Ann. Inst. Statist. Math.
Vol. 57, No. 4, 655–664 (2005)

ESTIMATION OF THE NUMBER OF COMPONENTS OF FINITE MIXTURES OF MULTIVARIATE DISTRIBUTIONS

JOGI HENNA

Faculty of Science, University of the Ryukyus, 1 Senbaru, Nishihara-cho, Okinawa 903-0213, Japan

(Received April 16, 2004; revised December 6, 2004)

Abstract. An estimator of the number of components of a finite mixture of k-dimensional distributions is given on the basis of a one-dimensional independent random sample obtained by a transformation of a k-dimensional independent random sample. A consistency of the estimator is shown. Some simulation results are given in a case of finite mixtures of two-dimensional normal distributions.

Key words and phrases: k-dimensional finite mixture, normal pdf, number of components, one-dimensional finite mixture, orthogonal matrix.

1. Introduction

Let $\mathcal{R}^\ell$ mean an ℓ-dimensional Euclidean space. Let $\mathcal{F} = \{f_{\boldsymbol{\theta}}(\boldsymbol{x}) : \boldsymbol{\theta} \in \Theta\}$ be a family of known k-dimensional probability density functions (pdf's), where the parameter space Θ is a compact subset of $\mathcal{R}^{d_1}$ for a d_1.

For a positive integer m, a pdf $f(\boldsymbol{x} \mid \mathcal{A}_m)$ given by

$$f(\boldsymbol{x} \mid \mathcal{A}_m) = \sum_{i=1}^{m} \alpha_i f_{\boldsymbol{\theta}_i}(\boldsymbol{x}) \tag{1.1}$$

is called a finite mixture of $f_{\boldsymbol{\theta}_1}(\boldsymbol{x}), f_{\boldsymbol{\theta}_2}(\boldsymbol{x}), \dots, f_{\boldsymbol{\theta}_m}(\boldsymbol{x})$ (Titterington *et al.* (1985)), where $\sum_{i=1}^m \alpha_i = 1$, $0 < \alpha_i \le 1$, $\boldsymbol{\theta}_i \in \Theta$ $(i = 1, 2, \dots, m)$ and $\mathcal{A}_m = (\alpha_1, \alpha_2, \dots, \alpha_m; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_m)$. So a single $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ in $\mathcal{F}$ is also considered a finite mixture for $m = 1$ as a special case. Each $f_{\boldsymbol{\theta}_i}(\boldsymbol{x})$ is called a component of $f(\boldsymbol{x} \mid \mathcal{A}_m)$ and each α_i a mixing ratio of $f_{\boldsymbol{\theta}_i}(\boldsymbol{x})$.

The purpose of this paper is to give an estimator $\widehat{m}_n$ of the number m of components on the basis of an independent random sample $(\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n)$ from the distribution (1.1). The importance to estimate the number m is described in McLachlan and Basford (1988), Titterington (1990) and others. Henna (1985), Feng and McCulloch (1994), Chen and Kalbfleisch (1996) and Richardson and Green (1997) have treated one-dimensional finite mixtures. Keribin (2000) has given a method which can be applied to a special multivariate normal mixture under the assumption that a superior value Q of m is known. Some methods to determine the number of components are described in McLachlan and Peel (2000). Chen *et al.* (2001) and Garel (2001) have given a test for m in a univariate case.

In this paper, a method which can be applied to k-dimensional finite mixture distributions is considered though the analysis is based on one-dimensional samples. For the purpose, we consider a real valued function T satisfying the following condition, that is, putting $Y_\xi = T(\boldsymbol{X}_\xi)$ $(\xi = 1, 2, \dots, n)$, then $(Y_1, Y_2, \dots, Y_n)$ can be regarded as an independent random sample from a finite mixture with m components such as

$$h(y \mid \boldsymbol{c}_m) = \sum_{i=1}^{m} \alpha_i h_{\boldsymbol{\delta}_i}(y), \tag{1.2}$$

where $h_{\boldsymbol{\delta}_i}(y)$ is a one-dimensional pdf with a parameter $\boldsymbol{\delta}_i$ and $\boldsymbol{c}_m = (\alpha_1, \alpha_2, \dots, \alpha_m; \boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \dots, \boldsymbol{\delta}_m)$. In other words, by transforming $(\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n)$, we obtain an independent random sample $(Y_1, Y_2, \dots, Y_n)$ which can be considered to have come from a finite mixture of m one-dimensional distributions. And then we construct an estimator of m on the basis of $(Y_1, Y_2, \dots, Y_n)$.

In Section 2, some notations and preliminary lemmas are given. In Section 3, an estimator is given and researched its consistency when $\mathcal{F}$ is a family of k-dimensional normal distributions. In Section 4, in particular, an estimator is researched when $\mathcal{F}$ is a finite family of k-dimensional normal distributions with known parameters. In Section 5, some simulation results are given.

