A new condition for identifiability of finite mixture distributions

N. Atienza, J. Garcia-Heras, J.M. Muñoz-Pichardo

Abstract In this paper a sufficient condition for the identifiability of finite mixtures is given. This condition is less restrictive than Teicher's condition (Teicher H, Ann Math Stat 34:1265–1269 (1963)) and therefore it can be applied to a wider range of families of mixtures. In particular, it applies to the classes of all finite mixtures of Log-gamma and of reversed Log-gamma distributions. These fami-lies have been already studied by Henna J Jpn Stat Soc 24:193–200 (1994) using another condition, different from Teicher's, but more difficult to check in many cases. Furthermore, the result given in this paper is very appropiated for the case of mixtures of the union of different distribution families. To illustrate this an applica-tion to the class of all finite mixtures generated by the union of Lognormal, Gamma and Weibull distributions is given, where Teicher's and Henna's conditions are not applicable.

Keywords Identifiability · Finite mixture · Log-normal distributions · Gamma distributions · Weibull distributions

1 Introduction and definitions

Finite mixtures of distributions are very useful for building probability models of a wide variety of random phenomena in the biological, physical and social sciences.

N. Atienza (✉)
Departamento Matemática Aplicada I, Escuela Técnica Superior de Ingeniería Informática
Universidad de Sevilla,
Av. Reina Mercedes, s/n
S.N. Sevilla 41012, Spain
E-mail: natienza@us.es

J. Garcia-Heras · J.M. Muñoz-Pichardo
Departamento Estadística e I.O., Facultad de Matemáticas,
Universidad de Sevilla,
c\Tarfia, s/n
S.N. Sevilla 41012, Spain

A broad spectrum of their applications can be found in many references (Böhning 2000; Lindsay 1995; McLachlan and Basford 1988; McLachlan and Peel 2000; Titterington et al. 1985). Let $\mathcal{F}$ be a family of m-dimensional cumulative distribution functions. A finite mixture of $F_1, \ldots, F_k \in \mathcal{F}$ is any convex combination $H = \pi_1 F_1 + \cdots + \pi_k F_k$, where

$$\{\pi_i\}_{i=1}^k \in \mathcal{C} = \left\{ \{\pi_i\}_{i=1}^h \middle| h \in \mathcal{Z}^+; \pi_i > 0, \forall i = 1, \ldots, h; \sum_{i=1}^h \pi_i = 1 \right\}$$

are the weights of the mixture ($\mathcal{Z}^+$ denotes the set of natural numbers). $F_1, \ldots, F_k$ are called the components of the mixture.

For the estimation procedures to be well-defined it is required that H is identifiable, in other words, the question about the existence of a unique characterization of any one of the class of models being considered. Particular difficulties with identifiability appear in mixture models.

To be more precise, given $\mathcal{F}$ a class of m-dimensional distribution functions, the class of finite mixtures of $\mathcal{F}$ is:

$$\mathcal{H} = \left\{ H \middle| \exists \{\pi_i\}_{i=1}^k \in \mathcal{C}, \exists F_1, \ldots, F_k \in \mathcal{F} : H(\cdot) = \sum_{i=1}^k \pi_i F_i(\cdot) \right\}.$$

In this context, $\mathcal{H}$ is identifiable if and only if for any $H, \hat{H} \in \mathcal{H}$,

$$H = \sum_{j=1}^k \pi_j F_j, \quad \hat{H} = \sum_{j=1}^{\hat{k}} \hat{\pi}_j \hat{F}_j,$$

the equality $H = \hat{H}$ implies $k = \hat{k}$ and that $(\pi_1, F_1), \ldots, (\pi_k, F_k)$'s are a permutation of $(\hat{\pi}_1, \hat{F}_1), \ldots, (\hat{\pi}_k, \hat{F}_k)$.

Identifiability problems concerning finite and countable mixtures have been widely investigated. Teicher (1963) gave a sufficient condition for a finite mixture to be identifiable and applied it to the Normal and Gamma families. Khalaf (1988) proved the identifiability of finite mixtures of Weibull, Lognormal, Chi, Pareto, and of power function distributions by applying a modification of Teicher's theorem given by Chandra (1977) to the moment generating function of $\log X$. Barndorff-Nielsen (1965) gave a sufficient condition for mixtures of exponential families. Henna (1994) gave a modification of Teicher's theorem, based on the assumption of the parametric space to be a product space, which involves a biparametric order with several conditions that are hard to check in many cases. He obtained the identifiability of mixtures of Log-gamma and reverse Log-gamma distributions as a consequence.

