Identifiability of Finite Mixture Models with underlying Normal Distribution

Cédric Noel[1] and Jang Schiltz[2]

[1] University of Luxembourg and IUT of Thionville-Yutz, University of Lorraine,
Espace Cormontaigne Impasse Alfred Kastler F-57970 Yutz, France
(E-mail: cedric.noel@univ-lorraine.fr>)

[2] Department of Finance, University of Luxembourg, 6, rue Richard
Coudenhove-Kalergi L-1359 Luxembourg, Luxembourg
(E-mail: jang.schiltz@uni.lu)

Abstract. In this paper, we show under which conditions generalized finite mixture with underlying normal distribution are identifiable in the sense that a given dataset leads to a uniquely determined set of model parameter estimations up to a permutation of the clusters.

Keywords: Identifiability, Finite Mixture Models.

1 Introduction

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 and numerical algorithms risk to find only part of these solutions. Worse, the researcher fitting the model might not even be aware that the solution his computer found is only one of many possibilities.

Identifiability of distributions has been an important research topic in the 1960s. Teicher ([6]) proved that the class of all mixtures of one-dimensional normal distributions is identifiable. Yakowitz and Spragins ([9]) extended this result five years later to the class of all Gaussian mixtures.

For a long time, it was believed that identifiability for linear regression mixtures with Gaussian errors follows directly from these results. DeSarbo and Cron ([1]) even make that claim explicitly. Hennig ([2]) only showed in 2000 that that statement is not correct in general by constructing counterexamples. Henning investigated the identifiability of the parameters of models for data generated by different linear regression distributions with Gaussian errors.

In this paper, we extend his results to finite mixture models in which the typical trajectories in the different clusters do not just follow a line, but a polynomial of any degree.

The remainder of this article is structured as follows. In section two, we present the class of finite mixture models we are interested in. In section three, we present some basic results about the identifiability of mixtures of distributions. In section four, finally, we prove under which conditions finite mixture models are identifiable.

2 Finite Mixture Models

Starting from a collection of individual trajectories, the aim of finite mixture models is to divide the population into a number of homogenous sub-populations and to estimate, at the same time, a typical trajectory for each sub-population (Nagin [3]).

More, precisely, consider a population of size N and a variable of interest Y. Let $Y_i = y_{i_1}, y_{i_2}, ..., y_{i_T}$ be T measures of the variable Y, taken at times $t_1, ..., t_T$ for subject number i. To estimate the parameters defining the shape of the trajectories, we need to fix the number K of desired subgroups. Denote the probability of a given subject to belong to group number k by π_k.

The objective is to estimate a set of parameters $\Omega = \{\pi_k, \beta_0^k, \beta_1^k, ...; k = 1, ..., K\}$ which allow to maximize the probability of the measured data. The particular form of Ω is distribution specific, but the β parameters always perform the basic function of defining the shapes of the trajectories. In Nagin's finite mixture model (Nagin [3]), the shapes of the trajectories are described by a polynomial function of age or time. Assume that for a subject in group k

$$y_{i_t} = \sum_{j=1}^{s} \beta_j^k a_{it}^j + \varepsilon_{it}, \tag{1}$$

where a_{it} denotes the age of subject i at time t, s the degree of the polynomial describing the trajectories in the different groups and ε_{it} is a disturbance assumed to be normally distributed with a zero mean and a constant standard deviation σ. The likelihood of the data is then given by

$$L = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \prod_{t=1}^{T} g_k(y_{i_t}), \tag{2}$$

where $g_k(y_{i_t})$ is the probability distribution function of y_{i_t} given membership in group k. In this paper we restrict ourselves to normal distributions.
The disadvantage of the basic model is that the trajectories are static and do not evolve in time. Thus, Nagin introduced several generalizations of his model in his book (Nagin [3]). Among others, he introduced a model allowing to add covariates to the trajectories. Let $z_1, ..., z_M$ be M covariates potentially influencing Y. We are then looking for trajectories

$$y_{i_t} = \sum_{j=0}^{s} \beta_j^k a_{it}^j + \alpha_1^j z_1 + ... + \alpha_M^j z_M + \varepsilon_{it}, \tag{3}$$

where ε_{it} is normally distributed with zero mean and a constant standard deviation σ. The covariates z_m may depend or not upon time t.

But even this generalized model still has two major drawbacks. First, the influence of the covariates in this model is unfortunately limited to the intercept of the trajectory. This implies that for different values of the covariates, the corresponding trajectories will always remain parallel by design, which does not necessarily correspond to reality.

