Statistica Sinica **30** (2020), 1881-1903
doi:https://doi.org/10.5705/ss.202018.0266

FINITE MIXTURE MODELING, CLASSIFICATION AND STATISTICAL LEARNING WITH ORDER STATISTICS

Armin Hatefi, Nancy Reid, Mohammad Jafari Jozani and Omer Ozturk

Memorial University, University of Toronto,
University of Manitoba and The Ohio State University

Abstract: We propose a unified approach to maximum likelihood estimation, classification, and statistical learning in the context of finite mixture models, based on observations that can be considered a collection of order statistics. We consider both supervised and unsupervised learning approaches. New missing-data mechanisms and expectation-maximization (EM) algorithms are developed to exploit the structure of the observed data in the estimation process under each learning strategy. In addition, we present model-based classification criteria, and show how they can be used to conduct better inferences about rarely observed components in finite mixture models. Using simulation studies, we evaluate the performance of the estimation and classification methodologies. Finally the proposed methods are applied to data from a fishery study to estimate the age structure of Spot, a short-lived fish species.

Key words and phrases: Classification, EM algorithm, finite mixture models, latent variables, order statistics, ranked-set sampling.

1. Introduction

Consider a population of M subpopulations, and suppose we are interested in a random phenomenon, X, with a probability density function (pdf) that can be written as a finite mixture model (FMM). Lastly, we randomly select n sampling units from the population. In many situations, some observations may be missing, possibly at random, but not necessarily; however, we can easily assign ranks to the observed values, and thus retain order statistics. A typical situation occurs in life testing. Here, an experiment is terminated after the first r out of n items under the test have failed, where each item is composed of M components, each with its own lifetime distribution. Observations of this kind are called censored samples, and can lead to the selection of various types of order statistics from samples of size n. A collection of order statistics may also be available when finding the final measurements on all the sampling units is expensive, perhaps owing to budgetary and/or other constraints. In such a case,

an experiment can be scaled back to select a subset of the sampled units for the final study. For example, in studies that need to determine the age of a fish population, it is common practice to first catch a large number of fish, and then to use a subsample to determine the age. In this case, researchers might use systematic sampling to generate the subsample after the larger sample has been ordered by length of fish. For example, they may opt to use every third fish in the ordered sample, which is easy to explain and for field workers to follow. We use the term *selected order statistics* when observations are obtained from specific designs that lead to specific choices of order statistics, for example:

- Single-censored samples from FMMs, where either the r_1 smallest (left-censored) X values or the r_2 largest (right-censored) X values are not observed, with r_1 and r_2 fixed by design (Miyata (2011); Mendenhall and Hader (1958)).
- Doubly censored samples from FMMs, where the r_1 smallest and r_2 largest X values are not observed, with fixed values of r_1 and r_2 (Sindhu, Feroze and Aslam (2016); Saleem, Aslam and Economou (2010)).
- Compressed data from FMMs, where a large number of data points are replaced by a small number of selected order statistics (Bishop (2006)).
- Systematic subsamples, with auxiliary information enabling the ordering of sampled units, as in the fish example described above.

We also use the term *induced order statistics* when, after observing a simple random sample with missing observations, auxiliary information is used to assign a rank to each observation. In all of these examples, observations can be considered collections of order statistics for a sample of size n from an FMM, whether labeled or unlabeled. In other words, we might or might not know the subpopulation from which the data are observed. Then, we can estimate the unknown parameters of the underlying FMM using these data.

Several variations of rank-based sampling (RBS) designs lead to independent order statistics. Inferences for FMMs in these settings are discussed in Hatefi, Jafari Jozani and Ziou (2014); Hatefi, Jafari Jozani and Ozturk (2015). In this study, the order statistics are correlated and finite mixture modeling is a more challenging problem. Thus, we provide a unified approach to statistical inferences for FMMs based on various collections of order statistics. We consider the problem under both supervised and unsupervised learning methods. To obtain maximum likelihood (ML) estimates of the parameters, we introduce new

missing-data mechanisms and expectation-maximization (EM) algorithms that accommodate the dependence structure among the order statistics. This imposes several difficulties in the estimation process, because the log-likelihood function contains terms that are convex combinations of survival functions, which typically do not have a closed form for many statistical distributions. Moreover, we develop new model-based classification criteria for an FMM with rarely observed components.

Section 2 discusses likelihood functions based on unlabeled order statistics of FMMs. The associated EM algorithm and its modified version are explained in Section 3. Section 4 presents various model-based classification criteria. In Section 5, we study estimators of the parameters of FMMs under the supervised learning method. Section 6 compares the performance of several estimation procedures using numerical studies. Then, in Section 7, the proposed estimation procedures are applied to data from a fishery study to determine the age structure of fish. Finally, Section 8 concludes the paper. All proofs, some further remarks, and additional simulation study are provided in the online Supplementary Material.

2. Order Statistics of the FMM

Suppose that the pdf of a random variable of interest X follows a mixture of M component densities

$$f(x;\boldsymbol{\Psi}) = \pi_1 f_1(x;\theta_1) + \cdots + \pi_M f_M(x;\theta_M), \tag{2.1}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$ is a vector of unknown mixing proportions, with $\pi_j > 0$ and $\sum_{j=1}^{M} \pi_j = 1$, and $f_j(\cdot;\theta_j)$, for $j = 1, \ldots, M$, refers to the pdf of the jth component of the FMM, specified up to a vector θ_j of unknown parameters, known a priori to be distinct. Let $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{M-1}, \boldsymbol{\xi})^\top$ denote a vector of all unknown parameters, where $\boldsymbol{\xi} = (\theta_1^\top, \ldots, \theta_M^\top)^\top$, and the superscript $\top$ refers to the vector transpose. The cumulative distribution function (cdf) of X is given by $F(x;\boldsymbol{\Psi}) = \sum_{j=1}^{M} \pi_j F_j(x;\theta_j)$, where $F_j(\cdot;\theta_j)$ represents the cdf of the jth component. For further information on the theory and applications of FMMs, see McLachlan and Peel (2004).

Suppose $\tilde{\mathbf{X}} = \{X_{(i_1)}, X_{(i_2)}, \ldots, X_{(i_k)}\}$, where $1 \le i_1 \le i_2 \le \cdots \le i_k \le n$, is a collection of k, for $k = 2, \ldots, n-1$, order statistics from a random sample of size n from (2.1), where $X_{(i_l)}$ is the i_lth smallest observation in the sample.

According to the theory of order statistics, the log-likelihood function of $\boldsymbol{\Psi}$

based on $\tilde{\mathbf{X}} = \tilde{\mathbf{x}}$ is

$$l(\boldsymbol{\Psi}|\tilde{\mathbf{x}}) \propto \sum_{r=1}^{k} \log f(x_{i_r};\boldsymbol{\Psi}) + (i_1-1)\log F(x_{i_1};\boldsymbol{\Psi}) + (n-i_k)\log \bar{F}(x_{i_k};\boldsymbol{\Psi}) + \sum_{s=2}^{k}(i_s - i_{s-1}-1)\log\left[F(x_{i_s};\boldsymbol{\Psi}) - F(x_{i_{s-1}};\boldsymbol{\Psi})\right], \tag{2.2}$$

and the maximum likelihood estimator (MLE) of $\boldsymbol{\Psi}$, $\widehat{\boldsymbol{\Psi}}_{MLE}$, is obtained as the solution to $\partial l(\boldsymbol{\Psi}|\tilde{\mathbf{x}})/\partial\boldsymbol{\Psi} = 0$ in $\boldsymbol{\Psi}$. The complexity of (2.2) typically makes this intractable, owing to the presence of convex combinations of components of the form $\log f(x_{i_r};\boldsymbol{\Psi})$, $\log F(x_{i_1};\boldsymbol{\Psi})$, $\log\left[F(x_{i_s};\boldsymbol{\Psi}) - F(x_{i_{s-1}};\boldsymbol{\Psi})\right]$, and $\log \bar{F}(x_{i_k};\boldsymbol{\Psi})$. To solve this problem, we model $\tilde{\mathbf{X}} = \tilde{\mathbf{x}}$ as incomplete data. The likelihood and log-likelihood functions based on $\tilde{\mathbf{X}} = \tilde{\mathbf{x}}$ are then called incomplete likelihood and log-likelihood functions, respectively.

