Biometrika (2005), **92,** 3, *pp.* 667–678

Printed in Great Britain

Nonparametric inference in multivariate mixtures

BY PETER HALL, AMNON NEEMAN, REZA PAKYARI AND RYAN ELMORE

Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia

peter.hall@maths.anu.edu.au amnon.neeman@maths.anu.edu.au
reza.pakyari@maths.anu.edu.au ryan.elmore@maths.anu.edu.au

SUMMARY

We consider mixture models in which the components of data vectors from any given subpopulation are statistically independent, or independent in blocks. We argue that if, under this condition of independence, we take a nonparametric view of the problem and allow the number of subpopulations to be quite general, the distributions and mixing proportions can often be estimated root-n consistently. Indeed, we show that, if the data are k-variate and there are p subpopulations, then for each $p \geqslant 2$ there is a minimal value of k, k_p say, such that the mixture problem is always nonparametrically identifiable, and all distributions and mixture proportions are nonparametrically identifiable when $k \geqslant k_p$. We treat the case $p = 2$ in detail, and there we show how to construct explicit distribution, density and mixture-proportion estimators, converging at conventional rates. Other values of p can be addressed using a similar approach, although the methodology becomes rapidly more complex as p increases.

Some key words: Bandwidth; Curve estimation; Independent marginals; Kernel methods; Nonparametric density estimation.

1. INTRODUCTION

Suppose a population consists of p different subpopulations, and that the sampled data from each subpopulation are vectors of length k. It is of interest to estimate the k-variate distributions of the subpopulation and the values of the mixing proportions. In the conventional, parametric approach, models are fitted to the distributions of the p subpopulations, and model parameters, as well as the mixing proportions, are estimated, for example by maximum likelihood; see Everitt & Hand (1981, Ch. 2), Titterington et al. (1985, Ch. 4), McLachlan & Basford (1988, Ch. 2, 4), Lindsay (1995, Ch. 3) and McLachlan & Peel (2000, Ch. 2).

In the present paper we shall show that, for each $p \geqslant 2$, there is a minimal $k = k_p$ such that, provided $k \geqslant k_p$ and the marginals are independent, and the mixing proportions are all distinct, the marginal distributions and the mixing probabilities are identifiable in a nonparametric sense. Moreover, they are estimable root-n consistently. Implications of the assumption of independent marginals, and ways in which it can be relaxed, will be discussed shortly. These results imply that, from at least one point of view, the 'curse of dimensionality' works in reverse.

One portion of our proof of this result is explictly constructive. We suggest a general method that might be used to construct, from the mixture distribution, all the unknown

marginal distributions of subpopulation components, and the mixing proportions, if the problem does in fact have a solution. Furthermore, we prove that this method must lead to a unique solution if k is at least as large as some finite k_p. The definition of k_p, here and below, refers to our particular method; different methods may have different minimal values of k for which nonparametric identifiability is feasible. However, finiteness of the minimal k for one method implies finiteness for any method that gives the least possible k.

On the other hand, showing that k_p is always finite is not so straightforward. Our approach to solving this problem is based on algebraic geometry and uses classical invariant theory, that is the theory of polynomial functions invariant under a group action.

Next we address the assumption that all component populations have independent marginals. If the models are Gaussian then, even if p is as small as 2, there are $k^2 + 3k + 1$ unknowns to be estimated. To reduce this number it is typically assumed that the marginal distributions are independent. In operational terms, this amounts to accepting a degree of bias in return for a reduction in variance; see for example Rindskopt & Rindskopt (1986), Thompson & Walter (1988), Walter (1988), Valenstein (1990), Torrance-Rynard & Walter (1997) and Hui & Zhou (1998). The condition of independence can be imposed on the same pragmatic grounds in a nonparametric setting.

However, the condition holds exactly in some contexts. For example, it has been argued that observed dependencies in genetic behaviour are caused by populations being mixtures, rather than comprising a single type; and that, when an appropriate mixture model is employed, properties of different genes will indeed be independent within each subpopulation. The most suitable model for genetic behaviour in such a population would therefore be the one discussed above, where each subpopulation has independent marginals; see for example Cardon & Palmer (2003).

In some approaches to latent class analysis in sociology, a degree of dependence is permitted within classes which are otherwise assumed to be independent. This context motivates a generalisation of our model for completely independent marginals, which we now discuss.

Suppose that the set of indices $\{1, \ldots, k\}$ can be partitioned into disjoint subsets $\mathscr{S}_1, \ldots, \mathscr{S}_{k'}$, where $2 \leqslant k' \leqslant k$, and that, for each subpopulation, this partition decomposes a data vector into k' mutually independent subvectors. Then, provided $k' \geqslant k_p$, the following is true. For each subpopulation and each $1 \leqslant \ell \leqslant k'$, the joint distribution of components with indices in $\mathscr{S}_\ell$ can be estimated root-n consistently from data from the mixture. Also, if $\mathscr{S}_\ell$ contains $s_\ell \geqslant 1$ indices, the joint density of components with indices in $\mathscr{S}_\ell$ can be estimated at the standard rate pertaining to nonparametric density estimation in s_ℓ dimensions. In particular, if the density of the s_ℓ-variate data subvector has t_ℓ bounded derivatives, then the L_2 convergence rate of a density estimator, computed from a sample of size n, is $n^{-2t_\ell/(s_\ell + 2t_\ell)}$. For a given subpopulation, by multiplying together distribution or density estimators corresponding to the k' subsets of indices, we obtain estimators that converge at the L_2 rate n^{-1}, in the distribution case, or rate $\max_\ell n^{-2t_\ell/(s_\ell + 2t_\ell)}$, for the density. Furthermore, the mixing proportions can be estimated root-n consistently. For the sake of brevity we shall not explore this setting explicitly, although doing so is straightforward.

