# Large Cluster Approximation to the Finite Mixture Information Matrix with an Application to Meta-Analysis

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

Department of Mathematics and Statistics,
University of Maryland, Baltimore County

**Abstract**

Finite mixtures have been adopted in a wide range of statistical applications. Their utility comes at a computational cost and a loss of tractability for common inference techniques. The Fisher information matrix (FIM) is often used with maximum likelihood estimation but does not have a simple analytical form in the finite mixture setting. Raim, Neerchal, and Morel (2014, submitted) have recently shown that, in some finite mixture settings, a certain block-diagonal matrix becomes close to the FIM as the sample size increases. This block-diagonal matrix is the FIM of the complete data: the observed data along with a latent indicator of the subpopulation from the mixture from which an observation is drawn. The convergence requires that the sample is "grouped" so that individual observations are known to be drawn from a common subpopulation. One application where this kind of sampling can naturally be justified is meta-analysis. We consider model-based clustering of studies in a meta-analysis to explore the nature of their heterogeneity. A simulation study is presented to illustrate the closeness of the complete data FIM and actual FIM. Use of the complete data FIM is also demonstrated on an example dataset measuring selenium content in nonfat milk powder. We conclude that the complete data FIM can serve as a reasonable approximation to the actual FIM for this kind of application.

**Key Words:** Fisher Information; Complete Data; Model-Based Clustering; Exponential Family.

## 1. Introduction

A finite mixture density $f(\boldsymbol{x} \mid \boldsymbol{\theta}) = \sum_{\ell=1}^{J} \pi_\ell f(\boldsymbol{x} \mid \boldsymbol{\phi}_\ell)$ is a weighted sum of $J$ densities. We will assume they share a common parametric form, which need not be the case in general. Finite mixtures are commonly used to analyze data that exhibit multiple modes or extra variation — more variation than a simpler density such as $f(\boldsymbol{x} \mid \boldsymbol{\phi}_\ell)$ is capable of modeling. This flexibility comes at the price of computational and theoretical complication. The Fisher information matrix (FIM) of $\boldsymbol{X} \sim f(\boldsymbol{x} \mid \boldsymbol{\theta})$ has the general form

$$\mathcal{I}(\boldsymbol{\theta}) = \mathrm{E}\left[\left\{\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{X} \mid \boldsymbol{\theta})\right\}\left\{\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{X} \mid \boldsymbol{\theta})\right\}^T\right] \tag{1}$$

In many applications, the inverse of the FIM is the large sample covariance matrix of the maximum likelihood estimator (MLE). Therefore, the FIM is routinely used to compute standard errors and confidence intervals when the MLE is taken as the estimator and its large sample distribution is used for inference. The FIM is also commonly used to carry out Wald and Score tests. In most finite mixture settings,

Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD, 21250, U.S.A, Email: `araim1@umbc.edu`. Andrew Raim is now with the U.S. Census Bureau.

the general FIM expression (1) does not reduce to something with a simple closed form.

In this work, we illustrate a simple block-diagonal matrix, denoted $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$, that can provide a very close approximation to the FIM under certain finite mixture models. This matrix is the FIM of the "complete data" in the sense of the Expectation-Maximization (EM) algorithm, which is commonly considered for finite mixtures (McLachlan and Peel, 2000, Section 2.8). The observed $\boldsymbol{X}$ is augmented with a random variable $Z$ that labels which of the $J$ mixed densities truly generated $\boldsymbol{X}$. Suppose that $Z$ assumes values $1, \ldots, J$ with corresponding probabilities $\pi_1, \ldots, \pi_J$; we notate this by $Z \sim \text{Discrete}(1, \ldots, J; \boldsymbol{\pi})$. Suppose also that $\boldsymbol{X}$ follows $f(\boldsymbol{x} \mid \boldsymbol{\phi}_\ell)$ given $Z = \ell$. Then marginally $\boldsymbol{X}$ is distributed according to the finite mixture $f(\boldsymbol{x} \mid \boldsymbol{\theta})$.

The first work using the block diagonal approximation appears to be from Blischke (1962, 1964), who proposed a method-of-moments estimator for the finite mixture of binomial densities. A block-diagonal matrix is proposed as a substitute for the FIM to study asymptotic efficiency of the estimator. In later work, Morel and Nagaraj (1993) extended the block-diagonal matrix to the setting of multinomial finite mixtures when considering estimation for the random-clumped multinomial distribution. In both the binomial and multinomial finite mixture cases, the block-diagonal matrix was shown to become close to the FIM as the number of trials increases; this justified its use as an approximation. Raim et al. (2014a) noted that the block-diagonal matrix is obtained as the complete data FIM. This allows the matrix to be formulated for any missing data problem in great generality, but increasing the sample size may not cause it to become close to the FIM. In binomial and multinomial finite mixtures, there is something special about the structure of the sample — observations which are composed of multiple trials — that induces closeness between the actual and complete data FIM.

Raim et al. (2014b) extended the convergence result from the multinomial finite mixture setting to the finite mixture of exponential family densities. This justifies its use outside of the multinomial mixture setting. The major caveat is that we require a certain "grouped sampling" scheme for the convergence to hold. To introduce this idea, suppose $T$ follows a binomial finite mixture with $m$ trials; then all $m$ of its Bernoulli trials are generated from the same (unknown) density within the finite mixture. Similarly, if $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ are distributed according to a finite mixture of exponential family densities (possibly multivariate), we will require that they are independent and identically distributed from the same (unknown) density within the finite mixture. Then, just as in the binomial case, we may work with the sufficient statistic $\boldsymbol{T}$, whose complete data FIM will become close to the actual information when $m$ is sufficiently large.

