Queen's Economics Department Working Paper No. 1153

Nonparametric Identification and Estimation of Multivariate Mixtures

Hiroyuki Kasahara
University of Western Ontario

Katsumi Shimotsu
Queen's University

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

12-2007

Nonparametric Identification and Estimation of Multivariate Mixtures

Hiroyuki Kasahara
Department of Economics
University of Western Ontario
hkasahar@uwo.ca

Katsumi Shimotsu*
Department of Economics
Queen's University
shimotsu@econ.queensu.ca

December 30, 2007

Abstract

We study nonparametric identifiability of finite mixture models of k-variate data with M subpopulations, in which the components of the data vector are independent conditional on belonging to a subpopulation. We provide a sufficient condition for nonparametrically identifying M subpopulations when $k \geq 3$. Our focus is on the relationship between the number of values the components of the data vector can take on, and the number of identifiable subpopulations. Intuition would suggest that if the data vector can take many different values, then combining information from these different values helps identification. Hall and Zhou (2003) show, however, when $k = 2$, two-component finite mixture models are not nonparametrically identifiable regardless of the number of the values the data vector can take. When $k \geq 3$, there emerges a link between the variation in the data vector, and the number of identifiable subpopulations: the number of identifiable subpopulations increases as the data vector takes on additional (different) values. This points to the possibility of identifying many components even when $k = 3$, if the data vector has a continuously distributed element. Our identification method is constructive, and leads to an estimation strategy. It is not as efficient as the MLE, but can be used as the initial value of the optimization algorithm in computing the MLE. We also provide a sufficient condition for identifying the number of nonparametrically identifiable components, and develop a method for statistically testing and consistently estimating the number of nonparametrically identifiable components. We extend these procedures to develop a test for the number of components in binomial mixtures.

Key words and phrases: finite mixture; binomial mixture; model selection; number of components; rank estimation

*Mailing address: Hiroyuki Kasahara, Department of Economics, University of Western Ontario, London, Ontario N6A 5C2, Canada. Katsumi Shimotsu, Department of Economics, Queen's University, Kingston, Ontario K7L 3N6, Canada.

The financial support from SSHRC and Royal Bank of Canada Fellowship is gratefully acknowledged.

1 Introduction: finite mixture models with independent marginals

Consider the following M-component finite mixture model of a k-vector $X = (X_1, \dots, X_k)$, where the elements of X are independently distributed within each component:

$$F(x) = F(x_1, \dots, x_k) = \sum_{m=1}^{M} \pi^m \prod_{j=1}^{k} F^{jm}(x_j), \quad \pi^m > 0, \quad \sum_{m=1}^{M} \pi^m = 1, \tag{1}$$

where $F(x)$ is the distribution function of X, π^m is the mixture proportion of the mth subpopulation, and $F^{jm}(x_j)$ is the distribution function of X_j conditional on being from the mth subpopulation.

We study the nonparametric identifiability of this mixture model, i.e., whether the information from $F(x)$ can uniquely determine π^m and $F^{jm}(x_j)$'s when no parametric restrictions are imposed on them. Analyzing nonparametric identification is relevant for applied work, because there is rarely theory to guide the specification of component distributions. For example, Cruz et al. (2004) report a simulation result in which imposing incorrect parametric restrictions on component distributions leads to erroneous inference. Zhou et al. (2005) develop a nonparametric maximum likelihood method for $M = 2$ to estimate ROC curves in the absence of a gold standard.

Nonparametric identifiability of finite mixtures has recently attracted increasing attention. Hall and Zhou (2003) and Hall et al. (2005) analyze nonparametric identifiability of the above mixture model with two components ($M = 2$). Hettmansperger and Thomas (2000) and Cruz-Medina et al. (2004) analyze nonparametric identification of models analogous to (1). Their approach involves reducing multivariate data to binomial or multinomial variables, and applying the identification theory for binomial and multinomial mixtures of Blischke (1964) and Elmore and Wang (2003).

Hall and Zhou (2003) show that $k \geq 3$ is necessary and sufficient for nonparametric identification when $M = 2$. Somewhat surprisingly, the non-identifiability for $k = 2$ by Hall and Zhou (2003) holds regardless of the number of values the X_j's can take. If X_j takes J different values, then considering $F(x)$ for all possible values of X provides $J^k - 1$ restrictions, whereas the number of unknowns is $kJM + M - 1$. Hence, as J increases, the number of restrictions increases at an exponential rate, whereas the number of unknowns increases only linearly. Intuition would suggest that this additional information helps identification. This is not the case, however, when $k = 2$.

When $k \geq 3$, combining information from different x's changes the picture substantially. Now the model (1) is nonparametrically identifiable. We provide a sufficient condition that enables one to identify up to M components. Furthermore, we show that, when $k \geq 3$, there emerges a link between the variation in X and the number of identifiable components: the number of identifiable components increases as X takes more (different) values. If X is continuously distributed, one can identify as many components as desired. Hall et al. (2005) show that the model in (1) is nonparametrically identifiable when $k \geq k_M = (1 + o(1))6M \ln(M)$ as $M \to \infty$. Our results imply that $k = 3$ is sufficient if the X_j's have sufficient variation. Our identification method is constructive, and leads to an estimation strategy. It is not as efficient as the MLE, but can be used as the initial value of the optimization algorithm in computing the MLE.

Testing the number of components in finite mixtures has long been a challenging problem. The asymptotic distribution of the likelihood ratio statistic has been derived recently (Dacunha-Castelle and Gassiat, 1999; Liu and Shao, 2003) but is nonstandard, and not easy to tabulate. There is also a growing literature on consistent estimation of the number of components, including Henna (1985), Leroux (1992), Chen and Kalbfleisch (1996), Dacunha-Castelle and Gassiat, (1997, 1999), Keribin (2000), and James et al. (2001). In these papers, the component distributions are assumed to belong to a parametric family. Little is known of identifiability of the number of components in a nonparametric setting.

We provide a sufficient condition for nonparametrically identifying the number of components. This condition is stated in terms of $F(x)$, and hence testable by using empirical distribution functions. Using this fact, we develop a procedure to statistically test, and consistently estimate the number of nonparametrically identifiable components. It is based on an estimate of the rank of a matrix constructed from the empirical distribution of X. Since our procedure does not require estimating a mixture model, it is computationally easy to implement. Extending this framework, we also develop a procedure to statistically test and consistently estimate the number of components in mixtures of binomial distributions. Simulations illustrate our procedure performs well.

Kasahara and Shimotsu (2007) study nonparametric identification of finite mixture dynamic discrete choice models widely used in econometrics using a similar approach to this paper. This paper analyzes nonparametric identifiability in a more general context of multivariate mixtures, and provides a clearer intuition behind the identification results.

We assume the elements of X (or blocks of the elements of X) are independent conditional on being from a subpopulation, as in Hall and Zhou (2003) (and other existing papers on nonparametric identification mentioned above). Hall and Zhou (2003, section 2.3) and Hall et al. (2005, p. 668) discuss the validity of the assumption of independent marginals, and list the body of work that employs it. Elmore et al. (2004) and Zhou et al. (2005) also employ a model with independent marginals.

The reminder of the paper is organized as follows. Section 2 briefly reviews the non-identifiability for $k = 2$ shown by Hall and Zhou (2003). Section 3 discusses the identifiability under $k \geq 3$ and provides a sufficient condition for nonparametric identification. Section 4 gives a sufficient condition for nonparametrically identifying the number of components, and section 5 introduces a test of the number of mixture components. Section 6 reports simulation results, and proofs are collected in the Appendix.

2 Non-identifiability of finite mixture models under $k = 2$

In this section, we consider a two-component mixture model of a k-dimensional variable $X = (X_1, \ldots, X_k)$:

$$F(x) = F(x_1, \ldots, x_k) = \pi \prod_{j=1}^{k} F^{j1}(x_j) + (1-\pi) \prod_{j=1}^{k} F^{j2}(x_j), \tag{2}$$

where $\pi \in (0,1)$. $F(x)$ denotes the distribution function of the observed data, and $F^{jm}(x_j)$ denotes the univariate distribution function of X_j conditional on being from the mth subpopulation. Let Q be the primitive parameter of this model; $Q = \{\pi, \{\{F^{jm}(x_j)\}_{j=1}^{k}\}_{m=1}^{2}\}$. Q is nonparametrically identified if it is uniquely determined from $F(x)$, and its marginals.

Hall and Zhou (2003) show that this model is nonparametrically non-identifiable if $k = 2$. Somewhat surprisingly, this non-identifiability for $k = 2$ holds regardless of the number of values the X_j's can take. Suppose both X_1 and X_2 can take at least J distinct values, $\{\xi_1, \ldots, \xi_J\}$. Then, considering $F(x)$ for all possible values of X provides $J^2 - 1$ restrictions, whereas the number of unknowns in $Q = \{\pi, \{F^{11}(\xi_l), F^{21}(\xi_l), F^{12}(\xi_l), F^{22}(\xi_l)\}_{l=1}^{J}\}$ is $4J + 1$. This suggests that it may be possible to nonparametrically identify Q if J is sufficiently large. However, the additional restrictions from $F(x)$ at different values of x cancel with each other, and the effective number of restrictions is always smaller than the number of unknowns.

