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

August 5, 2008

Abstract

This article analyzes the identifiability of k-variate, M-component finite mixture models without making parametric assumptions on the component distributions. We consider the identifiability of both the number of components and the component distributions. Under the assumption of conditionally independent marginals that have been used in the existing literature, we reveal an important link between the number of variables (k), the number of values each variable can take, and the number of identifiable components. The number of components (M) is nonparametrically identifiable if $k \geq 2$ and each element of the variables takes at least M different values. The mixing proportions and the component distributions are nonparametrically identified if $k \geq 3$ and each element of the variables takes at least M different values. Our requirement on k substantially improves the existing work, which requires either $k \geq 2M - 1$ or $k \geq 6M \log M$. The number of components is identified by the rank of a matrix constructed from the distribution function of the data. Exploiting this property, we propose a procedure to nonparametrically estimate the number of components.

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

*Address for correspondence: Katsumi Shimotsu, Department of Economics, Queen's University, Kingston, Ontario K7L 3N6, Canada.

1 Introduction

Finite mixture models provide flexible ways to model unobserved population heterogeneity. Because of their flexibility, finite mixtures have been used in numerous applications in diverse fields such as biological, physical, and social sciences. For example, empirical researchers in economics often use finite mixtures to control unobserved individual-specific effects (cf., Keane and Wolpin 1997; Cameron and Heckman 1998). Comprehensive theoretical account and examples of applications can be found in Everitt and Hand (1981), Titterington et al. (1985), McLachlan and Basford (1988), Lindsay (1995), and McLachlan and Peel (2000).

A finite mixture model is characterized by three main determinants; the component distributions, the number of components, and the mixing proportions. Despite their key role in the specification of mixtures, there is often little theoretical guidance for selecting the form of the component distributions and/or the number of components. In many applications, the component distributions are assumed to belong to a certain parametric family, such as normal, even though it may be unrealistic to assume so. The number of components then is either fixed or determined by the fit of the model to the data.

However, the shape of the component distributions and the number of components are related to each other. For example, the skewness in a sample distribution can be attributed either to a mixture of normals or a single skewed distribution (Schork et al. 1990). Further, it has been known that the estimates of the number of components are sensitive to the choice of the component distributions. For example, Roeder (1994) fits normal mixture models to red blood cell sodium-lithium countertransport (SLC) activity data. For the raw data, her test supports a three-component normal mixture, whereas a square root transformation of the data pulls in the large values, and supports a two-component normal mixture. Cruz-Medina et al. (2004) report a simulation result in which imposing incorrect parametric restrictions on the component distributions leads to erroneous inference on the number of components.

This article analyzes the nonparametric identifiability of finite mixture models. We establish sufficient conditions under which the true model (the component distributions, the number of components, and the mixing proportions) is identified from the distribution function of the data when no parametric assumptions are imposed on the component distributions. Nonparametric identifiability of finite mixtures has recently attracted increasing attention. Hall and Zhou (2003) and Hall et al. (2005) analyze nonparametric identifiability of k-variate finite mixture models in which the marginal distributions are independent conditional on belonging to a subpopulation. Hettmansperger and Thomas (2000) and Cruz-Medina et al. (2004) provide sufficient conditions for the nonparametric identification of models analogous to that of Hall and Zhou (2003) by reducing the data to binomial or multinomial responses.

We impose the same assumption as Hall and Zhou (2003) and the papers mentioned earlier: we assume the data are k-variate, and the marginal distributions are independent conditional on belonging to a subpopulation. This independent marginal assumption is a

key assumption, and it is certainly strong. However, it is applicable to many cases in practice. For example, Cruz-Medina et al. (2004) employ this assumption in analyzing data from six repeated measurements of children's reaction time to a cognitive task. It is universally recognized that there are large differences in the ways children approach cognitive tasks. Cruz-Medina et al. (2004) model these differences by a finite mixture, in which each subpopulation represents a child's solution strategy. Repeated measurements of each child may then be assumed as independent, provided that the experiments are designed properly. Hettmansperger and Thomas (2000) use the same model as Cruz-Medina et al. (2004) to analyze data from eight repeated measurements of reactions of college-age women to a cognitive task. Zhou et al. (2005) use a similar model to estimate ROC curves. Further, as argued by Hall et al. (2005), a practical consideration may necessitate imposing independence when modeling multivariate data. For example, a two-component, k-dimensional normal mixture has $k^2 + 3k + 1$ parameters to be estimated.

We make the following contributions. Let X denote the k-vector variable of interest. First, we show that the number of components is nonparametrically identified if $k \geq 2$ and some regularity conditions are satisfied. In our model, the variation in X provides a source of identification, and the identifiable number of components is related to the number of different values X can take. For example, if $k = 2$ and each element of X takes M different values, it is possible to identify the existence of up to M components. Second, we show that, in addition to the number of components, the mixing proportions and the component distributions are nonparametrically identified if $k \geq 3$ and some regularity conditions are satisfied. Here, the requirement on k is stronger than in identifying only the number of components. Similarly, the number of identifiable components is determined by the number of values X can take. For example, if $k = 3$ and each element of X takes M different values, it is possible to identify up to M component distributions. If X is continuously distributed, one can identify as many component distributions as desired.

