in this paper was developed strictly for the former case. As a concrete example, suppose $X_1, \ldots, X_m$ are Normal random variables. Let $\mathcal{I}_m(\boldsymbol{\theta})$ denote the information matrix of $T = \sum_{i=1}^m X_i$, where $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_s, \pi_1, \ldots, \pi_{s-1})$ and 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

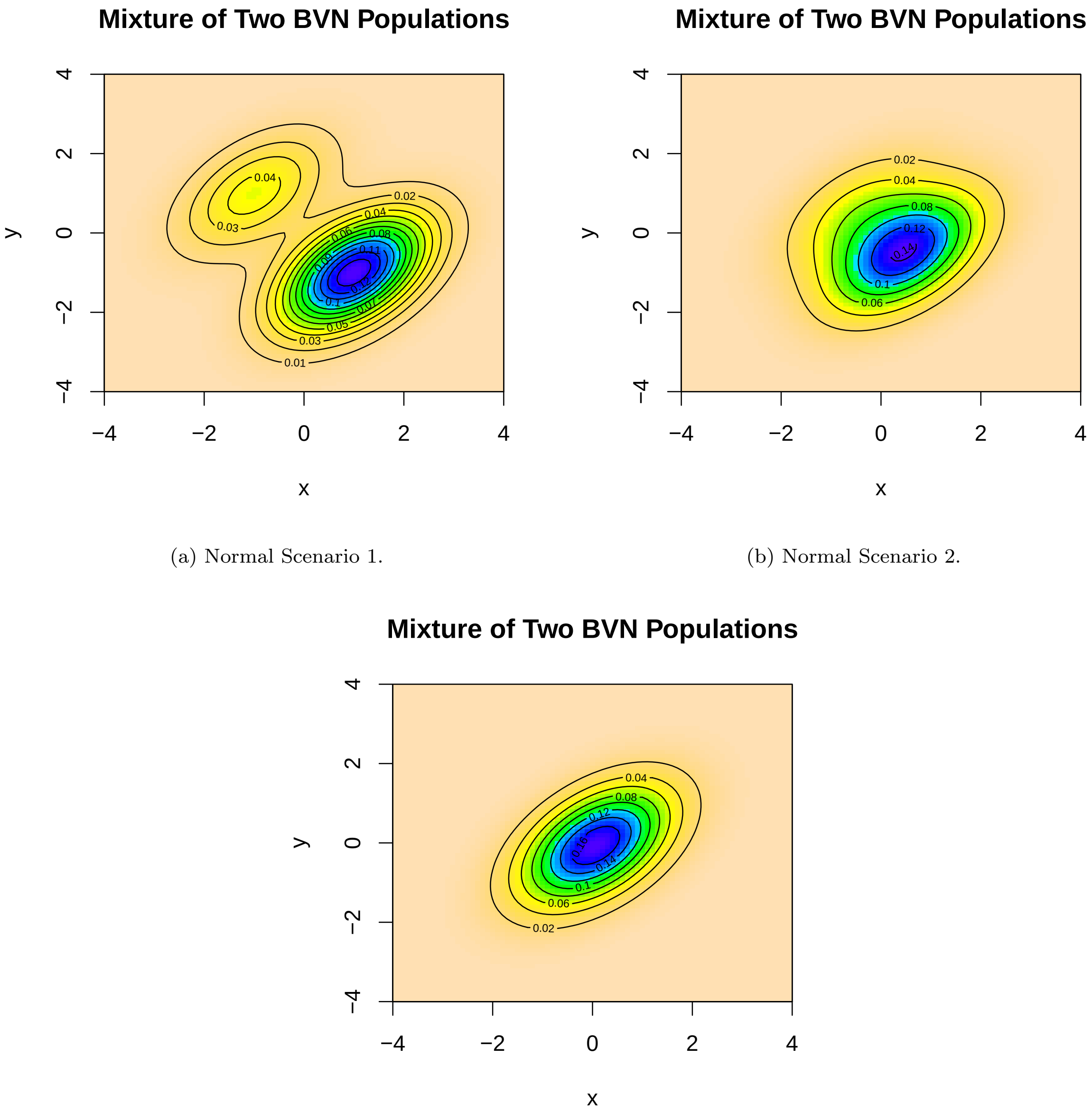
$$f(x \mid \boldsymbol{\theta}) = \sum_{\ell=1}^{s} \pi_\ell \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x - \mu_\ell)^2\right\},$$