Hall and Zhou (2003) prove the non-identifiability by showing that there exists a continuum of Q's that satisfy (2) for a given $F(x)$ when $k = 2$. In the following, we provide an additional insight to this problem by showing that only $4J - 1$ of these $J^2 - 1$ restrictions are effective.

First, we introduce the *irreducibility* condition used by Hall and Zhou (2003, p. 215). Let $F^j(x_j)$ denote the marginal distribution function of X_j.

Assumption 1 (irreducibility) *$F(x_1, x_2)$ is not identical to $F^1(x_1)F^2(x_2)$ for any x_1 and x_2.*

Note that if the irreducibility condition fails, then we have $F^{j1}(x_j) = F^{j2}(x_j) = F^j(x_j)$ in (2), and that the right hand side of (2) is not uniquely determined with respect to π.

The following proposition shows that all the $J^2 - 1$ restrictions implied by $F(x_1, x_2)$ can be constructed from a set of $4J - 1$ restrictions and, therefore, the number of unknowns in Q is strictly larger than the number of effective restrictions when $k = 2$.

Proposition 1 *Suppose that the distribution function of (X_1, X_2) is given by (2) with $k = 2$, and $F(x_1, x_2)$ satisfies Assumption 1. Suppose $\tilde{Q} = \{\tilde{\pi}, \{\tilde{F}^{11}(\xi_l), \tilde{F}^{21}(\xi_l), \tilde{F}^{12}(\xi_l), \tilde{F}^{22}(\xi_l)\}_{l=1}^{J}\}$ satisfies*

$$F(x_1, x_2) = \tilde{\pi} \prod_{j=1}^{2} \tilde{F}^{j1}(x_j) + (1-\tilde{\pi}) \prod_{j=1}^{2} \tilde{F}^{j2}(x_j), \tag{3}$$

for $(x_1, x_2) = (\xi_1, \xi_1), (\xi_1, \xi_2), \ldots, (\xi_1, \xi_J)$ and $(\xi_2, \xi_1), (\xi_3, \xi_1), \ldots, (\xi_J, \xi_1)$, and $\tilde{Q}$ satisfies

$$F^1(x_1) = \tilde{\pi}\tilde{F}^{11}(x_1) + (1-\tilde{\pi})\tilde{F}^{12}(x_1), \quad F^2(x_2) = \tilde{\pi}F^{21}(x_1) + (1-\tilde{\pi})\tilde{F}^{22}(x_2), \tag{4}$$

for $x_1 = \xi_1, \ldots, \xi_J$*, and* $x_2 = \xi_1, \ldots, \xi_J$*. Then* $\tilde{Q}$ *satisfies (3) for all* $(x_1, x_2) \in \{\xi_1, \ldots, \xi_J\} \times \{\xi_1, \ldots, \xi_J\}$.

3 Sufficient conditions for nonparametric identification when $k \geq 3$

When $k \geq 3$, the restrictions from $F(x)$ at different values of x help identification. The number of identifiable components increases as the number of values the X_j's can take increases. We focus on the case $k = 3$, but the following argument is also valid for $k \geq 3$. The distribution function of X is

$$F(x) = \pi^1 \prod_{j=1}^{k} F^{j1}(x_j) + \cdots + \pi^M \prod_{j=1}^{k} F^{jM}(x_j), \tag{5}$$

where $\pi^m > 0$ and $\sum_{m=1}^{M} \pi^m = 1$. Let $\mathcal{X}_j$ denote the support of X_j. Consider a partition of $\mathcal{X}_j$ into M subsets, $\Xi_1^j, \ldots, \Xi_M^j$. Define, for $a, b, c = 1, \ldots, M$,

$$\begin{aligned}
p_a^{jm} &= \mathbb{P}(X_j \in \Xi_a^j | X_j \text{ is from the } m\text{th subpopulation}) = \int 1\{x_j \in \Xi_a^j\} dF^{jm}(x_j), && (6)\\
P_{a,b}^{12} &= \mathbb{P}(X_1 \in \Xi_a^1, X_2 \in \Xi_b^2) = \sum_{m=1}^{M} \pi^m p_a^{1m} p_b^{2m}, && (7)\\
P_{a,b,c}^{123} &= \mathbb{P}(X_1 \in \Xi_a^1, X_2 \in \Xi_b^2, X_3 \in \Xi_c^3) = \sum_{m=1}^{M} \pi^m p_a^{1m} p_b^{2m} p_c^{3m}. && (8)
\end{aligned}$$

Arrange the $p_\cdot^{jm}$'s into two $M \times M$ matrices as

$$L_j = \begin{bmatrix} p_1^{j1} & \cdots & p_M^{j1} \\ \vdots & \ddots & \vdots \\ p_1^{jM} & \cdots & p_M^{jM} \end{bmatrix}, \quad j = 1, 2. \tag{9}$$

The mth row of L_j represents the distribution function of X_j conditional on being from the mth subpopulation. Define, for $h \in \{1, \ldots, M\}$, two $M \times M$ matrices as

$$P = \begin{bmatrix} P_{1,1}^{12} & \cdots & P_{1,M}^{12} \\ \vdots & \ddots & \vdots \\ P_{M,1}^{12} & \cdots & P_{M,M}^{12} \end{bmatrix}, \quad P_h = \begin{bmatrix} P_{1,1,h}^{123} & \cdots & P_{1,M,h}^{123} \\ \vdots & \ddots & \vdots \\ P_{M,1,h}^{123} & \cdots & P_{M,M,h}^{123} \end{bmatrix}. \tag{10}$$

Define $V = \text{diag}(\pi^1, \ldots, \pi^M)$ and $D_h = \text{diag}(p_h^{31}, \ldots, p_h^{3M})$. Then P and P_h are expressed as

$$P = L_1' V L_2, \quad P_h = L_1' D_h V L_2 = L_1' V D_h L_2. \tag{11}$$

The following proposition and corollary provide a sufficient condition for nonparametrically identifying L_1, L_2, V, and D_h. Here, P and P_h are functions of the observables, while L_1, L_2, V, and D_h are unknowns. The restrictions from P alone are not sufficient to determine L_1, L_2 and V uniquely - additional information from P_h enables the identification.

Proposition 2 *Suppose P is nonsingular and we can find h such that the characteristic roots of $P_h P^{-1}$ are distinct. Then L_1, L_2, D_h, and V are uniquely determined from P and P_h.*

Corollary 1 *Suppose L_1 and L_2 are nonsingular and that there exists h such that $p_h^{3m} \neq p_h^{3n}$ for any $m \neq n$. Then, L_1, L_2, D_h, and V are uniquely determined from P and P_h.*

Once L_1 and V are identified, we can identify

$$p_S^{3m} = \mathbb{P}(X_3 \in S | X_3 \text{ is from the } m\text{th subpopulation})$$

for any subset S of $\mathcal{X}_3$. To see why, define $P_{a,S}^{13} = \mathbb{P}(X_1 \in \Xi_a^1, X_3 \in S) = \sum_{m=1}^{M} \pi^m p_a^{1m} p_S^{3m}$, and

$$P_S = \begin{bmatrix} P_{1,S}^{13} \\ \vdots \\ P_{M,S}^{13} \end{bmatrix}, \quad L_S = \begin{bmatrix} p_S^{31} \\ \vdots \\ p_S^{3M} \end{bmatrix}.$$

Then, $P_S = L_1' V L_S$ holds, and L_S is determined uniquely by $L_S = V^{-1}(L_1')^{-1} P_S$. Using the same argument, we can identify p_S^{jm} for any subset S of $\mathcal{X}_j$ for $j = 1, 2$.

Remark 1

1. *Identification requires both L_1 and L_2 to be nonsingular. Therefore, for identifying M components, all the elements of X need to take at least M distinct values. If X is continuously distributed, it is possible to identify as many components as desired.*

2. *The sufficient condition of Proposition 2 relaxes the identification condition by Hall et al. (2005), which requires $k \geq k_M = (1 + o(M)) 6M \log(M)$ as M increases. As long as X has sufficient variation, and L_1 and L_2 are nonsingular, $k = 3$ suffices for identification.*

3. *Hall and Zhou (2003, section 4.2) show the nonparametric non-identifiability of the following model with a continuously distributed random effect: $\psi(x) = \int \{\prod_{j=1}^{k} F_j(x_j|\lambda)\} \phi(\lambda) d\lambda$, where ϕ is the density of the random effect Λ, and $F_j(x_j|\lambda)$ is the distribution function of X_j conditional on the realization λ of Λ. Our results show that, if the random effect has a discrete distribution with finite support, then it is possible to nonparametrically identify $F_j(x_j|\lambda)$, and the distribution function of the random effect.*

4. *Hettmansperger and Thomas (2000) analyze nonparametric identification and inference of the model (5) with conditionally iid marginals ($F^{1m}(x_1) = \cdots = F^{km}(x_k)$) by defining $Y = \sum_{j=1}^{k} 1\{X_j \leq c\}$, and reducing the data to a mixture of binomials. Cruz-Medina et al. (2004) consider splitting the support of X_j further and reducing the data to a mixture of multinomials. In both cases, identification requires $k \geq 2M - 1$. Our results imply that treating each X_j separately, and not reducing the data help identification.*

5. *When $k \geq 4$, and X can be decomposed into $k' \geq 3$ conditionally independent subvectors, we can apply Proposition 2 to these subvectors. For example, assume k is odd, let $Z_1 = (X_1, \ldots, X_{(k-1)/2})$, $Z_2 = (X_{(k-1)/2+1}, \ldots, X_{k-1})$, and assume Z_1, Z_2, and X_k are independent conditional on belonging to a subpopulation. Partition the support of Z_1, Z_2, and X_k to construct P and P_h. When the X_j's have J distinct support points, it is possible to identify up to $J^{(k-1)/2}$ components.*

The proof of Proposition 2 uses a similar idea to that of Proposition 1 of Kasahara and Shimotsu (2007), which in turn is developed starting from the contributions to the analysis of latent class models by Anderson (1954) and Gibson (1955).

