Nonparametric Identification and Estimation of the Number of Components in Multivariate Mixtures

Hiroyuki Kasahara
Department of Economics
University of British Columbia
hkasahar@mail.ubc.ca

Katsumi Shimotsu*
Faculty of Economics
University of Tokyo
shimotsu@e.u-tokyo.ac.jp

January 10, 2013

Abstract

This article analyzes the identifiability of the number of components in k-variate, M-component finite mixture models in which each component distribution has independent marginals, including models in latent class analysis. Without making parametric assumptions on the component distributions, we investigate how one can identify the number of components from the distribution function of the observed data. When $k \geq 2$, a lower bound on the number of components (M) is nonparametrically identifiable from the rank of a matrix constructed from the distribution function of the observed variables. Building on this identification condition, we develop a procedure to consistently estimate a lower bound on the number of components.

Keywords: finite mixture; latent class analysis; nonnegative rank; rank estimation

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. Comprehensive theoretical accounts 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 number of components, the component distributions, and the mixing proportions. As emphasized in Hettmansperger and Thomas (2000), there is often little theoretical guidance for selecting

*Address for correspondence: Katsumi Shimotsu, Faculty of Economics, University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.

the number of components and/or the form of the component distributions despite their key role in the specification of mixtures. Furthermore, it has been known that the estimates of the number of components are sensitive to the choice of the component distributions (see, for example, Schork et al. (1990) and Roeder (1994)), and that imposing incorrect parametric restrictions on the component distributions may lead to erroneous inference on the number of components (Cruz-Medina et al. 2004).

This article analyzes the nonparametric identifiability of the number of components in k-variate, $\tilde{M}$-component finite mixture models of $W = (W_1, \ldots, W_k)'$ under the assumption that the W_j's are independently (but not necessarily identically) distributed within each component:

$$F(w) = F(w_1, \ldots, w_k) = \sum_{m=1}^{\tilde{M}} \pi^m F_1^m(w_1) F_2^m(w_2) \cdots F_k^m(w_k), \quad \pi^m > 0, \quad \sum_{m=1}^{\tilde{M}} \pi^m = 1. \quad (1)$$

Here, $F(w)$ is the distribution function of W, π^m is the mixture proportion of the m-th subpopulation, and $F_j^m(w_j)$ is the distribution function of W_j conditional on being from the m-th subpopulation, respectively. The number of components in $F(w)$, M, is defined as the smallest positive integer $\tilde{M}$ for which a finite mixture representation (1) can be found.

We analyze how one can recover the number of components M from the exact knowledge of the distribution function of observed variables $F(w_1, \ldots, w_k)$ when no parametric assumptions are imposed on the component distributions. Nonparametric identifiability and estimation of finite mixtures has recently attracted increasing attention. Hall and Zhou (2003), Hall et al. (2005), and Allman et al. (2009) analyze nonparametric identifiability of component distributions and mixing proportions in model (1) under known M.[1] In particular, Allman et al. (2009) show that if $k \geq 3$, model (1) is nonparametrically identifiable for any M if the $F_j^m(w_j)$'s are linearly independent. Hettmansperger and Thomas (2000) and Cruz-Medina et al. (2004) analyze the nonparametric identification and estimation of model (1) with iid marginals by partitioning the support of W_j into bins and transforming the data to multinomial vectors. Benaglia et al. (2009) and Levine et al. (2011) develop algorithms for estimating model (1) nonparametrically using kernels. However, no theoretical results on the identification of the number of components in model (1) are provided in the existing literature.

We show that a lower bound on the number of components M is identified without imposing any parametric assumptions if $k \geq 2$. Interestingly, this result holds despite the fact that the component distributions are not identifiable when $k = 2$ (see Clogg 1981; Hall and Zhou 2003). The lower bound is stated in terms of the rank of a matrix constructed from the (multinomial) distribution function of the observed data, where for continuous variables,

[1] Kasahara and Shimotsu (2009) study nonparametric identification of finite mixture dynamic discrete choice models widely used in econometrics.

we transform each element of W to a discrete random variable by partitioning its support as in Elmore et al. (2004). We also illustrate the cases in which the bound is tight except possibly for a set of mixture models with zero Lebesgue measure, and therefore, the bound is tight generically in the sense of Allman et al. (2009, p. 3106). By estimating the rank of its empirical analogue, we develop a procedure to consistently estimate a lower bound on the number of components. Simulations illustrate that our procedure performs well.

The mixture model (1) assumes that the marginal distributions are independent conditional on belonging to a subpopulation. The conditional independence assumption may be viewed as a version of a standard repeated measures random effects model, in which multivariate observations on an individual are often assumed to be independent conditional on the identity of the individual. The model (1) has important applications as demonstrated in some recent works on nonparametric mixture models as well as those on multinomial mixtures (e.g., Zhou et al. 2005; Dunson and Xing 2011; Bhattacharya and Dunson 2011), and encompasses models in latent class analysis that has been widely used in many fields including sociology, psychology, and biostatistics (Lazarsfeld and Henry 1968; Clogg 1995; Hagenaars and McCutcheon 2002; Magidson and Vermunt 2004; Skrondal and Rabe-Hesketh 2004). Once an estimate of a lower bound of M is obtained, one can use algorithms such as Benaglia et al. (2009) and Levine et al. (2011) to nonparametrically estimate the mixture model (1), provided that the mixing proportions and the component distributions are identifiable.

Numerous methods to select the number of components have been proposed in a parametric setting (see, for example, Henna 1985; Leroux 1992; Lindsay and Roeder 1992; Windham and Cutler 1992; Roeder 1994; Chen and Kalbfleisch 1996; Dacunha-Castelle and Gassiat 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. Furthermore, our selection procedure is based on a statistic whose asymptotic distribution is chi-squared or can be easily simulated, and it does not require the estimation of a mixture model with a different number of components.

The remainder of the article is organized as follows. Section 2 discusses the nonparametric identifiability of a lower bound on the number of components under $k \geq 2$. Section 3 introduces a procedure to test a lower bound on the number of mixture components. Section 4 reports simulation results, and empirical examples are provided in section 5. The supplementary appendix contains the proofs, mathematical details, and detailed results from simulations and empirical examples.

2 Nonparametric identification of a lower bound on the number of components

2.1 Two-variable case

We first analyze the nonparametric identification of a *lower bound* on the number of components for the mixture model (1) with $k = 2$. For notational clarity, we use X and Y in place of W_1 and W_2. Specifically, consider the following finite mixture models of variable (X, Y):

$$F(x, y) = \sum_{m=1}^{\tilde{M}} \pi^m F_x^m(x) F_y^m(y), \quad \pi^m > 0, \quad \sum_{m=1}^{\tilde{M}} \pi^m = 1, \tag{2}$$

where $F_x^m(x)$ and $F_y^m(y)$ are the distribution functions of X and Y conditional on being from the m-th subpopulation. No assumptions are imposed on $F_x^m(x)$'s and $F_y^m(y)$'s except that they are distribution functions. Define the number of components in $F(x, y)$, M, as the smallest positive integer $\tilde{M}$ for which a finite mixture representation (2) can be found.