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

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

[2] The numerical integration sometimes produced inaccurate results, which we believe were caused by the very large limits of integration we provided to the software. For example, in Scenario 1 when $m = 8$ and in Scenario 2 when $m = 26$, $\mathcal{I}_{55} = 4.6667$ instead of the expected 5.3333. These results have been omitted from the tables and plots.



(a) Normal Scenario 1.

(b) Normal Scenario 2.

(c) Normal Scenario 3.

Figure 1: Densities for the bivariate normal finite mixture under the three scenarios.

Table 1: Results for bivariate normal mixture. The diagonals $\widetilde{\mathcal{I}}_{ii}$ are given with corresponding $\mathcal{I}_{ii}$ in parentheses. The last column shows Frobenius norm of the matrix difference $\widetilde{\mathcal{I}} - \mathcal{I}$.

(a) Scenario 1

m	$\widetilde{\mathcal{I}}_{11}$	$\widetilde{\mathcal{I}}_{22}$	$\widetilde{\mathcal{I}}_{33}$	$\widetilde{\mathcal{I}}_{44}$	$\widetilde{\mathcal{I}}_{55}$	$\Vert\widetilde{\mathcal{I}} - \mathcal{I}\Vert_{\mathrm{F}}$
1	0.333 (0.276)	0.333 (0.273)	1 (0.910)	1 (0.920)	5.333 (4.914)	0.6486
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

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

(c) Scenario 3

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

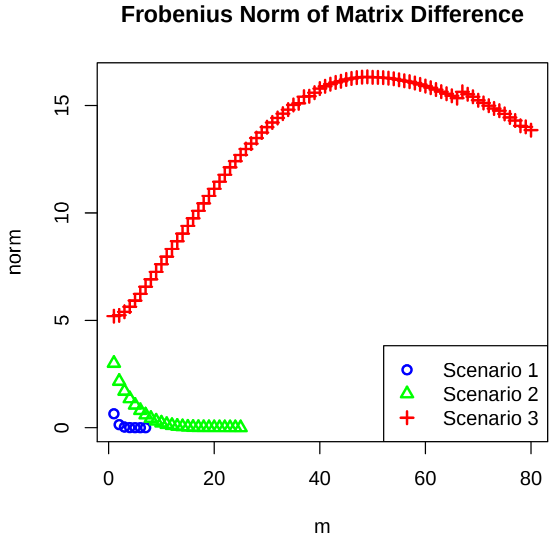


Figure 2: Frobenius norm of $\widetilde{\mathcal{I}}_m - \mathcal{I}_m$, as m varies, for the three normal scenarios.

- 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 the convergence may not occur when the sample is not drawn in a clustered manner.

Remark 5.4 (Dirichlet-Multinomial)**.** The distribution of a Dirichlet-Multinomial random variable T may be written as

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

which is a continuous mixture of multinomial. Then the complete data density of $(\boldsymbol{T}, \boldsymbol{\mu})$ is

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

Let $J = k + 1$ to ensure the parameter space of the multinomial family contains an open set in $\mathbb{R}^k$. The Dirichlet-Multinomial density is obtained by finding the marginal distribution of $\boldsymbol{T}$, as

$$f(\boldsymbol{t} \mid \boldsymbol{\alpha}) = \frac{m!}{t_1! \cdots t_J!} \frac{\prod_{j=1}^J \Gamma(\alpha_j + t_j)}{\Gamma(\sum_{j=1}^J \alpha_j)} \frac{\Gamma(\sum_{j=1}^J \alpha_j + m)}{\prod_{j=1}^J \Gamma(\alpha_j)}. \tag{5.1}$$

