Nonparametric Identification of Multivariate Mixtures

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

Katsumi Shimotsu[*]
Department of Economics
Hitotsubashi University
shimotsu@econ.hit-u.ac.jp

August 30, 2010

Abstract

This article analyzes the identifiability of 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 and the component distributions from the distribution function of the observed data.

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. A lower bound on the number of components (M) is nonparametrically identifiable if $k \geq 2$, and the maximum identifiable number of components is determined by the number of different values each variable takes. When M is known, the mixing proportions and the component distributions are nonparametrically identified from matrices constructed from the distribution function of the data if (i) $k \geq 3$, (ii) two of k variables take at least M different values, and (iii) these matrices satisfy some rank and eigenvalue conditions.

For the unknown M case, we propose an algorithm that possibly identifies M and the component distributions from data. We discuss a condition for nonparametric identification and its observable implications. In case M cannot be identified, we use our identification condition to develop a procedure that consistently estimates a lower bound on the number of components by estimating the rank of a matrix constructed from the distribution function of observed variables.

Key words and phrases: finite mixture; latent class analysis; latent class model; model selection; number of components; rank estimation

---

[*]Address for correspondence: Katsumi Shimotsu, Department of Economics, Hitotsubashi University.

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 (e.g., Keane and Wolpin 1997; Cameron and Heckman 1998). 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 the number of components and/or the form of the component distributions despite their key role in the specification of mixtures. In many applications, the component distributions are assumed to belong to a certain parametric family, such as normal, and the number of components is then 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. 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)). Further, 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 k-variate, $\tilde{M}$-component finite mixture models of $W = (W_1, ..., W_k)$ 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. When $F(w)$ can be expressed as (1), it is also possible to write $F(w)$ as a mixture with more than $\tilde{M}$ components. Therefore, we define the number of components in $F(w)$, M, 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, the component distributions (F_j^m's), and the mixing proportions (π^m's) 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. Identification problems differ from problems of statistical

inference in that we assume hypothetical access to infinite data; identifiability is a prerequisite for statistical inference since consistent estimation is not possible without identifiability (Koopmans and Reiersøl 1950; Koopmans 1950; Allman et al. 2009). For example, Goodman (1974b) analyzes a finite mixture model with four binary variables and shows that the model is not identifiable so that consistently estimating such a model is not possible without further restrictions. Nonparametric identifiability 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 k-variate finite mixture models (1). Hettmansperger and Thomas (2000) and Cruz-Medina et al. (2004) provide sufficient conditions for the nonparametric identification of finite mixtures with iid marginals.

The mixture model (1) assumes that the marginal distributions are independent conditional on belonging to a subpopulation. The independence assumption is a key assumption, and it is certainly strong. However, it is applicable to many cases in practice (Hettmansperger and Thomas 2000; Cruz-Medina et al. 2004; Zhou et al. 2005), and the model (1) encompasses models in latent class analysis (Lazarsfeld and Henry, 1968) that has been widely used in many fields including sociology, psychology, and biostatistics (Clogg 1995; Hagenaars and McCutcheon 2002; Magidson and Vermunt 2004; Skrondal and Rabe-Hesketh 2004). Further, as argued by Hall et al. (2005), a practical consideration associated with the curse of dimensionality may necessitate imposing independence when modeling multivariate data.

We make the following contributions. We identify the objects of our interest by transforming each element of W to a discrete random variable through partitioning its support and then analyzing the resulting (multiway) contingency table. First, 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 variation within each variable provides information on the number of components, and the maximum identifiable number of components is limited by the number of different values each variable takes.

Second, we establish that, when M is known, the mixing proportions and the component distributions are nonparametrically identified from matrices constructed from the distribution function of data if (i) $k \geq 3$, (ii) two of k variables take at least M different values, and (iii) these matrices satisfy some rank and eigenvalue conditions. These sufficient conditions are, in principle, testable from the observed data. Here, the requirement on the number of variables k is stronger than in identifying only a lower bound on the number of components. For the unknown M case, we develop an algorithm that possibly identifies both M and the component distributions from data; we provide a sufficient condition for nonparametric identification under unknown M and discuss its observable implications.

Our sufficient conditions for nonparametric identification when M is known substantially improve the requirement on the number of variables, k, in the existing literature while few

identification results exist for the case M is unknown. Using model (1) with known M and assuming additionally the W_j's are identically distributed within each component (i.e., $F_j^m(w_j) = F^m(w_j)$ for all j's), 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). Their transformation achieves robustness against parametric misspecification as we do, but their sufficient condition requires $k \geq 2M - 1$. Hence, for instance, if $k = 3$, at the most, two components are identifiable. In contrast, our analysis shows that, even when $k = 3$, a large number of components can be identified using the variation in W. Further, their approach relies on the additional assumption of identically distributed marginals, and thus our approach is applicable to a wider class of mixture models than theirs. On the other hand, one has to be cautious of using our identification algorithm for statistical inference because using higher order partitions of W might make the data thinner and the inference more difficult in finite samples. Our sufficient conditions are also applicable to latent class analysis and improve the previously established identification conditions by Anderson (1954), Gibson (1955) and Madansky (1960), which require $2^{(k-1)/2} \geq M$. Hall, Neeman, Pakyari, and Elmore (2005) analyze model (1) but their sufficient condition requires $k \geq (1+o(1))6M \log M$ as $M \to \infty$.

In a recent study, Allman, Matias, and Rhodes (2009) use the same model as ours and analyze nonparametric identification when $k \geq 3$ and M is known. Applying the result of Kruskal's theorem (Kruskal, 1976, 1977), Allman et al. (2009) approach the problem by finding sufficient conditions in terms of the unobservable component distributions, whereas we approach the problem by finding sufficient conditions in terms of the distribution function of observable data. Our identification conditions are stronger than those in Allman et al. (2009) in some cases but weaker in other cases, hence our results for $k \geq 3$ and those of Allman et al. (2009) are complementary to each other.

Our identification condition on the number of components is stated in terms of the rank of a matrix constructed from the distribution function of observed variables W. By estimating the rank of its empirical analogue, we develop a procedure to consistently estimate a lower bound on 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.

It has also been known that the likelihood ratio test does not lead to the standard chi-square distribution when applied to testing the number of components because the parameter value specified under the null hypothesis lies on the boundary of the parameter space. In

contrast, our selection procedure is based on a statistic that has the asymptotic chi-squared distribution and is easy to implement without requiring the estimation of a mixture model with a different number of components. Simulations illustrate that our procedure performs well.

Kasahara and Shimotsu (2009) 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 a lower bound on the number of components under $k \geq 2$. Section 3 discusses sufficient conditions for nonparametric identification of the mixing proportions and the component distributions under $k \geq 3$. Section 4 introduces a procedure to test a lower bound on the number of mixture components. Section 5 reports simulation results, and empirical examples are provided in section 6. Proofs are collected in the Appendix.

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

2.1 Two-variable case

We first analyze 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 s and t mutually exclusive and exhaustive subsets, respectively, as $\Delta^x = \{\delta_1^x, \ldots, \delta_s^x\}$ and $\Delta^y = \{\delta_1^y, \ldots, \delta_t^y\}$. Define $\Delta = \Delta^x \times \Delta^y$, and let $\mathcal{D}$ be the set of all finite partitions of $\mathcal{X}$ and $\mathcal{Y}$. Given a choice of partition $\Delta \in \mathcal{D}$, collect the distributions of

X and Y conditional on being from the m-th subpopulation into a vector as

$$p_x^m = (\Pr(x \in \delta_1^x|m), \ldots, \Pr(x \in \delta_s^x|m))' \quad \text{and} \quad p_y^m = (\Pr(y \in \delta_1^y|m), \ldots, \Pr(y \in \delta_t^y|m))',$$
(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 $(a, b) = (1, 1), \ldots, (s, t)$ into an $s \times t$ bivariate probability matrix as

$$P_\Delta = \begin{bmatrix} \Pr(X \in \delta_1^x, Y \in \delta_1^y) & \cdots & \Pr(X \in \delta_1^x, Y \in \delta_t^y) \\ \vdots & \ddots & \vdots \\ \Pr(X \in \delta_s^x, Y \in \delta_1^y) & \cdots & \Pr(X \in \delta_s^x, Y \in \delta_t^y) \end{bmatrix}.$$
(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

$$P_\Delta = \sum_{m=1}^{\tilde{M}} \pi^m p_x^m (p_y^m)', \quad \pi^m > 0, \quad \sum_{m=1}^{\tilde{M}} \pi^m = 1.$$
(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. Since 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 nonnegative rank of P_Δ is no larger than M, but could be strictly smaller than M when a single partition Δ does not fully reveal the information for identifying the number of components in $F(x, y)$. M is identified with the maximum value of $\text{rank}_+(P_\Delta)$'s over all possible finite partitions, i.e., $M = \max_{\Delta \in \mathcal{D}} \text{rank}_+(P_\Delta)$.

The following proposition, originally due to 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 (Cohen and Rothblum, 1993) *(a) $rank(P_\Delta) \leq rank_+(P_\Delta) \leq \min\{s, t\}$. (b) If $rank(P_\Delta) \leq 2$, then $rank(P_\Delta) = rank_+(P_\Delta)$. (c) If $s \leq 3$ or $t \leq 3$, then $rank_+(P_\Delta) = rank(P_\Delta)$.*

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 $s \leq |\mathcal{X}|$ and $t \leq |\mathcal{Y}|$, where $|\mathcal{S}|$ denotes the number of elements in a set $\mathcal{S}$. It follows from Proposition 1 that $\text{rank}_+(P_\Delta) = \text{rank}(P_\Delta)$ if $\text{rank}_+(P_\Delta) \leq 3$, whereas $\text{rank}_+(P_\Delta)$ may be strictly larger than $\text{rank}(P_\Delta)$ when $\text{rank}_+(P_\Delta) > 3$.

The rank of P_Δ could be different from the nonnegative rank of P_Δ 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. For example, suppose that $s = t = 4$ and $P_\Delta = \sum_{m=1}^4 \pi^m p_x^m (p_y^m)'$, where $\pi^m > 0$ and p_x^m's are linearly independent but $p_y^1 + p_y^2 - p_y^3 - p_y^4 = 0$, so that the rank of P_Δ is 3. Writing one p_y^m in terms of the other p_y^m's 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 the nonnegative rank of P_Δ is strictly larger than 3.

The nonnegative rank of P_Δ equals the number of components in P_Δ. However, determining the nonnegative rank of a matrix is computationally difficult[1], and it is still a subject of on-going 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. The tightest lower bound on M we may construct from the rank of P_Δ's is the maximal rank of P_Δ's over all possible finite partitions of $\mathcal{X} \times \mathcal{Y}$.

Corollary 1 *The number of components in $F(x, y)$, M, is no smaller than the maximal rank of P_Δ over all possible finite partitions of $\mathcal{X} \times \mathcal{Y}$, i.e., $M \geq \max_{\Delta \in \mathcal{D}} \text{rank}(P_\Delta)$.*

When both X and Y have finite support points, we may choose $\Delta = \mathcal{X} \times \mathcal{Y}$, and the rank of $P_{\mathcal{X} \times \mathcal{Y}}$ gives the tightest lower bound on M. In Section 4, we develop a procedure to estimate a lower bound on M by estimating the rank of P_Δ.

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. First, we group k variables in $W = (W_1, \ldots, W_k)$ into two groups. Since there are multiple ways to group the variables in W, a tighter lower bound on M is obtained by combining the information across different groupings rather than using only one grouping. We index the groupings by α, and let X^α and Y^α denote the first and second group of variables. For example, when k is even, we may have $X^\alpha = (W_1, \ldots, W_{k/2})$ and $Y^\alpha = (W_{k/2+1}, \ldots, W_k)$ for some α. Let M^α denote the number of components in $F(x^\alpha, y^\alpha)$, defined as the smallest number of mixture components for which the joint distribution of (X^α, Y^α) admits a finite mixture representation as

$$F(x^\alpha, y^\alpha) = \sum_{m=1}^{M^\alpha} \pi^m F_{x^\alpha}^m(x^\alpha) F_{y^\alpha}^m(y^\alpha). \tag{6}$$

---

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

Let $\mathcal{A}$ be the set of indices α's for all the possible groupings. The relation $M \geq \max_{\alpha \in \mathcal{A}} M^\alpha$ then holds because the factorization in (1) is the factorization in (6) with an additional constraint that the elements of X^α and Y^α are conditionally independent. On the other hand, M could be strictly larger than $\max_{\alpha \in \mathcal{A}} M^\alpha$ because grouping several variables into two could lead to a loss of information.

Let Δ denote a partition of the support of (X^α, Y^α). Constructing the matrix P_Δ^α from the distribution of (X^α, Y^α), the number of components in P_Δ^α is given by the nonnegative rank of P_Δ^α. Taking its maximum across different partitions gives $M^\alpha = \max_{\Delta \in \mathcal{D}^\alpha} \mathrm{rank}_+(P_\Delta^\alpha)$, where $\mathcal{D}^\alpha$ denotes the set of all possible finite partitions of the support of (X^α, Y^α). The tightest lower bound on M in terms of $\mathrm{rank}_+(P_\Delta^\alpha)$'s is obtained by repeating this procedure for different groupings and taking the maximum of M^α over $\alpha \in \mathcal{A}$, i.e., $M \geq \max_{\alpha \in \mathcal{A}} \max_{\Delta \in \mathcal{D}^\alpha} \mathrm{rank}_+(P_\Delta^\alpha)$.

