

**DEPARTMENT OF STATISTICS**  
The Pennsylvania State University  
University Park, PA 16802 U.S.A.

**TECHNICAL REPORTS AND PREPRINTS**

**Number 03-04: April 2003**

**Identifiability and Estimation in Finite Mixture Models  
with Multinomial Components**

**Ryan Elmore and Shaoli Wang**

# Identifiability and Estimation in Finite Mixture Models with Multinomial Components

RYAN ELMORE and SHAOLI WANG<sup>†</sup>

*Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA*

---

## Abstract

We will present several results related to the identifiability of multinomial finite mixture models. In particular, a necessary and sufficient condition is given for this model to be identifiable. In the special case of binomial mixtures, a “strong” necessary condition for identifiability is proved. Estimation procedures are discussed at the end of the paper.

*Keywords:* multinomial, finite mixture, identifiable, expectation-maximization

---

## 1 Introduction

In this paper, we will discuss issues related to the identifiability and estimation in finite mixture distributions having multinomial component densities. The ideas presented in this paper emanate from a series of presentations given to Professor Thomas P. Hettmansperger’s nonparametric mixture model seminar by Ryan Elmore and Shaoli Wang during the Fall semester of 2002. Our intentions are to codify those notes and subsequent thoughts into an easily readable and understandable manuscript.

The identifiability of mixture models has received a fair amount of attention over the years, see *e.g.* Teicher (1961, 1963), Yakowitz and Spragins (1968), however these articles

<sup>†</sup>Authors are listed in alphabetical order.

tend to be very general in nature regarding the classes of distributions considered. Lindsay (1995) and Titterington et al. (1985) offer detailed accounts of mixture models, both of which discuss issues related to the identifiability of these models. McLachlan and Peel (2000) is an up-to-date account of finite mixture models which focuses on the modeling issues.

The paper is organized as follows. In sections 2, we will introduce the concept of identifiability in finite mixture models. We will present identifiability results for binomial mixture models and multinomial mixture models in sections 3 and 4, respectively. Parameter estimation is discussed in section 5 along with a detailed example. A summary of these notes and some caveats are given in section 6. Finally, a few supplemental results are presented in the appendix.

## 2 Identifiability

We begin with some definitions. We first note that identifiability is defined in terms of a class of distributions. A parametric class of distribution functions is given by

$$\mathcal{H} = \{H(\mathbf{y}; \boldsymbol{\theta}) : \mathbf{y} \in \mathbb{R}^m, \boldsymbol{\theta} \in \Theta\},$$

where  $\boldsymbol{\theta}$  is the parameter of the distribution  $H$  and  $\Theta$  is the parameter space. The class of  $K$ -component finite mixture distributions on the class  $\mathcal{H}$  is then

$$\mathcal{F}_K = \left\{ \sum_{k=1}^K \pi_k H(\mathbf{y}; \boldsymbol{\theta}_k) : \pi_k \geq 0, \boldsymbol{\theta}_k \in \Theta, 1 \leq k \leq K, \sum_{k=1}^K \pi_k = 1 \right\}.$$

In other words,  $\mathcal{F}$  is the collection of all mixtures of the form  $\int H(\mathbf{y}; \boldsymbol{\theta}) dG(\boldsymbol{\theta})$  where  $G(\boldsymbol{\theta})$  is a discrete distribution on  $\Theta$  having at most  $K$  points of support.

**Definition 2.1 (Identifiability).** *The class of  $K$ -component finite mixture distributions  $\mathcal{F}_K$  is said to be identifiable if each member in  $\mathcal{F}_K$  has exactly one representation. More precisely, if*

$$\begin{aligned} \sum_{k=1}^{K_1} \pi_{1k} H(\mathbf{y}; \boldsymbol{\theta}_{1k}) &= \sum_{k=1}^{K_2} \pi_{2k} H(\mathbf{y}; \boldsymbol{\theta}_{2k}), \\ \boldsymbol{\theta}_{1k} \in \Theta, \boldsymbol{\theta}_{2k} \in \Theta, \pi_{1k} > 0, \pi_{2k} > 0, \sum_{k=1}^{K_1} \pi_{1k} &= 1, \sum_{k=1}^{K_2} \pi_{2k} = 1, \\ \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{K_1} \text{ are distinct, } K_2 \leq K_1 \leq K, \end{aligned} \quad (1)$$

then  $K_1 = K_2$  and an appropriate arrangement of the parameters is obtained so that  $\pi_{1k} = \pi_{2k}$  and  $\boldsymbol{\theta}_{1k} = \boldsymbol{\theta}_{2k}$ , for  $k = 1, 2, \dots, K_1$ , then  $\mathcal{F}_K$  is identifiable.

From this definition it follows that a class of  $K$ -component finite mixture distributions  $\mathcal{F}_K$  is not identifiable if at least one member  $F \in \mathcal{F}_K$  has an additional representation in the class. An even stronger non-identifiability is defined as follows.

**Definition 2.2 (Strong Non-identifiability).** *We say a class of  $K$ -component finite mixture distributions  $\mathcal{F}_K$  is not identifiable in a strong sense if every member in  $\mathcal{F}_K$  has at least two different representations as in (1).*

### 3 Binomial mixture models

Let  $Y$  have a binomial distribution with  $m$  trials and probability of success  $\theta$ . The probability mass function of  $Y$  is

$$f(y; \theta) = \binom{m}{y} \theta^y (1 - \theta)^{m-y}, \quad y = 0, 1, \dots, m.$$

Denote  $\mathbf{f}(\theta) = (f(0, \theta), f(1, \theta), \dots, f(m, \theta))^T$ , where  $T$  stands for transposition. Define

$$U = \begin{pmatrix} 1 & -m & \binom{m}{2} & \cdots & (-1)^m \\ \binom{m}{1} & -m(m-1) & \cdots & m(-1)^{m-1} & \\ & \binom{m}{2} & \cdots & \binom{m}{2}(-1)^m & \\ & & \ddots & \vdots & \\ 0 & & & & 1 \end{pmatrix}$$

so that  $U$  is a nonsingular, upper-triangular matrix such that

$$\mathbf{f}(\theta) = U \begin{pmatrix} 1 \\ \theta \\ \theta^2 \\ \vdots \\ \theta^m \end{pmatrix} \quad (2)$$

We define the class of  $K$ -component binomial mixture models as

$$\mathcal{B}_{m,K} = \left\{ \sum_{k=1}^K \pi_k \mathbf{f}(\theta_k) : \pi_k \geq 0, 0 \leq \theta_k \leq 1, 1 \leq k \leq K, \sum_{k=1}^K \pi_k = 1 \right\}$$

Many authors have contributed results on the identifiability of  $\mathcal{B}_{m,K}$ , see *e.g.* Teicher (1961), Blischke (1964), Titterington et al. (1985), and Lindsay (1995). These results can be summarized in the following lemma.

