Nonparametric Identification of Finite Mixture Models

Hiroyuki Kasahara[1] Katsumi Shimotsu[2]

[1]Department of Economics
University of British Columbia,
Hitotsubashi Institute for Advanced Study
Hitotsubashi University

[2]Faculty of Economics
University of Tokyo

Finite Mixture Models

Given a **latent variable** $Z^* \in \{1, 2, \ldots, M\}$,

$$F(\boldsymbol{y}|\boldsymbol{x}) = \sum_{m=1}^{M} \underbrace{\Pr(Z^* = m|\boldsymbol{X} = \boldsymbol{x})}_{:=\lambda^m} \underbrace{F(\boldsymbol{y}|\boldsymbol{x}, Z^* = m)}_{:=F^m(\boldsymbol{y}|\boldsymbol{x})}$$

$$= \sum_{m=1}^{M} \lambda^m(\boldsymbol{x}) F^m(\boldsymbol{y}|\boldsymbol{x})$$

- F is a cumulative distribution function (CDF) of an observed random variable $\boldsymbol{Y}$ conditional on $\boldsymbol{X} = \boldsymbol{x}$.
- The superscript m represents the m-th **component.**
- $\{\lambda^m(\cdot)\}_{m=1}^{M}$ is called the **mixing weights.**
- $\{F^m(\cdot|\cdot)\}_{m=1}^{M}$ is called the **component distributions.**

Non-parametric identification

- The parameter $\boldsymbol{\theta} = \{\{\lambda^m(\boldsymbol{x}), F^m(\boldsymbol{y}|\boldsymbol{x})\}_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{X}\times\mathcal{Y}}\}_{m=1}^{M}$.
- $\boldsymbol{\theta}$ is said to be nonparametrically identified (or identifiable) if it is uniquely determined by the distribution function $F(\boldsymbol{y}|\boldsymbol{x})$ without making any parametric assumption on $\{\lambda^m(\boldsymbol{x}), F^m(\boldsymbol{y}|\boldsymbol{x})\}_{m=1}^{M}$.
- While we don't impose parametric assumptions on $\{\lambda^m(\boldsymbol{x}), F^m(\boldsymbol{y}|\boldsymbol{x})\}_{m=1}^{M}$, we consider various non-parametric assumptions.

Non-parametric identification is important!

- A finite mixture model provides a flexible way to control for unobserved heterogeneity.
- Choosing a parametric family for the component distributions is often difficult because of a lack of guidance from economic theory.
- Even if you estimate a parametric mixture model, understanding the source of non-parametric identification is important.
  - You don't want to rely on parametric form assumption.
  - The identification analysis of parametric finite mixture models becomes transparent once mixing weights and component distributions are nonparametrically identified.

This presentation

- We review different approaches for establishing non-parametric identification.
- We also discuss the identification of the number of components.
- Empirical examples. (In progress: I really appreciate if you suggest me empirical examples!)

Example 1: Clinical tests (Hall and Zhou, 2003)

$$
\begin{aligned}
F(y_1, y_2, \ldots, y_J) &= \lambda F^1(y_1, y_2, \ldots, y_J) + (1-\lambda) F^2(y_1, y_2, \ldots, y_J) \\
&= \lambda \prod_{j=1}^{J} F_j^1(y_j) + (1-\lambda) \prod_{j=1}^{J} F_j^2(y_j)
\end{aligned}
$$

- a patient has a disease ($Z^* = 1$) or not ($Z^* = 2$).
- $\lambda = \Pr(\text{patient has a disease})$
- $\boldsymbol{Y} = (Y_1, \ldots, Y_J)$: outcome of J clinical tests
- Conditional independence assumption (CI):

$$
F(y_1, y_2, \ldots, y_J | Z = z) = \prod_{j=1}^{J} F_j(y_j | Z = z)
$$

We are interested in identifying the **model parameter** $\boldsymbol{\theta} = \left\{ \lambda, \{F_j^1(\cdot)\}_{j=1}^J, \{F_j^2(\cdot)\}_{j=1}^J \right\}$ from $F(y_1, y_2, \ldots, y_J)$.

Example 2: Endogeneity by unobserved ability

The model

$$Y = \alpha(\boldsymbol{X}, U^*) + \beta(\boldsymbol{X}, U^*) T + \varepsilon, \quad \varepsilon \perp\!\!\!\perp T \,|\boldsymbol{X}, U^*.$$

- Y: log-wage, $\quad T$: education
- Unobserved ability $U^* \in \mathcal{U}^* := \{u_1^*, u_2^*, \ldots, u_M^*\}$.
- Two proxies for U^*: U_1 and U_2 (e.g., ASVAB of NLSY79).

Fix $\boldsymbol{X}$. Assume: $(Y, T) \perp\!\!\!\perp U_1 \perp\!\!\!\perp U_2 \mid U^*$.

$$\begin{aligned} &F(y, t, u) \\ &= \sum_{u^* \in \mathcal{U}^*} \Pr(U^* = u^*) \Pr(Y \leq y, T \leq t | u^*) \Pr(U_1 \leq u_1 | u^*) \Pr(U_2 \leq u_2 | u^*) \\ &= \sum_{m=1}^{M} \lambda^m F_{y,t}^m(y, t) F_{u_1}^m(u_1) F_{u_2}^m(u_2). \end{aligned}$$

Example 3: Misclassified and endogenous regressor

$$Y = \alpha(\boldsymbol{X}) + \beta(\boldsymbol{X})T^* + \varepsilon, \quad \varepsilon \perp\!\!\!\perp Z|X, T^*$$

- Y: outcome (e.g., log-wage)
- T^*: true years of education with Corr$(T^*, \varepsilon) \neq 0$
- T: reported years of education
- Z: an instrument for T^* (e.g., college proximity)