We proceed to construct a partition, Δ, of the support of (X, Y), and form a matrix that represents the distribution of (X, Y) over Δ. Let $\mathcal{X}$ and $\mathcal{Y}$ denote the support of X and Y. Partition $\mathcal{X}$ and $\mathcal{Y}$ into $|\Delta_x|$ and $|\Delta_y|$ mutually exclusive and exhaustive subsets, respectively, as $\Delta_x = \{\delta_1^x, \ldots, \delta_{|\Delta_x|}^x\}$ and $\Delta_y = \{\delta_1^y, \ldots, \delta_{|\Delta_y|}^y\}$, where $|\mathcal{S}|$ denotes the number of elements in a set $\mathcal{S}$. Define $\Delta = \Delta_x \times \Delta_y$. Given a partition Δ, collect the distributions of X and Y conditional on being from the m-th subpopulation into a vector as

$$p_x^m = (\Pr(X \in \delta_1^x | m), \ldots, \Pr(X \in \delta_{|\Delta_x|}^x | m))' \quad \text{and} \quad p_y^m = (\Pr(Y \in \delta_1^y | m), \ldots, \Pr(Y \in \delta_{|\Delta_y|}^y | m))', \tag{3}$$

respectively. The vectors p_x^m and p_y^m implicitly depend on Δ_x and Δ_y.

Arrange $\Pr(X \in \delta_a^x, Y \in \delta_b^y)$ for partition level $(a, b) = (1, 1), \ldots, (|\Delta_x|, |\Delta_y|)$ into a $|\Delta_x| \times |\Delta_y|$ bivariate probability matrix as

$$P_\Delta = \begin{bmatrix} \Pr(X \in \delta_1^x, Y \in \delta_1^y) & \cdots & \Pr(X \in \delta_1^x, Y \in \delta_{|\Delta_y|}^y) \\ \vdots & \ddots & \vdots \\ \Pr(X \in \delta_{|\Delta_x|}^x, Y \in \delta_1^y) & \cdots & \Pr(X \in \delta_{|\Delta_x|}^x, Y \in \delta_{|\Delta_y|}^y) \end{bmatrix}. \tag{4}$$

Then, P_Δ represents the distribution of (X, Y) on the partition Δ and can be expressed in terms of π^m's, p_x^m's, and p_y^m's as

$$P_\Delta = \sum_{m=1}^{\tilde{M}} \pi^m p_x^m (p_y^m)', \quad \pi^m > 0, \quad \sum_{m=1}^{\tilde{M}} \pi^m = 1. \tag{5}$$

Equation (5) is a finite mixture model (2) that is restricted to the partition Δ.

For a partition Δ, define *the number of components in* P_Δ as the smallest integer $\tilde{M}$ such that the finite mixture representation (5) is possible. The number of components in P_Δ is closely related to the concept of *nonnegative rank* developed by Cohen and Rothblum (1993). For a nonnegative matrix A, its nonnegative rank is denoted by $\text{rank}_+(A)$ and defined as the smallest number of nonnegative rank-one matrices such that A equals their sum. Because P_Δ is a nonnegative matrix and the right-hand side of equation (5) is the sum of nonnegative rank-one matrices, by definition, the number of components in P_Δ is the nonnegative rank of P_Δ.

The following proposition, originally from Cohen and Rothblum (1993), states the properties of the nonnegative rank of P_Δ and its relation to the rank of P_Δ. Proposition 1(a),(b), and (c) correspond to Lemma 2.3, Theorem 4.1, and Corollary 4.2 of Cohen and Rothblum (1993), respectively.

Proposition 1 (Cohen and Rothblum, 1993) *(a)* $\textit{rank}(P_\Delta) \leq \textit{rank}_+(P_\Delta) \leq \min\{|\Delta_x|, |\Delta_y|\}$. *(b) If* $\textit{rank}(P_\Delta) \leq 2$, *then* $\textit{rank}(P_\Delta) = \textit{rank}_+(P_\Delta)$. *(c) If* $|\Delta_x| \leq 3$ *or* $|\Delta_y| \leq 3$, *then* $\textit{rank}_+(P_\Delta) = \textit{rank}(P_\Delta)$.

From Proposition 1(a), $\text{rank}(P_\Delta)$ gives a lower bound on the number of components in P_Δ whereas the number of support points of X and Y gives an upper bound on the number of identifiable components since $|\Delta_x| \leq |\mathcal{X}|$ and $|\Delta_y| \leq |\mathcal{Y}|$. It follows from Proposition 1 that $\text{rank}_+(P_\Delta) = \text{rank}(P_\Delta)$ if $\text{rank}_+(P_\Delta) \leq 3$.

The number of components in P_Δ is identified with the nonnegative rank of P_Δ. Determining the nonnegative rank of a matrix is computationally difficult[2], however, and is still a subject of ongoing research (see, for example, Dong, Lin, and Chu 2009). Therefore, it is useful to characterize a *lower bound* on the number of components in P_Δ in terms of the rank of P_Δ.

An obvious limitation of the lower bound based on the rank of P_Δ is a possible discrepancy between the lower bound and the actual number of components. This is because the latter requires that the components π^m's, p_x^m's, and p_y^m's in (5) to be nonnegative while the former does not.[3] We investigate the size of a set of mixture models wherein $\text{rank}_+(P_\Delta) > \text{rank}(P_\Delta)$. Given a positive integer M_0, define the space of M_0-component mixture models $\theta = \{p_x^m, p_y^m, \pi^m\}_{m=1}^{M_0}$ by $\Theta \subset (\mathcal{S}_{|\Delta_x|-1})^{M_0} \times (\mathcal{S}_{|\Delta_y|-1})^{M_0} \times \mathcal{S}_{M_0-1}$ as in Allman et al. (2009, p. 3107), where $\mathcal{S}_k$ denotes the standard k-simplex. The following proposition shows that if we randomly draw a mixture model θ from Θ and construct a matrix $P(\theta) = \sum_{m=1}^{M_0} \pi^m p_x^m (p_y^m)'$, then we have $\text{rank}(P(\theta)) = M_0$ with probability one. This

[2]Vavasis (2009) shows that determining the nonnegative rank of a matrix is NP-hard.

[3]For example, suppose that $|\Delta_x| = |\Delta_y| = 4$ and $P_\Delta = \sum_{m=1}^{4} \pi^m p_x^m (p_y^m)'$, where $\pi^m > 0$ and p_x^ms are linearly independent but $p_y^1 + p_y^2 - p_y^3 - p_y^4 = 0$, so that $\text{rank}(P_\Delta)$ is at most 3. Writing one p_y^m in terms of the other p_y^ms and substituting into P_Δ will give a three-term mixture representation of P_Δ. However, if $-\pi^1 p_x^1 + \pi^2 p_x^2$ and $-\pi^3 p_x^3 + \pi^4 p_x^4$ have both positive and negative elements, then the resulting three-term mixture representation necessarily contains negative components, and $\text{rank}_+(P_\Delta)$ is strictly larger than 3.

result holds because, when $\text{rank}(P(\theta)) < M_0$, either the vectors $\{p_x^m\}_{m=1}^{M_0}$ or $\{p_y^m\}_{m=1}^{M_0}$ are linearly dependent, but the set of linearly dependent $\{p_x^m\}_{m=1}^{M_0}$'s (or $\{p_y^m\}_{m=1}^{M_0}$'s) has Lebesgue measure zero in $\mathbb{R}^{\Delta_x \times M_0}$ (or $\mathbb{R}^{\Delta_y \times M_0}$).