**Lemma 3.1.** *The class of  $K$ -component binomial mixture models  $\mathcal{B}_{m,K}$  is identifiable if and only if  $m \geq 2K - 1$ .*

Lemma 3.1 says that the condition  $2K - 1 \leq m$  is necessary and sufficient for the identifiability of  $K$ -component binomial models. Note that the necessity of this lemma states that if  $m < 2K - 1$ , then at least one member of  $\mathcal{B}_{m,K}$  has an additional representation in this class. In fact, we will show that if this condition fails, then every  $K$ -component

binomial mixture model has infinitely many distinct representations provided at most one component parameter is 0 or 1. Before presenting our non-identifiability results, we need the following lemma.

**Lemma 3.2.** *If  $0 \leq \theta_1 < \theta_2 < \dots < \theta_{2K} \leq 1$ , then the linear system of equations*

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ \theta_1 & \theta_2 & \cdots & \theta_{2K} \\ \theta_1^2 & \theta_2^2 & \cdots & \theta_{2K}^2 \\ \vdots & \vdots & & \vdots \\ \theta_1^{2K-2} & \theta_2^{2K-2} & \cdots & \theta_{2K}^{2K-2} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{2K} \end{pmatrix} = \mathbf{0} \quad (3)$$

has a unique solution  $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_{2K})^T \neq \mathbf{0}$  up to a constant multiplier. Furthermore,  $\alpha_i \neq 0$  and if  $\alpha_1 < 0$ , then the  $\alpha_i$ 's have alternating signs,

$$\alpha_{2j-1} < 0, \quad \alpha_{2j} > 0, \quad j = 1, 2, \dots, K.$$

*Proof.* Consider a matrix

$$W = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \theta_1 & \theta_2 & \cdots & \theta_{2K} \\ \theta_1^2 & \theta_2^2 & \cdots & \theta_{2K}^2 \\ \vdots & \vdots & & \vdots \\ \theta_1^{2K-2} & \theta_2^{2K-2} & \cdots & \theta_{2K}^{2K-2} \\ \theta_1^{2K-1} & \theta_2^{2K-1} & \cdots & \theta_{2K}^{2K-1} \end{pmatrix}.$$

$W$  is a  $2K$  by  $2K$  Vandermonde matrix and hence is invertible. Note that  $WW^{-1} = I_{2K}$ , where  $I_{2K}$  is the identity matrix, and hence the last column of  $W^{-1}$  is a nonzero solution to (3). Because the left most matrix in (3) has rank  $2K - 1$ , this is the only solution to (3)

up to a constant multiplier. The last column  $\alpha$  of  $W^{-1}$  has elements

$$\begin{aligned}
\alpha_i &= \frac{(-1)^i}{\det(W)} \det \begin{pmatrix} 1 & \cdots & 1 & 1 & \cdots & 1 \\ \theta_1 & \cdots & \theta_{i-1} & \theta_{i+1} & \cdots & \theta_{2K} \\ \theta_1^2 & \cdots & \theta_{i-1}^2 & \theta_{i+1}^2 & \cdots & \theta_{2K}^2 \\ \vdots & & \vdots & \vdots & & \vdots \\ \theta_1^{2K-2} & \cdots & \theta_{i-1}^{2K-2} & \theta_{i+1}^{2K-2} & \cdots & \theta_{2K}^{2K-2} \end{pmatrix} \\
&= (-1)^i \frac{\prod_{1 \leq l < j \leq 2K, l, j \neq i} (\theta_j - \theta_l)}{\prod_{1 \leq l < j \leq 2K} (\theta_j - \theta_l)} \\
&= \frac{-1}{\prod_{j \neq i} (\theta_j - \theta_i)}.
\end{aligned}$$

■

Now suppose that we are given a  $K$ -component binomial mixture model

$$\begin{aligned}
&\sum_{i=1}^K \pi_{2i-1} \mathbf{f}(\theta_{2i-1}), \quad \pi_1, \dots, \pi_K > 0, \quad \sum_{i=1}^K \pi_i = 1, \\
&0 < \theta_1 < \theta_3 < \cdots < \theta_{2K-1} < 1
\end{aligned}$$

We are looking for another representation of this model,

$$\sum_{i=1}^K \pi_{2i} \mathbf{f}(\theta_{2i}) = \sum_{i=1}^K \pi_{2i-1} \mathbf{f}(\theta_{2i-1}). \tag{4}$$

