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

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

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

The aim of this paper is to propose a new method of estimation for the order r having the following advantages:

(1) It leads to consistent estimators, even for non-dominated families where likelihood methods do not exist. The method applies to a wide class of mixture models.

(2) There is no need to have a priori upper bounds for the order of the mixture.

(3) It has numerical interest since it avoids the computation of estimates of the mixing parameters.

Let us first briefly survey the history of order estimation to see why it is difficult and how previous workers proposed to deal with the problem. There exists a very large literature on this subject and in particular that dedicated to clustering.

The first natural idea, when all G_θ are dominated by a common measure, is to use likelihood methods. Since the pioneering work of Akaike (1974) it is known that a penalization term has to be added to the maximum-likelihood statistic to obtain a good contrast function for the order. In the context of mixture models, this method implies various difficulties, at the theoretical as well as at the practical level. First of all, the asymptotic distribution of the maximum-likelihood statistic was not known in general. Ghosh and Sen (1985) and later Self and Lieng (1987) gave the asymptotic behaviour of likelihood ratios under a very restrictive assumption: some separability condition for the parameter values, in order to avoid difficulties due to non-identifiability. This, in particular, excluded the possibility of knowing the asymptotic of the maximum-likelihood statistic for an overestimated order. However, very recently, we gave a general likelihood theory which applies to finite mixture models. So, from a theoretical point of view, penalized likelihood techniques can be used to estimate r without any condition except that the model is dominated (Dacunha-Castelle and Gassiat 1995), and the parameters are bounded. Indeed, if they are not bounded, the likelihood statistic may not converge (see Hartigan (1985) for counterexamples). The use of these techniques requires the computation of maximum-likelihood estimators of the parameters, which appears to be very difficult. Bootstrapping may be a useful technique to study the maximum distribution (McLachlan 1987). However, for overestimated orders, it is not possible to have consistent estimators of the mixture's natural parameters. This is due to the non-identifiability of the model.

Various workers have proposed to replace the likelihood by other contrast functions. For instance, Ranneby (1984) has given a method which allows one to estimate the Kullback distance when the likelihood is not relevant. Many papers on clustering include a comparison of penalized likelihood techniques with other techniques using another kind of divergence function or distance function. This was done by Bozdogan (1994), where a long bibliography on clustering and ideas of information theory can be found.

The problem of the choice of the penalty term is of course not limited to mixture models. It is well known from mixture theory but also from other popular theories such as autoregressive moving-average models that it is difficult to choose the value of the penalty term. Different workers give experimental or heuristic justifications for this choice; for example Bock (1994) and Rissanen and Ristad (1994) used stochastic complexity as a criterion.

Another approach is that of nonparametric techniques. Izenman and Sommer (1988)

linked the order of the mixture with the number of modes of the distribution. An estimator of the order is then based on a nonparametric estimator of the number of modes. Roeder (1994) proposed a graphical technique based on sign changes of the differences of the densities. All of this applies of course only to particular mixtures. Lindsay (1989) has given a consistent moment method to estimate the order, but he did not use all the properties of this simple idea: if μ is the mixing distribution on Θ, it is possible to know the number of points of its support using Hankel moment matrices.

The organization of the paper is as follows. In Section 2, we present our general estimation method: Hankel matrices, characterization of order, and the penalty criterion. We then prove consistency and exponential convergence of the estimator. The method requires the possibility of consistently estimating algebraic moments of the mixing distribution. In Section 3 we given several examples: translation or scale mixtures, translation *and* scale mixtures, and exponential mixtures. Section 4 contains some further considerations; other criteria are presented, generalizing the idea of Hankel matrices. They could be used when prior information is available (for instance, bounds on the support of the mixing distribution μ). In particular, we propose an alternative method using the Toeplitz matrices of trigonometric moments of the mixture. We also discuss other models where this method could be exploited, as well as a discussion of likelihood methods. In Section 5, we give some numerical examples to illustrate the effectiveness of our method, especially using Toeplitz matrices. Proofs are given in the final section.

2. General method of estimation

Let us first describe the method for a one-dimensional parameter space, $\Theta \subset \mathbb{R}$. The idea is to characterize the discrete measures μ with at most r points of support using functions of their algebraic moments. Define then

$$\Phi_p = (\theta^j)_{1 \leqslant j \leqslant 2p}.$$

For any measure μ on Θ, $\mu(\Phi_p) = \int_\Theta \Phi_p(\theta)\, \mathrm{d}\mu(\theta)$. For any vector c in $\mathbb{R}^\infty$, define c^p as the vector in $\mathbb{R}^{2p}$ of the first $2p$ components of c: $c^p = (c_1, \ldots, c_{2p})$. Define $H(c^p, p)$, the Hankel matrix of order p, as the $(p+1) \times (p+1)$ matrix given by

$$H(c^p, p)_{i,j} = c^p_{i+j-2}, \qquad 1 \leqslant i, j \leqslant p+1,$$

where $c^p_0 = 1$. If μ is a measure with finite Φ_p moments, $H(\mu(\Phi_p), p)$ is said to be the Hankel matrix of μ. The use of Hankel matrices in the truncated moment problem can be found in the books of Karlin and Studden (1966) and of Krein and Nudel'man (1977). We just recall the fundamental properties that we shall use in the sequel.

Let K_p be the subset of $\mathbb{R}^{2p}$ of those vectors c^p such that there exists a positive measure μ such that $c^p = \mu(\Phi_p)$. Define on $\mathbb{R}^{2p}$ a function $L(., p)$ by

$$L(c^p, p) = \det H(c^p, p).$$

The following result characterizes the probability measures with r points of support.

Proposition 1. *$H(c^p, p)$ is a non-negative if and only if $c^p \in K_p$. Moreover, a non-negative $H(c^p, p)$ degenerates in at least one direction if and only if every probability measure μ satisfying $\mu(\Phi_p) = c^p$ is discrete and supported by at most p points.*

The proof of this proposition is very easy and relies on the following identity:

$$\forall u \in \mathbb{R}^{p+1}, \qquad u^{\mathrm{T}} \cdot H(c^p,\ p) \cdot u = \int_{\mathbb{R}} \left(\sum_{i=0}^{p} u_{i+1} \cdot x^i \right)^2 \mathrm{d}\nu(x)$$

for any (positive or non-positive) measure ν such that $\nu(\Phi_p) = c^p$.

If μ is a discrete probability measure having r points of support and if, for every integer p, $c^p = \mu(\Phi_p)$, it is then obvious that the order r is characterized by

$$r = \inf \{ p,\ L(c^p,\ p) = 0 \}. \tag{1}$$

Assume now that we have a consistent estimator $\hat{c}_n^p$ of $c^p = \mu(\Phi_p)$, based on the observations $X_1, \ldots, X_n$, supposed to be independent with common distribution Q. This is the case in many useful situations and will be developed in the forthcoming sections. Then (1) could lead us to choose the estimator of the order r as the minimizer in p of $|L(\hat{c}_n^p,\ p)|$. This in turn would lead us to choose r larger than the true value, since $|L(\hat{c}_n^p,\ p)|$ is close to 0 for sufficiently large n as soon as r is larger than the true value. To overcome this problem, we introduce a penalty term, based on Akaike's idea. Let $l(n)$ be a positive function of n, such that $\lim_{n \to +\infty} l(n) = 0$, and $A(p)$ a positive, strictly increasing function. Define the empirical penalized objective function by

$$J_n(p) = |L(\hat{c}_n^p,\ p)| + A(p) l(n). \tag{2}$$

The estimator $\hat{r}_n$ is now defined as the minimizer of J_n over all $\mathbb{N}$.