In some cases, we have an access to two different samples with different mixing probabilities but the same component distributions. The distribution function of the first and second sample is respectively given by,

$$F(x) = \sum_{m=1}^{M} \pi^m \prod_{j=1}^{k} F^{jm}(x_j), \qquad \bar{F}(x) = \sum_{m=1}^{M} \bar{\pi}^m \prod_{j=1}^{k} F^{jm}(x_j).$$

For example, suppose we have the results of k diagnostic tests from two different groups of patients, whose disease status is unknown. The fraction of patients with disease ($m = 1$) differs across two groups of patients, so $\pi^1 \neq \bar{\pi}^1$. But the distributions of the test outcomes are the same across groups once one conditions on the true disease status, so that the $F^{jm}(x_j)$'s are common.

In this case, we may nonparametrically identify the model even when $k = 2$. Define $V = \text{diag}(\pi^1, \ldots, \pi^M)$ and $\bar{V} = \text{diag}(\bar{\pi}^1, \ldots, \bar{\pi}^M)$, and consider a decomposition similar to (11): $P = L_1' V L_2$ and $\bar{P} = L_1' \bar{V} L_2$. It follows that $P(\bar{P})^{-1} = L_1' V(\bar{V})^{-1}(L_1')^{-1}$. Consequently, $V(\bar{V})^{-1}$ and L_1' are identified with the characteristic roots and characteristic vectors of $P(\bar{P})^{-1}$. Similarly, the characteristic vectors of $\bar{P}P^{-1}$ identify L_2, and we in turn identify V and $\bar{V}$. This result is useful in the context of diagnostic tests (cf., Hall and Zhou, 2003), making it possible to determine the distributional properties of diagnostic tests even when only two tests are available.

4 Identifying the number of components

In this section, we provide a sufficient condition to nonparametrically identify the number of mixture components, M. Section 4.1 analyzes a general case while Section 4.2 studies binomial mixtures.

4.1 General case

In this subsection, we provide a sufficient condition to nonparametrically identify M when the distribution function of X is given by (5). Here, we are interested in identifying M, but not the component distributions such as $F^{jm}(x_j)$. The requirement in k becomes weaker than in Section 3: it is possible to identify M even when $k = 2$.

Let R_1 and R_2 be integers such that $R_1, R_2 \geq M$. We may set R_1 and R_2 to be the same, but it is not necessary to do so. For each $j = 1, 2$, consider a partition of $\mathcal{X}_j$ into R_j subsets, $\Xi^j_1, \ldots, \Xi^j_{R_j}$. Following (6)-(7), define p^{1m}_a, p^{2m}_b and $P^{12}_{a,b}$ for $a = 1, \ldots, R_1$, and $b = 1, \ldots, R_2$. Arrange p^{1m}_a's and p^{2m}_b's into $M \times R_1$ and $M \times R_2$ matrices as

$$L_1^* = \begin{bmatrix} p_1^{11} & \cdots & p_{R_1}^{11} \\ \vdots & \ddots & \vdots \\ p_1^{1M} & \cdots & p_{R_1}^{1M} \end{bmatrix}, \qquad L_2^* = \begin{bmatrix} p_1^{21} & \cdots & p_{R_2}^{21} \\ \vdots & \ddots & \vdots \\ p_1^{2M} & \cdots & p_{R_2}^{2M} \end{bmatrix}. \tag{12}$$

Arrange P^{12}_{ab}'s into a $R_1 \times R_2$ matrix as

$$P^* = \begin{bmatrix} P_{1,1}^{12} & \cdots & P_{1,R_2}^{12} \\ \vdots & \ddots & \vdots \\ P_{R_1,1}^{12} & \cdots & P_{R_1,R_2}^{12} \end{bmatrix}, \tag{13}$$

The intuition behind our identification result is simple. Suppose there is only one component, so that $M = 1$. Then, the joint distribution of X_1 and X_2 is a product of their marginal distributions, and we have $P^* = (L_1^*)' L_2^*$. Consequently, the rank of P^* equals one, which is the number of components. For $M \geq 2$, we may write P^* as $P^* = (L_1^*)' V L_2^*$. Then, the rank of P^* provides information on the rank of L_1^* and L_2^*, which is related to the number of components in our finite mixture model.

Proposition 3 *Define L_1^* and L_2^* as in (12), and define P^* as in (13). Then $M \geq rank(P^*)$. Furthermore, if both L_1^* and L_2^* have rank M, then $M = rank(P^*)$.*

The rank of P^* corresponds to the number of nonparametrically identifiable components from the joint distribution of X_1 and X_2. When L_1^* has only rank $M - 1$, then two components have the same marginal distribution for X_1 with respect to the partition $\Xi^1_1, \ldots, \Xi^1_{R_1}$. Consequently, the variation of X_1 is not sufficient to identify M. The proof of Proposition 3 is essentially the same as the proof of Proposition 3 of Kasahara and Shimotsu (2007), but we provide it in the Appendix for completeness.

When $k \geq 3$, we can group variables into two groups and apply Proposition 3, similarly to Remark 1.5. For example, when k is even, we may let $Z_1 = (X_1, \ldots, X_{k/2})$, $Z_2 = (X_{k/2+1}, \ldots, X_k)$, and partition the support of Z_1 and Z_2 to construct P^*. Reducing the data into bivariate vectors is another option. For example, as our simulation study illustrates, we may define $Z_1 = X_1 + \cdots + X_{k/2}$ and $Z_2 = X_{k/2+1} + \cdots + X_k$, and partition the support of Z_1 and Z_2 to construct P^*.

4.2 Binomial mixtures

Suppose X follows a mixture of binomial distributions, $B(K, p_m)$, in which p_m is the parameter of the mth component distribution:

$$\mathbb{P}(X = k) = \sum_{m=1}^{M} \pi^m (1 - p_m)^{K-k} p_m^k, \quad k = 0, \ldots, K \tag{14}$$

where $0 < p_1 < \cdots < p_M < 1$, $\pi^m > 0$, and $\sum_{m=1}^{M} \pi^m = 1$.

In this subsection, we provide a necessary and sufficient condition to identify M. It has been known that $K \geq 2M - 1$ is both necessary and sufficient to identity the parameters of the model, $\{\pi^m, p^m\}_{m=1}^{M}$ (Teicher, 1961, 1963; Blischke, 1964). However, little is known about the identifiability of M itself. Provided $K \geq 2M - 1$, it is not clear if we can identify how many components are present in this model. In the following, we show that M is identified as the rank of a matrix of the factorial moments of the data.

Similar to Blischke (1964), define the kth (normalized) population factorial moment as

$$f(k) = E\left[\frac{X(X-1)\cdots(X-k+1)}{K(K-1)\cdots(K-k+1)}\right],$$

for $k = 1, \ldots, K$, and define $f(0) = 1$. Then, as shown in Blischke (1962, 1964),

$$f(k) = \sum_{m=1}^{M} \pi^m p_m^k.$$

Let K^* be an even number no larger than K. Define the following $(K^*/2 + 1) \times (K^*/2 + 1)$

matrix

$$P_B = \begin{bmatrix} f(0) & f(1) & \cdots & f(K^*/2) \\ f(1) & f(2) & \cdots & f(K^*/2+1) \\ \vdots & \vdots & \ddots & \vdots \\ f(K^*/2) & f(K^*/2+1) & \cdots & f(K^*) \end{bmatrix}, \tag{15}$$

as well as $V = \text{diag}(\pi^1, \ldots, \pi^M)$ and[1]

$$L_B = \begin{bmatrix} 1 & p_1 & \cdots & p_1^{K^*/2} \\ \vdots & \vdots & & \vdots \\ 1 & p_M & \cdots & p_M^{K^*/2} \end{bmatrix}.$$

Then, it follows that $P_B = L_B' V L_B$, and the rank of P_B provides the information on M via the rank of L_B. Using an analogous argument to the proof of Proposition 3, we obtain the following corollary that identifies M. Its proof can be found in the Appendix.

Corollary 2 *Suppose X follows (14), and assume $K^* \geq 2M - 2$. Define P_B as in (15). Then $M = rank(P_B)$.*

Note that the condition on K is $K^* \geq 2M - 2$. This condition is weaker than $K \geq 2M - 1$, the necessary and sufficient condition for identifying $\{\pi^m, p^m\}_{m=1}^M$. Hence, in order to identify only M, we need one less variation in X.

5 Testing the number of identifiable components

Proposition 3 shows that the rank of P^* gives the lower bound of the number of mixture components. If, in addition, both L_1^* and L_2^* have rank M, then the rank of P^* equals the number of components. Therefore, we may test the (lower bound of the) number of components by estimating P^* and testing its rank.