Fix $\boldsymbol{X}$. Assume (i) $T \perp\!\!\!\perp Z|T^*, \boldsymbol{X}$ and (ii) $Y \perp\!\!\!\perp Z|T^*, \boldsymbol{X}$.

$$F(y, t, z) = \sum_{t^* \in \mathcal{T}^*} \underbrace{\Pr(T^* = t^*|z)}_{=\lambda^m(z)} \Pr(Y \leq y|t^*) \Pr(T \leq t|t^*)$$

$$= \sum_{m=1}^{M} \lambda^m(z) F_y^m(y) F_t^m(t)$$

Example 4: Dynamic Panel Data Models (Kasahara and Shimotsu, 2009; Hu and Shum, 2012)

- Dynamic panel data: $\{\boldsymbol{y}_i, \boldsymbol{x}_i\}$ for $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})'$ and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iT})'$. T fixed, $N \to \infty$.
- Non-stationarity is allowed.
- $\boldsymbol{X}_i$ is exogenous to latent variable Z_i^* (or assume that $\boldsymbol{X}$ follows the first order Markov process).
- **Y_{it} follows the first order Markov process:**

$$F(\boldsymbol{y}|\boldsymbol{x}) = \sum_{m=1}^{M} \lambda^m F_1^m(y_1|\boldsymbol{x}) \prod_{t=2}^{T} F_t^m(y_t|\{y_s\}_{s=1}^{t-1}, \boldsymbol{x})$$
$$= \sum_{m=1}^{M} \lambda^m F_1^m(y_1|\boldsymbol{x}) \prod_{t=2}^{T} F_t^m(y_t|y_{t-1}, \boldsymbol{x}).$$

Other Examples for Finite Mixtue Models

- Models with unobserved heterogeneity in which multiple proxies for unobserved heterogeneity are observed
- Structural dynamic programming models with unobserved heterogeneity (e.g., Keane and Wolpin (1997))
- Duration models with multiple spells
- Multiple equilibria in discrete game with incomplete information
- Hidden Markov Models (Allman et al., 2009)

Approaches for establishing non-parametric identification of finite mixture models

1. Solving a system of equations: Hall and Zhou (2003) and Hall et al. (2005)
2. Kruskal's theorem: Kruskal (1977), Sidiropoulos and Bro (2000), Allman et al. (2009)
3. Eigen-decomposition: Green (1951), Anderson (1954), Gibson (1955), Leurgans et al. (1993), Chang (1996), De Lathauwer (2006), Hu (2008), Kasahara and Shimotsu (2009), Carroll et al. (2010), Hu and Shum (2012), Bonhomme et al. (2016)
4. Other identifying restrictions: exclusion restrictions, tail conditions, support variation, symmetry.

Hall and Zhou (2003)

Hall and Zhou (2003): Two-components mixture under conditional independence

$$F(y_1, y_2, \dots, y_J) = \lambda \prod_{j=1}^{J} F_j^1(y_j) + (1-\lambda) \prod_{j=1}^{J} F_j^2(y_j)$$

with $\boldsymbol{y} = (y_1, \dots, y_J)' \in \mathcal{Y}^J$ with $\mathcal{Y} = \{1, 2, \dots., |\mathcal{Y}|\}$.

- $|\mathcal{Y}|^J - 1$ restrictions for $1 + J(|\mathcal{Y}| - 1)$ unknowns.
- When $J = 1$, identification is hopeless.
- When $J = 2$, $|\mathcal{Y}|^2 - 1 > 1 + 2(|\mathcal{Y}| - 1)$ when $|\mathcal{Y}| \geq 3$.
- Can we identify the model parameter when $J = 2$?

Hall and Zhou (2003): Non-identification when $J = 2$

The mixture density and their marginal densities are

$$\begin{aligned} f(y_1, y_2) &= \lambda p_1(y_1) p_2(y_2) + (1-\lambda) q_1(y_1) q_2(y_2) \\ f_1(y_1) &= \lambda p_1(y_1) + (1-\lambda) q_1(y_1) \\ f_2(y_2) &= \lambda p_2(y_2) + (1-\lambda) q_2(y_2). \end{aligned}$$

where $p_j := f_j^1$ and $q_j := f_j^2$ for $j = 1, 2$.

Solving for p_1, p_2, q_1, q_2 gives a continuum of solutions indexed by two scalar parameters.

$\Rightarrow$ Non-identification

Hall and Zhou (2003): Identification when $J = 3$

The mixture density and their marginal densities are

$$f(y_1, y_2, y_3) = \lambda p_1(y_1)p_2(y_2)p_3(y_3) + (1-\lambda)q_1(y_1)q_2(y_2)q_3(y_2)$$

$$f(y_j, y_k) = \lambda p_j(y_j)p_k(y_k) + (1-\lambda)q_j(y_j)q_k(y_k)$$

$$\text{for } (j,k) \in \{(1,2),(1,3),(2,3)\}$$

$$f_j(y_j) = \lambda p_j(y_j) + (1-\lambda)q_j(y_j) \quad \text{for } j = 1,2,3$$

We can uniquely solve for p_1, p_2, p_3, q_1, q_2, q_3, λ as functionals of $f(y_1, y_2, y_3)$ under an irreducibility condition.

$\Rightarrow$ Identification

Example 1: Clinical tests (Hall and Zhou, 2003)

$$F(y_1, y_2, \dots, y_J) = \lambda \prod_{j=1}^{J} F_j^1(y_j) + (1-\lambda) \prod_{j=1}^{J} F_j^2(y_j)$$

- a patient has a disease ($Z^* = 1$) or not ($Z^* = 2$).
- $\lambda = \Pr(\text{patient has a disease})$
- $\boldsymbol{Y} = (Y_1, \dots, Y_J)$: outcome of J clinical tests

We can identify $\boldsymbol{\theta} = \left\{\lambda, \{F_j^1(\cdot)\}_{j=1}^J, \{F_j^2(\cdot)\}_{j=1}^J\right\}$ from $F(y_1, y_2, \dots, y_J)$ when $J \geq 3$.