In view of the difficulty of determining nonnegative rank, an alternative lower bound is obtained from taking the maximum of $\mathrm{rank}(P_\Delta^\alpha)$ over Δ and α. Namely, we have $M \geq \max_{\alpha \in \mathcal{A}} \max_{\Delta \in \mathcal{D}^\alpha} \mathrm{rank}(P_\Delta^\alpha)$. This lower bound is the tightest lower bound on M in terms of the rank of P_Δ^α's but may not be as tight as the one based on the nonnegative rank.

2.3 Relation to latent class analysis

Consider a special case in which an observation vector $W = (W_1, \dots, W_k)$ consists of k dichotomous or polytomous responses, typically answers to questions or results of diagnoses. In this case, our model (1) becomes identical to the model used in *latent class analysis* (Lazarsfeld and Henry 1968). For recent surveys and applications of latent class analysis, see Clogg (1995), Hagenaars and McCutcheon (2002), Magidson and Vermunt (2004), Skrondal and Rabe-Hesketh (2004), and the references therein.

In latent class analysis, it is assumed that the observations belong to one of the M latent classes, with the probability of being in class $m \in \{1, \dots, M\}$ equal to π^m. The responses are assumed to be conditionally independent given membership in a given latent class. Let $\xi = (\xi_1, \dots, \xi_k)'$ denote a possible value of W, then latent class analysis formulates the distribution function of W as

$$\Pr(W = \xi) = \sum_{m=1}^{M} \pi^m \Pr(W_1 = \xi_1 | m) \cdots \Pr(W_k = \xi_k | m). \tag{7}$$

Therefore, we can identify a lower bound on M in a latent class model (7) by grouping the variables in W into two groups X^α and Y^α and computing the rank of P_Δ^α.

The latent class analysis with $k = 2$ (two-way contingency table) is also known as *latent budget analysis* (Goodman 1974a; Clogg 1981; de Leeuw and van der Heijden 1988). Because the parameters in a latent budget model are not identifiable, applied researchers impose *a priori* restrictions on the model's parameters to make it identifiable and fit the model to data.

However, the validity of such restrictions is not always clear. Our result indicates that it is possible to identify a lower bound on M without imposing restrictions on the parameters.

3 Nonparametric identification of finite mixture models

When $k = 2$, the mixture model (1) is not identified regardless of the number of values the W_j's can take. In a latent class model of a two-dimensional contingency table, Clogg (1981, p. 847) shows that two degrees of freedom are lost and the model is not identified unless two restrictions are imposed on the parameters. When $M = 2$, Hall and Zhou (2003, Theorem 4.1) solve the model (1) for the unknown parameters, $\{\pi^1, F_1^1(w_1), F_1^2(w_1), F_2^1(w_2), F_2^2(w_2)\}$, and show that there is a two-parameter continuum of solutions to (1).

This section considers nonparametric identification of the mixing proportions and the component distributions of a finite mixture model (1) with $k = 3$:

$$F(x, y, z) = \sum_{m=1}^{\tilde{M}} \pi^m F_x^m(x) F_y^m(y) F_z^m(z), \quad \pi^m > 0, \quad \sum_{m=1}^{\tilde{M}} \pi^m = 1, \tag{8}$$

where $F(x, y, z)$ is the distribution function of variable (X, Y, Z), and $F_x^m(x)$, $F_y^m(y)$, and $F_z^m(z)$ are the distribution functions of X, Y, and Z conditional on being from the m-th subpopulation, respectively. Similar to Section 2, define the number of components in $F(x, y, z)$, M, as the smallest positive integer $\tilde{M}$ for which a finite mixture representation (8) can be found.

We first provide sufficient conditions for nonparametric identification of the finite mixture model (8) when the value of M is known. We then extend our identification analysis to the unknown M case, and discuss identification in a $k > 3$ variable model.

3.1 Identification when M is known

We first transform the distribution function $F(x, y, z)$ into an $M \times M \times 2$ contingency table, and identify the component distributions associated with this contingency table. We then show that, once the component distributions are identified with respect to this contingency table, it is possible to identify $F_x^m(x)$, $F_y^m(y)$, and $F_z^m(z)$ at any support point (x, y, z).

Denote the support of X, Y, and Z by $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$, respectively. Partition $\mathcal{X}$ and $\mathcal{Y}$ into M mutually exclusive and exhaustive subsets. With a slight abuse of notation, let $\Delta^x = \{\delta_1^x, \ldots, \delta_M^x\}$ and $\Delta^y = \{\delta_1^y, \ldots, \delta_M^y\}$ denote these partitions, and let $\Delta = \Delta^x \times \Delta^y$. Similarly, partition $\mathcal{Z}$ into 2 mutually exclusive and exhaustive subsets as $\Delta^z = \{\delta_1^z, \delta_2^z\}$.

Define p_x^m, p_y^m, and P_Δ as in (3) and (4) where the number of partitions on the variables (X, Y) is set equal to the number of components, i.e., $s = t = M$. For partition level

9

$h \in \{1, 2\}$ of the variable Z, define an $M \times M$ matrix

$$P_{\Delta,h} = \begin{bmatrix} \Pr(X \in \delta_1^x, Y \in \delta_1^y, Z \in \delta_h^z) & \cdots & \Pr(X \in \delta_1^x, Y \in \delta_M^y, Z \in \delta_h^z) \\ \vdots & \ddots & \vdots \\ \Pr(X \in \delta_M^x, Y \in \delta_1^y, Z \in \delta_h^z) & \cdots & \Pr(X \in \delta_M^x, Y \in \delta_M^y, Z \in \delta_h^z) \end{bmatrix}. \tag{9}$$

Note that $P_{\Delta,1}$ and $P_{\Delta,2}$ summarize the distribution of (X, Y, Z) on the partition $\Delta \times \Delta^z$. Let $p_z^m(h) = \Pr(Z \in \delta_h^z | m)$ be the m-th component distribution of Z for partition level $h = 1, 2$. Then, $P_{\Delta,h}$ can be written as, similar to (5),

$$P_{\Delta,h} = \sum_{m=1}^{M} (\pi^m p_z^m(h)) p_x^m (p_y^m)'. \tag{10}$$

We proceed to write P_Δ and $P_{\Delta,h}$ in matrix form. Collect the component distributions of X and Y restricted to the partition Δ, p_x^m's and p_y^m's, into two $M \times M$ matrices as

$$L_x = \begin{bmatrix} p_x^1, \ldots, p_x^M \end{bmatrix} \text{ and } L_y = \begin{bmatrix} p_y^1, \ldots, p_y^M \end{bmatrix}, \tag{11}$$

respectively, where L_x and L_y implicitly depend on the choice of partition Δ. Collect π^m's and $p_z^m(h)$'s into $M \times M$ diagonal matrices as $V = \mathrm{diag}(\pi^1, \ldots, \pi^M)$ and $D_h = \mathrm{diag}(p_z^1(h), \ldots, p_z^M(h))$. With this notation at hand, P_Δ and $P_{\Delta,h}$ in (5) and (10) can be expressed in matrix forms as

$$P_\Delta = L_x V (L_y)', \quad P_{\Delta,h} = L_x D_h V (L_y)' = L_x V D_h (L_y)'. \tag{12}$$

Here, P_Δ and $P_{\Delta,h}$ are functions of the observable variables, whereas L_x, L_y, D_h, and V represent the unknown component distributions and the unknown mixture probabilities. The following proposition provides a sufficient condition for identifying L_x, L_y, D_h, and V from P_Δ and $P_{\Delta,h}$ as well as for identifying $F_x^m(\cdot)$'s, $F_y^m(\cdot)$'s, and $F_z^m(\cdot)$'s from $F(x, y, z)$.

Proposition 2 *Suppose that M is known and that there exists a partition $\Delta \times \Delta^z$ on the variables (X, Y, Z) for which the matrix P_Δ is nonsingular and the eigenvalues of $P_{\Delta,h}(P_\Delta)^{-1}$ are distinct for partition level $h = 1$ of the variable Z. Then, for such a partition $\Delta \times \Delta^z$, we may uniquely determine L_x, L_y, D_h, and V from P_Δ and $P_{\Delta,h}$. Further, we may uniquely determine the component distributions $F_x^m(\cdot)$, $F_y^m(\cdot)$, and $F_z^m(\cdot)$ for $m = 1, \ldots, M$ in (8) from the distribution function of (X, Y, Z), $F(x, y, z)$.*

Remark 1

1. *Since $P_{\Delta,1} + P_{\Delta,2} = P_\Delta$, the above sufficient condition can be stated equivalently in terms of the eigenvalues of $P_{\Delta,2}(P_\Delta)^{-1}$ for partition level $h = 2$.*

2. *The proof of Proposition 2 is constructive. Namely, the proof provides an algorithm to compute L_x, L_y, V, and D_h from $P_{\Delta,h}$ and P_Δ. Under the stated assumptions in Proposition 2, we have $P_{\Delta,h}(P_\Delta)^{-1} = L_x D_h (L_x)^{-1}$ and $(P_{\Delta,h})'((P_\Delta)')^{-1} = L_y D_h (L_y)^{-1}$, and we can compute L_x and L_y from the eigenvectors of $P_{\Delta,h}(P_\Delta)^{-1}$ and $(P_{\Delta,h})'((P_\Delta)')^{-1}$, while their eigenvalues identify D_h. Finally, V is computed as $(L_x)^{-1} P_\Delta (L_y')^{-1}$.*

   Once L_x, L_y, and V are identified for some partition Δ, $F_x^m(\cdot)$, $F_y^m(\cdot)$, and $F_z^m(\cdot)$ are identified from $F(x, y, z)$ without any additional assumptions. For example, for any $x \in \mathcal{X}$, define a $1 \times M$ vector

$$P_{x, \Delta^y} = \left(\Pr(X \leq x, Y \in \delta_1^y), \cdots, \Pr(X \leq x, Y \in \delta_M^y) \right), \tag{13}$$

   which can be computed from the distribution function of the data. Define $q_x = (F_x^1(x), \ldots, F_x^M(x))$. Then, since $P_{x, \Delta^y} = q_x V(L_y)'$ holds, q_x is identified as $q_x = P_{x, \Delta^y}((L_y)')^{-1} V^{-1}$. Potentially, we can use this algorithm to estimate mixture models or to obtain initial values for other estimation algorithms.

3. *The non-singularity of P_Δ requires that X and Y take at least M distinct values. Hence, the number of support points in X and Y provides the upper bound for the identifiable number of components.*

Our sufficient condition in Proposition 2 is new in the literature, aside from a recent contribution by Allman et al. (2009). Under the assumption of known M, Allman et al. (2009, Section 7) analyze the same model as our model but via Kruskal's theorem (Kruskal, 1976, 1977). Theorems 8 and 9 (and the extension on p. 3116) of Allman et al. (2009) establish a stronger result than ours, in that it is possible to identify more than M types from an $M \times M \times M$ contingency table whereas our results do not improve for an $M \times M \times M$ table. However, their sufficient conditions are stated in terms of the component distribution (L_x, L_y and L_z in our notation), which is not observable. In contrast, our sufficient conditions in Proposition 2 are stated in terms of what we observe. Corollary 11 of Allman et al. (2009) gives a sufficient condition in terms of the observables, but it requires an $M \times M \times M$ contingency table in order to identify M types, whereas we require only an $M \times M \times 2$ contingency table. Further, our proof is constructive, whereas the proof of Allman et al. (2009) is not.

Proposition 2 makes a contribution to latent class analysis. The existing identification results in latent class analysis (Anderson (1954), Gibson (1955), and Madansky (1960)) focus on dichotomous response variables, and, consequently, relate the number of variables (k) with the number of identifiable components (M). Under the assumption of known M, Madansky (1960) obtains the weakest sufficient condition, which requires $2^{(k-1)/2} \geq M$.[2] On the other

---

[2]The same condition, $2^{(K-1)/2} \geq M$, is given by Corollary 5 of Allman et al. (2009).

hand, our Proposition 2 shows that the variation within the variables plays a key role for identification; even when $k = 3$ in latent class model (7), we may identify the number of components potentially up to the numbers of support points in W_1 and W_2, provided that the relevant rank condition in Proposition 2 is satisfied.

The sufficient condition in Proposition 2 includes a condition on the eigenvalues of $P_{\Delta,1}(P_\Delta)^{-1}$. Since $P_{\Delta,1}(P_\Delta)^{-1} = L_x D_1 (L_x)^{-1}$, the eigenvalues of $P_{\Delta,1}(P_\Delta)^{-1}$ are distinct if and only if $\Pr(Z \in \delta_1^z | i) \neq \Pr(Z \in \delta_1^z | j)$ for any pair of components $i \neq j$. If $\Pr(Z \in \delta_1^z | i) = \Pr(Z \in \delta_1^z | j)$ for some $i \neq j$, then the partition δ_1^z provides no information on distinguishing between components i and j. In such a case, however, we may potentially identify p_x^i and p_x^j separately by partitioning $\mathcal{Z}$ into $u > 2$ subsets, $\delta_1^z, \ldots, \delta_u^z$, and applying the algorithm to some other partition δ_ℓ^z if the eigenvalues of $P_{\Delta,\ell}(P_\Delta)^{-1}$ that correspond to components i and j are distinct. The following corollary shows that, repeating the algorithm across different partitions of Z, identification is possible even when the eigenvalue condition of Proposition 2 does not hold. Proposition 2 and Corollary 2 are equivalent if Z has only two support points.