Proposition 2 *If $M_0 \leq \min\{|\Delta_x|, |\Delta_y|\}$, then $M_0 = rank_+(P(\theta)) = rank(P(\theta))$ holds for all the points in Θ except possibly for a set of Lebesgue measure zero.*

When X and Y are discrete, taking the support of (X, Y) as Δ and applying Proposition 2 gives that $\text{rank}(P_\Delta) = \text{rank}_+(P_\Delta) = M$ with probability one if we draw an M-component bivariate mixture model with conditionally independent marginals. Hence, the bound is tight with probability one. When X and Y are continuous, there is no obvious choice of a single partition Δ. The nonnegative rank of P_Δ could be strictly smaller than M when a single partition Δ does not fully reveal the information for identifying the number of components in $F(x, y)$. A tighter lower bound of M may be obtained by taking the maximum value of the rank of P_Δs across different partitions.

Proposition 3 *Suppose that in model (2), the distribution of (X, Y) is continuous. (a) If $\{F_x^m(\cdot)\}_{m=1}^M$ are linearly independent and $\{F_y^m(\cdot)\}_{m=1}^M$ are linearly independent, then there exists a partition Δ with $|\Delta_x| = |\Delta_y| = M$ such that $rank(P_\Delta) = M$. (b) For any fixed Δ, the probability that $rank_+(P_\Delta) = rank(P_\Delta)$ is one when a mixture model is drawn randomly.*

Proposition 3(a) gives a sufficient condition under which the rank of P_Δ is equal to M for some choice of Δ; in this case, M is identified with the maximum value of $\text{rank}(P_\Delta)$'s over all possible partitions of $\mathcal{X} \times \mathcal{Y}$. Proposition 3(b) implies that whether $\text{rank}(P_\Delta) = M$ holds depends on whether $\text{rank}_+(P_\Delta) = M$ holds.

2.2 General k-variable case

We now illustrate how our approach in Section 2.1 can be applied to the mixture model (1) with $k \geq 3$ to obtain a lower bound on M. Consider a hyperrectangle partition $\Delta = \Delta_1 \times \cdots \times \Delta_k$ of $\mathcal{W}$. Let P_Δ denote $F(w)$ on Δ. Note that P_Δ is a k-dimensional array. P_Δ is written as a weighted average of k-dimensional tensors as follows:

$$P_\Delta = \sum_{m=1}^{\text{rank}_+(P_\Delta)} \pi^m P^m, \quad P^m = \otimes_{j=1}^k p_j^m, \quad \pi^m > 0, \quad \sum_{m=1}^{\text{rank}_+(P_\Delta)} \pi^m = 1, \tag{6}$$

where $\otimes$ denotes a tensor product and p_j^m is a $|\Delta_j| \times 1$ vector. Here, $\text{rank}_+(P_\Delta)$ is defined as the smallest positive integer for which a representation (6) holds and called the nonnegative (tensor) rank of P_Δ (see, for example, Lim and Common 2009). As in the two-variable case, $\text{rank}_+(P_\Delta)$ provides a lower bound on M.

We construct a matrix from P_Δ by grouping the variables in $W = (W_1, \ldots, W_k)'$ into two groups. We index the groupings by α. For the grouping α, let X^α and Y^α be the grouped variables. Let $\Delta_{x^\alpha} = \{\delta_1^{x^\alpha}, \ldots, \delta_{|\Delta_{x^\alpha}|}^{x^\alpha}\} = \prod_{j \in S_x(\alpha)} \Delta_j$ be the partition of the support of X^α, where $S_x(\alpha)$ is the set of indices such that $W_j \in X^\alpha$, and define Δ_{y^α} similarly. Then, we construct a $|\Delta_{x^\alpha}| \times |\Delta_{y^\alpha}|$ bivariate probability matrix P_Δ^α by arranging $\Pr(X^\alpha \in \delta_a^{x^\alpha}, Y^\alpha \in \delta_b^{y^\alpha})$ for partition level $(a, b) = (1, 1), \ldots, (|\Delta_{x^\alpha}|, |\Delta_{y^\alpha}|)$ as in (4).

A lower bound on M can be obtained in terms of $\text{rank}_+(P_\Delta)$, $\text{rank}_+(P_\Delta^\alpha)$, and $\text{rank}(P_\Delta^\alpha)$ as $M \geq \text{rank}_+(P_\Delta) \geq \text{rank}_+(P_\Delta^\alpha) \geq \text{rank}(P_\Delta^\alpha)$. Taking the maximum value of $\text{rank}_+(P_\Delta)$, $\text{rank}_+(P_\Delta^\alpha)$, and $\text{rank}(P_\Delta^\alpha)$ across different partitions, Δ's, and different groupings, α's, gives tighter lower bounds. Such bounds may still be, however, strictly smaller than M.

We investigate when $\text{rank}_+(P_\Delta) = \text{rank}(P_\Delta^\alpha)$ holds. With a slight abuse of notation, given a positive integer M_0, define the space of M_0-component mixture models $\theta = \{p_1^m, \ldots, p_k^m, \pi^m\}_{m=1}^{M_0}$ as $\Theta \subset (\mathcal{S}_{|\Delta_1|-1})^{M_0} \times \cdots \times (\mathcal{S}_{|\Delta_k|-1})^{M_0} \times \mathcal{S}_{M_0-1}$. Given an element θ of Θ, group W into X^α and Y^α and construct a bivariate probability matrix $P^\alpha(\theta) = \sum_{m=1}^{M_0} \pi^m p_{x^\alpha}^m (p_{y^\alpha}^m)'$, where $p_{x^\alpha}^m$ is $|\Delta_{x^\alpha}| \times 1$ and $p_{y^\alpha}^m$ is $|\Delta_{y^\alpha}| \times 1$. The following proposition shows that if the grouped variables have sufficiently large state spaces relative to M_0, an analogous result to Proposition 2 holds for a k-variable model.

Proposition 4 *If $M_0 \leq \min\{|\Delta_{x^\alpha}|, |\Delta_{y^\alpha}|\}$, then $M_0 = \text{rank}(P^\alpha(\theta))$ holds for all the points in Θ except possibly for a set of Lebesgue measure zero.*

In the continuous variable case, we have a simple corollary of Proposition 3(a).

Corollary 1 *Suppose that in model (1), the distribution of W is continuous. If there exists a grouping α such that the two families $\{F_{x^\alpha}^m(\cdot)\}_{m=1}^M$ and $\{F_{y^\alpha}^m(\cdot)\}_{m=1}^M$ are respectively composed of linearly independent vectors, then $\text{rank}(P_\Delta^\alpha) = M$.*

2.3 Relation to latent class analysis

Consider a special case in which an observation vector $W = (W_1, \ldots, W_k)'$ consists of k dichotomous or polytomous responses, which are typically answers to questions or results of diagnoses. In this case, our model (1) becomes identical to the model used in *latent class analysis*. For recent surveys and applications of latent class analysis, see the references in the introduction.

The latent class analysis with $k = 2$ (two-way contingency table) is also known as *latent budget analysis* (Goodman 1974; Clogg 1981; de Leeuw and van der Heijden 1988). Testing the number of components in a latent budget model is particularly difficult because the parameters of the model are not identified unless some restrictions are imposed. Using our result, it is possible to identify a lower bound on M without imposing restrictions on the parameters, even though identifying a lower bound of M does not solve the problem of parameter non-identification.

3 Estimating a lower bound on the number of components

Proposition 1 in Section 2 shows that the rank of a matrix P_Δ in (5) gives a lower bound on the number of mixture components. In this section, we develop procedures to estimate the rank of P_Δ for a given partition Δ and extend these procedures to the case when there are more than two variables.