The following theorem states sufficient conditions for the consistency of $\hat{r}_n$.

Theorem 1. *Assume that*

$$\text{(C)} \qquad \forall p \in \mathbb{N},\ \hat{c}_n^p \to c^p \ \text{ and } \ \frac{1}{l(n)} [L(\hat{c}_n^p,\ p) - L(c^p,\ p)] \to 0 \ \text{ a.s.}$$

Then $\hat{r}_n \to r$ a.s.

The same result holds replacing everywhere the almost sure convergence by convergence in probability.

Remarks.

(a) We do not need an upper bound for the order because of the positivity of the contrast function $|L(.,\ p)|$, as becomes clear in the proof. Also, the estimator exists since obviously for any integer m

$$\hat{r}_n \leqslant A^{-1} \left(\frac{|L(\hat{c}_n^m,\ m)|}{l(n)} + A(m) \right),$$

where A^{-1} is the inverse function of A (A is extended here as a strictly increasing function on

the positive real numbers). In particular, this gives a recursive way to reduce the set where J_n has to be minimized. This is an important consequence of taking the absolute value of L. Indeed, with a given upper bound on the order, the analogous estimator obtained by minimizing $L(\hat{c}_n^p, p) + A(p)l(n)$ (with no absolute value) over a bounded set of integers is also consistent.

(b) The method requires the choice of the penalty term $A(p)l(n)$. As we already said in the introduction it is well known that such a choice is difficult. If one chooses a suitable Bayesian criterion (see, for example, Schwarz (1978)) or a stochastic complexity criterion, then $(p \log n)/n^{1/2}$ can be considered as optimal. However, the problem cannot be considered to be well solved from a theoretical point of view. Numerical investigations are also necessary, but they are not included in this paper.

Assumption (C) concerns the asymptotic convergence of $\hat{c}_n^p$ as an estimator of c^p, it relates $l(n)$ and the almost sure (respectively in probability) speed of convergence of $\hat{c}_n^p$ to c^p. Indeed, using a Taylor expansion of order 1, (C) holds as soon as $\{1/l(n)\}(\hat{c}_n^p - c^p) \to 0$ almost surely (respectively in probability)

Section 3 will be devoted to the choice of the estimator $\hat{c}_n^p$ of c^p. For different situations it is possible to construct $\hat{c}_n^p$ easily. An important situation is the case where $\hat{c}_n^p$ is a function of the empirical moments of the observations; if

$$c^p = f_p(\mathrm{E}\psi_p(X_t)), \tag{3}$$

then

$$\hat{c}_n^p = f_p\left(\frac{1}{n}\sum_{t=1}^{n}\psi_p(X_t)\right),$$

where ψ_p and f_p are multidimensional real functions, say that ψ_p takes values in $\mathbb{R}^{q_p}$ for an integer q_p. In general, ψ_p will be exactly Φ_p. In this case, we may determine exactly the speed of convergence, which will be exponentially fast.

Theorem 2. *Denote by F_p the function $L(f_p(.), p)$. Assume that for every integer $p \leqslant r$ the function F_p is Lipschitz with respect to some norm in $\mathbb{R}^{q_p}$, and that the generating functions $\int \exp\langle t, \psi_p(x)\rangle \, \mathrm{d}Q(x)$ are defined in a neighbourhood of 0. Assume also that*

$$n^{1/2}l(n) \to +\infty.$$

Then there exists a positive constant d such that, for every integer n larger than n_0 (which may be explicitly computed and depends on $A(.)$ and $l(.)$, and on the underlying distribution Q):

$$P(\hat{r}_n \neq r) \leqslant 2q_r \exp\{-dnl^2(n)\}.$$

Remarks. If the ψ_p are the algebraic functions, the assumption requires the distribution Q to have a bounded support. However, in Section 4 we extend the method using trigonometric moments instead of algebraic moments, and the Lipschitz assumption together with the generating functions assumption hold when the ψ_p are the trigonometric functions.

The constants involved in the theorem may be approximated using (14) given in Section 6. Of course, d is larger for well-separated populations.

The next proposition may be used to reduce the multidimensional case $\Theta \subset \mathbb{R}^s$ to the real one $\Theta \subset \mathbb{R}$.

Proposition 2. *Let μ be a discrete probability on $\mathbb{R}^s$ with r support points. For every unitary vector $v \in \mathbb{R}^s$, let $v\mu$ be the distribution of $\langle v, \theta \rangle$ ($\langle ., . \rangle$ is the usual scalar product in $\mathbb{R}^s$). Then $v\mu$ is discrete with r points of support except for at most $r(r-1)/2$ values of v.*

Proposition 2 may be used in the following way. Choose $q = (r(r-1)/2) + 1$ different unitary vectors v_i, $i = 1, \ldots, q$. Assume that estimators of the algebraic moments of all $\langle v_i, \theta \rangle$ are available. Call them $\hat{c}_n^p(v_i)$. Define then $\hat{r}_n$ as the minimizer over $\mathbb{N}$ of

$$\sum_{i=1}^{q} |L(\hat{c}_n^p(v_i),\ p)| + A(p)l(n).$$

Then, if the estimators of the involved algebraic moments satisfy the assumptions, Theorems 1 and 2 hold.

Note also that being able to estimate the algebraic moments of all possible $\langle v, \theta \rangle$, $v \in \mathbb{R}^s$ up to order p is the same as being able to estimate the moments of all monomials of coordinates of θ of degree less or equal to p, since

$$\mathrm{E}([\langle v, \theta \rangle]^k) = \sum_{k_1 + \ldots + k_s = k} \frac{k!}{k_1! \ldots k_s!} \prod_{i=1}^{s} v_i^{k_i} \mathrm{E}\left(\prod_{i=1}^{s} \theta_i^{k_i} \right).$$

3. Examples

We first propose examples where the estimation of $c^p = \mu(\Phi_p)$ is obvious, owing to the structure of the family $(G_\theta)_{\theta \in \Theta}$, such as translation mixtures and scale mixtures.

Let us again emphasize the fact that, at least for our three first examples, the family (G_θ) does not have to be dominated, so that our method applies in cases where it is not possible to apply likelihood methods.