To obtain $\widehat{\boldsymbol{\Psi}}_{MLE}$, we construct a new EM algorithm, following the work of Dempster, Laird and Rubin (1977). Let $\boldsymbol{\Delta} = \{\mathbf{Z}_1,\ldots,\mathbf{Z}_k,\mathbf{W}_1,\ldots,\mathbf{W}_{k+1}\}$ be a collection of $2k+1$ latent vectors, each of length M. For each order statistic $X_{(i_r)}$, for $r = 1,\ldots,k$, we define $\mathbf{Z}_r = (Z_{r1},\ldots,Z_{rM})$, with $\mathbf{Z}_r \overset{i.i.d}{\sim} Mult(1,\boldsymbol{\pi})$. We also introduce the following:

- $\mathbf{W}_1 = (W_{11},\ldots,W_{1M})$, with $\mathbf{W}_1 \sim Mult(i_1 - 1, \boldsymbol{\pi})$,
- $\mathbf{W}_s = (W_{s1},\ldots,W_{sM})$, with $\mathbf{W}_s \sim Mult(i_s - i_{s-1} - 1, \boldsymbol{\pi})$, for $s = 2,\ldots,k$, and
- $\mathbf{W}_{k+1} = (W_{k+1\,1},\ldots,W_{k+1\,M})$, with $\mathbf{W}_{k+1} \sim Mult(n - i_k, \boldsymbol{\pi})$.

The complete likelihood function is given by the following lemma; the proof is provided in the Supplementary Material.

Lemma 1. *Let $\tilde{\mathbf{X}} = \{X_{(i_1)}, X_{(i_2)},\ldots,X_{(i_k)}\}$ be a collection of $k = 2,\ldots,n-1$ order statistics from a random sample of size n from (2.1); and let $\boldsymbol{\Delta} = (\mathbf{Z}_1,\ldots,\mathbf{Z}_k,\mathbf{W}_1,\ldots,\mathbf{W}_{k+1})$ be a collection of latent vectors, as defined above. Then the complete-data likelihood function based on $(\tilde{\mathbf{X}},\boldsymbol{\Delta})$ is given by*

$$f(\tilde{\mathbf{x}},\boldsymbol{\delta};\boldsymbol{\Psi}) \propto \prod_{j=1}^{M}\left\{\pi_j F_j(x_{i_1};\theta_j)\right\}^{w_{1j}}\left\{\pi_j \bar{F}_j(x_{i_k};\theta_j)\right\}^{w_{k+1\,j}}\prod_{r=1}^{k}\left\{\pi_j f_j(x_{i_r};\theta_j)\right\}^{z_{rj}} \times\left(\prod_{s=2}^{k}\left[\pi_j\{F_j(x_{i_s};\theta_j) - F_j(x_{i_{s-1}};\theta_j)\}\right]^{w_{sj}}\right).$$

Using Lemma 1, the joint distribution of $(\tilde{\mathbf{X}}, \mathbf{Z}_r)$, for $r = 1, \ldots, k$, is

$$\begin{aligned} f(\tilde{\mathbf{x}}, \mathbf{z}_r) \propto & \{F(x_{i_1}; \boldsymbol{\Psi})\}^{i_1 - 1} \prod_{j=1}^{M} \{\pi_j f_j(x_{i_r}; \theta_j)\}^{z_{rj}} \prod_{\substack{s=1 \\ s \neq r}}^{k} f(x_{i_s}; \boldsymbol{\Psi}) \\ & \times \prod_{s=2}^{k} \left\{F(x_{i_s}; \boldsymbol{\Psi}) - F(x_{i_{s-1}}; \boldsymbol{\Psi})\right\}^{i_s - i_{s-1} - 1} \left\{\bar{F}(x_{i_k}; \boldsymbol{\Psi})\right\}^{n - i_k} \end{aligned} \tag{2.3}$$

In the Supplementary Material, we provide further remarks on the joint pdf of the order statistics and their latent variables.
From (2.3) and the pdf of the order statistics, we can easily show that

$$f_{\mathbf{Z}_r | \tilde{\mathbf{X}}}(\mathbf{z}_r | \tilde{\mathbf{x}}) = \prod_{j=1}^{M} \left\{ \frac{\pi_j f_j(x_{i_r}; \theta_j)}{f(x_{i_r}; \boldsymbol{\Psi})} \right\}^{z_{rj}}, \tag{2.4}$$

and conclude that $\mathbf{Z}_r | \tilde{\mathbf{X}} = \tilde{\mathbf{x}} \sim Mult(1, \pi_1 f_1(x_{i_r}; \theta_1)/f(x_{i_r}; \boldsymbol{\Psi}), \ldots, \pi_M f_M(x_{i_r}; \theta_M) /f(x_{i_r}; \boldsymbol{\Psi}))$, for each $r = 1, \ldots, k$.

Lemma 2. *Let $\mathbf{Z}_r$ be the latent vector associated with $X_{(r)}$, for $r = 1, \ldots, k$. For given order statistics, $\mathbf{Z}_r$ are independent and identically distributed (i.i.d.).*

The proof, taken from Yang (1977), is given in the Supplementary Material.

Based on Remark 5 in the Supplementary Material and the pdf of the order statistics, we have

$$f_{\mathbf{W}_1 | \tilde{\mathbf{X}}}(\mathbf{w}_1 | \tilde{\mathbf{x}}) = \binom{i_1 - 1}{w_{11}, \ldots, w_{1M}} \prod_{j=1}^{M} \left(\frac{\pi_j F_j(x_{i_1}; \theta_j)}{F(x_{i_1}; \boldsymbol{\Psi})} \right)^{w_{1j}}; \tag{2.5}$$

that is, $\mathbf{W}_1 | \tilde{\mathbf{X}} = \tilde{\mathbf{x}} \sim Mult(i_1 - 1, \pi_1 F_1(x_{i_1}; \theta_1)/F(x_{i_1}; \boldsymbol{\Psi}), \ldots, \pi_M F_M(x_{i_1}; \theta_M) /F(x_{i_1}; \boldsymbol{\Psi}))$. Similarly, from Remark 6 in the Supplementary Material, we have

$$f(\mathbf{w}_r | \tilde{\mathbf{x}}) = \prod_{j=1}^{M} \binom{i_r - i_{r-1} - 1}{w_{r1}, \ldots, w_{rM}} \left(\frac{\pi_j [F_j(x_{i_r}; \theta_j) - F_j(x_{i_{r-1}}; \theta_j)]}{F(x_{i_r}; \boldsymbol{\Psi}) - F(x_{i_{r-1}}; \boldsymbol{\Psi})} \right)^{w_{rj}}, \tag{2.6}$$

for each $r = 2, \ldots, k$. Finally, from Remark 7 in the Supplementary Material, we have

$$f(\mathbf{w}_{k+1} | \tilde{\mathbf{x}}) = \binom{n - i_k}{w_{k+1\,1}, \ldots, w_{k+1\,M}} \prod_{j=1}^{M} \left(\frac{\pi_j \bar{F}_j(x_{i_k}; \theta_j)}{\bar{F}(x_{i_k}; \boldsymbol{\Psi})} \right)^{w_{k+1\,j}}. \tag{2.7}$$

From Lemma 1, the complete-data log-likelihood function is

$$
\begin{aligned}
l(\boldsymbol{\Psi}|\tilde{\mathbf{x}},\boldsymbol{\delta}) \propto \sum_{j=1}^{M} \Bigg\{ & w_{1j} \log\left[\pi_j F_j(x_{i_1};\theta_j)\right] + w_{k+1\,j} \log\left[\pi_j \bar{F}_j(x_{i_k};\theta_j)\right] \\
& + \sum_{r=1}^{k} z_{rj} \log\left[\pi_j f_j(x_{i_r};\theta_j)\right] \\
& + \sum_{s=2}^{k} w_{sj} \log\{\pi_j[F_j(x_{i_s};\theta_j) - F_j(x_{i_{s-1}};\theta_j)]\} \Bigg\}.
\end{aligned} \tag{2.8}
$$