2. Some notations and preliminary lemmas

Let $\boldsymbol{X}$ be a random vector with a pdf $f_{\boldsymbol{\theta}}(\boldsymbol{x}) \in \mathcal{F}$. Let us consider a transformation $\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{X} + \boldsymbol{\rho}$ with an orthogonal matrix $\boldsymbol{M}$ and a column vector $\boldsymbol{\rho}$. Assume that $\boldsymbol{Y}$ has a pdf $g_{\boldsymbol{\omega}}(\boldsymbol{y})$ with a parameter $\boldsymbol{\omega}$ when $\boldsymbol{X}$ has $f_{\boldsymbol{\theta}}(\boldsymbol{x})$. Let

$$\mathcal{G} = \{g_{\boldsymbol{\omega}}(\boldsymbol{y}) : \boldsymbol{\omega} \in \Omega\}, \tag{2.1}$$

where Θ corresponds to Ω, which is assumed to be a compact subset of $\mathcal{R}^{d_2}$ for a d_2, through $\boldsymbol{Y}$. Then the correspondence of $\mathcal{F}$ to $\mathcal{G}$ through $\boldsymbol{Y}$ is one-to-one because $g_{\boldsymbol{\omega}}(\boldsymbol{y}) = f_{\boldsymbol{\theta}}(\boldsymbol{M}^{-1}(\boldsymbol{y} - \boldsymbol{\rho}))$ holds (Billingsley (1986)). So it can be easily seen that a necessary and sufficient condition for $f(\boldsymbol{x} \mid \mathcal{A}_m)$ to be the finite mixture (1.1) is that $g(\boldsymbol{y} \mid \mathcal{B}_m)$ to be the finite mixture

$$g(\boldsymbol{y} \mid \mathcal{B}_m) = \sum_{i=1}^{m} \alpha_i g_{\boldsymbol{\omega}_i}(\boldsymbol{y}), \tag{2.2}$$

where $f(\boldsymbol{x} \mid \mathcal{A}_m)$ and $f_{\boldsymbol{\theta}_i}(\boldsymbol{x})$ correspond to $g(\boldsymbol{y} \mid \mathcal{B}_m)$ and $g_{\boldsymbol{\omega}_i}(\boldsymbol{y})$ $(j = 1, 2, \dots, m)$, respectively, through $\boldsymbol{Y}$ with $\mathcal{B}_m = (\alpha_1, \alpha_2, \dots, \alpha_m; \boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \dots, \boldsymbol{\omega}_m)$.

Let $h_{\boldsymbol{\delta}_j}(y_j)$ be the marginal pdf with a parameter $\boldsymbol{\delta}_j$ obtained by

$$h_{\boldsymbol{\delta}_j}(y_j) = \int \cdots \int_{\boldsymbol{R}^{k-1}} g_{\boldsymbol{\omega}}(\boldsymbol{y}) dy_1 \cdots (dy_j) \cdots dy_k \qquad (j = 1, 2, \dots, k), \tag{2.3}$$

where the multiple integral is calculated with respect to the variables $(y_1, y_2, \dots, y_k)$ except y_j.

Let the parameter space $\Delta_j = \{\boldsymbol{\delta}_j : \boldsymbol{\omega} \in \Omega\}$ obtained by the integration (2.3) be a compact subset of $\mathcal{R}^{k_j}$ for a k_j and the component parameter $\boldsymbol{\omega}_i$ of the mixture (2.2) correspond to $\boldsymbol{\delta}_{ji} \in \Delta_j$ $(j = 1, 2, \dots, k)$. Then some of $\boldsymbol{\delta}_{j1}, \boldsymbol{\delta}_{j2}, \dots, \boldsymbol{\delta}_{jm}$ may equal as can be seen from the example of normal mixture of Henna (2001). So we denote here the different members of $\boldsymbol{\delta}_{j1}, \boldsymbol{\delta}_{j2}, \dots, \boldsymbol{\delta}_{jm}$ by $\boldsymbol{\pi}_{j1}, \boldsymbol{\pi}_{j2}, \dots, \boldsymbol{\pi}_{jm_j}$ anew. Let β_{ji} be the sum of all mixing ratios $\{\alpha_s\}$ of $\{g_{\boldsymbol{\omega}_s}(\boldsymbol{y})\}$ in (2.2), where $g_{\boldsymbol{\omega}_s}(\boldsymbol{y})$ has the same marginal pdf $h_{\boldsymbol{\pi}_{ji}}(y_j)$. Of course, if all of $\boldsymbol{\delta}_{j1}, \boldsymbol{\delta}_{j2}, \ldots, \boldsymbol{\delta}_{jm}$ are different, then $m_j = m$, $\beta_{ji} = \alpha_i$ and $\boldsymbol{\pi}_{ji} = \boldsymbol{\delta}_{ji}$ hold.

Accordingly, if $\boldsymbol{X}$ has the finite mixture (1.1), then $\boldsymbol{Y}$ has the finite mixture (2.2). Furthermore, letting $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_k)'$, then Y_j has the finite mixture

$$h_j(y_j \mid \boldsymbol{c}_{jm_j}) = \sum_{i=1}^{m_j} \beta_{ji} h_{\boldsymbol{\pi}_{ji}}(y_j) \quad (j = 1, 2, \ldots, k), \tag{2.4}$$

where $\boldsymbol{c}_{jm_j} = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jm_j}; \boldsymbol{\pi}_{j1}, \boldsymbol{\pi}_{j2}, \ldots, \boldsymbol{\pi}_{jm_j})$.

Now we adopt $T_j(\boldsymbol{X}) = \boldsymbol{a}_j \boldsymbol{X} + \rho_j$ for the real valued function T mentioned in Section 1, where $\boldsymbol{a}_j$ is the j-th row of $\boldsymbol{M}$ and ρ_j the j-th coordinate of $\boldsymbol{\rho}$. Then $Y_{j\xi} = T_j(\boldsymbol{X}_\xi)$ is the j-th coordinate of $\boldsymbol{Y}_\xi = \boldsymbol{M}\boldsymbol{X}_\xi + \boldsymbol{\rho}$. From the above arguments, $(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n)$ can be considered an independent random sample from the distribution (2.2). Furthermore $(Y_{j1}, Y_{j2}, \ldots, Y_{jn})$ can be considered an independent random sample from the distribution (2.4).