3.1. Translation mixtures

We assume here that $\Theta \subset \mathbb{R}$. The family (G_θ) is given by

$$\mathrm{d}G_\theta(x) = \mathrm{d}G(x - \theta),$$

where G is a known probability distribution on $\mathbb{R}$. Note that the random variables X_t may be described as

$$X_t = Y_t + M_t, \tag{4}$$

where Y_t and M_t are mutually independent, Y_t has distribution G and M_t has distribution μ. We then have

$$Q(x^k) = \sum_{l=0}^{k} \frac{k!}{l!(k-l)!} G(x^l)\mu(x^{k-l}).$$

Since $G(x^l)$ is known, this relation leads to a triangular linear system for the computation of $\mu(x^k)$, $k = 1, \ldots, 2p$. If $Q(x^k)$ is estimated by

$$\hat{Q}_n(x^k) = \frac{1}{n}\sum_{t=1}^{n} X_t^k,$$

the triangular linear system may be solved with $\hat{Q}_n(x^k)$ in place of $Q(x^k)$, leading to a consistent estimate $\hat{c}_n^p$ of $(\mu(x^k))_{k=1,\ldots,2p}$, with the same asymptotic properties as $\hat{Q}_n(x^k)$: central limit theorem and law of the iterated logarithm. Then we have the following.

Proposition 3. *As soon as, for all integer p,* $\int x^p \, \mathrm{d}G(x) < +\infty$,

(i) (C) *holds with* $n^{1/2}(\log\log n)^{-1/2} l(n) \to +\infty$ *for the a.s. convergence,*

(ii) (C) *holds with* $n^{1/2} l(n) \to +\infty$ *for the convergence in probability and Theorem 1 applies and*

(iii) *Theorem 2 applies as soon as for all integer p,* $\mathrm{E}(\exp \gamma_p X^{2p}) < \infty$ *for some* $\gamma_p > 0$, *where X has distribution G.*

When G is an s-dimensional distribution as well as $\Theta \subset \mathbb{R}^s$, Proposition 2 may be used by taking a scalar product in (4) with any s-dimensional vector v.

3.2. Scale mixtures

Here, $\Theta = \mathbb{R}^+$ and the family (G_θ) is given by

$$\mathrm{d}G_\theta(x) = \mathrm{d}G\left(\frac{x}{\theta}\right),$$

where G is a known probability distribution over $\mathbb{R}$. Note that the random variables X_t may now be described by

$$X_t = \sigma_t Y_t$$

where σ_t and Y_t are mutually independent, Y_t has distribution G and σ_t has distribution μ. Then we have

$$\forall k \in \mathbb{N},\ \mathrm{E}(X_t^k) = \mathrm{E}(\sigma_t^k)\cdot\mathrm{E}(Y_t^k),$$

$$Q(x^k) = G(x^k)\cdot\mu(x^k).$$

If, for all integer k, $G(x^k) \neq 0$, $\mu(x^k)$ is estimated by $\hat{Q}_n(x^k)/G(\varphi_k)$ and Proposition 3 holds. If, for some integer k, $G(x^k) = 0$, we take squares everywhere, so that all $G(x^k)$ are replaced by $G(x^{2k})$ which are now always positive. The method is the same, replacing G and μ by their images using the application $x \to x^2$, and Proposition 3 holds again.

3.3. Covariance mixtures

Here we assume that Q is a distribution on $\mathbb{R}^s$, and $\Theta \subset \mathscr{M}_s$, where $\mathscr{M}_s$ is the space of $s \times s$ symmetric matrices. Q is given by

$$Q = \sum_{i=1}^{r} \pi_i \cdot G(\Sigma_i^{-1} \cdot)$$

where Σ_i^{-1}, $i = 1, \ldots, r$ are r different covariance matrices and G a given distribution in $\mathbb{R}^s$. With no loss of generality we may assume that the covariance matrix of G is the identity. The model may be described by

$$X_t = S_t \cdot Y_t,$$

where Y_t and S_t are mutually independent, Y_t is a random vector with distribution G, S_t is a random matrix with distribution μ on a space of dimension $s(s+1)/2$. It is easy to see that

$$\mathrm{E}(X_1 \cdot X_1^{\mathrm{T}}) = \mathrm{E}(S_1^2)$$

and for every integer k

$$\mathrm{E}((X_1 \cdot X_1^{\mathrm{T}})^k) = \mathrm{E}(S_1^{2k})\mathrm{E}((Y_1 \cdot Y_1^{\mathrm{T}})^k)$$

so that all moments of S_1^2 can be estimated. Now, the distribution of S_1^2 has the same number of points of support as that of S_1. Our method thus applies using Proposition 2.

3.4. Translation mixtures with unknown scale

Let now Q be a mixture of a translation family $(G(. - \theta))$ with an unknown scaling σ_0:

$$Q(\mathrm{d}x) = \sum_{i=1}^{r} \pi_i \, \mathrm{d}G\left(\frac{x - \theta_i}{\sigma_0}\right) \tag{5}$$

Here again, G is known. We cannot directly use Hankel's criterion for this model. So we have to generalize the method for unknown scale parameters. This generalization holds when the distribution G is in a certain class that we define now.

Definition. *$\mathscr{L}$ is the set of distributions G with the following property. If U is a random variable with distribution G, for any real c in $]0, 1[$, there exist two independent random variables $\tilde{U}$ and V_c, where $\tilde{U}$ has distribution G and $U = c\tilde{U} + V_c$.*

The class $\mathscr{L}$ is characterized at the end of this section in Proposition 5. Just note here that Gaussian distributions are in class $\mathscr{L}$.

Let (5) hold with G in class $\mathscr{L}$. We have the following triangular system relating the algebraic moments of μ to those of Q:

$$\forall k = 1, \ldots, 2p, \; Q(x^k) = \sum_{l=0}^{k} \frac{k! \sigma_0^{\,l}}{l!(k-l)!} G(x^l)\mu(\theta^{k-l}).$$

If σ is assumed to be a value of the scaling parameter, define $\mu(\theta^j, \sigma)$, $j = 1, \ldots, 2p$ as the solution of the following triangular system:

$$\forall k = 1, \ldots, 2p,\ Q(x^k) = \sum_{l=0}^{k} \frac{k!\sigma^l}{l!(k-l)!} G(x^l)\mu(\theta^{k-l}, \sigma). \tag{6}$$

Now let $H(\sigma, p)$ be the Hankel matrix built with the pseudo-moments $\mu(\theta^j, \sigma)$, $j = 1, \ldots, 2p$. (We call them pseudo-moments since it may happen that they are not moments of a positive measure.) We have the following characterization of the scaling factor σ_0 and of the order r.

Theorem 3.

(i) $\forall \sigma < \sigma_0,\ \forall p$: $\det\{H(\sigma, p)\} > 0$.

(ii) *For* $\sigma = \sigma_0$, $\det\{H(\sigma, p)\} = 0$ *if and only if* $p \geqslant r$.

Theorem 3 is proved in Section 6. Its interpretation is the following: (σ_0, r) is the smallest root of the equation $|\det\{H(\sigma, p)\}| = 0$, where 'smallest' means smallest in all directions. It results in the idea of estimating σ_0 by compensating the criterion, leading to the contrast function K_n:

$$K_n(p, \sigma) = |\det\{\hat{H}_n(\sigma, p)\}| + A(p)l_1(n) + \sigma M l_1(n).$$

Here, $l_1(n)$ is a positive functions decreasing to 0 when n goes to infinity and M a positive real number. $\hat{H}_n(\sigma, p)$ is the Hankel matrix built with the solution of the triangular system (6) where $Q(x^k)$ is replaced by $\hat{Q}_n(x^k)$. Then $\hat{r}_n$ and $\hat{\sigma}_n$ are defined by

$$K_n(\hat{r}_n, \hat{\sigma}_n) = \min\{K_n(p, \sigma);\ p \in \mathbb{N}, \sigma \in \mathbb{R}^+\}.$$

We have the following.