Our sufficient conditions for nonparametric identification substantially improve the requirement on the number of elements, k, in the existing literature. Under an additional assumption of identically distributed variables, Hettmansperger and Thomas (2000) and Cruz-Medina et al. (2004) transform the data into binomial or multinomial variables and apply the results on the identifiability of binomial and multinomial mixtures of Blischke (1964) and Elmore and Wang (2003). The reduction to binomial and multinomial mixtures, however, leads to a loss of information. The sufficient condition of Hettmansperger and Thomas (2000) and Cruz-Medina et al. (2004) requires $k \geq 2M - 1$, hence the maximum number of identifiable components is quite limited when k is small; for instance, if $k = 3$, at the most, two components are identifiable. In contrast, our analysis suggests that, even when $k = 3$, a large number of components can be identified using the variation in X. Hall et al. (2005) show identifiability using a different approach from ours. The sufficient condition by Hall et

al. (2005) requires $k \geq (1 + o(1))6M \log M$ as $M \to \infty$, which they state "is undoubtedly larger than the minimal value."

Our identification condition of the number of components is stated in terms of the rank of a matrix constructed from the distribution function of X. By estimating the rank of its empirical analogue, we develop a procedure to statistically test this identification condition, and consistently estimate the number of components. Numerous methods to select the number of components have been proposed in a parametric setting (see Henna 1985; Leroux 1992; Lindsay and Roeder 1992; Windham and Cutler 1992; Roeder 1994; Chen and Kalbfleisch 1996; Dacunha-Castelle and Gassiat 1997, 1999; Keribin 2000; James et al. 2001; Woo and Sriram 2006). Our proposed procedure requires the conditional independence assumption but makes no distributional assumptions on the components. Further, most of the existing selection procedures require repeated estimation of a mixture model with a different number of components (e.g., Leroux 1992; Lindsay 1995; Chen and Kalbfleisch 1996), and can be computationally intensive because of possible multiple local maxima in criterion function; on the other hand, our procedure is easy to implement without the requirement of the estimation of a mixture model. We also develop a procedure to statistically test and consistently estimate the number of components in mixtures of binomial distributions. Simulations illustrate that our procedure performs well.

Kasahara and Shimotsu (2008) study nonparametric identification of finite mixture dynamic discrete choice models widely used in econometrics using a similar approach to this article. This article analyzes nonparametric identifiability in a more general context of multivariate mixtures.

The remainder of the article is organized as follows. Section 2 discusses the nonparametric identifiability of the number of components under $k \geq 2$. Section 3 provides a sufficient condition for nonparametric identification of the mixing proportions and the component distributions under $k \geq 3$. Section 4 introduces a test of the number of mixture components. Section 5 reports simulation results, and an empirical example with the same dataset as in Hettmansperger and Thomas (2000). Proofs are collected in the Appendix.

2 Nonparametric identification of the number of components

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

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

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.

In this section, we provide a sufficient condition to nonparametrically identify the number of mixture components, M. Section 2.1 analyzes a general case, whereas Section 2.2 studies binomial mixtures.

2.1 General case

Suppose the distribution function of X is given by (1) with $k = 2$. Here, we are interested in identifying M, but not the component distributions such as $F^{jm}(x_j)$. Let $\mathcal{X}_j$ denote the support of X_j for $j = 1, 2$. Consider a partition of $\mathcal{X}_j$ into R_j subsets, $\Xi_1^j, \ldots, \Xi_{R_j}^j$. We may set R_1 and R_2 to be the same, but it is not necessary to do so. Define, for $j = 1, 2$ and $\alpha = 1, \ldots, R_j$,

$$p_\alpha^{jm} = \Pr(X_j \in \Xi_\alpha^j | X_j \text{ is from the } m\text{th subpopulation}) = \int 1\{x_j \in \Xi_\alpha^j\} dF^{jm}(x_j). \tag{2}$$

Define, for $a = 1, \ldots, R_1$ and $b = 1, \ldots, R_2$,

$$P_{a,b}^{12} = \Pr(X_1 \in \Xi_a^1, X_2 \in \Xi_b^2) = \sum_{m=1}^{M} \pi^m p_a^{1m} p_b^{2m}. \tag{3}$$

Arrange p_a^{1m}'s and p_b^{2m}'s into $M \times R_1$ and $M \times R_2$ matrices as

$$L_1 = \begin{bmatrix} p_1^{11} & \cdots & p_{R_1}^{11} \\ \vdots & \ddots & \vdots \\ p_1^{1M} & \cdots & p_{R_1}^{1M} \end{bmatrix}, \qquad L_2 = \begin{bmatrix} p_1^{21} & \cdots & p_{R_2}^{21} \\ \vdots & \ddots & \vdots \\ p_1^{2M} & \cdots & p_{R_2}^{2M} \end{bmatrix}. \tag{4}$$

The mth row of L_j represents the distribution function of X_j with respect to the partition $\Xi_1^j, \ldots, \Xi_{R_j}^j$ conditional on being from the mth subpopulation. Arrange the P_{ab}^{12}'s into a $R_1 \times R_2$ matrix as

$$P = \begin{bmatrix} P_{1,1}^{12} & \cdots & P_{1,R_2}^{12} \\ \vdots & \ddots & \vdots \\ P_{R_1,1}^{12} & \cdots & P_{R_1,R_2}^{12} \end{bmatrix}, \tag{5}$$

The following proposition establishes that the rank of P gives the lower bound of M.