Secondly, in Nagin's model, the standard deviation of the disturbance is the same for all the groups. That too is quite restrictive. One can easily imagine situations in which in some of the groups all individual are quite close to the mean trajectory of their group, whereas in other groups there is a much larger dispersion. To address and overcome these two drawbacks, Schiltz ([5]) proposed the following generalization of Nagin's model.

Let $x_1, ..., x_M$ and $z_{i_1}, ..., z_{i_T}$ be covariates potentially influencing Y. Here the x variables are covariates not depending on time like gender or cohort membership in a multicohort longitudinal study and the z variable is a covariate depending on time like being employed or unemployed. They can of course also designate time-dependent covariates not depending on the subjects of the data set which still influence the group trajectories, like GDP of a country in case of an analysis of salary trajectories.

The trajectories in group k will then be written as

$$y_{i_t} = \sum_{j=0}^{s} \left(\beta_j^k + \sum_{m=1}^{M} \alpha_m^k x_m + \gamma_j^k z_{i_t} \right) a_{it}^j + \varepsilon_{it}^k, \tag{4}$$

where the disturbance ε_{it}^k is normally distributed with mean zero and a standard deviation σ_k, constant inside group k, but different from one group to another. Since, for each group, this model is just a classical fixed effects model for panel data regression (see Wooldridge ([8])), it is well defined and we can get consistent estimates for the model parameters.

That model allows obviously to overcome the drawbacks of Nagin's model. The standard deviation of the uncertainty can vary across groups and the trajectories depend in a nonlinear way on the covariates.

Whereas the basic model is usually identified under very mild conditions, it is obvious that this is no longer true in all generality for the two generalized models. We will investigate this in the remainder of this paper.

3 Identifiability

In 1963, Teicher ([6]) showed the following result for mixtures of normal distributions.

Proposition 1. *The class of all mixtures of one-dimensional normal distributions is identifiable.*

We will use that proposition to prove under which conditions finite mixture models are identifiable.

Consider the distribution f of a finite mixture model.

$$f(y_i; \Omega) = \sum_{k=1}^{K} \pi_k g_k(y_i; \theta^k), \tag{5}$$

which is equivalent to

$$F(y_i; \Omega) = \sum_{k=1}^{K} \pi_k G_k(y_i; \theta^k), \tag{6}$$

where F and G_k denote the cumulative distribution functions (cdf's) of f and g_k respectively.

Let $\mathcal{F} = \left\{F(y;\omega),\ y \in \mathbb{R}^T,\ \omega \in \mathbb{R}_K^{s+2}\right\}$ be a family of T-dimensional cdf's indexed by a parameter set ω, such that $F(y;\omega)$ is measurable in $\mathbb{R}^T \times \mathbb{R}_K^{s+2}$. The the $s+2$-dimensional cdf $H(x) = \int_{\mathbb{R}_K^{s+2}} F(y;\omega)dG(\omega)$ 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 said to be identifiable if Q is a one-to-one map from $\mathcal{G}$ onto $\mathcal{H}$.

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

$$\mathcal{H} = \left\{ H(y) : H(y) = \sum_i c_i F(y,\omega_i),\ c_i > 0, \sum_i c_i = 1,\ F(y,\omega_i) \in \mathcal{F} \right\}. \quad (7)$$

In this context, the definition of identifiability implies that $\mathcal{F}$ generates an identifiable finite mixture model if and only if

$$\sum_{i=1}^{N} c_i F_i = \sum_{i=1}^{M} c'_i F'_i \quad (8)$$

implies that $N = M$ and for each i, $1 \leq i \leq N$ there is some j, $1 \leq j \leq N$, such that $c_i = c'_j$ and $F_i = F'_j$.

We can then easily prove the following characterization of identifiability.

Theorem 1. *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.*

We denote by $< A >$ the span of A over the real numbers.

Proof. Necessity.
Suppose that the family $\mathcal{F}$ is not linearly independent. Then, there exist an integer N and N real numbers a_i, at least one of them not being zero, such that, $\sum_{i=1}^N a_i F_i = 0$. Without loss of generality, we can suppose that $a_i < 0 \Leftrightarrow i \leq M$. Thus, $\sum_{i=1}^M |a_i| F_i = \sum_{i=M+1}^N |a_i| F_i$.