As a preliminary to give an estimator of m, we first construct an estimator of the number m_j of components of (2.4) on the basis of $(Y_{j1}, Y_{j2}, \ldots, Y_{jn})$. For the purpose, assume that $h_{\boldsymbol{\delta}_j}(y)$ is continuous in $\boldsymbol{\delta}_j$ on Δ_j for each y. Let us define parameter spaces by

$$\begin{aligned} \mathcal{C}_\ell^{(j)} = \Bigg\{ & \boldsymbol{c}_\ell : \boldsymbol{c}_\ell = (\beta_1, \beta_2, \ldots, \beta_\ell; \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_\ell), \\ & \sum_{i=1}^{\ell} \beta_i = 1, 0 \le \beta_i \le 1, \boldsymbol{\pi}_i \in \Delta_j, i = 1, 2, \ldots, \ell \Bigg\}, \\ & \qquad (j = 1, 2, \ldots, k; \ell = 1, 2, \ldots). \end{aligned} \tag{2.5}$$

Let $\widehat{\boldsymbol{c}}_{\ell,n} = (\widehat{\beta}_{1,n}, \widehat{\beta}_{2,n}, \ldots, \widehat{\beta}_{\ell,n}; \widehat{\boldsymbol{\pi}}_{1,n}, \widehat{\boldsymbol{\pi}}_{2,n}, \ldots, \widehat{\boldsymbol{\pi}}_{\ell,n})$ be any $\boldsymbol{c}_\ell$ on $\mathcal{C}_\ell^{(j)}$ which minimizes

$$\begin{aligned} S_n(\boldsymbol{c}_\ell) &= \int_{-\infty}^{+\infty} \{H(y \mid \boldsymbol{c}_\ell) - F_n(y)\}^2 dF_n(y) \\ &= \frac{1}{n} \sum_{q=1}^{n} \left\{ \sum_{i=1}^{\ell} \beta_i H_{\boldsymbol{\pi}_i}(Y_{(q)}) - \frac{q}{n} \right\}^2, \end{aligned} \tag{2.6}$$

where $F_n(y)$, $Y_{(q)}$ and $H(y \mid \boldsymbol{c}_\ell)$ are the empirical distribution function, the q-th order statistic of $(Y_{j1}, Y_{j2}, \ldots, Y_{jn})$ and

$$H(y \mid \boldsymbol{c}_\ell) = \sum_{i=1}^{\ell} \beta_i H_{\boldsymbol{\pi}_i}(y) = \sum_{i=1}^{\ell} \beta_i \int_{-\infty}^{y} h_{\boldsymbol{\pi}_i}(t) dt, \tag{2.7}$$

respectively. The existence of $\widehat{\boldsymbol{c}}_{\ell,n}$ is guaranteed since $\mathcal{C}_\ell^{(j)}$ is a compact set and $S_n(\boldsymbol{c}_\ell)$ continuous in $\boldsymbol{c}_\ell$ on $\mathcal{C}_\ell^{(j)}$ from the assumption.

Let us now give an estimater of m_j as follows:

$$\widehat{m}_{j,n} = \text{the minimum integer } \ell \text{ such that } S_n(\widehat{\boldsymbol{c}}_{\ell,n}) < \lambda^2(n)/n, \tag{2.8}$$

where $\lambda(n) \uparrow \infty$, $\lambda^2(n)/n \to 0$ as $n \to \infty$ and $\sum\{\lambda^2(n)/n\}e^{-2\lambda^2(n)} < \infty$.

The existence of $\widehat{m}_{j,n}$ for all n sufficiently large is guaranteed with probability one by Lemma 4.3 of Henna (1985).

The following lemma can be obtained from Theorem 4.1 of Henna (1985) under an identifiability condition (Teicher (1963)).

LEMMA 2.1. *Assume that, for any two finite mixtures* $h_j(y_j \mid \boldsymbol{c}_{\ell_1}^{(1)})$ *and* $h_j(y_j \mid \boldsymbol{c}_{\ell_2}^{(2)})$, *the relationship* $h_j(y_j \mid \boldsymbol{c}_{\ell_1}^{(1)}) = h_j(y_j \mid \boldsymbol{c}_{\ell_2}^{(2)})$ *implies that* $\ell_1 = \ell_2$ *and* $\boldsymbol{c}_{\ell_1}^{(1)} = \boldsymbol{c}_{\ell_2}^{(2)}$, *where* $\boldsymbol{c}_{\ell_1}^{(1)} = \boldsymbol{c}_{\ell_2}^{(2)}$ *means for a permutation of parameter labels. Then we have*

$$\mathrm{P}_{\mathcal{A}_m}^{(\infty)}\{\widehat{m}_{j,n} = m_j \text{ for all } n \text{ sufficiently large}\} = 1. \tag{2.9}$$

Furthermore we can obtain the following immediately from the above lemma.

COROLLARY 2.1. *Assume that the assumption of the last lemma holds for all* $j = 1, 2, \dots, k$. *Then we have*

$$\mathrm{P}_{\mathcal{A}_m}^{(\infty)}\{\widehat{m}_{j,n} = m_j \ (j = 1, 2, \dots, k) \text{ for all } n \text{ sufficiently large}\} = 1. \tag{2.10}$$

The estimator

$$\widehat{m}_n = \max_{1 \le j \le k} \widehat{m}_{j,n} \tag{2.11}$$

could be a good candidate for the estimation of m, but might unfortunately under-estimate the number of components (see the example of Henna (2001) and the following section).