Corollary 2 *Suppose that M is known and that there exists a partition $\Delta \times \Delta^z$ with $\Delta^z = \{\delta_1^z, \ldots, \delta_u^z\}$ such that P_Δ is nonsingular and there are M linearly independent eigenvectors in the set of eigenvectors of $P_{\Delta,1}(P_\Delta)^{-1}, \ldots, P_{\Delta,u}(P_\Delta)^{-1}$. Then, we may uniquely determine π^m, $F_x^m(\cdot)$, $F_y^m(\cdot)$, and $F_z^m(\cdot)$ for $m = 1, \ldots, M$ in (8) from $F(x, y, z)$.*

Hall and Zhou (2003, Theorem 4.3 and Appendix) show that model (8) with $M = 2$ is identifiable if and only if $F(x, y, z)$ is *irreducible*, namely, if none of its bivariate marginals factorizes into the product of univariate marginals. In model (8) with $M \geq 3$, irreducibility is also necessary for the conditions of Proposition 2 to hold but not sufficient for identification. We may consider a model such that $M = 3$ and $\text{rank}(P_\Delta) = 2$, so that $F(x, y, z)$ is irreducible but the mixture model is not identifiable.

The following proposition provides a necessary condition for nonparametric identification when $M \geq 3$. Part (a) provides a condition that corresponds to the irreducibility condition while part (b) shows that the eigenvector condition in Corollary 2 (and the eigenvalue condition in Proposition 2 if Z has only two support points) is also necessary.

Proposition 3 *(a) Suppose that $F_z^i(\cdot) = F_z^j(\cdot)$ for some pair of components $i \neq j$ in model (8). Then, it is not possible to uniquely determine $\{\pi^i, \pi^j, F_x^i(\cdot), F_x^j(\cdot), F_y^i(\cdot), F_y^j(\cdot)\}$ in (8) from $F(x, y, z)$. (b) Suppose that M is known but the eigenvector condition of Corollary 2 does not hold for any partition $\Delta \times \Delta^z$ such that P_Δ is nonsingular. Then, it is not possible to uniquely determine π^m, $F_x^m(\cdot)$, $F_y^m(\cdot)$, and $F_z^m(\cdot)$ for $m = 1, ..., M$ in (8) from $F(x, y, z)$.*

12

3.2 Identification when M is unknown

The assumption of known M is important in our Proposition 2 and the other existing identification studies discussed above.[3] When M is unknown, currently no identification results are available.

In this subsection, we develop an algorithm that can potentially identify both M and the component distributions from the distribution function of the data. We extend our constructive approach of the known M case in Proposition 2 to the unknown M case where the dimension of P_Δ is not restricted to $M \times M$. Since P_Δ is not invertible in general, we compute the generalized inverse of P_Δ, multiply it with $P_{\Delta,h}$, and compute the eigenvectors and eigenvalues of the product. Under some conditions, these eigenvectors and eigenvalues identify M and component distributions if and only if $\text{rank}(P_\Delta) = M$.

Given a partition $\Delta^x = \{\delta_1^x, \ldots, \delta_s^x\}$, $\Delta^y = \{\delta_1^y, \ldots, \delta_t^y\}$, $\Delta^z = \{\delta_1^z, \delta_2^z\}$, define P_Δ by (4), and let $P_{\Delta,h}$ be an $s \times t$ matrix defined similar to (9) but the partition index (i,j) in (δ_i^x, δ_j^y) runs from $(1,1)$ to (s,t). Let p_x^m and p_y^m be $s \times 1$ and $t \times 1$ vectors defined by (3), respectively. Then, P_Δ and $P_{\Delta,h}$ can be expressed as $P_\Delta = L_x V (L_y)'$ and $P_{\Delta,h} = L_x D_h V (L_y)' = L_x V D_h (L_y)'$ as in (12), where the dimension of L_x and L_y are $s \times M$ and $t \times M$, respectively, while $V = \text{diag}(\pi^1, \ldots, \pi^M)$ and $D_h = \text{diag}(p_z^1(h), \ldots, p_z^M(h))$.

Our goal is to recover M, L_x, L_y, V, D_h, $F_x^m(\cdot)$, $F_y^m(\cdot)$, and $F_z^m(\cdot)$ from $F(x,y,z)$. To this end, we consider the following algorithm, which is similar to the one in Remark 1.2 but uses the Moore-Penrose generalized inverse in place of the ordinary inverse. Let A^+ denote the Moore-Penrose generalized inverse (henceforth M-P inverse) of A.

Step 1. Compute the eigenvalues and eigenvectors of $P_{\Delta,h}(P_\Delta)^+$. Let $\hat{M}$ denote the number of nonzero eigenvalues, and let $e^1, \ldots, e^{\hat{M}}$ denote these eigenvalues. Normalize the eigenvectors associated with $e^1, \ldots, e^{\hat{M}}$ so that the elements of each eigenvector sum to one, and collect them into an $s \times \hat{M}$ matrix $\hat{L}_x$.

Step 2. Compute the eigenvalues and eigenvectors of $P'_{\Delta,h}(P'_\Delta)^+$.[4] Similar to Step 1, construct a $t \times \hat{M}$ matrix $\hat{L}_y$ from the normalized eigenvectors of $P'_{\Delta,h}(P'_\Delta)^+$.

Step 3. Compute an $\hat{M} \times \hat{M}$ matrix $\hat{V} = (\hat{L}_x)^+ P_\Delta (\hat{L}'_y)^+$.

Step 4. For any $x \in \mathcal{X}$, define a $1 \times t$ vector P_{x,Δ^y} as in (13) except that M is replaced with t. Compute $(\hat{F}_x^1(x), \ldots, \hat{F}_x^{\hat{M}}(x)) = P_{x,\Delta^y}((\hat{L}_y)')^+ (\hat{V})^+$. For any $y \in \mathcal{Y}$, define a $s \times 1$ vector $P_{\Delta^x,y} = (\Pr(X \in \delta_1^x, Y \le y), \cdots, \Pr(X \in \delta_s^x, Y \le y))'$, and compute $(\hat{F}_y^1(y), \ldots, \hat{F}_y^{\hat{M}}(y))' = (\hat{V})^+ (\hat{L}_x)^+ P_{\Delta^x,y}$. Similarly, for any $z \in \mathcal{Z}$, compute $(\hat{F}_z^1(z), \ldots, \hat{F}_z^{\hat{M}}(z))'$ using $P_{\Delta^x,z}$ in place of $P_{\Delta^x,y}$.

---

[3]For example, Allman et al. (2009, p. 3105) write "we always assume the number of latent classes is known, which is crucial in using Kruskal's approach."

[4]Note that $P_{\Delta,h}(P_\Delta)^+$ and $P'_{\Delta,h}(P'_\Delta)^+$ have the same $\hat{M}$ nonzero eigenvalues.

This algorithm takes $\{\Delta^x, \Delta^y, \Delta^z, P_{\Delta,h}, P_\Delta\}$, $\{P_{x,\Delta^y}\}_{x\in\mathcal{X}}$, $\{P_{\Delta^x,y}\}_{y\in\mathcal{Y}}$, and $\{P_{\Delta^x,z}\}_{z\in\mathcal{Z}}$ as its input and generates $\{\hat{M}, \hat{L}_x, \hat{L}_y, \hat{V}\}$ and $\{e^m, \hat{F}_x^m(x), \hat{F}_y^m(y), \hat{F}_z^m(z)\}_{m=1}^{\hat{M}}$ for $(x, y, z) \in$ $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ as its output. The following proposition shows that the output of this algorithm identifies M and the component distributions if and only if $\mathrm{rank}(P_\Delta) = M$.

Proposition 4 *Suppose that the data are generated by the model (8) with M components. Further, suppose there exists a partition $\Delta^x \times \Delta^y \times \Delta^z$ such that the non-zero eigenvalues of $P_{\Delta,h}(P_\Delta)^+$ are distinct and $rank(P_\Delta) = rank(P_{\Delta,h})$ for some choice of partition level $h \in \{1, 2\}$. Then, the following (a) and (b) hold.*

(a) If $rank(P_\Delta) = M$, then $\hat{V}$ is a diagonal matrix whose elements are positive and sum to one. Further, the number of components, the component distributions, and the mixing proportions are uniquely determined from the algorithm as $M = \hat{M}$, $L_x = \hat{L}_x$, $L_y = \hat{L}_y$, $V = \hat{V}$, $D_h = diag(e^1, \ldots, e^{\hat{M}})$, $F_x^m(\cdot) = \hat{F}_x^m(\cdot)$, $F_y^m(\cdot) = \hat{F}_y^m(\cdot)$, and $F_z^m(\cdot) = \hat{F}_z^m(\cdot)$ for $m = 1, \ldots, M$.

(b) Only if $rank(P_\Delta) = M$, both $P_\Delta = \hat{L}_x \hat{V} \hat{L}_y'$ and $F(x, y, z) = \sum_{m=1}^{\hat{M}} \hat{\pi}^m \hat{F}_x^m(x) \hat{F}_y^m(y) \hat{F}_z^m(z)$ give valid finite mixture representations, where $\hat{\pi}^m$ is the m-th diagonal element of $\hat{V}$. Namely, (i) the elements of every column of $\hat{L}_x$ and $\hat{L}_y$ are nonnegative and sum to one, (ii) $\hat{V}$ is a diagonal matrix whose elements are positive and sum to one, and (iii) $\hat{F}_x^m(\cdot)$, $\hat{F}_y^m(\cdot)$, and $\hat{F}_z^m(\cdot)$ are valid distribution functions.

The eigenvalues of $P_{\Delta,h}(P_\Delta)^+$ correspond to $\Pr(Z \in \delta_h^z | m)$ when $rank(P_\Delta) = M$. If $rank(P_\Delta) \neq rank(P_{\Delta,h})$, then $\Pr(Z \in \delta_h^z | m) = 0$ for some component m, and the algorithm fails to identify M because $P_{\Delta,h}$ has fewer than M effective components. In view of Corollary 2, the condition on the nonzero eigenvalues of $P_{\Delta,h}(P_\Delta)^+$ can be weakened to the eigenvector condition similar to the one in Corollary 2.

Given a partition $\Delta \times \Delta^z$ such that the non-zero eigenvalues of $P_{\Delta,h}(P_\Delta)^+$ are distinct and $\mathrm{rank}(P_\Delta) = \mathrm{rank}(P_{\Delta,h})$, this algorithm generates valid component distributions if and only if $\mathrm{rank}(P_\Delta) = M$. The "only if" part gives a testable implication of $\mathrm{rank}(P_\Delta) = M$; if the algorithm produces $\hat{\pi}^m$, $\hat{F}_x^m(\cdot)$, $\hat{F}_y^m(\cdot)$, and $\hat{F}_z^m(\cdot)$ that give a valid finite mixture representation of $F(x, y, z)$ with $\hat{M}$ components, then M is equal to the rank of P_Δ, and hence M and component distributions are jointly identified from the distribution function of the data.

On the other hand, when at least one of the conditions (i)-(iii) in Proposition 4(b) is violated for any choice of partitions $\Delta \times \Delta^z$, we learn that $\mathrm{rank}(P_\Delta) < M$ for any partition Δ. In such a case, however, Proposition 4 does not tell us how to identify M and the component distributions. This is a limitation of our algorithm. Also, given our identification result, how to make statistical inference on the condition $\mathrm{rank}(P_\Delta) = M$ and how to nonparametrically estimate a mixture model from finite data are important future research topics that are not explored in this paper.

3.3 General k-variate case

We now extend our approach to a k-variate finite mixture model (1) with $k > 3$. We only consider the case of known M here but, for the case of unknown M, it is also possible to extend the identification result of Propositions 4 to a general k-variate mixture model.

Consider grouping W into three groups, $(X^\alpha, Y^\alpha, Z^\alpha)$, with the grouping index α. Let $\mathcal{A}$ be the set of indices α's for all possible groupings. For example, if k is odd, we can choose $X^\alpha = (W_1, \ldots, W_{(k-1)/2})$, $Y^\alpha = (W_{(k-1)/2+1}, \ldots, W_{k-1})$, and $Z^\alpha = W_k$ for some $\alpha \in \mathcal{A}$.

We apply Proposition 2 to X^α, Y^α, and Z^α as follows. Let $(\mathcal{X}^\alpha, \mathcal{Y}^\alpha, \mathcal{Z}^\alpha)$ denote the support of $(X^\alpha, Y^\alpha, Z^\alpha)$. As in Section 3.1, partition $\mathcal{X}^\alpha$ and $\mathcal{Y}^\alpha$ into M mutually exclusive and exhaustive subsets. Let $\Delta^{x,\alpha}$ and $\Delta^{y,\alpha}$ denote these partitions, and define $\Delta^\alpha = \Delta^{x,\alpha} \times \Delta^{y,\alpha}$. Similarly, partition $\mathcal{Z}^\alpha$ into 2 mutually exclusive and exhaustive subsets, and let $\Delta^{z,\alpha}$ denote this partition. Given Δ^α and $\Delta^{z,\alpha}$, we may construct P_{Δ^α} and $P_{\Delta^\alpha, h}$ analogously to P_Δ and $P_{\Delta, h}$ in (5) and (10), respectively.