Since the F_i are cdf's, this implies that

$$\lim_{y \to (+\infty, \cdots, +\infty)} \sum_{i=1}^{M} |a_i| F_i(y) = \lim_{y \to (+\infty, \cdots, +\infty)} \sum_{i=M+1}^{N} |a_i| F_i(y), \quad (9)$$

hence

$$\sum_{i=1}^{M} |a_i| = \sum_{i=M+1}^{N} |a_i|. \quad (10)$$

Now, define c_i for each i by

$$c_i = \frac{|a_i|}{\sum_{i=M+1}^{N} |a_i|}.$$

Then,$\sum_{i=1}^{M} c_i = \sum_{i=M+1}^{N} c_i = 1$ and

$$\sum_{i=1}^{M} c_i F_i = \sum_{i=M+1}^{N} c_i F_i.$$

Thus we have two different distinct representations of the same mixture and therefore $\mathcal{H}$ is not identifiable.

Sufficiency.
If $\mathcal{F}$ is a linearly independent family, there exists a basis of $< \mathcal{F} >$. If we suppose that $\mathcal{H}$ is non identifiable there exist two distinct representations of the same mixture. Therefore $\mathcal{H} \subset < \mathcal{F} >$ which contradicts the uniqueness of the representation property of bases. □

We will now analyze the identifiability of some classes of generalized finite mixture models.

4 Identifiability of a class of finite mixture models

We will prove the identifiability of a big subclass of the generalized finite mixture model presented in section 2. Consider indeed the model defined by

$$Y_{it} = f(a_{it}; \beta^k, \delta^k) + \varepsilon_{it}^k = \beta^k A_{it} + \delta^k W_{it} + \varepsilon_{it}^k, \tag{11}$$

that we can write as

$$Y_i = \beta^k A_i + \delta^k W_i + \varepsilon_i^k, \tag{12}$$

with $Y_i = (Y_{i1}, \cdots, Y_{iT})$, $A_i = (A_{i1}, \cdots, A_{iT})$, $W_i = (W_{i1}, \cdots, W_{iT})$ and $\varepsilon_i^k \sim \mathcal{N}(0; \sigma_k I_T)$.

Thus, $Y_i \sim \mathcal{N}\left(\beta^k A_i + \delta^k W_i, \sigma_k I_T\right)$.

Hennig ([2]) showed the identiability of clusterwise linear regression models in the case of a one-dimensional normal distribution. We extend this results to the case of multi-dimensional normal distributions and polynomial trajectories.

We can write

$$\mathcal{L}\left((Y_i)_{i \in I}\right) = \bigotimes_{i \in I} F_{A_i, W_i, J}, \tag{13}$$

where $F_{A_i,W_i,J}(Y_i) = \int_{T_1} \Phi_{0,\Sigma}(Y_i - \beta_k A_i - \delta_k W_i)\, dJ\left(\beta, \sigma^2\right)$ with $T_1 = \mathbb{R}^{s+1} \times \mathbb{R}_0^+$, $J \in \Omega_1 = \mathcal{J}(T_1)$ and $\Sigma = \sigma I_T$.

$\mathcal{J}(T_1)$ denotes the set of mixing distributions with finite support on the parameter set T. $S(J)$ is the support set of $J \in \mathcal{J}(T_1)$. Thus, $K = |S(J)|$ is the number of mixture components and the elements of $\mathcal{J}(T_1)$ are distributions generating parameter values $(\beta^1, \sigma_1^2), \cdots, (\beta^K, \sigma_K^2)$ for K clusters with probability $J(\beta^1, \sigma_1^2), \cdots, J(\beta^K, \sigma_K^2)$. I is some index set, here $I = \{1, \cdots N\}$ since we suppose that we analyze data from a population of size N. $\bigotimes$ denotes the independent product of distributions.
Identifiability of a model means that knowing the data distribution $\mathcal{L}(Y_i), i \in I$, one can identify uniquely the mixing distribution J. That is, no two distinct sets of parameters lead to the same data distribution.

4.1 Nagin's base model

Nagin's base model can be written as

$$\mathcal{C}_1 = \left(F_{A,J} \ : \ F_{A,J} = \bigotimes_{i \in I} F_{A_i,J} \right)_{J \in \Omega_1}$$

In that case, identifiability means that, knowing the data distributions $\mathcal{L}(Y_i)_{i \in I}$, we can uniquely identify the mixing distribution J and two distinct sets of parameters $(\beta^1, \sigma_1^2, J(\beta^1, \sigma_1^2)), \cdots, (\beta^K, \sigma_K^2, J(\beta^K, \sigma_K^2))$ and
$(\beta'^1, \sigma_1^{'2}, J(\beta'^1, \sigma_1^{'2})), \cdots, (\beta'^K, \sigma_K^{'2}, J(\beta'^K, \sigma_K^{'2}))$ lead to different data distributions.