In applications following a grouped sampling structure, we would observe a sample $\boldsymbol{T}_1, \ldots, \boldsymbol{T}_n$ from the finite mixture so that each $\boldsymbol{T}_i$ is a sufficient statistic based on $m_i$ individual observations $\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{im_i}$. Each $\boldsymbol{T}_i$ is assumed to have an unobserved subpopulation indicator $Z_i$, with $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \text{Discrete}(1, \ldots, J; \boldsymbol{\pi})$. The actual FIM and complete data FIM for the sample are obtained by summing the corresponding information matrices with respect to each observation as

$$\mathcal{I}(\boldsymbol{\theta}) = \mathcal{I}_{m_1}(\boldsymbol{\theta}) + \cdots + \mathcal{I}_{m_n}(\boldsymbol{\theta}) \quad \text{and} \quad \widetilde{\mathcal{I}}(\boldsymbol{\theta}) = \widetilde{\mathcal{I}}_{m_1}(\boldsymbol{\theta}) + \cdots + \widetilde{\mathcal{I}}_{m_n}(\boldsymbol{\theta}).$$

The closeness of $\mathcal{I}(\boldsymbol{\theta})$ and $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ therefore depends on all $m_i$ being sufficiently large. For the rest of the paper, we use a subscripted $\mathcal{I}_m(\boldsymbol{\theta})$ and $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta})$ to denote actual and complete data information for a sufficient statistic $\boldsymbol{T}$ of $m$ individuals. We will omit the subscript for information taken with respect to a sample $\boldsymbol{T}_1, \ldots, \boldsymbol{T}_n$.

To state the result formally, suppose $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ are independent and identically distributed from one of the $J$ exponential family densities $f(\boldsymbol{x} \mid \boldsymbol{\eta}_1), \ldots, f(\boldsymbol{x} \mid \boldsymbol{\eta}_J)$. Assume that the $J$ densities have the same parametric form, indexed by natural parameter $\boldsymbol{\eta}_\ell$, $\ell = 1, \ldots J$. Denote $\boldsymbol{\theta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_J, \pi_1, \ldots, \pi_{J-1})$ as the full parameter of the finite mixture.[1] Note that $\pi_J = 1 - \sum_{\ell=1}^{J-1} \pi_\ell$ and is therefore not included in the unknown $\boldsymbol{\theta}$. The density of the sufficient statistic $\boldsymbol{T}$ can then be written as

$$f(\boldsymbol{t} \mid \boldsymbol{\theta}) \propto \sum_{\ell=1}^{J} \pi_\ell \exp\left\{\boldsymbol{\eta}_\ell^T \boldsymbol{t} + m \cdot a(\boldsymbol{\eta}_\ell)\right\}.$$

Subsequently, the complete data FIM of $(\boldsymbol{T}, Z)$ is

$$\begin{aligned}\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) &= \text{Blockdiag}\left(\pi_1 \boldsymbol{F}_1, \ldots, \pi_J \boldsymbol{F}_J, \boldsymbol{F}_\pi\right), \quad \text{where}\\ \boldsymbol{F}_\ell &= \text{Var}(\boldsymbol{T} \mid Z = \ell),\\ \boldsymbol{F}_\pi &= \boldsymbol{D}_\pi^{-1} + \pi_J^{-1} \mathbf{1}\mathbf{1}^T,\end{aligned}$$

and where $\boldsymbol{D}_\pi = \text{Diag}(\pi_1, \ldots, \pi_{J-1})$. Notice that $\boldsymbol{F}_\ell$ is the FIM under the density $f(\boldsymbol{t} \mid \boldsymbol{\eta}_\ell)$ and $\boldsymbol{F}_\pi$ is the FIM under the distribution $\text{Mult}_J(1, \boldsymbol{\pi})$.

Raim et al. (2014b) show that $\widetilde{\mathcal{I}}_m(\boldsymbol{\theta}) - \mathcal{I}_m(\boldsymbol{\theta}) \to \mathbf{0}$ as $m \to \infty$. The rate of convergence is seen to depend on the amount of overlap between the mixed densities; the rate may be very slow if some of the $\boldsymbol{\eta}_\ell$ are similar, or very fast if they are all distinct. In practice, if some $\boldsymbol{\eta}_\ell$ are similar, the slow convergence may be avoided by representing them by a common mixture component. It can also be shown that if $\mathcal{I}_m(\boldsymbol{\theta})$ 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 $m \to \infty$.

Analysis with finite mixtures often favors quantities of observed information, such as the Hessian of the log-likelihood or outer product of the score vector, to the expected Fisher information. The Hessian is especially convenient when Newton-Raphson iterations are used for estimation; the inverse Hessian is a byproduct of the iterations and provides an estimate of standard errors. The EM algorithm is another popular approach for estimation in finite mixtures; EM in its pure form does not produce standard errors as a byproduct, but several methods based on observed information have been recommended (McLachlan and Peel, 2000, Chapter 2). For the case of multivariate normal finite mixtures, expressions for the observed information are given by Boldea and Magnus (2009). An easily computed approximation to the FIM may still be of interest despite accessibility to observed information. In theoretical work, the FIM is a quantity that can be of interest itself. Practically, the nonsingularity of $\mathcal{I}(\boldsymbol{\theta})$ depends on the sample only through its argument, while the observed FIM is more vulnerable to singularity if an unlucky sample is obtained.

In the remainder of this paper, we consider model-based clustering in the setting of meta-analysis. Model-based clustering is a natural application of finite mixtures where the densities $f(\boldsymbol{x} \mid \boldsymbol{\eta}_1), \ldots, f(\boldsymbol{x} \mid \boldsymbol{\eta}_J)$ represent subgroups of the overall population which characterize $J$ potential clusters for observations. The corresponding

[1] In the finite mixture setting, there is commonly a lack of identifiability due to label-switching unless special restrictions are made to parameters. It can be seen that the value of $\sum_{\ell=1}^{J} \pi_{\rho(\ell)} f(\boldsymbol{x} \mid \boldsymbol{\phi}_{\rho(\ell)})$ is unchanged for any given $\boldsymbol{x}$ when taking $\rho(\cdot)$ to be any permutation of $(1, \ldots, J)$. The general criteria for identifiability can be modified to allow label switching under finite mixtures and still be meaningful (McLachlan and Peel, 2000, Section 1.14). Label-switching is not a problem for this work, so we make no attempt to prevent it.

$\pi_1, \dots, \pi_J$ represent prior probabilities of belonging to the clusters. Fraley and Raftery (2002) provide more discussion on model-based clustering. Our goal is to determine a clustering for the studies of a meta-analysis. The grouped sampling assumption is quite reasonable in this setting if subjects within a study follow a common distribution.