Several statistics for testing the rank of a matrix have been proposed: the LDU decomposition-based statistic by Gill and Lewbel (1992) and Cragg and Donald (1996), the minimum chi-squared type statistic by Cragg and Donald (1997), the characteristic root-based statistics by Robin and Smith (2000), and the statistics using the singular value decomposition by Kleibergen and Paap (2006). We use the test statistic by Robin and Smith (2000), because it does not need the covariance matrix of the estimate of the matrix to be of full rank.

In the following, we review the statistic by Robin and Smith (2000), and discuss two procedures to estimate the rank of a matrix: sequential hypothesis testing, and a model selection approach.

[1] F_B, V, and L_B corresponds to D, A, and P in Blischke (1964, pp. 513-514).

5.1 Statistic by Robin and Smith (2000)

Let B be a $p \times q$ matrix with $p \geq q$. The matrix B corresponds to P^* or P_B in Section 4. Suppose the rank of B is r_0, where $0 \leq r_0 < q$. Our interest is to estimate r_0 and test H_0 : $\text{rank}(B) = r_0$ against $H_1 : \text{rank}(B) > r_0$, using a consistent estimate of B.

The procedure by Robin and Smith (2000) is based on the estimates of the characteristic roots of BB'. Let $\lambda_1 \geq \cdots \geq \lambda_{r_0} > 0$ and $\lambda_{r_0+1} = \cdots = \lambda_p = 0$ denote the ordered characteristic roots of BB'. Let c_i denote the characteristic vector of BB' associated with λ_i, and collect them into a $p \times p$ matrix $C = (c_1, \ldots, c_p)$. For $0 \leq r < p$, partition C as $C = (C_r, C_{p-r})$, where $C_r = (c_1, \ldots, c_r)$ and $C_{p-r} = (c_{r+1}, \ldots, c_p)$. An alternative representation for the characteristic roots $\lambda_1 \geq \cdots \geq \lambda_{r_0} > 0$ and $\lambda_{r_0+1} = \cdots = \lambda_q = 0$ is obtained as the ordered characteristic roots of $B'B$. Let d_i denote the characteristic vector of $B'B$ associated with λ_i, and collect them into a $q \times q$ matrix $D = (d_1, \ldots, d_q)$. For $0 \leq r < q$, partition D as $D = (D_r, D_{q-r})$, where $D_r = (d_1, \ldots, d_r)$ and $D_{q-r} = (d_{r+1}, \ldots, d_q)$. For unique characteristic roots, the corresponding characteristic vectors are identified up to a normalization of its length, whereas for multiple roots, including zero roots, the corresponding characteristic vectors are identified up to an orthonormal matrix of dimension equal to the multiplicity of the roots.

Let $\hat{B}$ be a root-N consistent estimator of B. The test statistic by Robin and Smith (2000) is based on the characteristic roots of $\hat{B}\hat{B}'$. Let $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ be the ordered characteristic roots of $\hat{B}\hat{B}'$. Robin and Smith (2000) consider the following test statistic:

$$CRT(r) = N \sum_{i=r+1}^{q} \hat{\lambda}_i.$$

Following Robin and Smith (2000), we introduce the following assumptions:

Assumption 2 $\sqrt{N} vec(\hat{B} - B) \to_d N(0, \Omega)$ *where* Ω *is finite and rank* s, $0 < s \leq pq$.

Assumption 3 *If* $r_0 < q \leq p$, *the* $(p - r_0)(q - r_0) \times (p - r_0)(q - r_0)$ *matrix* $(D_{q-r_0} \otimes C_{p-r_0})'\Omega(D_{q-r_0} \otimes C_{p-r_0})$ *is nonzero; that is,* $rk\left[(D_{q-r_0} \otimes C_{p-r_0})'\Omega(D_{q-r_0} \otimes C_{p-r_0})\right] > 0$.

Assumption 4 *There exists* $\hat{\Omega}$ *such that* $\hat{\Omega} \to_p \Omega$.

Assumption 3 requires that at least one column of $D_{q-r_0} \otimes C_{p-r_0}$ is not in the null space of Ω. If Ω has full rank, this assumption is automatically satisfied. Assumption 3 is a weak assumption, and we expect it to hold in most cases we consider.[2] However, in general, this assumption is empirically nonverifiable without an explicit knowledge of Ω and $D_{q-r_0} \otimes C_{p-r_0}$.

Robin and Smith (2000) derive the asymptotic distribution of $CRT(r_0)$ when $r_0 < q$:

[2]Kleibergen and Paap (2006) need a stronger assumption (Assumption 2, p.104) on the rank of a matrix involving Ω, which may be violated, for instance, when we apply it to binomial mixtures.

Proposition 4 *(Robin and Smith, 2000, Theorem 3.2 and Corollary 3.2) If $r_0 < q$ and Assumptions 2-3 hold, $CRT(r_0)$ has an asymptotic distribution described by $\sum_{i=1}^{t} \gamma_i Z_i^2$, where $t \leq \min\{s, (p-r_0)(q-r_0)\}$, and $\gamma_1 \geq \cdots \geq \gamma_t$ are the nonzero ordered characteristic roots of the matrix $(D_{q-r_0} \otimes C_{p-r_0})'\Omega(D_{q-r_0} \otimes C_{p-r_0})$, and $\{Z_i\}_{i=1}^{t}$ are independent standard normal variates.*

Let $\hat{C}$ and $\hat{D}$ be the estimates of C and D derived from $\hat{B}$, and let $\hat{\gamma}_i$ be the estimate of γ_i constructed from $\hat{C}$, $\hat{D}$ and $\hat{\Omega}$. Robin and Smith (2000, Theorem 4.1) show that we can estimate the asymptotic distribution function of $CRT(r_0)$ consistently by $\hat{F}_{r_0}^{CRT}(\cdot)$, the distribution function of $\sum_{i=1}^{(p-r_0)(q-r_0)} \hat{\gamma}_i Z_i^2$. We can approximate this distribution function to any desired degree by simulations, and test $H_0 : \text{rank}(B) = r_0$ against $H_1 : \text{rank}(B) > r_0$.

5.2 Sequential hypothesis testing

We now discuss estimation of r_0. Robin and Smith (2000) consider sequential hypothesis testing: we sequentially test $H_0 : \text{rank}(B) = r$ against $H_1 : \text{rank}(B) > r$ for $r = 0, 1, \ldots, q$,[3] and stop at the first value for r that leads to a nonrejection of H_0.[4] By allowing the significance level of the test to change with the sample size N, it is possible to estimate r_0 consistently. For $r = 0, \ldots, q$, let $\hat{c}_{1-\alpha_N}^{r}$ denote the $100(1-\alpha_N)$ percentile of the cdf $\hat{F}_{r}^{CRT}(\cdot)$, and define

$$\hat{r} = \min_{r \in \{0,\ldots,q\}} \{r : CRT(r) \geq \hat{c}_{1-\alpha_N}^{i}, i = 0, \ldots, r-1, CRT(r) < \hat{c}_{1-\alpha_N}^{r}\}. \quad (16)$$

By letting α_N go to zero at a sufficiently slow rate as the sample size increases, $\hat{r}$ converges to the rank of B.

Proposition 5 *(Robin and Smith, 2000, Theorem 5.2) If the conditions of Proposition 4 and Assumption 4 hold, and if $\alpha_N = o(1)$ and $-N^{-1} \ln \alpha_N = o(1)$ as $N \to \infty$, then $\hat{r} - r_0 = o_p(1)$.*

5.3 Model selection procedure

We propose to employ a model selection procedure to estimate r_0 consistently. Consider the following criterion function

$$S(r) = CRT(r) - f(N)g(r),$$

where $g(r)$ is a (possibly stochastic) penalty function, which is bounded in probability. Define

$$\tilde{r} = \arg\min_{1 \leq r \leq q} S(r).$$

Under a standard condition on $f(N)$ and $g(N)$, this gives a consistent estimate of r_0:

[3] Robin and Smith (2000) propose to test the null for $r = 0, 1, \ldots, p$, but it is not necessary to test the null for $r > q$ because $\text{rank}(B)$ cannot be larger than q.

[4] Cragg and Donald (1997) also use sequential hypothesis testing with their estimator.

Proposition 6 *Suppose that $f(N) \to \infty$, $f(N)/N \to 0$, and $\mathbb{P}(g(r) - g(r_0) < 0) \to 1$ for all $r > r_0$ as $N \to \infty$. Then $\tilde{r} \to_p r_0$.*

If the asymptotic distribution of $S(r_0)$ were chi-squared with $(p-r_0)(q-r_0)$ degrees of freedom, then using $f(N) = 1$ and $g(r) = 2(p-r)(q-r)$ would give an AIC-type criterion, while using $f(N) = \log(N)$ and $g(r) = (p-r)(q-r)$ would give a BIC-type criterion.