Corollary 3 *Suppose that M is known and, for some choice of grouping $\alpha \in \mathcal{A}$, there exists a partition $\Delta^\alpha \times \Delta^{z,\alpha}$ such that P_{Δ^α} is nonsingular and the eigenvalues of $P_{\Delta^\alpha, h}(P_{\Delta^\alpha})^{-1}$ are distinct for partition level $h = 1$ of the variable Z^α. Then, we may uniquely determine π^m, $F_1^m(\cdot), \ldots, F_k^m(\cdot)$ for $m = 1, \ldots, M$ in (1) from $F(w_1, \ldots, w_k)$.*

Thus, the k-variate mixture model (1) is nonparametrically identified if the assumptions corresponding to those in Proposition 2 hold for at least one grouping in $\mathcal{A}$.

4 Estimating a lower bound on the number of components

Proposition 1 in Section 2 shows that the rank of an $s \times t$ matrix P_Δ in (5) gives a lower bound on the number of mixture components. In this section, we develop two procedures to estimate the rank of P_Δ for a given partition Δ: sequential hypothesis testing and model selection. These procedures are based on the test statistic proposed by Kleibergen and Paap (2006). We also extend these procedures to estimate the maximum rank of P_Δ's across different groupings of variables when there are more than two variables.

4.1 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. Write the singular value decomposition of an $s \times t$ matrix P_Δ as

$$P_\Delta = USV' = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix} \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix} \begin{pmatrix} V_{11}' & V_{12}' \\ V_{21}' & V_{22}' \end{pmatrix}, \tag{14}$$

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 test statistic by Kleibergen and Paap (2006) is based on an orthogonal transformation of S_2 as

$$\Lambda_r = (U_{22}U_{22}')^{-1/2}U_{22}S_2V_{22}'(V_{22}V_{22}')^{-1/2} = A_{r,\perp}'P_\Delta B_{r,\perp}',$$

where $A_{r,\perp}' = (U_{22}U_{22}')^{1/2}(U_{22}')^{-1}[U_{12}'\vdots U_{22}']$ and $B_{r,\perp} = (V_{22}V_{22}')^{1/2}(V_{22}')^{-1}[V_{12}'\vdots V_{22}']$. Unlike S_2, Λ_r is not restricted to be non-negative.[5] 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.

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 estimator 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 $\hat{\lambda}_r = vec(\hat{\Lambda}_r)$, as summarized below.

Proposition 5 *(Kleibergen and Paap, 2006, Theorem 1) Suppose that Assumptions 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 following test statistic called the *rk-statistic*:

$$\text{rk}(r) = N\hat{\lambda}_r'\hat{\Omega}_r^{-1}\hat{\lambda}_r. \tag{15}$$

where $\hat{\Omega}_r$ is a consistent estimator for Ω_r. If the assumptions of Proposition 5 hold, then $\text{rk}(r)$ converges in distribution to a $\chi^2((s-r)(t-r))$ random variable under $H_0 : \text{rank}(P_\Delta) = r$. The nonsingularity assumption on Ω_r can be relaxed by using the M-P inverse as discussed in Section 4.4.

The choice of Δ is left to the researcher. From the perspective of pure identification, if one's goal is to identify as many components as possible, then it is desirable to use a partition that is as fine as possible. From the perspective of estimating $rank(P_\Delta)$ from data, however, using too fine a partition may cause problems because some cells may have few or

---

[5]Robin and Smith (2000) propose a rank statistic based on a consistent estimator of S_2. But, because S_2 is nonnegative, the asymptotic distribution of their estimator of S_2 is not Gaussian when $S_2 = 0$.

no observations. 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.

4.2 Sequential hypothesis testing

Denote the population rank of P_Δ by r_0. To estimate r_0, we sequentially test $H_0 : \text{rank}(P_\Delta) = r$ against $H_1 : \text{rank}(P_\Delta) > r$ 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

$$\hat{r} = \min_{r \in \{0, \ldots, t^*\}} \{r : \text{rk}(i) \geq c^i_{1-\alpha_N}, i = 0, \ldots, r - 1, \text{rk}(r) < c^r_{1-\alpha_N}\}. \tag{16}$$

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 $-N^{-1} \ln \alpha_N = o(1)$.

4.3 Model selection procedure

We also propose a model selection procedure based on the statistic $\text{rk}(r)$ to estimate r_0 consistently. Consider the following criterion function

$$Q(r) = \text{rk}(r) - f(N)g(r),$$

where $g(r)$ is a (possibly stochastic) penalty function. Define

$$\tilde{r} = \arg \min_{1 \leq r \leq t^*} Q(r).$$

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

Proposition 6 *Suppose the conditions of Proposition 5 hold, and $\hat{\Omega}_r$ converges to a nonsingular matrix for any $r \geq r_0$. 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. On the other hand, the AIC is not necessarily consistent and tends to overestimate r_0 with a large sample size.

4.4 The case of multiple variables

So far, our proposed procedures are based on a two-variable test. We now discuss how to extend our method to the case with more than 2 variables.

The proposed approach parallels the way we identify a lower bound on M in k-variate models in Section 2.2. We divide the variables into two groups, construct a matrix from their joint distribution, and examine its rank. Since there is more than one way to divide k variables in two groups, we combine information from different groupings into one test statistic, and test the null hypothesis that all the matrices have rank no larger than r.

Suppose $W = (W_1, \ldots, W_k)'$ with $k \geq 3$ follows the distribution function (1). For simplicity, we assume W_j for $j = 1, \ldots, k$ has a finite support $\mathcal{W}_j = \{1, \ldots, |\mathcal{W}_j|\}$.[6] As in Section 2, group the variables in W into two groups X^α and Y^α, with the grouping index α, and let $\mathcal{X}^\alpha$ and $\mathcal{Y}^\alpha$ denote their support. Let $P_{\mathcal{X}^\alpha \mathcal{Y}^\alpha}$ denote a bivariate probability matrix derived from the joint distribution of X^α and Y^α. We test the null hypothesis that $\mathrm{rank}(P_{\mathcal{X}^\alpha \mathcal{Y}^\alpha}) \leq r$ for all $\alpha \in \mathcal{A}_0$.

Let $\mathcal{A}_0 = \{1, \ldots, |\mathcal{A}_0|\}$ be a set of indices for the α's over which we construct test statistics. It is convenient to assume that all the variables in W are included in the first grouping $\{X^1, Y^1\}$. For instance, we can choose $X^1 = (W_1, \ldots, W_{[k/2]})$ and $Y^1 = (W_{[k/2]+1}, \ldots, W_k)$. Observe that $P_{\mathcal{X}^1 \mathcal{Y}^1}$ contains $|\mathcal{X}^1| \times |\mathcal{Y}^1| = (\prod_{j=1}^{[k/2]} |\mathcal{W}_j|) \times (\prod_{j=[k/2]+1}^{k} |\mathcal{W}_j|) = \prod_{j=1}^{k} |\mathcal{W}_j|$ elements and that the elements of $P_{\mathcal{X}^1 \mathcal{Y}^1}$ exhaust all the possible values of W. Therefore, for every $\alpha \in \mathcal{A}_0$, the elements of the probability matrix $P_{\mathcal{X}^\alpha \mathcal{Y}^\alpha}$ can be expressed as a linear combination of the elements of $P_{\mathcal{X}^1 \mathcal{Y}^1}$, and there exists a matrix Π^α such that $\mathrm{vec}(P_{\mathcal{X}^\alpha \mathcal{Y}^\alpha}) = \Pi^\alpha \mathrm{vec}(P_{\mathcal{X}^1 \mathcal{Y}^1})$.

Define $A_{r,\perp}^\alpha$, $B_{r,\perp}^\alpha$, and λ_r^α analogously to $A_{r,\perp}$, $B_{r,\perp}$, and λ_r in Section 4.1 using $P_{\mathcal{X}^\alpha \mathcal{Y}^\alpha}$ in place of P_Δ. Define $\hat{\lambda}_r^\alpha = \mathrm{vec}((\hat{A}_{r,\perp}^\alpha)' \hat{P}_{\mathcal{X}^\alpha \mathcal{Y}^\alpha} (\hat{B}_{r,\perp}^\alpha)') = (\hat{B}_{r,\perp}^\alpha \otimes (\hat{A}_{r,\perp}^\alpha)') \Pi^\alpha \mathrm{vec}(\hat{P}_{\mathcal{X}^1 \mathcal{Y}^1})$ using the estimators of $P_{\mathcal{X}^1 \mathcal{Y}^1}$, $A_{r,\perp}^\alpha$ and $B_{r,\perp}^\alpha$. To test the null hypothesis that $\mathrm{rank}(P_{\mathcal{X}^\alpha \mathcal{Y}^\alpha}) \leq r$ for all $\alpha \in \mathcal{A}_0$, we stack $\hat{\lambda}_r^\alpha$'s into a vector as $\hat{\lambda}_r(\mathcal{A}_0) = ((\hat{\lambda}_r^1)', \ldots, (\hat{\lambda}_r^{|\mathcal{A}_0|})')'$ 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 4 *Suppose that $\sqrt{N} \mathrm{vec}(\hat{P}_{\mathcal{X}^1 \mathcal{Y}^1} - P_{\mathcal{X}^1 \mathcal{Y}^1}) \to_d N(0, \Sigma_{\mathcal{X}^1 \mathcal{Y}^1})$ and that $\Omega_r(\mathcal{A}_0)$ defined in (17) below is nonsingular. If $\mathrm{rank}(P_{\mathcal{X}^\alpha \mathcal{Y}^\alpha}) \leq r$ 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*

$$\Omega_r(\mathcal{A}_0) = \begin{bmatrix} \Psi^1 \Sigma_{\mathcal{X}^1 \mathcal{Y}^1} (\Psi^1)' & \cdots & \Psi^1 \Sigma_{\mathcal{X}^1 \mathcal{Y}^1} (\Psi^{|\mathcal{A}_0|})' \\ \vdots & \ddots & \vdots \\ \Psi^{|\mathcal{A}_0|} \Sigma_{\mathcal{X}^1 \mathcal{Y}^1} (\Psi^1)' & \cdots & \Psi^{|\mathcal{A}_0|} \Sigma_{\mathcal{X}^1 \mathcal{Y}^1} (\Psi^{|\mathcal{A}_0|})' \end{bmatrix}, \tag{17}$$

---

[6]When W_j is continuously distributed, we may discretize the support of W_j into a finite number of subsets that is strictly larger than the number of components under the null hypothesis.

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

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

$$\text{ave-rk}(r, \mathcal{A}_0) = N(\hat{\lambda}_r(\mathcal{A}_0))'(\hat{\Omega}_r(\mathcal{A}_0))^{-1}\hat{\lambda}_r(\mathcal{A}_0), \tag{18}$$

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 using the inverse of their covariance matrix as the weight. Under the assumptions in Corollary 4, ave-rk$(r, \mathcal{A}_0)$ converges in distribution to a $\chi^2(\nu(\mathcal{A}_0))$ random variable, where $\nu(\mathcal{A}_0) \equiv \sum_{\alpha \in \mathcal{A}_0}(|\mathcal{X}^\alpha| - r)(|\mathcal{Y}^\alpha| - r)$ is the number of elements in $\hat{\lambda}_r(\mathcal{A}_0)$. When the number of variables is very large, however, calculating $\hat{\lambda}^\alpha_r$ for all the possible groupings will become computationally challenging.

$\Omega_r(\mathcal{A}_0)$ is a $\nu(\mathcal{A}_0) \times \nu(\mathcal{A}_0)$ matrix, but its rank cannot be larger than the rank of $\Sigma_{\mathcal{X}^1 \mathcal{Y}^1}$ because all the $\hat{\lambda}^\alpha_r$'s are functions of $\text{vec}(\hat{P}_{\mathcal{X}^1 \mathcal{Y}^1})$. When $|\mathcal{A}_0|$ is very large, $\nu(\mathcal{A}_0)$ may become larger than the rank of $\Sigma_{\mathcal{X}^1 \mathcal{Y}^1}$, and, consequently, the covariance matrix $\Omega_r(\mathcal{A}_0)$ is singular and the assumption of Corollary 4 is violated. In such a case, if $\text{Pr}(\text{rank}(\hat{\Omega}_r(\mathcal{A}_0)) = \text{rank}(\Omega_r(\mathcal{A}_0))) \to 1$, using the M-P inverse of $\hat{\Omega}_r(\mathcal{A}_0)$ in the ave-rk statistic (18) 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 generalized inverse 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

$$\text{ave-rk}^+(r, \mathcal{A}_0) = N(\hat{\lambda}_r(\mathcal{A}_0))'(\hat{\Omega}_{r,c}(\mathcal{A}_0))^+\hat{\lambda}_r(\mathcal{A}_0). \tag{19}$$

The asymptotic distribution of ave-rk$^+(r, \mathcal{A}_0)$ is $\chi^2(J_c)$, where J_c is the number of eigenvalues of $\Omega_r(\mathcal{A}_0)$ which 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 $\Omega_r(\mathcal{A}_0)$.[7]

We also consider alternative statistics that are less subject to the singularity problem. In the average rk-statistic, we stack the rk-statistic $\hat{\lambda}^\alpha_r$ for all $\alpha \in \mathcal{A}_0$ into one large vector $\hat{\lambda}_r(\mathcal{A}_0)$ and take its quadratic form. 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 average rk-statistic ave-rk$(r, \mathcal{A}_j)$ as in (18) but using $\mathcal{A}_j$ in place of $\mathcal{A}_0$. If $\nu(\mathcal{A}_j)$ is not too large, then each ave-rk$(r, \mathcal{A}_j)$ is less subject to the singularity problem than ave-rk$(r, \mathcal{A}_0)$. We then combine the information

---

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

in ave-rk$(r, \mathcal{A}_j)$ for $j = 1, \ldots, K$ into the max-rk and the sum-rk statistics defined as

$$\text{max-rk}(r) = \max_{j=1,\ldots,K} \text{ave-rk}(r, \mathcal{A}_j), \quad \text{sum-rk}(r) = \sum_{j=1,\ldots,K} \text{ave-rk}(r, \mathcal{A}_j). \tag{20}$$

We can apply the sequential hypothesis testing procedure to max-rk(r) and sum-rk(r). While their asymptotic null distributions are not chi-square, they can be easily simulated using the relation $\sqrt{N}\hat{\lambda}_r^\alpha = \hat{\Psi}^\alpha \sqrt{N}(\text{vec}(\hat{P}_{\mathcal{X}^1 \mathcal{Y}^1}) - \text{vec}(P_{\mathcal{X}^1 \mathcal{Y}^1}))$, because it is easy to simulate the asymptotic distribution of $\sqrt{N}(\text{vec}(\hat{P}_{\mathcal{X}^1 \mathcal{Y}^1}) - \text{vec}(P_{\mathcal{X}^1 \mathcal{Y}^1}))$.