3. EM Algorithm

Here, we use the EM algorithm of Dempster, Laird and Rubin (1977) to obtain $\widehat{\boldsymbol{\Psi}}_{MLE}$ using (2.8). To this end, let $\boldsymbol{\Psi}^{(0)}$ be an initial value for $\boldsymbol{\Psi}$.
E-Step: Given $\tilde{\mathbf{X}} = \tilde{\mathbf{x}}$, the conditional expectation of the complete-data log-likelihood function is $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(0)}) = E_{\boldsymbol{\Psi}^{(0)}}[l(\boldsymbol{\Psi}|\tilde{\mathbf{x}})]$, where the expectation is taken under $\boldsymbol{\Psi}^{(0)}$. In the $(p+1)$th iteration, $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(p)})$ is computed in the E-step, where $\boldsymbol{\Psi}^{(p)}$ is the estimate of $\boldsymbol{\Psi}$ obtained from the pth iteration. From (2.4), (2.5), (2.6), and (2.7), we have

$$
\tau_{r,j}(\boldsymbol{\Psi}) = \mathbb{E}\left(Z_{rj}|\tilde{\mathbf{x}}\right) = \frac{\pi_j f_j(x_{i_r};\theta_j)}{f(x_{i_r};\boldsymbol{\Psi})},\ r = 1,\ldots,k;\ j = 1,\ldots,M. \tag{3.1}
$$

$$
\beta_{1,j}(\boldsymbol{\Psi}) = \mathbb{E}\left(W_{1j}|\tilde{\mathbf{x}}\right) = (i_1 - 1)\frac{\pi_j F_j(x_{i_1};\theta_j)}{F(x_{i_1};\boldsymbol{\Psi})},\ j = 1,\ldots,M. \tag{3.2}
$$

$$
\begin{aligned}
\beta_{s,j}(\boldsymbol{\Psi}) = \mathbb{E}\left(W_{sj}|\tilde{\mathbf{x}}\right) &= (i_s - i_{s-1} - 1)\frac{\pi_j[F_j(x_{i_s};\theta_j) - F_j(x_{i_{s-1}};\theta_j)]}{[F(x_{i_s};\boldsymbol{\Psi}) - F(x_{i_{s-1}};\boldsymbol{\Psi})]}, \\
& s = 2,\ldots,k;\ j = 1,\ldots,M.
\end{aligned} \tag{3.3}
$$

$$
\beta_{k+1,j}(\boldsymbol{\Psi}) = \mathbb{E}\left(W_{k+1\,j}|\tilde{\mathbf{x}}\right) = (n - i_k)\frac{\pi_j \bar{F}_j(x_{i_k};\theta_j)}{\bar{F}(x_{i_k};\boldsymbol{\Psi})},\ j = 1,\ldots,M. \tag{3.4}
$$

Combining these with (2.8), the expectation at the $(p+1)$th iteration is

$$
Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(p)}) = Q_1(\boldsymbol{\pi}, \boldsymbol{\Psi}^{(p)}) + Q_2(\boldsymbol{\xi}, \boldsymbol{\Psi}^{(p)}), \tag{3.5}
$$

$$
Q_1(\boldsymbol{\pi}, \boldsymbol{\Psi}^{(p)}) = \sum_{j=1}^{M} \log \pi_j \left\{ \sum_{r=1}^{k} \tau_{r,j}(\boldsymbol{\Psi}^{(p)}) + \sum_{s=1}^{k+1} \beta_{s,j}(\boldsymbol{\Psi}^{(p)}) \right\},
$$

$$
Q_2(\boldsymbol{\xi}, \boldsymbol{\Psi}^{(p)}) = \sum_{j=1}^{M} \Bigg[\beta_{1,j}(\boldsymbol{\Psi}^{(p)}) \log F_j(x_{i_1};\theta_j) + \beta_{k+1,j}(\boldsymbol{\Psi}^{(p)}) \log \bar{F}_j(x_{i_k};\theta_j)
$$

$$+\sum_{r=1}^{k}\tau_{r,j}(\boldsymbol{\Psi}^{(p)})\log f_j(x_{i_r};\theta_j)$$
$$+\sum_{s=2}^{k}\beta_{s,j}(\boldsymbol{\Psi}^{(p)})\log\{F_j(x_{i_s};\theta_j)-F_j(x_{i_{s-1}};\theta_j)\}\Bigg].$$

M-Step: In the $(p+1)$th iteration of the M-step, $Q(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(p)})$ is maximized with respect to $\boldsymbol{\Psi}$ to obtain $\boldsymbol{\Psi}^{(p+1)}$. From (3.5), the estimate $\hat{\boldsymbol{\pi}}^{(p+1)}$ is updated by maximizing $Q_1(\boldsymbol{\pi},\boldsymbol{\Psi}^{(p)})$ with respect to $\boldsymbol{\pi}$. Owing to the constraint $\sum_{j=1}^{M}\pi_j=1$, we use the Lagrangian multiplier to update the mixing proportions π_j, for $j=1,\ldots,M-1$, as follows:

$$\hat{\pi}_j^{(p+1)}=\frac{1}{n}\left\{\sum_{s=1}^{k}\tau_{s,j}(\boldsymbol{\Psi}^{(p)})+\sum_{s=1}^{k+1}\beta_{s,j}(\boldsymbol{\Psi}^{(p)})\right\}. \tag{3.6}$$

Using $Q_2(\boldsymbol{\xi},\boldsymbol{\Psi}^{(p)})$ in (3.5), we obtain $\boldsymbol{\xi}^{(p+1)}$ as the solution to

$$\boldsymbol{\xi}^{(p+1)}=\arg\max_{\boldsymbol{\xi}}Q_2(\boldsymbol{\xi},\boldsymbol{\Psi}^{(p)}). \tag{3.7}$$

Finally, the $\widehat{\boldsymbol{\Psi}}_{MLE}$ of FMM (2.1) is computed by iterating the the E-step and the M-step until the algorithm converges.

3.1. Modified EM algorithm

In the algorithm proposed above, each M-step requires finding a solution to (3.7). Thus, updating $\boldsymbol{\xi}$ is cumbersome, computationally expensive, and affects the convergence rate of the algorithm. This intractability is due to the terms of $\partial\log F_j(x_{(i_1)};\theta_j)/\partial\boldsymbol{\xi}$, $\partial\log(1-F_j(x_{(i_k)};\theta_j))/\partial\boldsymbol{\xi}$, and $\partial\log\{F_j(x_{(i_s)};\theta_j)-F_j(x_{(i_{s-1})};\theta_j)\}/\partial\boldsymbol{\xi}$ in the log-likelihood function. When the cdf of the component densities does not have a closed form, which is the case for most commonly used distributions, the dependence structures among the order statistics make the computations extensive and time consuming. To solve this problem, Johnson and Mehrotra (1972) and Mehrotra and Nanda (1974) proposed a modification technique in which the expectation of the likelihood function is maximized to obtain the MLE. Recently, Hatefi, Jafari Jozani and Ozturk (2015) employed this modified approach for FMM analyses under various RBS designs. Using the properties of the RBS, where the order statistics are independent, they showed that the M-step for $\boldsymbol{\xi}$ in the EM algorithm reduces to the M-step in the usual simple random sampling EM algorithm. Unfortunately, owing to the dependence

structure among the order statistics, this is not the case in the EM algorithm under correlated order statistics. Based on their work, we propose computing the M-step of the EM algorithm for estimating $\boldsymbol{\xi}$ using the M-step for $\boldsymbol{\xi}$ of an EM algorithm for SRS data. However, despite the similarity in updating $\boldsymbol{\xi}$, note that the observations are order statistics of the FMMs. Accordingly, instead of equation (3.7), the following modified estimating equation is used to update $\boldsymbol{\xi}$:

$$\hat{\boldsymbol{\xi}}^{(p+1)} = \arg\max_{\boldsymbol{\xi}} \sum_{s=1}^{k} \sum_{j=1}^{M} \left\{ \tau_{s,j}(\boldsymbol{\Psi}^{(p)}) \log f_j(x_{i_s}; \theta_j) \right\}, \tag{3.8}$$

where $\tau_{s,j}(\boldsymbol{\Psi}^{(p)})$ is defined in (3.1). This updating step for $\boldsymbol{\xi}$ is the same as that under SRS data, but we still take advantage of the information in the order statistics and their latent variables when updating the mixing proportions in each step. This indirectly affects the estimation of $\boldsymbol{\xi}$.

4. Classification

Once the parameters of the FMM are estimated, we can determine the component membership of each observation. Based on the characteristics of the order statistics of the FMM, we propose several model-based classification criteria. These criteria enable us to determine the component membership of the observations, and to make probabilistic inferences about rarely observed component(s) in FMMs. We first focus on the classification of a sample of order statistics from an FMM.

Suppose we have observed $X_{(r)} = x_{(r)}$. To classify $x_{(r)}$, we estimate its component membership vector $\mathbf{Z}_r = (Z_{r1}, \ldots, Z_{rM})$ by $\widehat{\mathbf{Z}}_r$, where

$$\hat{Z}_{rj} = \begin{cases} 1, \text{ if } j = \text{argmax}_h \eta_h(x_{(r)}; \boldsymbol{\Psi}), \\ 0, \text{ otherwise,} \end{cases}$$

for $j = 1, \ldots, M$, and $\eta_h(x_{(r)}; \boldsymbol{\Psi}) = \mathbb{P}(Z_{rh} = 1 | x_{(r)}; \boldsymbol{\Psi})$. From (2.4), the posterior distribution of $\mathbf{Z}_r$ given $X_{(r)} = x_{(r)}$, is given by

$$\mathbb{P}(\mathbf{Z}_r = \mathbf{z}_r | x_{(r)}) = \binom{1}{z_{r1}, \ldots, z_{rM}} \prod_{h=1}^{M} \left\{ \frac{\pi_h f_h(x_{(r)}; \theta_h)}{f(x_{(r)}; \boldsymbol{\Psi})} \right\}^{z_{rh}};$$

thus,

$$\eta_h(x_{(r)};\boldsymbol{\Psi}) = \frac{\pi_h\, f_h(x_{(r)};\theta_h)}{f(x_{(r)};\boldsymbol{\Psi})}. \tag{4.1}$$

The posterior probabilities $\eta_h(x_{(r)};\boldsymbol{\Psi})$ are then estimated by $\eta_h(x_{(r)};\hat{\boldsymbol{\Psi}}_{MLE})$. Using the classifier in (4.1), we assign each observation to the component that has the highest estimated posterior probability. Note that the expression obtained in (4.1) as the posterior probability of component membership of each order statistic is equal to the commonly used expression for the SRS design. However, the parameters are estimated using the order statistics of the FMM in (2.1).