In section 2 we give a new version of Teicher's condition with weaker assumptions, which allows us to study identifiability problems in a wider context. It is particularly useful in the study of finite mixtures of the union of different families of distributions, widely used in the modelling of biological, medical and social phenomena. The study of the identifiability of finite mixtures of the union of Lognormal, Gamma and Weibull distributions families is given in section 3. These mixtures are used in modelling the variable 'time of hospital stay' (length of stay).

2 A sufficient condition for the identifiability

The problem of identifiability of mixtures has been the focus of interest in the last decades. The paper of Teicher (1963) is the starting point for investigation of identifiability of finite mixture models. However, identifiability is stated under some restrictive conditions which are not applicable to some multiparametric families. Next, we propose a new result which relaxes the requirements of Teicher's theorem.

In the sequel, A' will denote the accumulation set of $A \subset \mathcal{R}^d$, consisting of all points for which every neighborhood contains infinitely many distinct points of A.

Theorem 1 *Let $\mathcal{F}$ be a family of distributions. Let M be a linear mapping which transforms any $F \in \mathcal{F}$ into a real function ϕ_F with domain $S(F) \subset \mathcal{R}^d$. Let $S_0(F) = \{t \in S(F) : \phi_F(t) \neq 0\}$. Suppose that there exists a total order $\prec$ on $\mathcal{F}$, such that for any $F \in \mathcal{F}$ there exists $t(F) \in S_0(F)'$ verifying:*

(a) If $F_1, F_2, \ldots, F_m \in \mathcal{F}$ with $F_1 \prec F_i$ for $2 \leq i \leq m$, then

$$t(F_1) \in \left[S_0(F_1) \cap [\cap_{i=2}^m S(F_i)]\right]'.$$

(b) If $F_1 \prec F_2$, then $\lim_{t \to t(F_1)} \dfrac{\phi_{F_2}(t)}{\phi_{F_1}(t)} = 0$.

Then, the class $\mathcal{H}$ of all finite mixture distributions of $\mathcal{F}$ is identifiable.

Proof Suppose

$$\sum_{i=1}^{k} \pi_i F_i = \sum_{j=1}^{\hat{k}} \hat{\pi}_j \hat{F}_j \tag{1}$$

for some $\{\pi_i\}_{i=1}^k, \{\hat{\pi}_i\}_{j=1}^{\hat{k}} \in \mathcal{C}$ and $F_1, \ldots, F_k, \hat{F}_1, \ldots, \hat{F}_k \in \mathcal{F}$. Without any loss of generality, we can assume that $k \leq \hat{k}$, $F_i \prec F_j$, $\hat{F}_i \prec \hat{F}_j$ for $i < j$ and $F_1 \preceq \hat{F}_1$. Hence $F_1 \prec \hat{F}_j$ for all $j = 2 \ldots \hat{k}$. Set:

$$A = S_0(F_1) \bigcap [\cap_{i=2}^k S(F_i)] \bigcap [\cap_{i=1}^{\hat{k}} S(\hat{F}_i)].$$

From (a) there exists $t(F_1) \in A'$. For any $t \in A$,

$$\sum_{i=1}^{k} \pi_i \phi_{F_i}(t) = \sum_{j=1}^{\hat{k}} \hat{\pi}_j \phi_{\hat{F}_j}(t).$$

Since $t \in S_0(F_1)$, we can divide by $\phi_{F_1}(t)$. Letting $t \to t(F_1)$, we obtain

$$\sum_{i=1}^{k} \pi_i \lim_{t \to t(F_1)} \frac{\phi_{F_i}(t)}{\phi_{F_1}(t)} = \sum_{i=1}^{\hat{k}} \hat{\pi}_i \lim_{t \to t(F_1)} \frac{\phi_{\hat{F}_i}(t)}{\phi_{F_1}(t)}.$$

From (b),

$$\pi_1 = \hat{\pi}_1 \lim_{t \to t(F_1)} \frac{\phi_{\hat{F}_1}(t)}{\phi_{F_1}(t)}.$$

If $F_1 \prec \hat{F}_1$ held, we would have $\pi_1 = 0$, a contradiction. Thus $F_1 = \hat{F}_1$, and $\pi_1 = \hat{\pi}_1$. The corresponding summands in (1) cancel, and we can continue in this fashion to obtain $\pi_i = \hat{\pi}_i$ and $F_i = \hat{F}_i$ for $i = 1, 2, \ldots \min\{k, \hat{k}\}$. Finally $k = \hat{k}$. Indeed, if $\hat{k} > k$ held, we would have $\sum_{j=k+1}^{\hat{k}} \hat{\pi}_j \hat{F}_j(x) = 0$, and therefore $\hat{\pi}_j = 0$ for $k + 1 \leq j \leq \hat{k}$. Thus $\mathcal{H}$ is identifiable. □