Section 2 introduces the meta-analysis setup. A simulation study in Section 3 investigates closeness between variance estimates from the complete data FIM and actual FIM under varying study sizes, numbers of studies, and arrangements of subpopulations. Section 4 presents an analysis of an example dataset reporting the content of selenium in nonfat milk powder. Finally, Section 5 concludes the paper.

## 2. Application to Meta-Analysis

Suppose $n$ studies are carried out to evaluate a certain treatment; e.g., a new drug. Let $y_{ij}$ represent the outcome of the $j$th individual in the $i$th study for $i = 1, \dots, n$ and $j = 1, \dots, m_i$. Meta-analysis considers the $n$ studies together to determine whether they are compatible in supporting a common conclusion. If so, they can combine to make a much stronger inference than any one of the studies alone. Often, the data are obtained from published studies which provide summary statistics, test statistics, or p-values, but do not report individual observations. Hartung et al. (2008) present a comprehensive introduction to the subject of meta-analysis.

Testing for homogeneity is recommended as a first step in meta-analysis practice before attempting to combine studies to make an inference. Although apparently not common in the meta-analysis literature, it is possible to use the technique of model-based clustering to explore heterogeneity in the data. Fitting a finite mixture of $J$ densities to the $n$ studies may identify clusters of studies which are similar within cluster and dissimilar between clusters. We may want to be aware of these patterns before attempting to combine studies. Aitkin (1999) has also considered applying finite mixture models to meta-analysis, but with a different interpretation. He assumes a model with a study-level random intercept and uses finite mixtures as a robust alternative against the common assumption that the random intercept follows a normal distribution.

Suppose the observed data are the summary statistics

$$\bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} \quad \text{and} \quad s_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2,$$

along with a study size $m_i$, for studies $i = 1, \dots, n$. Suppose there are $J$ latent populations of normal densities. Conditional on the $i$th study belonging to the $\ell$th population, assume its observations follow

$$y_{ij} \overset{\text{iid}}{\sim} \mathrm{N}(\mu_\ell, \sigma_\ell^2), \quad j = 1, \dots, m_i. \tag{2}$$

We furthermore assume unobserved subpopulation indicators $z_i \overset{\text{iid}}{\sim} \text{Discrete}(1, \dots J; \boldsymbol{\pi})$. In a standard fixed effects model, each study might be characterized by its own $(\mu_i, \sigma_i)$, and homogeneity can be assumed if determined appropriate (Hartung et al., 2008, Chapter 6). Let $\mathbf{1}_m$ denote the $m$-dimensional vector of ones, and $\boldsymbol{I}_m$ denote the $m \times m$ identity matrix. Given $z_i = \ell$, we have the exponential family

$$(y_{i1}, \dots, y_{im_i}) \sim \mathrm{N}\left(\mu_\ell \mathbf{1}_{m_i}, \sigma_\ell^2 \boldsymbol{I}_{m_i}\right). \tag{3}$$

Let $\mathcal{D}_i = (\bar{y}_i, s_i^2)$ represent the observed data for the $i$th study, which is a sufficient statistic for the distribution (3). Also denote $\boldsymbol{\theta}_\ell = (\mu_\ell, \sigma_\ell)$ as the unknown parameter under the $\ell$th density of the mixture. Given $z_i = \ell$, we have the familiar result that the statistics

$$\bar{y}_i \sim \mathrm{N}\left(\mu_\ell, \frac{\sigma_\ell^2}{m_i}\right) \quad \text{and} \quad \frac{(m_i-1)s_i^2}{\sigma_\ell^2} \sim \chi^2_{m_i-1}$$

are independent. This leads to the conditional density

$$f(\mathcal{D}_i \mid z_i = \ell) = \mathrm{N}\left(\bar{y}_i \;\middle|\; \mu_\ell, \frac{\sigma_\ell^2}{m_i}\right) \times \left[\frac{m_i-1}{\sigma_\ell^2}\right] \chi^2\left(\frac{(m_i-1)s_i^2}{\sigma_\ell^2} \;\middle|\; m_i - 1\right),$$

where $\mathrm{N}(\cdot \mid \mu, \sigma^2)$ represents the normal density with mean $\mu$ and variance $\sigma^2$, and $\chi^2(\cdot \mid \nu)$ represents the chi-square density with $\nu$ degrees of freedom. The likelihood of all the observed data with respect to $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J, \boldsymbol{\pi})$ is then

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathcal{D}_i), \quad \text{where } f(\mathcal{D}_i) = \sum_{\ell=1}^{J} \pi_\ell f(\mathcal{D}_i \mid z_i = \ell). \tag{4}$$

We emphasize that (4) will cluster the $n$ studies based on membership in the $J$ normal distributions, as opposed to clustering by test statistics or p-values.

We will consider inference by maximization of the likelihood (4). We will make use of the matrix $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ in two ways: by considering a two-stage hybrid scoring algorithm for estimation and in approximating FIM-based standard errors. For estimation, we first proceed with iterations

$$\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)} + \widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta}^{(g)}) S(\boldsymbol{\theta}^{(g)}), \quad g = 0, 1, \ldots$$

until an initial convergence $|\log L(\boldsymbol{\theta}^{(g)}) - \log L(\boldsymbol{\theta}^{(g-1)})| < \varepsilon_0$ is attained at some iteration $g = g^*$. We then continue with iterations

$$\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)} + [-H(\boldsymbol{\theta}^{(g)})]^{-1} S(\boldsymbol{\theta}^{(g)}), \quad g = g^*, g^*+1, \ldots$$

until the final desired convergence $|\log L(\boldsymbol{\theta}^{(g)}) - \log L(\boldsymbol{\theta}^{(g-1)})| < \varepsilon$. This hybrid scoring approach has a built in robustness to poor starting values from which Newton-Raphson or Fisher scoring may not progress toward a solution (Raim et al., 2014a). The Hessian $H(\boldsymbol{\theta})$ of the log-likelihood may be computed by numerical differentiation. The score function $S(\boldsymbol{\theta})$ consists of the entries

$$\frac{\partial}{\partial \mu_\ell} \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\pi_\ell f(\mathcal{D}_i \mid z_i = \ell)}{f(\mathcal{D}_i)} \left[m_i \frac{\bar{y}_i - \mu_\ell}{\sigma_\ell^2}\right] \quad \text{and}$$
$$\frac{\partial}{\partial \sigma_\ell} \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\pi_\ell f(\mathcal{D}_i \mid z_i = \ell)}{f(\mathcal{D}_i)} \left[-\frac{m_i}{\sigma_\ell} + m_i \frac{(\bar{y}_i - \mu_\ell)^2}{\sigma_\ell^3} + \frac{(m_i-1)s_i^2}{\sigma_\ell^3}\right],$$

for $\ell = 1, \ldots J$, and

$$\frac{\partial}{\partial \pi_\ell} \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{f(\mathcal{D}_i \mid z_i = \ell) - f(\mathcal{D}_i \mid z_i = J)}{f(\mathcal{D}_i)},$$

for $\ell = 1, \ldots J - 1$.