Proposition 1 *The number of components is no smaller than the rank of P; i.e., $M \geq rank(P)$. Furthermore, if both L_1 and L_2 have rank M, then $M = rank(P)$.*

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 $P = (L_1)' L_2$, where L_1 and L_2 are row vectors. Consequently, the rank of P equals one, which is the number of components. For $M \geq 2$, we may write P as

$P = (L_1)'VL_2$ with $V = \text{diag}(\pi^1, \ldots, \pi^M)$. 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 Section 4, we test the (lower bound of the) number of components by estimating P and testing its rank.

The condition on the rank of L_1 and L_2 is not empirically testable. Recall that the mth row of L_j represents the marginal distribution of X_j conditional on being from the mth subpopulation. If this rank condition fails, a component distribution of X_j can be expressed as a liner combination of the other component distributions, which reduces the effective number of components. Therefore, this rank condition means that the marginal distribution of the X_j's needs to have M different components for $M = \text{rank}(P)$ to hold. The rank condition also requires that $R_1, R_2 \geq M$. Hence, in order to identify M, all the elements of X need to take at least M distinct values. The proof of Proposition 1 follows closely the proof of Proposition 3 of Kasahara and Shimotsu (2008).

When $k \geq 3$, we can group variables into two groups and apply Proposition 1. 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 instance, as our real data example in Section 5.3 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.

2.2 Binomial mixtures

We can use the idea in Section 2.1 to identify the number of components in binomial mixtures. Suppose Y follows an M-component mixture of binomial distributions, $B(K, p_m)$, in which p_m is the parameter of the mth component distribution. It has been known that $K \geq 2M-1$ is both necessary and sufficient to identity the parameters of the model (Teicher, 1961, 1963; Blischke, 1964). However, little is known about the identifiability of M itself. In the following, we show that M is identified as the rank of a matrix of the factorial moments of the data.

When Y follows an M-component mixture of binomial distributions, $B(K, p_m)$, the distribution function of Y is

$$\Pr(Y = k) = \sum_{m=1}^{M} \pi^m (1 - p_m)^{K-k} p_m^k, \quad k = 0, \ldots, K, \tag{6}$$

where $0 < p_1 < \cdots < p_M < 1$, $\pi^m > 0$, and $\sum_{m=1}^{M} \pi^m = 1$. Similar to Blischke (1964), define the kth (normalized) population factorial moment as

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

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

$$P_B = \begin{bmatrix} f(0) & f(1) & \cdots & f(K^*/2) \\ f(1) & f(2) & \cdots & f(K^*/2+1) \\ \vdots & \vdots & \ddots & \vdots \\ f(K^*/2) & f(K^*/2+1) & \cdots & f(K^*) \end{bmatrix}, \tag{7}$$

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 1, we obtain the following corollary that identifies M.

Corollary 1 *Suppose Y follows (6), and assume $K^* \geq 2M-2$. Define P_B as in (7). Then $M = \text{rank}(P_B)$.*

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 Y.

3 Nonparametric identification of the component distributions

In this section, we assume M is known, and provide sufficient conditions for nonparametrically identifying the mixing proportions and the component distributions.

When $k = 2$, Hall and Zhou (2003) prove that there exists a continuum of component distributions that satisfy (1) for a given $F(x)$. Hence, the component distributions in model (1) is nonparametrically non-identifiable if $k = 2$. Somewhat surprisingly, this non-identifiability 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

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

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 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.

When $k \geq 3$, the restrictions from $F(x)$ at different values of x help identification. The mixing proportions and the component distributions are nonparametrically identified, and the number of identifiable components increases as the X_j's take more different values. We focus on the case $k = 3$, but the following argument is also valid for $k \geq 3$. The distribution function of X is given by (1). Consider a partition of $\mathcal{X}_j$ into M subsets, $\Xi_1^j, \ldots, \Xi_M^j$, and define $p_\cdot^{jm}$'s for $j = 1, \ldots, 3$ as in (2). Define, for $a, b, c = 1, \ldots, M$,

$$P_{a,b,c}^{123} \quad = \quad \Pr(X_1 \in \Xi_a^1, X_2 \in \Xi_b^2, X_3 \in \Xi_c^3) = \sum_{m=1}^{M} \pi^m p_a^{1m} p_b^{2m} p_c^{3m}. \tag{8}$$

We use a similar notation to Section 2 but set $R_1 = R_2 = M$. Arrange the $p_\cdot^{1m}$'s and $p_\cdot^{2m}$'s into two $M \times M$ matrices as

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

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 $V = \mathrm{diag}(\pi^1, \ldots, \pi^M)$ and $D_h = \mathrm{diag}(p_h^{31}, \ldots, p_h^{3M})$. 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 2 *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

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

for any subset S of $\mathcal{X}_3$. To see why, define $P_{a,S}^{13} = \Pr(X_1 \in \Xi_a^1, X_3 \in S) = \sum_{m=1}^{M} \pi^m p_a^{1m} p_S^{3m}$, 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, provided that the relevant rank condition is satisfied.*

2. *Hettmansperger and Thomas (2000) analyze nonparametric estimation and inference of the model (1) with conditionally iid marginals 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 $k \geq 2M - 1$.*