The following remark describes the classification of unobserved X_l, given observed order statistics X_r, where $l \leq r$; other classification scenarios are summarized as Remarks 8 and 9 in the Supplementary Material.

Remark 1. Given $X_{(r)} = x_{(r)}$ and its label $\mathbf{Z}_{(r)} = \mathbf{z}_{(r)}$, suppose we are now interested in classifying an unobserved order statistic $X_{(l)}$, for $l \leq r$. To this end, the component membership vector $\mathbf{Z}_l = (Z_{l1}, \ldots, Z_{lM})$ can be estimated by $\widehat{\mathbf{Z}}_l$, where

$$\hat{Z}_{lj} = \begin{cases} 1, \text{ if } j = \text{argmax}_h \alpha_h(x_{(l)};\boldsymbol{\Psi}), \\ 0, \text{ otherwise,} \end{cases}$$

and $\alpha_h(x_{(l)};\boldsymbol{\Psi}) = \mathbb{P}(Z_{lj} = 1 | x_{(r)}, \mathbf{z}_{(r)};\boldsymbol{\Psi})$. From Remark 2 in the Supplementary Material, the posterior distribution of $\mathbf{Z}_l$ is given by

$$\mathbb{P}(\mathbf{Z}_l = \mathbf{z}_l | \mathbf{Z}_r = \mathbf{z}_r, x_{(r)}) = \binom{1}{z_{l1}, \ldots, z_{lM}} \prod_{h=1}^{M} \left\{ \frac{\pi_h F_h(x_{(r)};\theta_h)}{F(x_{(r)};\boldsymbol{\Psi})} \right\}^{z_{lh}};$$

consequently, $\alpha_h(x_{(r)};\boldsymbol{\Psi}) = \pi_h\, F_h(x_{(r)};\theta_h)/F(x_{(r)};\boldsymbol{\Psi})$. In other words, given the observed value y for the the rth order statistic $X_{(r)}$ selected from a sample of size n from the FMM, missing (unselected) order statistics smaller than y are classified into the jth component of the FMM if $\alpha_j(y;\hat{\boldsymbol{\Psi}}) > \alpha_h(y;\hat{\boldsymbol{\Psi}})$, for all $h = 1, \ldots, M; j \neq h$.

Next we investigate how to use the properties of the order statistics of FMMs with rarely observed component(s). In other words, we determine the probability of observing at least m observations from these rare components. These probabilities are studied in Lemmas 4, 5, and 6; the proofs are provided in the

Supplementary Material. We first state the following result from David and Nagaraja (1981).

Lemma 3. *Let X be a random variable with cdf $F(\cdot;\mathbf{\Psi})$. Then,*

$$\sum_{i=r}^{n}\binom{i}{n}[F(x;\mathbf{\Psi})]^i[\bar{F}(x;\mathbf{\Psi})]^{n-i}=\mathbb{I}_{F(x;\mathbf{\Psi})}(r,n-r+1), \tag{4.2}$$

where $\mathbb{I}_{F(x;\mathbf{\Psi})}(r,n-r+1)=(1/B(r,n-r+1))\int_0^{F(x;\mathbf{\Psi})}t^{r-1}(1-t)^{n-r}dt$, and $B(a,b)=\Gamma(a+b)/(\Gamma(a)\Gamma(b))$.

Lemma 4. *Let $X_{(r)}=x_r$ be the observed rth order statistic from the FMM in* (2.1)*, based on a random sample of size n. For $m=1,\ldots,r-1$, let $T^1_{m,j}$ denote the event of observing at least m sample points less than $X_{(r)}$ from component j; then, we have $\mathbb{P}(T^1_{m,j}|x_r)=\mathbb{I}_{G_1(x_r)}(m,r-m)$, where $G_1(x_r)=\pi_jF_j(x_r;\theta_j)/F(x_r;\mathbf{\Psi})$ and $j=1,\ldots,M$. In addition, let S^1_j denote the event of observing no sample points less than $X_{(r)}$ from component j; then, we have $\mathbb{P}(S^1_j|x_r)=1-\mathbb{I}_{G_1(x_r)}(1,r-1)$.*

Lemma 5. *Let $X_{(r)}=x_r$ and $X_{(l)}=x_l$ be the observed rth and lth order statistics, respectively, for $r<l$, for the FMM in* (2.1) *from a sample of size n. Let $T^2_{m,j}$ denote the event of observing at least m sample points between $X_{(r)}$ and $X_{(l)}$ from component j; then, we have $\mathbb{P}(T^2_{m,j}|x_r,x_l)=\mathbb{I}_{G_2(x_r,x_l)}(m,l-r-m)$, for $m=1,\ldots,l-r-1$, where $G_2(x_r,x_l)=\pi_j[F_j(x_l;\theta_j)-F_j(x_r;\theta_j)]/[F(x_l;\mathbf{\Psi})-F(x_r;\mathbf{\Psi})]$ and $j=1,\ldots,M$. Therefore, let S^2_j denote the event of observing no sample points between $X_{(r)}$ and $X_{(l)}$ from component j; then, we have $\mathbb{P}(S^2_j|x_r,x_l)=1-\mathbb{I}_{G_2(x_r,x_l)}(1,l-r-1)$.*

Lemma 6. *Let $X_{(l)}=x_l$ be the observed lth order statistic from the FMM in* (2.1) *based on a random sample of size n. For m; $m=1,\ldots,n-l-1$, let $T^3_{m,j}$ denote the event of observing at least m sample points greater than $X_{(l)}$ from component j; then, we have $\mathbb{P}(T^3_{m,j}|x_l)=\mathbb{I}_{G_3(x_l)}(m,n-l-m+1)$, where $G_3(x_l)=\pi_j\bar{F}_j(x_l;\theta_j)/\bar{F}(x_l;\mathbf{\Psi})$ and $j=1,\ldots,M$. Further, let S^3_j denote the event of observing no sample points greater than $X_{(l)}$ from component j; then, we have $\mathbb{P}(S^3_j|x_l)=1-\mathbb{I}_{G_3(x_l)}(1,n-l)$.*

As mentioned in Section 1, in many environmental, ecological, and medical studies, measuring the variable of interest is difficult and time-consuming. However, rank information can usually be obtained easily, as in the example of determining the age of fish based on their length, as described in the introduction. Hatefi, Jafari Jozani and Ozturk (2015) exploited the properties of a ranked

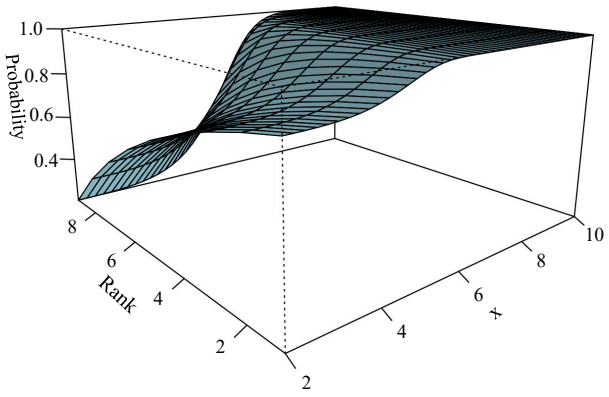

Figure 1. The probability of observing at least one observation from the second component when the set size is 10.

set sampling (RSS) design under perfect ranking to analyze the age of fish based on length frequency data. To obtain a sample of k fish, a simple random sample of k^2 fish is selected first, these fish are then divided randomly into k sets of size k. Then, in each set, fish are ranked based on their length and, finally, the ith smallest fish from set i is selected for age determination. In the following example, we use Lemma 6, for a perfect RSS (i.e., there is no ranking error in the sampling process) as an example of order statistics of FMMs.

Example 1. Consider a perfect RSS, with set size $H = 10$, from a mixture of two normal distributions with $\mathbf{\Psi} = \{\pi, \mu_1, \mu_2, \sigma_1, \sigma_2\} = \{0.8, 4.87, 8, 1, 2\}$. Figure 1 shows the probability of observing at least one observation from the second component. For example, Given $x_{(5)} = 4$, the probability of observing at least $m = 3$ units of $H = 10$ sampling units from a rare population (second component with $\pi = 0.2$) is 0.0856. Figure 1 shows that if the rank is fixed, then, as the value of x increases, the probability of observing a sample from the rare component increases. Furthermore, if x is fixed, then as the rank of x increases, the probability of observing a sample from the rare event decreases.