In light of the non-standard asymptotic distribution of $CRT(r_0)$, we propose the following penalty function $g(r)$ for a BIC-type criterion:

$$g(r) = (p-r)(q-r)\bar{\gamma}(r) \tag{17}$$

where $\bar{\gamma}(r) = \frac{\sum_{i=1}^{(p-r)(q-r)} \hat{\gamma}_i}{(p-r)(q-r)}$ is the average of the characteristic roots of $(\hat{D}_{q-r} \otimes \hat{C}_{p-r})'\hat{\Omega}(\hat{D}_{q-r} \otimes \hat{C}_{p-r})$. In an AIC-type criterion, $g(r)$ is multiplied by 2. The term $\bar{\gamma}(r)$ in (17) makes our model selection procedure invariant to a rescaling of B.[5] Further, the asymptotic distribution of $CRT(r_0)/\bar{\gamma}(r_0)$ has the same mean as a chi-squared random variable with $(p-r_0)(q-r_0)$ degrees of freedom.

To apply Proposition 6 with $g(r)$ defined in (17), we need additional assumptions to guarantee that $g(r)$ becomes strictly decreasing in r as $N \to \infty$. Using the relation $tr(AB) = tr(BA)$, and the properties of the Kronecker product, we obtain

$$\begin{aligned} g(r) - g(r+1) &= tr[(\hat{d}_{r+1} \otimes \hat{c}_{r+1})'\hat{\Omega}(\hat{d}_{r+1} \otimes \hat{c}_{r+1})] + \sum_{j=r+2}^{p} tr[(\hat{d}_{r+1} \otimes \hat{c}_j)'\hat{\Omega}(\hat{d}_{r+1} \otimes \hat{c}_j)] \\ &\quad + \sum_{i=r+2}^{q} tr[(\hat{d}_i \otimes \hat{c}_{r+1})'\hat{\Omega}(\hat{d}_i \otimes \hat{c}_{r+1})]. \end{aligned} \tag{18}$$

Since $\hat{\Omega}$ is positive semidefinite, it follows that $g(r)$ is nonincreasing in r. $g(r)$ becomes strictly decreasing as $N \to \infty$ if the right hand side of (18) becomes strictly positive for any r. This holds, for example, if $(d_r \otimes c_r)'\Omega(d_r \otimes c_r) > 0$ for $1 \le r \le q$, or if for any $1 \le r \le q$ there exists a pair (i,j) such that $(d_i \otimes c_j)'\Omega(d_i \otimes c_j) > 0$ where $r+1 \le i \le p$ and $r+1 \le j \le q$.

6 Simulation study

6.1 General case: an example with normal mixtures

We conduct Monte Carlo simulation experiments with normal mixtures to assess the finite sample performance of our proposed procedures for selecting the number of components. The

[5] Alternatively, we may consider a BIC-type criterion function of the form $S(r) = CRT(r)/\bar{\gamma}(r) - f(N)g(r)$ with $f(N) = \log(N)$ and $g(r) = (p-r)(q-r)$. These two versions of $S(r)$ performed similarly in simulations that are not reported here.

reported results are based on 10,000 simulated samples. Regarding the number of components, we experiment with $M = 2$ and 3.

While the simulated DGP is a parametric (normal) model, our selection procedures do not assume the knowledge of parametric structures. We partition the support of X_j into R_j subsets such that $\mathbb{P}(X_j \in \Xi_l^j) = 1/R_j$ for $l = 1, \ldots, R_j$. Specifically, let $\bar{x}_\beta^j$ denote the β quantiles of X_j. Let $\beta_l = l/R_j$ for $l = 0, 1, \ldots, R_j$, and define $\Xi_l^j = (\bar{x}_{\beta_{l-1}}^j, \bar{x}_{\beta_l}^j]$ for $l = 1, \ldots, R_j - 1$ and $\Xi_{R_j}^j = (\bar{x}_{\beta_{R_j-1}}, \infty)$.

We construct a consistent estimator of the covariance matrix of $\sqrt{N}\text{vec}(\hat{P}^* - P^*)$ as follows. With a slight abuse of notation, let $X_1, \ldots, X_N$ denote N iid draws of X, and let $X_{t,j}$ denote the jth element of X_t. Let $\hat{P}^*$ be the empirical distribution estimator of P^*: for $a = 1, \ldots, R_1$ and $b = 1, \ldots, R_2$, the (a, b)th element of $\hat{P}^*$ is $\hat{P}_{a,b}^{*12} = N^{-1} \sum_{t=1}^N 1\{X_{t,1} \in \Xi_a^1, X_{t,2} \in \Xi_b^2\}$. Because $\{N\hat{P}_{a,b}^{*12}\}_{a=1,\ldots,R_1,b=1,\ldots,R_2}$ follows a multinomial distribution with the parameter $\{P_{a,b}^{*12}\}$, we can easily see

$$\begin{aligned} E\hat{P}_{a,b}^{*12} &= P_{a,b}^{*12}, \qquad \text{var}(\hat{P}_{a,b}^{*12}) = P_{a,b}^{*12}(1 - P_{a,b}^{*12})/N, \\ \text{cov}(\hat{P}_{a,b}^{*12}, \hat{P}_{c,d}^{*12}) &= -P_{a,b}^{*12} P_{c,d}^{*12}/N, \qquad (a, b) \neq (c, d). \end{aligned}$$

Let Ω denote the $(R_1R_2) \times (R_1R_2)$ covariance matrix of $\sqrt{N}\text{vec}(\hat{P}^* - P^*)$. Note that the rank of Ω is $R_1R_2 - 1$ because $\sum_{a=1}^{R_1} \sum_{b=1}^{R_2} \hat{P}_{a,b}^{*12} = 1$. Let $\theta = \text{vec}(P^*)$, then the ith diagonal element of Ω is given by $\theta_i(1 - \theta_i)$, and the (i, j)th off-diagonal element of Ω is given by $-\theta_i\theta_j$.

We first consider a bivariate normal mixture

$$F(x) = \sum_{m=1}^M \pi^m F^m(x), \tag{19}$$

where $x = (x_1, x_2)'$, and $F^m(x)$ is $N_2(\mu^m, I)$. We set $\mu^1 = (0, 0)'$ and $\mu^2 = (2.0, 1.0)'$ for $M = 2$. For $M = 3$, we set, in addition, $\mu^3 = (4.0, 3.0)'$. The mixing probabilities are equal across subpopulations, so that $\pi^1 = \pi^2 = 1/2$ for $M = 2$, while $\pi^1 = \pi^2 = \pi^3 = 1/3$ for $M = 3$. R_1 and R_2 are chosen to $R_1 = R_2 = M + 1$.[6] In simulations, we use the sample quantiles of X_j's to determine the boundaries of Ξ_l^j. This introduces additional variation, and may affect the asymptotic distribution of $CRT(r)$ statistic, but the consistency of our procedure is not affected. We experimented bootstrapping $CRT(r)$ statistic, however it did not improve the results substantially.

Table 1 reports the result of experiments when the data is generated from the model with two components ($M = 2$). For the sequential hypothesis testing procedure (SHT), the smaller the significance level α is, the more likely the procedure underestimates the number of components. The performance of the SHT improves at all the significance levels as the sample size increases.

[6]We also experimented with $R_1 = R_2 = M + 2$ (not reported here) and found that the procedures with $R_1 = R_2 = M + 1$ performed better than those with $R_1 = R_2 = M + 2$.

Furthermore, the "optimal" choice of significance level, i.e., α that selects $M = 2$ most frequently, decreases from 0.1 to 0.05, and then to 0.01 as the sample size increases from $N = 50$ to 200, and then to 1000, respectively. These results are in agreement with Proposition 5. Overall, the SHT performs well in reasonably sized samples. The performance of BIC is somewhat disappointing, despite its theoretic superiority to the AIC.

The lower two panels of Table 1 report the performance of the AIC and BIC. With a small sample size of $N = 50$, the AIC performs better than the SHT. With a larger sample size of $N = 200$ however, the AIC substantially overestimates the number of components, highlighting its inconsistency. On the other hand, the BIC performs worse than both the SHT and AIC when $N = 50$, but the performance of BIC is comparable to that of the SHT when $N = 1000$.

Table 2 reports the simulation results when the data is generated from the model with three components ($M = 3$). The overall pattern is similar to Table 1, but the tendency to underestimate M is more pronounced. For the SHT and BIC, the frequency of choosing $M = 3$ approaches one as the sample size increases. The AIC performs better than the SHT and BIC when $N = 100$ and $N = 400$, but overestimates the number of components more often than the SHT and BIC when $N = 2000$.

Next, we consider a trivariate normal mixture of the form (19) where $x = (x_1, x_2, x_3)'$ and $F^m(x)$ is $N_3(\mu^m, I)$. To apply our selection procedure to trivariate mixtures, we group the second and the third variables into one group as $Z_2 = (X_2, X_3)'$. We consider a partition of $\mathcal{X}_1$ into $R_1 = M + 1$ subsets while $\mathcal{X}_2$ and $\mathcal{X}_3$ are partitioned into $R_2 = R_3 = M$ subsets and, thus, the support of Z_2 is partitioned into M^2 subsets.[7] For instance, for the model with two components, we estimate the rank of the following matrix (see (13)):

$$P^* = \begin{bmatrix} P_{1,(1,1)} & P_{1,(1,2)} & P_{1,(2,1)} & P_{1,(2,2)} \\ P_{2,(1,1)} & P_{2,(1,2)} & P_{2,(2,1)} & P_{2,(2,2)} \\ P_{3,(1,1)} & P_{3,(1,2)} & P_{3,(2,1)} & P_{3,(2,2)} \end{bmatrix},$$

where $P_{a,(b,c)} = \mathbb{P}(X_1 \in \Xi_a^1, Z_2 \in \Xi_b^2 \times \Xi_c^3)$.