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 $\boldsymbol{T} \mid \boldsymbol{\mu}$ is free of $\boldsymbol{\alpha}$ so that $\frac{\partial}{\partial \boldsymbol{\alpha}} \log f(\boldsymbol{t}, \boldsymbol{\mu} \mid \boldsymbol{\alpha}) = \frac{\partial}{\partial \boldsymbol{\alpha}} \log f(\boldsymbol{\mu} \mid \boldsymbol{\alpha})$; therefore, the complete data information matrix is just the FIM with respect to

$\text{Dirichlet}_J(\boldsymbol{\alpha})$. This is analogous to the finite mixture case, where the first s diagonal blocks correspond to the support points of the mixing distribution $\text{Discrete}(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_s; \boldsymbol{\pi})$, and the lower-right block of the matrix corresponds to $\boldsymbol{\pi}$. Now the mixing process follows a Dirichlet distribution whose support is the probability simplex in $\mathbb{R}^J$ (i.e. which is known and does not have corresponding entries in the information matrix). (Neerchal and Morel, 1998, Theorem 1) shows that the FIM of $\boldsymbol{T}$ converges to the FIM of $\text{Dirichlet}_k(\boldsymbol{\alpha})$ as $m \to \infty$. Therefore, the results in this paper may extend to more general settings than when the latent mixing process follows a finite mixture distribution.

Remark 5.5 (Normal-Normal)**.** Let us consider a second continuous mixture along the lines of Example 5.4. 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

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

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 $\mathcal{I}_m(\theta) = (\sigma^2/m + \tau^2)^{-1}$. As in Example 5.4, 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 $m \to \infty$.

Remark 5.6 (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 which can be written as $T = NY + (X \mid N)$, where

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

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

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

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

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

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 matrix $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta})$ may be formulated by first forming the approximate information matrix,

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

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

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

The convergence of $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta})$ to zero follows from Theorem 3.10.

Remark 5.7 (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 can be written as

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

and the Fisher information is therefore found by computing

$$\mathcal{I}(\beta) = \int_0^\infty \left\{\frac{1}{\beta} - \left[1 - \left(\frac{x}{\lambda}\right)^\beta\right] \log\left(\frac{x}{\lambda}\right)\right\}^2 f(x \mid \beta, \lambda) dx. \tag{5.2}$$

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 $\boldsymbol{X} = (X_1, \ldots, X_m)$ given Z are a random sample from $\text{Weibull}(\beta_Z, \lambda_Z)$ and $Z \sim \text{Discrete}(1, \ldots, s; \boldsymbol{\pi})$. The marginal density of $\boldsymbol{X}$ is then given by

$$f(\boldsymbol{x} \mid \boldsymbol{\theta}) = \sum_{\ell=1}^s \pi_\ell \left[\left(\frac{\beta_\ell}{\lambda_\ell}\right)^m \left(\prod_{i=1}^m \frac{x_i}{\lambda_\ell}\right)^{\beta_\ell - 1} \exp\left\{-\sum_{i=1}^m (x_i/\lambda_\ell)^{\beta_\ell}\right]\right\} \tag{5.3}$$

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

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

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

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

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

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

where F_ℓ is given by multiplying the $\text{Weibull}(\beta_\ell, \lambda_\ell)$ information (5.2) by m, and $\boldsymbol{F}_\pi = \boldsymbol{D}_\pi^{-1} + \pi_s^{-1} \boldsymbol{1}\boldsymbol{1}^T$ as usual for finite mixtures.

Consider two scenarios using densities of the form

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

with

Table 2: Results for Weibull mixture. The diagonals $\widetilde{\mathcal{I}}_{ii}$ are given with corresponding $\mathcal{I}_{ii}$ in parentheses. The last column shows Frobenius norm of the matrix difference $\widetilde{\mathcal{I}} - \mathcal{I}$. All entries of $\mathcal{I}$ were approximated by Monte Carlo simulation.