Theorem 2. *Let* $h_j = \min\left\{q \ : \ \{A_{ij}, i \in I\} \subseteq \cup_{i=1}^{q} H_i \ \ H_i \in \mathcal{H}_{n-1}\right\}$.

If there exist j *such that* $|S(J)| < h_j, \ \forall J$ *then* $\mathcal{C}_1$ *is identifiable.*

Proof. We need to show only that $F_{A_i,J} = F_{A_i,\tilde{J}} \Rightarrow J = \tilde{J}$ because J contains all information to define the common distribution $F_{A_i,J}$ of $(Y_i)_{i \in I}$.

Suppose that $F_{A_i,J} = F_{A_i,\tilde{J}}$ and $J \neq \tilde{J}$. Without loss of generality we can assume that $|S(J)| \geq |S(\tilde{J}), |$. Thus there exists $(\beta^1, \sigma_1) \in S(\tilde{J})$ such that

$$J\{(\beta^1, \sigma_1^2)\} \neq \tilde{J}\{(\beta^1, \sigma_1^2)\}. \tag{14}$$

$F_{A_i,J} = F_{A_i,\tilde{J}}$ implies the equality of the marginal Gaussian mixtures for all $A_i, \ i \in I$ and

$$F_{A_i,J}(Y_i) = \int_{T_1} \Phi_{\beta_k A_i, \Sigma}\left(Y_i\right) dJ\left(\beta, \sigma^2\right) \tag{15}$$

$$= F_{A_i,\tilde{J}}(Y_i) = \int_{T_1} \Phi_{\beta_k A_i, \Sigma}\left(Y_i\right) d\tilde{J}\left(\beta, \sigma^2\right). \tag{16}$$

The identifiabilty of finite Gaussian mixtures then implies, for $i \in I$

$$J\left\{(\beta,\sigma^2) : \left(\beta A_i,\sigma^2\right) = \left(\beta^1 A_i,\sigma_1^2\right)\right\} = \tilde{J}\left\{(\tilde{\beta},\tilde{\sigma}^2) : \left(\tilde{\beta} A_i,\tilde{\sigma}^2\right) = \left(\beta^1 A_i,\sigma_1^2\right)\right\} \tag{17}$$

The idea of the proof is that the restriction to $|S(\tilde{J})|$ ensures the existence of a matrix A_i whose marginal mixture $\mathcal{N}(\beta_1 A_i,\sigma_1^2)$, parameterized by J, cannot be explained by $(\tilde{\beta},\tilde{\sigma}^2) \in S(\tilde{J})$ if $\tilde{\beta} \neq \beta_1$. Therefore $S(\tilde{J})$ must contain (β_1,σ_1^2).

Suppose that for all $(\beta,\sigma^2) \in S(J)$, and in particular for (β^1,σ_1^2), there exists $i(\beta) \in I$ such that

$$\forall(\tilde{\beta},\tilde{\sigma}^2) \in S(\tilde{J}) \ : \ \beta A_{i(\beta)} = \tilde{\beta} A_{i(\beta)} \Rightarrow \beta = \tilde{\beta}. \tag{18}$$

The definition of $A_i = A_{i(\beta^1)}$ implies that

$$\forall S(\tilde{J}) \ni (\tilde{\beta},\tilde{\sigma}^2) \neq (\beta^1,\sigma_1^2) : \ (\tilde{\beta} A_i,\tilde{\sigma}^2) \neq (\beta^1 A_i,\sigma_1^2). \tag{19}$$

Thus, using (27) and (17),

$$\tilde{J}\{(\beta^1,\sigma_1^2)\} = J\{(\beta,\sigma) : \ (\beta A_i,\sigma^2) = (\beta^1 A_i,\sigma_1^2)\}. \tag{20}$$

But $J\{(\beta,\sigma) : \ (\beta A_i,\sigma^2) = (\beta^1 A_i,\sigma_1^2)\} \neq 0$ because it contains (β^1,σ_1^2). For the same reason, $\tilde{J}\{(\beta^1,\sigma_1^2)\} \neq 0$.
Hence (19) implies that $(\beta^1 A_i,\sigma_1^2) \in S(\tilde{J})$.