Let us consider the following:

- The argument given in the proof also works when some of the components of $t(F)$ are infinite.
- this new condition includes multivariate distributions, whereas Teicher's result can only be applied to univariate distributions.
- Application of Teicher's result may be restrictive because the condition $S(F_1) \subset S(F_2)$ is imposed by the relation $F_1 \prec F_2$. This condition is relaxed in (a) of Theorem 1.
- The new theorem proposed here can be applied to a wider range of families of distributions than Henna's (1994) result (see section 3.2). But, moreover, in cases where it is useful, Theorem 1 can be checked in a simpler way because there is no need to consider the parametric space as a product space, and then computations are easier (see section 3.1).

Finally, let us state the following consequence, which simplifies the assumptions in cases where the point $t(F) = t_0$ does not depend on $F \in \mathcal{F}$.

Corollary 1 *Let $\mathcal{F}$ be a family of distributions. Let M be a linear mapping which transforms any $F \in \mathcal{F}$ into a real function ϕ_F with domain $S(F) \subseteq \mathcal{R}^d$. Let $S_0(F) = \{t \in S(F) : \phi_F(t) \neq 0\}$ and suppose that there exists a point t_0 verifying*

$$t_0 \in \left[\bigcap_{1 \leq i \leq k} S_0(F_i) \right]' \tag{2}$$

for any finite collection of distributions $F_1, \ldots, F_k \in \mathcal{F}$. If the order

$$F_1 \prec F_2 \text{ if and only if } \lim_{t \to t_0} \frac{\phi_{F_2}(t)}{\phi_{F_1}(t)} = 0 \tag{3}$$

is a total ordering on $\mathcal{F}$, then the class $\mathcal{H}$ of all finite mixture distributions of $\mathcal{F}$ is identifiable.

This corollary does simplify the verification of identifiability for some classes considerably, in particular, when $S_0(F)$ is of the form $(a(F), +\infty)$, $(-\infty, b(F))$ (or $(-\infty, +\infty)$). In these cases, we can consider t_0 to be $+\infty$ or $-\infty$, respectively.

3 Applications

In this section, two different applications of Theorem 1 are given. First, it is applied to two families of distributions, already studied by Henna (1994): the class of finite mixtures of Log-gamma distributions and the class of finite mixtures of reversed

Log-gamma distributions. The goal is to illustrate, with both examples, the usefulness of our result in cases where Teicher's requirements are not satisfied. It is also shown that our condition for identifiability is easier to verify than Henna's, since it avoids to define two subfamilies of parameters and the definitions of a total order in each of them.

The second application deals with a family which fails to verify both Teicher's and Henna's conditions for the usual mappings: the union of Lognormal, Gamma and Weibull distributions families. In particular it is illustrating the usefulness of our result in the study of identifiability of finite mixtures of the union of different families of distributions.

3.1 Application to the Log-gamma and reversed Log-gamma families

Let $\mathcal{F}_S$ be the family of Log-gamma distributions, as follows:

$$\mathcal{F}_S = \left\{ F : F(x;\mu,\sigma,k) = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{\exp[ks - \exp(s)]}{\Gamma(k)}\, \mathrm{d}s, \quad \mu \in \mathcal{R}, \sigma, k > 0 \right\},$$

where $\Gamma(k) = \int_0^\infty t^{k-1}e^{-t}\, \mathrm{d}t$ is the Gamma function.

Theorem 2 *The class $\mathcal{H}_{\mathcal{F}_S}$ of all finite mixtures of $\mathcal{F}_S$ is identifiable.*

Proof Let M be the map which transforms a distribution function $F \in \mathcal{F}_S$ into its moment generating function. M is linear and,

$$M[F(\cdot;\mu,\sigma,k)] = \phi_{(\mu,\sigma,k)}(t) = e^{\mu t}\Gamma(\sigma t + k)/\,\Gamma(k) \quad t \in (-k/\sigma, +\infty)$$

In this case $S_0(F(\cdot;\mu,\sigma,k)) = (-k/\sigma, +\infty)$, and therefore $+\infty$ verify (2) in Corollary 1. The proof is completed by showing that (3) defines a total order on $\mathcal{F}_S$.