M-components finite mixture models under conditional independence

$$F(y_1, y_2, \ldots, y_J) = \sum_{m=1}^{M} \lambda^m \prod_{j=1}^{J} F_j^m(y_j) \tag{1}$$

- Extending the identification argument of Hall and Zhou (2003) to M-components mixture models is difficult (Hall et al., 2005).
- We may identify (1) by Kruskal's theorem and eigen-decomposition

Notations

- For the sake of clarity, we assume that the support of ***Y*** is discrete.

$$y \in \mathcal{Y} = \{1, 2, \ldots., |\mathcal{Y}|\}.$$

- M-components finite mixture models:

$$P(y_1, y_2, \ldots, y_J) = \sum_{m=1}^{M} \underbrace{\Pr(Z^* = m)}_{:=\lambda^m} \prod_{j=1}^{J} \underbrace{\Pr(Y_j = y_j | Z^* = m)}_{:=p_j^m(y_j)}$$

$$= \sum_{m=1}^{M} \lambda^m \prod_{j=1}^{J} p_j^m(y_j).$$

Kruskal's theorem

Kruskal's theorem

Suppose that $\boldsymbol{y} = (y_1, y_2, y_3)'$ with $y_j \in \{1, 2, \ldots., |\mathcal{Y}_j|\}$.

A tensor representation of the probility mass function:

$$\mathbb{P} = \sum_{m=1}^{M} \lambda^m \boldsymbol{p}_1^m \otimes \boldsymbol{p}_2^m \otimes \boldsymbol{p}_3^m = \sum_{m=1}^{M} \boldsymbol{p}_1^m \otimes \boldsymbol{p}_2^m \otimes (\lambda^m \boldsymbol{p}_3^m),$$

where

$$\boldsymbol{p}_j^m := \begin{bmatrix} p_j^m(1) \\ \vdots \\ p_j^m(|\mathcal{Y}_j|) \end{bmatrix} \quad \text{with} \quad p_j^m(i) := \Pr(Y_j = i | Z^* = m).$$

Define, for $j = 1, 2$,

$$\boldsymbol{L}_j := \begin{bmatrix} \boldsymbol{p}_j^1 & \cdots & \boldsymbol{p}_j^M \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} \lambda^1 \boldsymbol{p}_3^1 & \cdots & \lambda^M \boldsymbol{p}_3^M \end{bmatrix}.$$

Kruskal's theorem

Definition (Kruskal rank)

The Kruskal rank of matrix $\boldsymbol{L}$, denoted by $k_{\boldsymbol{L}}$, is the largest value of positive integer k such that every subset of k columns of the matrix $\boldsymbol{L}$ is linearly independent.

Theorem (Kruskal's Theorem (Kruskal, 1977))

Suppose that

$$k_{\boldsymbol{L}_1} + k_{\boldsymbol{L}_2} + k_{\boldsymbol{D}} \geq 2M + 2. \tag{2}$$

Then, $\boldsymbol{L}_1$, $\boldsymbol{L}_2$, $\boldsymbol{D}$ are uniquely identified from a 3-dimensional tensor $\mathbb{P}$ up to permutation and scaling of columns.

Kruskal's theorem

- Because columns of stochastic matrices sum to 1, $\boldsymbol{\theta} = \{\lambda^m, \boldsymbol{p}_1^m, \boldsymbol{p}_2^m, \boldsymbol{p}_3^m\}_{m=1}^M$ is uniquely identified (Allman et al., 2009).
- Kruskal's sufficient condition

$$k_{\boldsymbol{L}_1} + k_{\boldsymbol{L}_2} + k_{\boldsymbol{D}} \geq 2M + 2$$

  is also necessary when $M = 2$ or 3 but it is not necessary when $M > 3$ (Ten Berge and Sidiropoulos, 2002; Stegeman and Ten Berge, 2006).
- The proof is not constructive.
- Sidiropoulos and Bro (2000) extends the Kruskal's sufficient condition for $J > 3$: $\sum_{j=1}^J k_{\boldsymbol{L}_j} \geq 2M + (J-1)$

Eigen-decomposition

Eigen-decomposition

Consider the following matrix representation:

$$\boldsymbol{P}_k(i,j) = \sum_{m=1}^{M} \lambda^m p_1^m(i) p_2^m(j) p_3^m(k), \quad \boldsymbol{Q}(i,j) = \sum_{m=1}^{M} \lambda^m p_1^m(i) p_2^m(j)$$

so that

$$\boldsymbol{P}_k = \boldsymbol{L}_1 \boldsymbol{D}_k \boldsymbol{\Lambda} (\boldsymbol{L}_2)^\top, \quad \boldsymbol{Q} = \boldsymbol{L}_1 \boldsymbol{\Lambda} (\boldsymbol{L}_2)^\top,$$

with

$$\boldsymbol{L}_\ell := \begin{bmatrix} p_\ell^1(1) & \dots & p_\ell^M(1) \\ \vdots & \ddots & \vdots \\ p_\ell^1(|\mathcal{Y}_\ell|) & \dots & p_\ell^M(|\mathcal{Y}_\ell|) \end{bmatrix}, \quad \boldsymbol{D}_k = \begin{bmatrix} p_3^1(k) & & 0 \\ & \ddots & \\ 0 & & p_3^M(k) \end{bmatrix},$$

$$\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda}) = \begin{bmatrix} \lambda^1 & & 0 \\ & \ddots & \\ 0 & & \lambda^M \end{bmatrix}.$$