3. An estimator $\widehat{m}_n$ when $\mathcal{F}$ is a family of normal pdf's

Let $\mathcal{F} = \{n(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) : (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \Theta\}$, where $n(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a k-dimensional normal pdf with a mean vector $\boldsymbol{\mu}$ and a variance-covariance matrix $\boldsymbol{\Sigma}$. Consider a finite normal mixture

$$f(\boldsymbol{x} \mid \mathcal{A}_m^\circ) = \sum_{i=1}^{m} \alpha_i^\circ n(\boldsymbol{x} \mid \boldsymbol{\mu}_i^\circ, \boldsymbol{\Sigma}_i^\circ), \tag{3.1}$$

as a special case of (1.1). Let $(\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n)$ be an independent random sample from the distribution (3.1).

In order to give an estimator of the number m, we first construct a sequence $\{\boldsymbol{M}_\gamma\}$ of proper orthogonal matrices as follows:

(i) For $\gamma = 1$, $\boldsymbol{M}_1 = (\boldsymbol{e}_1^{(1)}, \boldsymbol{e}_2^{(1)}, \dots, \boldsymbol{e}_k^{(1)})$ is the $k \times k$ identity matrix.

(ii) For $\gamma \ge 2$, $\boldsymbol{M}_\gamma = (\boldsymbol{e}_1^{(\gamma)}, \boldsymbol{e}_2^{(\gamma)}, \dots, \boldsymbol{e}_k^{(\gamma)})$ is a $k \times k$ orthogonal matrix such that $\boldsymbol{e}_1^{(\gamma)}$ is linearly independent of any $k-1$ vectors in $\{\boldsymbol{e}_i^{(\ell)} : 1 \le i \le k, 1 \le \ell \le \gamma - 1\}$, and $\boldsymbol{e}_j^{(\gamma)}$ is linearly independent of any $k-1$ vectors in $\{\boldsymbol{e}_i^{(\ell)} : 1 \le i \le k, 1 \le \ell \le \gamma - 1\} \cup \{\boldsymbol{e}_1^{(\gamma)}, \boldsymbol{e}_2^{(\gamma)}, \dots, \boldsymbol{e}_{j-1}^{(\gamma)}\}$ when $2 \le j \le k$.

Repeating the arguments of the last section by replacing $\boldsymbol{M}$ with $\boldsymbol{M}_\gamma$, then we can see that $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = n(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ corresponds to $g_{\boldsymbol{\omega}^{(\gamma)}}(\boldsymbol{y}) = n(\boldsymbol{y} \mid \boldsymbol{\mu}^{(\gamma)}, \boldsymbol{\Sigma}^{(\gamma)})$ through $\boldsymbol{Y}^{(\gamma)} = \boldsymbol{M}_\gamma \boldsymbol{X} + \boldsymbol{\rho}$, where $\boldsymbol{\omega}^{(\gamma)} = (\boldsymbol{\mu}^{(\gamma)}, \boldsymbol{\Sigma}^{(\gamma)})$ with $\boldsymbol{\mu}^{(\gamma)} = \boldsymbol{M}_\gamma \boldsymbol{\mu} + \boldsymbol{\rho}$ and $\boldsymbol{\Sigma}^{(\gamma)} = \boldsymbol{M}_\gamma \boldsymbol{\Sigma} \boldsymbol{M}'_\gamma$ (Anderson (1984)). Hence $\Omega^{(\gamma)} = \{\boldsymbol{\omega}^{(\gamma)} : (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \Theta\}$ is a compact subset of $\mathcal{R}^{\frac{1}{2}k(k+1)+k}$. Furthermore we have $h_{\boldsymbol{\delta}_j^{(\gamma)}}(y_j) = n(y_j \mid \mu_j^{(\gamma)}, (\sigma_j^{(\gamma)})^2)$ with $\boldsymbol{\delta}_j^{(\gamma)} = (\mu_j^{(\gamma)}, (\sigma_j^{(\gamma)})^2)$, where $\mu_j^{(\gamma)}$ and $(\sigma_j^{(\gamma)})^2$ are the j-th coordinate of $\boldsymbol{\mu}^{(\gamma)}$ and the (j,j)-th element of $\boldsymbol{\Sigma}^{(\gamma)}$, respectively. Therefore $\Delta_j^{(\gamma)} = \{\boldsymbol{\delta}_j^{(\gamma)} : \boldsymbol{\omega}^{(\gamma)} \in \Omega^{(\gamma)}\}$ is a compact subset of $\mathcal{R}^2$.

Let $Y_{j\xi}^{(\gamma)}$ be the j-th coordinate of $\boldsymbol{Y}_\xi^{(\gamma)} = \boldsymbol{M}_\gamma \boldsymbol{X}_\xi + \boldsymbol{\rho}$, then $(Y_{j1}^{(\gamma)}, Y_{j2}^{(\gamma)}, \dots, Y_{jn}^{(\gamma)})$ can be considered an independent random sample from the distribution

$$h_j(y_j \mid \boldsymbol{c}_{jm_j^{(\gamma)}}^{(\gamma)}) = \sum_{i=1}^{m_j^{(\gamma)}} \beta_{ji}^{(\gamma)} n(y_j \mid \nu_{ji}^{(\gamma)}, (v_{ji}^{(\gamma)})^2) \qquad (j = 1, 2, \dots, k), \tag{3.2}$$