The complete data FIM $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ is readily computed as the $(3J-1) \times (3J-1)$ matrix

$$\widetilde{\mathcal{I}}(\boldsymbol{\theta}) = \text{Blockdiag}(\pi_1 \boldsymbol{F}_1, \ldots, \pi_J \boldsymbol{F}_J, \boldsymbol{F}_\pi),$$

where

$$\boldsymbol{F}_\ell = \text{Diag}\left(\frac{M}{\sigma_\ell^2}, \frac{2M}{\sigma_\ell^2}\right) \quad \text{and} \quad \boldsymbol{F}_\pi = n\left[\boldsymbol{D}_\pi^{-1} + \pi_J^{-1}\mathbf{1}\mathbf{1}^T\right],$$

denoting $M = \sum_{i=1}^n m_i$. Note that $\widetilde{\mathcal{I}}^{-1}(\boldsymbol{\theta})$ is readily obtained in closed form by inverting each block of $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$.

To study the average effect of the treatment, define $\mu_{\text{avg}} = \sum_{\ell=1}^J \pi_\ell \mu_\ell$. This quantity can be estimated by $\sum_{\ell=1}^J \hat{\pi}_\ell \hat{\mu}_\ell$ using the invariance property of maximum likelihood. An estimate of its variance can be obtained by applying the delta method to the large sample variance approximated by $\widetilde{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}})$. Because this variance estimate is derived from the complete data FIM, it requires similar conditions to provide a useful result.

## 3. Simulation Study

Based on the meta-analysis setup discussed in Section 2, we present a simulation to compare the large sample variance obtained from $\widetilde{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}})$ with that obtained from $\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})$. Consider four scenarios with $J = 2$ mixture components:

- Scenario 1: $\boldsymbol{\mu} = (-1, 1)$, $\boldsymbol{\pi} = (0.5, 0.5)$,
- Scenario 2: $\boldsymbol{\mu} = (0, 1)$, $\boldsymbol{\pi} = (0.5, 0.5)$,
- Scenario 3: $\boldsymbol{\mu} = (-1, 1)$, $\boldsymbol{\pi} = (0.9, 0.1)$,
- Scenario 4: $\boldsymbol{\mu} = (0, 1)$, $\boldsymbol{\pi} = (0.9, 0.1)$,

and four scenarios with $J = 3$ mixture components:

- Scenario 1: $\boldsymbol{\mu} = (-1, 0, 1)$, $\boldsymbol{\pi} = (1/3, 1/3, 1/3)$,
- Scenario 2: $\boldsymbol{\mu} = (-1, 0.5, 1)$, $\boldsymbol{\pi} = (1/3, 1/3, 1/3)$,
- Scenario 3: $\boldsymbol{\mu} = (-1, 0, 1)$, $\boldsymbol{\pi} = (0.9, 0.05, 0.05)$,
- Scenario 4: $\boldsymbol{\mu} = (-1, 0.5, 1)$, $\boldsymbol{\pi} = (0.9, 0.05, 0.05)$.

For both settings of $J$, Scenarios 1 and 3 feature better separation between mixture subpopulations, while Scenarios 2 and 4 have subpopulations with more overlap. On the other hand, Scenarios 1 and 2 have subpopulations with equal proportions, while in Scenarios 3 and 4 one group represents a large majority. For all scenarios, we take $\sigma_1 = \sigma_2 = \sigma_3 = 1$, and a common $m_i = m$ for the sample sizes within study. We consider $m \in \{10, 50, 100\}$ and $n \in \{10, 50, 100\}$.

For each scenario and each setting of $m$ and $n$, we repeat the following steps for $r = 1, \ldots, R$, where $R = 500$ is the number of repetitions.

1. Draw $z_i \overset{\text{iid}}{\sim} \text{Discrete}(1, \ldots, J; \boldsymbol{\pi})$. If $\boldsymbol{z} = (z_1, \ldots, z_n)$ does not contain at least one representative from each group $1, \ldots, J$, redraw it.[2]
2. Draw $\bar{y}_i \sim \text{N}\left(\mu_\ell, \sigma_\ell^2/m\right)$. and $s_i^2 \sim \sigma_\ell^2 (m-1)^{-1} \chi_{m-1}^2$ given $z_i = \ell$. Repeat for $i = 1, \ldots, n$.

[2] This is done to avoid problems fitting a $J$ component mixture. Without this provision, we are very likely to draw problematic samples during the course of a simulation, especially for smaller $n$.

3. Compute the MLE $\hat{\boldsymbol{\theta}}^{(r)}$ using the hybrid scoring algorithm.[3]
4. Compute the complete data FIM $\widetilde{\mathcal{I}}(\hat{\boldsymbol{\theta}}^{(r)})$.
5. Compute $\mathcal{I}(\hat{\boldsymbol{\theta}}^{(r)})$ by Monte-Carlo approximation using 20,000 draws from the density (4) using $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(r)}$.
6. Compute the Frobenius norm distance $d_{\text{VAR}}^{(r)}$ between $\widetilde{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}}^{(r)})$ and $\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}}^{(r)})$.
7. Compute the Euclidean distance $d_{\text{SE}}^{(r)}$ between standard errors computed from the diagonal of $\widetilde{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}}^{(r)})$ with those from $\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}}^{(r)})$.

Once these steps are complete, the average distances

$$\bar{d}_{\text{VAR}} = \frac{1}{R}\sum_{r=1}^{R} d_{\text{VAR}}^{(r)} \quad \text{and} \quad \bar{d}_{\text{SE}} = \frac{1}{R}\sum_{r=1}^{R} d_{\text{SE}}^{(r)}$$

are taken to summarize the outcome.