(a) Scenario 1

m	$\widetilde{\mathcal{I}}_{11}$	$\widetilde{\mathcal{I}}_{22}$	$\widetilde{\mathcal{I}}_{33}$	$\lVert\widetilde{\mathcal{I}} - \mathcal{I}\rVert_{\mathrm{F}}$
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

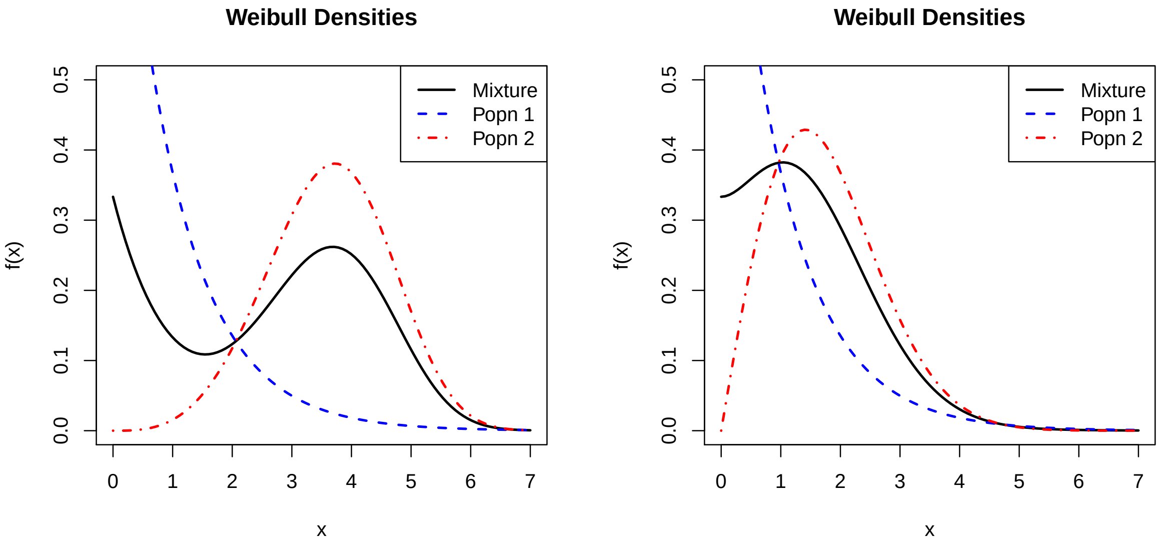
m	$\widetilde{\mathcal{I}}_{11}$	$\widetilde{\mathcal{I}}_{22}$	$\widetilde{\mathcal{I}}_{33}$	$\lVert\widetilde{\mathcal{I}} - \mathcal{I}\rVert_{\mathrm{F}}$
1	0.6079 (0.3919)	0.3039 (0.1696)	4.5 (1.0642)	3.4731
2	1.2158 (0.8718)	0.6079 (0.3840)	4.5 (1.7997)	2.8164
3	1.8237 (1.3980)	0.9118 (0.6135)	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

- 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 for these scenarios, respectively. Evaluation of the approximate information matrix requires computing (5.2), which we compute by numerical integration. The true FIM is computed by Monte Carlo simulation using 100,000 samples. We have elected to use a basic Monte Carlo method, and while a more accurate method could be used, there is clear evidence of the convergence in Table 2. As expected, it is faster in Scenario 1 where the subpopulations are further apart.

6 Conclusions

This paper extended (Raim et al., 2014) from multinomial finite mixtures to the more general class of exponential family finite mixtures, making it relevant to statistical analysis beyond binomial and multinomial data. The main convergence result showed that the true 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 provides a justification for 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



(a) Scenario 1.

(b) Scenario 2.

Figure 3: Densities for the Weibull finite mixture under the two scenarios.

the similarity between subpopulations. Example 5.3 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 assumes this setting (e.g. the $R_i(\cdot)$ and $Q_i(\cdot)$ functions are critical to the proof), but Examples 5.4 and 5.5 provide evidence of the convergence even when the latent mixing process has a continuous distribution. Example 5.7 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 have a reliable method of correcting accuracy in the approximate information when m is not large or the subpopulations are not well-separated.

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 (`www.umbc.edu/hpcf`) at the university. The first author thanks the facility for financial support as an RA.