Proposition 4. *For sufficiently large M, Theorem 1 holds for* $\hat{r}_n$. *Moreover,* σ_n *is a consistent estimator of* σ_0 *under the same conditions as for* $\hat{r}_n$.

The proof of the consistency of $\hat{\sigma}_n$ follows the same lines as that of $\hat{r}_n$ and will be omitted.

Remark. The characterization given in Theorem 3 may be used for mixture models with additive noise and with unknown signal-to-noise ratio. That is, when the observations are given by

$$Z_t = X_t + \sigma_0 U_t = M_t + Y_t + \sigma_0 U_t,$$

X_t has a translation mixture distribution, and U_t is independent of X and has a distribution in class $\mathscr{L}$; σ_0 is unknown. This model is of course useful in signal theory.

We recall now a different characterization of class $\mathscr{L}$ (Petrov 1975, p. 83, Lemma 12).

Proposition 5. *$\mathscr{L}$ is the set of infinitely divisible real distributions with arbitrary Gaussian part and with absolutely continuous Levy measure, with density f, such that $xf(x)$ decreases on $\mathbb{R}^+$ and $\mathbb{R}^-$.*

3.5. Mixtures of exponential families

Let us first give some general considerations for $\Theta \subset \mathbb{R}^s$. To estimate the multidimensional moments of μ, we can try to find these moments using the structure of the statistical model $G_\theta(\mathrm{d}x)$. This means that, for every $k \in \mathbb{N}^s$, $k = (k_1 \ldots, k_s)$, we need a function ψ_k such that

$$\theta^k = \int \psi_k(x)\, \mathrm{d}G_\theta(x).$$

Here θ^k means $\theta_1^{k_1} \ldots \theta_s^{k_s}$. This relation implies that

$$\int \theta^k \,\mathrm{d}\mu(\theta) = \int \psi_k(x)\, \mathrm{d}Q(x),$$

so that $\mu(\theta_k)$ can be obtained as a linear functional of the distribution Q.

Consider the situation where we can write

$$G_\theta(\mathrm{d}x) = \sum_{k \in \mathbb{N}^s} \theta^k \psi_k(x) \nu(\mathrm{d}x), \tag{7}$$

where the series converges in $L^1(\nu) \cap L^2(\nu)$. Suppose moreover that (ψ_k) is a free and total system in $L^2(\nu)$. Then the classical theory of biorthogonal systems (Brezinski 1992) allows us to obtain (ψ_k) using the relations $\int \psi_k(x)\psi_l(x)\,\mathrm{d}\nu(x) = \delta_k^l$, $k, l \in \mathbb{N}^s$. Here δ_k^l is the Kronecker symbol.

In general, the ψ_k need not be distinct for distinct k. We give an example below. If $k \neq k'$ implies that $\psi_k \neq \psi_{k'}$, then all moments $\mu(\theta^k)$ can be estimated and our method applies.

Now let us discuss the case where (G_θ) is an exponential family, in order to have a description of what can happen:

$$G_\theta(\mathrm{d}x) = \exp(-\phi(\theta)) \exp(\langle \theta, x \rangle) S(\mathrm{d}x),$$

where S is a positive measure on $\mathbb{R}^s$. We suppose that the support of S has a non-empty interior. Then $\Theta \subset \mathbb{R}^s$ is defined by

$$\Theta = \left\{ \theta, \int_{\mathbb{R}^s} \exp(\langle \theta, x \rangle)\, \mathrm{d}S(x) = \exp\{\phi(\theta)\} < \infty \right\}.$$

We assume that the interior of Θ is not empty. Then we have

$$\exp\{\phi(\theta)\} G_\theta(\mathrm{d}x) = \sum_{n=0}^{\infty} \frac{\langle \theta, x \rangle^n}{n!} S(\mathrm{d}x)$$

$$= \sum_{n=0}^{\infty} \sum_{h_1 + \ldots + h_s = n} C(h_1, \ldots, h_s) \theta_1^{h_1} \ldots \theta_s^{h_s} \cdot x_1^{h_1} \ldots x_s^{h_s} S(\mathrm{d}x),$$

for suitable constants $C(h_1, \ldots, h_s)$. Now let $\tilde{\mu}$ be defined by $(\mathrm{d}\tilde{\mu}/\mathrm{d}\mu)(\theta) = \exp\{-\phi(\theta)\}$. Then μ is discrete with r points of support if and only if the same holds for $\tilde{\mu}$. We estimate $\tilde{\mu}(\theta^k)$ by $\hat{Q}_n(x^k)/S(x^k\psi_k(x))$. The set of functions x^k is a free and total system but, to recover all the moments of θ, we require that they be different, which is not necessarily true. We can illustrate the difficulty by the case of the general mixture of Gaussian distributions $\mathcal{N}(m, \sigma)$. Let $\theta = (\theta_1, \theta_2)$, $\theta_1 = -m/2\sigma^2$, $\theta_2 = 1/2\sigma^2$, $x = (x_1, x_2)$, $x_1 = u$, $x_2 = u^2$. To compute for example all the moments of θ of order two, we have to compute five moments (θ_1, θ_1^2, θ_2, θ_2^2 and $\theta_1\theta_2$) based on four moments of u (u, u^2, u^3 and u^4), which is impossible. Here, our method does not apply.

For full exponential families, the linear dimension of the support of S equals the dimension of Θ, and our method applies. The same remains true for curved families if there is a unique representation of x^k. This is true for instance for the beta family:

$$B^{-1}(\alpha, \beta)x^{\alpha}(1 - x)^{\beta}1_{0,1}(x)\,\mathrm{d}x.$$

4. Further considerations

It appears that the proposed method applies in many other situations. Indeed, if one wishes to estimate a discrete number, which may be modelled as the number of points of support of some variables, if the observations allow one to estimate Φ moments of this variable, then one may use methods similar to those proposed here. Deconvolution of discrete signals may be solved using Hankel methods (Gamboa and Gassiat 1994), even with additive noise (Gassiat and Gautherat 1994). Source separation of discrete signals or circular complex signals may also be performed via Hankel methods (Gamboa and Gassiat 1995).

4.1. Generalization of the method to other sets of moments

Here Θ is assumed to be a bounded interval. By translation and scaling, it is sufficient to consider the case $\Theta = [0, 2\pi]$. We then replace the algebraic functions by the set of functions:

$$\Phi_p = (\varphi_k(x))_{1\leqslant k\leqslant p} = (\exp \mathrm{i}kx)_{1\leqslant k\leqslant p}.$$

For any complex vector c^p in $\mathbb{C}^p$, let $T(c^p, p)$ be the $(p + 1) \times (p + 1)$ Toeplitz matrix of order p, i.e.,

$$\forall i, j \in \{1, \ldots, p + 1\}\, T(c^p, p)_{i,j} = \begin{cases} c^p_{i-j}, & \text{if } i - j \geqslant 0, \\ c^p_{j-i}, & \text{if } i - j < 0. \end{cases}$$

Let $\tilde{K}_p$ be the subspace of $\mathbb{C}^p$ defined by $c \in \tilde{K}_p$ if and only if there exists a positive measure μ on $[0, 2\pi]$ such that $c = \mu(\Phi_p)$. Now let $\tilde{L}(c^p, p) = \det T(c^p, p)$. We then have the following proposition, similar to Proposition 1.

