Identifiability of Finite Mixture Models

Jang SCHILTZ (University of Luxembourg)

joint work with

Cédric NOEL (University of Lorraine & University of Luxembourg)

SMTDA 2020
June 4, 2020
1 Introduction
Outline

1. Introduction

2. Characterization of Identifiability
1 Introduction

2 Characterization of Identifiability

3 Identifiability of finite mixture models
Outline

1 Introduction

2 Characterization of Identifiability

3 Identifiability of finite mixture models
Identifiability

Definition: A finite mixture model is identifiable if a given dataset leads to a uniquely determined set of model parameter estimations up to a permutation of the clusters.

Identifiability of the parameters is a necessary condition for the existence of consistent estimators for any statistical model. Without identifiability, there might be several solutions for the parameter estimation problem.
Identifiability

**Definition:**
A finite mixture model is identifiable if a given dataset leads to a uniquely determined set of model parameter estimations up to a permutation of the clusters.

Identifiability of the parameters is a necessary condition for the existence of consistent estimators for any statistical model.

Without identifiability, there might be several solutions for the parameter estimation problem.
Teicher (1963): The class of all mixtures of one-dimensional normal distributions is identifiable.
Literature Review

- Teicher (1963): The class of all mixtures of one-dimensional normal distributions is identifiable.

- Yakowitz and Spragins (1968): Extension to the class of all Gaussian mixtures.
Literature Review

- Teicher (1963): The class of all mixtures of one-dimensional normal distributions is identifiable.

- Yakowitz and Spragins (1968): Extension to the class of all Gaussian mixtures.

Finite Mixture Models

Data:
Variable of interest $Y_i = y_{i1}, y_{i2}, \ldots, y_{iT}$
Covariants $x_1, \ldots, x_M$ and $z_{i1}, \ldots, z_{iT}$
$a_{it}$ age of subject $i$ at time $t$
Finite Mixture Models

Data:
Variable of interest $Y_i = y_{i1}, y_{i2}, \ldots, y_{iT}$
Covariants $x_1, \ldots, x_M$ and $z_{i1}, \ldots, z_{iT}$

$a_{it}$ age of subject $i$ at time $t$

Model:

$K$ groups of size $\pi_k$ with trajectories

$$y_{it} = \sum_{j=0}^{s_k} \left( \beta^k_j + \sum_{m=1}^{M} \alpha^k_m x_m + \gamma^k_j z_{it} \right) a^j_{it} + \varepsilon^k_{it},$$

where $\varepsilon^k_{it} \sim \mathcal{N}(0, \sigma^k)$.
Outline

1 Introduction

2 Characterization of Identifiability

3 Identifiability of finite mixture models
Notations

Distribution $f$ of a finite mixture model:

$$f(y_i; \Omega) = \sum_{k=1}^{K} \pi_k g_k(y_i; \theta^k).$$

Cumulative distribution function $F$ of a finite mixture model:

$$F(y_i; \Omega) = \sum_{k=1}^{K} \pi_k G_k(y_i; \theta^k).$$
Mixtures and mixing distributions

Let $\mathcal{F} = \{ F(y; \omega), \ y \in \mathbb{R}^T, \ \omega \in \mathbb{R}^{s+2}_K \}$ be a family of $T$-dimensional cdf’s indexed by a parameter set $\omega$, such that $F(y; \omega)$ is measurable in $\mathbb{R}^T \times \mathbb{R}^{s+2}_K$.

The the $s + 2$-dimensional cdf $H(x) = \int_{\mathbb{R}^K} F(y; \omega) dG(\omega)$ is the image of the above mapping, of the $s + 2$-dimensional cdf $G$.

The distribution $H$ is called the mixture of $\mathcal{F}$ and $G$ its mixing distribution.

Let $\mathcal{G}$ denote the class of all $s + 2$-dimensional cdf’s $G$ and $\mathcal{H}$ the induced class of mixtures $H$.