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. *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.*

Example 1 *Example 3 of Hettmansperger and Thomas (2000) considers the following experimental data. In the experiment, college-age women are repeatedly asked to adjust a luminous rod to the vertical position in a darkened room, where the rod is surrounded by a luminous square frame tilted 28^o to the right or left. There are $k = 8$ repeated measurements for each woman. The rod's error deviation in degrees from the vertical is used as the measurement in task, i.e., X_j in our notation with $j = 1, \ldots, 8$.*

Hettmansperger and Thomas define the response variable for each subject $Y = \sum_{j=1}^{8} I\{|X_j| \leq 5^o\}$, which is assumed to follow a mixture of binomials. Because $k = 8$, they can identify up to four components. Our Proposition 2 implies, however, that more than four components are identifiable because X_j is continuously distributed. A part of identifying information is lost upon summing the indicator functions, $I\{|X_j| \leq 5^o\}$, over j.

The proof of Proposition 2 uses the idea in the analysis of latent class models by Anderson (1954) and Gibson (1955). Finite mixture models have been studied and widely used in psychometrics and biostatistics under the name of *latent class models*. In latent class models, an observation vector $X = (X_1, \ldots, X_k)$ consists of k dichotomous or polychotomous responses, typically answers to questions or results of diagnoses. The observations are assumed to belong to one of M classes, with the probability of being in class $m \in \{1, \ldots, M\}$ equal to π^m and unknown. The responses are assumed to be conditionally independent given membership in a given latent class.

Anderson (1954) and Gibson (1955) analyze nonparametric identification of latent class models with k dichotomous responses, which is a special case of our model (1) where the X_j's take only two values. Anderson (1954) and Gibson (1955) derive a sufficient condition for nonparametric identification under the assumption $k \geq 2M - 1$. Madansky (1960) extends their analyses to obtain a sufficient condition under the assumption $2^{(k-1)/2} \geq M$, thus relaxing the assumption of Anderson and Gibson. Lazarsfeld and Henry (1968) summarize early contributions to latent class models, and its Chapter 4 nicely summarizes the results of Anderson (1954), Gibson (1955), and Madansky (1960).

Since both Anderson and Gibson consider only dichotomous responses, their sufficient conditions require a large k if M is not very small. Our analysis highlights that the variation in the X_j's provides an important source of identification and helps identification even when k is not very large.

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 =diag$(\pi^1, \ldots, \pi^M)$ and $\bar{V}$ =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 Estimating the number of identifiable components

Proposition 1 in Section 2 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. This suggests that we may estimate (the lower bound of) the number of components by estimating P and its rank.

Several statistics for testing the rank of a matrix have been proposed (see Gill and Lewbel 1992; Cragg and Donald 1996, 1997; Robin and Smith 2000; Kleibergen and Paap 2006). We use the test statistic by Robin and Smith (2000), because it does not require the covariance matrix of the estimate of the matrix to be of full rank. In the following, we briefly review the statistic by Robin and Smith (2000), and then propose two procedures to estimate the number of components by estimating the rank of a matrix: a model selection procedure and a sequential hypothesis testing procedure.

4.1 Statistic by Robin and Smith (2000)

Robin and Smith (2000) propose a test statistic for the rank of a matrix based on the characteristic roots. The basic idea is simple. The rank of a matrix is equal to the number of non-zero characteristic roots. Thus, given an estimate of a matrix, we may test the rank of a matrix by testing the number of characteristic roots that are non-zero.

With a slight abuse of notation, let P be a $p \times q$ matrix with $p \geq q$. Suppose the rank of P is r_0, where $0 \leq r_0 < q$. Our interest is to test $H_0 : \text{rank}(P) = r_0$ against $H_1 : \text{rank}(P) > r_0$, using a consistent estimate of P, denoted by $\hat{P}$.

Given that the rank of P is the same as the rank of PP', Robin and Smith (2000) develop a procedure based on the estimate of the characteristic roots of PP'. Let $\lambda_1 \geq \cdots \geq \lambda_p$ denote

the ordered characteristic roots of PP'. Let $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ be the ordered characteristic roots of $\hat{P}\hat{P}'$. When $p > q$, it is always the case that $\hat{\lambda}_{q+1} = ... = \hat{\lambda}_p = 0$ even in finite sample. Furthermore, if r_0 is the true rank, then we expect that $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_{r_0} > 0$ while $\sum_{i=r_0+1}^{q} \hat{\lambda}_i \to_p 0$ as the sample size increases. Thus, Robin and Smith (2000) propose the following test statistic

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

If $r < r_0$, then $\sum_{i=r+1}^{q} \lambda_i > 0$ and $CRT(r)$ goes to infinity as the sample size increases. When $r = r_0$, the asymptotic distribution of $CRT(r)$ is given by a weighted sum of squared independent standard normal variates as the following Proposition 3 states.

We first introduce the assumptions. Let $C = (c_1, ..., c_p)$ be a $p \times p$ matrix, where c_i denotes the characteristic vector of PP' associated with the i-th largest characteristic root λ_i. For $0 \leq r < p$, partition C as $C = (C_r, C_{p-r})$, where $C_r = (c_1, \ldots, c_r)$ and $C_{p-r} = (c_{r+1}, \ldots, c_p)$. Similarly, let $D = (d_1, ..., d_q)$ be a $q \times q$ matrix, and partition D as $D = (D_r, D_{q-r})$, where d_i denotes the characteristic vector of $P'P$ associated with the i-th largest characteristic root of $P'P$.