3.1 Statistic by Kleibergen and Paap (2006)

Kleibergen and Paap (2006) develop a procedure to test the null hypothesis that the rank of P_Δ is equal to r as described below. For notational brevity, let $s = |\Delta_x|$ and $t = |\Delta_y|$, and write the singular value decomposition of an $s \times t$ matrix P_Δ as

$$P_\Delta = USV' = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix} \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix} \begin{pmatrix} V_{11}' & V_{12}' \\ V_{21}' & V_{22}' \end{pmatrix},$$

where U is an $s \times s$ orthogonal matrix, V is a $t \times t$ orthogonal matrix, and S is an $s \times t$ matrix that contains the singular values of P_Δ in decreasing order on its main diagonal and is equal to zero elsewhere. In the partition of U, S, and V on the right-hand side, U_{11}, S_1, and V_{11} are $r \times r$, and the dimensions of the other submatrices are defined conformably. Then, the null hypothesis $H_0 : \text{rank}(P_\Delta) = r$ is equivalent to $H_0 : S_2 = 0$ because the rank of a matrix is equal to the number of non-zero singular values.

The statistic by Kleibergen and Paap is based on an orthogonal transformation of S_2 given by $\Lambda_r = A_{r,\perp}' P_\Delta B_{r,\perp}'$, where $A_{r,\perp}' = (U_{22}U_{22}')^{1/2}(U_{22}')^{-1}[U_{12}'\, U_{22}']$ and $B_{r,\perp} = (V_{22}V_{22}')^{1/2}(V_{22}')^{-1}[V_{12}'\, V_{22}']$. Unlike S_2, Λ_r is not restricted to be non-negative. Then, the null hypothesis $H_0 : rank(P_\Delta) = r$ is equivalent to $H_0 : \Lambda_r = 0$. Let $\hat{P}_\Delta$ be an estimator of the matrix P_Δ with sample size N. We assume that $\text{vec}(\hat{P}_\Delta)$ is asymptotically normally distributed.

Assumption 1 *$\sqrt{N}vec(\hat{P}_\Delta - P_\Delta) \to_d N(0, \Sigma)$ as $N \to \infty$, where Σ is an $st \times st$ covariance matrix.*

When the distribution of W is discrete or Δ is predetermined, $\text{vec}(\hat{P}_\Delta)$ follows a multinomial distribution, and a formula for Σ is easily available. If W has a continuous distribution and the empirical quantiles of the W_j's are used to construct Δ, then $\text{vec}(\hat{P}_\Delta)$ follows the empirical multivariate quantile-partitioned (EMQP) distribution (Borkowf, 2000) described in the supplemental appendix, and one can use bootstrap to estimate Σ.

We estimate Λ_r by $\hat{\Lambda}_r = \hat{A}_{r,\perp}' \hat{P}_\Delta \hat{B}_{r,\perp}'$ and test $H_0 : \Lambda_r = 0$, where $\hat{A}_{r,\perp}$ and $\hat{B}_{r,\perp}$ are the estimators of $A_{r,\perp}$ and $B_{r,\perp}$ obtained from the singular value decomposition of $\hat{P}_\Delta$. Kleibergen and Paap (2006) derive the asymptotic distribution of $\hat{\lambda}_r = \text{vec}(\hat{\Lambda}_r)$, as summarized below.

Proposition 5 *(Kleibergen and Paap, 2006, Theorem 1) Suppose that Assumption 1 holds and that* $\Omega_r = (B_{r,\perp} \otimes A'_{r,\perp})\Sigma(B_{r,\perp} \otimes A'_{r,\perp})'$ *is nonsingular. If rank*$(P_\Delta) \leq r$*, then* $\sqrt{N}\hat{\lambda}_r \to_d N(0, \Omega_r)$ *as* $N \to \infty$*.*

Kleibergen and Paap (2006, Corollary 1) propose the statistic called the *rk statistic*:

$$\mathrm{rk}(r) = N\hat{\lambda}_r'\hat{\Omega}_r^{-1}\hat{\lambda}_r, \tag{7}$$

where $\hat{\Omega}_r$ is a consistent estimator for Ω_r. If the assumptions of Proposition 5 hold, then $\mathrm{rk}(r)$ converges in distribution to a $\chi^2((s-r)(t-r))$ random variable under $H_0 : \mathrm{rank}(P_\Delta) = r$. The nonsingularity assumption on Ω_r can be relaxed by using the Moore-Penrose (M-P) pseudoinverse as discussed in Section 3.4.

The choice of Δ is left to the researcher. As for the number of partitions, it is desirable to use a partition that is as fine as possible from the perspective of pure identification, but using a finer partition increases the variance of $\hat{P}_\Delta$. In practice, we suggest setting the number of partitions equal to one plus the maximum number of components we want to allow for in modeling the data. As for the choice of partitions, a natural choice would be to use equiprobable intervals as in Pearson's chi-squared test, but there may be cases where using a non-equiprobable partition gives a stronger power because mixture models often have fat tails. The optimal choice of partitions remains an open question.

3.2 Sequential hypothesis testing

Denote the population rank of P_Δ by r_0. To estimate r_0, we sequentially test $H_0 : \mathrm{rank}(P_\Delta) = r$ against $H_1 : \mathrm{rank}(P_\Delta) > r$ starting from $r = 0$, and then $r = 1, \ldots, t^*$, where $t^* = \min\{s, t\}$. The first value for r that leads to a nonrejection of H_0 gives our estimate for r_0.

For $r = 0, \ldots, t^*$, let $c^r_{1-\alpha_N}$ denote the $100(1-\alpha_N)$ percentile of the cumulative distribution function of a $\chi^2\left((s-r)(t-r)\right)$ random variable. Then, our estimator based on sequential hypothesis testing (SHT, hereafter) is defined as

$$\hat{r} = \min_{r \in \{0,\ldots,t^*\}} \{r : \mathrm{rk}(i) \geq c^i_{1-\alpha_N}, i = 0, \ldots, r-1, \mathrm{rk}(r) < c^r_{1-\alpha_N}\}. \tag{8}$$

The estimator $\hat{r}$ depends on the choice of the significance level α_N. As shown by Robin and Smith (2000, Theorem 5.2), $\hat{r}$ converges to r_0 in probability as $N \to \infty$ if we choose α_N such that $\alpha_N = o(1)$ and $-N^{-1}\ln\alpha_N = o(1)$.

3.3 Information criteria

We also consider a selection procedure by information criteria to estimate r_0 consistently. Consider the criterion function $Q(r) = \mathrm{rk}(r) - f(N)g(r)$, where $g(r)$ is a (possibly stochastic)

penalty function. Define $\tilde{r} = \arg\min_{1\le r\le t^*} Q(r)$. Under a standard condition on $f(N)$ and $g(r)$, this gives a consistent estimate of r_0:

Proposition 6 *Suppose that the conditions of Proposition 5 hold, and $\hat{\Omega}_r$ converges to a nonsingular matrix for any $r \ge r_0$. 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, $\tilde{r} \to_p r_0$.*

For the choice of $f(N)$ and $g(r)$, we consider the penalty terms in the Akaike (AIC), Bayesian (BIC), and Hannan-Quinn (HQ) information criteria. We choose $g(r) = (s-r)(t-r)$ with $f(N) = 2$ for AIC, $f(N) = \log(N)$ for BIC, and $f(N) = 2\log(\log(N))$ for HQ. The BIC and HQ model selection procedures provide a consistent estimate of r_0 since their choice of $f(N)$ and $g(r)$ satisfies the conditions in Proposition 6. In contrast, AIC is not necessarily consistent and tends to overestimate r_0 with a large sample size.