By (14), $\tilde{J}\{(\beta^1,\sigma_1^2)\} \neq J\{(\beta^1,\sigma_1^2)\}$. Consequently, equation (20) implies that

$$\exists S(J) \ni (\beta^2,\sigma_2^2) \neq (\beta^1,\sigma_1^2) : \ (\beta^2 A_i,\sigma_2^2) = (\beta^1 A_i,\sigma_1^2). \tag{21}$$

Consider $A_i = A_{i(\beta^2)}$ and apply the same arguments than above to get $(\beta^2 A_i,\sigma_2^2) \in S(\tilde{J})$. This result leads to a contradiction between (19) and (21) . Indeed, $(\beta^2,\sigma_2^2) \in S(\tilde{J})$ and $(\beta^2,\sigma_2^2) \neq (\beta^1,\sigma_1^2)$. By (19), $(\beta^2 A_i,\sigma_2^2) \neq (\beta^1 A_i,\sigma_1^2)$ and by (21), $(\beta^2 A_i,\sigma_2^2) = (\beta^1 A_i,\sigma_1^2)$.

Thus there exists some $(\beta,\sigma) \in S(J)$ such that $\forall i \in I \ \forall(\tilde{\beta},\tilde{\sigma}^2) \in S(\tilde{J}) \ : \beta A_i = \tilde{\beta} A_i \Rightarrow \beta \neq \tilde{\beta}$.

Hence

$$\{A_{ij}, i \in I, \ j = 1 \cdots T\} \subset \cup_{(\tilde{\beta},\tilde{\sigma}^2):\tilde{\beta}\neq\beta}\{x : \ \beta x = \tilde{\beta} x\}. \tag{22}$$

Therefore $\cup_{(\tilde{\beta},\tilde{\sigma}^2):\tilde{\beta}\neq\beta}\{x : \ \beta x = \tilde{\beta} x\}$ is composed by $|S(\tilde{J})|$ different hyperplanes.
So for $j = 1 \cdots T$, $h_j \leq |S(\tilde{J})| \leq |S(J)|$. □

4.2 Addition of covariates independent of the clusters

Let us now add covariates to the model that are independent of the K groups. Define

$$\mathcal{C}_2 = \left(F_{A,J} \ : \ F_{A,J} = \bigotimes_{i \in I} F_{A_i, W_i, J} \right)_{J \in \Omega_1} , \tag{23}$$

$$\mathcal{C}_{2A} = \left(F_{A,J} \ : \ F_{A,J} = \bigotimes_{i \in I} F_{A_i, J} \right)_{J \in \Omega_1} , \tag{24}$$

$$\mathcal{C}_{2W} = \left(F_{A,J} \ : \ F_{A,J} = \bigotimes_{i \in I} F_{W_i, J} \right)_{J \in \Omega_1} . \tag{25}$$

We have then the following identifiability result.

Theorem 3. *If $\mathcal{C}_{2A}$ and $\mathcal{C}_{2W}$ are identifiable and W_{ij} is not a multiple of A_{ij}, for all i, j, then $\mathcal{C}_2$ is identifiable.*

Since the covariates are just a linear addition to the model, the proof follows directly from the 2 following propositions.

Proposition 2. *$\mathcal{C}_{2A}$ is identifiable if and only if $d_k < T$ for all $1 \leq k \leq K$ and the a_{it} are distinct, for all values of i and t.*

Proof. We need to show that $F_{A_i,J} = F_{A_i,\tilde{J}} \Leftrightarrow J = \tilde{J}$. Suppose that

$$F_{A_i,J}(Y_i) = \int_{T_1} \Phi_{\beta^k A_i, \Sigma}(Y_i) \, dJ\left(\beta, \sigma^2\right) \tag{26}$$

$$= F_{A_i,\tilde{J}}(Y_i) = \int_{T_1} \Phi_{\beta^k A_i, \Sigma}(Y_i) \, d\tilde{J}\left(\beta, \sigma^2\right) . \tag{27}$$

By identifiabilty of finite Gaussian mixtures, the equality above is equivalent, for $i \in I$, to:

$$J\left\{(\beta, \sigma^2) : \left(\beta A_i, \sigma^2\right) = \left(\mu_1, \sigma_1^2\right)\right\} = \tilde{J}\left\{(\tilde{\beta}, \tilde{\sigma}^2) : \left(\tilde{\beta} A_i, \tilde{\sigma}^2\right) = \left(\mu_1, \sigma_1^2\right)\right\} \tag{28}$$

for some (μ_1, σ_1).

Assume that there exists $\tilde{\beta}$ in $S(\tilde{J})$ such that $\beta A_i = \tilde{\beta} A_i$ for some $\beta \in S(J)$. This means that $1 \leq t \leq T$, $\beta A_{it} = \tilde{\beta} A_{it}$ but $\tilde{\beta} \neq \beta$ for all $\beta \in S(J)$.
If $d_k < T$ and a_{it} are different $\forall t \leq T$, we have 2 different polynomials of degree strictly smaller than T that intersect in T points.
Thus $\beta = \tilde{\beta}$.