Proposition 6. *$T(c^p, p)$ is non-negative if and only if $c \in \tilde{K}_p$. Moreover, $T(c^p, p)$ is non-negative and degenerates in at least one direction if and only if every probability measure μ satisfying $\mu(\Phi_p) = c$ is discrete and supported by at most p points.*

Proof (see Krein and Nudel'man (1977)). A similar criterion may then be used, replacing L by $\tilde{L}$ everywhere, with the same properties. □

Let us just show how it applies for the translation mixture model

$$X_t = Y_t + M_t,$$

where M_t is a discrete random variable known to lie in $[0, 2\pi]$, and Y_t has distribution G. We obviously have $Q(\varphi_k) = G(\varphi_k) \cdot \mu(\varphi_k)$. Assume that $\forall k \in \mathbb{N}$, $G(\varphi_k) \neq 0$. Then $Q(\varphi_k)$ is estimated by

$$\hat{Q}_n(\varphi_k) = \frac{1}{n} \sum_{t=1}^{n} \varphi_k(X_t),$$

so that the estimation $\hat{c}_n^p$ of $c^p = \mu(\Phi_p)$ will be defined as

$$\hat{c}^p = \left(\frac{\hat{Q}_n(\varphi_k)}{G(\varphi_k)} \right)_{1 \leqslant k \leqslant p}$$

and $\hat{c}_n^p$ is obviously consistent. We then have the following.

Proposition 7. *If $G(\varphi_k) \neq 0$, then*

(i) (C) *holds as soon as $n^{1/2} (\log \log n)^{-1/2} l(n) \to +\infty$ for the a.s. convergence and*
(ii) (C) *holds as soon as $n^{1/2} l(n) \to +\infty$ for the convergence in probability and Theorem 1 applies.*

Moreover, Theorem 2 applies without any restriction.

Proof. For all p, $\tilde{L}(\cdot, p)$ is infinitely differentiable. Assumption (C) results from a Taylor expansion of order 1, and from the law of the iterated logarithm for independent sequences for the a.s. convergence, or from the central limit theorem for the convergence in probability. □

A closer look at Toeplitz and Hankel methods shows that it involves two different kinds of argument. The first argument concerns the moments that are used, the fact that polynomials have a particular behaviour with respect to the number of zeros. This property generalizes to Chebyshev systems of functions. The second argument is the possibility of exhibiting a computational functional to discriminate particular points of the closed convex hull of the moment functions (determinant of the Hankel or Toeplitz forms). In the case of a bounded interval, various examples of other criteria may be given. A large family of such criteria may be found in the work of Gamboa and Gassiat (1993).

4.2. Extension of the method to other models

As another example, let us consider the case of centred stationary Gaussian processes X with a finite discrete spectral measure μ, where the number of frequencies has to be estimated (in this example the observations are correlated). So we can write

$$\mu(\mathrm{d}\lambda) = \sum_{l=1}^{r} p_l \delta_{\theta_l}(\mathrm{d}\lambda),\ 0 \leqslant p_l,\ \theta_l \in [0,\ 2\pi],\ \theta_l \neq \theta_{l'} \text{ for } l \neq l'.$$

Then X_j, $j \in \mathbb{N}$, has the representation

$$X_j = \sum_{l=1}^{r} A_l \, \mathrm{e}^{ij\theta_l},$$

where A_l, $l = 1, \ldots, r$, are independent centred Gaussian random variables with variance p_l. The parameter of interest r, the number of different frequencies, has to be estimated on the basis of the observations $X_1, \ldots, X_n$. Let ν be the random measure:

$$\nu(\mathrm{d}\lambda) = \sum_{l=1}^{r} A_l^2 \delta_{\theta_l}.$$

Let $\hat{e}_n(k)$ be the empirical covariance of X:

$$\hat{e}_n(k) = \frac{1}{n-k} \sum_{j=1}^{n-k} X_{j+k} \bar{X}_j,$$

$$\hat{e}_n(k) = \sum_{l=1}^{r} A_l^2 \, \mathrm{e}^{ik\theta_l} + \mathrm{O}\left(\frac{1}{n}\right).$$

$\hat{e}_n$ gives an estimator of the Fourier coefficients of the random measure ν. Now, A_l^2 is a.s. positive, so that ν is a.s. a discrete random measure with r points of support, and the Toeplitz criterion may be used to estimate r.

4.3. Penalized likelihood methods

Assume that there exists a positive measure ν that dominates the family $(G_\theta)_{\theta \in \Theta}$, so that, if $\mu = \sum_{i=1}^{r} \pi_i G_{\theta i}$,

$$Q_{\theta^r, \pi^r}(.) = \sum_{i=1}^{r} \pi_i G_{\theta_i} \ll \nu,$$

where $\theta^r = (\theta_i)_{1 \leqslant i \leqslant r}$, $\pi^r = (\pi_i)_{1 \leqslant i \leqslant r}$, and

$$\frac{\mathrm{d}Q_{\theta^r, p^r}}{\mathrm{d}\nu}(.) = \sum_{i=1}^{r} \pi_i \frac{\mathrm{d}G_{\theta i}}{\mathrm{d}\nu}(.) = q(\theta^r, \pi^r, .).$$

Define the log-likelihood L_n as

$$L_n(r, \theta^r, \pi^r) = \sum_{t=1}^{n} \log q(\theta^r, \pi^r, X_t).$$

The penalized maximum likelihood can be defined as usual:

$$J_n(p) = -k(n) \max_{\theta^p, \pi^p} l_n(p, \theta^p, \pi^p) + A(p),$$

where $k(n)$ is a positive function tending to 0 when n goes to infinity, and $A(p)$ is a strictly increasing function of the order p.

As we already mentioned in the introduction, penalized likelihood methods have been extensively studied (although not with general theoretical results). Of course, the problem is that the estimators of θ^p and π^p are not consistent when we are near the boundary, i.e., for $p > r$. We suggest that a good procedure could be the following.

(1) Estimate the order using the Hankel criterion.

(2) Then test the order r against $r+1$ for instance using likelihood tests as they are developed in our paper (Dacunha-Castelle and Gassiat 1995). The likelihood statistic has a distribution which is asymptotically a function of the supremum of a Gaussian process. This distribution has to be tabulated using simulations.