Table 3 shows the result of trivariate mixtures for a two-components model.[8] We set the first two variables, (X_1, X_2), to have the same distribution as the bivariate case, thus $\mu^1 = (0, 0)'$ and $\mu^2 = (2.0, 1.0)'$. We experiment with two different distributions of X_3. The first panel of Table 3 reports the case where X_3 has the same distribution as X_2, i.e., $E[X_3|m = 1] = 0$ and $E[X_3|m = 2] = 1$. Comparing the first panel of Table 3 with Table 1, we find that the our selection procedures perform better with trivariate mixtures than with bivariate mixtures across different procedures and sample sizes. Thus, the additional information from the third variable can improve the performance of our selection procedures.

[7] We also experimented a partition of $\mathcal{X}_2$ and $\mathcal{X}_3$ into $M - 1$ or $M + 1$ subsets, but the results did not improve for either two components models or three components models.

[8] The results of trivariate mixtures for three components model are similar and, thus, not reported here.

This is not necessarily the case, however, when the third variable contains little information for distinguishing different subpopulations. The second panel of Table 3 reports the case in which the distribution of the third variable similar across different subpopulations; specifically, $E[X_3|m=1]=0$ and $E[X_3|m=2]=0.5$. Comparing them with the result of Table 1, we notice that our procedure performs worse with trivariate mixtures than with bivariate mixtures in these cases.

Instead of grouping the second and third variables into one group as $Z_2=(X_2,X_3)'$, we may consider a sum of the second and the third variables: $Z_2=X_2+X_3$. The results are reported in Table 4. Comparing it with the result of Table 1, our procedure now performs better with trivariate mixtures than with bivariate mixtures even under the assumption that $E[X_3|m=1]=0$ and $E[X_3|m=2]=0.5$. In this case, the means of both X_2 and X_3 are higher when $m=2$ than when $m=1$ and, as a result, the information for distinguishing different subpopulations is augmented by summing up X_2 and X_3.

We have to be cautious, however, of applying this method blindly because it is possible that the summation operation could reduce the information for distinguishing different subpopulations. The second panel of Table 4 illustrates this point under the alternative assumption that $E[X_3|m=1]=0.5$ and $E[X_3|m=2]=0$; in this case, if we use $Z_2=X_2+X_3$ instead of grouping variable $Z_2=(X_2,X_3)'$, our procedure performs worse.

6.2 Binomial mixtures

We also conduct Monte Carlo simulations for mixtures of binomial distributions, $B(K,p_m)$, as defined in (14) with $M=2,3$, and 4. We set $(p_1,p_2)=(0.2,0.5)$, $(p_1,p_2,p_3)=(0.2,0.5,0.9)$, and $(p_1,p_2,p_3,p_4)=(0.05,0.3,0.7,0.95)$ for models with two, three and four components, respectively. The value of K is chosen to $K=2M$ so that the maximum identifiable number of components is the true number of components plus one. As before, the mixing probabilities are set to equal to each other across subpopulations.

For binomial mixtures, we construct a consistent estimate of Ω from an estimate of the covariance matrix of the sample factorial moments. Define $\nu(X,k)=\frac{X(X-1)\cdots(X-k+1)}{K(K-1)\cdots(K-k+1)}$ so that $f(k)=E(\nu(X,k))$. We estimate $f(k)$ by $\hat{f}(k)=N^{-1}\sum_{i=1}^{N}\nu(X_i,k)$. Hence, $N\text{cov}(\hat{f}(j),\hat{f}(k))$ is equal to $E(\nu(X,j)\nu(X,k))-E(\nu(X,j))E(\nu(X,k))$, which is a linear function of $EX,\ldots,EX^{j+k}$ and, thus, can be estimated from sample moments of X.

Tables 5, 6, and 7 show the results for models with two, three, and four components, respectively. Across three different models, as the sample size increases, the frequency to select the true number of components approaches one in the SHT and BIC; on the other hand, the AIC tends to overestimate the true number of components. It is also seen that a relatively large number of observations is required to estimate M accurately when M is large.

7 Appendix

7.1 Proof of Proposition 1

First, note that, if the joint distribution function of (X_1, X_2) is given by (2) with $k = 2$, then the marginal distribution function of X_1 and X_2 is given by $F^1(x_1) = \pi F^{11}(x_1) + (1-\pi)F^{12}(x_1)$ and $F^2(x_2) = \pi F^{21}(x_1) + (1-\pi)F^{22}(x_2)$, respectively. In light of $F(x_1, x_2) = \pi F^{11}(x_1)F^{21}(x_2) + (1-\pi)F^{12}(x_1)F^{22}(x_2)$, it follows that

$$F(x_1, x_2) - F^1(x_1)F^2(x_2) = \pi(1-\pi)[F^{11}(x_1) - F^{12}(x_1)][F^{21}(x_2) - F^{22}(x_2)]. \qquad (20)$$

Using the irreducibility, we have, *for any* $x_a, x_b, x_c \in \{\xi_1, \dots, \xi_J\}$,

$$F(x_a, x_b) - F^1(x_a)F^2(x_b) = \frac{[F(x_a, x_c) - F^1(x_a)F^2(x_c)][F(x_c, x_b) - F^1(x_c)F^2(x_b)]}{F(x_c, x_c) - F^1(x_c)F^2(x_c)}.$$

Let $(x_a, x_b, x_c) = (\xi_i, \xi_j, \xi_1)$, then

$$F(\xi_i, \xi_j) - F^1(\xi_i)F^2(\xi_j) = \frac{[F(\xi_i, \xi_1) - F^1(\xi_i)F^2(\xi_1)][F(\xi_1, \xi_j) - F^1(\xi_1)F^2(\xi_j)]}{F(\xi_1, \xi_1) - F^1(\xi_1)F^2(\xi_1)}. \qquad (21)$$

Since $\tilde{Q}$ satisfies (3) and (4) for $(x_1, x_2) = \{(\xi_1, \xi_1), (\xi_1, \xi_i), (\xi_j, \xi_1)\}$, the relation (20) holds for these pairs of (x_1, x_2). Therefore, the right hand side of (21) equals $\tilde{\pi}(1-\tilde{\pi})[\tilde{F}^{11}(\xi_i) - \tilde{F}^{12}(\xi_i)][\tilde{F}^{21}(\xi_j) - \tilde{F}^{22}(\xi_j)]$, and hence $\tilde{Q}$ satisfies (3) for $(x_1, x_2) = (\xi_i, \xi_j)$. Repeating the above for all pairs of (ξ_i, ξ_j) gives the stated result. □

7.2 Proof of Proposition 2 and Corollary 1

Since P is nonsingular, we can construct a matrix $B_h = P_h P^{-1} = L_1' D_h (L_1')^{-1}$. Because $B_h L_1' = L_1' D_h$, the characteristic roots of B_h determine the diagonal elements of D_h, and the characteristic vectors of B_h determine the columns of L_1' uniquely up to multiplicative constants. Since $p_1^{1m} + \cdots + p_M^{1m} = 1$ for each m, each column of L_1' must sum to one, and hence the columns of L_1' are uniquely determined. Having determined L_1', we can recover the rows of L_2 uniquely up to multiplicative constants from $(L_1')^{-1}P$ because $(L_1')^{-1}P = VL_2$. Since $p_1^{2m} + \cdots + p_M^{2m} = 1$ for each m, each row of L_2 must sum to one, and hence the rows of L_2 are uniquely determined. Then V is determined as $V = (L_1')^{-1}P(L_2)^{-1}$.

Corollary 1 is proven by observing that P is nonsingular and the characteristic roots of $P_h P^{-1}$ are distinct when the conditions of Corollary 1 are satisfied. □

7.3 Proof of Proposition 3

Let $V = \text{diag}(\pi^1, \ldots, \pi^M)$, then $P^* = (L_1^*)' V L_2^*$. It follows that $\text{rank}(P^*) \leq \min\{\text{rank}(L_1^*), \text{rank}(L_2^*), \text{rank}(V)\}$. Since $\text{rank}(V) = M$, it follows that $\text{rank}(P^*) \leq M$ where the inequality becomes strict when $\text{rank}(L_1^*)$ or $\text{rank}(L_2^*)$ is smaller than M.

When $\text{rank}(L_1^*) = \text{rank}(L_2^*) = M$, multiplying both sides of $P^* = (L_1^*)' V L_2^*$ from the right by $(L_2^*)'(L_2^*(L_2^*)')^{-1}$ gives $P^*(L_2^*)'(L_2^*(L_2^*)')^{-1} = (L_1^*)'V$. There are M linearly independent columns in $(L_1^*)'V$, because $(L_1^*)'$ has M linearly independent columns while V is a diagonal matrix with strictly positive elements. Thus, $\text{rank}(P^*(L_2^*)'(L_2^*(L_2^*)')^{-1}) = M$. Hence, $M \leq \min\{\text{rank}(P^*), \text{rank}(L_2^*), \text{rank}(L_2^*(L_2^*)')^{-1}\} \leq \text{rank}(P^*)$, and it follows that $\text{rank}(P^*) = M$. $\square$

7.4 Proof of Corollary 2

Since $P_B = L_B' V L_B$, if follows from the proof of Proposition 3 that $\text{rank}(P_B) \leq M$. In view of the proof of Proposition 3, $\text{rank}(P_B) = M$ follows if we show $\text{rank}(L_B) = M$.