3.4 Case of multiple variables

Suppose that $W = (W_1, \ldots, W_k)'$ with $k \ge 3$ follows the distribution function (1). As in Section 2.2, we group the variables in W into two groups X^α and Y^α, with the grouping index α, and let P^α_Δ denote a $|\Delta_{x^\alpha}| \times |\Delta_{y^\alpha}|$ bivariate probability matrix derived from the joint distribution of X^α and Y^α on a partition Δ. We test the null hypothesis that $\mathrm{rank}(P^\alpha_\Delta) \le r$ for all $\alpha \in \mathcal{A}_0$, where $\mathcal{A}_0$ is a set of the αs over which we construct test statistics.

We assume that all the variables in W are included in the first grouping $\{X^1, Y^1\}$. Then, for every $\alpha \in \mathcal{A}_0$, the elements of the probability matrix P^α_Δ can be expressed as a linear combination of the elements of P^1_Δ, and therefore, there exists a matrix Π^α such that $\mathrm{vec}(P^\alpha_\Delta) = \Pi^\alpha \mathrm{vec}(P^1_\Delta)$.

Define $A^\alpha_{r,\perp}$, $B^\alpha_{r,\perp}$, and λ^α_r analogously to $A_{r,\perp}$, $B_{r,\perp}$, and λ_r in Section 4.1 using P^α_Δ in place of P_Δ. Define $\hat{\lambda}^\alpha_r = \mathrm{vec}((\hat{A}^\alpha_{r,\perp})'\hat{P}^\alpha_\Delta(\hat{B}^\alpha_{r,\perp})') = (\hat{B}^\alpha_{r,\perp} \otimes (\hat{A}^\alpha_{r,\perp})')\Pi^\alpha \mathrm{vec}(\hat{P}^1_\Delta)$ using the estimators of P^1_Δ, $A^\alpha_{r,\perp}$, and $B^\alpha_{r,\perp}$. To test the null hypothesis that $\mathrm{rank}(P^\alpha_\Delta) \le r$ for all $\alpha \in \mathcal{A}_0$, we stack $\hat{\lambda}^\alpha_r$'s into a vector as $\hat{\lambda}_r(\mathcal{A}_0) = ((\hat{\lambda}^1_r)', \ldots, (\hat{\lambda}^{|\mathcal{A}_0|}_r)')'$ and test the null hypothesis $\lambda_r(\mathcal{A}_0) = 0$. Extending Proposition 5, the following corollary establishes the asymptotic normality of $\hat{\lambda}_r(\mathcal{A}_0)$. We omit its proof to save space, because it is a straightforward consequence of Slutsky's theorem.

Corollary 2 *Suppose that $\sqrt{N}vec(\hat{P}^1_\Delta - P^1_\Delta) \to_d N(0, \Sigma^1_\Delta)$ and that $\Omega_r(\mathcal{A}_0)$ defined in (9) below is nonsingular. If $rank(P^\alpha_\Delta) \le r$ for all $\alpha \in \mathcal{A}_0$, we have $\sqrt{N}\hat{\lambda}_r(\mathcal{A}_0) \to_d N(0, \Omega_r(\mathcal{A}_0))$ as $N \to \infty$, where*

$$\Omega_r(\mathcal{A}_0) = \begin{bmatrix} \Psi^1\Sigma^1_\Delta(\Psi^1)' & \cdots & \Psi^1\Sigma^1_\Delta(\Psi^{|\mathcal{A}_0|})' \\ \vdots & \ddots & \vdots \\ \Psi^{|\mathcal{A}_0|}\Sigma^1_\Delta(\Psi^1)' & \cdots & \Psi^{|\mathcal{A}_0|}\Sigma^1_\Delta(\Psi^{|\mathcal{A}_0|})' \end{bmatrix} \tag{9}$$

and $\Psi^\alpha = (B^\alpha_{r,\perp} \otimes (A^\alpha_{r,\perp})')\Pi^\alpha$.

We can test the null hypothesis $H_0 : \text{rank}(P^\alpha_\Delta) \leq r$ for all $\alpha \in \mathcal{A}_0$ by the *average rk statistic* defined as

$$\text{ave-rk}(r, \mathcal{A}_0) = N(\hat{\lambda}_r(\mathcal{A}_0))'(\hat{\Omega}_r(\mathcal{A}_0))^{-1}\hat{\lambda}_r(\mathcal{A}_0), \tag{10}$$

where $\hat{\Omega}_r(\mathcal{A}_0)$ is a consistent estimator of $\Omega_r(\mathcal{A}_0)$. Thus, ave-rk$(r, \mathcal{A}_0)$ combines information from $\hat{\lambda}^\alpha_r$'s across different α's using the inverse of their covariance matrix as the weight. Under the assumptions in Corollary 2, ave-rk$(r, \mathcal{A}_0)$ converges in distribution to a $\chi^2(\nu(\mathcal{A}_0))$ random variable, where $\nu(\mathcal{A}_0) \equiv \sum_{\alpha\in\mathcal{A}_0}(|\Delta_{x^\alpha}|-r)(|\Delta_{y^\alpha}|-r)$ is the number of elements in $\hat{\lambda}_r(\mathcal{A}_0)$. We note, however, that the average rk statistic may give a slack lower bound when enumerating sufficiently many of the groupings and partitions of the data is not computationally feasible.

When $\nu(\mathcal{A}_0)$ is larger than the rank of Σ^1_Δ, the covariance matrix $\Omega_r(\mathcal{A}_0)$ becomes singular and the assumption of Corollary 2 is violated. In such a case, if $\Pr(\text{rank}(\hat{\Omega}_r(\mathcal{A}_0)) = \text{rank}(\Omega_r(\mathcal{A}_0))) \to 1$, using the M-P pseudoinverse of $\hat{\Omega}_r(\mathcal{A}_0)$ in the ave-rk statistic (10) gives a test statistic whose asymptotic distribution is $\chi^2(\text{rank}(\Omega_r(\mathcal{A}_0)))$ (Andrews, 1987). However, in finite samples, if $\hat{\Omega}_r(\mathcal{A}_0)$ has a very small but nonzero eigenvalue, its pseudoinverse may take a very large value and behave erratically. To deal with the singularity of $\Omega_r(\mathcal{A}_0)$, we follow Lütkepohl and Burda (1997) to use a suitable reduced rank estimator in place of $\hat{\Omega}_r(\mathcal{A}_0)$. Given a small constant c, we apply a singular decomposition to $\hat{\Omega}_r(\mathcal{A}_0)$ and replace the eigenvalues smaller than c with zero. Let $\hat{\Omega}_{r,c}(\mathcal{A}_0)$ denote this low-rank approximation of $\hat{\Omega}_r(\mathcal{A}_0)$, and define the *modified* average rk statistic as

$$\text{ave-rk}^+(r, \mathcal{A}_0) = N(\hat{\lambda}_r(\mathcal{A}_0))'(\hat{\Omega}_{r,c}(\mathcal{A}_0))^+\hat{\lambda}_r(\mathcal{A}_0). \tag{11}$$

The asymptotic distribution of ave-rk$^+(r, \mathcal{A}_0)$ is $\chi^2(J_c)$, where J_c is the number of eigenvalues of $\Omega_r(\mathcal{A}_0)$ that are no smaller than c. The behavior of ave-rk$^+(r, \mathcal{A}_0)$ could be sensitive to the choice of c. In the simulations in Section 5, we set c equal to 0.01 times the largest eigenvalue of $\Omega_r(\mathcal{A}_0)$.[4]