Eigen-decomposition

Consider the case where $|\mathcal{Y}_1| = |\mathcal{Y}_2| = M$. Then,

$$\underbrace{\boldsymbol{P}_k}_{\text{observable}} = \boldsymbol{L}_1 \boldsymbol{D}_k \boldsymbol{\Lambda} (\boldsymbol{L}_2)^\top, \qquad \underbrace{\boldsymbol{Q}}_{\text{observable}} = \boldsymbol{L}_1 \boldsymbol{\Lambda} (\boldsymbol{L}_2)^\top$$

Then, when $\boldsymbol{Q}$ is non-singular

$$\underbrace{\boldsymbol{P}_k \boldsymbol{Q}^{-1}}_{\text{observable}} = \boldsymbol{L}_1 \boldsymbol{D}_k \boldsymbol{L}_1^{-1}.$$

The eigenvalues of $\boldsymbol{P}_k \boldsymbol{Q}^{-1}$ identify $\boldsymbol{D}_k$ and the eigenvectors of $\boldsymbol{P}_k \boldsymbol{Q}^{-1}$ identify $\boldsymbol{L}_1$ up to a scaling and permutation.

Eigen-decomposition

Theorem (Eigen-decomposition)

Suppose that

1. $|\mathcal{Y}_1|, |\mathcal{Y}_2| \ \geq M$.
2. *The column vectors of* $\boldsymbol{L}_j = [\boldsymbol{p}_j^1, \cdots, \boldsymbol{p}_j^M]$ *are linearly independent for* $j = 1, 2$
3. *The elements of* $\{p_3^m(k)\}_{m=1}^M$ *are distinct for some* $k \in \{1, \ldots, \kappa_3\}$.

Then, $\boldsymbol{L}_1$, $\boldsymbol{L}_2$, $\boldsymbol{L}_3$, *and* $\boldsymbol{\lambda}$ *are uniquely determined from* $\{\boldsymbol{P}_k\}_{k=1}^{\kappa_3}$ *and* $\boldsymbol{Q}$.

Kruskal's theorem vs. Eigen-decomposition

- For $M = 2$ or 3, Kruskal's theorem provides necessary and sufficient conditions while eigen-decomposition only provides sufficient conditions.
- For $M \geq 4$, they are complementary.
- The proof for Kruskal's theorem is rather inaccessible while the proof for eigen-decomposition is straightforward.
- Eigen-decomposition suggests an explicit algorithm for identification (c.f., simultaneous matrix decomposition).
- Eigen-decomposition is useful for identifying models with dependency as we discuss below.

Generic Identifiability

- Sidiropoulos and Bro (2000)'s sufficient condition $\sum_{j=1}^{J} k_{\boldsymbol{L}_j} \geq 2M + (J - 1)$ implies that the number of identifiable types M increases only linearly with the dimension J of $\boldsymbol{Y}$.
- We can identify more types by considering *generic identifiability* (Allman et al., 2009).
- A property is called generic when it holds everywhere except for a set of Lebesgue measure 0.
- For example, in the set of $J \times J$ matrices, the set of singular matrices has Lebesgue measure 0.
- Using eigen-decomposition, we can show that the number of generically identifiable types M increases exponentially with J.

Generic Identifiability

- Let

$$P(y_1, y_2, \ldots, y_J) = \sum_{m=1}^{M} \lambda^m \prod_{j=1}^{J} p_j^m(y_j),$$

  where $y_j \in \mathcal{Y} = \{1, \ldots, |\mathcal{Y}|\}$.

- Define $\boldsymbol{L}_j$ and $\boldsymbol{\lambda}$ as before.

- Suppose J is odd.

Theorem

Suppose that $|\mathcal{Y}|^{(J-1)/2} \geq M$. Then, $\boldsymbol{L}_1, \ldots, \boldsymbol{L}_J$, and $\boldsymbol{\lambda}$ are generically uniquely identified from $\mathbb{P}$ up to label swapping.

Examples

Example 2: Endogeneity

The model

$$Y = \alpha(\boldsymbol{X}, U^*) + \beta(\boldsymbol{X}, U^*)T + \varepsilon, \quad \varepsilon \perp\!\!\!\perp T \mid \boldsymbol{X}, U^*.$$

- Y: log-wage, T: education
- The unobserved ability $U^* \in \mathcal{U}^* := \{u_1^*, u_2^*, \dots, u_M^*\}$.
- Two test scores: U_1 and U_2 (e.g., ASVAB of NLSY79).

Assume: $(Y, T) \perp\!\!\!\perp U_1 \perp\!\!\!\perp U_2 \mid \boldsymbol{X}, U^*$. Then

$$P(y, t, u|\boldsymbol{x}) = \sum_{m=1}^{M} \lambda^m p_{y,t}^m(y, t|\boldsymbol{x}) p_{u_1}^m(u_1|\boldsymbol{x}) p_{u_2}^m(u_2|\boldsymbol{x}).$$

$\Rightarrow$ We can apply Kruskal's theorem / eigen-decomposition

Applying Kruskal's theorem / eigen-decomposition

The key is to have a mathematical expression of the form:

$$P(x, y, z, w) = \sum_{m=1}^{M} q_x^m(x, w) q_y^m(y, w) q_z^m(z, w).$$

$\Rightarrow$ 3 independent variations within each component.

Regularity conditions for eigen-decomposition:

1. $|\mathcal{X}|, |\mathcal{Y}| \geq M$ and $|\mathcal{Z}| \geq 2$.
2. Full column rank for $\boldsymbol{L}_x = [\boldsymbol{q}_x^1, \ldots, \boldsymbol{q}_x^M]$ and $\boldsymbol{L}_y$
   $\Rightarrow$ e.g., we cannot have $\boldsymbol{q}_x^1 = \pi \boldsymbol{q}_x^2 + (1 - \pi) \boldsymbol{q}_x^3$.
3. For some $k \in \mathcal{Z}$, $q_z^m(k) \neq q_z^{m'}(k)$ for all $m \neq m'$.