According to equation (2), equation (4) requires that

$$\alpha = (\pi_1, -\pi_2, \dots, \pi_{2K-1}, -\pi_{2K})^T$$

is a solution to the linear system (3). By Lemma 3.2, we need to find  $\theta_2, \theta_4, \dots, \theta_{2K}$  satisfying

$$0 < \theta_1 < \theta_2 < \theta_3 < \theta_4 < \cdots < \theta_{2K-1} < \theta_{2K} < 1$$

or

$$0 < \theta_2 < \theta_1 < \theta_4 < \theta_3 < \cdots < \theta_{2K} < \theta_{2K-1} < 1,$$

such that

$$\begin{aligned}\pi_{2i-1} &= c \left( \prod_{j=1, j \neq 2i-1}^{2K} (\theta_j - \theta_{2i-1}) \right)^{-1} \\ &= c \left( \prod_{j=1}^K (\theta_{2j} - \theta_{2i-1}) \prod_{j=1, j \neq i}^K (\theta_{2j-1} - \theta_{2i-1}) \right)^{-1}\end{aligned}\quad (5)$$

$$i = 1, 2, \dots, K,$$

where  $c \neq 0$  is a constant. Equation (5) is equivalent to

$$\frac{\pi_{2K-1}}{\pi_{2i-1}} = (-1) \prod_{j=1}^K \frac{\theta_{2j} - \theta_{2i-1}}{\theta_{2j} - \theta_{2K-1}} \prod_{j=1, j \neq i}^{K-1} \frac{\theta_{2j-1} - \theta_{2i-1}}{\theta_{2j-1} - \theta_{2K-1}}$$

$$i = 1, 2, \dots, K-1.$$

or

$$\begin{aligned}& (-1)^{K-i-1} \frac{\pi_{2K-1}}{\pi_{2i-1}} \prod_{j=1, j \neq i}^{K-1} \frac{\theta_{2j-1} - \theta_{2K-1}}{\theta_{2j-1} - \theta_{2i-1}} \\ &= (-1)^{K-i} \prod_{j=1}^K \frac{\theta_{2j} - \theta_{2i-1}}{\theta_{2j} - \theta_{2K-1}}\end{aligned}\quad (6)$$

$$i = 1, 2, \dots, K-1.$$

Define

$$\gamma_i = (-1)^{K-i-1} \frac{\pi_{2K-1}}{\pi_{2i-1}} \prod_{j=1, j \neq i}^{K-1} \frac{\theta_{2j-1} - \theta_{2K-1}}{\theta_{2j-1} - \theta_{2i-1}}\quad (7)$$

$$i = 1, 2, \dots, K-1.$$

These are arbitrary nonzero constants with the same sign. Now equation (6) becomes

$$\gamma_i = (-1)^{K-i} \left( \prod_{j=1, j \neq i}^{K-1} \frac{\theta_{2j} - \theta_{2i-1}}{\theta_{2j} - \theta_{2K-1}} \right) \frac{\theta_{2K} - \theta_{2i-1}}{\theta_{2i} - \theta_{2K-1}} \frac{\theta_{2i} - \theta_{2i-1}}{\theta_{2K} - \theta_{2K-1}}$$

$$i = 1, 2, \dots, K-1,$$

or

$$\theta_{2i} - \theta_{2i-1} = \gamma_i (-1)^{K-i} \left( \prod_{j=1, j \neq i}^{K-1} \frac{\theta_{2j} - \theta_{2K-1}}{\theta_{2j} - \theta_{2i-1}} \right) \frac{\theta_{2i} - \theta_{2K-1}}{\theta_{2K} - \theta_{2i-1}} (\theta_{2K} - \theta_{2K-1})\quad (8)$$

$$i = 1, 2, \dots, K-1.$$

Define

$$x_i = \theta_{2i} - \theta_{2i-1}, \quad i = 1, 2, \dots, K,$$

and consider

$$\theta_{2i-3} - \theta_{2i-1} < x_i < \theta_{2i+1} - \theta_{2i-1}, \quad i = 1, 2, \dots, K,$$

where  $\theta_{-1} = 0$  and  $\theta_{2K+1} = 1$ . Then equation (8) becomes

$$x_i = x_K g_i(x_1, \dots, x_K), \quad i = 1, \dots, K-1, \quad (9)$$

where  $g_i$  are continuously differentiable functions ( $C^1$  functions).

Define

$$L = \min_{0 \leq i \leq K} \frac{\theta_{2i+1} - \theta_{2i-1}}{2}.$$

Then on region

$$[-L, L]^K = [-L, L] \times [-L, L] \times \dots \times [-L, L]$$

$g_i$  and  $\partial g_i / \partial x_j$  are continuous and bounded.

Let  $M > 0$  be an upper bound of these functions, i.e.,

$$|g_i| \leq M, \quad |\partial g_i / \partial x_j| \leq M, \quad i = 1, \dots, K-1, \quad j = 1, \dots, K.$$

Let

$$h_i(x_1, \dots, x_K) = x_i - x_K g_i(x_1, \dots, x_K), \quad i = 1, \dots, K-1$$

$$\mathbf{g} = (g_1, \dots, g_{K-1})^T, \quad \mathbf{h} = (h_1, \dots, h_{K-1})^T.$$

Then equation (9) becomes

$$\mathbf{h}(x_1, \dots, x_{K-1}, x_K) = \mathbf{0}. \quad (10)$$

Let  $D$  denote the differential with respect to  $(x_1, \dots, x_{K-1})^T$ . Then

$$D\mathbf{h} = I - x_K D\mathbf{g},$$

where  $I$  is  $(K - 1) \times (K - 1)$  identity matrix. In particular,

$$D\mathbf{h}(0, \dots, 0) = I.$$

Since  $\mathbf{h}$  is continuously differentiable, it follows from the Implicit Function Theorem that in a neighborhood of  $(x_1, \dots, x_K)^T = \mathbf{0}$ , equation (10) has a unique solution

$$x_i = x_i(x_K), \quad i = 1, \dots, K - 1, \quad |x_K| \text{ is small.} \quad (11)$$

This solution is  $C^1$ . Moreover, if  $x_K \neq 0$ , then

$$x_i(x_K)x_K > 0, \quad i = 1, \dots, K - 1.$$

In other words, they have the same sign. Note that if we have either  $\theta_1 = 0$  or  $\theta_{2K-1} = 1$ , but not both, then we still have the second representation for the given binomial mixture model. Therefore, we have shown the following theorem.