5. Statistical Learning with Order Statistics

In this section, we study how the notion of order statistics can be incorporated into supervised and unsupervised learning in the context of FMMs. As in the previous section, we use the properties of order statistics to make inferences about FMMs in the context of unsupervised learning, where information about

the component membership of the order statistics is not available. Because the cost of obtaining k order statistics is the same as that of ordering the entire sample, we examine the order statistics under unsupervised learning for the sake of completeness, in the context of estimation, classification, and the consistency of the results. This enables us to better compare the performance of the proposed methods with that of their counterparts under supervised learning, particularly in settings in which measuring the labeled data is difficult. In this section, we study the order statistics of FMMs in the context of supervised learning. In this case, both measured values of the order statistics and their component memberships are available.

In Subsection 5.1, we revisit the results of Section 2 for the order statistics of FMMs in an unsupervised learning setting, after which, we examine the order statistics of an FMM for supervised learning. Suppose $\mathbf{X} = (X_1, \ldots, X_k)$ represents a collection of unlabeled SRS data of size k from the FMM in (2.1). In the case of labeled SRS data, for each observation X_i for $i = 1, \ldots, k$, let $\mathbf{Z}_i^* = \{z_{i1}^*, \ldots, z_{iM}^*\}$ be the observed label, such that $z_{ij}^* = 1$ if X_i is from component j and is zero otherwise.

5.1. Unsupervised learning using ordered statistics from FMMs

Suppose we only have access to the unlabeled SRS data $\mathbf{x} = (x_1, \ldots, x_k)$; in this case, the likelihood function becomes $L_{un}(\boldsymbol{\Psi}|\mathbf{x}) = \prod_{i=1}^{k} \sum_{j=1}^{M} \pi_j f_j(x_i; \theta_j)$. As in Section 2, we introduce the latent variables $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iM})$, for $i = 1, \ldots, k$ and for each x_i, such that $Z_{ij} = 1$ if x_i comes from component j of the FMM, and $Z_{ij} = 0$ otherwise. Now, let $\mathcal{Y}_{un} = (\mathbf{X}, \mathbf{Z})$ denote the complete data with likelihood function

$$L_{un}(\boldsymbol{\Psi}|\mathcal{Y}_{un}) = \prod_{i=1}^{k} \prod_{j=1}^{M} \{\pi_j f_j(y_i; \theta_j)\}^{z_{ij}}. \tag{5.1}$$

As in Section 3, we obtain ML estimates of the parameters using the EM algorithm. The conditional expectation of $Z_{ij}|\mathbf{y}$, computed in the E-step, is used in the $(p+1)$th step to update $\boldsymbol{\Psi}_{un}^{(p+1)} = (\pi_{un}^{(p+1)}, \boldsymbol{\xi}_{un}^{(p+1)})$, as follows:

$$\hat{\pi}_{un,j}^{(p+1)} = \frac{1}{k} \sum_{i=1}^{k} \tau_{r,j}(\boldsymbol{\Psi}^{(p)}),\ j = 1, \ldots, M, \tag{5.2}$$

$$\hat{\boldsymbol{\xi}}_{un}^{(p+1)} = \arg\max_{\boldsymbol{\xi}} \sum_{i=1}^{k} \sum_{j=1}^{M} \left\{ \tau_{r,j}(\boldsymbol{\Psi}^{(p)}) \log f_j(y_i; \theta_j) \right\}, \tag{5.3}$$

where $\tau_{r,j}(\mathbf{\Psi}^{(p)}) = \mathbb{E}(Z_{rj}|\mathbf{y})$, for $r = 1, \ldots, k$.

Let $\tilde{\mathbf{X}}_{ou} = \{X_{(i_1)}, \ldots, X_{(i_k)}\}$ be the collection of order statistics of unlabeled data $\mathbf{X}$ from a sample of size n. Let $\mathcal{Y}_{ou} = (\tilde{\mathbf{X}}_{ou}, \mathbf{\Delta})$ denote the complete order statistics, consisting of the unlabeled order statistics and their latent variables. According to Lemma 1, the likelihood function based on $\mathcal{Y}_{ou}$ can be written as

$$L(\mathbf{\Psi}|\mathcal{Y}_{ou}) \propto L(\mathbf{\Psi}|\mathcal{Y}_{un})\, \kappa(\mathbf{\Psi}|\mathcal{Y}_{ou}), \tag{5.4}$$

where

$$\begin{aligned}\kappa(\mathbf{\Psi}|\mathcal{Y}_{ou}) = &\prod_{j=1}^{M} \Big\{ \{\pi_j F_j(y_{(i_1)};\theta_j)\}^{w_{1j}} \{\pi_j \bar{F}_j(y_{(i_k)};\theta_j)\}^{w_{m+1,j}} \\ &\times \prod_{s=2}^{k} \left\{\pi_j [F_j(y_{(i_s)};\theta_j) - F_j(y_{(i_{s-1})};\theta_j)]\right\}^{w_{sj}} \Big\}. \end{aligned} \tag{5.5}$$

From (5.1), it is apparent the $\kappa(\mathbf{\Psi}|\mathcal{Y}_{ou})$ is the contribution of k order statistics to the unsupervised learning of FMMs.

5.2. Supervised learning with ordered statistics of FMMs

In this subsection, we analyze FMMs using labeled data. For SRS supervised learning, we estimate the parameters based on the labeled data. The likelihood function based on these observations is

$$L_{us}(\mathbf{\Psi}|\mathbf{x}, \mathbf{z}^*) = \prod_{i=1}^{k} \prod_{j=1}^{M} \{\pi_j f_j(x_i;\theta_j)\}^{z^*_{ij}}. \tag{5.6}$$

Using (5.6), the ML estimate $\hat{\mathbf{\Psi}}_{us}$ is

$$\hat{\pi}_{us,j} = \frac{1}{k} \sum_{i=1}^{k} z^*_{ij}, \tag{5.7}$$

$$\hat{\theta}_j = \arg\max_{\theta_j} \sum_{i=1}^{k} \log f_j(x_i;\theta_j),\ j = 1, \ldots, M. \tag{5.8}$$

Here, we show how to exploit the properties of order statistics to make inferences for FMMs using labeled data. Let $\tilde{\mathbf{X}}_{os} = \{X_{(i_1)}, \ldots, X_{(i_k)}\}$ be the collection of k order statistics for the labeled data $\mathbf{X}$ from a sample of size n, with labels $\mathbf{Z}^* = \{Z^*_1, \ldots, Z^*_k\}$. Using the pdf of the order statistics, the likelihood

function based on $(\tilde{\mathbf{X}}_{os}, \mathbf{Z}^*)$ is

$$\begin{aligned} L_{os}(\boldsymbol{\Psi}|\tilde{\mathbf{x}}_{os}, \mathbf{z}^*) \propto & \left\{F(x_{(i_1)}; \boldsymbol{\Psi})\right\}^{i_1-1} \left\{\bar{F}(x_{(i_k)}; \boldsymbol{\Psi})\right\}^{n-i_k} \\ & \times \prod_{s=2}^{k} \left\{F(x_{(i_s)}; \boldsymbol{\Psi}) - F(x_{(i_{s-1})}; \boldsymbol{\Psi})\right\}^{i_s - i_{s-1}-1} \\ & \times \prod_{r=1}^{k} \prod_{j=1}^{M} \left\{\pi_j f_j(x_{(i_r)}; \theta_j)\right\}^{z^*_{rj}}. \end{aligned} \tag{5.9}$$