Assumption 1 $\sqrt{N}vec(\hat{P} - P) \to_d N(0, \Omega)$ *where* Ω *is finite and rank* s, $0 < s \leq pq$.

Assumption 2 *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; i.e., rank*$[(D_{q-r_0} \otimes C_{p-r_0})'\Omega(D_{q-r_0} \otimes C_{p-r_0})] > 0$.

Assumption 3 *There exists* $\hat{\Omega}$ *such that* $\hat{\Omega} \to_p \Omega$.

Assumption 2 is a weak assumption, requiring 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.

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

Proposition 3 *(Robin and Smith, 2000, Theorem 3.2 and Corollary 3.2) If* $r_0 < q$ *and Assumptions 1-2 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.*

As shown by Robin and Smith (2000, Theorem 4.1), we can estimate the asymptotic distribution function of $CRT(r_0)$ consistently by $\hat{F}_{r_0}^{CRT}(\cdot)$, the distribution function of $\sum_{i=1}^{(p-r_0)(q-r_0)} \hat{\gamma}_i Z_i^2$, where $\hat{\gamma}_i$ is the consistent estimate of γ_i obtained from $\hat{P}$. The distribution function $\hat{F}_{r_0}^{CRT}(\cdot)$ can be easily computed by simulations once the γ_i's are estimated.

4.2 Model selection procedure

We first propose a model selection procedure based on the statistic $CRT(r)$ 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

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

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

Proposition 4 *Suppose that* $f(N) \to \infty$, $f(N)/N \to 0$, *and* $\Pr(g(r) - g(r_0) < 0) \to 1$ *for all* $r > r_0$ *as* $N \to \infty$. *Then* $\tilde{r} \to_p r_0$.

If the asymptotic distribution of $CRT(r_0)$ 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:

$$g(r) = (p - r)(q - r)\bar{\gamma}(r) \tag{12}$$

where $\bar{\gamma}(r) = \sum_{i=1}^{(p-r)(q-r)} \hat{\gamma}_i / \{(p - r)(q - r)\}$ 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 (12) makes our model selection procedure invariant to a rescaling of P. 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 4 with $g(r)$ defined in (12), 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

$$\begin{aligned} g(r) - g(r+1) &= tr\{(\hat{d}_{r+1} \otimes \hat{c}_{r+1})'\hat{\Omega}(\hat{d}_{r+1} \otimes \hat{c}_{r+1})\} + \sum_{j=r+2}^{p} tr\{(\hat{d}_{r+1} \otimes \hat{c}_j)'\hat{\Omega}(\hat{d}_{r+1} \otimes \hat{c}_j)\} \\ &\quad + \sum_{i=r+2}^{q} tr\{(\hat{d}_i \otimes \hat{c}_{r+1})'\hat{\Omega}(\hat{d}_i \otimes \hat{c}_{r+1})\}. \end{aligned} \tag{13}$$

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 (13) 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 $r+1 \leq i \leq p$ and $r+1 \leq j \leq q$.

4.3 Sequential hypothesis testing

Using the consistent estimate for the asymptotic distribution function of $CRT(r_0)$, $\hat{F}^{CRT}_{r_0}(\cdot)$, we may test $H_0 : \text{rank}(P) = r$ against $H_1 : \text{rank}(P) > r$ for any integer value of r. To estimate r_0, we sequentially test $H_0 : \text{rank}(P) = r$ against $H_1 : \text{rank}(P) > r$ starting from $r = 0$, and then $r = 1, \ldots, q$. The first value for r that leads to a nonrejection of H_0 gives our estimate for r_0.

For $r = 0, \ldots, q$, let $\hat{c}^r_{1-\alpha_N}$ denote the $100(1-\alpha_N)$ percentile of the cdf $\hat{F}^{CRT}_r(\cdot)$. Then, our estimator based on sequential hypothesis testing is defined as

$$\hat{r} = \min_{r \in \{0,\ldots,q\}} \{r : CRT(r) \geq \hat{c}^i_{1-\alpha_N}, i = 0, \ldots, r-1, CRT(r) < \hat{c}^r_{1-\alpha_N}\}. \tag{14}$$

The estimator $\hat{r}$ depends on the choice of the significance level α_N. To achieve consistency, we allow the significance level of the test to decrease with the sample size N. The following proposition states that, by letting α_N go to zero at a sufficiently slow rate as the sample size increases, $\hat{r}$ converges to the rank of P.

Proposition 5 *(Robin and Smith, 2000, Theorem 5.2) If the conditions of Proposition 3 and Assumption 3 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 Simulation study

5.1 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 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 $\Pr(X_j \in \Xi^j_l) = 1/R_j$ for $l = 1, \ldots, R_j$. Specifically, let $\bar{x}^j_\beta$ denote the β quantiles of X_j. Let $\beta_l = l/R_j$ for $l = 0, 1, \ldots, R_j$, and define $\Xi^j_l = (\bar{x}^j_{\beta_{l-1}}, \bar{x}^j_{\beta_l}]$ for $l = 1, \ldots, R_j - 1$ and $\Xi^j_{R_j} = (\bar{x}_{\beta_{R_j-1}}, \infty)$.