In related work, Hettmansperger & Thomas (2000) and Thomas & Hettmansperger (2001) treated inference in mixtures by reducing multivariate data to binomial or multinomial responses. The latter cases are effectively parametric, and can be addressed in relatively conventional ways. Woodward et al. (1995) introduced minimum Hellinger distance methods for estimating mixture proportions. Leroux (1992) and Chen &

Kalbfleisch (1996) discussed maximum-penalised likelihood methods for inference about mixtures. Lindsay (1994, 1995) described minimum Hellinger distance methods and likelihood-based methods. There have been many studies of methods for estimating mixture proportions; see for example Windham & Cutler (1992), who drew connections to cluster analysis. Hall & Zhou (2003) discussed the case $p = 2$ in the context of the present paper, but proposed only implicit methods that are awkward to implement. By way of contrast, the techniques here are explicit and easy to use. The case $p = 3$ is addressed in the 2005 Australian National University Ph.D thesis of R. Pakyari

2. Deterministic recovery of component distributions

2·1. *Inversion of mixture models*

Let $(\pi_1, \ldots, \pi_p)$ denote a p-variate multinomial distribution with none of the component probabilities vanishing, and let F_{ji}, for $1 \leqslant i \leqslant k$ and $1 \leqslant j \leqslant p$, be continuous univariate distribution functions. The mixture model,

$$\pi_1 \prod_{i=1}^{k} F_{1i} + \ldots + \pi_p \prod_{i=1}^{k} F_{pi} = \Phi, \tag{2·1}$$

implies a set of lower-dimensional submodels,

$$\pi_1 \prod_{m=1}^{\ell} F_{1i_m} + \ldots + \pi_p \prod_{m=1}^{\ell} F_{pi_m} = \Phi_{i_1 \ldots i_\ell}, \tag{2·2}$$

where $1 \leqslant \ell \leqslant k$, $1 \leqslant i_1 < \ldots < i_\ell \leqslant k$, and $\Phi_{i_1 \ldots i_\ell}$ denotes the ℓ-variate 'marginal distribution of Φ corresponding to vector components with indices $i_1, \ldots, i_\ell$. Our ultimate goal is to show how to estimate the univariate distributions F_{ji}, and the mixing probabilities π_j, in (2·1), using only data from the k-variate distribution Φ and making no parametric assumption about the distributions F_{ji}. In § 2·2, however, our aim is to show how (2·1) may be 'inverted' to express the F_{ji}'s and the π_j's in terms of the functions $\Phi_{i_1 \ldots i_\ell}$.

In the sense that the order of the p populations can always be permuted, there are always $p!$ solutions to this problem. Assuming that no two of the π_i's are identical, we can remove this redundancy by insisting that $\pi_1 < \ldots < \pi_p$. Nevertheless, the potential redundancy will always make an appearance in terms of solutions for the distributions F_{ji} and probabilities π_j.

Our approach is to view equation (2·2) as representing kp unknown functions F_{ji}, expressed in terms of the estimable functions $\Phi_{i_1 \ldots i_\ell}$, and to solve the equations for the unknowns. If $\pi_1, \ldots, \pi_p$ are given, we require at least kp such equations, and, if we are to estimate $\pi_1, \ldots, \pi_p$ as well, we need at least $kp + 1$ equations in all. The number of different equations of the type (2·2) is $2^k - 1$, and so we need $2^k - 1 \geqslant kp + 1$. The least value of k, k'_p say, for which this is possible is given by $k'_p = 3, 4, 5, 5, 5, 6, \ldots, 6$ for $p = 2, 3, 4, \ldots, 10$, respectively. The least value of k, k_p say, for which equations (2·2) have a unique solutions in the F_{ji}'s and the π_j's, must satisfy $k_p \geqslant k'_p$. In the Appendix we derive a bound in the other direction, outlining the route taken by a proof that k_p is no larger than a quantity which equals $\{1 + o(1)\} 6p \log p$ as p increases.

Let $\Delta_{ji} = F_{ji} - \Phi_i$ and

$$\Psi_{i_1 \ldots i_r} = \Phi_{i_1 \ldots i_r} - \sum_{s=2}^{r-1} \Psi_{i_1 \ldots i_s} \Phi_{i_{s+1} \ldots i_r} \left[\binom{r}{s} \right] - \Phi_{i_1} \ldots \Phi_{i_r},$$

where the tensor-like notation

$$\left[\binom{r}{s}\right]$$

indicates that the corresponding term $\Psi_{i_1\ldots i_s}\Phi_{i_{s+1}\ldots i_r}$, and all

$$\binom{r}{s}-1$$

other terms of like construction, are included at that point. Then it may be shown from (2·2) that

$$\pi_1\prod_{m=1}^{\ell}\Delta_{1i_m}+\ldots+\pi_p\prod_{m=1}^{\ell}\Delta_{pi_m}=\Psi_{i_1\ldots i_\ell}, \tag{2·3}$$

where $\Psi_{i_1\ldots i_\ell}$ is an explicitly-defined functional of $\Phi_{r_1\ldots r_s}$, for $1\leqslant s\leqslant\ell$, $r_1<\ldots<r_s$ and $\{r_1,\ldots,r_s\}\subseteq\{i_1,\ldots,i_\ell\}$. The simplest case is $\ell=2$, for which $\Psi_{i_1i_2}=\Phi_{i_1i_2}-\Phi_{i_1}\Phi_{i_2}$.

If we eliminate the pth population, using $\Delta_{pi}=-\pi_p^{-1}\sum_{j\leqslant p-1}\pi_j\Delta_{ji}$, then we obtain, for $\ell\geqslant 2$,

$$\pi_1\prod_{m=1}^{\ell}\Delta_{1i_m}+\ldots+\pi_{p-1}\prod_{m=1}^{\ell}\Delta_{p-1,i_m}-(-\pi_p)^{-(\ell-1)}\prod_{m=1}^{\ell}\left(\sum_{j=1}^{p-1}\pi_j\Delta_{ji_m}\right)=\Psi_{i_1\ldots i_\ell}. \tag{2·4}$$

Thus, we have reduced the 2^k-1 equations (2·2), involving kp unknowns F_{ji}, for $1\leqslant j\leqslant p$ and $1\leqslant i\leqslant k$, to the 2^k-k-1 equations (2·4), involving $k(p-1)$ unknowns Δ_{ji}, for $1\leqslant j\leqslant p-1$ and $1\leqslant i\leqslant k$, without losing the essential character of (2·2), which is that all the unknowns are on the left-hand side and only directly estimable quantities are on the right. Further 'simplifications,' based on other low-dimensional versions of (2·2) for $\ell\geqslant 2$, are algebraically very complex, however.