First, $\text{rank}(L_B) \leq M$ because L_B is a $M \times (K^*/2+1)$ matrix. To show $\text{rank}(L_B) \geq M$, first note that the condition $K^* \geq 2M - 2$ guarantees that $K^*/2 \geq M - 1$. Consider the following $M \times M$ submatrix of L_B:

$$L_B^* = \begin{bmatrix} 1 & p_1 & \cdots & p_1^{M-1} \\ \vdots & \vdots & & \vdots \\ 1 & p_M & \cdots & p_M^{M-1} \end{bmatrix}.$$

Since L_B^* is a Vendermonde matrix, its determinant is given by $\prod_{i<j}(p_j - p_i)$, which is nonzero by definition. Hence, $\text{rank}(L_B^*) = M$. Since L_B^* is a submatrix of L_B, $\text{rank}(L_B) \geq \text{rank}(L_B^*) = M$. It follows that $\text{rank}(L_B) = M$. $\square$

7.5 Proof of Proposition 6

First, we show $\mathbb{P}(\tilde{r} < r_0) \to 0$. If $\tilde{r} < r_0$, this implies $S(r) < S(r_0)$ for some $r < r_0$. Thus $\mathbb{P}(\tilde{r} < r_0) \leq \sum_{r=1}^{r_0-1} \mathbb{P}(S(r) < S(r_0))$. Now, for any $r < r_0$,

$$\begin{aligned} \mathbb{P}(S(r) < S(r_0)) &= \mathbb{P}(CRT(r) - CRT(r_0) - f(N)g(r) + f(N)g(r_0) < 0) \\ &\leq \mathbb{P}\left(N \sum_{i=r+1}^{r_0} \hat{\lambda}_i + f(N)[g(r_0) - g(r)] < 0\right). \end{aligned}$$

This probability tends to 0 as $N \to \infty$ because $f(N)/N \to 0$ and $\sum_{i=r+1}^{r_0} \hat{\lambda}_i \to_p \sum_{i=r+1}^{r_0} \lambda_i > 0$ since the λ_i's are continuous functions of the elements of B.

Second, we show $\mathbb{P}(\tilde{r} > r_0) \to 0$. Similarly as above, we have $\mathbb{P}(\tilde{r} > r_0) \leq \sum_{r=r_0+1}^{q} \mathbb{P}(S(r) <$

$S(r_0)$). Now, for any $r > r_0$,

$$\mathbb{P}(S(r) < S(r_0)) \leq \mathbb{P}\left(-N \sum_{i=r_0+1}^{r} \hat{\lambda}_i + f(N)[g(r_0) - g(r)] < 0\right).$$

This probability tends to 0 as $N \to \infty$ because $N \sum_{i=r_0+1}^{r} \hat{\lambda}_i$ converges to a weighted sum of chi-squared variables, $f(N) \to \infty$, and $\mathbb{P}(g(r_0) - g(r) > 0) \to 1$ as $N \to \infty$. □

References

Anderson, T. W. (1954), "On estimation of parameters in latent structure analysis," *Psychometrika*, 19, 1-10.

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

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

Chen, J. and Kalbfleisch, J.D. (1996), "Penalized minimum-distance estimates in finite mixture models," *Canadian Journal of Statistics*, 24, 167-175.

Cragg, J. G. and Donald, S. G. (1996). "On the asymptotic proper ties of LDU-based tests of the rank of a matrix," *Journal of the American Statistical Association*, 91, 1301-1309.

Cragg, J. G. and Donald, S. G. (1997). "Inferring the rank of a matrix," *Journal of Econometrics*, 76, 223-250.

Cruz-Medina, I. R., Hettmansperger, T. P. and Thomas, H. (2004), "Semiparametric mixture models and repeated measures: the multinomial cut point model," *Applied Statistics*, 53, 463-474.

Dacunha-Castelle, D. and Gassiat, E. (1997). "The estimation of the order of a mixture model," *Bernoulli*, 3, 279-299.

Dacunha-Castelle, D. and Gassiat, E. (1999). "Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes," *The Annals of Statistics*, 27, 1178-1209.

Elmore, R. T., Hettmansperger, T. P. and Thomas, H. (2004), "Estimating component cumulative distribution functions in finite mixture models," *Communications in Statistics-Theory and Methods*, 33, 2075-2086.

Elmore, R. T. and Wang, S. (2003), "Identifiability and estimation in finite mixture models with multinomial components," Technical Report 03-04. Department of Statistics, Pennsylvania State University, University Park.

Gibson, W. A. (1955), "An extension of Anderson's solution for the latent structure equations," *Psychometrika*, 20, 69-73.

Gill, L. and Lewbel, A. (1992), "Testing the rank and definiteness of estimated matrices with applications to factor, state-space, and ARMA models," *Journal of the American Statistical Association*, 87, 766-776.

Hall, P. and Zhou, X.-H. (2003), "Nonparametric estimation of component distributions in a multivariate mixture," *The Annals of Statistics*, 31, 201-224.

Hall, P., Neeman, A., Pakyari, R. and Elmore, R. (2005), "Nonparametric inference in multivariate mixtures," *Biometrika*, 92, 667-678.

Henna, J. (1985), "On estimating of the number of constituents of a finite mixture of continuous distributions," *The Annals of the Institute of Statistical Mathematics*, 37, 235-240.

Hettmansperger, T. P. and Thomas, H. (2000), "Almost nonparametric inference for repeated measures in mixture models," *Journal of the Royal Statistical Society, Ser. B*, 62, 811-825.

James, L. F. , Priebe, C. E., and Marchette, D. J. (2001), "Consistent estimation of mixture complexity," *The Annals of Statistics*, 29, 1281-1296.

Kasahara, H. and Shimotsu, K. (2007), "Nonparametric identification of finite mixture models of dynamic discrete choices," Queen's University Working Paper. Available at http://www.econ.queensu.ca/faculty/shimotsu/papers/ident_revise.pdf

Keribin, C. (2000), "Consistent estimation of the order of mixture models," *Sankhyā Series A*, 62, 49-62.

Kleibergen, F. and Paap, R. (2006), "Generalized reduced rank tests using the singular value decomposition," *Journal of Econometrics*, 133, 97-126.

Leroux, B. G. (1992), "Consistent estimation of a mixing distribution," *The Annals of Statistics*, 20, 1350-1360.

Liu, X. and Shao, Y. (2003), "Asymptotics for likelihood ratio tests under loss of identifiability," *The Annals of Statistics*, 31, 807-832.

Robin, J-M. and Smith, R. (2000), "Tests of rank." *Econometric Theory*, 16: 151-175.

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.

Zhou, X. H., Castelluccio, P. and Zhou, C. (2005), "Nonparametric estimation of ROC curves in the absence of a gold standard," *Biometrics*, 61, 600-609.

Table 1: Selection Frequency for the Number of Components: Bivariate Normal with $M = 2$

		$N = 50$			$N = 200$			$N = 1000$		
		Significance level α			Significance level α			Significance level α		
		.10	05	.01	.10	.05	.01	.10	.05	.01
SHT	$M = 1$	0.4936	0.6394	0.8544	0.0245	0.0494	0.1528	0.0000	0.0000	0.0000
	$M = 2$	0.4480	0.3403	0.1437	0.8896	0.9083	0.8389	0.9023	0.9527	0.9902
	$M \geq 3$	0.0584	0.0203	0.0019	0.0859	0.0423	0.0083	0.0977	0.0473	0.0098
AIC	$M = 1$		0.3956			0.0125			0.0000	
	$M = 2$		0.5128			0.8474			0.8448	
	$M \geq 3$		0.0916			0.1401			0.1552	
BIC	$M = 1$		0.7887			0.2807			0.0000	
	$M = 2$		0.2010			0.7044			0.9921	
	$M \geq 3$		0.0103			0.0149			0.0079	

Notes: The parameter values are: $\pi^1 = \pi^2 = 1/2$, $\mu^1 = (0, 0)'$ and $\mu^2 = (2, 1)$.

Table 2: Selection Frequency for the Number of Components: Bivariate Normal with $M = 3$

		$N = 100$			$N = 400$			$N = 2000$		
		Significance level α			Significance level α			Significance level α		
		.10	05	.01	.10	.05	.01	.10	.05	.01
SHT	$M = 1$	0.0000	0.0001	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$M = 2$	0.7090	0.8184	0.9462	0.1650	0.2590	0.5148	0.0000	0.0000	0.0000
	$M = 3$	0.2610	0.1713	0.0522	0.7663	0.7090	0.4811	0.9039	0.9538	0.9900
	$M \geq 4$	0.0300	0.0102	0.0014	0.0687	0.0320	0.0041	0.0961	0.0462	0.0100
AIC	$M = 1$		0.0000			0.0000			0.0000	
	$M = 2$		0.5759			0.0941			0.0000	
	$M = 3$		0.3747			0.7935			0.8456	
	$M \geq 4$		0.0494			0.1124			0.1544	
BIC	$M = 1$		0.0006			0.0000			0.0000	
	$M = 2$		0.9453			0.7275			0.0022	
	$M = 3$		0.0517			0.2685			0.9920	
	$M \geq 4$		0.0024			0.0040			0.0058	

Notes: The parameter values are: $\pi^1 = \pi^2 = \pi^3 = 1/3$, $\mu^1 = (0, 0)'$, $\mu^2 = (2, 1)$, and $\mu^3 = (4, 3)$.