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 $\hat{P}^{12}_{a,b} = N^{-1}\sum_{t=1}^N 1\{X_{t,1} \in \Xi^1_a, X_{t,2} \in \Xi^2_b\}$. Because $\{N\hat{P}^{12}_{a,b}\}_{a=1,\ldots,R_1, b=1,\ldots,R_2}$ follows a multinomial distribution with the

parameter $\{P^{12}_{a,b}\}$, we can easily see

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

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 $\sum_{a=1}^{R_1}\sum_{b=1}^{R_2}\hat{P}^{12}_{a,b} = 1$. 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

$$F(x) = \sum_{m=1}^{M} \pi^m F^m(x), \tag{15}$$

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 $R_1 = R_2 = M + 1$.[2] In simulations, we use the sample quantiles of X_j's to determine the boundaries of Ξ^j_l. 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. 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 last two rows 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 the BIC is comparable to that of the SHT when $N = 1000$. The performance of the BIC is somewhat disappointing, despite its theoretical superiority to the AIC.

Table 2 reports the simulation results when the data is generated from the model with

[2] 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 $R_1 = R_2 = M + 2$.

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 (15) with two components ($M = 2$), where $x = (x_1, x_2, x_3)'$ and $F^m(x)$ is $N_3(\mu^m, I)$. We set the first two variables, (X_1, X_2), to have the same distribution as the bivariate case while the third variable X_3 has the same distribution as X_2. 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 partition $\mathcal{X}_1$ into 3 subsets while $\mathcal{X}_2$ and $\mathcal{X}_3$ are partitioned into $R_2 = R_3 = 2$ subsets and, thus, the support of Z_2 is partitioned into $2^2 = 4$ subsets. Accordingly, we estimate the rank of the following matrix (see (5)):

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

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

Table 3 shows the result of this model. Comparing Table 3 with Table 1, we find that 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.

5.2 Binomial mixtures

We also conduct Monte Carlo simulations for mixtures of binomial distributions, $B(K, p_m)$, as defined in (6) 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 $\nu(Y, k) = \frac{Y(Y-1)\cdots(Y-k+1)}{K(K-1)\cdots(K-k+1)}$ so that $f(k) = E(\nu(Y, k))$. We estimate $f(k)$ by $\hat{f}(k) = N^{-1}\sum_{i=1}^{N}\nu(Y_i, k)$. Hence, $N\text{cov}(\hat{f}(j), \hat{f}(k))$ is equal to $E(\nu(X, j)\nu(Y, k)) - E(\nu(Y, j))E(\nu(Y, k))$, which is a linear function of $EY, \ldots, EY^{j+k}$ and, thus, can be estimated from sample moments of Y.

Tables 4, 5, and 6 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.

5.3 Example 3 of Hettmansperger and Thomas (2000)

The data consists of $n = 83$ college-age women each with eight replications of Witkin's rod-and-frame task. The response variable, measured as the rod's error deviation in degrees from the vertical, is continuously distributed. We denote the eight response variables for each woman by $\{X_j : j = 1, \ldots, 8\}$. Hettmansperger and Thomas apply various tests of the number of components for binomial mixtures to the transformed data. Lindsay's (1995) gradient function method suggests $M = 4$, the Hellinger and the Pearson penalized distances suggest $M = 2$, and the bootstrapped likelihood ratio test suggests $M = 3$.

We apply our method by taking the average of the first and the last four variables for each woman and defining $Z_1 = \sum_{j=1}^{4} |X_j|/4$ and $Z_2 = \sum_{j=5}^{8} |X_j|/4$. We then partition the space of Z_1 and Z_2 into 4 regions using their quantiles, construct a P matrix corresponding to (5), and estimate the rank of P. The result is presented in Table 7. When $M = 4$, the criterion function takes the value of zero because $q = r$. In this example, the SHT and BIC estimate $M = 3$ while the AIC selects $M \geq 4$.

More generally, there are ${}_8C_4/2 = 35$ ways to construct a pair (Z_1, Z_2) (up to permutation of Z_1 and Z_2) by choosing 4 response variables out of 8 variables. By estimating the rank of P for all the pairs of (Z_1, Z_2), we obtain 35 different estimates of the number of components. Table 8 presents the frequencies of the estimated number of components across these 35 estimates. $M = 3$ is most frequently chosen for all three procedures, followed by $M \geq 4$.

Following Hettmansperger and Thomas, we also construct the response variable $Y = \sum_{j=1}^{8} 1\{|X_j| \leq 5^o\}$, which is viewed as a mixture of binomial distributions, and then estimate the number of components by our method in Section 4. As shown in Table 9, all of the SHT, AIC, and BIC indicate that there are $M = 3$ components.