Independent variation may come from conditioning variables rather than outcome variables.

Example 3: Misclassified and endogenous regressor

$$Y = \alpha(\boldsymbol{X}) + \beta(\boldsymbol{X})T^* + \varepsilon, \quad \varepsilon \perp\!\!\!\perp Z|X, T^*$$

- Y: outcome (e.g., log-wage)
- T^*: true years of education with $\mathrm{Corr}(T^*, \epsilon) \neq 0$
- T: reported years of education
- Z: an instrument for T^* (e.g., college proximity)

Assume (i) $T \perp\!\!\!\perp Z|T^*, \boldsymbol{X}$ and (ii) $Y \perp\!\!\!\perp Z|T^*, \boldsymbol{X}$.

$$P(y, t|z, \boldsymbol{x}) = \sum_{t^* \in \mathcal{T}^*} \underbrace{\Pr(T^* = t^*|z, \boldsymbol{x})}_{:=\lambda^m(z,\boldsymbol{x})} \Pr(Y = y|t^*, \boldsymbol{x}) \Pr(T = t|t^*, \boldsymbol{x}$$

$$= \sum_{m=1}^{M} \lambda^m(z, \boldsymbol{x}) p_y^m(y|\boldsymbol{x}) p_t^m(t|\boldsymbol{x})$$

$\Rightarrow$ We can apply Kruskal's theorem / eigen-decomposition

Example 3: Misclassified and endogenous regressor

- Suppose that $T \perp\!\!\!\perp Z | T^*$ does not hold, e.g., your incentive to lie about your education qualification depends on college proximity.

$$P(y, t | z, \boldsymbol{x}) = \sum_{m=1}^{M} \lambda^m(\boldsymbol{z}, \boldsymbol{x}) p_y^m(y | \boldsymbol{x}) p_t^m(t | \boldsymbol{z}, \boldsymbol{x})$$

$\Rightarrow$ $\boldsymbol{z}$ is in both λ^m and p_t^m.

$\Rightarrow$ We cannot apply Kruskal's theorem / eigen-decomposition

Example 4: Dynamic Panel Data Models

- Dynamic panel data with $T = 3$: $\{\boldsymbol{y}_i, \boldsymbol{x}_i\}$ for $\boldsymbol{y}_i = (y_{i1}, y_{i2}, y_{i3})'$ and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3})'$.

- **Markovian assumption**

$$P(\boldsymbol{y}|\boldsymbol{x}) = \sum_{m=1}^{M} \lambda^m p_1^m(y_1|\boldsymbol{x}) p_2^m(y_2|y_1, \boldsymbol{x}) p_3^m(y_3|y_2, \boldsymbol{x}).$$

$\Rightarrow$ y_1 is in both p_1^m and p_2^m, and y_2 is in both p_2^m and p_3^m.

$\Rightarrow$ We cannot apply Kruskal's theorem and/or eigen-decomposition.

- Dynamic panel data with $T = 5$: $\{\boldsymbol{y}_i, \boldsymbol{x}_i\}$ for $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i5})'$ and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{i5})'$.
- Fix $y_2 = \bar{y}_2$ and $y_4 = \bar{y}_4$. Fix and drop $\boldsymbol{x}$.

$$P(\boldsymbol{y}) = \sum_{m=1}^{M} \lambda^m \underbrace{p_1^m(y_1) p_2^m(\bar{y}_2|y_1)}_{=p_{12}^m(y_1, \bar{y}_2)} \underbrace{p_3^m(y_3|\bar{y}_2) p_4^m(\bar{y}_4|y_3)}_{=p_{34}^m(y_3, \bar{y}_4|\bar{y}_2)} p_5^m(y_5|\bar{y}_4)$$

$$= \sum_{m=1}^{M} \lambda^m p_{12}^m(y_1, \bar{y}_2) p_{34}^m(y_3, \bar{y}_4|\bar{y}_2) p_5^m(y_5|\bar{y}_4)$$

$\Rightarrow$ We can apply Kruskal's theorem / eigen-decomposition to establish identification

Williams (2018)

- Dynamic panel data with $T = 3$: $\{\boldsymbol{y}_i, \boldsymbol{x}_i\}$.
- Suppose that $\boldsymbol{X}_i = (\tilde{\boldsymbol{X}}_i', \boldsymbol{V}_i)'$ with $\boldsymbol{V}_i = (V_{i1}, \dots, V_{iT})'$.
- $V_{it} \perp\!\!\!\perp Z_i^* | (V_{i1}, \dots, V_{it-1}), \tilde{\boldsymbol{X}}_i$.

$$P(\boldsymbol{y}|\tilde{\boldsymbol{x}}, \boldsymbol{v}_i) = \sum_{m=1}^{M} \lambda^m p_1^m(y_1, \tilde{\boldsymbol{x}}, v_1) p_2^m(y_2|y_1, \tilde{\boldsymbol{x}}, v_2) p_3^m(y_3|y_2, \tilde{\boldsymbol{x}}, v_3)$$

$\Rightarrow$ We can apply Kruskal's theorem / eigen-decomposition to establish identification

Hu and Shum (2012)

- Dynamic panel data with $T = 4$ with discrete support.
- Fix $(y_2, y_3) \in \{(\bar{y}_2, \bar{y}_3), (y_2^\dagger, y_3^\dagger), (y_2^\dagger, \bar{y}_3), (\bar{y}_2, y_3^\dagger)\}$.

$$p(y_1, \bar{y}_2, \bar{y}_3, y_4) = \sum_{m=1}^{M} \lambda^m \underbrace{p_{12}^m(y_1, \bar{y}_2)}_{\bar{y}_3 \text{ is excluded}} p_3^m(\bar{y}_3|\bar{y}_2) \underbrace{p_4^m(y_4|\bar{y}_3)}_{\bar{y}_2 \text{ is excluded}} .$$