**Theorem 3.1.** *If  $2K - 1 > m$ , then each  $K$ -component binomial mixture model*

$$\sum_{i=1}^K \pi_i \mathbf{f}(\theta_i), \quad \pi_1, \dots, \pi_K > 0, \quad \sum_{i=1}^K \pi_i = 1, \quad 0 \leq \theta_1 < \theta_2 < \dots < \theta_K \leq 1$$

*has infinitely many distinct representations, provided  $\theta_0 \neq 0$  or  $\theta_K \neq 1$ . In particular, the class  $\mathcal{B}_{m,K}$  is not identifiable.*

## 4 Multinomial mixture models

Let  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_r, Y_{r+1})^T$  have a multinomial distribution with  $m$  trials and probability  $\theta_j$  of being in the  $j^{\text{th}}$  category,  $j = 1, 2, \dots, r + 1$ . The probability mass function of  $\mathbf{Y}$  can be written as

$$f(\mathbf{y}; \boldsymbol{\theta}, m) = \binom{m}{y_1 y_2 \dots y_r} \theta_1^{y_1} \theta_2^{y_2} \dots \theta_{r+1}^{y_{r+1}}, \quad \mathbf{y} \in S, \quad \boldsymbol{\theta} \in \Theta, \quad (12)$$

where  $y_{r+1} = m - \sum_{i=1}^r y_i$ ,  $\theta_{r+1} = 1 - \sum_{i=1}^r \theta_i$ , and

$$S = \left\{ (y_1, \dots, y_r)^T \in \mathbb{R}^r : y_i \text{ are integers, } y_i \geq 0, \sum_{i=1}^r y_i \leq m \right\},$$

$$\Theta = \left\{ (\theta_1, \dots, \theta_r)^T \in \mathbb{R}^r : \theta_i \geq 0, \sum_{i=1}^r \theta_i \leq 1 \right\}.$$

Notice that there are  $\binom{m+r}{r}$  distinct vectors in  $S$ . We can stack the  $\binom{m+r}{r}$  corresponding values of the probability mass function into a single  $\binom{m+r}{r}$  by 1 vector as

$$\mathbf{f}(\boldsymbol{\theta}, m) = \begin{pmatrix} \theta_{r+1}^m \\ m\theta_1\theta_{r+1}^{m-1} \\ \vdots \\ m\theta_r\theta_{r+1}^{m-1} \\ \vdots \\ \binom{m}{2}\theta_1^2\theta_{r+1}^{m-2} \\ \vdots \\ \binom{m}{2}\theta_r^2\theta_{r+1}^{m-2} \\ \vdots \\ \theta_1^m \\ \vdots \\ \theta_r^m \end{pmatrix}, \quad (13)$$

where  $\theta_{r+1} = 1 - \sum_{i=1}^r \theta_i$ .

Denote

$$\boldsymbol{\delta} = (1, \theta_1, \dots, \theta_r, \theta_1^2, \dots, \theta_r^2, \dots, \theta_1^m, \dots, \theta_r^m)^T \quad (14)$$

and

$$A = \begin{pmatrix} 1 & -m & \cdots & -m & \cdots & \binom{m}{2} & \cdots & \binom{m}{2} & \cdots & (-1)^m & \cdots & (-1)^m \\ & m & & & & & & & & & & \\ & & \ddots & & & & & & & & & \\ & & & m & & & & & & & & \\ & & & & \ddots & & & & & & & \\ & & & & & \binom{m}{2} & & & & & & \\ & & & & & & \ddots & & & & & \\ & & & & & & & \binom{m}{2} & & & & \\ & & & & & & & & \ddots & & & \\ & & & & & & & & & 1 & & \\ & & & & & & & & & & \ddots & \\ & & & & & & & & & & & 1 \\ 0 & & & & & & & & & & & \end{pmatrix}. \quad (15)$$

Obviously,  $A$  is a nonsingular, upper-triangular matrix, independent of the parameter  $\theta$ . It can be shown that (13) can be rewritten in another form, as stated in the following lemma.

**Lemma 4.1.**

$$\mathbf{f}(\theta, m) = A\delta.$$

The following definition is not necessary, however it will make the formulation of multinomial mixture models easier.

**Definition 4.1.** Let  $\theta, \eta \in \mathbb{R}^r$ . We say that  $\theta \prec \eta$  if  $\theta_i = \eta_i$  for  $i = 1, \dots, k < r$ , and  $\theta_{k+1} < \eta_{k+1}$ .

Therefore, we have that if  $\theta \neq \eta$ , then either  $\theta \prec \eta$  or  $\eta \prec \theta$ .

We define the class of  $K$ -component multinomial mixture models as

$$\mathcal{M}_{m,K} = \left\{ \sum_{k=1}^K \pi_k \mathbf{f}(\theta_k, m) : \pi_k \geq 0, \theta_k \in \Theta, 1 \leq k \leq K, \sum_{k=1}^K \pi_k = 1 \right\}.$$

Note that we will order the parameters in a multinomial mixture model so that  $\theta_1 \prec \theta_2 \prec \dots \prec \theta_K$ .

**Lemma 4.2.** *Let  $\theta_1 \prec \theta_2 \prec \dots \prec \theta_K$  be in  $\Theta$ . If  $K \leq m + 1$ , then  $f(\theta_1, m), \dots, f(\theta_K, m)$  are linearly independent.*

*Proof.* Consider

$$H = \{\boldsymbol{\eta} \in \mathbb{R}^r : \boldsymbol{\eta} = \boldsymbol{\theta}_j - \boldsymbol{\theta}_i, 1 \leq i < j \leq K\}.$$

Since  $\boldsymbol{\theta}_i$  are distinct,  $\mathbf{0} \notin H$ . Moreover, there are at most  $\binom{K}{2}$  distinct vectors in  $H$ . Let  $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots, \boldsymbol{\eta}_n$  be the distinct vectors in  $H$ , i.e.,

$$H = \{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots, \boldsymbol{\eta}_n\}.$$

For  $\boldsymbol{\eta} \in \mathbb{R}^r$ ,  $\boldsymbol{\eta} \neq \mathbf{0}$ , define

$$\boldsymbol{\eta}^\perp = \{\boldsymbol{x} \in \mathbb{R}^r : \boldsymbol{x}^T \boldsymbol{\eta} = 0\}.$$