where $\beta_{ji}^{(\gamma)}$ and $(\nu_{ji}^{(\gamma)}, (v_{ji}^{(\gamma)})^2)$ are the parameters obtained by considering $\boldsymbol{\delta}_{ji}^{(\gamma)} = (\mu_{ji}^{(\gamma)}, (\sigma_{ji}^{(\gamma)})^2)$ for $\boldsymbol{\delta}_{ji}$ in construction of (2.4) with $\mu_{ji}^{(\gamma)}$ and $(\sigma_{ji}^{(\gamma)})^2$ being the j-th coordinate of $\boldsymbol{\mu}_i^{(\gamma)} = \boldsymbol{M}_\gamma \boldsymbol{\mu}_i^\circ + \boldsymbol{\rho}$ and the (j,j)-th element of $\boldsymbol{\Sigma}_i^{(\gamma)} = \boldsymbol{M}_\gamma \boldsymbol{\Sigma}_i^\circ \boldsymbol{M}'_\gamma$, respectively. Here all of $(\nu_{j1}^{(\gamma)}, (v_{j1}^{(\gamma)})^2), (\nu_{j2}^{(\gamma)}, (v_{j2}^{(\gamma)})^2), \dots, (\nu_{jm_j^{(\gamma)}}^{(\gamma)}, (v_{jm_j^{(\gamma)}}^{(\gamma)})^2)$ which correspond to $\boldsymbol{\pi}_{j1}, \boldsymbol{\pi}_{j2}, \dots, \boldsymbol{\pi}_{jm_j}$ of (2.4) are different.

We give an estimator $\widehat{m}_{j,n}^{(\gamma)}$ of $m_j^{(\gamma)}$ on the basis of $(Y_{j1}^{(\gamma)}, Y_{j2}^{(\gamma)}, \dots, Y_{jn}^{(\gamma)})$ in a similar way to (2.8). Yakowitz and Spragins (1968) showed that the family of all finite mixtures of k-dimensional normal distributions is identifiable. Hence the condition of Corollary 2.1 is satisfied. Accordingly we have the following provided that $\boldsymbol{M}_1, \boldsymbol{M}_2, \dots, \boldsymbol{M}_s$ are matrices constructed by the procedure (i) and (ii).

LEMMA 3.1. *For any given positive integer s, we have*

$$\mathrm{P}_{\mathcal{A}_m^\circ}^{(\infty)}\{\widehat{m}_{j,n}^{(\gamma)} = m_j^{(\gamma)} \ (j = 1, 2, \dots, k; \gamma = 1, 2, \dots, s) \text{ for all } n \text{ sufficiently large}\} = 1. \tag{3.3}$$

Unfortunately, for any γ and j, the number $m_j^{(\gamma)}$ is not necessarily equal to m as can be seen from the example of normal mixture of Henna (2001). However we can obtain the following.

THEOREM 3.1. *Assume that all of $\boldsymbol{\mu}_1^\circ, \boldsymbol{\mu}_2^\circ, \dots, \boldsymbol{\mu}_m^\circ$ are different and $m(m-1)(k-1) < 2ks$ holds. Then there exist $\gamma(\le s)$ and j such that*

$$\mathrm{P}_{\mathcal{A}_m^\circ}^{(\infty)}\{\widehat{m}_{j,n}^{(\gamma)} = m \text{ for all } n \text{ sufficiently large}\} = 1. \tag{3.4}$$

PROOF. From Theorem 3.1 of Henna (2001), there exist γ and j such that $(Y_{j1}^{(\gamma)}, Y_{j2}^{(\gamma)}, \dots, Y_{jn}^{(\gamma)})$ can be considered an independent random sample from a finite mixture of m one-dimensional normal pdf's with different means $\mu_{j1}^{(\gamma)}, \mu_{j2}^{(\gamma)}, \dots, \mu_{jm}^{(\gamma)}$, where $\mu_{ji}^{(\gamma)}$ is the j-coordinate of $\boldsymbol{\mu}_i^{(\gamma)} = \boldsymbol{M}_\gamma \boldsymbol{\mu}_i^\circ + \boldsymbol{\rho}$. Accordingly we have the conclusion by Lemma 2.1. □

However, we cannot know which γ and j satisfy $m_j^{(\gamma)} = m$. So we cannot construct $\widehat{m}_{j,n}^{(\gamma)}$ which satisfies (3.4) actually. Hence we need to give another estimator. As can be seen from the construction of (3.2), $m_j^{(\gamma)} \leq m$ holds for any γ and j. So, we give an estimator of m as follows:

$$\widehat{m}_n(s) = \max_{1\leq\gamma\leq s}\left\{\max_{1\leq j\leq k}\widehat{m}_{j,n}^{(\gamma)}\right\} \qquad (s = 1, 2, \dots). \tag{3.5}$$

If $m(m-1)(k-1) < 2ks$, then at least one of $\{m_j^{(\gamma)} : j = 1, 2, \dots, k; \gamma = 1, 2, \dots, s\}$ equals to m by Lemma 3.1 of Henna (2001). Hence the following is an immediate consequence of Lemma 3.1 and Theorem 3.1.

THEOREM 3.2. *Under the assumption of the last theorem, we have*

$$\mathrm{P}_{\mathcal{A}_m^\circ}^{(\infty)}\{\widehat{m}_n(s) = m \text{ for all } n \text{ sufficiently large}\} = 1. \tag{3.6}$$

But as m is unknown, we cannot know at a given step s whether the condition $m(m-1)(k-1) < 2ks$ holds or not. So we cannot know when stopping the algorithm to give $\widehat{m}_n(s)$ which satisfies (3.6) actually. Hence we need to give another estimator. For the purpose, again by $m_j^{(\gamma)} \leq m$ for any γ and j, we can obtain the following from Lemma 3.1.