Then,

$$\boldsymbol{P}_{\bar{y}_2,\bar{y}_3} = \boldsymbol{L}_{1,\bar{y}_2} \boldsymbol{D}_{\bar{y}_2,\bar{y}_3} \boldsymbol{\Lambda} \left(\boldsymbol{L}_{2,\bar{y}_3}\right)^\top ,$$

with

$$\boldsymbol{L}_{1,\bar{y}_2} := \begin{bmatrix} p_{12}^1(1, \bar{y}_2) & \dots & p_{12}^M(1, \bar{y}_2) \\ \vdots & \ddots & \vdots \\ p_{12}^1(M, \bar{y}_2) & \dots & p_{12}^M(M, \bar{y}_2) \end{bmatrix}, \boldsymbol{D}_{\bar{y}_2,\bar{y}_3} = \begin{bmatrix} p_3^1(\bar{y}_3|\bar{y}_2) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & p_3^M(\bar{y}_3|\bar{y}_2) \end{bmatrix} \text{ etc.}$$

Hu and Shum (2012)

Evaluating at $(y_2, y_3) \in \{(\bar{y}_2, \bar{y}_3), (y_2^\dagger, y_3^\dagger), (y_2^\dagger, \bar{y}_3), (\bar{y}_2, y_3^\dagger)\}$,

$$\boldsymbol{P}_{\bar{y}_2,\bar{y}_3} = \boldsymbol{L}_{1,\bar{y}_2} \boldsymbol{D}_{\bar{y}_2,\bar{y}_3} \boldsymbol{\Lambda} \left(\boldsymbol{L}_{2,\bar{y}_3}\right)^\top, \quad \boldsymbol{P}_{y_2^\dagger,y_3^\dagger} = \boldsymbol{L}_{1,y_2^\dagger} \boldsymbol{D}_{y_2^\dagger,y_3^\dagger} \boldsymbol{\Lambda} \left(\boldsymbol{L}_{2,y_3^\dagger}\right)^\top,$$

$$\boldsymbol{P}_{y_2^\dagger,\bar{y}_3} = \boldsymbol{L}_{1,y_2^\dagger} \boldsymbol{D}_{y_2^\dagger,\bar{y}_3} \boldsymbol{\Lambda} \left(\boldsymbol{L}_{2,\bar{y}_3}\right)^\top, \quad \boldsymbol{P}_{\bar{y}_2,y_3^\dagger} = \boldsymbol{L}_{1,\bar{y}_2} \boldsymbol{D}_{\bar{y}_2,y_3^\dagger} \boldsymbol{\Lambda} \left(\boldsymbol{L}_{2,y_3^\dagger}\right)^\top.$$

Then,

$$\begin{aligned} & \boldsymbol{P}_{\bar{y}_2,\bar{y}_3} \left(\boldsymbol{P}_{y_2^\dagger,\bar{y}_3}\right)^{-1} \boldsymbol{P}_{y_2^\dagger,y_3^\dagger} \left(\boldsymbol{P}_{\bar{y}_2,y_3^\dagger}\right)^{-1} \\ &= \boldsymbol{L}_{1,\bar{y}_2} \left[\boldsymbol{D}_{\bar{y}_2,\bar{y}_3} (\boldsymbol{D}_{y_2^\dagger,\bar{y}_3})^{-1} \boldsymbol{D}_{y_2^\dagger,y_3^\dagger} (\boldsymbol{D}_{\bar{y}_2,y_3^\dagger})^{-1}\right] (\boldsymbol{L}_{1,\bar{y}_2})^{-1}. \end{aligned}$$

$\Rightarrow$ We may apply eigen-decomposition to identify $\boldsymbol{L}_{1,\bar{y}_2}$ and $\boldsymbol{D}_{\bar{y}_2,\bar{y}_3} (\boldsymbol{D}_{y_2^\dagger,\bar{y}_3})^{-1} \boldsymbol{D}_{y_2^\dagger,y_3^\dagger} (\boldsymbol{D}_{\bar{y}_2,y_3^\dagger})^{-1}$.

Identification argument in Hu and Shum (2012) and Carroll et al. (2010)

The key is to have a mathematical expression of the form:

$$P(x, y, z, v) = \sum_{m=1}^{M} q_{zv}^m(z, v) \underbrace{q_x^m(x, z)}_{v \text{ is excluded}} \underbrace{q_y^m(y, v)}_{z \text{ is excluded}} .$$

2 independent variation (z and y) and some exclusion restrictions for other variables (z and v).

Evaluating at $(z, v) \in \{(\bar{z}, \bar{v}), (z^\dagger, v^\dagger), (\bar{z}, v^\dagger), (z^\dagger, \bar{v})\}$:

$$\boldsymbol{P}_{\bar{z},\bar{v}} = \boldsymbol{L}_{1,\bar{z}} \boldsymbol{D}_{\bar{z},\bar{v}} \boldsymbol{\Lambda} \left(\boldsymbol{L}_{2,\bar{v}}\right)^\top, \quad \boldsymbol{P}_{z^\dagger,v^\dagger} = \boldsymbol{L}_{1,z^\dagger} \boldsymbol{D}_{z^\dagger,v^\dagger} \boldsymbol{\Lambda} \left(\boldsymbol{L}_{2,v^\dagger}\right)^\top,$$

$$\boldsymbol{P}_{\bar{z},v^\dagger} = \boldsymbol{L}_{1,\bar{z}} \boldsymbol{D}_{\bar{z},v^\dagger} \boldsymbol{\Lambda} \left(\boldsymbol{L}_{2,v^\dagger}\right)^\top, \quad \boldsymbol{P}_{z^\dagger,\bar{v}} = \boldsymbol{L}_{1,z^\dagger} \boldsymbol{D}_{z^\dagger,\bar{v}} \boldsymbol{\Lambda} \left(\boldsymbol{L}_{2,\bar{v}}\right)^\top,$$