If $\hat{\Omega}_r(\mathcal{A}_j)$ is singular for some j, we may also construct a modified average rk-statistic for $\mathcal{A}_j$, denoted by ave-rk$^+(r, \mathcal{A}_j)$, as in (19) using $\mathcal{A}_j$ in place of $\mathcal{A}_0$, and make an inference based on the modified max-rk and the modified sum-rk statistics. By choosing $\mathcal{A}_j$'s so that the degree of freedom $\nu(\mathcal{A}_j)$ is sufficiently small, the modified max-rk and sum-rk statistics would be less sensitive to the choice of c than the modified average rk-statistics ave-rk$^+(r, \mathcal{A}_0)$.

5 Simulation Study

We conduct Monte Carlo simulation experiments to assess the finite sample performance of our proposed procedures for selecting the number of components. We generate samples with normal mixtures and $M = 3$ components. The reported results are based on $1,000$ simulated samples with three different sample sizes: $N = 500, 2000$, and 8000.

In the first experiment, we consider a two-variable normal mixture with three components, thus the distribution function of $W = (W_1, W_2)'$ is $\sum_{m=1}^M \pi^m N(\mu^m, I_2)$, where $\mu^m = (\mu_1^m, \mu_2^m)'$. We experiment with two different parameterizations of μ^m. The first design sets $\mu^1 = (0, 0)'$, $\mu^2 = (1.0, 2.0)'$, and $\mu^3 = (2.0, 1.0)'$, whereas the second design sets $\mu^1 = (0, 0)'$, $\mu^2 = (0.5, 1.0)'$, and $\mu^3 = (1.0, 0.5)'$. Hence, the component distributions of W are further from each other in the first design compared with the second design. In both designs, the mixing probabilities are set to $\pi^1 = \pi^2 = \pi^3 = 1/3$. Regarding the number of partitions, we choose $t = s = 4$ so that we can sequentially test the null hypothesis rank$(P_\Delta) = 1, 2$, and 3. We partition the support of W_i into 4 equiprobable subsets as $\Delta^{w_i} = \{\delta_1^{w_i}, \ldots, \delta_4^{w_i}\}$ so that $\Pr(W_i \in \delta_a^{w_i}) = 1/4$ for $a = 1, \ldots, 4$.

Table 1 reports the result of experiments in which we estimate a lower bound on M by rank(P_Δ) with sequential hypothesis testing (SHT), AIC, BIC, and HQ. The first panel of Table 1 shows the results with the first design. The performance of all the procedures improves as the sample size increases from 500 to 2000, and then to 8000. In SHT, the "optimal" choice of significance level, i.e., α that selects $M = 3$ most frequently, decreases from 0.1 to 0.01 when the sample size increases from $N = 2000$ to 8000. With the sample size of 500 and 2000, the AIC outperforms other statistics. With a larger sample size of 8000, however, the AIC overestimates the number of components and is outperformed by SHT and

HQ, highlighting its inconsistency. The performance of the BIC is the worst among all of the methods even at $N = 8000$; while the BIC is consistent, its finite sample bias is substantial in this setup. Our result shows that the HQ is a better choice than the BIC but the HQ is outperformed by the AIC and SHT when the sample size is small at $N = 500$ and 2000. The second panel of Table 1 reports the results with the second design. The overall performance of our methods is substantially worse than the first design, reflecting the difficulty of estimating the number of components when the component distributions are close to each other.

Next, we consider a four-variable normal mixture with three components. The distribution function of $W = (W_1, \ldots, W_4)'$ is $\sum_{m=1}^{M} \pi^m N(\mu^m, I_4)$, where $\mu^m = (\mu_1^m, \ldots, \mu_4^m)'$. We set $\mu^1 = (0, 0, 0, 0)'$, $\mu^2 = (1.0, 2.0, 0.5, 1.0)'$, and $\mu^3 = (2.0, 1.0, 1.0, 0.5)'$ with mixing probabilities $\pi^1 = \pi^2 = \pi^3 = 1/3$. Thus, (W_1, W_2) and (W_3, W_4) have the same distribution as the first and second design in Table 1, respectively. Following the approach in Sections 4.4, the variables in W are divided into two groups, X^α and Y^α, each containing two variables. There are three different ways to choose 2 variables out of 4, hence $\alpha = \{1, 2, 3\}$. We then estimate the probability matrix $P_{\mathcal{X}^\alpha \mathcal{Y}^\alpha}$ for each α, and construct the average rk-statistic (18). The support of W_i is partitioned into 2 equiprobable subsets, so that the dimension of $P_{\mathcal{X}^\alpha \mathcal{Y}^\alpha}$ is 4×4. For example, when $X^1 = (W_1, W_2)'$ and $Y^1 = (W_3, W_4)'$, then each element of $P_{\mathcal{X}^\alpha \mathcal{Y}^\alpha}$ is given by $\Pr(X^1 \in \delta_a^{w_1} \times \delta_b^{w_2}, Y^1 \in \delta_c^{w_3} \times \delta_d^{w_4})$ for $a, b, c, d = 1$ or 2.

The first panel of Table 2 reports the results with the average rk-statistic (18). When testing the null hypothesis that a lower bound on $M = 1$, the covariance matrix $\Omega_r(\mathcal{A}_0)$ becomes singular, and thus we use the M-P inverse to construct the average rk-statistic; in effect, we use the modified average rk-statistic (19) in which c is equal to machine epsilon. SHT performs better than the model selection procedures across all sample sizes, and the HQ appears to perform the best within the model selection procedures. Note that the four-variable test substantially outperforms the two-variable tests reported in Table 1. The variable (W_3, W_4) may not provide a good signal for separating the components when used alone, but it provides significant additional information when used in conjunction with (W_1, W_2).

The second panel of Table 2 reports the performance of the maximum likelihood estimator (MLE)-based parametric model selection procedure with AIC, BIC and HQ. Each component distribution is correctly specified as a 4-dimensional normal distribution with a diagonal covariance matrix.[8] Our proposed methods outperform the MLE-based model selection procedures when $N = 500$ and are at least comparable to the MLE in other sample sizes. This is somewhat surprising because our selection methods do not use parametric restrictions of the normal mixture model. The relatively poor performance of the MLE-based procedure could be due to the difficulties in estimating the large number of parameters in the normal mixtures.[9] For instance, a four-variable normal mixture model with 3 component distributions

---

[8]We do not implement SHT based on the likelihood ratio statistic here because the likelihood ratio statistic is not asymptotically chi-square distributed.

[9]The maximum likelihood estimates are obtained by numerically maximizing the normal mixture likelihood

has 26 parameters even when the covariance matrix is restricted to be diagonal.

The third panel of Table 2 reports the results using information from only one grouping out of three. Using $\{X^1, Y^1\} = \{(W_1, W_2), (W_3, W_4)\}$ and $\{X^3, Y^3\} = \{(W_1, W_4), (W_2, W_3)\}$ outperforms Table 1, but using $\{X^2, Y^2\} = \{(W_1, W_3), (W_2, W_4)\}$ performs poorly. Thus, using information from all four variables can potentially improve the performance of our selection methods, but it is not clear how to choose the best grouping *a priori* in practice. Further, note that the average rk-statistic in the first panel outperforms any one of the 3 rk-statistics in the third panel. Therefore, it is important to combine information from different groupings for our selection procedures.

In the third experiment, we consider a 3-component normal mixture with 8 variables to examine the max-rk and sum-rk statistics developed in (20). The distribution function of W is given by $F(w_1, \ldots, w_8) = \sum_{m=1}^{3} \pi^m N(\mu^m, I_8)$, where $\mu^m = (\mu_1^m, \ldots, \mu_8^m)'$. We choose $\pi^1 = \pi^2 = \pi^3 = 1/3$, $\mu^1 = (0, 0, 0, 0, 0, 0, 0, 0)'$, $\mu^2 = (0.5, 1.0, 0.25, 0.5, 0.75, 0.25, 1.0, 0.25)'$, and $\mu^3 = (1.0, 0.5, 0.5, 0.25, 0.25, 0.75, 0.25, 1.0)'$. To calculate the max-rk and sum-rk statistics, we first choose 4 variables out of 8. From the 4 chosen variables, we construct the modified average rk-statistic by the procedure for the four-variable model in Table 2, namely, by dividing 4 variables into two bivariate groups and estimating the probability matrix $P_{\mathcal{X}^\alpha \mathcal{Y}^\alpha}$ for three different groupings.[10] These three groupings correspond to $\mathcal{A}_j$ in (20). Since there are $_8C_4 = 70$ ways to choose 4 variables out of 8, there are 70 different rk-statistics. Finally, we combine information from these 70 modified average rk-statistics into the modified max-rk and modified sum-rk statistics defined as in (20) but using ave-rk$^+(r, \mathcal{A}_j)$ in place of ave-rk$(r, \mathcal{A}_j)$.

The first panel of Table 3 reports the performance of SHT with the modified max-rk and sum-rk statistics. Both statistics perform well, but the max-version chose $M = 1$ more frequently when $N = 500$. Overall, the modified sum-rk statistic tends to perform better than the modified max-rk statistic. The second panel of Table 3 reports the mean selection frequencies by the SHT and the AIC/BIC/HQ across 70 different modified average rk-statistics. As shown in the second panel of Table 3, both the max-rk and the sum-rk statistics perform substantially better than individual average rk-statistics. Thus, combining information from different average rk-statistics improves the performance of our procedures.

In the fourth experiment, we employ a challenging setup. The distribution of $W_1, \ldots, W_5$ is the same as the third experiment, but W_6, W_7, and W_8 are set to have identical distributions across sub-populations. Specifically, we set $\mu^1 = (0, 0, 0, 0, 0, 0, 0, 0)'$, $\mu^2 = (0.5, 1.0, 0.25, 0.5, 0.75, 0, 0, 0)'$, and $\mu^3 = (1.0, 0.5, 0.5, 0.25, 0.25, 0, 0, 0)'$. This is a challenging setup in that only five out of

---

using a quasi-Newton method with BFGS updating. For each simulated data set, we use 5 different randomized initial parameter values.

[10]We use the modified average rk-statistic because Ω_r becomes singular when we test the null hypothesis of $M = 1$. We choose $c = 0.01 \times \hat{s}_1$ for the modified average rk-statistic, where $\hat{s}_1$ is the estimated largest singular value of Ω_r.

eight variables can be used to identify the number of components and the researcher does not have any prior knowledge as to which variables should be used. The maximum number of identifiable components depends on the choice of variables included in X^α and Y^α. For example, if either X^α or Y^α contains $\{W_6, W_7, W_8\}$, then $P_{\mathcal{X}^\alpha \mathcal{Y}^\alpha}$ using such a $\{X^\alpha, Y^\alpha\}$ can identify only up to two types.

Table 4 shows the results of the fourth experiment. The performance in Table 4 is generally worse than that in Table 3 because it is more difficult to identify the number of components. On the other hand, both the max-rk and the sum-rk statistics choose the correct M well when $N \geq 2000$. The mean performance of the 70 rk-statistic is much worse than in Table 3. This reflects the lack of power of some choices of $\{X^\alpha, Y^\alpha\}$ discussed above. We also note that, when $N = 500$ and $\alpha = 0.10, 0.05$, the max-rk statistics selects the correct M more often than the sum-rk statistics, but the max-rk statistics also chooses $M = 1$ more frequently than the sum-rk statistics.

6   Examples

Empirical analysis using latent class models involves determining the number of latent classes that are needed to give an adequate description of the data. Choosing the number of latent classes is often a challenge in practice because the parameters of some latent class models are not identifiable without imposing further restrictions. In addition, the likelihood ratio statistic does not have the standard chi-square limiting distribution when applied to testing the number of components. This section provides three illustrative empirical examples, focusing on how to apply our procedures in such cases.