Using Stirling's formula, which establishes that $\Gamma(z+1) \sim \sqrt{2\pi z}(z/e)^z$ for $z \to +\infty$ (i.e. the limit of the quotient is 1),

$$\begin{aligned}\phi_{(\mu,\sigma,k)}(t) &\sim \frac{e^{\mu t}}{\Gamma(k)}\sqrt{2\pi}(\sigma t)^{\sigma t+k-1/2}\left[1+\frac{k-1}{\sigma t}\right]^{\sigma t+k-1/2}\exp\{-\sigma t - k + 1\} \\ &\sim \frac{\sqrt{2\pi}}{\Gamma(k_i)}\exp\left(\mu_i t - \sigma_i t\right)\exp\left[(\sigma_i t + k_i - 1/2)(\log\sigma_i + \log t)\right].\end{aligned}$$

Therefore, for $t \to +\infty$

$$\begin{aligned}\frac{\phi_{(\mu_2,\sigma_2,k_2)}(t)}{\phi_{(\mu_1,\sigma_1,k_1)}(t)} \sim C\exp\Big\{&\left[(\mu_2-\mu_1)-(\sigma_2-\sigma_1)+(\sigma_2\log\sigma_2-\sigma_1\log\sigma_1)\right]t \\ &+(\sigma_2-\sigma_1)t\log t + (k_2-k_1)\log t\Big\}\end{aligned}$$

with C some positive constant.

Hence $F(\cdot;\mu_1,\sigma_1,k_1) \prec F(\cdot;\mu_2,\sigma_2,k_2)$ if and only if $[\sigma_2 < \sigma_1]$, or $[\sigma_2 = \sigma_1$ and $\mu_2 < \mu_1]$ or $[\sigma_2 = \sigma_1,\ \mu_2 = \mu_1$ and $k_2 < k_1]$, which is obviously a total order in $\mathcal{F}_S$. □

Analogously, considering $t(F) = -\infty$, we can prove that the class $\mathcal{H}_{\mathcal{F}_R}$ of all finite mixtures of $\mathcal{F}_R$ is identifiable, where $\mathcal{F}_R$ is the family of reversed log-gamma distributions, that is:

$$\mathcal{F}_R = \left\{ R : R(x;\mu,\sigma,k) = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{\exp\{-ks - \exp(-s)\}}{\Gamma(k)} ds,\ \mu \in \mathcal{R},\ \sigma, k > 0 \right\}.$$

3.2 Application to the union of Lognormal, Gamma and Weibull families

In most theoretic results in the literature about finite mixtures, their components belong to the same parametric family. However, there exist experiences which can be modelled with mixtures of different parametric families (see e.g. Ashton 1971; Cohen 1965). For instance, the Lognormal, Gamma and Weibull models have been used for fitting the length of hospital stay (Marazzi et al. 1998). In this direction, Atienza (2003) proposes mixture models in which the previous families are involved, improving the goodness of fit. In this subsection, the identifiability of these kinds of mixtures is considered.

Let $\mathcal{F}_L$, $\mathcal{F}_G$ and $\mathcal{F}_W$ be, respectively, the Lognormal, Gamma and Weibull distributions families:

$$\begin{aligned}
\mathcal{F}_L &= \left\{ F : F(x;\mu,\sigma) = \int_0^x \frac{\exp\left\{-(\log u - \mu)^2/2\sigma^2\right\}}{\sqrt{2\pi}\sigma u} du,\ \mu \in \mathcal{R},\ \sigma > 0, x > 0 \right\} \\
\mathcal{F}_G &= \left\{ F : F(x;a,b) = \int_0^x \frac{b^{-a}}{\Gamma(a)} u^{a-1} \exp\{-u/b\}\, du,\ a, b > 0, x > 0 \right\} \\
\mathcal{F}_W &= \left\{ F : F(x;c,d) = \int_0^x \frac{c}{d^c} u^{c-1} \exp\{-u^c/d^c\}\, du,\ c, d > 0, x > 0 \right\}
\end{aligned}$$

Write $\mathcal{U} = \mathcal{F}_L \cup \mathcal{F}_G \cup \mathcal{F}_W$.