Tables 1 and 2 show the results for $J = 2$ and $J = 3$ respectively. We expect that the distances $\bar{d}_{\text{VAR}}$ and $\bar{d}_{\text{SE}}$ will be decreasing when $m$ and $n$ are increased and when the means are more separated. However, this pattern appears to be broken in several of the cases when $m = 10$ and $J = 3$. Upon closer inspection of the results used to build the tables, $\bar{d}_{\text{VAR}}$ and $\bar{d}_{\text{SE}}$ became large when the MLE converged to a "degenerate" solution. The largest distances in cases $\{J = 3, \text{Scenario } 1, m = 10, n = 50\}$ and $\{J = 3, \text{Scenario } 3, m = 10, n = 10\}$ coincided with a very small component of $\hat{\boldsymbol{\pi}}$ or with components of $\hat{\boldsymbol{\mu}}$ which were very close together. Therefore, in addition to overlap between subpopulations, very rare subpopulations also have a detrimental effect on the approximation. Indeed, the expression $\pi_\ell^{-1}$ can be seen as a multiplier while obtaining rates of convergence in Raim et al. (2014b), but it is de-emphasized as a constant term.

For $J = 2$ and $J = 3$ cases where $m > 10$, degenerate solutions appear not to be an issue and a pattern becomes more clear. Comparing the corresponding entries across the four scenarios for $J = 2$, distances appear to be increasing from Scenario 1 to Scenario 2 to Scenario 3 to Scenario 4. This indicates that distance between subpopulations may not be a serious issue even in Scenarios 2 and 4. Comparing the corresponding entries for $J = 3$ across the four scenarios, distances appear to be increasing from Scenario 1 to Scenario 3 to Scenario 2 to Scenario 4. This reflects that overlap between subpopulations 2 and 3 has an adverse effect on the closeness of $\widetilde{\mathcal{I}}(\boldsymbol{\theta})$ and $\mathcal{I}(\boldsymbol{\theta})$. For all cases, there appears to be some increase in $\bar{d}_{\text{VAR}}$ and $\bar{d}_{\text{SE}}$ when moving from an equal proportion scenario to the corresponding scenario with one prevalent group.

## 4. Finite Mixture Analysis of Selenium Data

Hartung et al. (2008) discuss an example meta-analysis with $n = 4$ studies measuring selenium in nonfat milk powder. The data are shown in Table 3 and were originally reported by Eberhardt et al. (1989). The response $y_{ij}$ represents selenium content in the $j$th observation of the $i$th study.

We fit finite mixtures by maximum likelihood for $J = 2$ and $J = 3$, and also fit the $J = 1$ (single subpopulation, no mixture) case for comparison. Table 4

[3] If the negative Hessian evaluated at $\hat{\boldsymbol{\theta}}^{(r)}$ is not positive semidefinite, we discard the $r$th repetition. This indicates that the solution of the likelihood optimization was not a maximum. We noticed that such repetitions sometimes yielded elements of $\widetilde{\mathcal{I}}(\hat{\boldsymbol{\theta}}^{(r)})$ and $\mathcal{I}(\hat{\boldsymbol{\theta}}^{(r)})$ which differed by orders of magnitude. This and Footnote 2 are related to our coming discussion on "degenerate" solutions.

**Table 1**: Simulation results for $J = 2$.

(a) Scenario 1

| m | n | $\bar{d}_{\texttt{VAR}}$ | $\bar{d}_{\texttt{SE}}$ |
|---|---|---|---|
| 10 | 10 | 0.001549 | 0.003046 |
| | 50 | 0.000206 | 0.000942 |
| | 100 | 0.000097 | 0.000634 |
| 50 | 10 | 0.000292 | 0.000904 |
| | 50 | 0.000060 | 0.000401 |
| | 100 | 0.000028 | 0.000269 |
| 100 | 10 | 0.000230 | 0.000753 |
| | 50 | 0.000049 | 0.000339 |
| | 100 | 0.000024 | 0.000240 |

(b) Scenario 2

| m | n | $\bar{d}_{\texttt{VAR}}$ | $\bar{d}_{\texttt{SE}}$ |
|---|---|---|---|
| 10 | 10 | 0.079266 | 0.080574 |
| | 50 | 0.007517 | 0.029456 |
| | 100 | 0.003574 | 0.020157 |
| 50 | 10 | 0.000328 | 0.001050 |
| | 50 | 0.000061 | 0.000419 |
| | 100 | 0.000029 | 0.000279 |
| 100 | 10 | 0.000230 | 0.000753 |
| | 50 | 0.000049 | 0.000339 |
| | 100 | 0.000024 | 0.000240 |

(c) Scenario 3

| m | n | $\bar{d}_{\texttt{VAR}}$ | $\bar{d}_{\texttt{SE}}$ |
|---|---|---|---|
| 10 | 10 | 0.005189 | 0.007406 |
| | 50 | 0.001454 | 0.003672 |
| | 100 | 0.000549 | 0.002125 |
| 50 | 10 | 0.000379 | 0.001168 |
| | 50 | 0.000086 | 0.000503 |
| | 100 | 0.000038 | 0.000329 |
| 100 | 10 | 0.000243 | 0.000914 |
| | 50 | 0.000052 | 0.000392 |
| | 100 | 0.000022 | 0.000254 |

(d) Scenario 4

| m | n | $\bar{d}_{\texttt{VAR}}$ | $\bar{d}_{\texttt{SE}}$ |
|---|---|---|---|
| 10 | 10 | 0.180094 | 0.148739 |
| | 50 | 0.064043 | 0.097045 |
| | 100 | 0.023615 | 0.062747 |
| 50 | 10 | 0.000617 | 0.001914 |
| | 50 | 0.000135 | 0.000787 |
| | 100 | 0.000052 | 0.000456 |
| 100 | 10 | 0.000243 | 0.000915 |
| | 50 | 0.000052 | 0.000392 |
| | 100 | 0.000022 | 0.000254 |