Then $\mathcal{H}$ is identifiable if $Q$ is a one-to-one map from $\mathcal{G}$ onto $\mathcal{H}$.
Characterization of identifiability

The set $\mathcal{H}$ of all finite mixtures of class $\mathcal{F}$ of distributions is the convex hull of $\mathcal{F}$.

$$\mathcal{H} = \left\{ H(y) : H(y) = \sum_i c_i F(y, \omega_i), \; c_i > 0, \sum_i c_i = 1, \; F(y, \omega_i) \in \mathcal{F} \right\}. \quad (2)$$

**Theorem**

A necessary and sufficient condition for the class $\mathcal{H}$ of all finite mixtures of the family $\mathcal{F}$ to be identifiable is that $\mathcal{F}$ is a linearly independent family over the field of real numbers.
The Model

\[ Y_{it} = f(a_{it}; \beta^k, \delta^k) + \varepsilon_{it} = \beta^k A_{it} + \delta^k W_{it} + \varepsilon_{it}. \]  

(3)

We can write

\[ \mathcal{L}((Y_i)_{i \in I}) = \bigotimes_{i \in I} F_{A_i, W_i, J}. \]  

(4)

Identifiability of a model means that knowing the data distribution \( \mathcal{L}(Y_i), i \in I \), one can uniquely identify the mixing distribution \( J \).

That is, no two distinct sets of parameters lead to the same data distribution.
Nagin’s base model

\[ C_1 = \left( F_{A,J} : F_{A,J} = \bigotimes_{i \in I} F_{A_i,J} \right)_{J \in \Omega_1} \]

**Theorem**

Let \( h_j = \min \left\{ q : \{ A_{ij}, i \in I \} \subseteq \bigcup_{i=1}^q H_i \quad H_i \in \mathcal{H}_{n-1} \right\} \).

If there exist \( j \) such that \( |S(J)| < h_j, \quad \forall J \) then \( C_1 \) is identifiable.
Adding covariates independent of cluster membership

\[ C_2 = \left( F_{A,J} : F_{A,J} = \bigotimes_{i \in I} F_{A_i,W_i,J} \right)_{J \in \Omega_1} \]  \hspace{1cm} (5)

\[ C_{2A} = \left( F_{A,J} : F_{A,J} = \bigotimes_{i \in I} F_{A_i,J} \right)_{J \in \Omega_1} \]  \hspace{1cm} (6)

\[ C_{2W} = \left( F_{A,J} : F_{A,J} = \bigotimes_{i \in I} F_{W_i,J} \right)_{J \in \Omega_1} \]  \hspace{1cm} (7)

**Theorem**

If \( C_{2A} \) and \( C_{2W} \) are identifiable and \( W_{ij} \) is not a multiple of \( A_{ij} \), for all \( i, j \), then \( C_2 \) is identifiable.
Numerical Example

- Two clusters with sizes $\pi_1 = \pi_2 = \frac{1}{2}$.
- Two time-points 1 and 2.
- Same variability in both clusters $\sigma = 0.1$

We simulate 50 samples of 100 trajectories with parameters

- $\beta^1 = (3, -2)$ and $\beta^2 = (0, 2)$ (linear model)
- $\beta^1 = (10, -12.5, 3.5)$ and $\beta^2 = (-2, 5, -1)$ (polynomial model).
Parallel coordinate plots of the estimated parameter

Linear Model

Parabolic Model
The generalized model

Theorem

The model is identifiable if

- $d_k < T$ for all $1 \leq k \leq K$ and all $a_{it}$ are distinct, for all $i_t$.
- $W_k$ has full rank for all $1 \leq k \leq K$.
- $rk(A_k, W_k) = rk(A_k) + rk(W_k)$, for all $1 \leq k \leq K$. 
Numerical Example

- Two clusters with sizes $\pi_1 = \pi_2 = \frac{1}{2}$.
- Two time-points 1 and 2.
- Same variability in both clusters $\sigma = 0.1$
- Shape description parameters $\beta_1 = (3, -2)$, $\beta_2 = (0, 2)$, $\delta_1 = 2$ and $\delta_2 = -3$.

We simulate 50 samples of 100 trajectories for 3 types of models:
- The covariate is independent of time and only takes values 0 or 1
- The covariate is time dependent but in a nonlinear way
- The covariate is time dependent in a linear way
Parallel coordinate plots of the estimated parameter values:

- **Model 1**
  - Parameters: \( \pi, \text{intercept}, \text{slope}, \sigma, \delta \)
  - Values: \( 0.35, 0.65, -0.075, 3.079, -2.03, 2.05, 0.0883, 0.1084, -3.05, 2.04 \)

- **Model 2**
  - Parameters: \( \pi, \text{intercept}, \text{slope}, \sigma, \delta \)
  - Values: \( 0.589, 3.1001, 2.04, 0.1114, 2.01, 1.00e+00, -0.0577, 3.1001, -2.06, 2.04 \)

- **Model 3**
  - Parameters: \( \pi, \text{intercept}, \text{slope}, \sigma, \delta \)
  - Values: \( 6.34e-08, 1.00e+00, 1.00e+00, 1.251, 36.1, 7.60, 7.76, -124.6, 67.8 \)