2·2. *The case $p=2$*

Put $\Psi_{i_1i_2}=\Phi_{i_1i_2}-\Phi_{i_1}\Phi_{i_2}$, for $1\leqslant i_1,\ i_2\leqslant k$ with $i_1\neq i_2$, and assume that, for each $1\leqslant i\leqslant k$, there exist i_1 and i_2, with neither value equal to i, such that $\Psi_{i_1i_2}$ does not vanish identically. Then

$$F_{1i}=\pm\left(\frac{\pi_2\Psi_{ii_1}\Psi_{ii_2}}{\pi_1\Psi_{i_1i_2}}\right)^{\frac{1}{2}}+\Phi_i,\quad F_{2i}=\mp\left(\frac{\pi_1\Psi_{ii_1}\Psi_{ii_2}}{\pi_2\Psi_{i_1i_2}}\right)^{\frac{1}{2}}+\Phi_i. \tag{2·5}$$

The $+$ and $-$ signs in (2·5) are of course chosen respectively; switching from $(+,-)$ to $(-,+)$ amounts only to interchanging the two populations in the mixture. The quantities of which we take the square root at (2·5) are always nonnegative, and in fact $\Psi_{ii_1}\Psi_{ii_2}/\Psi_{i_1i_2}=\pi_1\pi_2(F_{1i}-F_{2i})^2$.

Formula (2·5) implies that we may express F_{1i} and F_{2i} as

$$F_{1i}=\left(\frac{1-\pi_1}{\pi_1}\right)^{\frac{1}{2}}\chi_{1i}+\Phi_i,\quad F_{2i}=\left(\frac{\pi_1}{1-\pi_1}\right)^{\frac{1}{2}}\chi_{2i}+\Phi_i, \tag{2·6}$$

where χ_{1i} and χ_{2i} are known functionals of $\Phi_1,\ldots,\Phi_k$ and of $\Phi_{i_1i_2}$, for $1\leqslant i_1,\ i_2\leqslant k$. If F_{1i} and F_{2i} are given by (2·6), then equations (2·2) with $\ell=1$ and $\ell=2$ become

$$\chi_{1i}+\chi_{2i}\equiv 0,\quad \chi_{1i_1}\chi_{1i_2}\equiv\Psi_{i_1i_2}, \tag{2·7}$$

respectively. Result (2·7) has two consequences. First, no matter how large the value of k, the univariate and bivariate forms of (2·2) contain no information about π_1 or π_2. Secondly, $\Psi_{i_1 i_2}$ factorises into the product of its 'marginals'.

As a prelude to determining π_1, and hence $\pi_2 = 1 - \pi_1$, from the trivariate distributions defined by taking $\ell = 3$ at (2·2), we make the following assumption.

Assumption 1. There exists a triple (i_1, i_2, i_3), and a point $(x_{i_1}, x_{i_2}, x_{i_3})$, such that $\Psi_{i_1 i_2}(x_{i_1}, x_{i_2})\Psi_{i_2 i_3}(x_{i_2}, x_{i_3})\Psi_{i_1 i_3}(x_{i_1}, x_{i_3}) \neq 0$.

Note that the product $\Psi_{i_1 i_2}\Psi_{i_2 i_3}\Psi_{i_1 i_3}$ is always nonnegative; it equals

$$(\pi_1 \pi_2)^3 (F_{1 i_1} - F_{2 i_1})^2 (F_{1 i_2} - F_{2 i_2})^2 (F_{1 i_3} - F_{2 i_3})^2.$$

For any such triple (i_1, i_2, i_3), and for the respective choices of the $+$ and $-$ signs in the definition of F_{ji} at (2·5),

$$\left(\frac{1-\pi_1}{\pi_1}\right)^{\frac{1}{2}} (2\pi_1 - 1) = \pm \frac{\Phi_{i_1}|\Psi_{i_2 i_3}| + \Phi_{i_2}|\Psi_{i_1 i_3}| + \Phi_{i_3}|\Psi_{i_1 i_2}| + \Phi_{i_1}\Phi_{i_2}\Phi_{i_3} - \Phi_{i_1 i_2 i_3}}{(\Psi_{i_1 i_2}\Psi_{i_2 i_3}\Psi_{i_1 i_3})^{\frac{1}{2}}}, \tag{2·8}$$

where Φ_i and $\Psi_{i_1 i_2}$ are interpreted as $\Phi_i(x_i)$ and $\Psi_{i_1 i_2}(x_{i_1}, x_{i_2})$, respectively. The left-hand side of (2·8) is strictly increasing in $\pi \in (0, \frac{1}{4}(1 + 5^{\frac{1}{2}})]$. Therefore, provided we choose the $+$ or $-$ sign so that the right-hand side of (2·8) is not strictly positive, π_1 is uniquely determined as an element of $(0, \frac{1}{2}]$. In this way we determine the lesser of π_1 and π_2, as well as the sign we should take at (2·5) in order that this lesser value should equal π_1.