5. Numerical experiments

The experiments were conducted for Gaussian translation mixtures with known variance. For a small true number of populations, i.e., up to six populations, the method seems to be very efficient, and not very sensitive to the choice of the penalty term. For three different populations, where the density is

$$\frac{\mathrm{d}Q(x)}{\mathrm{d}x} = \sum_{i=1}^{3} \frac{\pi_i}{0.5(2\pi)^{1/2}} \exp\left(-\frac{(x-\theta_i)^2}{0.5}\right),$$

where the probability mixture is

$$0.3\delta_{-0.5} + 0.5\delta_0 + 0.2\delta_{1.3}.$$

The density of Q has the representation given in Figure 1. With 100 independent realizations and with 500 observations we obtained the following estimators: $\hat{r} = 2$ 13 times, $\hat{r} = 3$ 72 times and $\hat{r} = 4$ 15 times.

Let us discuss experiments with a large number of different populations. First of all, we chose to apply the method using Toeplitz matrices as described in Section 4. Indeed, for Hankel methods, we have to compute algebraic moments at least up to order $2r+2$. The empirical moments then have a variance which is governed by the moment of order $4r+4$, and the moments of a Gaussian variable grow exponentially rapidly. The number of observations that are required to have a good accuracy on the estimators of the moment is

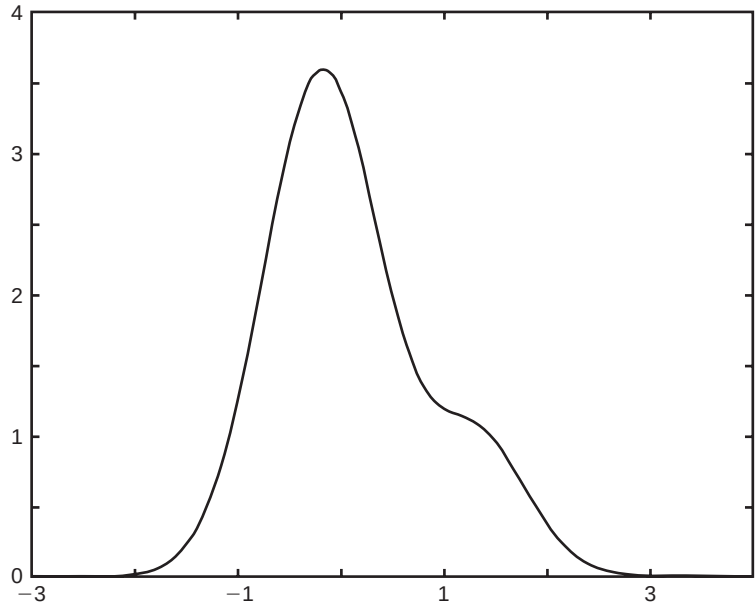

Figure 1. Density of the mixture.

then very large if empirical moments are used as estimators. This is not the case for trigonometric moments. Another possibility could be to refine the estimation using bootstrap techniques; we leave this for further work. Here we just want to show that, for simple examples, our method gives good results with very fast computations.

We simulated a mixture with eight different Gaussian populations, with the following density:

$$\frac{\mathrm{d}Q(x)}{\mathrm{d}x} = \sum_{i=1}^{8} \frac{\pi_i}{0.5(2\pi)^{1/2}} \exp\left(-\frac{(x-\theta_i)^2}{0.5}\right),$$

where the probability mixture is

$$0.1\delta_{-4} + 0.15\delta_{-2} + 0.30\delta_{-0.5} + 0.1\delta_1 + 0.1\delta_{2.5} + 0.07\delta_3 + 0.08\delta_3 + 0.08\delta_{3.7} + 0.1\delta_{4.5}.$$

The density of Q has the representation given in Figure 2.

The number of populations is not related to the number of modes. First of all, the choice of the penalty term appears to be quite important. We noticed that the distance between the points of the mixture distribution had an influence on the estimation; in some sense, there is a need to compromise between good separation of the points in the mixing distribution (which can be increased by rescaling) and good accuracy in the empirical estimation of the trigonometric moments. We then considered a change in the scale of the observations. Indeed, whereas scaling with a large parameter leads to better separation of the populations,

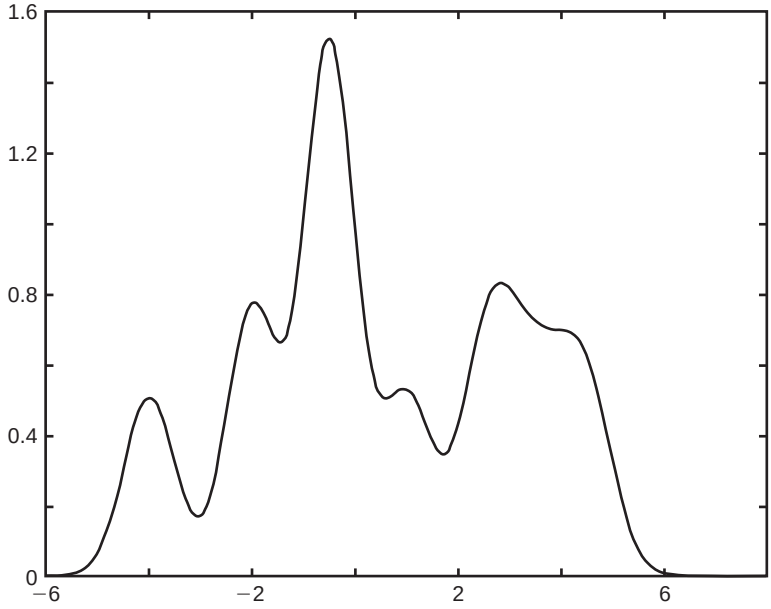

Figure 2. Density of the mixture.

scaling with a small parameter leads to better accuracy of the estimators. The choice of the rescaling factor is a result of a preliminary numerical investigation that will be fully developed elsewhere. However, it leads us here to multiply the observations by a factor of 0.39.

This scaling factor is a first attempt to good setting of the penalty term. Indeed we observed that good rescaling made the estimation process less sensitive to the choice of the penalty term. This is a qualitative decision rule for the scaling factor (here 0.39). We then considered 20 independent realizations for $n = 500$, 1000, 1500 and 2000 observations. The penalty function is given in Figure 3.

The estimation of the number of populations is as follows:

for $n = 500$, [8; 8; 8; 8; 8; 8; 8; 8; 10; 8; 8; 8; 8; 8; 8; 8; 8; 7; 7; 8],
for $n = 1000$, [8; 8; 9; 8; 8; 8; 8; 8; 8; 8; 8; 8; 8; 8; 8; 9; 8; 8; 8; 8],
for $n = 1500$, [8; 8; 8; 8; 8; 8; 8; 8; 9; 8; 8; 8; 8; 8; 8; 8; 8; 8; 7; 8],
for $n = 2000$, [8; 8; 8; 8; 8; 8; 7; 8; 8; 8; 8; 8; 9; 8; 9; 7; 8; 8; 8; 8].

The choice of the penalty term requires obviously strong computational investigations, which will be the object of a complete numerical further work. Our preliminary experiments just make it obvious that it has to be carefully chosen when the number of populations increases, and that a choice of some rescaling number may help.