Table 3: Selection Frequency for the Number of Components: Trivariate Normal with $M = 2$

$E[x_3 \mid m=1]=0$ and $E[x_3 \mid m=2]=1$										
		$N = 50$			$N = 200$			$N = 1000$		
		Significance level α			Significance level α			Significance level α		
		.10	05	.01	.10	.05	.01	.10	.05	.01
SHT	$M=1$	0.3864	0.5284	0.7809	0.0020	0.0052	0.0298	0.0000	0.0000	0.0000
	$M=2$	0.5448	0.4410	0.2149	0.8942	0.9396	0.9582	0.8880	0.9396	0.9863
	$M \geq 3$	0.0688	0.0306	0.0042	0.1038	0.0552	0.0120	0.1120	0.0604	0.0137
AIC	$M=1$		0.3084			0.0011			0.0000	
	$M=2$		0.5988			0.8570			0.8501	
	$M \geq 3$		0.0928			0.1419			0.1499	
BIC	$M=1$		0.7792			0.1342			0.0000	
	$M=2$		0.2130			0.8600			0.9990	
	$M \geq 3$		0.0078			0.0058			0.0010	

$E[x_3 \mid m=1]=0$ and $E[x_3 \mid m=2]=0.5$										
		$N = 50$			$N = 200$			$N = 1000$		
		Significance level α			Significance level α			Significance level α		
		.10	05	.01	.10	.05	.01	.10	.05	.01
SHT	$M=1$	0.5766	0.7135	0.9065	0.0384	0.0746	0.2170	0.0000	0.0000	0.0000
	$M=2$	0.3671	0.2625	0.0905	0.8503	0.8681	0.7719	0.8692	0.9296	0.9822
	$M \geq 3$	0.0563	0.0240	0.0030	0.1113	0.0573	0.0111	0.1308	0.0704	0.0178
AIC	$M=1$		0.4910			0.0230			0.0000	
	$M=2$		0.4356			0.8251			0.8288	
	$M \geq 3$		0.0734			0.1519			0.1712	
BIC	$M=1$		0.9144			0.5183			0.0000	
	$M=2$		0.0816			0.4774			0.9976	
	$M \geq 3$		0.0040			0.0043			0.0024	

Notes: The parameter values are: $\pi^1 = \pi^2 = 1/2$, $\mu^1 = (0,0)'$ and $\mu^2 = (2,1)'$.

Table 4: Selection Frequency for the Number of Components: Trivariate Normal with $M = 2$ and using $Z_2 = X_2 + X_3$

$E[x_3\|m=1] = 0$ and $E[x_3\|m=2] = 0.5$ using $Z_2 = X_2 + X_3$										
		$N = 50$			$N = 200$			$N = 1000$		
		Significance level α			Significance level α			Significance level α		
		.10	05	.01	.10	.05	.01	.10	.05	.01
SHT	$M = 1$	0.4483	0.5922	0.8170	0.0132	0.0315	0.1060	0.0000	0.0000	0.0000
	$M = 2$	0.4896	0.3853	0.1799	0.8987	0.9271	0.8875	0.9030	0.9481	0.9899
	$M \geq 3$	0.0621	0.0225	0.0031	0.0881	0.0414	0.0065	0.0970	0.0519	0.0101
AIC	$M = 1$		0.3538			0.0062			0.0000	
	$M = 2$		0.5464			0.8515			0.8425	
	$M \geq 3$		0.0998			0.1423			0.1575	
BIC	$M = 1$		0.7498			0.2030			0.0000	
	$M = 2$		0.2372			0.7843			0.9921	
	$M \geq 3$		0.0130			0.0127			0.0079	

$E[x_3\|m=1] = 0.5$ and $E[x_3\|m=2] = 0$ using $Z_2 = X_2 + X_3$										
		$N = 50$			$N = 200$			$N = 1000$		
		Significance level α			Significance level α			Significance level α		
		.10	05	.01	.10	.05	.01	.10	.05	.01
SHT	$M = 1$	0.8316	0.9096	0.9775	0.6838	0.7899	0.9255	0.1163	0.1920	0.4018
	$M = 2$	0.1467	0.0840	0.0221	0.2832	0.1982	0.0734	0.8052	0.7731	0.5926
	$M \geq 3$	0.0217	0.0064	0.0004	0.0330	0.0119	0.0011	0.0785	0.0349	0.0056
AIC	$M = 1$		0.7537			0.5823			0.0720	
	$M = 2$		0.2118			0.3642			0.7994	
	$M \geq 3$		0.0345			0.0535			0.1286	
BIC	$M = 1$		0.9651			0.9738			0.7829	
	$M = 2$		0.0324			0.0254			0.2153	
	$M \geq 3$		0.0025			0.0008			0.0018	

Notes: The parameter values are: $\pi^1 = \pi^2 = 1/2$, $\mu^1 = (0, 0)'$ and $\mu^2 = (2, 1)'$.

Table 5: Selection Frequency for the Number of Components: Binomial with $M = 2$

		$N = 50$			$N = 200$			$N = 1000$		
		Significance level α			Significance level α			Significance level α		
		.10	05	.01	.10	.05	.01	.10	.05	.01
SHT	$M = 1$	0.7494	0.8752	0.9763	0.2423	0.4083	0.7493	0.0000	0.0001	0.0013
	$M = 2$	0.1887	0.0960	0.0189	0.6860	0.5633	0.2471	0.9104	0.9586	0.9907
	$M \geq 3$	0.0619	0.0288	0.0048	0.0717	0.0284	0.0036	0.0896	0.0413	0.0080
AIC	$M = 1$		0.6279			0.1564			0.0000	
	$M = 2$		0.2748			0.7181			0.8554	
	$M \geq 3$		0.0973			0.1255			0.1446	
BIC	$M = 1$		0.9051			0.6644			0.0025	
	$M = 2$		0.0754			0.3290			0.9904	
	$M \geq 3$		0.0195			0.0066			0.0071	

Notes: The parameter values are $\pi^1 = \pi^2 = 1/2$, $(p^1, p^2) = (0.2, 0.5)$, and $K = 4$.

Table 6: Selection Frequency for the Number of Components: Binomial with $M = 3$

		$N = 100$			$N = 400$			$N = 2000$		
		Significance level α			Significance level α			Significance level α		
		.10	05	.01	.10	.05	.01	.10	.05	.01
SHT	$M = 1$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$M = 2$	0.6918	0.7867	0.9202	0.1949	0.2918	0.5148	0.0000	0.0000	0.0008
	$M = 3$	0.2813	0.2042	0.0791	0.7327	0.6750	0.4810	0.9061	0.9541	0.9910
	$M \geq 4$	0.0269	0.0091	0.0007	0.0724	0.0332	0.0042	0.0939	0.0459	0.0082
AIC	$M = 1$		0.0000			0.0000			0.0000	
	$M = 2$		0.6176			0.1447			0.0000	
	$M = 3$		0.3286			0.7305			0.8492	
	$M \geq 4$		0.0538			0.1248			0.1508	
BIC	$M = 1$		0.0000			0.0000			0.0000	
	$M = 2$		0.8635			0.5155			0.0019	
	$M = 3$		0.1322			0.4778			0.9941	
	$M \geq 4$		0.0043			0.0067			0.0040	

Notes: The parameter values are $\pi^1 = \pi^2 = \pi^3 = 1/3$, $(p^1, p^2, p^3) = (0.2, 0.5, 0.9)$, and $K = 6$.

Table 7: Selection Frequency for the Number of Components: Binomial with $M = 4$

		$N = 200$			$N = 800$			$N = 4000$		
		Significance level α			Significance level α			Significance level α		
		.10	05	.01	.10	.05	.01	.10	.05	.01
SHT	$M = 1$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$M = 2$	0.0061	0.0145	0.0504	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$M = 3$	0.6625	0.7405	0.8378	0.1919	0.2816	0.4917	0.0000	0.0000	0.0004
	$M = 4$	0.3017	0.2338	0.1108	0.7463	0.6933	0.5042	0.9093	0.9575	0.9916
	$M \geq 5$	0.0297	0.0112	0.0010	0.0618	0.0251	0.0041	0.0907	0.0425	0.0080
AIC	$M = 1$		0.0000			0.0000			0.0000	
	$M = 2$		0.0032			0.0000			0.0000	
	$M = 3$		0.5997			0.1393			0.0000	
	$M = 4$		0.3440			0.7515			0.8491	
	$M \geq 5$		0.0531			0.1092			0.1509	
BIC	$M = 1$		0.0000			0.0000			0.0000	
	$M = 2$		0.0325			0.0000			0.0000	
	$M = 3$		0.8222			0.5257			0.0024	
	$M = 4$		0.1426			0.4706			0.9945	
	$M \geq 5$		0.0027			0.0037			0.0031	

Notes: The parameter values are $\pi^1 = \pi^2 = \pi^3 = \pi^4 = 1/4$, $(p^1, p^2, p^3, p^4) = (0.05, 0.3, 0.7, 0.095)$, and $K = 8$.