**Table 2**: Simulation results for $J = 3$.

(a) Scenario 1

| m | n | $\bar{d}_{\texttt{VAR}}$ | $\bar{d}_{\texttt{SE}}$ |
|---|---|---|---|
| 10 | 10 | 0.888527 | 0.260816 |
| | 50 | 0.024406 | 0.073086 |
| | 100 | 0.009997 | 0.048623 |
| 50 | 10 | 0.000774 | 0.002072 |
| | 50 | 0.000115 | 0.000628 |
| | 100 | 0.000057 | 0.000429 |
| 100 | 10 | 0.000459 | 0.001218 |
| | 50 | 0.000089 | 0.000493 |
| | 100 | 0.000045 | 0.000349 |

(b) Scenario 2

| m | n | $\bar{d}_{\texttt{VAR}}$ | $\bar{d}_{\texttt{SE}}$ |
|---|---|---|---|
| 10 | 10 | 0.431405 | 0.280111 |
| | 50 | 0.143226 | 0.202966 |
| | 100 | 0.067697 | 0.165858 |
| 50 | 10 | 0.018799 | 0.037173 |
| | 50 | 0.001805 | 0.011061 |
| | 100 | 0.000855 | 0.007510 |
| 100 | 10 | 0.001524 | 0.005118 |
| | 50 | 0.000198 | 0.001559 |
| | 100 | 0.000096 | 0.001069 |

(c) Scenario 3

| m | n | $\bar{d}_{\texttt{VAR}}$ | $\bar{d}_{\texttt{SE}}$ |
|---|---|---|---|
| 10 | 10 | 0.247619 | 0.229594 |
| | 50 | 0.823684 | 0.341970 |
| | 100 | 0.165029 | 0.216319 |
| 50 | 10 | 0.001732 | 0.004162 |
| | 50 | 0.000681 | 0.002535 |
| | 100 | 0.000290 | 0.001398 |
| 100 | 10 | 0.000499 | 0.001477 |
| | 50 | 0.000170 | 0.000829 |
| | 100 | 0.000083 | 0.000547 |

(d) Scenario 4

| m | n | $\bar{d}_{\texttt{VAR}}$ | $\bar{d}_{\texttt{SE}}$ |
|---|---|---|---|
| 10 | 10 | 0.392268 | 0.252544 |
| | 50 | 0.318579 | 0.269270 |
| | 100 | 0.291981 | 0.254594 |
| 50 | 10 | 0.019179 | 0.035859 |
| | 50 | 0.031873 | 0.040131 |
| | 100 | 0.006490 | 0.022828 |
| 100 | 10 | 0.003068 | 0.008994 |
| | 50 | 0.001557 | 0.006725 |
| | 100 | 0.000636 | 0.003828 |

shows the AIC, BIC, AICC, and log-likelihood for the three fitted models. Table 5 shows estimates, standard errors, and 95% confidence intervals. The intervals for $\pi_1$ and $\pi_2$ are very wide; this is likely due to the large standard error resulting from modest sample sizes. Table 6 shows plug-in estimates of the posterior probabilities $\mathrm{P}(Z_i = \ell \mid \mathcal{D}_i)$ that the $i$th study belongs to the $\ell$th cluster. Table 7 shows standard errors for the MLE $\hat{\boldsymbol{\theta}}$ computed by four methods: the actual information matrix $\mathcal{I}(\hat{\boldsymbol{\theta}})$, the complete data information matrix $\widetilde{\mathcal{I}}(\hat{\boldsymbol{\theta}})$, the Hessian $H(\hat{\boldsymbol{\theta}})$ of the log-likelihood, and the parametric bootstrap variance estimator $\hat{V}_{\text{boot}}$. The matrix $\mathcal{I}(\hat{\boldsymbol{\theta}})$ is computed by Monte-Carlo approximation using 400,000 draws from the density (4) using $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. The Hessian $H(\hat{\boldsymbol{\theta}})$, or more specifically the negative of its inverse, is available as a byproduct of the hybrid scoring method. The matrix $\hat{V}_{\text{boot}}$ is computed by drawing 10,000 bootstrap samples from the density (4) using $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, fitting an MLE $\hat{\boldsymbol{\theta}}^{(b)}_{\text{boot}}$ to each bootstrap sample $b$, and computing the sample variance of the 10,000 bootstrapped MLEs.[4]

We notice that the posterior probability of each observation belonging to one of the $J$ clusters is very close to 1. This can be explained by a result from Raim et al. (2014b), showing that $\mathrm{P}(Z_i = \ell \mid \mathcal{D}_i)$ converges to 1 rapidly when $\ell$ is the "true" cluster, and converges rapidly to 0 otherwise. When $J = 2$ clusters are assumed, cluster 2 consists only of study 3, while cluster 1 consists of the other studies. When $J = 3$ clusters are assumed, studies 2 and 4 are moved from cluster 1 to the new cluster 3.[5] These cluster assignments are reflected in the estimates and standard errors, which do not change between $J = 2$ and $J = 3$ for cluster 2. In cluster 1, the estimate for $\sigma_1$ increases when studies 2 and 4 are moved, and $\pi_1$ is decreased. The estimate of the mean $\mu_1$ changes to reflect that cluster 1 represents study 1. Referring back to the data in Table 3, it is seen that $s_1^2$ is very large, $s_2^2$ and $s_4^2$ are medium, and $s_3^2$ is very small. Therefore, our likelihood-based clustering is primarily based on the sample variance. The reader may have been able to produce this clustering by eyeballing the dataset; however, a more complicated dataset may have treatment and control groups, multiple outcomes, or many more studies. In these cases, model-based clustering can help to identify patterns that are not so obvious.

Table 7 shows that $\widetilde{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}})$ and $\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})$ yield very similar standard errors when $J = 2$. For this case, there is little disadvantage in using the complete data FIM in place of the actual FIM; the advantage being greatly simplified computation. In the $J = 3$ case, the differences in the two types of standard errors are more noticeable. Figure 1 plots the densities $\mathrm{N}(\hat{\mu}_\ell, \hat{\sigma}_\ell^2)$ for $\ell = 1, \ldots, J$. In the case of $J = 3$, we can see that Clusters 1 and 3 are somewhat overlapped which may affect the closeness of $\widetilde{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}})$ and $\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})$. We also notice that the FIM-based standard errors sometimes differ from those computed using $H(\hat{\boldsymbol{\theta}})$ and $\hat{V}_{\text{boot}}$. The goal of this study is not necessarily to suggest which standard errors are preferred. If we are confident about the specification of the model and the method of estimation, and the greatly increased computing workload is not an issue, we might trust the standard errors from $\hat{V}_{\text{boot}}$ above the others.