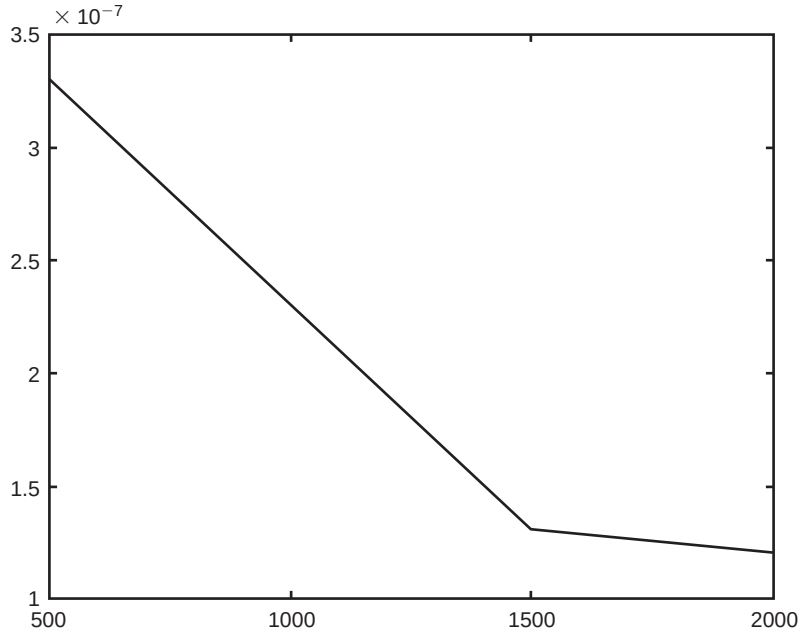

Figure 3. Penalty function of n.

6. Proofs

Proof of Theorem 1. We give the proof of the almost sure convergence, the convergence in probability being the same replacing everywhere 'a.s.' by 'in probability':

$$J_n(p) - J_n(r) = |L(\hat{c}_n^p,\ p)| - |L(\hat{c}_n^r,\ r)| + l(n)\{A(p) - A(r)\}.$$

Now, if $p \in \{1, \ldots, r-1\}$,

$$J_n(p) - J_n(r) = |L(\hat{c}_n^p,\ p)| - |L(\hat{c}_n^r,\ r) - L(c,\ r)| + l(n)\{A(p) - A(r)\},$$

$L(.,\ p)$ being continuous, $\hat{c}_n^p \to c^p$ a.s., $\lim_{n\to+\infty} L(\hat{c}_n,\ p) = L(c,\ p)$ a.s., and $L(c,\ p) > 0$; see Proposition 1. Thus

$$\forall p \in \{1, \ldots, r-1,\ \},\ \lim_{n\to+\infty} \frac{J_n(p) - J_n(r)}{l(n)} = \lim_{n\to+\infty} \frac{|L(\hat{c}_n^p,\ p)|}{l(n)} = +\infty \text{ a.s.},$$

which clearly implies that

$$\liminf_{n\to\infty} \hat{r}_n \geqslant r \text{ a.s.} \tag{8}$$

Now, $\hat{r}_n \geqslant r$ implies that $J_n(\hat{r}_n) \leqslant J_n(r)$, which in turns implies that

$$A(\hat{r}_n) \leqslant \frac{|L(\hat{c}_n^r,\ r)|}{l(n)} + A(r)$$

and using Assumption (C) and the fact that A is strictly increasing

$$\limsup_{n\to\infty} \hat{r}_n \leqslant r \text{ a.s.} \tag{9}$$

Now (8) and (9) imply obviously the a.s. convergence of $\hat{r}_n$. □

Proof of Theorem 2. Let us now prove the exponential convergence:

$$P(\hat{r}_n \neq r) \leqslant \sum_{p=1}^{r-1} P(J_n(p) < J_n(r)) + P(\hat{r}_n > r).$$

For any $p < r$ we have

$$P(J_n(p) < J_n(r)) = P(|L(\hat{c}_n^p,\ p)| - L(c^p,\ p) < l(n)\{A(r) - A(p)\} + |L(\hat{c}_n^r,\ r)| - L(c^p,\ p)).$$

Take $n \geqslant n_p$ such that

$$l(n_p)\{A(r) - A(p)\} \leqslant \frac{L(c^p,\ p)}{3};$$

we have

$$P(J_n(p) < J_n(r)) \leqslant P\left(L(\hat{c}_n^p,\ p) - L(c^p,\ p) \leqslant \frac{L(c^p,\ p)}{3}\right) + P\left(|L(\hat{c}_n^r,\ r)| \geqslant \frac{L(c^p,\ p)}{3}\right).$$

Now, for any integer n,

$$P(\hat{r}_n > r) \leqslant P(|L(\hat{c}_n^r,\ r)| \geqslant l(n)\{A(r+1) - A(r)\}).$$

Define $n^* = \max_{1\leqslant p\leqslant r-1} n_p$. We have for $n \geqslant n^*$

$$\begin{aligned} P(\hat{r}_n \neq r) \leqslant \sum_{p=1}^{r-1} \left\{ P\left(L(\hat{c}_n^p,\ p) - L(c^p,\ p) \leqslant -\frac{L(c^p,\ p)}{3}\right) + P\left(|L(\hat{c}_n^r,\ r)| \geqslant \frac{L(c^p,\ p)}{3}\right)\right\} \\ + P(|L(\hat{c}_n^r,\ r)| \geqslant l(n)\{A(r+1) - A(r)\}). \end{aligned} \tag{10}$$

Since all norms in $\mathbb{R}^{q_p}$ are equivalent, now let M_p denote the Lipschitz factor of F_p with respect to the l_1 norm:

$$\left|F_p\left(\frac{1}{n}\sum_{t=1}^{n} \psi_p(X_t)\right) - F_p(Q(\psi_p))\right| \leqslant M_p \left\|\frac{1}{n}\sum_{t=1}^{n} \psi_p(X_t) - Q(\psi_p)\right\|,$$

where $\|\cdot\|$ is the l_1 norm in $\mathbb{R}^{q_p}$. We have for any positive real number u

$$P\left(\left\|\frac{1}{n}\sum_{t=1}^{n} \psi_p(X_t) - Q(\psi_p)\right\| \geqslant u\right) \leqslant \sum_{k=1}^{q_p} P\left(\left|\frac{1}{n}\sum_{t=1}^{n} \psi_{p,k}(X_t) - Q(\psi_{p,k})\right| \geqslant \frac{u}{q_p}\right), \tag{11}$$

where $\psi_{p,k}$ is the kth component of ψ_p. Then usual Markov inequality gives for any positive real number u, and any integers p and k,

$$P\left(\frac{1}{n}\sum_{t=1}^{n}\psi_{p,k}(X_t) - Q(\psi_{p,k}) \geqslant u\right) \leqslant \exp\{-nh_{p,k}(u)\}, \tag{12}$$

$$P\left(-\frac{1}{n}\sum_{t=1}^{n}\psi_{p,k}(X_t) - Q(\psi_{p,k}) \leqslant -u\right) \leqslant \exp\{-nh_{p,k}(-u)\}, \tag{13}$$

where $h_{p,k}$ is the Cramer transform of the logarithm of the generating function of $\psi_{p,k}(X) - \mathrm{E}\psi_{p,k}(X)$ given by

$$h_{p,k}(u) = \inf_{\delta\in\mathbb{R}} [\log E \exp\{\delta(\psi_{p,k}(X) - \mathrm{E}\psi_{p,k}(X))\} - \delta\cdot u].$$

Moreover, since the generating functions $\int \exp < t,\ \psi_p(x) > \mathrm{d}Q(s)$ are defined on a neighbourhood of 0, we have for all p and k: $h_{p,k}(0) = 0,\ h'_{p,k}(0) = 0,\ h''_{p,k}(0) > 0$.