If we know cluster membership for each value Y_i, we can write

$$Y_k = A_k \beta_k^t,$$

where $Y_k = \begin{pmatrix} Y_{k11} \\ \vdots \\ Y_{kn_kT} \end{pmatrix}$ and $A_k = \begin{pmatrix} 1 & a_{11} & \cdots & a_{11}^{d_k-1} \\ \vdots & & & \vdots \\ 1 & a_{n_kT} & \cdots & a_{n_kT}^{d_k-1} \end{pmatrix}$.

Since $\begin{pmatrix} 1 & a_{11} & \cdots & a_{11}^{d_k-1} \\ \vdots & & & \vdots \\ 1 & a_{n_kT} & \cdots & a_{n_kT}^{d_k-1} \end{pmatrix}$ is a Vandermonde matrix and $d_k < T$, which is required for the matrix to be invertible, the invertibility condition is guaranteed to hold if all the a_{it} values are distinct.

So

$$\beta_k = Y_k \left(A_k^t A_k\right)^{-1} A_k. \tag{29}$$

□

Proposition 3. *If for all $1 \leq t, t' \leq T$ and for $i, j \in I$, $a_{it} = a_{jt}$ and $a_{it} \neq a'_{it}$, $\mathcal{C}_{2A}$ is identifiable if and only if $d_k < T$ for all $1 \leq k \leq K$.*

Proof. In this case the matrix $\begin{pmatrix} 1 & a_{11} & \cdots & a_{11}^{d_k-1} \\ \vdots & & & \vdots \\ 1 & a_{nT} & \cdots & a_{nT}^{d_k-1} \end{pmatrix}$ becomes $\begin{pmatrix} 1 & a_{11} & \cdots & a_{11}^{d_k-1} \\ \vdots & & & \vdots \\ 1 & a_{1T} & \cdots & a_{1T}^{d_k-1} \end{pmatrix}$ and is invertible if $d_k < T$ and $a_{it} \neq a_{it'}$, $1 \leq t, t' \leq T$. □

Numerical example

Let us illustrate the two previous propositions by an example. To keep everything the easiest possible, we consider an example with just two clusters with sizes $\pi_1 = \pi_2 = \frac{1}{2}$ and two time-points 1 and 2. For the sake of simplicity, we also suppose that the variability of the error term is the same for both groups and we take $\sigma = 0.1$
To emphasize the difference between the identifiability of a mixture of probability distributions and the identifiability of finite mixture models, we point out that proposition 1 implies that the mixture distribution $\frac{1}{2}\mathcal{N}(\mu_{1t}; 0.1) + \frac{1}{2}\mathcal{N}(\mu_{2t}; 0.1)$ is always identifiable.

Proposition 3 tells us that finite mixture models with polynomial trajectories will be identifiable as long as the degree of the polynomials is at most 1, since $T = 2$. To illustrate this, we simulate 50 samples of 100 observations, once with linear trajectories and once with polynomials of degree 2. More precisely, we use the following parameter values

- $\beta^1 = (3, -2)$ and $\beta^2 = (0, 2)$ for the linear model ;
- $\beta^1 = (10, -12.5, 3.5)$ and $\beta^2 = (-2, 5, -1)$ for the polynomial model.

We then use our R package `trajeR` (Noel and Schiltz [4]) to fit these 100 samples and illustrate the result by means of parallel coordinate plots (Wegman [7]).

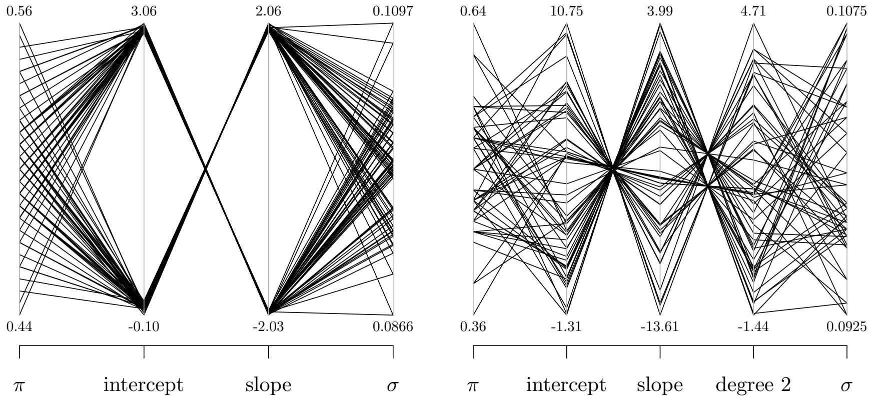

Fig. 1: Parallel coordinate plots of the estimated parameters for 50 simulated samples for linear and parabolic trajectories.