In order to obtain the ML estimate of $\boldsymbol{\Psi}$, we introduce the latent vectors $\mathbf{W_s} = (W_{s1}, \ldots, W_{sM})$, for $s = 1, \ldots, k+1$. Let $\mathcal{Y}_{os} = (\tilde{\mathbf{X}}_{os}, \mathbf{Z}^*, \mathbf{W})$ denote the complete labeled order statistics. Similarly to Lemma 1, the complete likelihood function version of (5.9) is given by

$$L(\boldsymbol{\Psi}|\mathcal{Y}_{os}) \propto L(\boldsymbol{\Psi}|\mathcal{Y}_{us})\, \kappa(\boldsymbol{\Psi}|\mathcal{Y}_{os}), \tag{5.10}$$

where $\kappa(\boldsymbol{\Psi}|\mathcal{Y}_{os})$ is defined in (5.5) by replacing y_{i_j} with $x_{(i_j)}$. From (5.10), it is apparent that $\kappa(\boldsymbol{\Psi}|\mathcal{Y}_{os})$ shows the contribution of k order statistics from a sample of size n to the supervised FMM. Now, we estimate the parameters of the FMM using the EM algorithm presented in Section 3. The E-step requires only the conditional expectation of the latent variables $\mathbf{W}_s$, for $s = 1, \ldots, n$, given $\tilde{\mathbf{x}}_{os}, \mathbf{z}^*$. As in Section 3, using (3.2), (3.3), and (3.4), the parameters are updated on the $(p+1)$th step using

$$\hat{\pi}^{(p+1)}_{os,j} = \frac{1}{n} \left\{ \sum_{s=1}^{k} z^*_{sj} + \sum_{s=1}^{k+1} \beta_{s,j}(\boldsymbol{\Psi}^{(p)}) \right\}, \tag{5.11}$$

where $j = 1, \ldots, M-1$ and, on the $(p+1)$th iteration of the M-step, the estimates of the component parameters $\boldsymbol{\xi}^{(p+1)}_{os}$ are updated using

$$\boldsymbol{\xi}^{(p+1)}_{os} = \arg\max_{\boldsymbol{\xi}} Q_{os}(\boldsymbol{\xi}, \boldsymbol{\Psi}^{(p)}), \tag{5.12}$$

where

$$\begin{aligned} Q_{os}(\boldsymbol{\xi}, \boldsymbol{\Psi}^{(p)}) = & \sum_{j=1}^{M} \Big\{ \beta_{1,j}(\boldsymbol{\Psi}^{(p)}) \log F_j(x_{i_1}; \theta_j) + \beta_{k+1,j}(\boldsymbol{\Psi}^{(p)}) \log \bar{F}_j(x_{i_k}; \theta_j) \\ & + \sum_{r=1}^{k} z^*_{rj} \log f_j(x_{i_r}; \theta_j) \end{aligned}$$

$$+\sum_{s=2}^{k}\beta_{s,j}(\boldsymbol{\Psi}^{(p)})\log[F_j(x_{i_s};\theta_j)-F_j(x_{i_{s-1}};\theta_j)]\Big\}.$$

Then, the E-step and M-step are repeated until the algorithm converges.

6. Numerical Studies

In this section, we empirically study the performance of the MLEs of the FMM parameters under various order statistics designs D_i, for $i = 1, \ldots, 6$, as shown in Table 1. In all designs, the original simple random sample size is assumed to be $n = 30$, where we observe only k order statistics, for $k \in \{6, 8, 10\}$. We select D_i such that the performance of $\widehat{\boldsymbol{\Psi}}_{MLE}$ can be evaluated under different scenarios, including right- and left- censoring schemes (D_1, D_2), a modified version of maxima-minima nominated sampling (D_3, D_4), and systematic sampling (D_5). The MLEs of the parameters of the FMMs are computed assuming we have labeled order statistics, unlabeled order statistics, labeled SRS data, and unlabeled SRS data. We used the modified EM algorithm to compute $\widehat{\boldsymbol{\Psi}}_{MLE}$. The underlying FMM is assumed to be a mixture of two univariate normal distributions,

$$f(x;\boldsymbol{\Psi}) = \pi\phi(x;\mu_1,\sigma) + (1-\pi)\phi(x;\mu_2,\sigma). \tag{6.1}$$

with parameters $\boldsymbol{\Psi} = \{\pi, \mu_1, \mu_2, \sigma\}$. Owing to the key role of mixing the proportion parameters in mixture modeling, we investigate two simulation studies. The first, described in Subsection 6.1, estimates the mixing proportion, where the component parameters are assumed to be known. The second, provided in the Supplementary Material, estimates all parameters of the model. We investigate the performance of the estimation and classification procedures based on designs D_i, and compare it with the case in which observations are simple random samples. Note that we do not necessarily suggest using order statistics for finite mixture modeling as a sampling scheme to replace SRS, but rather as a natural setting that happens in many real-world applications. The goal is to show how the rank information provided by different collections of order statistics can affect the estimation and classification processes. To generate observations using D_i, for each simulation, we take a sample of size $n = 30$ from (6.1). After ranking the observations, we select the order statistics using the designs shown in Table 1. When using an unsupervised approach, we consider only the value of the selected order statistics, whereas in a in supervised approach, we observe both the

Table 1. Collections of order statistics.

Design	Collection of Order Statistics	Experiment (k=size)
D_1	{1, 2, 3, 4, 5, 6}	Right censored data (6)
D_2	{23, 24, 25, 26, 27, 28, 29, 30}	Left censored data (8)
D_3	{1, 2, 3 , 28, 29, 30}	Modified MMN sample (6)
D_4	{1, 2, 3, 4, 5, 26, 27, 28, 29, 30}	Modified MMN sample (10)
D_5	{1, 5, 10, 20, 25, 30}	Systematic selection (6)

selected order statistics and their component memberships.

6.1. Simulation study 1

We first estimate π and evalute the classification performance when the component parameters of the FMM are assumed to be known. Using Table 1, we generate samples from model (6.1). We consider $(\mu_1, \mu_2, \sigma) = (9.01, 11.70, 1.15)$ and $\pi \in \{0.35, 0.50, 0.60, 0.67, 0.80\}$, such that the component parameters are the same as those for Spot data analyzed in Section 7. The modified EM algorithm, described in Subsection 3.1 is carried out 5,000 times, with initial value 0.5, for π, with stopping criteria $|\pi^{(k+1)} - \pi^{(k)}| < 10^{-6}$.

Tables 2 and 3 provide the biases, square root of the mean squared errors ($\sqrt{MSE}$), classification precisions (CLP%), and convergence rates (CVR%) for all estimation procedures. The classification precision rate (CLP%) is the average proportion of correct classification rates over 5,000 simulations. The simulation studies are devised so that we have access to the true component membership of the sampling units under all estimation procedures. Comparing the true and predicted memberships of the test data, we compute the correct classification rate of the classifiers for each estimator in each simulation. The rate of convergence (CVR%) is calculated as the average number of times that the estimation procedure converged over 5,000 replications. Comparing the ML estimates of π under each design D_i, we clearly observe the impact of various collections of order statistics on the estimation and classification procedures. For instance, from Table 2, when $\pi = 0.8$, design D_1 practically fails to capture the rare event (i.e., the second component), yielding a convergence rate for the estimation procedure of about 1%. On the other hand, using the collection of upper order statistics (design D_2) guarantees that we will observe data from the rare component and, consequently, improves the convergence rate of the estimation procedures by 93%.

The relative efficiency (RE) of the proposed estimator depends on the sampling design D_i. The estimator based on design D_5 provides a substantial im-

Table 2. Bias, $\sqrt{MSE}$, (CLP%), and (CVR%) under supervised learning, based on the designs of Table 1, against those of SRS data of the same size, when π is the only unknown parameter of model (6.1).

		OS					SRS				
	π	0.35	0.50	0.60	0.67	0.80	0.35	0.50	0.60	0.67	0.80
D_1	Bias	−0.09	−0.15	−0.19	−0.24	−0.30	0.02	−0.00	−0.01	−0.03	−0.07
	$\sqrt{MSE}$	0.15	0.24	0.28	0.35	0.44	0.18	0.19	0.18	0.17	0.17
	CLP%	87.7	86.6	87.1	84.7	87.7	85.2	84.0	84.6	85.5	87.7
	CVR%	31.1	9.5	4.6	2.8	1.2	92.1	97.0	94.8	91.1	72.6
D_2	Bias	0.17	0.11	0.07	0.04	0.00	0.01	0.00	−0.00	−0.02	−0.04
	$\sqrt{MSE}$	0.26	0.18	0.13	0.10	0.08	0.16	0.17	0.17	0.16	0.13
	CLP%	87.4	87.4	87.5	88.4	90.8	86.0	85.2	85.6	86.2	88.5
	CVR%	7.3	21.1	40.8	61.8	93.8	97.0	99.3	98.2	96.1	83.9
D_3	Bias	0.04	−0.00	−0.03	−0.04	−0.05	0.03	−0.00	−0.02	−0.03	−0.07
	$\sqrt{MSE}$	0.16	0.15	0.16	0.16	0.15	0.18	0.19	0.18	0.18	0.16
	CLP%	87.7	86.9	87.2	88.1	90.4	85.2	84.1	84.6	85.5	87.9
	CVR%	100	100	100	99.9	99.5	92.4	96.7	94.9	90.8	74.0
D_4	Bias	0.02	−0.00	−0.02	−0.03	−0.02	0.01	−0.00	−0.00	−0.01	−0.02
	$\sqrt{MSE}$	0.13	0.13	0.13	0.13	0.11	0.15	0.16	0.15	0.15	0.12
	CLP%	88.2	87.2	87.4	88.2	90.5	86.4	85.6	85.8	86.6	88.9
	CVR%	99.9	100	100	100	99.7	98.7	99.8	99.4	98.3	88.4
D_5	Bias	0.00	−0.00	−0.00	−0.00	−0.00	0.03	0.00	−0.02	−0.04	−0.07
	$\sqrt{MSE}$	0.11	0.11	0.11	0.11	0.09	0.18	0.19	0.18	0.18	0.16
	CLP%	88.0	87.3	87.8	88.4	90.7	85.1	84.1	84.7	85.3	87.8
	CVR%	99.8	100	99.9	99.8	97.7	92.6	97.1	94.5	90.8	73.3

provement over the MLE of the SRS design. For example, the relative efficiencies RE = MSE(SRS)/MSE(D_5) from Table 2 are $(0.18^2/0.11^2 =)2.7$, 2.98, 2.7, 2.7, and 3.16, for $\pi = 0.35, 0.50, 0.60, 0.67$, and 0.8, respectively. These empirical results show that the MLE based on design D_5 is at least 2.7 times more efficient than the corresponding SRS estimator. The same efficiencies under unsupervised learning in Table 3 are $4.76, 4.69, 4.34, 4.76, 4.41$. These RE values indicate that design D_5 is much better suited to unsupervised learning.