3. Estimating π_j, F_{ji} and f_{ji} when $p = 2$

3·1. *Estimators of π_j and F_{ji}*

Suppose that we observe k-variate data $X_m = (X_{m1}, \ldots, X_{mk})$, for $1 \leqslant m \leqslant n$, drawn from the mixture distribution Φ defined at (2·1) with $p = 2$, and suppose for definiteness that $\pi_1 < \pi_2$. Our estimators of F_{ji} and π_j are based on replacing $\Phi_{j_1 \ldots j_\ell}$ and $\Psi_{j_1 \ldots j_\ell}$ at (2·5) and (2·8) by their canonical estimators, and averaging over points of the sample space for which the resulting denominators are not too close to zero. In particular, our estimator of π_1 is $\hat{\pi}_1$, the unique solution in $(0, \frac{1}{2}]$ of the equation

$$\left(\frac{1-\hat{\pi}_1}{\hat{\pi}_1}\right)^{\frac{1}{2}} (2\hat{\pi}_1 - 1) = -\frac{6}{k(k-1)(k-2)\|\mathscr{S}_1(\varepsilon_1)\|} \times \left| \sum\sum\sum_{i_1 < i_2 < i_3} \int_{\mathscr{S}_1(\varepsilon_1)} \frac{\hat{\rho}_1(x_{i_1}, x_{i_2}, x_{i_3})}{\hat{\rho}_2(x_{i_1}, x_{i_2}, x_{i_3})} dx_{i_1} dx_{i_2} dx_{i_3} \right|, \tag{3·1}$$

where the series is taken over all

$$\binom{k}{3}$$

triples $\{i_1, i_2, i_3\} \subseteq \{1, \ldots, k\}$ with $i_1 < i_2 < i_3$, $\mathscr{S}_1(\varepsilon_1)$ denotes the set of $(x_{i_1}, x_{i_2}, x_{i_3})$ such that $\hat{\rho}_2(x_{i_1}, x_{i_2}, x_{i_3}) > \varepsilon_1$, $\|\mathscr{S}\|$ denotes the ℓ-variate content of an ℓ-variate set $\mathscr{S}$, $\varepsilon_1 > 0$ is a small positive constant,

$$\hat{\rho}_1 = \hat{\Phi}_{i_1}|\hat{\Psi}_{i_2 i_3}| + \hat{\Phi}_{i_2}|\hat{\Psi}_{i_1 i_3}| + \hat{\Phi}_{i_3}|\hat{\Psi}_{i_1 i_2}| + \hat{\Phi}_{i_1}\hat{\Phi}_{i_2}\hat{\Phi}_{i_3} - \hat{\Phi}_{i_1 i_2 i_3},$$

$\hat\rho_2 = |\hat\Psi_{i_1i_2}\hat\Psi_{i_2i_3}\hat\Psi_{i_1i_3}|^{\frac12}$, $\hat\Phi_{i_1\dots i_\ell}$ is the empirical distribution function of the ℓ-variate data $(X_{mi_1},\dots,X_{mi_\ell})$, for $1\leqslant m\leqslant n$, and $\hat\Psi_{i_1i_2}=\hat\Phi_{i_1i_2}-\hat\Phi_{i_1}\hat\Phi_{i_2}$. Our estimator of π_2 is of course $\hat\pi_2=1-\hat\pi_1$. Section 4 will suggest empirical methods for choosing thresholds, and Theorem 1 will show that the rate of convergence of estimators is largely unaffected by threshold choice.

If the sign of the triple series at (3·1) is positive then our estimator of $F_{1i}(x_i)$ is $\hat F_{1i}(x_i)=\hat\Phi_i(x_i)-\hat G_i(x_i)$, where

$$\hat G_i(x_i)=\frac{2}{(k-1)(k-2)}\sum_{i_1<i_2:i_1\ne i\ne i_2}\sum\frac{1}{\|\mathscr{S}_{2i_1i_2}(\varepsilon_2)\|}\times\int_{\mathscr{S}_{2i_1i_2}(\varepsilon_2)}\left|\frac{\hat\pi_2\hat\Psi_{ii_1}(x_i,x_{i_1})\hat\Psi_{ii_2}(x_i,x_{i_2})}{\hat\pi_1\hat\Psi_{i_1i_2}(x_{i_1},x_{i_2})}\right|^{\frac12}dx_{i_1}dx_{i_2},$$

the series is taken over all

$$\binom{k-1}{2}$$

pairs $\{i_1,i_2\}\subseteq\{1,\dots,k\}$ with $i_1<i_2$ and $i_1\ne i\ne i_2$, $\mathscr{S}_{2i_1i_2}(\varepsilon_2)$ is the set of (x_{i_1},x_{i_2}) such that $|\hat\Psi_{i_1i_2}(x_{i_1},x_{i_2})|>\varepsilon_2$, and $\varepsilon_2>0$ is a small positive constant. In this case our estimator of F_{2i} is $\hat F_{2i}=\hat\Phi_i+(\hat\pi_1/\hat\pi_2)\hat G_i$. On the other hand, if the sign of the triple series at (3·1) is negative then our estimators of F_{1i} and F_{2i} are $\hat F_{1i}=\hat\Phi_i+\hat G_i$ and $\hat F_{2i}=\hat\Phi_i-(\hat\pi_1/\hat\pi_2)\hat G_i$. The complexity of the calculations will grow like k^2 as k increases.

These estimators will generally not themselves be distribution functions. This difficulty may be overcome by renormalising, as follows. Let $\hat F$ denote either $\hat F_{1i}$ or $\hat F_{2i}$, and put

$$\tilde F(u)=\max\left\{\inf_{v\geqslant u}\hat F(v),0\right\}\Big/\sup_{-\infty<v<\infty}\hat F(v).$$

Then, provided $\sup\hat F>0$, $\tilde F$ is a distribution function. In this manner we define $\tilde F_{1i}$ and $\tilde F_{2i}$.

The theorem below, proved in the Appendix, shows that $\hat F_{ji}$ and $\tilde F_{ji}$ are both uniformly consistent for F_{ji}, and converge at rate $O_p(n^{-\frac12})$.