We also consider an alternative statistic that is applicable even when $\nu(\mathcal{A}_0)$ is large. In the alternate statistic, we first choose K subsets of $\mathcal{A}_0$ as $\{\mathcal{A}_1, \ldots, \mathcal{A}_K\}$ so that $\mathcal{A}_0 = \bigcup_{j=1}^K A_j$, and construct the ave-rk$^+(r, \mathcal{A}_j)$ as in (11) but using $\mathcal{A}_j$ in place of $\mathcal{A}_0$. We then combine the information in ave-rk$^+(r, \mathcal{A}_j)$ for $j = 1, \ldots, K$ into the modified max-rk statistic defined as max-rk$^+(r) = \max_{j=1,\ldots,K}$ ave-rk$^+(r, \mathcal{A}_j)$. By choosing $\mathcal{A}_j$'s so that the degree of freedom $\nu(\mathcal{A}_j)$ is sufficiently small, max-rk$^+(r)$ would be less sensitive to the choice of c than ave-rk$^+(r, \mathcal{A}_0)$. We can apply the sequential hypothesis testing procedure to max-rk$^+(r)$. Its asymptotic null distribution is not chi-squared but it can be easily simulated using the relation $\sqrt{N}\hat{\lambda}^\alpha_r = \hat{\Psi}^\alpha\sqrt{N}(\text{vec}(\hat{P}^1_\Delta) - \text{vec}(P^1_\Delta))$.

[4] See Lütkepohl and Burda (1997) for other choices of c.

4 Simulation study

We conduct simulation experiments to assess the finite sample performance of our proposed procedures for selecting the number of components. The reported results are based on 1000 simulated samples from mixtures with $M = 3$ components with three different sample sizes: $N = 500$, 2000, and 8000. To construct the rk statistic (7) and the ave-rk$^+$ statistic (11) for each sample, we estimate Ω_r and $\Omega_r(\mathcal{A}_0)$ consistently by nonparametric bootstrap using 1000 random samples with replacement from empirical distributions.

In the first experiment, we generate samples of (X, Y) from a 3-component normal mixture $\sum_{m=1}^{3} \pi^m N_2(\mu^m, I_2)$ with $\mu^1 = (0, 0)'$, $\mu^2 = (1.0, 2.0)'$, $\mu^3 = (2.0, 1.0)'$, and $\pi^1 = \pi^2 = \pi^3 = 1/3$. We denote the $100 \times q$ percentile of empirical distributions of X and Y by z_q^x and z_q^y, respectively. We consider three different partitions of the form $\Delta^j = \{(-\infty, z_{q_1}^j], (z_{q_1}^j, z_{q_2}^j], \ldots, (z_{q_{t-2}}^j, z_{q_{t-2}}^j], (z_{q_{t-1}}^j, \infty)\}$ for $j = x, y$, where t is the number of partitions. We choose $t = 4$ with $(q_1, q_2, q_3) = (0.25, 0.5, 0.75)$ for Partition 1 and $(q_1, q_2, q_3) = (0.1, 0.5, 0.9)$ for Partition 2. Thus, the support of X or Y is partitioned into 4 (asymptotically) equiprobable subsets in Partition 1 while Partition 2 provides the finer partitions in the tail part of distributions than Partition 1. Partition 3 combines Partitions 1 and 2 as $(q_1, q_2, q_3, q_4, q_5) = (0.1, 0.25, 0.5, 0.75, 0.9)$.

Figures 1(a), (b), and (c) show the frequency that the SHT with $\alpha = 0.05$, AIC, BIC, and HQ based on the rk statistic correctly select $M = 3$ for Partitions 1, 2, and 3, respectively. Across different partitions, the performance of all the procedures improves as the sample size increases. With sample sizes 500 and 2000, the AIC outperforms other statistics, but overestimates the number of components at $N = 8000$ in Partitions 1 and 2. BIC exhibits the worst performance among all of the methods in this setup. HQ is a better choice than BIC but is outperformed by SHT in most cases. SHT in Partition 2 performs better than in Partition 1, even though the improvement is not substantial. This is probably because the tail part of distributions provides important information for separately identifying different components in this experiment. SHT in Partition 2 also performs better than SHT in Partition 3; using a larger number of partitions does not necessarily improve performance because $\text{var}(\hat{P}_\Delta)$ increases with the number of partitions.

Figures 1(d), (e), and (f) show the selection frequency across different partitions when we generate samples from 3-component chi-squared mixtures, where $\pi^m = 1/3$, $X^m \sim \chi^2(k_x^m)$, and $Y^m \sim \chi^2(k_y^m)$ with $(k_x^1, k_y^1) = (1, 1)$, $(k_x^2, k_y^2) = (3, 6)$, and $(k_x^3, k_y^3) = (6, 3)$. Figures 1(g), (h), and (i) report the selection frequency for 3-component gamma mixtures with component-specific shape parameters, where $\pi^m = 1/3$, $X^m \sim \text{Gamma}(k_1^m, 1)$, and $Y^m \sim \text{Gamma}(k_2^m, 1)$ with $(k_1^1, k_2^1) = (1, 1)$, $(k_1^2, k_2^2) = (1.5, 3)$, and $(k_1^3, k_2^3) = (3, 1.5)$. In both the chi-squared and gamma mixtures, the relative performance across SHT, AIC, BIC, and HQ and the relative performance of SHT across different partitions are qualitatively similar to the case of the normal mixture discussed above.