LEMMA 3.2. *For any given positive integer s_1, we have*

$$\mathrm{P}_{\mathcal{A}_m^\circ}^{(\infty)}\{\widehat{m}_n(s) \leq m \ (s = 1, 2, \dots, s_1) \text{ for all } n \text{ sufficiently large}\} = 1. \tag{3.7}$$

Let $s_\circ$ be the minimum positive integer s such as $m(m-1)(k-1) < 2ks$. Then the following is an immediate consequence of the last theorem.

LEMMA 3.3. *Assume that all of $\boldsymbol{\mu}_1^\circ, \boldsymbol{\mu}_2^\circ, \dots, \boldsymbol{\mu}_m^\circ$ are different. Then, for any given positive integer s_1 such as $s_\circ \leq s_1$, we have*

$$\mathrm{P}_{\mathcal{A}_m^\circ}^{(\infty)}\{\widehat{m}_n(s) = m \ (s = s_\circ, s_\circ + 1, \dots, s_1) \text{ for all } n \text{ sufficiently large}\} = 1. \tag{3.8}$$

If we construct $\boldsymbol{M}_1, \boldsymbol{M}_2, \dots$ sequentially, then we can necessarily obtain $\boldsymbol{M}_1, \boldsymbol{M}_2, \dots, \boldsymbol{M}_s$ such as $m(m-1)(k-1) < 2ks$ before long. Hence, if we construct $\widehat{m}_n(1), \widehat{m}_n(2), \dots$ sequentially, then we can necessarily obtain a consistent estimator $\widehat{m}_n(s)$ which satisfies (3.6) before long. As can be seen from the definition, for the given $(\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n)$, the estimator $\widehat{m}_n(s)$ is monotone increasing with respect to s. In addition, if all of $\boldsymbol{\mu}_1^\circ, \boldsymbol{\mu}_2^\circ, \dots, \boldsymbol{\mu}_m^\circ$ are different, it can be considered that $\widehat{m}_n(s) \leq m$ when $s \leq s_\circ - 1$ and $\widehat{m}_n(s) = m$ when $s_\circ \leq s \leq s_1$ for n sufficiently large by Lemmas 3.2 and 3.3, respectively. Hence, it can be considered that the sequence $\widehat{m}_n(1), \widehat{m}_n(2), \dots$

may become invariant soon for n sufficiently large. Taking into account of these, we give an estimator as follows:

$$\widehat{m}_n = \widehat{m}_n(s_\circ^*), \tag{3.9}$$

where $s_\circ^*$ is the minimum positive integer s such as $\widehat{m}_n(s) = \widehat{m}_n(s+1) = \cdots = \widehat{m}_n(s+s_1^*)$ with s_1^* a given positive integer.

The existence of $\widehat{m}_n$ is guaranteed with probability one, for n sufficiently large, by the last lemma. It can be seen that $\widehat{m}_n$ is given without any knowledge about m other than a fact that m is finite. If s_1^* is sufficiently large, then $m(m-1)(k-1) < 2k(s_\circ^* + s_1^*)$ may hold. Accordingly, we can easily have the following from the last two lemmas.

THEOREM 3.3. *Assume that all of $\boldsymbol{\mu}_1^\circ, \boldsymbol{\mu}_2^\circ, \dots, \boldsymbol{\mu}_m^\circ$ are different. Then, for s_1^* sufficiently large, we have*

$$\mathrm{P}_{\mathcal{A}_m^\circ}^{(\infty)}\{\widehat{m}_n = m \textit{ for all } n \textit{ sufficiently large}\} = 1. \tag{3.10}$$

The last theorem states the asymptotic behavior using $\widehat{m}_n$. No more reference to the $m(m-1)(k-1) < 2ks$ condition is needed. In fact, as m is finite, we are assure that there exists an integer s_1^* such that $m(m-1)(k-1) < 2k(s_\circ^* + s_1^*)$, so that at least one of $\{m_j^{(\gamma)} : j = 1, 2, \dots, k; \gamma = 1, 2, \dots, s_\circ^* + s_1^*\}$ is equal to m by Lemma 3.1 of Henna (2001).

When implementing the algorithm, the problem is to determine how long must be the invariance of the sequence to decide that the optimum is reached. And here, we have no way to do that except to consider an upper bound using applicable arguments or to define a priori a length $s_1^* = 5$ for example (but may be are there also linear algebra considerations that can lead to sufficient conditions?).

4. An estimator $\widehat{m}_n$ when $\mathcal{F}$ is a known finite family of normal pdf's

Let $\mathcal{F}$ be that of the last section with a known finite set Θ of L elements. Then, for any $\boldsymbol{M}$ and $\boldsymbol{\rho}$, $\Omega = \{\boldsymbol{\omega} : (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \Theta\}$ is a known finite subset of $\mathcal{R}^{\frac{1}{2}k(k+1)+k}$, where $\boldsymbol{\omega} = (\boldsymbol{\nu}, \boldsymbol{\Sigma}^*)$ with $\boldsymbol{\nu} = \boldsymbol{M}\boldsymbol{\mu} + \boldsymbol{\rho}$ and $\boldsymbol{\Sigma}^* = \boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}'$. Furthermore $\Delta_j = \{\boldsymbol{\delta}_j : \boldsymbol{\omega} \in \Omega\}$ is a known finite subset of $\mathcal{R}^2$ with $\boldsymbol{\delta}_j = (\nu_j, \sigma_j^2)$, where ν_j and σ_j^2 are the j-th coordinate of $\boldsymbol{\nu}$ and the (j,j)-th element of $\boldsymbol{\Sigma}^*$, respectively.