6.1   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, which is originally studied by Clogg (1981) using latent class models. With two variables, the unrestricted latent class model is not identifiable in this case. Clogg estimates the two-class and three-class models using this data by imposing *a priori* restrictions on a set of parameters. On the other hand, as our theoretical analysis shows, a lower bound on the number of latent classes is estimable without imposing any restrictions. For instance, we may test the null hypothesis that the data are generated from the two or the three class models as analyzed by Clogg.

Panel (1) of Table 5 presents the 8×8 table of social mobility in Great Britain taken from Table 1.C of Clogg (1981). Here, occupational categories are: 1=professional and high administrative; 2=managerial and executive; 3=inspectional, supervisory, and other non-manual (high grade); 4=inspectional, supervisory, and other non-manual (low grade); 5=routine grades of nonmanual; 6=skilled manual; 7=semi-skilled manual; 8=unskilled manual. Table

1.B of Clogg (1981) presents the 5×5 table in which categories 2 and 3, categories 5 and 6, and categories 7 and 8 were combined. We apply our procedures to both the 5×5 table and the 8×8 table.

Panel (2) of Table 5 presents the result of the SHT procedure applied to the 5×5 table, rejecting the null hypothesis that the number of latent classes is no more than 4 at any significance level. The AIC/BIC/HQ model selection procedures similarly indicate that the number of latent classes is at least 5 (not reported in Table 5). We further examine the number of latent classes in the 8×8 table starting from the null hypothesis of no more than five classes; the results are presented in Panels (3) and (4) of Table 5. The SHT and the HQ model selection procedures suggest that this intergenerational occupational mobility data could be generated from 6 or 7 latent classes while the BIC and the AIC suggests 5 and 8 latent classes, respectively. Overall, the results of our procedures suggest that there are more than 5 latent classes, rejecting the two or the three class models.

6.2 Types of Trades Started by Different Ethnic Groups in Amsterdam and Rotterdam

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. Van der Heijden, van der Ark, and Mooijaart (2002) study this data, which are presented in Panel (1) of Table 6. There are 6 types of trades and 5 ethnic groups for each of two cities.[11] The types of trade in Panel (1) are 1=wholesale trade; 2=retail trade; 3=producer services; 4=catering and restaurants; 5=personal services. Members of some ethnic groups are more likely to start certain types of trades because of factors such 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 three or not. Panels (2) and (3) of Table 6 show the estimated lower bound on the number of latent classes for Amsterdam and Rotterdam across different procedures. For Amsterdam, the SHT procedure suggests 3 or 4 latent classes, whereas the AIC, BIC, and HQ suggest 4, 2, and 3 latent classes, respectively. For Rotterdam, all of our procedures suggest 3 latent classes. Overall, the results of our procedures agree with the conclusion of van der Heijden et al. (2002).

---

[11]In the original table, there are 8 ethnic groups but we have merged the "Cape Verdeans" and the "Ghanaians" into the "Other" ethnic group because they are relatively small ethnic minorities.

6.3 Response Patterns in Five-item Subsets of LSAT and the Number of Latent Ability Distributions

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 max-rk and sum-rk statistics to this dataset by taking a similar approach to Table 3. Using the notation in Section 4.4, 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.[12] Given a choice of 4 items, we group the 4 items into two bivariate groups X^α and Y^α and estimate the probability matrix $P_{\mathcal{X}^\alpha \mathcal{Y}^\alpha}$. We then construct the average rk-statistic in (18) from the estimates of $P_{\mathcal{X}^\alpha \mathcal{Y}^\alpha}$ for three different groupings.[13] Finally, we construct the max-rk and the sum-rk statistics in (20) from 5 average rk-statistics.

The upper panel of Table 7 reports the results from the sum-rk and max-rk statistics. With these statistics, all of the procedures indicate that there are at least 3 latent ability classes in both LSAT-6 and LSAT-7. The lower panel of Table 7 reports the estimated lower bound on the number of latent ability distributions by SHT and the AIC/BIC/HQ across 5 different choices of 4 items. The selected number of latent classes differs across different choices of 4 items. The results are mixed even within the same choice of 4 items across different procedures; some indicate there are 2 latent classes while others suggest 3.

Acknowledgement

7 Appendix: proofs

7.1 Proof of Proposition 1

The proofs are given in Cohen and Rothblum (1993). Proposition 1.(a),(b), and (c) correspond to Lemma 2.3, Theorem 4.1, and Corollary 4.2 of Cohen and Rothblum (1993).

7.2 Proof of Proposition 2

Since P_Δ is nonsingular, both L_x and L_y are of full rank. It follows that $P_{\Delta,h}(P_\Delta)^{-1} = L_x D_h (L_x)^{-1}$. Because $[P_{\Delta,h}(P_\Delta)^{-1}]L_x = L_x D_h$ and the eigenvalues of $P_{\Delta,h}(P_\Delta)^{-1}$ are distinct, the eigenvalues of $P_{\Delta,h}(P_\Delta)^{-1}$ determine the elements of D_h, whereas its eigenvec-

---

[12]There are 5 ways to choose 4 items out of 5.

[13]There are 3 ways to group 4 items into two bivariate groups.

tors determine the columns of L_x uniquely up to a multiplicative constant. Then, L_x is uniquely determined since the elements of each column of L_x must sum to one. Using an analogous argument, the columns of L_y are uniquely determined from the eigenvectors of $(P_{\Delta,h})'((P_\Delta)')^{-1} = L_y D_h (L_y)^{-1}$. Having determined L_x and L_y, V is uniquely determined as $V = (L_x)^{-1} P_\Delta (L_y)'^{-1}$.

Given L_x, L_y, and V, we may uniquely determine $F_x^m(x)$, $F_y^m(y)$, and $F_z^m(z)$ from $F(x,y,z)$ as follows. For every $x \in \mathcal{X}$, define P_{x,Δ^y} by (13), and let $q_x = (F_x^1(x), \ldots, F_x^M(x))$. Since $P_{x,\Delta^y} = q_x V (L_y)'$, we may uniquely determine $\{F_x^m(x)\}_{m=1}^M$ by $q_x = P_{x,\Delta^y}((L_y)')^{-1} V^{-1}$. Defining $P_{\Delta^x,y}$ and $P_{\Delta^x,z}$ analogously and applying the same argument identifies $\{F_y^m(y)\}_{m=1}^M$ and $\{F_z^m(z)\}_{m=1}^M$ uniquely. $\square$

7.3   Proof of Corollary 2

Because $P_{\Delta,h}(P_\Delta)^{-1} = L_x D_h (L_x)^{-1}$ holds for all h, the set of eigenvectors of $P_{\Delta,1}(P_\Delta)^{-1}, \ldots, P_{\Delta,u}(P_\Delta)^{-1}$ contains the columns of L_x. Therefore, if there are M linearly independent eigenvectors, then L_x is determined uniquely. L_y is uniquely determined by a similar argument. Once L_x and L_y are identified, the rest of the proof follows the proof of Proposition 2. $\square$

7.4   Proof of Proposition 3

We prove part (a) first. Without loss of generality, let $i = 1$ and $j = 2$, and let $F_z^1(\cdot) = F_z^2(\cdot) = A_z(z)$. Then, we can write $F(x,y,z)$ as $F(x,y,z) = [\pi^1 F_x^1(x) F_y^1(y) + \pi^2 F_x^2(x) F_y^2(y)] A_z(z) + \sum_{m=3}^M \pi^m F_x^m(x) F_y^m(y) F_z^m(z)$. For any z such that $A_z(z) > 0$, rearrange this equality as

$$\frac{F(x,y,z) - \sum_{m=3}^M \pi^m F_x^m(x) F_y^m(y) F_z^m(z)}{(\pi^1 + \pi^2) A_z(z)} = \frac{\pi^1}{\pi^1 + \pi^2} F_x^1(x) F_y^1(y) + \frac{\pi^2}{\pi^1 + \pi^2} F_x^2(x) F_y^2(y). \tag{21}$$

The right hand side of (21) takes the form of the distribution function of a bivariate mixture model. Therefore, from Theorem 4.2 of Hall and Zhou (2003), $\{\pi^1/(\pi^1 + \pi^2), \pi^2/(\pi^1 + \pi^2), F_x^1(\cdot), F_y^1(\cdot), F_x^2(\cdot), F_y^2(\cdot)\}$ are not identifiable from the left hand side of (21). Hence, it is not possible to uniquely determine $\{\pi^1, \pi^2, F_x^1(\cdot), F_y^1(\cdot), F_x^2(\cdot), F_y^2(\cdot)\}$ from $F(x,y,z)$ even with the additional knowledge of $\{\pi^m, F_x^m(\cdot), F_y^m(\cdot)\}_{m=3}^M$, and part (a) follows.

We proceed to prove part (b). First, consider the case where Z is discrete with $\mathcal{Z} = \{1, \ldots, |\mathcal{Z}|\}$. Let $\delta_h^z = \{h\}$ for $h = 1, \ldots, |\mathcal{Z}|$. If there are less than M linearly independent eigenvectors in the set of eigenvectors of $P_{\Delta,1}(P_\Delta)^{-1}, \ldots, P_{\Delta,|\mathcal{Z}|}(P_\Delta)^{-1}$, then, in view of $P_{\Delta,h}(P_\Delta)^{-1} = L_x D_h (L_x)^{-1}$ and the property of eigenvectors, there exists a pair of components $i \neq j$ such that $\Pr(Z = h|i) = \Pr(Z = h|j)$ for all h. Therefore, the stated result follows from part (a). When Z is continuous, partition $\mathcal{Z}$ into u mutually exclusive and exhaustive subsets, $\delta_1^z, \ldots, \delta_u^z$. Then, there exists a pair of components $i \neq j$ such that $\Pr(Z \in \delta_h^z|i) = \Pr(Z \in \delta_h^z|j)$ for all h. Since $\delta_1^z, \ldots, \delta_u^z$ can be chosen arbitrarily, part (b)

follows from part (a). $\square$

7.5 Proof of Proposition 4

Because the data are generated by the model and $\mathrm{rank}_+(P_\Delta) = M$, we can factorize P_Δ and $P_{\Delta,h}$ as $P_\Delta = L_x V L_y'$ and $P_{\Delta,h} = L_x V D_h L_y'$.

We prove part (a) first. Because $\mathrm{rank}(P_\Delta) = \mathrm{rank}(P_{\Delta,h}) = M$, we have $\mathrm{rank}(L_x) = \mathrm{rank}(L_y) = \mathrm{rank}(D_h) = M$. Consequently, it follows from the property of the M-P inverse (e.g., Seber (2007, 7.54(d) and 7.65)) that $(P_\Delta)^+ = (L_y')^+ V^{-1} (L_x)^+$ and $(L_x)^+ L_x = (L_y)^+ L_y = I_M$. This gives $P_{\Delta,h}(P_\Delta)^+ = L_x D_h (L_x)^+$. Then, $\hat{M} = \mathrm{rank}(L_x D_h (L_x)^+) \geq \mathrm{rank}(L_x D_h) + \mathrm{rank}(D_h (L_x)^+) - \mathrm{rank}(D_h) = M$ from Frobenius inequality (e.g., Seber (2007, 3.18)) while $\hat{M} = \mathrm{rank}(L_x D_h (L_x)^+) \leq \min\{\mathrm{rank}(L_x), \mathrm{rank}(D_h)\} = M$. Hence, we obtain $\hat{M} = M$. Since $[P_{\Delta,h}(P_\Delta)^+] L_x = L_x D_h$ and the non-zero eigenvalues of $P_{\Delta,h}(P_\Delta)^+$ are distinct, we have $e^m = p_z^m(h)$ for $m = 1, \ldots, M$ and $\hat{L}_x = L_x$. Similarly, $\hat{L}_y = L_y$ follows from $P_{\Delta,h}'(P_\Delta')^+ = L_y D_h (L_y)^+$. Finally, $\hat{V} = V$ follows from $V = (L_x)^+ P_\Delta (L_y')^+$.

Once L_x, L_y, and V are identified, we can uniquely determine $F_x^m(\cdot)$, $F_y^m(\cdot)$, and $F_z^m(\cdot)$ by repeating the argument in the later part of the proof of Proposition 2 but using the M-P inverse of L_x and L_y in place of their inverse.

We proceed to prove part (b). We show $M \leq \mathrm{rank}(P_\Delta)$ first. Recall that M is defined as the smallest positive integer $\tilde{M}$ for which a trivariate finite mixture representation (8) can be found. Suppose $F(x, y, z) = \sum_{m=1}^{\hat{M}} \hat{\pi}^m \hat{F}_x^m(x) \hat{F}_y^m(y) \hat{F}_z^m(z)$ gives a valid trivariate finite mixture representation (8). Then we have $M \leq \hat{M}$ by the definition of M. Further, because $\hat{M} = \mathrm{rank}(P_{\Delta,h}(P_\Delta)^+) \leq \mathrm{rank}((P_\Delta)^+) = \mathrm{rank}(P_\Delta)$, we have $M \leq \mathrm{rank}(P_\Delta)$.