Apply eigen-decomposition to

$$\boldsymbol{P}_{\bar{z},\bar{v}} \left(\boldsymbol{P}_{z^\dagger,\bar{v}}\right)^{-1} \boldsymbol{P}_{z^\dagger,v^\dagger} \left(\boldsymbol{P}_{\bar{z},v^\dagger}\right)^{-1} .$$

Example 3: Misclassified and endogenous regressor

$$Y = \alpha(\boldsymbol{X}) + \beta(\boldsymbol{X})T^* + \varepsilon, \quad \varepsilon \perp\!\!\!\perp Z|X, T^*$$

- Recall that, if $T \perp\!\!\!\perp Z|T^*$ does not hold,

$$P(y, t|z, \boldsymbol{x}) = \sum_{m=1}^{M} \lambda^m(\boldsymbol{z}, \boldsymbol{x}) p_y^m(y|\boldsymbol{x}) p_t^m(t|\boldsymbol{z}, \boldsymbol{x})$$

$\Rightarrow$ $\boldsymbol{z}$ is in both λ^m and p_t^m.

$\Rightarrow$ We <u>cannot</u> apply Kruskal's theorem / eigen-decomposition

Example 3: Misclassified and endogenous regressor

- Now, suppose that $\boldsymbol{X} = (V, \tilde{X})$ and V that does not affect an incentive to lie (e.g., $V =$ gender). Fix $\tilde{X}$.

$$p(y, t|z, v, \tilde{x}) = \sum_{m=1}^{M} \lambda^m(z, v, \tilde{x}) \underbrace{p_y^m(y|v, \tilde{x})}_{z \text{ is excluded}} \underbrace{p_t^m(t|z, \tilde{x})}_{v \text{ is excluded}}.$$

Evaluating at $(z, v) \in \{(\bar{z}, \bar{v}), (z^\dagger, v^\dagger), (\bar{z}, v^\dagger), (z^\dagger, \bar{v})\}$

$\Rightarrow$ We can establish identification using the argument in Hu and Shum (2012) and Carroll et al. (2010) (Kasahara and Shimotsu, on-going project).

Identification of the Number of Components

Identification of the Number of Components (Kasahara and Shimotsu, 2009, 2014)

- M-components finite mixture models with $J = 2$:

$$P(x,y) = \sum_{m=1}^{M} \lambda^m p_x^m(x) p_y^m(y).$$

$\Rightarrow$ When $J = 2$, the mixture model is not identified (Hall and Zhou, 2003).

- Can we identify the number of components M?

Identification of the Number of Components

Collect the distribution of (X, Y) to a matrix:

$$\underset{(|\mathcal{X}|\times|\mathcal{Y}|)}{\boldsymbol{Q}} = \begin{bmatrix} \Pr(X=1, Y=1) & \cdots & \Pr(X=1, Y=|\mathcal{Y}|) \\ \vdots & \ddots & \vdots \\ \Pr(X=|\mathcal{X}|, Y=1) & \cdots & \Pr(X=|\mathcal{X}|, Y=|\mathcal{Y}|) \end{bmatrix}.$$

Define

$$\underset{(|\mathcal{X}|\times 1)}{\boldsymbol{p}_x^m} = (\Pr(X=1|Z^*=m), \ldots, \Pr(X=|\mathcal{X}||Z^*=m))',$$

$$\underset{(|\mathcal{Y}|\times 1)}{\boldsymbol{p}_y^m} = (\Pr(Y=1|Z^*=m), \ldots, \Pr(Y=|\mathcal{Y}||Z^*=m))'.$$

Then $\boldsymbol{Q}$ can be expressed as, for some $\widetilde{M}$,

$$\boldsymbol{Q} = \sum_{m=1}^{\tilde{M}} \lambda^m \boldsymbol{p}_x^m (\boldsymbol{p}_y^m)', \quad \boldsymbol{p}_x^m, \boldsymbol{p}_y^m \geq 0, \quad \lambda^m > 0, \quad \sum_{m=1}^{\tilde{M}} \lambda^m = 1.$$

Lower bound of the number of components

- Define the number of components in $\boldsymbol{Q}$, denoted by M, as the smallest integer $\tilde{M}$ such that the above finite mixture representation is possible.

- $M = \text{rank}_+(\boldsymbol{Q})$, i.e., the nonnegative rank of $\boldsymbol{Q}$

- For a nonnegative matrix A, its nonnegative rank ($\text{rank}_+(A)$) is the smallest number of nonnegative rank-one matrices such that A equals their sum.

Relation between rank and nonnegative rank

Proposition (Cohen and Rothblum (1993))

1. *rank*$(\boldsymbol{Q}) \leq M \leq \min\{|\mathcal{X}|, |\mathcal{Y}|\}$.
2. *If rank*$(\boldsymbol{Q}) \leq 2$*, then* $M =$ *rank*$(\boldsymbol{Q})$.
3. *If* $|\mathcal{X}| \leq 3$ *or* $|\mathcal{Y}| \leq 3$*, then* $M =$ *rank*$(\boldsymbol{Q})$.

Therefore,

$$\text{rank}(\boldsymbol{Q}) = M \quad \text{if} \quad M \leq 3.$$

In general, for $M \geq 4$,

$$\text{rank}(\boldsymbol{Q}) \leq \text{M}$$

Why does rank($\boldsymbol{Q}$) identify a lower bound?