Defining $\mathcal{C}_\ell^{(j)}$ and $S_n(\widehat{\boldsymbol{c}}_{\ell,n})$ in a similar way to those of Section 2, we give an estimator of m_j as follows:

$$\widehat{m}_{j,n} = \begin{cases} \text{the minimum integer } \ell \ (\le L-1) \text{ such as } S_n(\widehat{\boldsymbol{c}}_{\ell,n}) < \lambda^2(n)/n \\ \text{or} \\ L \text{ if } S_n(\widehat{\boldsymbol{c}}_{\ell,n}) \ge \lambda^2(n)/n \text{ for all } \ell \ (\le L-1), \end{cases} \tag{4.1}$$

where λ is that of (2.8). Then we have the following.

THEOREM 4.1. *Assume that all of $\boldsymbol{\mu}_1^\circ, \boldsymbol{\mu}_2^\circ, \dots, \boldsymbol{\mu}_m^\circ$ are different. Then, for any $\boldsymbol{\rho}$, there exists an $\boldsymbol{M}$, such that*

$$\mathrm{P}_{\mathcal{A}_m^\circ}^{(\infty)}\{\widehat{m}_{j,n} = m \ (j = 1, 2, \dots, k) \textit{ for all } n \textit{ sufficiently large}\} = 1. \tag{4.2}$$

PROOF. From the assumption, the parameters $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \dots, (\boldsymbol{\mu}_L, \boldsymbol{\Sigma}_L)$ of pdf's in $\mathcal{F}$ are known. So, without loss of generality, we may assume that all of $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_L$ are different. Then, for any $\boldsymbol{\rho}$, we can obtain an $\boldsymbol{M}$ such that all of $\nu_{j1}, \nu_{j2}, \dots, \nu_{jL}$ are different for $j = 1, 2, \dots, k$ from Lemma 2.1 of Henna (2001), where ν_{ji} is the j-th coordinate of $\boldsymbol{\nu}_i = \boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{\rho}$. So, if $\boldsymbol{\mu}_1^\circ, \boldsymbol{\mu}_2^\circ, \dots, \boldsymbol{\mu}_m^\circ$ are different, then $\boldsymbol{\delta}_{j1}^\circ, \boldsymbol{\delta}_{j2}^\circ, \dots, \boldsymbol{\delta}_{jm}^\circ$ are different, where $\boldsymbol{\delta}_{ji}^\circ = (\mu_{ji}^\circ, \sigma_{ji}^{\circ 2})$ with μ_{ji}° and $\sigma_{ji}^{\circ 2}$ being the j-th coordinate of $\boldsymbol{M}\boldsymbol{\mu}_i^\circ + \boldsymbol{\rho}$ and the (j,j)-th element of $\boldsymbol{M}\boldsymbol{\Sigma}_i^\circ \boldsymbol{M}'$, respectively. Hence we have $m_j = m$ for $j = 1, 2, \dots, k$. Accordingly, we have the conclusion from Corollary 2.1. □

An inequality $(\nu_{j1} - \nu_{j2})^2 \leq \sum_{i=1}^k (\mu_{i1} - \mu_{i2})^2$ holds, that is, a distance between $n(y_j \mid \nu_{j1}, \sigma_{j1}^2)$ and $n(y_j \mid \nu_{j2}, \sigma_{j2}^2)$ is smaller than that between $n(\boldsymbol{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $n(\boldsymbol{x} \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. So it can be considered that there is a case where the detection of distinction between $n(y_j \mid \nu_{j1}, \sigma_{j1}^2)$ and $n(y_j \mid \nu_{j2}, \sigma_{j2}^2)$ becomes difficult depending on $\boldsymbol{M}$. Hence it can be guessed that the estimation by (4.1) tends to give underestimates. So we give an estimator of m again as follows:

$$\widehat{m}_n = \max_{1 \leq j \leq k} \widehat{m}_{j,n}. \tag{4.3}$$

Then the following is an immediate consequence of the last theorem.

THEOREM 4.2. *Under the assumption of the last theorem, for any* $\boldsymbol{\rho}$, *there exists an* $\boldsymbol{M}$ *such that*

$$\mathrm{P}_{\mathcal{A}_m^\circ}^{(\infty)}\{\widehat{m}_n = m \textit{ for all } n \textit{ sufficiently large}\} = 1. \tag{4.4}$$

5. Some simulation results

Now we give some simulation results for Theorem 4.2. The family considered here is $\mathcal{F} = \{n(\boldsymbol{x} \mid \boldsymbol{\mu}_i, I) : i = 1, 2, 3, 4\}$, where $\boldsymbol{\mu}_1 = (0,0)'$, $\boldsymbol{\mu}_2 = (0,4)'$, $\boldsymbol{\mu}_3 = (4,4)'$, $\boldsymbol{\mu}_4 = (4,8)'$ and I is the 2×2 identity matrix. In order to obtain one-dimensional independent random samples, let us have a try with

$$\boldsymbol{M} = \begin{pmatrix} 0.717106 & 0.696964 \\ -0.696964 & 0.717106 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\rho} = \mathbf{0}.$$

Using random numbers produced by The Institute of Statistical Mathematics, 5000 two-dimensional samples of sizes $n = 200$, 300, 400 and 500 were generated from a single normal pdf and from various mixtures of $\mathcal{F}$, respectively.