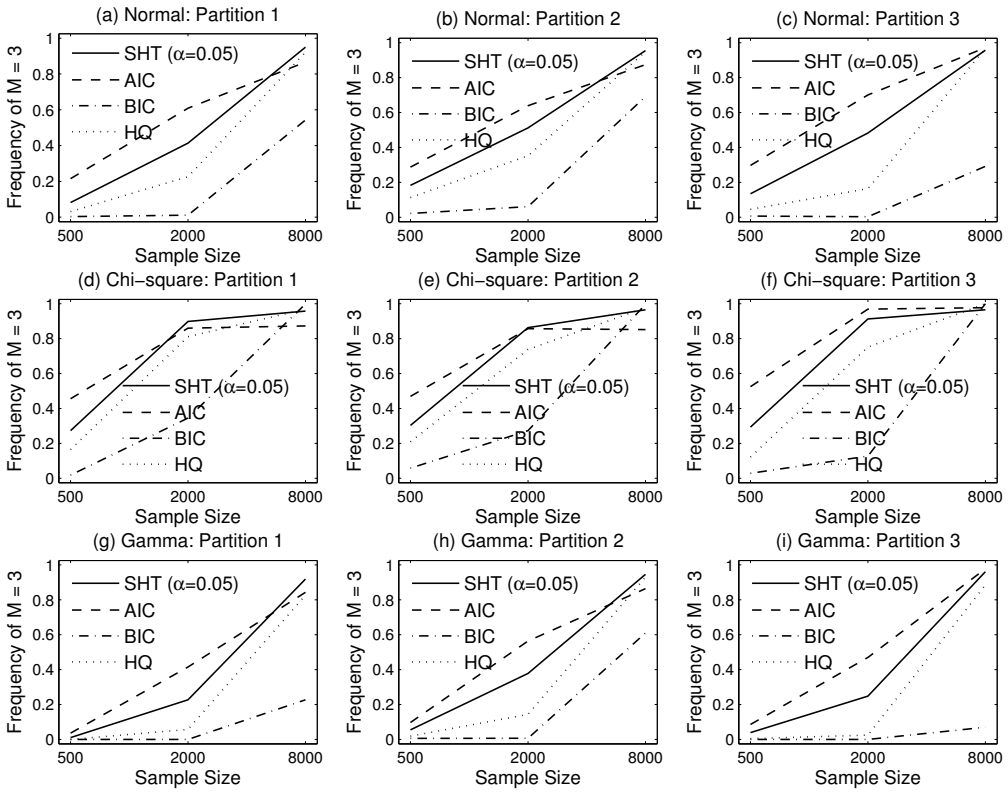

Figure 1: Selection Frequencies of the Number of Components: Two Variables

Next, we consider a 4-variable, 3-component normal mixture, where $W = (W_1, \ldots, W_4)'$ follows $\sum_{m=1}^{M} \pi^m N_4(\mu^m, I_4)$ with $\mu^1 = (0,0,0,0)'$, $\mu^2 = (1.0, 2.0, 0.5, 1.0)'$, $\mu^3 = (2.0, 1.0, 1.0, 0.5)'$, and $\pi^1 = \pi^2 = \pi^3 = 1/3$. Following the approach in Section 3.4, we consider three groupings: $\{X^1, Y^1\} = \{(W_1, W_2), (W_3, W_4)\}$, $\{X^2, Y^2\} = \{(W_1, W_3), (W_2, W_4)\}$, and $\{X^3, Y^3\} = \{(W_1, W_4), (W_2, W_3)\}$. We then estimate the probability matrix P_Δ^α for each $\alpha \in \{1, 2, 3\}$, and construct the ave-rk$^+$ statistic (11) by setting c equal to 0.01 times the largest eigenvalue of $\Omega_r(\mathcal{A}_0)$. The support of W_i is partitioned into 2 equiprobable subsets based on its empirical median, so that the dimension of P_Δ^α is 4×4. As an alternative method, we also consider the maximum likelihood estimator (MLE)-based parametric model selection procedure with AIC, BIC, and HQ, where each component distribution is correctly specified as a 4-dimensional normal distribution with unknown means and an unknown diagonal covariance matrix. Figure 2(a) reports the result. The MLE-based AIC substantially overestimates the number of components. While the MLE-based HQ outperforms the ave-rk$^+$-based SHT, their performances are comparable when $N \geq 2000$. This is encouraging given that our ave-rk$^+$-based methods do not use parametric restrictions of the normal mixture model.

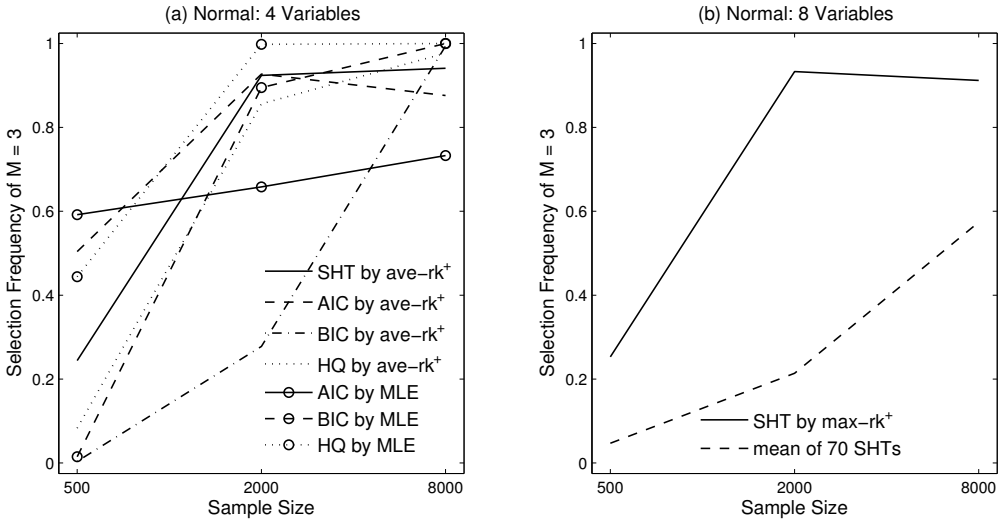

Figure 2: Selection Frequencies of the Number of Components: Four and Eight Variables

We also consider an 8-variable, 3-component normal mixture, where $W = (W_1, \ldots, W_8)'$ follows $\sum_{m=1}^{3} \pi^m N_8(\mu^m, I_8)$ with $\mu^1 = (0,0,0,0,0,0,0,0)'$, $\mu^2 = (1.0, 2.0, 0.5, 1.0, 0.75, 1.25, 0.25, 0.5)'$, $\mu^3 = (2.0, 1.0, 1.0, 0.5, 1.25, 0.75, 0.5, 0.25)'$, and $\pi^1 = \pi^2 = \pi^3 = 1/3$. We first choose 4 variables out of 8, and then construct the ave-rk$^+$ statistic using the procedure for the 4-variable model discussed in the previous paragraph. Because $\binom{8}{4} = 70$, we construct 70 ave-rk$^+$ statistics. Finally, we compute the max-rk$^+$ statistic from these 70 ave-rk$^+$ statistics. Figure 2(b) reports the selection frequency of SHT using the max-rk$^+$ statistic and the mean selection frequencies by SHT across 70 different ave-rk$^+$ statistics. The max-rk$^+$ statistic performs substantially better than individual ave-rk$^+$ statistics, suggesting that combining information from different ave-rk$^+$ statistics improves the performance of our procedures.

5 Examples

5.1 Intergenerational occupational mobility in Great Britain

We estimate the number of latent classes in the table of intergenerational mobility from father's occupation to subject's occupation in Great Britain, originally studied by Clogg (1981) using latent class models. Clogg estimates the 2-class and 3-class models using these data by imposing *a priori* restrictions on a set of parameters. Panel (1) of Table 1 presents the result of the SHT procedure applied to the 5×5 table of social mobility in Great Britain taken from Table 1.B of Clogg (1981); the null hypothesis that the number of latent classes is no more than 4 is rejected at any significance level. The AIC, BIC, and HQ procedures also indicate that the number of latent classes is at least 5 (not reported in the table). As

reported in Panel (2) of Table 1, we further examine the number of latent classes in the 8×8 table using Table 1.C of Clogg (1981) starting from the null hypothesis of no more than 5 classes. SHT suggests that this intergenerational occupational mobility data could be generated from 7 latent classes while BIC, AIC, and HQ suggest 5, 8, and 6 latent classes, respectively. Overall, the results of our procedures suggest that there are more than 5 latent classes, rejecting the 2- and 3-class models studied by Clogg.

Table 1: Intergenerational Social Mobility in Great Britain

	(1) 5×5 Table				(2) 8×8 Table		
Null hypothesis (H_0)	$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M = 5$	$M = 6$	$M = 7$
rk statistic	557.08	144.64	48.18	15.71	35.59	12.33	2.27
d.f.	16	9	4	1	9	4	1
p-value	0.000	0.000	0.000	0.000	0.000	0.015	0.132

Notes: The data are from Tables 1.B and 1.C of Clogg (1981).

Table 2: Type of Trade and Ethnic Group Data, Amsterdam and Rotterdam

Values of rk statistics and the degrees of freedom								
	Amsterdam				Rotterdam			
Null hypothesis (H_0)	$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M = 1$	$M = 2$	$M = 3$	$M = 4$
rk statistic	318.09	57.87	13.48	0.23	190.23	60.82	9.20	1.88
d.f.	20	12	6	2	20	12	6	2
p-value	0.000	0.000	0.036	0.891	0.000	0.000	0.163	0.391

Notes: The data are from Table 2a of van der Heijden et al. (2002).