A Additional Results

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

Lemma A.2. *Suppose* $\boldsymbol{A}, \boldsymbol{B}$ *are* $q \times q$ *symmetric positive definite matrices, and* $\boldsymbol{B} - \boldsymbol{A}$ *is positive definite. Then* $\boldsymbol{A}^{-1} - \boldsymbol{B}^{-1}$ *is positive definite.*

Lemma A.2. By Lemma A.1, $\boldsymbol{A}^{-1} - \boldsymbol{B}^{-1} = \boldsymbol{B}^{-1}(\boldsymbol{B} - \boldsymbol{A})\boldsymbol{A}^{-1}$. Suppose λ is an eigenvalue of $\boldsymbol{A}^{-1} - \boldsymbol{B}^{-1}$, then

$$\det(\boldsymbol{B}^{-1}(\boldsymbol{B}-\boldsymbol{A})\boldsymbol{A}^{-1}-\lambda I)=0 \quad \iff \quad \det(\boldsymbol{B}^{-1/2}(\boldsymbol{B}-\boldsymbol{A})^{1/2}\boldsymbol{A}^{-1}(\boldsymbol{B}-\boldsymbol{A})^{1/2}\boldsymbol{B}^{-1/2}-\lambda I)=0.$$

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

The following proof of Theorem 3.11 follows a similar argument to that of (Raim et al., 2014, Theorem 2.5), but is included in its entirety for completeness.

Theorem 3.11. Lemma A.1 gives $\mathcal{I}^{-1}(\boldsymbol{\theta})-\widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta})=\mathcal{I}^{-1}(\boldsymbol{\theta})\left[\mathcal{I}(\boldsymbol{\theta})-\widetilde{\mathcal{I}}(\boldsymbol{\theta})\right]\widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta})$ 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 $\boldsymbol{A}$, the matrix 2-norm $\|\cdot\|_2$ and the Frobenius norm $\|\cdot\|_{\mathrm{F}}$ are related by $\|\boldsymbol{A}\|_2\leq\|\boldsymbol{A}\|_{\mathrm{F}}\leq\sqrt{q}\|\boldsymbol{A}\|_2$. Therefore, in showing the convergence of $\|\boldsymbol{A}_m\|$ to zero, we may consider whichever norm is more convenient.

Recalling that $\|\boldsymbol{A}\|_{\mathrm{F}}^2=\sum_i\sum_j a_{ij}^2$, Proposition 3.3 and Theorem 3.10 give the simple bound

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

Next, we have

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

where $\widetilde{\mathcal{I}}_1(\boldsymbol{\eta}_\ell)=\mathrm{Var}(\boldsymbol{U}_1\mid Z=\ell)$ 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

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

Notice that

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

since both LHS and RHS are lower bounds for $\boldsymbol{x}^T\left[\mathcal{I}(\boldsymbol{\theta})-\widetilde{\mathcal{I}}(\boldsymbol{\theta})\right]\boldsymbol{x}+\boldsymbol{x}^T\widetilde{\mathcal{I}}(\boldsymbol{\theta})\boldsymbol{x}$, and the RHS is the greatest such bound. Therefore, 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 $0\geq\beta_1(m)\geq\cdots\geq\beta_q(m)$,

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

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 $\beta_q(m)\to0$. 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

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

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 (A.1) 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

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

which gives the result. □ □

References

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley-Interscience, 3rd edition, 2003.

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

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

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

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis.* Chapman and Hall/CRC, 2nd edition, 2003.

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

E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses.* Springer, 3rd edition, 2005.

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

Geoffrey McLachlan and David Peel. *Finite Mixture Models.* Wiley-Interscience, 2000.

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

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

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

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

Terence Orchard and Max A. Woodbury. A missing information principle: theory and applications. In L. M. Le Cam, J. Neyman, and E. L. Scott, editors, *Proceedings of the Sixth Berkely Symposium on Mathematics, Statistics and Probability, Volume 1.*, pages 697–715. Berkeley: University of California Press, 1972.

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

J. N. K. Rao. *Small Area Estimation*. Wiley-Interscience, 2003.

Sidney Resnick. *A Probability Path*. Birkhäuser, 1999.

Jun Shao. *Mathematical Statistics*. Springer, 2nd edition, 2008.