Because the data are generated by the model (8) with M components, $F(x, y) = \sum_{m=1}^M \pi^m F_x^m(x) F_y^m(y)$ gives a valid bivariate finite mixture representation, possibly with a redundant component. Since $\mathrm{rank}_+(P_\Delta)$ is no larger than the smallest positive integer $\tilde{M}$ for which a finite mixture representation of $F(x, y)$ is found, we have $\mathrm{rank}_+(P_\Delta) \leq M$. Further, since $\mathrm{rank}(P_\Delta) \leq \mathrm{rank}_+(P_\Delta)$ from the property of nonnegative rank, we obtain $\mathrm{rank}(P_\Delta) \leq M$. Because we have shown $M \leq \mathrm{rank}(P_\Delta)$ already, $\mathrm{rank}(P_\Delta) = M$ follows. $\square$

7.6 Proof of Corollary 3

Define L_x^α, L_y^α, and D_h^α similar to L_x, L_y, and D_h in Section 3.1 but using $(X^\alpha, Y^\alpha, Z^\alpha)$ and $(\mathcal{X}^\alpha, \mathcal{Y}^\alpha, \mathcal{Z}^\alpha)$ in place of (X, Y, Z) and $(\mathcal{X}, \mathcal{Y}, \mathcal{Z})$. Then, similar to (12), we have $P_{\Delta^\alpha} = L_x^\alpha V (L_y^\alpha)'$ and $P_{\Delta^\alpha, h} = L_x^\alpha V D_h^\alpha (L_y^\alpha)'$. Consequently, by applying the proof of Proposition 2, we can identify π^m's and the component distribution functions of $(X^\alpha, Y^\alpha, Z^\alpha)$. Finally, each of $F_1^m(\cdot)$, ..., $F_k^m(\cdot)$ is obtained from the m-th component distribution of $(X^\alpha, Y^\alpha, Z^\alpha)$ by integrating out the other elements. $\square$.

27

7.7 Proof of Proposition 5

The proof is given by the proof of Theorem 1 in Kleibergen and Paap (2006). $\square$

7.8 Proof of Proposition 6

First, we show $\Pr(\tilde{r} < r_0) \to 0$. If $\tilde{r} < r_0$, this implies $Q(r) < Q(r_0)$ for some $r < r_0$. Thus $\Pr(\tilde{r} < r_0) \leq \sum_{r=1}^{r_0-1} \Pr(Q(r) < Q(r_0))$. Observe that $\Pr(Q(r) < Q(r_0)) = \Pr(\mathrm{rk}(r) - \mathrm{rk}(r_0) - f(N)g(r) + f(N)g(r_0) < 0) = \Pr(N\hat{\lambda}_r'\hat{\Omega}_r^{-1}\hat{\lambda}_r - N\hat{\lambda}_{r_0}'\hat{\Omega}_{r_0}^{-1}\hat{\lambda}_{r_0} + f(N)(g(r_0) - g(r)) < 0)$. For any $r < r_0$, this probability tends to 0 as $N \to \infty$ because $f(N)/N \to 0$, $\hat{\lambda}_r'\hat{\Omega}_r^{-1}\hat{\lambda}_r \to_p \lambda_r'\Omega_r^{-1}\lambda_r > 0$, and $\hat{\lambda}_{r_0}'\hat{\Omega}_{r_0}^{-1}\hat{\lambda}_{r_0} \to_p \lambda_{r_0}'\Omega_{r_0}^{-1}\lambda_{r_0} = 0$.

Second, we show $\Pr(\tilde{r} > r_0) \to 0$. Similarly as above, we have $\Pr(\tilde{r} > r_0) \leq \sum_{r=r_0+1}^{t^*} \Pr(Q(r) < Q(r_0))$ and $\Pr(Q(r) < Q(r_0)) = \Pr(N\hat{\lambda}_r'\hat{\Omega}_r^{-1}\hat{\lambda}_r - N\hat{\lambda}_{r_0}'\hat{\Omega}_{r_0}^{-1}\hat{\lambda}_{r_0} + f(N)(g(r_0) - g(r)) < 0)$. For any $r > r_0$, this probability tends to 0 as $N \to \infty$ because both $N\hat{\lambda}_r'\hat{\Omega}_r^{-1}\hat{\lambda}_r$ and $N\hat{\lambda}_{r_0}'\hat{\Omega}_{r_0}^{-1}\hat{\lambda}_{r_0}$ converge to a chi-square random variable, $f(N) \to \infty$, and $\Pr(g(r_0) - g(r) > 0) \to 1$ as $N \to \infty$. $\square$

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.

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

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

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 Heckman, J. J. (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.

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., 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. (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.

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

de Leeuw, J. and P. G. M. van der Heijden (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.

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.

Goodman, L. A. (1974a) "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.

Goodman, L. A. (1974b). "Exploratory latent structure analysis using both identifiable and unidentifiable models," *Biometrika*, 61, 215-231.

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

Keane, M. P., and Wolpin, K. I. (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.

Koopmans, T. C. and Reiersøl, O. (1950). "The identification of structural characteristics." *The Annals of Mathematical Statistics*, 21, 165-181.

Koopmans, T. C. (ed). (1950). *Statistical Inference in Dynamic Economic Models.* Wiley, New York.

Kruskal, J. B. (1976). "More factors than subjects, tests and treatments: An indeterminacy theorem for canonical decomposition and individual differences scaling." *Psychometrika*, 41, 281-293.

Kruskal, J. B. (1977). "Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics." *Linear Algebra and its Applications*, 18, 95-138.

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.

Lindsay, B. G., Clogg, C. C., and Grego, J. (1991). "Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis." *Journal of the American Statistical Association*, 86, 96-107.

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

Madansky, A. (1960). "Determinantal methods in latent class analysis." *Psychometrika*, 25, 183-198.

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 Oaks: 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.

Seber, G. A. F. (2007). *A Matrix Handbook for Statisticians*. Wiley.

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

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 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.

Wood, G. R. (1999). "Binomial mixtures: geometric estimation of the mixing distribution." *The Annals of Statistics*, 27, 1706-1721.

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: Two Variables

	$N = 500$				$N = 2000$				$N = 8000$			
	$M=1$	$M=2$	$M=3$	$M\geq4$	$M=1$	$M=2$	$M=3$	$M\geq4$	$M=1$	$M=2$	$M=3$	$M\geq4$
$\alpha=0.10$	0.008	0.747	0.211	0.034	0.000	0.345	0.595	0.060	0.000	0.001	0.908	0.091
SHT $\alpha=0.05$	0.020	0.828	0.143	0.009	0.000	0.480	0.491	0.029	0.000	0.003	0.951	0.046
$\alpha=0.01$	0.063	0.892	0.045	0.000	0.000	0.703	0.296	0.001	0.000	0.029	0.963	0.008
AIC	0.004	0.668	0.282	0.046	0.000	0.269	0.623	0.108	0.000	0.001	0.852	0.147
BIC	0.458	0.542	0.000	0.000	0.000	0.961	0.039	0.000	0.000	0.417	0.582	0.001
HQ	0.083	0.849	0.063	0.005	0.000	0.686	0.302	0.012	0.000	0.033	0.932	0.035

Selection frequencies by rk-statistic using (W_1, W_2) with $t = 4$

first design: $\mu^1 = (0,0)'$, $\mu^2 = (1.0, 2.0)'$, $\mu^3 = (2.0, 1.0)'$

	$N = 500$				$N = 2000$				$N = 8000$			
	$M=1$	$M=2$	$M=3$	$M\geq4$	$M=1$	$M=2$	$M=3$	$M\geq4$	$M=1$	$M=2$	$M=3$	$M\geq4$
$\alpha=0.10$	0.746	0.227	0.023	0.004	0.354	0.594	0.044	0.008	0.000	0.901	0.092	0.007
SHT $\alpha=0.05$	0.842	0.147	0.011	0.000	0.480	0.502	0.018	0.000	0.002	0.945	0.050	0.003
$\alpha=0.01$	0.950	0.050	0.000	0.000	0.731	0.266	0.003	0.000	0.016	0.975	0.009	0.000
AIC	0.664	0.298	0.033	0.005	0.252	0.644	0.091	0.013	0.000	0.851	0.135	0.014
BIC	1.000	0.000	0.000	0.000	0.995	0.005	0.000	0.000	0.618	0.382	0.000	0.000
HQ	0.966	0.032	0.002	0.000	0.819	0.179	0.002	0.000	0.037	0.959	0.004	0.000
d.f.	9	4	1	—	9	4	1	—	9	4	1	—

second design: $\mu^1 = (0,0)'$, $\mu^2 = (0.5, 1.0)'$, $\mu^3 = (1.0, 0.5)'$

Notes: The true number of components is $M = 3$. $(W_1, W_2)'$ follows a three-component normal mixture distribution, where each component distribution is $N_2(\mu^m, I_2)$ for $m = 1, 2, 3$. The mixing proportions are $\pi^1 = \pi^2 = \pi^3 = 1/3$ in both designs.

Table 2: Selection Frequencies of the Number of Components: Four Variables

Selection frequencies by average rk-statistic constructed from simultaneously using 3 different groupings

	$N = 500$				$N = 2000$				$N = 8000$			
	$M=1$	$M=2$	$M=3$	$M \geq 4$	$M=1$	$M=2$	$M=3$	$M \geq 4$	$M=1$	$M=2$	$M=3$	$M \geq 4$
SHT $\alpha = 0.10$	0.002	0.173	0.803	0.022	0.000	0.002	0.946	0.052	0.000	0.000	0.942	0.058
ave- $\alpha = 0.05$	0.004	0.209	0.774	0.013	0.000	0.004	0.967	0.029	0.000	0.000	0.966	0.034
rk $\alpha = 0.01$	0.009	0.286	0.705	0.000	0.000	0.014	0.977	0.009	0.000	0.000	0.985	0.015
AIC by ave-rk	0.001	0.184	0.792	0.023	0.000	0.003	0.938	0.059	0.000	0.000	0.936	0.064
BIC by ave-rk	0.400	0.498	0.102	0.000	0.000	0.112	0.887	0.001	0.000	0.001	0.997	0.002
HQ by ave-rk	0.073	0.390	0.537	0.000	0.000	0.039	0.956	0.005	0.000	0.000	0.986	0.014
d.f.	15	12	3	—	15	12	3	—	15	12	3	—

Selection frequencies by MLE-based model selection under parametric multi-dimensional normal distribution

	$N = 500$				$N = 2000$				$N = 8000$			
	$M=1$	$M=2$	$M=3$	$M \geq 4$	$M=1$	$M=2$	$M=3$	$M \geq 4$	$M=1$	$M=2$	$M=3$	$M \geq 4$
AIC by MLE	0.000	0.015	0.371	0.614	0.000	0.000	0.366	0.634	0.000	0.000	0.437	0.563
BIC by MLE	0.003	0.976	0.021	0.000	0.000	0.102	0.897	0.001	0.000	0.000	0.987	0.013
HQ by MLE	0.000	0.569	0.424	0.007	0.000	0.001	0.985	0.014	0.000	0.000	0.984	0.016
d.f.	8	17	26	35	8	17	26	35	8	17	26	35

Selection frequencies by rk-statistic using a single grouping (X^α, Y^α)

$X^1 = (W_1, W_2)$	$N = 500$				$N = 2000$				$N = 8000$			
$Y^1 = (W_3, W_4)$	$M=1$	$M=2$	$M=3$	$M \geq 4$	$M=1$	$M=2$	$M=3$	$M \geq 4$	$M=1$	$M=2$	$M=3$	$M \geq 4$
$\alpha = 0.10$	0.000	0.530	0.415	0.055	0.000	0.028	0.886	0.086	0.000	0.000	0.907	0.093
SHT $\alpha = 0.05$	0.003	0.647	0.328	0.022	0.000	0.053	0.899	0.048	0.000	0.000	0.950	0.050
$\alpha = 0.01$	0.009	0.841	0.146	0.004	0.000	0.162	0.829	0.009	0.000	0.000	0.989	0.011
AIC	0.000	0.424	0.485	0.091	0.000	0.017	0.852	0.131	0.000	0.000	0.843	0.157
BIC	0.204	0.760	0.034	0.002	0.000	0.599	0.398	0.003	0.000	0.000	0.997	0.003
HQ	0.019	0.768	0.200	0.013	0.000	0.146	0.816	0.038	0.000	0.000	0.966	0.034
$X^2 = (W_1, W_3)$	$N = 500$				$N = 2000$				$N = 8000$			
$Y^2 = (W_2, W_4)$	$M=1$	$M=2$	$M=3$	$M \geq 4$	$M=1$	$M=2$	$M=3$	$M \geq 4$	$M=1$	$M=2$	$M=3$	$M \geq 4$
$\alpha = 0.10$	0.032	0.852	0.097	0.019	0.000	0.761	0.206	0.033	0.000	0.380	0.562	0.058
SHT $\alpha = 0.05$	0.073	0.864	0.058	0.005	0.000	0.854	0.137	0.009	0.000	0.524	0.448	0.028
$\alpha = 0.01$	0.195	0.781	0.023	0.001	0.000	0.957	0.042	0.001	0.000	0.739	0.257	0.004
AIC	0.020	0.798	0.153	0.029	0.000	0.683	0.272	0.045	0.000	0.283	0.612	0.105
BIC	0.712	0.287	0.001	0.000	0.001	0.998	0.001	0.000	0.000	0.991	0.009	0.000
HQ	0.222	0.749	0.026	0.003	0.000	0.953	0.043	0.004	0.000	0.774	0.219	0.007
$X^3 = (W_1, W_4)$	$N = 500$				$N = 2000$				$N = 8000$			
$Y^3 = (W_2, W_3)$	$M=1$	$M=2$	$M=3$	$M \geq 4$	$M=1$	$M=2$	$M=3$	$M \geq 4$	$M=1$	$M=2$	$M=3$	$M \geq 4$
$\alpha = 0.10$	0.000	0.589	0.370	0.041	0.000	0.056	0.865	0.079	0.000	0.000	0.905	0.095
SHT $\alpha = 0.05$	0.001	0.705	0.281	0.013	0.000	0.090	0.865	0.045	0.000	0.000	0.948	0.052
$\alpha = 0.01$	0.003	0.868	0.127	0.002	0.000	0.253	0.740	0.007	0.000	0.000	0.990	0.010
AIC	0.000	0.484	0.448	0.068	0.000	0.036	0.844	0.120	0.000	0.000	0.842	0.158
BIC	0.092	0.888	0.020	0.000	0.000	0.721	0.278	0.001	0.000	0.000	0.998	0.002
HQ	0.006	0.808	0.175	0.011	0.000	0.237	0.728	0.035	0.000	0.000	0.962	0.038
d.f.	9	4	1	—	9	4	1	—	9	4	1	—

Notes: The true number of components is $M = 3$. $W = (W_1, W_2, W_3, W_4)'$ follows a three-component normal mixture distribution, where each component distribution is $N_4(\mu^m, I_4)$ for $m = 1, 2, 3$. The parameter values are: $\pi^1 = \pi^2 = \pi^3 = 1/3$, $\mu^1 = (0, 0, 0, 0)'$, $\mu^2 = (1.0, 2.0, 0.5, 1.0)$, and $\mu^3 = (2.0, 1.0, 1.0, 0.5)$.