THEOREM 1. *Assume the mixture model* (2·1) *for* $p=2$ *and* $k\geqslant3$, *that each of the distributions* F_{ji} *is continuous, that Assumption* 1 *holds, and that* $\pi_1<\pi_2$. *Suppose too that the thresholds* ε_1 *and* ε_2 *satisfy*

$$0<\varepsilon_1<\max_{i_1<i_2<i_3}\max_{(x_{i_1},x_{i_2},x_{i_3})}|\Psi_{i_1i_2}(x_{i_1},x_{i_2})\Psi_{i_2i_3}(x_{i_2},x_{i_3})\Psi_{i_1i_3}(x_{i_1},x_{i_3})|^{\frac12},$$

$$0<\varepsilon_2<\min_{1\leqslant i\leqslant k}\max_{i_1<i_2:i_1\ne i\ne i_2}\max_{(x_{i_1},x_{i_2})}|\Psi_{i_1i_2}(x_{i_1},x_{i_2})|.$$

Then $|\hat\pi_1-\pi_1|=O_p(n^{-\frac12})$, *and, for* $1\leqslant i\leqslant k$ *and* $j=1,2$,

$$\sup_{-\infty<x<\infty}\{|\hat F_{ji}(x)-F_{ji}(x)|+|\tilde F_{ji}(x)-F_{ji}(x)|\}=O_p(n^{-\frac12}).$$

3·2. *Density estimation*

Note that, by (2·5),

$$f_{1i}=\pm\frac12\left(\frac{\pi_2}{\pi_1|\Psi_{i_1i_2}|}\right)^{\frac12}\left(\left|\frac{\Psi_{ii_2}}{\Psi_{ii_1}}\right|^{\frac12}\Psi_{ii_1}^{(1,0)}+\left|\frac{\Psi_{ii_1}}{\Psi_{ii_2}}\right|^{\frac12}\Psi_{ii_2}^{(1,0)}\right)+\phi_i,$$

where

$$\Psi^{(1,0)}_{i_1 i_2} = \frac{\partial}{\partial x_{i_1}} \Psi_{i_1 i_2}(x_{i_1}, x_{i_2}) = \Phi^{(1,0)}_{i_1 i_2}(x_{i_1}, x_{i_2}) - \phi_{i_1}(x_{i_1}) \Phi_{i_2}(x_{i_2}).$$

Estimators of $\phi_i(x_i)$ and $\Phi^{(1,0)}_{i_1 i_2}(x_{i_1}, x_{i_2})$ are

$$\hat{\phi}_i(x_i) = \frac{1}{nh} \sum_{m=1}^{n} K\left(\frac{x_i - X_{mi}}{h}\right),$$

$$\hat{\Phi}^{(1,0)}_{i_1 i_2}(x_{i_1}, x_{i_2}) = \frac{1}{nh} \sum_{m=1}^{n} K\left(\frac{x_{i_1} - X_{mi_1}}{h}\right) I(X_{mi_2} \leqslant x_{i_2}),$$

where K is a kernel and h a bandwidth. Our estimator of $\Psi^{(1,0)}_{i_1 i_2}$ is

$$\hat{\Psi}^{(1,0)}_{i_1 i_2} = \hat{\Phi}^{(1,0)}_{i_1 i_2} - \hat{\phi}_{i_1} \hat{\Phi}_{i_2},$$

giving the following estimator of $f_{1i}(x_i)$:

$$\begin{aligned} \hat{f}_{1i}(x_i) = \hat{\phi}_i(x_i) + \hat{S}_i(x_i) \frac{1}{(k-1)(k-2)} \sum_{i_1 < i_2 : i_1 \neq i \neq i_2} \frac{1}{\|\mathscr{S}_{ii_1 i_2}(\varepsilon_3)\|} \\ \times \int_{\mathscr{S}_{ii_1 i_2}(\varepsilon_3)} \left\{\frac{\hat{\pi}_2}{\hat{\pi}_1 |\hat{\Psi}_{i_1 i_2}(x_{i_1}, x_{i_2})|}\right\}^{\frac{1}{2}} \left\{ \left|\frac{\hat{\Psi}_{ii_2}(x_i, x_{i_2})}{\hat{\Psi}_{ii_1}(x_i, x_{i_1})}\right|^{\frac{1}{2}} \hat{\Psi}^{(1,0)}_{ii_1}(x_i, x_{i_1}) \right. \\ \left. + \left|\frac{\hat{\Psi}_{ii_1}(x_i, x_{i_1})}{\hat{\Psi}_{ii_2}(x_i, x_{i_2})}\right|^{\frac{1}{2}} \hat{\Psi}^{(1,0)}_{ii_2}(x_i, x_{i_2}) \right\} dx_{i_1} dx_{i_2}, \end{aligned}$$

where $\hat{S}_i(x_i) = \pm 1$ according as $\hat{F}_{1i}(x_i) = \hat{\Phi}_i(x_i) \pm \hat{G}_i(x_i)$, the summation is over all

$$\binom{k-1}{2}$$

pairs $\{i_1, i_2\} \subseteq \{1, \ldots, k\}$ with $i_1 < i_2$ and $i_1 \neq i \neq i_2$, and $\mathscr{S}_{ii_1 i_2}(\varepsilon_3)$ denotes the set of (x_{i_1}, x_{i_2}) such that

$$\min\{|\hat{\Psi}_{ii_1}(x_i, x_{i_1})|, |\hat{\Psi}_{ii_2}(x_i, x_{i_2})|, |\hat{\Psi}_{i_1 i_2}(x_{i_1}, x_{i_2})|\} > \varepsilon_3.$$

Thus, $\mathscr{S}_{ii_1 i_2}(\varepsilon_3)$ depends on the value of x_i.

The density estimator $\hat{f}_{1i}$ has asymptotic bias and variance properties similar to those of a conventional kernel-type estimator. In particular, its bias is of size h^2 and its variance is of size $(nh)^{-1}$. Details are given in a longer version of this paper, obtainable from the authors.

4. Numerical properties

4·1. *Simulation study*

The results reported here were all obtained using the following approach to choosing tuning parameters, including both thresholds and bandwidths. Fit a Gaussian model by maximum likelihood, assuming the components are independent; compute the resulting estimates of marginal means and variances, and of the mixing proportions; by simulation from the Gaussian model with parameters set equal to these estimated values, choose the optimal values of tuning parameters; and then apply the nonparametric method suggested in § 3.