Since  $\boldsymbol{\eta} \neq \mathbf{0}$ ,  $\boldsymbol{\eta}^\perp$  is a  $(r - 1)$ -dimensional linear subspace of  $\mathbb{R}^r$  and hence has a Lebesgue measure zero. Therefore,  $\cup_{i=1}^n \boldsymbol{\eta}_i^\perp$  also has a Lebesgue measure zero. Denote

$$\mathbb{R}_{++}^r = \{\boldsymbol{x} \in \mathbb{R}^r : \boldsymbol{x} = (x_1, \dots, x_r)^T, x_i > 0, i = 1, \dots, r\}.$$

Then the Lebesgue measure of  $\mathbb{R}_{++}^r - \cup_{i=1}^n \boldsymbol{\eta}_i^\perp$  is  $+\infty$ . In particular,  $\mathbb{R}_{++}^r - \cup_{i=1}^n \boldsymbol{\eta}_i^\perp$  is nonempty. Let  $\boldsymbol{v} = (v_1, \dots, v_r)^T$  be a vector in  $\mathbb{R}_{++}^r - \cup_{i=1}^n \boldsymbol{\eta}_i^\perp$ . Then

$$\boldsymbol{v}^T \boldsymbol{\eta} \neq 0, \quad \forall \boldsymbol{\eta} \in H.$$

It follows that  $\boldsymbol{v}^T \boldsymbol{\theta}_i$  are distinct for  $1 \leq i \leq K$ . Define



**Theorem 4.1.** *The distribution of  $Y_j$ , a marginal random variable of  $\mathbf{Y}$  is given by*

$$g(y_j) = \sum_{i=1}^{K'} \lambda_i b(y_j; m, \theta_{ij}) \quad (16)$$

where  $\sum_i \lambda_i = 1$ ,  $K' \leq K$ , and  $b(y_j; m, \theta_{ij})$  is the binomial mass function with  $m$  trials and probability of success  $\theta_{ij}$ .

*Proof.* Let  $F$  be the distribution function of  $\mathbf{Y}$ . Integrating over all margins except the  $j^{\text{th}}$  with respect to  $F$  is equivalent to summing over all margins except the  $j^{\text{th}}$  of  $f$  due to the discreteness. Therefore, we have

$$\begin{aligned} g(y_j) &= \int dF(\mathbf{y}_{[-j]}) \\ &= \sum \cdots \sum \sum_{i=1}^K \pi_i f(\mathbf{y}; m, \boldsymbol{\theta}_i) \end{aligned} \quad (17)$$

$$\begin{aligned} &= \sum_{i=1}^K \pi_i \sum \cdots \sum f(\mathbf{y}; m, \boldsymbol{\theta}_i) \quad (18) \\ &= \sum_{i=1}^{K'} \lambda_i b(y_j; m, \theta_{ij}). \end{aligned}$$

Note that if  $K = K'$ , the  $\lambda_i = \theta_i$  for all  $i$ . Note that the unmarked summations given in (17) and (18) are over all margins of  $f(\mathbf{y})$  except the  $j^{\text{th}}$ . ■

**Theorem 4.2 (Multinomial Identifiability).** *A necessary and sufficient condition for  $\mathcal{M}_{m,K}$  to be identifiable is  $m \geq 2K - 1$ .*

*Proof.* (Note: The authors found a proof of this theorem in Kim (1984) after the preparation of this manuscript, however the proof given below is quite different and will be presented for completeness.)

**Sufficiency:** By way of contradiction, suppose that the class is not identifiable, i.e.,

there is a multinomial mixture model

$$\pi_1 \mathbf{f}(\boldsymbol{\theta}_1) + \dots + \pi_K \mathbf{f}(\boldsymbol{\theta}_K)$$

which has at least two representations,

$$\lambda_1 \mathbf{f}(\boldsymbol{\theta}'_1) + \dots + \lambda_K \mathbf{f}(\boldsymbol{\theta}'_K) = \pi_1 \mathbf{f}(\boldsymbol{\theta}_1) + \dots + \pi_K \mathbf{f}(\boldsymbol{\theta}_K).$$

Without loss of generality, we assume that all the component weights  $\pi_i$  are positive. Then

$$[\pi_1 \mathbf{f}(\boldsymbol{\theta}_1) + \dots + \pi_K \mathbf{f}(\boldsymbol{\theta}_K)] - [\lambda_1 \mathbf{f}(\boldsymbol{\theta}'_1) + \dots + \lambda_K \mathbf{f}(\boldsymbol{\theta}'_K)] = \mathbf{0}.$$

Simplifying this equation leads to

$$\alpha_1 \mathbf{f}(\boldsymbol{\eta}_1) + \alpha_2 \mathbf{f}(\boldsymbol{\eta}_2) + \dots + \alpha_l \mathbf{f}(\boldsymbol{\eta}_l) = \mathbf{0}$$

so that  $\alpha_i \neq 0$  and  $\boldsymbol{\eta}_i$  are all distinct,  $\boldsymbol{\eta}_1 \prec \boldsymbol{\eta}_2 \prec \dots \prec \boldsymbol{\eta}_l$ ,  $\boldsymbol{\eta}_i$  equals some  $\boldsymbol{\theta}_j$  or  $\boldsymbol{\theta}'_j$ . It is ready to see that  $0 \leq l \leq 2K \leq m + 1$ .

Now we claim that  $l = 0$ . Otherwise,  $\mathbf{f}(\boldsymbol{\eta}_1), \dots, \mathbf{f}(\boldsymbol{\eta}_l)$  are linearly dependent and  $l \leq m + 1$ , which contradicts to Lemma 4.2. Therefore,  $l = 0$  and

$$\lambda_i = \pi_i, \quad \boldsymbol{\theta}'_i = \boldsymbol{\theta}_i, \quad 1 \leq i \leq K.$$

**Necessity:** To show the condition  $2K - 1 \leq m$  is also necessary, we only need to prove that some member of  $\mathcal{M}_{m,K}$  is not identifiable when  $2K - 1 > m$ . By the arguments used to show Theorem 3.1, the following two classes of  $K$ -component multinomial mixture models are not identifiable.