Figure 1 shows the result. On the left side, we see the different parameter estimations for the linear model. We see that in this case all parameter estimations give roughly the same result. There are 2 solutions for the different trajectory shape parameters, corresponding to the 2 clusters, one cluster with a trajectory defined by an intercept of 3 and a slope of -2 and one cluster with a trajectory defined by an intercept of 0 and a slope of 2. The estimation for the group sizes vary between 0.44 and 0.56 and the standard deviation of the error term are estimated as being between 0.087 and 0.110.
The right part of the graph shows the parameter estimation for the parabolical model. In this case the trajectory shape parameters cannot be precisely estimated and there is no indication of a two-group solution. This is a clear indication of the non identifiability of the model.

4.3 The generalized model

Now consider the generalized finite mixture model

$$Y_i = \beta^k A_i + \delta^k W_i + \epsilon_i.$$

We then have the following result.

Proposition 4. *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;*
- *$d_k < T$ for all $1 \leq k \leq K$;*
- *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 $1 \leq k \leq K$ where $rk(\cdot)$ denotes the rank of a matrix, A_k is defined as in the proof of proposition 2, and W_k are the elements W_i corresponding to A_k.*

Proof. If W_i does not depend on time, the trajectories of al clusters are just translations of each other. Thus, the first condition of the proposition implies that all trajectory parameters are identifiable and since $rk(A_k, W_k) =$

$rk(A_k) + rk(W_k)$ we can determine δ^k too.

In the general case, suppose that $d_k < T$ for all $1 \leq k \leq K$. Then for any integer c, a mixture of c components of the form $\sum_{k=1}^{c} \pi_k \mathcal{N}\left(\beta^k A_{it}, \sigma_k\right)$ is identifiable.
If we know the cluster membership of each value Y_i, we can determine β^k as in equation (29) by $\beta^k = Y_k \left(A_k^t A_k\right)^{-1} A_k$.

Denote $P_k = A_k^t \left(A_k A_k^t\right)^{-1} A_k$ and $R_k = I - P_k$. Then,

$$\begin{aligned}
\beta^k A_k + \delta^k W_k &= \beta^k A_k + \delta^k W_k P_k + \delta^k W_k \left(I - P_k\right) && (30)\\
&= \beta^k A_k + \delta^k W_k A_k^t \left(A_k A_k^t\right)^{-1} A_k + \delta^k W_k \left(I - P_k\right) && (31)\\
&= \left(\beta^k + \delta^k W_k A_k^t \left(A_k A_k^t\right)^{-1}\right) A_k + \delta^k W_k R_k && (32)\\
&= \left(\beta^k + \delta^k W_k A_k^t \left(A_k A_k^t\right)^{-1} \; \delta^k\right) \begin{pmatrix} A_k \\ W_k R_k \end{pmatrix} && (33)\\
&= \lambda_k V. && (34)
\end{aligned}$$

Suppose $\lambda_k V = 0$ for some λ_k. Then give $\beta^k A_k + \delta^k W_k = 0$, hence $\beta^k = \delta^k = 0$ by linear independence of the columns of A_k and W_k. So V is a $T + rk(W_k)$ matrix of full rank.
Since $Y_k = \lambda_k V + \varepsilon$, we have

$$\begin{aligned}
\hat{\lambda}_k &= Y_k \left(V V^t\right)^{-1} V^t && (35)\\
&= \left(A_k \, W_k R_k\right) \begin{pmatrix} A_k A_k^t & A_k R_k^t W_k^t \\ W_k R_k A_k^t & W_k R_k R_k^t W_k^t \end{pmatrix}^{-1} && (36)\\
&= \left(A_k \, W_k R_k\right) \begin{pmatrix} A_k A_k^t & A_k R_k W_k^t \\ W_k R_k A_k^t & W_k R_k R_k W_k^t \end{pmatrix}^{-1}. && (37)
\end{aligned}$$