We shall summarise simulation studies in three cases, in each of which $p = 2$, $k = 3$ and the two populations are identical except for a shift of location. Excepting location, each component population was either a product of three standard normal distributions, referred to below as the normal model, or a product of three Student's t distributions with 10 degrees of freedom, referred to as the $t(10)$ model, or a product of three double exponential distributions with density $\frac{1}{2}e^{-|x|}$, referred to as the Laplace model. In these respective cases, the difference between the mean vectors of the two 3-variate distributions were chosen to be (3, 4, 5), c(3, 4, 5) and (3, 3, 3), respectively, where c denotes the constant for converting the $t(10)$ noncentrality parameter into its mean.

Of course, only in the first case was our method for choosing tuning parameters applied under the correct model. In each setting we took $n = 500$. By averaging over 300 samples we computed numerical approximations to root mean integrated squared errors, shown in Figs 1 and 2, for estimators of the marginal distributions and marginal densities, respectively. In each case the value is depicted as a function of the mixing proportion, π_1, graphed on the horizontal axis in the interval $[0{\cdot}1, 0{\cdot}4]$. Performance of the density estimators was surprisingly constant, depending relatively little on choice of the type of marginal distribution or on the mixing proportion.

However, in the case of distribution estimation the method has somewhat greater difficulty with the Laplace distribution than with either of the other two. Also, when

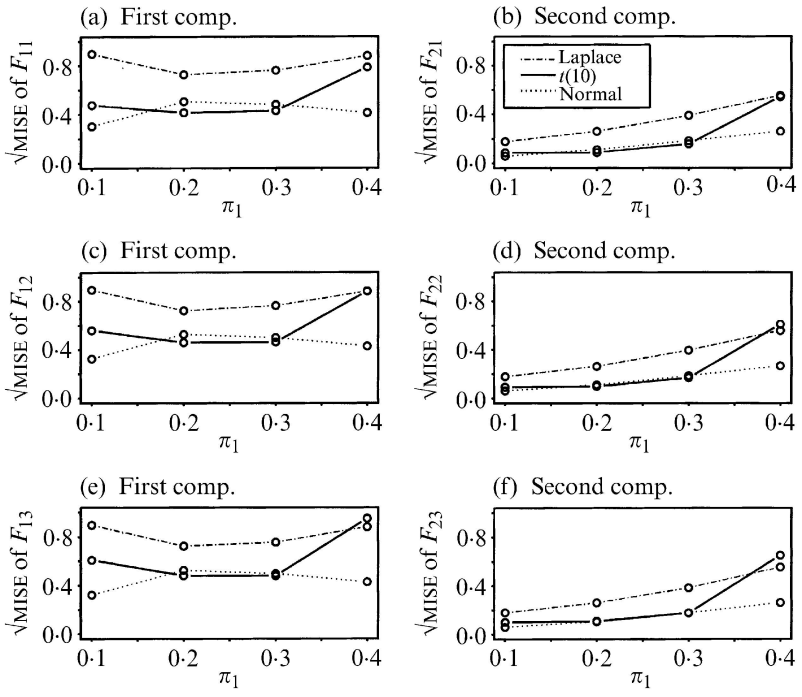

Fig. 1: Simulation study. Root mean integrated squared errors of non-parametric estimators of marginal distribution functions for normal, dotted lines, $t(10)$, solid lines, and Laplace, dot-dashed lines, models. Panels (a), (c) and (e) depict plots of root mean integrated squared errors against the mixing proportion π_1, for estimates of the three marginal distributions of the first component. Panels (b), (d) and (f) do the same for the second component.

the method is applied to the estimation of F_{2i}, performance tends to deteriorate as π_1 increases from 0·1 to 0·4. There are two reasons for this. First, as π_1 increases, the second subpopulation is observed less often, and so there is less information about it. Secondly, when π_1 is relatively close to 0·5, any estimator in this setting tends to confuse the two subpopulations; recall that the 'first' subpopulation is distinguished as the one that has the smaller mixing probability. As π_1 is increased beyond 0·5 this confusion diminishes, and performance improves a little, as long as π_1 is not too large. For $\pi_1 > 0{\cdot}75$, however, the scarcity of data from the second subpopulation becomes a major issue, and performance deteriorates badly.

The results discussed in the previous two sentences are apparent from Figs 1(a), (c) and (e), given the symmetry of the problem; the results mentioned in the earlier two sentences can be seen in Figs 1(b), (d) and (f).

For brevity we do not give plots of root mean squared errors of estimators of π_1. The plots would show relatively constant performance over all values of π_1. Indeed, given the different nature of the problem of estimating π_1, it is clear that $\hat{\pi}_1$ should be afflicted relatively little by the difficulties noted two paragraphs above. In the case of the Laplace model bias has little impact on the error of estimates of π_1, but for the other two models the errors arising from bias and error-about-the-mean are similar.

In the case of the normal model, our nonparametric estimator of π_1 performs very similarly to its parametric counterpart. When the subpopulation in question is sampled

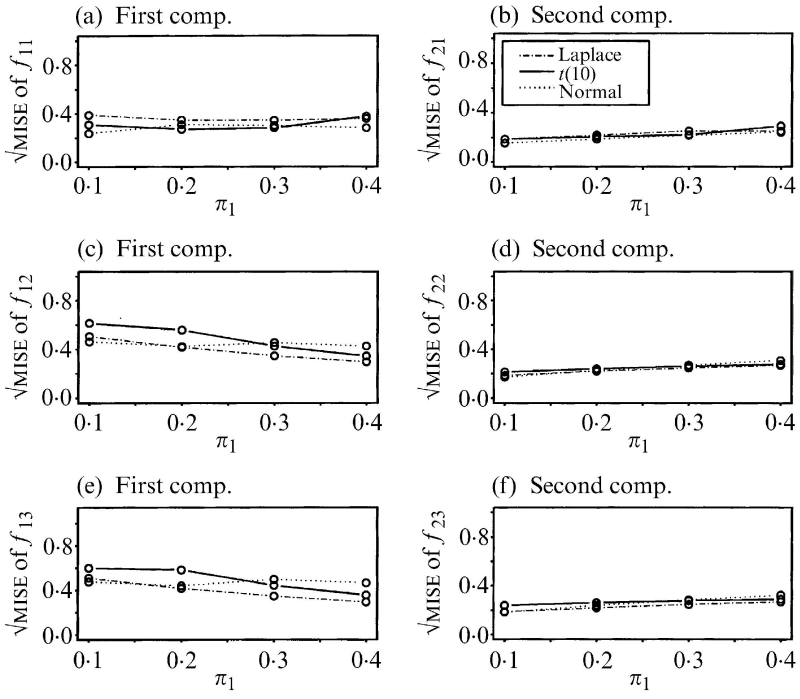

Fig. 2: Simulation study. Root mean integrated squared errors of nonparametric estimators of marginal density functions for normal, dotted lines, $t(10)$, solid lines, and Laplace, dot-dashed lines, models. Panels (a), (c) and (e) depict plots of root mean integrated squared errors against the mixing proportion π_1, for estimates of the three marginal distributions of the first component. Panels (b), (d) and (f) do the same for the second component.