(i)

$$\begin{aligned} & \pi_1 \mathbf{f}(\boldsymbol{\theta}_1) + \dots + \pi_K \mathbf{f}(\boldsymbol{\theta}_K), \quad \pi_i > 0, \quad i = 1, \dots, K, \quad \sum_{i=1}^K \pi_i = 1, \\ & \boldsymbol{\theta}_i = (\theta_i, \dots, \theta_i)^T, \quad i = 1, \dots, K, \quad 0 < \theta_1 < \theta_2 < \dots < \theta_K < 1/r. \end{aligned}$$

(ii)

$$\begin{aligned} \pi_1 \mathbf{f}(\boldsymbol{\theta}_1) + \cdots + \pi_K \mathbf{f}(\boldsymbol{\theta}_K), \quad \pi_i > 0, \quad i = 1, \dots, K, \quad \sum_{i=1}^K \pi_i = 1, \\ \boldsymbol{\theta}_i = (\delta_1, \delta_2, \dots, \delta_{r-1}, \theta_i)^T, \quad i = 1, \dots, K, \quad \delta_j > 0, \quad j = 1, \dots, r-1, \\ 0 < \theta_1 < \theta_2 < \cdots < \theta_K < 1, \quad \delta_1 + \cdots + \delta_{r-1} + \theta_K < 1. \end{aligned}$$

■

## 5 Estimation

In order to estimate the parameters in a multinomial mixture model, it is natural to cast the problem as a *missing data* problem. In doing this, the method of maximum likelihood can be used via an implementation of the expectation-maximization (EM) algorithm (Dempster et al., 1977). Let us consider the finite mixture model of multinomial distributions defined by

$$f(\mathbf{y}; m) = \sum_{g=1}^G \pi_g f(\mathbf{y}; \boldsymbol{\theta}_g, m) \tag{19}$$

with  $0 < \pi_g < 1$ ,  $\sum \pi_g = 1$ , and  $m \geq 2G - 1$ . Note that  $f(\mathbf{y}; \boldsymbol{\theta}_g, m)$  is of the form given in equation (12).

### 5.1 Complete-Data Formulation

Let us assume that a sample  $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$  is drawn from distribution (19). In this situation, the missing data are defined as the multinomial indicator vectors of component membership,  $\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_n$ . If the observation  $\mathbf{Y}_i$  is actually from the  $g^{\text{th}}$  component, then the vector,  $\mathbf{Z}_i$ , is a unit length vector in the direction of the  $g^{\text{th}}$  dimension for any  $g$ . For example, if the first observation actually comes from the third component, then

$\mathbf{Z}_1 = (0, 0, 1, 0, \dots, 0)^T$ . The complete-data distribution (*i.e.* joint distribution of  $\mathbf{Y}_i$  and  $\mathbf{Z}_i$ ) can be written as

$$f_c(\mathbf{y}_i, \mathbf{z}_i) = \prod_{g=1}^G [\pi_g f(\mathbf{y}_i; \boldsymbol{\theta}_g, m)]^{z_{ig}} \quad (20)$$

with complete-data log likelihood

$$\begin{aligned} l_c(\boldsymbol{\Psi}) &= \log \prod_{i=1}^n f_c(\mathbf{y}_i, \mathbf{z}_i) \\ &= \sum_{i=1}^n \sum_{g=1}^G z_{ig} [\log \lambda_g + \log f(\mathbf{y}_i; \boldsymbol{\theta}_g, m)], \end{aligned} \quad (21)$$

where  $\boldsymbol{\Psi} = (\boldsymbol{\pi}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_G)^T$ .

## 5.2 E-step

The E-step on the  $(k+1)^{st}$  iteration of the EM algorithm requires the computation of the conditional expectation of  $l_c(\boldsymbol{\Psi})$  given the observed data,  $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$  and the current value of the parameter, say  $\boldsymbol{\Psi}^{(k)}$ . This expectation is denoted  $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}) = E_{\boldsymbol{\Psi}^{(k)}}\{l_c(\boldsymbol{\Psi})\}$ . As the complete-data log likelihood is linear in the  $z_{ig}$ , the conditional expectation reduces to taking the expectation of  $Z_{ig}$  given  $\mathbf{y}_i$ . Because  $Z_{ig}$  given  $\mathbf{y}_i$  is a Bernoulli random variable with probability of success given by

$$\pi_{ig}^{*(k)} = \frac{\pi_g^{(k)} f(\mathbf{y}_i; \boldsymbol{\theta}_g^{(k)}, m)}{\sum_{g=1}^G \pi_g^{(k)} f(\mathbf{y}_i; \boldsymbol{\theta}_g^{(k)}, m)} \quad (22)$$

Therefore,  $E_{\boldsymbol{\Psi}^{(k)}}(Z_{ig} | \mathbf{y}_i) = P_{\boldsymbol{\Psi}^{(k)}}(Z_{ig} = 1 | \mathbf{y}_i) = \pi_{ig}^{*(k)}$ . Notice that  $\pi_{ig}^{*(k)}$  is the posterior conditional probability of the  $i^{th}$  sample member belonging to the  $g^{th}$  component given  $\mathbf{y}_i$  at the  $(k+1)^{th}$  iteration of the algorithm.

### 5.3 M-step

The  $(k + 1)^{th}$  iteration of the M-Step requires the maximization of  $Q(\Psi; \Psi^{(k)})$  in order to obtain updated estimates of the parameter,  $\Psi^{(k+1)}$ . By straightforward differentiation, it is easy to see that  $\pi_g^{(k+1)} = \sum_{i=1}^n \pi_{ig}^{*(k)} / n$ . In other words, each observation contributes its respective posterior probability of being in the  $g^{th}$  component to the estimate of the probability of membership in this component. In addition, differentiation with respect to  $\theta_g^{(k)}$  yields the weighted analogue to the non-mixture multinomial differentiation. This, in turn leads to the updated estimates given by

$$\theta_{gj}^{(k+1)} = \frac{\sum_{i=1}^n \pi_{ig}^{*(k)} y_{ij}}{m \sum_{i=1}^n \pi_{ig}^{*(k)}}$$

for  $j = 1, \dots, r$  and  $g = 1, \dots, G$ .