It is interesting to mention the tests of heterogeneity for the Selenium data reported by Hartung et al. (2008). Eight different procedures are compared to test

[4] As in the simulation study, we discard bootstrap repetitions where $-H(\hat{\boldsymbol{\theta}}^{(b)}_{\text{boot}})$ is not positive semidefinite.

[5] Recalling Footnote 1, it is not necessary that cluster labels $1, \ldots, J$ will retain similar meanings across multiple analyses of the data. In this case however, we can observe how the clusters are related across analyses.

**Table 3**: Selenium data.

| Study | $m$ | $\bar{y}$ | $s^2$ |
|---|---|---|---|
| 1 | 8 | 105.00 | 85.711 |
| 2 | 12 | 109.75 | 20.748 |
| 3 | 14 | 109.50 | 2.729 |
| 4 | 8 | 113.25 | 33.640 |

**Table 4**: Fit statistics for Selenium models.

| | $J = 1$ | $J = 2$ | $J = 3$ |
|---|---|---|---|
| LogLik | -35.9071 | -25.5588 | -23.2146 |
| AIC | 75.8143 | 61.1176 | 62.4293 |
| AICC | 87.8143 | 31.1176 | 33.6293 |
| BIC | 74.5869 | 58.0491 | 57.5196 |

equality of the means of the $n = 4$ studies. Of these, seven do not find significant evidence of heterogeneity. The test that does find evidence of inequality of the means is the ANOVA F-test, which assumes a common variance across all studies. In our clustering, we have similarly found that heterogeneity is due to the variances of the studies and not necessarily due to the means.

## 5. Conclusions

In this paper, we illustrated a model-based clustering application in meta-analysis where the complete data FIM serves as a reasonable approximation for the actual FIM. This setting is naturally compatible with our assumption of grouped sampling; that $m$ individuals from the same group are known to belong to a common subpopulation of the mixture. We find in the Selenium data analysis that the standard errors from the complete data FIM are very close to those from the actual FIM for a $J = 2$ component mixture, but not as close when $J = 3$. The best result will be achieved when all study sizes are large, densities of the finite mixture are distinct from one another, and none are associated with an extremely small proportion of the overall population.

In addition to meta-analysis, we are hopeful that the grouped sampling assumption can be justified in a variety of other kinds of statistical analyses. Users of finite mixtures in these settings could benefit from simplified computation with the complete data FIM.

## Acknowledgements

Thanks to Eric Slud, Thomas Mathew, Lauren Emanuel and Tommy Wright at the U.S. Census Bureau for their critical reading and feedback which has improved this manuscript. We also thank the High Performance Computing Facility at the University of Maryland, Baltimore County (www.umbc.edu/hpcf) for computational resources used in this work.

## References

M. Aitkin. Meta-analysis by random effect modelling in generalized linear models. *Statistics in Medicine*, 18(17–18):2343–2351, 1999.

**Table 5**: Estimates for finite mixture analysis of Selenium data. Standard errors were computed using the inverse of the complete data FIM. 95% confidence intervals are from the t-distribution with 4 degrees of freedom.

(a) $J = 1$ (no mixture).

| | Est. | SE | CI Lo | CI Hi |
|---|---|---|---|---|
| $\mu$ | 109.4286 | 0.8826 | 106.9780 | 111.8791 |
| $\sigma$ | 5.7201 | 0.6241 | 3.9872 | 7.4529 |

(b) $J = 2$.

| | Est. | SE | CI Lo | CI Hi |
|---|---|---|---|---|
| $\mu_1$ | 109.3929 | 1.3067 | 105.7649 | 113.0208 |
| $\sigma_1$ | 6.9143 | 0.9240 | 4.3490 | 9.4796 |
| $\mu_2$ | 109.5000 | 0.4254 | 108.3188 | 110.6812 |
| $\sigma_2$ | 1.5919 | 0.3008 | 0.7566 | 2.4271 |
| $\pi$ | 0.7500 | 0.2165 | 0.1489 | 1.3511 |
| $\mu_{\text{avg}}$ | 109.4196 | 0.9860 | 106.6820 | 112.1573 |

(c) $J = 3$.

| | Est. | SE | CI Lo | CI Hi |
|---|---|---|---|---|
| $\mu_1$ | 105.0950 | 3.1667 | 96.3029 | 113.8871 |
| $\sigma_1$ | 8.6574 | 2.1605 | 2.6589 | 14.6560 |
| $\mu_2$ | 109.5000 | 0.4255 | 108.3187 | 110.6813 |
| $\sigma_2$ | 1.5919 | 0.3008 | 0.7566 | 2.4271 |
| $\mu_3$ | 111.1463 | 1.1506 | 107.9517 | 114.3410 |
| $\sigma_3$ | 5.1109 | 0.8128 | 2.8541 | 7.3677 |
| $\pi_1$ | 0.2530 | 0.2193 | -0.3559 | 0.8620 |
| $\pi_2$ | 0.2450 | 0.2165 | -0.3511 | 0.8511 |
| $\mu_{\text{avg}}$ | 109.2037 | 1.5720 | 104.8391 | 113.5683 |

**Table 6**: For the Selenium analysis, posterior probabilities that each study belongs to clusters $1, \ldots, J$. The symbol $(*)$ marks the cluster with the highest probability for each study.