from with relatively high probability, the nonparametric estimators of marginal distributions tend to be superior to their parametric versions. This relationship is reversed, however, when the subpopulation is encountered only relatively rarely.

4·2. *Real-data example: Leptograpsus crabs*

Campbell & Mahon (1974) collected and analysed 100 Leptograpsus crabs from each of two species, in Fremantle, Western Australia. Five measurements of morphological characteristics were made for each crab. To simplify our analysis we discarded the last two of these measurements; the three measurements remaining were the width of the frontal lip of the carapace, the rear width of the carapace and the length along the midline of the carapace, the carapace being the outer, uppermost, hard shell of the crab.

We pooled the data from both species into a single sample of size 200, and repeatedly resampled datasets of size $n = 50$, without replacement, from the pooled sample. To each dataset obtained in this way we fitted a two-population mixture model, using our nonparametric methods to estimate the mixing proportion π and the marginal distribution functions. Since $p = 2$ and $k = k_p = 3$, we have sufficiently many components to justify a nonparametric approach.

We also fitted a Gaussian mixture model under the assumption of independent components, as well as a Gaussian model where the components were arbitrarily related. These two models involved 13 and 19 parameters, respectively.

Mean squared errors were then computed by comparison with the empirical 'truth' represented by the pooled dataset. Of course, the true value of π was $\frac{1}{2}$; the true marginal distribution functions were taken to be their empirical counterparts computed from all 200 data.

Figure 3 shows the mean integrated squared errors of estimators of marginal distributions for the nonparametric approach and for both parametric methods. The mean squared errors of estimators of the mixing proportion were 0·0011, 0·0083 and 0·0365 in the cases of the nonparametric, 13-parameter and 19-parameter normal fits, respectively.

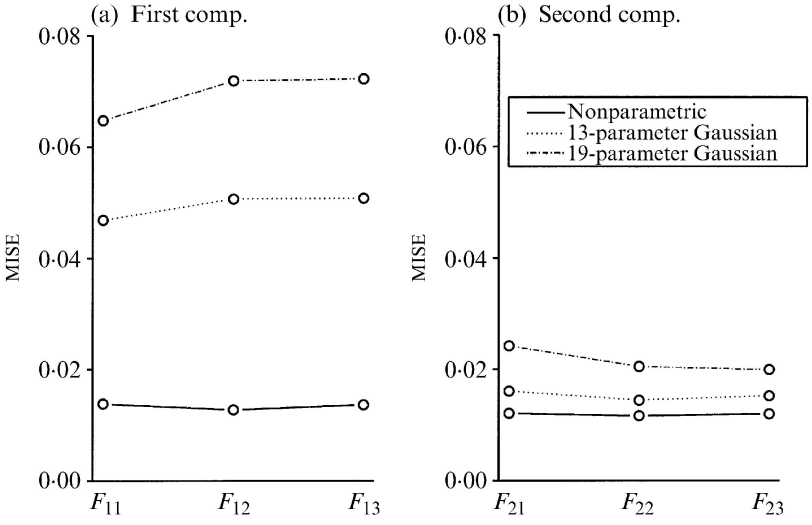

Fig. 3. Mean integrated squared errors of marginal distribution estimators by fitting three models for the crabs dataset; (a) first component, (b) second component.

The nonparametric method gave the best results overall, followed by the Gaussian model with independent components, and then by the more general Gaussian model. This order is preserved if, in our sampling experiment, the average proportion of males in the samples of size 50 is taken to lie anywhere between 0·1 and 0·9. Of course, the high variability of a 19-parameter fit, in the case of a sample of size only 50, has strongly influenced the result. As indicated in § 1, one of the motivations for our approach is to reduce dimension in problems such as this.

Acknowledgement

The authors are grateful to D. R. Cox for helpful comments.

Appendix

Technical details

Derivation of upper bound for k_p. The proof is based on properties of solutions of polynomial equations. In broad terms, such arguments are becoming more popular in statistics; see for example Pistone et al. (2001). In brief, our proof is as follows. Let $v = (v_0, \ldots, v_k)$ denote a vector for which $v_0 = 1$ and each other v_i is either 0 or 1. Consider the polynomial function

$$\psi_v(x) = \sum_{j=1}^{p} \prod_{i=0}^{k} x_{ji}^{v_i},$$

where x is the matrix of values x_{ji}. In view of the mixture model (2·1), we have in mind $x_{j0} = \pi_j$ and $x_{ji} = F_{ji}$ for $1 \leqslant i \leqslant k$ and $1 \leqslant j \leqslant p$. However, making this specialisation obscures the argument at this point. We consider ψ_v to be a function from $\mathbb{R}^{(k+1)p}$ to $\mathbb{R}$.

Our constraints on v imply that there are just 2^k functions ψ_v. Let Σ_p denote the set of all permutations of the integers $1, \ldots, p$. Write B for the set of all Σ_p-orbits in $\mathbb{R}^{(k+1)p}$; that is, to form B we identify in $\mathbb{R}^{(k+1)p}$ any two points which differ by a permutation $\sigma \in \Sigma_p$. Let ψ denote the mapping from B to $\mathbb{R}^{2^k}$ sending $x \in \mathbb{R}^{(k+1)p}$ to the vector of the 2^k entries $\psi_v(x)$. Then it may be proved that for each fixed value of p there exists a finite integer $k(p)$ with the property that, if $k \geqslant k(p)$, ψ is a birational transformation on to its image. In particular the coordinates of $y \in B$ are expressible, as quotients of polynomial functions, in terms of the coordinates of $\psi(y)$. Details are given by Elmore et al. (2005).