### 5.4 Convergence

There is an extensive literature on assessing the convergence of EM algorithms and several are summarized in McLachlan and Peel (2000). In this paper, we suggest using a relative difference stopping criterion. This is based on the absolute relative difference between parameter estimates at successive iterations of the algorithm. Formally, we suggest stopping the algorithm when

$$\frac{|\Psi_d^{(k)} - \Psi_d^{(k+1)}|}{\Psi_d^{(k)}} < \epsilon \quad \forall d, \tag{23}$$

given some small, prespecified value of  $\epsilon$ , say  $\epsilon = 10e-6$ . This stopping rule is discussed in Schafer (1995).

We close this section by noting that this algorithm is guaranteed to converge to a local maximum, not necessarily a global. Therefore, we recommend running the algorithm from several different sets of initial values to ensure that a maximum is indeed found.

## 6 Discussion

In this paper, we have provided an identifiability criterion to consider when analyzing finite multinomial mixture models. Namely, the number of trials,  $m$ , in a particular multinomial model, and the number of components of the finite mixture,  $K$ , must satisfy the condition  $m \geq 2K - 1$ . If this condition is not satisfied, *e.g.* if  $m < 2K - 1$ , all hope is not necessarily lost. That is, there may be a subclass of the class, say  $\mathcal{F}^* \subset \mathcal{F}$ , which is identifiable.

**Example 6.1 (Constrained Binomial Mixture).** *Consider the class of two-component binomial mixture models,  $\mathcal{B}_{2,2}$ . This class is not identifiable. However, the subclass of two-component binomial mixture models having probability parameters  $p_1$  and  $p_2 = 1 - p_1$ , is identifiable.*

*Proof.* To prove this result, consider the proof given in Appendix A with  $K = 2$  and  $m = 2$ . The result follows immediately. ■

Additional examples similar to 6.1 can be constructed for general multinomial mixtures by constraining the parameter space in an appropriate way. It should be noted that a strong non-identifiability result has *not* been given for the class of multinomial mixture models.

## Appendix A

An outline of the sufficiency proof given in Kim (1984) is presented below. The necessity is similar to what is presented in theorem (4.2)

*Proof. Sufficiency:* In this proof, we will use the fact stated in theorem (4.1). That is, every marginal random variable of  $\mathbf{Y}$  is distributed as a mixture of  $K' \leq K$  binomial

distributions. Therefore, we can apply the results given in Blischke (1964) (see the proof given below) to show that the marginals are identifiable. Kim (1984) points out a theorem given in Chandra (1977) which states that the identifiability of each marginal class implies the identifiability of the joint class. Therefore,  $\mathcal{M}_{m,K}$  is identifiable. ■

The proof given in Blischke (1964) in terms of multinomial parameters.

*Proof.* From theorem 4.1, every marginal random variable of  $Y$  will be distributed as a mixture of  $K' \leq K$  binomial distributions. First, consider any margin  $j$  having  $K$  components in the mixture. Because  $m \geq 2K - 1$ , we can define the  $2K - 1$   $j^{\text{th}}$ -marginal factorial moments ( $j = 1, 2, \dots, r$ ) as

$$\begin{aligned} f_{j1} &= \pi_1 \theta_{1j} + \pi_2 \theta_{2j} + \dots + \pi_k \theta_{Kj} \\ f_{j2} &= \pi_1 \theta_{1j}^2 + \pi_2 \theta_{2j}^2 + \dots + \pi_k \theta_{Kj}^2 \\ &\vdots \\ f_{j,2K-1} &= \pi_1 \theta_{1j}^{2K-1} + \pi_2 \theta_{2j}^{2K-1} + \dots + \pi_k \theta_{Kj}^{2K-1}. \end{aligned}$$

Note that each of these quantities differ from the usual factorial moments by a constant. We will use the moments to show that the parameters  $\pi_1, \dots, \pi_K, \theta_{1j}, \dots, \theta_{Kj}$  can be written in terms of  $f_{j1}, \dots, f_{j,2K-1}$ , uniquely.

Consider the  $K^{\text{th}}$  degree polynomial

$$\begin{aligned} g_j(x) &= (x - \theta_{1j})(x - \theta_{2j}) \dots (x - \theta_{Kj}) \\ &= B_K x^K + B_{K-1} x^{K-1} + \dots + B_2 x^2 + B_1 x + B_0 \end{aligned}$$

where  $B_K = 1$ . We will show that the coefficients,  $B_0, \dots, B_{K-1}$  are functions of the

$f_{j1}, \dots, f_{j,2K-1}$ . Because  $g_j(\theta_{ij}) = 0$  for  $i = 1, 2, \dots, K$ , we have

$$\begin{aligned}
0 &= B_0 + B_1\theta_{1j} + \dots + B_{K-1}\theta_{1j}^{K-1} + \theta_{1j}^K \\
0 &= B_0 + B_1\theta_{2j} + \dots + B_{K-1}\theta_{2j}^{K-1} + \theta_{2j}^K \\
&\vdots \\
0 &= B_0 + B_1\theta_{Kj} + \dots + B_{K-1}\theta_{Kj}^{K-1} + \theta_{Kj}^K.
\end{aligned} \tag{24}$$