Acknowledgement

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

6 Appendix: proofs

6.1 Proof of Proposition 1

Let $V = \text{diag}(\pi^1, \ldots, \pi^M)$, then $P = (L_1)'VL_2$. 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 $P = (L_1)'VL_2$ 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 $\text{rank}(P) = M$. □

6.2 Proof of Corollary 1

Since $P_B = L_B'VL_B$, if follows from the proof of Proposition 1 that $\text{rank}(P_B) \leq M$. In view of the proof of Proposition 1, $\text{rank}(P_B) = M$ 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 $\text{rank}(L_B) = M$. □

6.3 Proof of Proposition 2 and Corollary 2

Since P is nonsingular, we can construct a matrix $B_h = P_hP^{-1} = L_1'D_h(L_1')^{-1}$. Because $B_hL_1' = L_1'D_h$, the characteristic roots of B_h determine the 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 2 is proven by observing that P is nonsingular and the characteristic roots of P_hP^{-1} are distinct when the conditions of Corollary 2 are satisfied. □

6.4 Proof of Proposition 4

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

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

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 $\Pr(\tilde{r} > r_0) \to 0$. Similarly as above, we have $\Pr(\tilde{r} > r_0) \leq \sum_{r=r_0+1}^{q} \Pr(S(r) < S(r_0))$. Now, for any $r > r_0$,

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

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 $\Pr(g(r_0) - g(r) > 0) \to 1$ as $N \to \infty$. $\square$

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.

Cameron, S. V. and J. J. Heckman (1998). "Life cycle schooling and dynamic selection bias: models and evidence for five cohorts of American males," Journal of Political Economy 106: 262-333.

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.

Everitt, B. S. and Hand, D. J. (1981). *Finite Mixture Distributions*. New York: Wiley.

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. (2008), "Nonparametric identification of finite mixture models of dynamic discrete choices," Queen's University Working Paper, forthcoming in *Econometrica*. Available at http://www.econ.queensu.ca/faculty/shimotsu/papers/initialEM.pdf

Keane, M. P., and K. I. Wolpin (1997). "The career decisions of young men," Journal of Political Economy 105: 473-522.

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.

Lazarsfeld. P. F. and Henry, N. W. (1968), *Latent Structure Analysis*. Boston: Houghton Mifflin.

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

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry, and Applications*. Hayward: Institute of Mathematical Statistics.

Lindsay, B. G. and Roeder, K. (1992), "Residual diagnostics for mixture models," *Journal of the American Statistical Association*, 87, 785-794.

McLachlan. G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Dekker.

McLachlan, G. and Peel, D. (2000), *Finite Mixture Models*. New York: Wiley.

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

Roeder, K. (1994), "A graphical technique for detecting the number of components in a mixture of normals," *Journal of the American Statistical Association*, 89, 487-495.

Schork, N. J., Weder, A. B., Schork, A. (1990). On the asymmetry of biological frequency distributions. Genetic Epidemiology, 7, 427-446.

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.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

Windham, M. P. and Cutler, A. (1992), "Information ratios for validating mixture analysis," *Journal of the American Statistical Association*, 87, 1188-1192.

Woo, M-J. and Sriram, T. N. (2006), "Robust estimation of mixture complexity," *Journal of the American Statistical Association*, 101, 1475-1486.

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 Frequencies of the Number of Components: Bivariate Normal with $M = 2$

		$N = 50$			$N = 200$			$N = 1000$		
		$M = 1$	$M = 2$	$M \geq 3$	$M = 1$	$M = 2$	$M \geq 3$	$M = 1$	$M = 2$	$M \geq 3$
SHT	$\alpha = .10$	0.494	0.448	0.058	0.025	0.890	0.086	0.000	0.902	0.098
	$\alpha = .05$	0.639	0.340	0.020	0.049	0.908	0.042	0.000	0.953	0.047
	$\alpha = .01$	0.854	0.144	0.002	0.153	0.839	0.008	0.000	0.990	0.010
AIC		0.396	0.513	0.092	0.013	0.847	0.140	0.000	0.845	0.155
BIC		0.789	0.201	0.010	0.281	0.704	0.015	0.000	0.992	0.008

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

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

		$N = 100$				$N = 400$			
		$M = 1$	$M = 2$	$M = 3$	$M \geq 4$	$M = 1$	$M = 2$	$M = 3$	$M \geq 4$
SHT	$\alpha = .10$	0.000	0.709	0.261	0.030	0.000	0.165	0.766	0.069
	$\alpha = .05$	0.000	0.818	0.171	0.010	0.000	0.259	0.709	0.032
	$\alpha = .01$	0.000	0.946	0.052	0.001	0.000	0.515	0.481	0.004
AIC		0.000	0.576	0.375	0.049	0.000	0.094	0.794	0.112
BIC		0.000	0.945	0.052	0.002	0.000	0.728	0.269	0.004

		$N = 2000$			
		$M = 1$	$M = 2$	$M = 3$	$M \geq 4$
SHT	$\alpha = .10$	0.000	0.000	0.904	0.096
	$\alpha = .05$	0.000	0.000	0.954	0.046
	$\alpha = .01$	0.000	0.000	0.990	0.010
AIC		0.000	0.000	0.846	0.154
BIC		0.000	0.002	0.992	0.006

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 Frequencies of the Number of Components: Trivariate Normal with $M = 2$

		$N = 50$			$N = 200$			$N = 1000$		
		$M = 1$	$M = 2$	$M \geq 3$	$M = 1$	$M = 2$	$M \geq 3$	$M = 1$	$M = 2$	$M \geq 3$
SHT	$\alpha = .10$	0.386	0.545	0.069	0.002	0.894	0.104	0.000	0.888	0.112
	$\alpha = .05$	0.528	0.441	0.031	0.005	0.940	0.055	0.000	0.940	0.060
	$\alpha = .01$	0.781	0.215	0.004	0.030	0.958	0.012	0.000	0.986	0.014
AIC		0.308	0.599	0.093	0.001	0.857	0.142	0.000	0.850	0.150
BIC		0.779	0.213	0.008	0.134	0.860	0.006	0.000	0.999	0.001