Table 3: Selection Frequencies of the Number of Components: Eight Variables

	Selection frequencies based on the max or the sum of 70 modified average rk-statistics											
	$N = 500$				$N = 2000$				$N = 8000$			
	$M=1$	$M=2$	$M=3$	$M \geq 4$	$M=1$	$M=2$	$M=3$	$M \geq 4$	$M=1$	$M=2$	$M=3$	$M \geq 4$
max- $\alpha = 0.10$	0.158	0.016	0.826	0.000	0.000	0.000	0.995	0.005	0.000	0.000	0.981	0.019
rk$^+$ $\alpha = 0.05$	0.259	0.025	0.716	0.000	0.000	0.000	0.997	0.003	0.000	0.000	0.990	0.010
$\alpha = 0.01$	0.471	0.028	0.501	0.000	0.001	0.000	0.999	0.000	0.000	0.000	0.997	0.003
sum- $\alpha = 0.10$	0.008	0.130	0.862	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000
rk$^+$ $\alpha = 0.05$	0.013	0.147	0.840	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000
$\alpha = 0.01$	0.019	0.188	0.793	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000
	Mean of the selection frequencies across 70 modified average rk-statistics											
	$N = 500$				$N = 2000$				$N = 8000$			
	$M=1$	$M=2$	$M=3$	$M \geq 4$	$M=1$	$M=2$	$M=3$	$M \geq 4$	$M=1$	$M=2$	$M=3$	$M \geq 4$
mean $\alpha = 0.10$	0.493	0.255	0.251	0.002	0.069	0.287	0.636	0.007	0.000	0.141	0.842	0.017
of 70 $\alpha = 0.05$	0.613	0.210	0.177	0.001	0.108	0.314	0.575	0.003	0.001	0.165	0.827	0.008
SHT's $\alpha = 0.01$	0.799	0.122	0.078	0.000	0.220	0.337	0.442	0.000	0.002	0.216	0.780	0.001
mean of AIC's	0.067	0.689	0.242	0.002	0.015	0.347	0.629	0.009	0.000	0.146	0.834	0.020
mean of BIC's	0.085	0.915	0.000	0.000	0.170	0.815	0.015	0.000	0.065	0.597	0.338	0.000
mean of HQ's	0.110	0.867	0.022	0.000	0.104	0.671	0.225	0.000	0.007	0.346	0.647	0.000
mean of d.f.	11.02	11.99	3.00	—	11.00	12.00	3.00	—	11.00	12.00	3.00	—

Notes: The true number of components is $M = 3$. W follows a three-component normal mixture distribution $\sum_{m=1}^{3} \pi^m N_8(\mu^m, I_8)$. The parameter values are $\pi^1 = \pi^2 = \pi^3 = 1/3$, $\mu^1 = (0,0,0,0,0,0,0,0)'$, $\mu^2 = (0.5, 1.0, 0.25, 0.5, 0.75, 0.25, 1.0, 0.25)'$, $\mu^3 = (1.0, 0.5, 0.5, 0.25, 0.25, 0.75, 0.25, 1.0)'$. The modified rk-statistic with $c = 0.01 \times \hat{s}_1$ is used, where $\hat{s}_1$ is the estimated largest singular value of Ω_r.

Table 4: Selection Frequencies of the Number of Components: Eight Variables, where Three Variables do not have a mixture distribution

	Selection frequencies based on the max or the sum of 70 modified average rk-statistics											
	$N = 500$				$N = 2000$				$N = 8000$			
	$M=1$	$M=2$	$M=3$	$M \geq 4$	$M=1$	$M=2$	$M=3$	$M \geq 4$	$M=1$	$M=2$	$M=3$	$M \geq 4$
max- $\alpha = 0.10$	0.580	0.049	0.371	0.000	0.120	0.009	0.871	0.000	0.000	0.000	0.999	0.001
rk$^+$ $\alpha = 0.05$	0.678	0.050	0.272	0.000	0.175	0.008	0.817	0.000	0.000	0.000	1.000	0.000
$\alpha = 0.01$	0.844	0.034	0.122	0.000	0.366	0.011	0.623	0.000	0.000	0.000	1.000	0.000
sum- $\alpha = 0.10$	0.237	0.484	0.279	0.000	0.009	0.059	0.932	0.000	0.000	0.000	1.000	0.000
rk$^+$ $\alpha = 0.05$	0.264	0.492	0.244	0.000	0.012	0.075	0.913	0.000	0.000	0.000	1.000	0.000
$\alpha = 0.01$	0.328	0.472	0.200	0.000	0.022	0.095	0.883	0.000	0.000	0.000	1.000	0.000
	Mean of selection frequencies across the 70 modified average rk-statistics											
	$N = 500$				$N = 2000$				$N = 8000$			
	$M=1$	$M=2$	$M=3$	$M \geq 4$	$M=1$	$M=2$	$M=3$	$M \geq 4$	$M=1$	$M=2$	$M=3$	$M \geq 4$
mean $\alpha = 0.10$	0.765	0.126	0.108	0.001	0.553	0.171	0.275	0.001	0.251	0.206	0.541	0.002
of 70 $\alpha = 0.05$	0.849	0.086	0.065	0.000	0.646	0.149	0.205	0.000	0.296	0.214	0.489	0.001
SHT's $\alpha = 0.01$	0.945	0.034	0.021	0.000	0.794	0.101	0.105	0.000	0.385	0.223	0.392	0.000
mean of AIC's	0.072	0.825	0.103	0.001	0.073	0.655	0.271	0.001	0.035	0.424	0.539	0.003
mean of BIC's	0.044	0.956	0.000	0.000	0.082	0.917	0.000	0.000	0.125	0.836	0.040	0.000
mean of HQ's	0.069	0.927	0.004	0.000	0.113	0.862	0.025	0.000	0.092	0.689	0.218	0.000
mean of d.f.	11.00	11.99	3.00	—	11.00	12.00	3.00	—	11.00	12.00	3.00	—

Notes: The true number of components is $M = 3$. W follows a three-component normal mixture distribution $\sum_{m=1}^{3} \pi^m N_8(\mu^m, I_8)$. The parameter values are $\pi^1 = \pi^2 = \pi^3 = 1/3$, $\mu^1 = (0,0,0,0,0,0,0,0)'$, $\mu^2 = (0.5, 1.0, 0.25, 0.5, 0.75, 0, 0, 0)'$, $\mu^3 = (1.0, 0.5, 0.5, 0.25, 0.25, 0, 0, 0)'$. The modified rk-statistic with $c = 0.01 \times \hat{s}_1$ is used, where $\hat{s}_1$ is the estimated largest singular value of Ω_r.

Table 5: Intergenerational Social Mobility in Great Britain

(1) British Social Mobility Data (8 × 8 Table)								
Father's	Subject's Status							
Status	1	2	3	4	5	6	7	8
1	50	19	26	8	7	11	6	2
2	16	40	34	18	11	20	8	3
3	12	35	65	66	35	88	23	21
4	11	20	58	110	40	183	64	32
5	2	8	12	23	25	46	28	12
6	12	28	102	162	90	553	230	177
7	0	6	19	40	21	158	143	71
8	0	3	14	32	15	126	91	106

(2) rk-statistics for $H_0 : M = 1, 2, 3, 4$ (5×5 Table)				
The null hypothesis (H_0)	$M = 1$	$M = 2$	$M = 3$	$M = 4$
rk-statistic	557.09	144.64	48.18	15.71
d.f.	16	9	4	1
p-value	0.000	0.000	0.000	0.000

(3) rk-statistics for $H_0 : M = 5, 6, 7$ (8×8 Table)			
The null hypothesis (H_0)	$M = 5$	$M = 6$	$M = 7$
rk-statistic	35.59	12.33	2.27
d.f.	9	4	1
p-value	0.000	0.015	0.132

(4) The Selected Value of a Lower Bound on M (8×8 Table)			
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
Sequential Hypothesis Testing (SHT)	$M = 7$	$M = 7$	$M = 6$
	AIC	BIC	HQ
Model Selection by Information Criteria	$M = 8$	$M = 5$	$M = 6$
No. of Observations		3497	

Notes: The data are from Table 1.C of Clogg (1981). Occupational categories in Panel (1) are: 1=professional and high administrative; 2=managerial and executive; 3=inspectional, supervisory, and other non-manual (high grade); 4=inspectional, supervisory, and other non-manual (low grade); 5=routine grades of nonmanual; 6=skilled manual; 7=semi-skilled manual; 8=unskilled manual. Panel (2) reports the result from the 5 × 5 table in which categories 2 and 3, categories 5 and 6, and categories 7 and 8 were combined.

Table 6: Type of Trade and Ethnic Group data, Amsterdam and Rotterdam

(1) Cross-Classification by Ethnic Group and Type of Trade												
	Amsterdam						Rotterdam					
	Types of Trade						Types of Trade					
Ethnic Group	1	2	3	4	5	Total	1	2	3	4	5	Total
Dutch	382	367	788	113	28	1678	323	209	459	91	153	1235
Turks	14	21	3	8	10	56	29	30	2	15	14	90
Moroccans	12	36	2	5	7	62	8	17	2	13	5	45
Antilleans	8	6	2	1	2	19	5	4	3	4	3	19
Surinamese	44	33	33	17	24	151	35	31	28	19	33	146
Others	208	97	86	26	39	456	82	18	19	16	12	147
Total	668	560	914	170	110	2422	482	309	513	158	220	1682

(2) The values of rk-statistics and the degree of freedom								
	Amsterdam				Rotterdam			
The 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

(3) The Selected Value of a Lower Bound on M						
	Amsterdam			Rotterdam		
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
Sequential Hypothesis Testing	M=4	M=4	M=3	M=3	M=3	M=3
	AIC	BIC	HQ	AIC	BIC	HQ
Model Selection by Information Criteria	M=4	M=2	M=3	M=3	M=3	M=3
No. of Observations	2422			1682		

Notes: The data are from Table 2a of van der Heijden et al. (2002). Types of trade in Panel (1) are 1=wholesale trade; 2=retail trade; 3=producer services; 4=catering and restaurants; 5=personal services.

Table 7: Response Patterns in Five-item Subsets of LSAT and the Estimated Number of Latent Ability Distributions

	Number of Components Selected based on 5 items					
	LSAT 6			LSAT 7		
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
max-rk statistic	M=3	M=3	M=3	M=3	M=3	M=3
sum-rk statistic	M=3	M=3	M=3	M=3	M=3	M=3
	Number of Components Selected based on 4 items					
	LSAT 6			LSAT 7		
SHT	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
$\{W_1, W_2, W_3, W_4\}$	M = 2	M = 2	M = 2	M = 2	M = 2	M = 2
$\{W_2, W_3, W_4, W_5\}$	M = 3	M = 3	M = 2	M = 3	M = 2	M = 2
$\{W_1, W_3, W_4, W_5\}$	M = 3	M = 3	M = 3	M = 2	M = 2	M = 2
$\{W_1, W_2, W_4, W_5\}$	M = 2	M = 2	M = 2	M = 3	M = 3	M = 3
$\{W_1, W_2, W_3, W_5\}$	M = 3	M = 3	M = 3	M = 2	M = 2	M = 2
Model Selection	AIC	BIC	HQ	AIC	BIC	HQ
$\{W_1, W_2, W_3, W_4\}$	M = 2	M = 2	M = 2	M = 2	M = 2	M = 2
$\{W_2, W_3, W_4, W_5\}$	M = 3	M = 2	M = 2	M = 3	M = 2	M = 2
$\{W_1, W_3, W_4, W_5\}$	M = 3	M = 2	M = 2	M = 2	M = 2	M = 2
$\{W_1, W_2, W_4, W_5\}$	M = 2	M = 2	M = 2	M = 3	M = 2	M = 2
$\{W_1, W_2, W_3, W_5\}$	M = 3	M = 2	M = 3	M = 2	M = 2	M = 2
No. of observations	1000			1000		

Notes: The data are from Table 1 of Mislevy (1984).