Singular value decomposition of $\boldsymbol{Q}$ gives a representation:

$$\boldsymbol{Q} = \sum_{m=1}^{\widetilde{M}} \tilde{\lambda}^m \tilde{\boldsymbol{p}}_x^m (\tilde{\boldsymbol{p}}_y^m)',$$

- $\{\tilde{\lambda}^m\}_{m=1}^{M}$: non-zero singular values.
- $\{\tilde{\boldsymbol{p}}_x^m\}$ and $\{\tilde{\boldsymbol{p}}_y^m\}$: left- and right- singular vectors.
- $\widetilde{M} = \text{rank}(\boldsymbol{Q})$: the # of non-zero singular values.

Some elements of $\{\tilde{\boldsymbol{p}}_x^m\}$ and $\{\tilde{\boldsymbol{p}}_y^m\}$ may be negative.

$\Rightarrow$ the # of components M $>$ rank($\boldsymbol{Q}$).

Estimating the number of components

- Determining the nonnegative rank of a matrix is computationally difficult (NP-hard).
- Testing the rank of Q via the singular value decomposition (Kasahara and Shimotsu, 2014).
- Testing the number of components for parametric finite mixture models: not easy, but many existing papers, including Kasahara and Shimotsu (2015).
- Little existing work on testing the number of components for finite mixture models without imposing parametric assumption on components.

Other topics on identification of mixture models

- Models with continuous variables / continuous mixtures (Hu and Schennach, 2008; Allman et al., 2009; Hu and Shum, 2012)).
- Other identifying strategies:
    - exclusion restrictions: Compiani and Kitamura (2016), Henry, Kitamura, and Salanie (2014, QE)
    - tail conditions: Kitamura (2003), Henry, Kitamura, and Salanie (2010, working paper), Hohmann and Holzmann (2015), Jochmans, Henry, and Salanie (2017, ET)
    - support variation: D'Haultfœuille and Février (2015)
    - symmetry: Bordes, Mottelet, and Vandekerkhove (2006, AS), Hunter, Wang, Hettmansperger (2007, AS)

References

Allman, E. S., Matias, C., and Rhodes, J. A. (2009), "Identifiability of Latent Class Models with Many Observed Variables," *Annals of Statistics*, 37, 3099–3132.

Anderson, T. W. (1954), "On Estimation of Parameters in Latent Structure Analysis," *Psychometrika*, 19, 1–10.

Bonhomme, S., Jochmans, K., and Robin, J.-M. (2016), "Nonparametric Estimation of Finite Mixtures from Repeated Measurements," *Journal of the Royal Statistical Society Series B*, 78, 211–229.

Carroll, R. J., Chen, X., and Hu, Y. (2010), "Identification and Estimation of Nonlinear Models Using Two Samples with Nonclassical Measurement Errors," *Journal of Nonparametric Statististics*, 22, 379–399.

Chang, J. T. (1996), "Full Reconstruction of Markov Models on

Evolutionary Trees: Identifiability and Consistency," *Mathematical Biosciences*, 137, 51–73.

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

De Lathauwer, L. (2006), "A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization," *SIAM Journal on Matrix Analysis and Applications*, 28, 642–666.

Gibson, W. A. (1955), "An Extension of Anderson's Solution for the Latent Structure Equations," *Psychometrika*, 20, 69–73.

Green, B. F. (1951), "A General Solution for the Latent Class Model of Latent Structure Analysis," *Psychometrika*, 16, 151–166.

Hall, P., Neeman, A., Pakyari, R., and Elmore, R. T. (2005), "Nonparametric Inference in Multivariate Mixtures," *Biometrika*, 667–678.

Hall, P. and Zhou, X. H. (2003), "Nonparametric Estimation of Component Distributions in a Multivariate Mixture," *Annals of Statistics*, 201–224.

Hu, Y. (2008), "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution," *Journal of Econometrics*, 144, 27–61.

Hu, Y. and Schennach, S. (2008), "Instrumental Variable Treatment of Non-classical Measurement Error Models," *Econometrica*, 76, 195–216.

Hu, Y. and Shum, M. (2012), "Nonparametric Identification of Dynamic Models with Unobserved State Variables," *Journal of Econometrics*, 171, 32–44.

Kasahara, H. and Shimotsu, K. (2009), “Nonparametric identification of finite mixture models of dynamic discrete choices,” *Econometrica*, 77, 135–175.

— (2014), “Nonparametric Identification and Estimation of the Number of Components in Multivariate Mixtures,” *Journal of the Royal Statistical Society Series B*, 76, 97–111.

— (2015), “Testing the Number of Components in Normal Mixture Regression Models,” *Journal of the American Statistical Association*, 110, 1632–1645.

Keane, M. P. and Wolpin, K. I. (1997), “The Career Decisions of Young Men,” *Journal of Political Economy*, 105, 473–522.

Kruskal, J. B. (1977), “Three-Way Arrays: Rank and Uniqueness of Trillinear Decompositions, with Application to Arithmetic Complexity and Statistics,” *Linear Algebra and its Applications*, 18, 95–138.

Leurgans, S. E., Ross, R. T., and Abel, R. B. (1993), "A Decomposition for Three-way Arrays," *SIAM Journal on Matrix Analysis and Applications*, 14, 1064–1083.

Sidiropoulos, N. and Bro, R. (2000), "On the uniqueness of multilinear decomposiion of N-way arrays," *Journal of Chemometrics*, 14, 229–239.

Stegeman, A. and Ten Berge, J. (2006), "Kruskal's condition for uniqueness in Candecomp/Parafac when ranks and k-ranks coincide," *Computational Statistics and Data Analysis*, 50, 210–220.

Ten Berge, J. M. and Sidiropoulos, N. D. (2002), "On Uniqueness in CANDECOMP/PARAFAC," *Psychometrka*, 67, 399–409.

Williams, B. D. (2018), "Nonparametric Identification of Binary

Choice Models with Lagged Dependent Variables," George Washington University.