5.2 Types of trades started by different ethnic groups

The second example analyzes the difference across ethnic groups in the types of trades they start in two large cities in the Netherlands, Amsterdam and Rotterdam, studied by van der Heijden, van der Ark, and Mooijaart (2002). There are 6 types of trades and 5 ethnic groups for each of the two cities. The members of some ethnic groups are more likely to start certain types of trades because of such factors as the number of clients in the same ethnic group or their level of human capital, including knowledge of the Dutch language. From this viewpoint, each latent class could be reflecting a specific type of network and human capital. Based on likelihood ratio statistics, van der Heijden et al. (2002) conclude that the number of latent classes $M = 3$ "seems adequate" for both Amsterdam and Rotterdam. We apply our procedures to examine if the number of latent classes is at least 3 or not. Table 2 shows the rk statistics and the corresponding p-values from the SHT procedure. For Amsterdam, SHT suggests 3 or 4 latent classes, whereas AIC, BIC, and HQ suggest 4, 2, and 3 latent classes, respectively. For Rotterdam, all of our procedures suggest 3 latent classes.

5.3 Response patterns in five-item subsets of LSAT

In our third example, we analyze the response patterns in two different five-item subsets of LSAT, denoted by LSAT-6 and LSAT-7, originally studied by Mislevy (1984). We employ the max-rk$^+$ statistic to these data. The response to five items is represented by $\{W_1, W_2, W_3, W_4, W_5\}$ where $W_i \in \{0, 1\}$. We first choose 4 items out of 5 and then construct the ave-rk$^+$ statistic from the estimates of P^{α}_{Δ}s for three different groupings $\alpha = 1, 2, 3$, where we estimate the covariance matrix $\Omega_{r,c}(\mathcal{A}_0)$ using the asymptotic formula. Because there are $\binom{5}{4} = 5$ different ways of choosing 4 items out of 5, we construct the max-rk$^+$ statistic from the 5 ave-rk$^+$ statistics. SHT based on the max-rk$^+$ statistic suggests that $M \geq 2$ in LSAT-6 at $\alpha = 0.1$, 0.05, and 0.01, and that $M \geq 3$ at $\alpha = 0.1$ and 0.05 and $M \geq 2$ at $\alpha = 0.01$ in LSAT-7.

5.4 Example 3 of Hettmansperger and Thomas (2000)

We also apply our procedure to the data that consist of 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. Hettmansperger and Thomas apply various tests of the number of components to the data transformed to a binomial mixture: 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 $M = 3$. Following Hettmansperger and Thomas, we use the known cut-off point of 5 degree to define the 8 response variables $\{W_1, \ldots, W_8\}$ with $W_j = 1(|X_j| \leq 5^\circ)$, where X_j is the j-th error. We then calculate the max-rk$^+$ statistic from the 70 ave-rk$^+$ statistics with $c = 0.01$, each of which is constructed from the 4 chosen variables out of the 8 response variables using the covariance matrix $\Omega_r(\mathcal{A}_0)$ estimated by the asymptotic formula. The SHT based on the max-rk$^+$ statistic suggests that $M \geq 3$ at $\alpha = 0.1$, 0.05, and 0.01, consistent with the Hettmansperger and Thomas's result from the bootstrapped likelihood ratio test.

Acknowledgement

The authors are grateful to the Editor, Associate Editor, and anonymous referees whose comments immensely improved the paper. The authors thank Lealand Morin for the helpful comments. This work was supported by SSHRC, Royal Bank of Canada Fellowship, and JSPS Grant-in-Aid for Research Activity Start-up 21830036 and JSPS Grant-in-Aid for Scientific Research (C) 23530249.

References

Allman, E. S., Matias, C., and Rhodes, J. A. (2009). "Identifiability of parameters in latent structure models with many observed variables." *The Annals of Statistics*, 37, 3099-3132.

Andrews, D. W. K. (1987). "Asymptotic results for generalized Wald tests." *Econometric Theory*, 3, 348-358.

Benaglia, T. Chauveau, D., and Hunter, D. R. (2009). "An EM-like algorithm for semi- and non-parametric estimation in multivariate mixtures." *Journal of Computational and Graphical Statistics*, 18, 505-526.

Bhattacharya, A. and Dunson, D. B. (2011). "Simple factor models for multivariate unordered categorical data." Preprint, Duke University.

Borkowf, C. B. (2000). "On multidimensional contingency tables with categories defined by the empirical quantiles of the marginal data." *Journal of Statistical Planning and Inference*, 91, 33-51.

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

Clogg, C. C. (1981). "Latent Structure Models of Mobility." *The American Journal of Sociology*, 86, 836-868.

Clogg, C. C. (1995). "Latent class models." In: Arminger, G., Clogg, C. C., and Sobel, M. E. (Eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, Plenum Press, New York, 311-359.

Cohen, J. E. and Rothblum, U. G. (1993). "Nonnegative ranks, decompositions, and factorizations of nonnegative matrices." *Linear Algebra and its Applications*, 190, 149-168.

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

de Leeuw, J. and van der Heijden, P. G. M. (1988) The analysis of time-budgets with a latent time-budget model. In: E. Diday (Ed.), *Data Analysis and Informatics 5*, Amsterdam: North-Holland, 159-166.

Dong, B., Lin, M. M., and Chu., M. T. (2009). "Nonnegative rank factorization via rank reduction." Preprint, North Carolina State University.

Dunson, D. B., and Xing, C. (2009). "Nonparametric Bayes modeling of multivariate categorical data." *Journal of the American Statistical Association*, 104, 1042-1051.

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.

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

Goodman, L. A. (1974). "The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: a modified latent structure approach." *American Journal of Sociology*, 79, 1179-1259.

Hagenaars, J. A. and McCutcheon, A. L. (2002). *Applied Latent Class Analysis*, Cambridge: Cambridge University Press.

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. (2009). "Nonparametric identification of finite mixture models of dynamic discrete choices." *Econometrica*, 77, 135-175.

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.

Levine, M., Hunter, D. R., and Chauveau, D. (2011). "Maximum smoothed likelihood for multivariate mixtures." *Biometrika*, 98, 403-416.

Lim, L.-H. and Comon, P. (2009) "Nonnegative approximations of nonnegative tensors." *Journal of Chemometrics*, 23, 432-441.

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.

Lütkepohl, H. and Burda, M. M. (1997). "Modified Wald tests under nonregular conditions." *Journal of Econometrics*, 78, 315-332.

Magidson, J. and Vermunt, J. K. (2004). "Latent class models." In: Kaplan, D. (Ed.), *The Sage Handbook of Quantitative Methodology for the Social Sciences*, Thousand Oakes: Sage Publications, 175-198.

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.

Mislevy, R. J. (1984). "Estimating latent distribution." *Psychometrika*, 49, 359-381.

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., and Schork, A. (1990). "On the asymmetry of biological frequency distributions." Genetic Epidemiology, 7, 427-446.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton: Chapman & Hall/CRC.

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

van der Heijden, P. G. M., van der Ark, L. A., and Mooijaart, A. (2002). "Some examples of latent budget analysis and its extensions." In: Hagenaars, J. A. and McCutcheon, A. L. (Eds.), *Applied Latent Class Analysis*, Cambridge: Cambridge University Press, 107-136.

Vavasis, S. (2009). "On the complexity of nonnegative matrix factorization." *SIAM Journal of Optimization*, 20, 1364-1377.

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.