As a criterion, $\lambda(n) = (\log\log n)^2/5n$ was used, though there was no theorecical reason for this to be the optimum in the class of λ's satisfying the condition of (2.8). It seems that simulation results given below show that the criterion is fairly effective when the mixing ratios are nearly equal values. However, the questions of which is the optimum in the class of λ's satisfying the condition of (2.8) and which is the optimum in the class of orthogonal $\boldsymbol{M}$'s are worthy of further research.

Table 1 gives us, for various sample sizes, the percentages of exact estimate of m by the estimator $\widehat{m}_n = \max\{\widehat{m}_{1,n}, \widehat{m}_{2,n}\}$ for a single normal pdf $n(\boldsymbol{x} \mid \boldsymbol{\mu}_1, I)$, for two

Table 1. Percentages of $\hat{m}_n = m$.

	$m = 1$	$m = 2$	$m = 3$	$m = 4$
$n = 200$	94.9	100.0	99.9	30.0
$n = 300$	96.4	100.0	100.0	80.4
$n = 400$	96.7	100.0	100.0	97.6
$n = 500$	97.0	100.0	100.0	99.7

Table 2. Percentages of $\hat{m}_n = m$.

	$m = 2$	$m = 3$	$m = 4$
$n = 200$	100.0	98.9	9.8
$n = 300$	100.0	100.0	39.4
$n = 400$	100.0	100.0	70.7
$n = 500$	100.0	100.0	89.6

components $\frac{1}{2}n(\boldsymbol{x} \mid \boldsymbol{\mu}_1, I) + \frac{1}{2}n(\boldsymbol{x} \mid \boldsymbol{\mu}_2, I)$, for three components $\frac{1}{3}n(\boldsymbol{x} \mid \boldsymbol{\mu}_1, I) + \frac{1}{3}n(\boldsymbol{x} \mid \boldsymbol{\mu}_2, I) + \frac{1}{3}n(\boldsymbol{x} \mid \boldsymbol{\mu}_3, I)$ and for four components $\frac{1}{4}n(\boldsymbol{x} \mid \boldsymbol{\mu}_1, I) + \frac{1}{4}n(\boldsymbol{x} \mid \boldsymbol{\mu}_2, I) + \frac{1}{4}n(\boldsymbol{x} \mid \boldsymbol{\mu}_3, I) + \frac{1}{4}n(\boldsymbol{x} \mid \boldsymbol{\mu}_4, I)$, respectively.

Table 2 gives us the same to the above for two components $\frac{4}{10}n(\boldsymbol{x} \mid \boldsymbol{\mu}_1, I) + \frac{6}{10}n(\boldsymbol{x} \mid \boldsymbol{\mu}_2, I)$, for three components $\frac{3}{10}n(\boldsymbol{x} \mid \boldsymbol{\mu}_1, I) + \frac{3}{10}n(\boldsymbol{x} \mid \boldsymbol{\mu}_2, I) + \frac{4}{10}n(\boldsymbol{x} \mid \boldsymbol{\mu}_3, I)$ and for four components $\frac{2}{10}n(\boldsymbol{x} \mid \boldsymbol{\mu}_1, I) + \frac{2}{10}n(\boldsymbol{x} \mid \boldsymbol{\mu}_2, I) + \frac{3}{10}n(\boldsymbol{x} \mid \boldsymbol{\mu}_3, I) + \frac{3}{10}n(\boldsymbol{x} \mid \boldsymbol{\mu}_4, I)$, respectively.

Acknowledgements

The author wishes to express his hearty thanks to the referees for their valuable comments and suggestions.

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd ed., John Wiley, New York.

Billingsley, P. (1986). *Probability and Measure*, 2nd ed., John Wiley, New York.

Chen, H., Chen, J. and Kalbfleisch, J. D. (2001). A modified likelihood test for homogeneity in finite mixture models, *Journal of the Royal Statistical Society: Series B*, **63**, 19–29.

Chen, J. and Kalbfleisch, J. D. (1996). Penalized minimum-distance estimates in finite mixture models, *The Canadian Journal of Statistcs*, **24**, 167–175.

Feng, Z. D. and McCulloch, C. E. (1994). On the likelihood ratio test statistic for the number of components in a normal mixture with unequal variance, *Biometrics*, **50**, 1158–1162.

Garel, B. (2001). Likelihood ratio test for univariate Gaussian mixture, *Journal of Statistical Planning and Inference*, **96**, 325–350.

Henna, J. (1985). On estimating of the number of constituents of a finite mixture of continuous distributions, *Annals of the Institute of Statistical Mathematics*, **37**, 235–240.

Henna, J. (2001). Marginal distributions of finite mixtures of multivariate normal distributions, *Journal of The Japan Statistical Society*, **31**, 187–191.

Keribin, C. (2000). Consistent estimation of the order of mixture models, *Sankhyā*, **62**, 49–66.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models*, Marcel Dekker, New York.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*, John Wiley, New York.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components, *Journal of the Royal Statistical Society: Series B*, **59**, 731–792.

Teicher, H. (1963). Identifiablity of finite mixtures, *The Annals of Mathematical Statistics*, **34**, 1265–1269.

Titterington, D. M. (1990). Some recent research in the analysis of mixture distributions, *Statistics*, **21**, 619–641.

Titterington, D. M. *et al.* (1985). *Statistical Analysis of Finite Mixture Distributions*, John Wiley, New York.

Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures, *The Annals of Mathematical Statistics*, **39**, 209–214.