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

Table 4: Selection Frequencies of the Number of Components: Binomial with $M = 2$

		$N = 50$			$N = 200$			$N = 1000$		
		$M = 1$	$M = 2$	$M \geq 3$	$M = 1$	$M = 2$	$M \geq 3$	$M = 1$	$M = 2$	$M \geq 3$
SHT	$\alpha = .10$	0.749	0.189	0.062	0.242	0.686	0.072	0.000	0.910	0.090
	$\alpha = .05$	0.875	0.096	0.029	0.408	0.563	0.028	0.000	0.959	0.041
	$\alpha = .01$	0.976	0.019	0.005	0.749	0.247	0.004	0.001	0.991	0.008
AIC		0.628	0.275	0.093	0.001	0.857	0.142	0.000	0.850	0.150
BIC		0.779	0.213	0.008	0.134	0.860	0.006	0.000	0.999	0.001

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

Table 5: Selection Frequencies of the Number of Components: Binomial with $M = 3$

		$N = 100$				$N = 400$			
		$M = 1$	$M = 2$	$M = 3$	$M \geq 4$	$M = 1$	$M = 2$	$M = 3$	$M \geq 4$
SHT	$\alpha = .10$	0.000	0.692	0.281	0.027	0.000	0.195	0.733	0.072
	$\alpha = .05$	0.000	0.787	0.204	0.009	0.000	0.292	0.675	0.033
	$\alpha = .01$	0.000	0.920	0.079	0.001	0.000	0.515	0.481	0.004
AIC		0.000	0.618	0.329	0.054	0.000	0.145	0.731	0.125
BIC		0.000	0.864	0.132	0.004	0.000	0.516	0.478	0.007

		$N = 2000$			
		$M = 1$	$M = 2$	$M = 3$	$M \geq 4$
SHT	$\alpha = .10$	0.000	0.000	0.906	0.094
	$\alpha = .05$	0.000	0.000	0.954	0.046
	$\alpha = .01$	0.000	0.001	0.991	0.008
AIC		0.000	0.000	0.849	0.151
BIC		0.000	0.002	0.994	0.004

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 6: Selection Frequencies of the Number of Components: Binomial with $M = 4$

		$N = 200$					$N = 800$				
		$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M \geq 5$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M \geq 5$
SHT	$\alpha = .10$	0.000	0.006	0.663	0.302	0.030	0.000	0.000	0.192	0.746	0.062
	$\alpha = .05$	0.000	0.015	0.741	0.234	0.011	0.000	0.000	0.282	0.693	0.025
	$\alpha = .01$	0.000	0.050	0.838	0.111	0.001	0.000	0.000	0.492	0.504	0.004
AIC		0.000	0.003	0.600	0.344	0.053	0.000	0.000	0.139	0.752	0.109
BIC		0.000	0.033	0.822	0.143	0.003	0.000	0.000	0.526	0.471	0.004

		$N = 4000$				
		$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M \geq 5$
SHT	$\alpha = .10$	0.000	0.000	0.000	0.909	0.091
	$\alpha = .05$	0.000	0.000	0.000	0.958	0.043
	$\alpha = .01$	0.000	0.000	0.000	0.992	0.008
AIC		0.000	0.000	0.000	0.849	0.151
BIC		0.000	0.000	0.002	0.995	0.003

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 $K = 8$.

Table 7: Real Data Example: p-value of the SHT and the Value of the AIC/BIC Criterion Function

	$M = 1$	$M = 2$	$M = 3$	$M \geq 4$	No. of Components
SHT, p-value	0.000	0.000	0.106	–	3
AIC, S(r)	4.737	1.630	0.048	0.000	4
BIC, S(r)	3.537	1.010	-0.161	0.000	3

Notes: Based on $Z_1 = \sum_{j=1}^{4} X_j/4$ and $Z_2 = \sum_{j=5}^{8} X_j/4$.

Table 8: Real Data Example: Selection Frequency of the Number of Components

	$M = 1$	$M = 2$	$M = 3$	$M \geq 4$
SHT, $\alpha = .10$	0.000	0.000	0.657	0.342
SHT, $\alpha = .05$	0.000	0.029	0.714	0.257
SHT, $\alpha = .01$	0.000	0.171	0.686	0.143
AIC	0.000	0.000	0.600	0.400
BIC	0.000	0.171	0.629	0.200

Notes: The reported numbers are the frequencies across ${}_8C_4/2 = 35$ results.

Table 9: Real Data Example with Binomial Transformation: p-value of the SHT and the value of the AIC/BIC Criterion Function

	$M = 1$	$M = 2$	$M = 3$	$M \geq 4$	No. of Components
SHT, p-value	0.000	0.001	0.377	0.289	3
AIC, S(r)	3.508	0.017	-3e-5	-4e-7	3
BIC, S(r)	3.187	0.012	-1e-4	0.000	3

Notes: Based on binomially distributed variable $Y = \sum_{j=1}^{8} 1\{|X_j| \leq 5^o\}$.