Since $R_k = I - A_k^t \left(A_k A_k^t\right)^{-1} A_k$, we have $A_k R_k = R_k A_k^t = 0$ and $P_k^2 = P_k$. Moreover,

$$\begin{aligned}
\hat{\lambda}_k &= Y_k \left(V V^t\right)^{-1} V^t && (38)\\
&= Y_k \left(A_k \, W_k R_k\right) \begin{pmatrix} A_k A_k^t & 0 \\ 0 & W_k R_k W_k^t \end{pmatrix}^{-1} && (39)\\
&= Y_k \left(A_k \left(A_k A_k^t\right)^{-1} \; W_k R_k \left(W_k R_k W_k^t\right)^{-1}\right) && (40)\\
&= \left(Y_k A_k \left(A_k A_k^t\right)^{-1} \; Y_k W_k R_k \left(W_k R_k W_k^t\right)^{-1}\right). && (41)
\end{aligned}$$

Thus

$$\hat{\delta}^k = Y_k W_k R_k \left(W_k R_k W_k^t\right)^{-1}$$

and

$$\hat{\beta}^k = Y_k A_k \left(A_k A_k^t\right)^{-1} - \hat{\delta}^k W_k A_k^t \left(A_k A_k^t\right)^{-1}.$$

Hence all parameters are identified. □

Numerical example

Let us illustrate proposition 4 by an example. As in the previous example, we consider a model with just two clusters of sizes $\pi_1 = \pi_2 = \frac{1}{2}$, two timepoints 1 and 2 and a constant variability of the error term of $\sigma = 0.1$ Furthermore, we use the shape description parameters $\beta_1 = (3, -2)$ and $\beta_2 = (0, 2)$ and fix $\delta_1 = 2$ and $\delta_2 = -3$. We will study 3 types of models, defined by the following supplementary conditions.

- The covariate W 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;

To illustrate this, we simulate 50 samples of 100 observations, We then use our R package `trajeR` (Noel and Schiltz [4]) to fit these 100 samples and illustrate the result by means of parallel coordinate plots (Wegman [7]).

We can see on figure 2 that the two first model specifications shown on the two left graphs are identifiable. But the model represented on the right side is not. The linear dependence on time of the covariate has the effect that neither β nor δ can be uniquely determined.

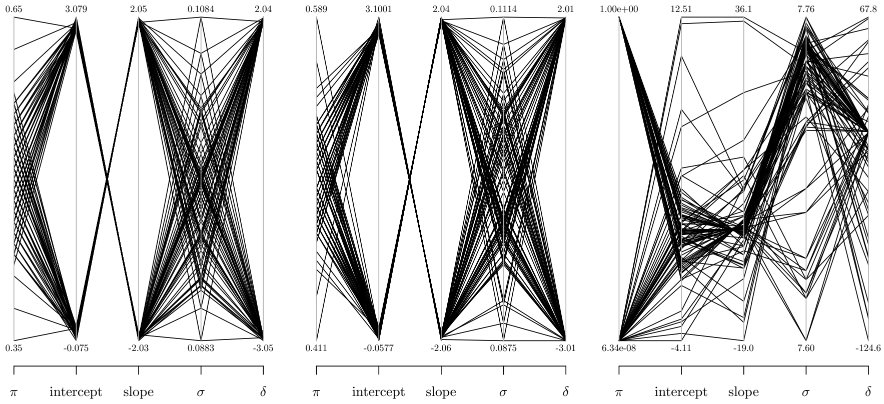

Fig. 2: Parallel coordinate plots of the estimated parameters for 50 simulated samples with different forms of the covariant.

References

1. W.S. Desarbo and W.L. Cron. A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5, 249–282, 1988.
2. C. Hennig. Identifiability of Models for Clusterwise Linear Regression. *Journal of Classification*, 17, 273–296, 2000.
3. D.S. Nagin. *Group-Based Modeling of Development*,Harvard University Press, Cambridge, 2005.
4. C. Noel and J. Schiltz. trajeR - an R package for finite mixture models, to appear, 2020.

5. J. Schiltz. A Generalization of Nagin's Finite Mixture Model. In: M. Stemmler, A. von Eye and W. Wiedermann. *Dependent Data in Social Sciences Research*. Springer, Heidelberg, 2015.
6. H.Teicher. Identifiability of Finite Mixtures. *Annals of Mathematical Statistics*, 34,4,1265–1269, 1963.
7. E.J. Wegman. Hyperdimensional Data Analysis Using Parallel Coordinates. *Journal of the American Statistical Association*,85, 411, 664–675, 1990.
8. J.M. Wooldridge. *Econometric Analysis of Cross-Section and Panel Data.* 2nd edition, MIT Press, Cambridge, 2010.
9. S.J. Yakowitz and J.D. Spragins. (1968), On the identifiability of finite mixtures. *Annals of Mathematical Statistics*, 39, 209–214, 1968.