Multiplying each equation in (24) by  $\pi_i\theta_{ij}^p$  for  $p = 0, 1, \dots, (K-1)$  for the appropriate  $i$ , we get

$$0 = \sum_{h=0}^K B_h \pi_i \theta_{ij}^{h+p} \quad \forall i, p$$

or written out explicitly for  $i = 1$

$$\begin{aligned}
0 &= B_0\pi_1 + B_1\pi_1\theta_{1j} + \dots + B_{K-1}\pi_1\theta_{1j}^{K-1} + \theta_{1j}^K \\
0 &= B_0\pi_1\theta_{1j} + B_1\pi_1\theta_{1j}^2 + \dots + B_{K-1}\pi_1\theta_{1j}^K + \theta_{1j}^{K+1} \\
&\vdots \\
0 &= B_0\pi_1\theta_{1j}^{K-1} + B_1\pi_1\theta_{1j}^K + \dots + B_{K-1}\pi_1\theta_{1j}^{2K-2} + \theta_{1j}^{2K-1}.
\end{aligned}$$

From this it follows that

$$\begin{aligned}
0 &= \sum_{h=0}^K B_h \sum_{i=1}^K \pi_i \theta_{ij}^{h+p} \quad p = 0, 1, \dots, (K-1) \\
&= \sum_{h=0}^K B_h f_{j,h+p} \quad p = 0, 1, \dots, (K-1)
\end{aligned}$$

which can be written in matrix form as

$$\underbrace{\begin{bmatrix} f_{j0} & f_{j1} & f_{j2} & \dots & f_{j,K-1} \\ f_{j1} & f_{j2} & f_{j3} & \dots & f_{j,K} \\ & & \vdots & & \\ f_{j,K-1} & f_{j,K} & f_{j,K+1} & \dots & f_{j,2K-2} \end{bmatrix}}_{\mathbf{D}_j} \underbrace{\begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_{K-1} \end{bmatrix}}_{\boldsymbol{\beta}} = \underbrace{\begin{bmatrix} -f_{j,K} \\ -f_{j,K+1} \\ \vdots \\ -f_{j,2K-1} \end{bmatrix}}_{\boldsymbol{\phi}}.$$

Here we define  $f_{j0} = \pi_1 + \pi_2 + \dots + \pi_K = 1$ . If  $\mathbf{D}_j$  is nonsingular, then there is a unique solution for  $\beta$  in terms of the  $j^{\text{th}}$  marginal factorial moments. To show that  $\mathbf{D}_j$  is nonsingular, notice that we can write  $\mathbf{D}_j = \mathbf{V}_j \mathbf{A} \mathbf{V}_j^T$  where

$$\mathbf{V}_j = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \theta_{1j} & \theta_{2j} & \dots & \theta_{Kj} \\ \theta_{1j}^2 & \theta_{2j}^2 & \dots & \theta_{Kj}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{1j}^{K-1} & \theta_{2j}^{K-1} & \dots & \theta_{Kj}^{K-1} \end{bmatrix}$$

is a Vandermonde matrix with determinant

$$|\mathbf{V}_j| = \prod_{i < i'} (\theta_{i'j} - \theta_{ij}) \neq 0.$$

Further,  $\mathbf{A} = \text{diag}(\pi)$ ,  $\pi = (\pi_1, \pi_2, \dots, \pi_K)^T$  which has determinant  $|\mathbf{A}| = \prod_{i=1}^K \pi_i \neq 0$ .

Therefore,

$$\begin{aligned} |\mathbf{D}_j| &= |\mathbf{V}_j| |\mathbf{A}| |\mathbf{V}_j^T| \\ &= \prod_{i=1}^K \pi_i \prod_{i < i'} (\theta_{i'j} - \theta_{ij})^2 \neq 0 \end{aligned}$$

and hence  $\beta = \mathbf{D}_j^{-1} \phi$ . Now, because the coefficients of  $g_j(x)$  are functions of the moments, so are the roots  $\theta_{1j}, \dots, \theta_{Kj}$ .

Next, consider the first  $K - 1$  moments written as

$$\mathbf{f}_j = \begin{bmatrix} f_{j0} \\ f_{j1} \\ \vdots \\ f_{j,K-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \theta_{1j} & \theta_{2j} & \dots & \theta_{Kj} \\ \theta_{1j}^2 & \theta_{2j}^2 & \dots & \theta_{Kj}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{1j}^{K-1} & \theta_{2j}^{K-1} & \dots & \theta_{Kj}^{K-1} \end{bmatrix} \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_K \end{bmatrix} = \mathbf{V}_j \pi.$$

As we have already seen,  $\mathbf{V}_j$  is invertible and hence a unique solution exists for  $\boldsymbol{\pi}$  in terms of the moments as well, or  $\boldsymbol{\pi} = \mathbf{V}_j^{-1} \mathbf{f}_j$ .

Because the above arguments hold for  $j = 1, 2, \dots, R$  and any  $K' \leq K$ , the class of models,  $\mathcal{M}_{m,K}$ , is identifiable. ■

#### ACKNOWLEDGEMENTS

The authors are extremely grateful to the other members of the nonparametric mixture model group: Thomas P. Hettmansperger, David Hunter, Hoben Thomas, Fengjuan Xuan, and Jennifer Hellman. Their insightful comments and contributions throughout the semester certainly made this manuscript possible.

## References

- Blischke, W. R. (1964). Estimating the parameters of mixtures of binomial distributions. *Journal of the American Statistical Association*, 59:510–528.
- Chandra, S. (1977). On the mixture of probability distributions. *The Scandinavian Journal of Statistics*, 4:105–112.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (c/r: P22-37). *Journal of the Royal Statistical Society, Series B, Methodological*, 39:1–22.
- Kim, B. (1984). *Studies of multinomial mixture models*. PhD thesis, University of North Carolina - Chapel Hill.
- Lindsay, B. G. (1995). *Mixture models: theory, geometry, and applications*. Hayward, Institute of Mathematical Statistics.

- McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley and Sons.
- Schafer, J. L. (1995). *Analysis of incomplete multivariate data by simulation*. Chapman & Hall Ltd.
- Teicher, H. (1961). Identifiability of mixtures. *The Annals of Mathematical Statistics*, 32:244–248.
- Teicher, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34:1265–1269.
- Titterton, D., Smith, A., and Makov, U. (1985). *Statistical analysis of finite mixture distributions*. John Wiley and Sons.
- Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39:209–214.