Bernoulli **3**(3), 1997, 279–299

The estimation of the order of a mixture model

DIDIER DACUNHA-CASTELLE[1] and ELISABETH GASSIAT[2*]

[1]*Equipe de Modélisation Stochastique et Statistique, Unité de Recherche associée au CNRS 743, Université Paris-Sud, 91405, Orsay Cédex, France*
[2]*Equipe d'Analyse et de Probabilités, Université d'Evry, 91025 Evry Cédex, France*

We propose a new method to estimate the number of different populations when a large sample of a mixture of these populations is observed. It is possible to define the number of different populations as the number of points in the support of the mixing distribution. For discrete distributions having a finite support, the number of support points can be characterized by Hankel matrices of the first algebraic moments, or Toeplitz matrices of the trigonometric moments. Namely, for one-dimensional distributions, the cardinality of the support may be proved to be the least integer such that the Hankel matrix (or the Toeplitz matrix) degenerates. Our estimator is based on this property. We first prove the convergence of the estimator, and then its exponential convergence under wide assumptions. The number of populations is not a priori bounded. Our method applies to a large number of models such as translation mixtures with known or unknown variance, scale mixtures, exponential families and various multivariate models. The method has an obvious computational advantage since it avoids any computation of estimates of the mixing parameters. Finally we give some numerical examples to illustrate the effectiveness of the method in the most popular cases.

Keywords: Hankel matrix; mixture models; order estimation; penalization

1. Introduction

The estimation of the number of populations that compose a mixture is a classical statistical problem.

Let $\{G_\theta, \theta \in \Theta\}$ be a parametric family of distributions. We consider the mixture model

$$Q = \sum_{i=1}^{r} \pi_i G_{\theta_i} = \int G_\theta \, d\mu(\theta), \qquad \theta_i \neq \theta_j \text{ for } i \neq j,$$

where r is the order of the mixture and $\mu = \sum_{i=1}^{r} \pi_i \delta_{\theta_i}$ is a probability distribution on Θ.

Assume that we observe an n sample $X_1, \ldots, X_n$ of the distribution Q, where the parameters $(\pi_1, \ldots, \pi_r; \theta_1, \ldots, \theta_r)$ are unknown, and we want to estimate the order r. Indeed, this may be either the first step for a complete estimation of the mixture, or the question of interest in itself. This problem is known to be in general quite hard partly because estimating the parameters of the mixture, given the order r, is difficult.

*To whom correspondence should be addressed. e-mail: Elisabeth.Gassiat@math.u-psud.fr

1350-7265