Our proof gives an explicit value for $k(p)$, satisfying $k(p) \sim 6p \log p$ as p increases. This is undoubtedly larger than the minimal value, k_p, but it nevertheless proves that $k_p < \infty$. The rational function, or quotient of polynomials, form of the functions, and in particular the functions' smoothness, implies that if we perturb the image by $O_p(n^{-1/2})$ then its inverse will be perturbed by $O_p(n^{-1/2})$. This implies the root-n consistency of our estimators of π_j and F_{ji}.

Sketch proof of Theorem 1. Conventional methods show that, for $\ell = 1, 2, 3$,

$$\sup|\hat{\Phi}_{i_1 \ldots i_\ell} - \Phi_{i_1 \ldots i_\ell}| = O_p(n^{-1/2}).$$

Therefore, $\sup|\hat{\Psi}_{i_1 i_2} - \Psi_{i_1 i_2}| = O_p(n^{-1/2})$ and $\sup|\hat{\rho}_j - \rho_j| = O_p(n^{-1/2})$, where ρ_1 and ρ_2 denote respectively the numerator and the denominator in the ratio on the right-hand side of (2·8). From this property and (2·8) it may be proved that the right-hand side of (3·1) equals

$$\{(1 - \pi_1)/\pi_1\}^{1/2}(2\pi_1 - 1) + O_p(n^{-1/2}).$$

This result, and the definition of $\hat{\pi}_1$, imply that $\hat{\pi}_1 = \pi_1 + O_p(n^{-1/2})$.

It follows from the definition of $\hat{G}_i$ that $\sup|\hat{G}_i - G_i| = O_p(n^{-1/2})$, where $G_i = \pi_2|F_{1i} - F_{2i}|$. This property, and the fact that $F_{1i} = \Phi_i \pm G_i$, where the $+$ and $-$ signs are taken according as these signs are needed at (2·8) to ensure that the right-hand side there is not strictly positive, may be used to prove, first, that $\sup|\hat{F}_{ji} - F_{ji}| = O_p(n^{-1/2})$ and thence that $\sup|\tilde{F}_{ji} - F_{ji}| = O_p(n^{-1/2})$. □

REFERENCES

CAMPBELL, N. A. & MAHON, R. J. (1974). A multivariate study of variation in two species of rock crab of the genus leptograpsus. *Aust. J. Zool.* **22**, 417–25.

CARDON, L. R. & PALMER, L. J. (2003). Population stratification and spurious allelic association. *Lancet* **361**, 598–604.

CHEN, J. & KALBFLEISH, J. D. (1996). Penalized minimum-distance estimates in finite mixture models. *Can. J. Statist.* **24**, 167–75.

ELMORE, R. T., HALL, P. & NEEMAN, A. (2005). An application of classical invariant theory to identifiability in nonparametric mixtures. *Ann. Inst. Fourier.* **55**, 1–28.

EVERITT, B. S. & HAND, D. J. (1981). *Finite Mixture Distributions.* London: Chapman and Hall.

HALL, P. & ZHOU, X. H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.* **31**, 201–24.

HETTMANSPERGER, T. P. & THOMAS, H. (2000). Almost nonparametric inference for repeated measures in mixture models. *J. R. Statist. Soc.* B **62**, 811–25.

HUI, S. L. & ZHOU, X. H. (1998). Evaluation of diagnostic tests without gold standards. *Statist. Meth. Med. Res.* **7**, 337–53.

LEROUX, B. G. (1992). Consistent estimation of a mixing distribution. *Ann. Statist.* **20**, 1350–60.

LINDSAY, B. G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *Ann. Statist.* **22**, 1081–114.

LINDSAY, B. G. (1995). *Mixture Models: Theory, Geometry, and Applications.* Hayward, CA.: Institute of Mathematical Statistics.

MCLACHLAN, G. J. & BASFORD, K. E. (1988). *Mixture Models: Inference and Applications to Clustering.* New York: Wiley.

MCLACHLAN, G. J. & PEEL, D. (2000). *Finite Mixture Models.* New York: Wiley.

PISTONE, G., RICCOMAGNO, E. & WYNN, H. (2001). *Algebraic Statistics: Computational Commutative Algebra in Statistics.* London: CRC Press.

RINDSKOPT, D. & RINDSKOPT, W. (1986). The value of latent class analysis in medical diagnosis. *Statist. Med.* **5**, 21–7.

THOMAS, H. & HETTMANSPERGER, T. P. (2001). Modelling change in cognitive understanding with finite mixtures. *Appl. Statist.* **50**, 435–48.

THOMPSON, W. D. & WALTER, S. D. (1988). A reappraisal of the kappa coefficient. *J. Clin. Epidemiol.* **41**, 949–58.

TITTERINGTON, D. M., SMITH, A. F. M. & MAKOV, U. (1985). *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley.

TORRANCE-RYNARD, V. L. & WALTER, S. D. (1997). Effects of dependent errors in the assessment of diagnostic test performance. *Statist. Med.* **16**, 2157–75.

VALENSTEIN, P. N. (1990). Evaluating diagnostic tests with imperfect standards. *Am. J. Clin. Pathol.* **93**, 252–8.

WALTER, S. D. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *J. Clin. Epidemiol.* **41**, 923–37.

WINDHAM, M. P. & CUTLER, A. (1992). Information ratios for validating mixture analyses. *J. Am. Statist. Assoc.* **87**, 1188–92.

WOODWARD, W. A., WHITNEY, P. & ESLINGER, P. W. (1995). Minimum hellinger distance estimation of mixture proportions. *J. Statist. Plan. Infer.* **48**, 303–19.

[*Received March* 2004. *Revised February* 2005]