7. Data Analysis

The age structure of fish is an important part of many fishery studies, because it provides valuable information about age of recruitment, maturity, and so on. As a result, estimations of the age structure play a key role in stock assessments and in the dynamics of a fish population. In this section, we examine the age

Table 3. Bias, $\sqrt{MSE}$, (CLP%), and (CVR%) under unsupervised learning, based on the designs of Table 1, against those of SRS data of the same size, when π is the only unknown parameter of model (6.1).

		OS					SRS				
	π	0.35	0.50	0.60	0.67	0.80	0.35	0.50	0.60	0.67	0.80
D_1	Bias	0.04	0.04	0.01	−0.01	−0.04	0.00	0.00	−0.00	0.00	−0.01
	$\sqrt{MSE}$	0.17	0.19	0.18	0.18	0.19	0.24	0.26	0.25	0.24	0.21
	CLP%	87.3	85.2	85.0	85.6	87.0	82.6	81.5	82.2	83.1	86.0
	CVR%	96.9	91.0	83.1	76.8	68.7	99.2	99.2	99.2	99.3	98.9
D_2	Bias	−0.02	−0.04	−0.04	−0.02	−0.01	0.00	−0.00	−0.00	−0.00	−0.01
	$\sqrt{MSE}$	0.17	0.18	0.16	0.15	0.11	0.21	0.22	0.22	0.21	0.18
	CLP%	85.6	86.0	87.0	87.8	90.3	84.8	84.3	84.5	84.9	87.3
	CVR%	86.8	95.8	98.3	99.5	99.8	99.5	99.6	99.4	99.5	99.2
D_3	Bias	0.03	−0.00	−0.03	−0.04	−0.04	0.01	−0.01	−0.00	−0.00	−0.01
	$\sqrt{MSE}$	0.16	0.15	0.16	0.16	0.15	0.25	0.26	0.25	0.24	0.21
	CLP%	87.6	86.9	87.2	87.9	90.1	82.9	82.0	82.3	83.5	86.1
	CVR%	99.9	100	100	99.9	99.7	99.2	99.5	99.5	99.2	99.1
D_4	Bias	0.02	0.00	−0.02	−0.02	−0.02	0.00	−0.00	−0.00	−0.00	−0.00
	$\sqrt{MSE}$	0.14	0.13	0.14	0.14	0.12	0.19	0.20	0.20	0.19	0.16
	CLP%	88.0	87.2	87.3	88.1	90.2	86.1	85.3	85.4	86.2	88.2
	CVR%	99.9	100	100	100	99.8	99.7	99.8	99.6	99.4	99.2
D_5	Bias	0.01	−0.00	−0.00	0.00	0.00	0.00	0.00	−0.00	0.00	−0.01
	$\sqrt{MSE}$	0.11	0.12	0.12	0.11	0.10	0.24	0.26	0.25	0.24	0.21
	CLP%	88.0	87.3	87.7	88.2	90.3	82.8	81.7	82.5	83.0	86.0
	CVR%	99.9	100	99.9	100	99.7	99.4	99.4	99.2	99.3	99.1

determination of Spot, as a short-lived fish species, using frequency data on the length of the fish. Owing to its commercial and recreational purposes and food source for other fish, Spot represent one of the most important and frequently caught fish in the Chesapeake Bay area. The existence of several environmental studies on such short-lived fish species (Thomas (1990); Rickabaugh and Capossela (2011)) has increased the importance of analyzing the age structures of Spot.

Recently, several fishery studies have tried different sampling designs based on ranks and order statistics. Among other things, these studies examine the mercury level of fish (Nourmohammadi, Jafari Jozani and Johnson (2015)), the stock abundance of fish (Wang, Ye and Milton (2009)), and RBS designs for age structure determination (Hatefi, Jafari Jozani and Ozturk (2015)).

Here, we employ an ML estimation for the parameters of the FMM in a

fishery study to determine the age structure of Spot. Owing to the cost of determining the age of fish, researchers may first capture and examine a large sample, from which they then draw a subsample for the age determination. Because the length of a fish is correlated to its age, length is often used as a concomitant to select the final sample. In this section, we consider the length and age determined by otoliths of 403 Virginia–Chesapeake Bay Spot as our population of interest. The data set is available online in the FSAdata package (Ogle (2013)). In this study, we focus on two classes of Spot: ages zero and one year, which are sexually immature and usually smaller; and fish that are two years and older, which are sexually mature and usually longer. A statistical analysis of the two groups is important because the second group plays a vital role in the current reproductivity of the current population, and the first group influences the dynamics and reproduction of the future population. Hatefi, Jafari Jozani and Ozturk (2015) showed that the length distribution of Spot is well-modeled by a mixture of two normal distributions with parameters $\mathbf{\Psi} = (\pi, \mu_1, \mu_2, \sigma) = (0.67, 9.01, 11.70, 1.15)$.

We perform a simulation study with 5,000 repetitions by generating samples using two common approaches to selecting a final sample. We generate samples of size $n = 30$, and then select the following ordered elements (rank collections) for each sample for the age determination. The 30 fish in the original sample are modeled according to their length, which is readily obtained. These collections include $D_1^* = \{1, 4, 7, 10, 13, 16, 19, 22, 25, 28\}$, $D_2^* = \{2, 5, 8, 11, 14, 17, 20, 23, 26, 29\}$, $D_3^* = \{3, 6, 9, 12, 15, 18, 21, 24, 27, 30\}$, $D_4^* = \{5, 10, 15, 20, 25\}$, and $D_5^* = \{1, 5, 10, 15, 20, 25, 30\}$. Then, we employ the proposed methods to estimate and classify the observations in order to determine the age structure of Spot. We study the effect of various collections of order statistics in the observed samples using D_i^*, for $i = 1, 2, 3, 4, 5$.

Tables 4 and 6 present the bias and square root of the MSE for the estimates of $\mathbf{\Psi}$ based on D_i^* under supervised and unsupervised learning approaches, respectively. Tables 5 and 7 present the computational aspects of the estimation procedures in the analysis of the Spot data set. The estimate $\hat{\pi}_{MLE}$, whether using either labeled or unlabeled order statistics, almost always outperforms the SRS-based estimate. This is because $\hat{\pi}_{MLE}$ takes full and direct advantage of rank information of the order statistics in these approaches. Note that an estimation of the component parameters of an FMM based on order statistics using the modified EM algorithm can not take full advantage of rank information. However, it does do so indirectly $\hat{\pi}$. Tables 5 and 7 show that the estimation procedures under the supervised and unsupervised approaches both outperform

Table 4. Bias and $\sqrt{MSE}$ of Spot data under the supervised learning approach, based on designs D_i^*, for $i = 1, \ldots, 5$, against those of SRS data of the same size.

		OS				SRS			
		π	μ_1	μ_2	σ	π	μ_1	μ_2	σ
D_1^*	Bias	-0.03	-0.13	-0.15	-0.08	-0.00	-0.01	-0.01	-0.16
	$\sqrt{MSE}$	0.11	0.36	0.56	0.23	0.14	0.45	0.69	0.32
D_2^*	Bias	-0.01	-0.01	0.04	-0.12	-0.01	0.01	-0.01	-0.16
	$\sqrt{MSE}$	0.10	0.32	0.52	0.26	0.14	0.45	0.67	0.32
D_3^*	Bias	0.01	0.09	0.26	-0.11	-0.01	-0.01	-0.00	-0.16
	$\sqrt{MSE}$	0.11	0.34	0.64	0.25	0.14	0.45	0.69	0.32
D_4^*	Bias	-0.06	0.12	-0.50	-0.37	-0.05	0.00	-0.01	-0.33
	$\sqrt{MSE}$	0.13	0.42	0.93	0.57	0.19	0.69	0.89	0.57
D_5^*	Bias	0.00	-0.29	0.52	0.01	-0.02	0.01	-0.01	-0.23
	$\sqrt{MSE}$	0.13	0.55	0.97	0.21	0.16	0.56	0.79	0.43

Table 5. Computational aspects of the estimators for the Spot data under supervised learning, based on designs D_i^*, for $i = 1, \ldots, 5$, against those of SRS data of the same size.