(a) $J = 2$.

| Study | Group 1 | Group 2 |
|---|---|---|
| 1 | 1.00E+00$*$ | 5.72E−58 |
| 2 | 1.00E+00$*$ | 3.97E−12 |
| 3 | 2.66E−06 | 1.00E+00$*$ |
| 4 | 1.00E+00$*$ | 2.61E−24 |

(b) $J = 3$.

| Study | Group 1 | Group 2 | Group 3 |
|---|---|---|---|
| 1 | 1.00E+00$*$ | 2.10E−58 | 2.31E−04 |
| 2 | 4.32E−03 | 1.77E−12 | 9.96E−01$*$ |
| 3 | 7.22E−09 | 1.00E+00$*$ | 4.32E−05 |
| 4 | 7.95E−03 | 1.53E−24 | 9.92E−01$*$ |

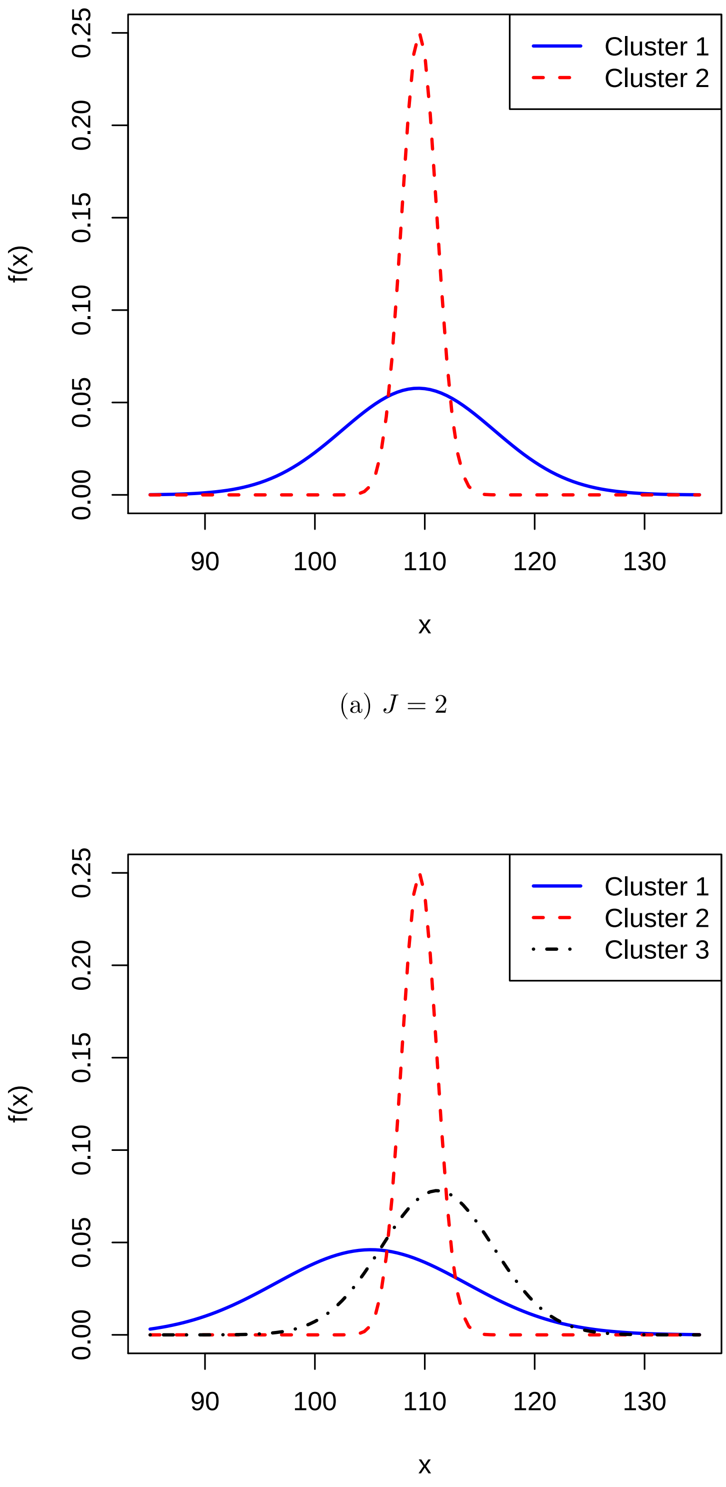
**Table 7**: Standard errors for Selenium data analysis. Each column displays the square roots of the diagonals of the given matrix.

(a) $J = 2$.

| | $-H^{-1}(\hat{\boldsymbol{\theta}})$ | $\widetilde{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}})$ | $\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})$ | $\hat{V}_{\text{boot}}$ |
|---|---|---|---|---|
| $\mu_1$ | 1.3067 | 1.2319 | 1.2329 | 1.4041 |
| $\sigma_1$ | 0.9240 | 0.8711 | 0.8740 | 0.9841 |
| $\mu_2$ | 0.4254 | 0.4913 | 0.4938 | 0.4510 |
| $\sigma_2$ | 0.3008 | 0.3474 | 0.3560 | 0.3473 |
| $\pi$ | 0.2165 | 0.2165 | 0.2172 | 0.1546 |

(b) $J = 3$.

| | $-H^{-1}(\hat{\boldsymbol{\theta}})$ | $\widetilde{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}})$ | $\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})$ | $\hat{V}_{\text{boot}}$ |
|---|---|---|---|---|
| $\mu_1$ | 3.1667 | 2.6558 | 2.9707 | 2.7138 |
| $\sigma_1$ | 2.1605 | 1.8779 | 1.9368 | 1.8407 |
| $\mu_2$ | 0.4255 | 0.4913 | 0.4980 | 0.4774 |
| $\sigma_2$ | 0.3008 | 0.3474 | 0.3661 | 0.3755 |
| $\mu_3$ | 1.1506 | 1.1187 | 1.1941 | 1.5026 |
| $\sigma_3$ | 0.8128 | 0.7910 | 0.8795 | 1.0150 |
| $\pi_1$ | 0.2193 | 0.2174 | 0.2338 | 0.1164 |
| $\pi_2$ | 0.2165 | 0.2165 | 0.2187 | 0.1123 |



(a) $J = 2$

(b) $J = 3$

**Figure 1**: Fitted subpopulation densities $\mathrm{N}(\hat{\mu}_1, \hat{\sigma}_1^2), \ldots, \mathrm{N}(\hat{\mu}_J, \hat{\sigma}_J^2)$ for Selenium data using $J = 2$ and $J = 3$.

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.

K. R. Eberhardt, C. P. Reeve, and C. H. Spiegelman. A minimax approach to combining means, with practical examples. *Chemometrics and Intelligent Laboratory Systems*, 5(2):129–148, 1989.

C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458): 611–631, 2002.

J. Hartung, G. Knapp, and B. K. Sinha. *Statistical Meta-Analysis with Applications*. Wiley, 2008.

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

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

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

A. M. Raim, N. K. Neerchal, and J. G. Morel. An approximation to the information matrix of exponential family finite mixtures, 2014b. Submitted.