Now, using (10)–(13) we have, for $n \geqslant n^*$,

$$\begin{aligned} P(\hat{r}_n \neq r) \leqslant \sum_{p=1}^{r-1}\Bigg[\sum_{k=1}^{q_p}\exp\left\{-nh_{p,k}\left(\frac{L(c^p,\ p)}{2M_pq_p}\right)\right\} + \sum_{k=1}^{q_r}\exp\left\{-nh_{r,k}\left(\frac{L(c^p,\ p)}{M_rq_r}\right)\right\} \\ + \sum_{k=1}^{q_r}\exp\left\{-nh_{r,k}\left(\frac{L(c^p,\ p)}{M_rq_r}\right)\right\}\Bigg] + \sum_{k=1}^{q_r}\exp\left\{-nh_{r,k}\left(l(n)\frac{A(r+1)-A(r)}{2M_rq_r}\right)\right\} \\ + \sum_{k=1}^{q_r}\exp\left\{-nh_{r,k}\left(l(n)\frac{A(r+1)-A(r)}{2M_rq_r}\right)\right\}. \end{aligned}$$

Define

$$m = \min\left\{h_{p,k}\left(\frac{L(c^p,\ p)}{3M_pq_p}\right),\ h_{p,k}\left(\frac{L(c^p,\ p)}{3M_rq_r}\right),\ h_{p,k}\left(-\frac{L(c^p,\ p)}{3M_rq_r}\right)\right\},$$

and

$$d_0 = \frac{\{A(r+1)-A(r)\}^2}{M_rq_r}\cdot\inf_{k\leqslant q_r} h''_{r,k}(0); \tag{14}$$

we then have

$$\begin{aligned} P(\hat{r}_n \neq r) \leqslant \left(\sum_{p=1}^{r-1} q_p + 2(r-1)q_r\right)\exp(-nm) \\ + 2q_r\exp\{-nl(n)^2d_0\}\{1+\mathrm{o}(1)\}. \end{aligned}$$

The dominant term in the inequality is the last term, which has order

$$\exp\{-nl(n)^2d_0\}.$$

This gives the theorem, by taking an appropriate $d < d_0$ to take all the first terms into

account, and $n \geqslant n_0$ where $n_0 \geqslant n^*$ is such that all terms are not larger than $\exp\{-nl(n)^2 d\}$. □

Proof of Theorem 3. We have

$$X_t = M_t + \sigma_0 \cdot Y_t,$$

where M_t has distribution μ, Y_t has distribution G and X_t has distribution Q.

For any real $c \in]0, 1[$ using the definition of class $\mathscr{L}$, the following equality holds in distribution:

$$X_t = M_t + c\sigma_0 Y_t + \sigma_0 V_c$$

so that $\det H(c\sigma_0, p)$ is the determinant of the Hankel matrix of the moments of the variable $\theta_t + \sigma_0 V_c$, which has strictly more than r points of support, and the theorem follows. □

Acknowledgements

The authors wish to thank L. Traverse for conducting the numerical experiments, and anonymous referees for constructive remarks on a first manuscript.

References

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Automat. Control*, **19**, 716–723.

Bock, H.H. (1994). Information and entropy in cluster analysis. In H. Bozdogan (ed.), *Proceedings of the First US–Japan Conference on the Frontiers of Statistical Modeling*, pp. 115–147. Deventer: Kluwer.

Bozdogan, H. (1994) Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. In H. Bozdogan (ed.) *Proceedings of the First US–Japan Conference on the Frontiers of Statistical Modeling*, pp. 69–113. Deventer: Kluwer.

Brezinski, C. (1992). *Biorthogonality and its Application to Numerical Analysis*. New York: Marcel Dekker.

Dacunha-Castelle, D. and Gassiat, E. (1997). Testing in locally conic models. *ESAIM Probab. Statist.* To appear.

Gamboa, F. and Gassiat, E. (1997). Bayesian methods and Maximum entropy for ill posed inverse problems. *Ann. Statist.*, **25**, 328–350.

Gamboa, F. and Gassiat, E. (1996). Blind deconvolution of discrete linear systems. *Ann. Statist.*, **24**, 1964–1981.

Gamboa, F. and Gassiat, E. (1995). Source separation when the input sources are discrete or have constant modulus. *IEEE Trans. Signal Processing*. Submitted.

Gassiat, E. and Gautherat, E. (1997). Identification of noisy linear systems with discrete random input. *IEEE Trans. Inf. Theory.* To appear.

Ghosh, J.K. and Sen, P.K. (1985). On the asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results. In L. Le Cam and R. Olshem (eds), *Proceedings of the*

Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II, pp. 789–806. VT: Wadsworth.

Hartigan, J. (1985). A failure of likelihood asymptotics for normal mixtures. In L. Le Cam and R. Olshen (eds), *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, Vol. II, pp. 807–811. VT: Wadsworth.

Izenman, A.J. and Sommer, C. (1988). Philatelic mixtures and multivariate densities. *J. Amer. Math. Soc.*, **83**, 94.

Karlin, S. and Studden, W.J. (1966). *Tchebychev Systems with Applications in Analysis and Statistics.* New York: Wiley.

Krein, M.G. and Nudel'man, A. (1977) *The Markov Moment Problem and Extremal Problems.* Providence, RI: American Mathematical Society.

Lindsay, B.G. (1989) Moment matrices: application in mixtures. *Ann. Statist.*, **17**, 722–740.

McLachlan, G.J. (1987). On bootstrapping the likelihood ratio test statistics for the number of components in a normal mixture. *Appl. Statist.*, **36**, 318–324.

Petrov, V.V. (1975) *Sums of Independent Random Variables.* Berlin: Springer-Verlag.

Ranneby, B.O. (1984). The maximum spacing method. An estimation method related to the maximum likelihood method. *Scand. J. Statist.*, **11**, 93–112.

Rissanen, J. and Ristad, E.S. (1994) Unsupervised classification with stochastic complexity. In H. Bozdogan (ed.), *Proceedings of the First US–Japan Conference on the Frontiers of Statistical Modeling*, pp. 171–182. Deventer: Kluwer.

Roeder, K. (1994). A graphical technique for determining the number of components in a mixture of normals. *J. Amer. Statist. Assoc.*, **89**, 487–495.

Schwarz, G. (1978). Estimation of the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Self, S.G. and Lieng, K.L. (1987). Asymptotic properties of maximum likelihood and maximum ratio tests under non standard conditions. *J. Amer. Math. Soc. (Theory and Method)*, **82**, 605–610.

Received May 1994 and revised October 1996