	OS				SRS			
	interation	CLP%	time	Conv.	iteration	CLP%	time	Conv.
D_1^*	4.26	86.40	0.0049	98.86	1.00	86.60	0.0004	98.06
D_2^*	3.00	86.83	0.0036	99.84	1.00	86.51	0.0004	98.10
D_3^*	3.69	86.89	0.0042	99.98	1.00	86.54	0.0004	97.96
D_4^*	4.60	85.37	0.0035	87.80	1.00	84.84	0.0003	85.72
D_5^*	3.58	85.86	0.0030	99.92	1.00	85.66	0.0003	93.52

their SRS counterparts in terms of classification precision and convergence rate.

8. Conclusion

We propose estimation and classification methods based on order statistics of FMMs. This study differs in terms of focus and structure from two recent works on order statistics in FMMs, namely, Hatefi, Jafari Jozani and Ziou (2014); Hatefi, Jafari Jozani and Ozturk (2015). The main objective of this study is to develop a statistical inference for classifying labeled and/or unlabeled current or future observations, based on correlated order statistics. In contrast, Hatefi, Jafari Jozani and Ziou (2014); Hatefi, Jafari Jozani and Ozturk (2015) estimate the parameters of an FMM, and classify the observations into subpopulations using

Table 6. Bias, $\sqrt{MSE}$ of the Spot data under unsupervised learning, based on designs D_i^*, for $i = 1, \ldots, 5$, against those of the SRS data of the same size.

		OS				SRS			
		π	μ_1	μ_2	σ	π	μ_1	μ_2	σ
D_1^*	Bias	-0.21	-0.62	-0.64	-0.19	-0.10	-0.35	-0.05	-0.34
	$\sqrt{MSE}$	0.35	0.97	1.12	0.36	0.23	0.77	0.85	0.53
D_2^*	Bias	-0.07	-0.24	-0.07	-0.23	-0.09	-0.33	-0.03	-0.34
	$\sqrt{MSE}$	0.16	0.48	0.54	0.36	0.23	0.77	0.86	0.53
D_3^*	Bias	0.06	0.19	0.56	-0.12	-0.10	-0.35	-0.06	-0.34
	$\sqrt{MSE}$	0.18	0.56	1.01	0.32	0.24	0.78	0.86	0.53
D_4^*	Bias	-0.15	-0.25	-0.72	-0.54	-0.13	-0.42	-0.20	-0.56
	$\sqrt{MSE}$	0.23	0.48	1.13	0.78	0.27	1.00	1.13	0.83
D_5^*	Bias	-0.01	-0.33	0.48	0.02	-0.11	-0.39	-0.12	-0.45
	$\sqrt{MSE}$	0.19	0.68	1.02	0.27	0.25	0.88	1.00	0.67

Table 7. Computational aspects of the estimators of the Spot data under unsupervised learning, based on designs D_i^*, for $i = 1, \ldots, 5$, against those of the SRS data of the same size.

	OS				SRS			
	interation	CLP%	time	Conv.	iteration	CLP%	time	Conv.
D_1^*	18.34	75.12	0.0220	92.30	12.43	82.17	0.0043	99.06
D_2^*	14.78	85.02	0.0186	99.52	12.64	81.91	0.0044	98.90
D_3^*	21.04	83.75	0.0251	94.88	12.46	81.91	0.0043	99.00
D_4^*	9.95	81.85	0.0080	99.88	8.20	79.91	0.0028	98.82
D_5^*	21.68	82.59	0.0189	93.76	10.34	80.83	0.0035	99.30

independent order statistics in ranked-set sampling designs. In this study, the order statistics are correlated, which requires different latent structures, missing data mechanisms, and EM algorithms to those in Hatefi, Jafari Jozani and Ziou (2014); Hatefi, Jafari Jozani and Ozturk (2015).

We used the properties of the correlated order statistics to estimate and classify FMMs using both supervised and unsupervised learning methods. Using the correlation structure of the order statistics, we obtained various model-based classification criteria. These criteria help us to determine the group membership of the data, and enable inferences about rarely observed components. Our framework is general enough to apply to several sampling designs from FMMs, including left censoring, right censoring, double censoring, minimal-maximal nomina-

tion sampling, and systematic sampling. Empirical evidence shows that selecting an appropriate collection of order statistics provides a substantial improvement over the SRS option in both supervised and unsupervised learning. For example, systematic sampling can be two or three times more efficient than its SRS counterpart when estimating the mixing proportion in supervised and unsupervised learning, respectively. The proposed methodologies were employed to determine the age structure of Spot fish using length frequency data. Numerical results illustrate that the procedures under the supervised and unsupervised approaches almost always outperform their SRS counterparts in terms of estimation and classification precision.

Supplementary Material

All proofs, eight remarks, an additional simulation study are provided in the online Supplementary Material.

Acknowledgements

The authors would like to thank the Co-Editor Professor Hans-Georg Muller, the AE and the referees for insightful comments, which improved the quality of the paper.

References

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, Singapore.

David, H. A. and Nagaraja, H. N. (1981). *Order Statistics*. Wiley Online Library, New Jersey.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B. (Statistical Methodology)* **39**, 1–38.

Furman, W. D. and Lindsay, B. G. (1994). Measuring the relative effectiveness of moment estimators as starting values in maximizing likelihoods. *Comput. Statist. Data Anal.* **17**, 493–507.

Hatefi, A., Jafari Jozani, M. and Ozturk, O. (2015). Mixture model analysis of partially rank-ordered set samples: Age groups of fish from length-frequency data. *Scandinavian Journal of Statistics* **42**, 848–871.

Hatefi, A., Jafari Jozani, M. and Ziou, D. (2014). Estimation and classification gor finite mixture models under ranked set sampling. *Statistica Sinica* **24**, 675–698.

Johnson, R. A. and Mehrotra, K. G. (1972). Locally most powerful rank tests for the two-sample problem with censored data. *The Annals of Mathematical Statistics* **43**, 823–831.

McLachlan, G. and Peel, D. (2004). *Finite Mixture Models*. Wiley, New York.

Mehrotra, K. and Nanda, P. (1974). Unbiased estimation of parameters by order statistics in the case of censored samples. *Biometrika* **61**, 601–606.

Mendenhall, W. and Hader, R. (1958). Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. *Biometrika* **45**, 504–520.

Miyata, Y. (2011). Maximum likelihood estimators in finite mixture models with censored data. *Journal of Statistical Planning and Inference* **141**, 56–64.

Nourmohammadi, M., Jafari Jozani, M. and Johnson, B. C. (2015). Distribution-free tolerance intervals with nomination samples: Applications to mercury contamination in fish. *Statistical Methodology* **26**, 16–33.

Ogle, D. (2013). Fisheries stock assessment (fsa) methods package for r. R package version 0.4.13.

Rickabaugh, H. and Capossela, K. (2011). Evaluation of the status of spot in maryland 2010. *Maryland DNR Fisheries Services Doc 6-23-2011*.

Saleem, M., Aslam, M. and Economou, P. (2010). On the bayesian analysis of the mixture of power function distribution using the complete and the censored sample. *Journal of Applied Statistics* **37**, 25–40.

Sindhu, T., Feroze, N. and Aslam, M.(2016). Doubly censored data from two-component mixture of inverse weibull distributions: Theory and applications. *Journal of Modern Applied Statistical Methods* **15**, 1–21.

Thomas, P.(1990). Teleost model for studying the effects of chemicals on female reproductive endocrine function. *The journal of experimental zoology* **256**, 126–128.

Wang, Y.-G., Ye, Y. and Milton, D. A.(2009). Efficient designs for sampling and subsampling in fisheries research based on ranked sets. *ICES Journal of Marine Science: Journal du Conseil* **66**, 928–934.

Yang, S.-S.(1977). General distribution theory of the concomitants of order statistics. *The Annals of Statistics* **5**, 996–1002.

Department of Mathematics and Statistics, Memorial University, St. John's, NL, Canada.

E-mail: ahatefi@mun.ca, hatefi.ar@gmail.com

Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada.

E-mail: reid@utstat.utoronto.ca

Department of Statistics, University of Manitoba, Winnipeg, MB, Canada.

E-mail: m_jafari_jozani@umanitoba.ca

Department of Statistics, The Ohio State University, Columbus, OH 43210, USA.

E-mail: omer@stat.osu.edu

(Received May 2017; accepted October 2018)