Theorem 3 *The class $\mathcal{H}_{\mathcal{U}}$ of all finite mixtures of $\mathcal{U}$ is identifiable.*

Proof Let M be the map which transforms a distribution function F into the moment generating function of $\log X$. M is linear:

$$\begin{aligned}
M[F_L(\cdot;\mu,\sigma)] &= \exp\left(t\mu + \frac{t^2\sigma^2}{2}\right), \quad t \in (-\infty, +\infty), \text{ for } F_L \in \mathcal{F}_L \\
M[F_G(\cdot;a,b)] &= \frac{b^t \Gamma(t+a)}{\Gamma(a)}, \quad t \in (-a, +\infty), \text{ for } F_G \in \mathcal{F}_G \\
M[F_W(\cdot;c,d)] &= d^t \Gamma\left(\frac{t}{c} + 1\right), \quad t \in (-c, +\infty), \text{ for } F_W \in \mathcal{F}_W
\end{aligned}$$

The point $t_0 = +\infty$ verifies condition (2) in Corollary 1. What is left is to prove that the ordering given by (3) is total on $\mathcal{U}$. Using Stirling's formula again it is easy to see:

(a) if $F(\mu_1,\sigma_1), G(\mu_2,\sigma_2) \in \mathcal{F}_L$ then:

$$G \prec F \quad \Leftrightarrow \quad [\sigma_1 < \sigma_2] \quad \text{or} \quad [\sigma_1 = \sigma_2, \mu_1 < \mu_2]$$

(b) if $F(a_1,b_1), G(a_2,b_2) \in \mathcal{F}_G$ then:

$$G \prec F \quad \Leftrightarrow \quad [b_1 < b_2] \quad \text{or} \quad [b_1 = b_2, a_1 < a_2]$$

(c) if $F(c_1, d_1), G(c_2, d_2) \in \mathcal{F}_W$ then:

$$G \prec F \quad \Leftrightarrow \quad [c_1 > c_2] \quad \text{or} \quad [c_1 = c_2, d_1 < d_2]$$

(d) if $G(\mu, \sigma) \in \mathcal{F}_L$ and $F \in (\mathcal{F}_G \cup \mathcal{F}_W)$ then:

$$G \prec F$$

(e) if $G(a, b) \in \mathcal{F}_G$ and $F(c, d) \in \mathcal{F}_W$ then:

$$G \prec F \quad \Leftrightarrow \quad [c > 1] \quad \text{or} \quad [c = 1, b > d] \quad \text{or} \quad [c = 1, b = d, a > 1]$$

(f) if $G(c, d) \in \mathcal{F}_W$ and $F(a, b) \in \mathcal{F}_G$ then:

$$G \prec F \quad \Leftrightarrow \quad [c < 1] \quad \text{or} \quad [c = 1, b < d] \quad \text{or} \quad [c = 1, b = d, a < 1]$$

Finally, if $c = 1, b = d$ and $a = 1$, the functions $F(a, b)$ and $G(c, d)$ are the same. This completes the proof. □

Acknowledgements This paper has been carried out with support by the Ministerio de Ciencia y Tecnología of Spain. Plan Nacional: I+D+I. Reference: BFM2001-3844.

References

Atienza N (2003) Mixturas de distribuciones: Modelización de experiencias con asimetría en los datos. Ph. D. Thesis. University of Seville, Spain

Ashton WD (1971) Distribution for gaps in road traffic. J Inst Math Appl 7:37–46

Barndorff-Nielsen O (1965) Identifiability of mixtures of exponential families. J Math Anal Appl 12:115–121

Böhning D (2000) Computer-assisted analysis of mixtures and applications: Meta-analysis, disease mapping and others. Chapman and Hall, BocaRaton, London

Chandra S (1977) On the mixtures of probability distributions. Scand J Stat 4:105–112

Cohen AC (1965) Estimation in mixtures of discrete distributions. In: Patil GP (ed) Classical and contagious discrete distributions. Pergamon, New York, pp 373–378

Henna J (1994) Examples of identifiable mixture. J Jpn Stat Soc 24:193–200

Khalaf EA (1988) Identifiability of finite mixtures using a new transform. Ann Inst Stat Math 40:261–265

Lindsay BG (1995) Mixture models: theory, geometry and applications. Institute of Mathematical Statistics, Hayward

Marazzi A, Paccaud F, Ruffieux C, Beguin C (1998) Fitting the distributions of length of stay by parametric models. Medical-Care 36(6):915–927

McLachlan GJ, Basford KE (1988) Mixture models. Marcel Dekker, New York

McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York

Teicher H (1963) Identifiability of finite mixtures. Ann Math Stat 34:1265–1269

Titterington DM, Smith AFM, Makov UE (1985) Statistical analysis of finite mixture distributions. Wiley, New York