Semiparametric estimation of a two-component mixture model where one component is known

Laurent Bordes, Céline Delmas, Pierre Vandekerkhove

doi: 10.1111/j.1467-9469.2006.00515.x
 Published by Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA Vol 33: 733–752, 2006

Semiparametric Estimation of a Two-component Mixture Model where One Component is known

LAURENT BORDES
LMAC, Université de Technologie de Compiègne

CÉLINE DELMAS
SAGA, INRA

PIERRE VANDEKERKHOVE
LAMA, Université de Marne-la-Vallée

ABSTRACT. We consider a two-component mixture model where one component distribution is known while the mixing proportion and the other component distribution are unknown. These kinds of models were first introduced in biology to study the differences in expression between genes. The various estimation methods proposed till now have all assumed that the unknown distribution belongs to a parametric family. In this paper, we show how this assumption can be relaxed. First, we note that generally the above model is not identifiable, but we show that under moment and symmetry conditions some 'almost everywhere' identifiability results can be obtained. Where such identifiability conditions are fulfilled we propose an estimation method for the unknown parameters which is shown to be strongly consistent under mild conditions. We discuss applications of our method to microarray data analysis and to the training data problem. We compare our method to the parametric approach using simulated data and, finally, we apply our method to real data from microarray experiments.

Key words: identifiability, microarray data, mixture, multiple test hypothesis, semiparametric, training data

1. Introduction

In this work, we consider the two-component mixture model defined by

$$g(x)=(1-p)f_0(x)+pf(x-\mu), \quad \forall x\in\mathbb{R}, \tag{1}$$

where the probability density function (pdf) f_0 is known and the unknown parameters are the mixing proportion $p\in(0,1)$, the non-null location parameter $\mu\in\mathbb{R}$ and an even pdf f. Such mixture models, semiparametric or non-parametric, have been recently studied by Hall & Zhou (2003), Bordes *et al.* (2006), Cruz-Medina & Hettmansperger (2004), Hunter *et al.* (2004) and can be situated between fully parametric mixture models and non-parametric mixture models (for an overview of classical mixture models we refer the reader to McLachlan & Peel, 2000).

The introduction of model (1) is motivated by the problem of detection of differentially expressed genes under two or more conditions in microarray data (conditions might be, e.g. 'healthy tissue versus diseased tissue', 'brain versus kidney', etc.). For this purpose a test statistic is built for each gene. Under the null hypothesis, corresponding to a lack of difference in expression, it has a known distribution (in general Student's or Fisher). We then observe the response of thousands of genes, which corresponds in practice to thousands of observations from statistical tests. The sample obtained in this way comes from a mixture of two

distributions: the known distribution f_0 (for the genes under the null hypothesis) and another distribution corresponding to $f(\cdot - \mu)$, which is the unknown distribution of the test statistics under the alternative hypothesis. Once the parameters p, μ and f have been estimated we can estimate the probability that a gene belongs to the null component of the mixture distribution conditionally on the observations. Therefore, using a classification criterion we allocate each gene to a component and then we distinguish the genes differentially expressed from the genes non-differentially expressed.

Model (1) appears as an alternative to parametric mixture models (working paper by C. Delmas, 2005), where the law under the alternative hypothesis is unknown. For a survey of these methods and a discussion of the issues at stake in these kinds of applications we refer the reader to Dudoit *et al.* (2002) and McLachlan *et al.* (2004).

Another important issue is the *training data* problem. We look at the problem of estimating all the parameters in (1), i.e. p, μ, f and f_0 (which this time is unknown) when, in addition to a sample of g-distributed random variables, a sample of f_0-distributed random variables is available (training data from the first component). In the classical training data problem data are available from each component (e.g. see Titterington *et al.*, 1985), and so the novelty here is that training data are available for only one of the two components of model (1).

The paper is organized as follows. The following section is devoted to the identifiability problem. First we show that model (1) is not identifiable *in general* even if it is locally identifiable. Then we give some sufficient conditions for achieving identifiability. In section 3, we propose an inference procedure based on the symmetry of the unknown component of the model, and then in section 4 we show that by solving the moment equations we can also estimate the unknown Euclidean part of the model. In section 5, we show that if model (1) is identifiable, estimators of unknown parameters are strongly consistent. Section 6 is devoted to a precise description of the two applications we introduced above, and section 7 presents simulation results and an application to a real data set. Future issues concerning such kind of semiparametric mixture models are finally discussed in section 8.

2. Identifiability

2.1. Some non-identifiable cases

From a general point of view model (1) is not identifiable, as the two following examples show.

$$\tfrac{3}{4}u_{-1,1}(x) + \tfrac{1}{4}u_{-3,3}(x-4) = \tfrac{2}{3}u_{-1,1}(x) + \tfrac{1}{3}u_{-4,4}(x-3), \quad \forall x \in \mathbb{R}, \tag{2}$$

where $u_{a,b}$ is the uniform pdf on (a,b) with a and b two real parameters such that $a<b$, and

$$(1-p)\varphi(x) + pf(x-1) = \left(1-\frac{p}{2}\right)\varphi(x) + \frac{p}{2}\varphi(x-2), \quad \forall x \in \mathbb{R}, \tag{3}$$

where φ is any even pdf, $p \in (0,1)$ and $f(x) = (\varphi(x-1) + \varphi(x+1))/2$.

Clearly, the two above examples show that without any additional assumptions on the model we cannot obtain an identifiability result. However, in the next section we shall see that there are some limitations to non-identifiability.

2.2. Local identifiability via moment equations

The previous examples show that identifiability of model (1) cannot be expected for $(p, \mu, f) \in (0,1) \times \mathbb{R}^* \times \mathcal{F}$, where $\mathcal{F}$ is the set of even pdf defined on $\mathbb{R}$. However, if we assume that f_0 has a third-order moment and that f belongs to

$$\mathcal{F}_3=\{f\in\mathcal{F};\int_{\mathbb{R}}|x|^3f(x)\,\mathrm{d}x<+\infty\},$$

then the moment equations lead to local identifiability of the model. Let us consider the equation

$$(1-p)f_0(x)+pf(x-\mu)=(1-p_1)f_0(x)+p_1f_1(x-\mu_1),\quad \forall x\in\mathbb{R}, \tag{4}$$

for fixed values of $(p,\mu,f)\in(0,1)\times\mathbb{R}\backslash\{\mu^{(0)}\}\times\mathcal{F}_3$. We denote by $\mu^{(0)}$ the mean of the pdf f_0.

Proposition 1
Equation (4) has at most two solutions $(p_1,\mu_1,f_1)\in(0,1)\times\mathbb{R}\backslash\{\mu^{(0)}\}\times\mathcal{F}_3$ if f_0 is a symmetric pdf and at most three solutions otherwise.

Proof. As f_0 is known we can, up to a translation, assume that f_0 has a null first moment. Therefore, we assume from now on that μ and μ_1 belong to $\mathbb{R}^*$. The first three moment equations are:

$$\begin{cases} p\mu=p_1\mu_1 \\ (1-p)\theta_0+p(\mu^2+\theta)=(1-p_1)\theta_0+p_1(\mu_1^2+\theta_1) \\ p(3\mu\theta+\mu^3)=p_1(3\mu_1\theta_1+\mu_1^3), \end{cases} \tag{5}$$

where θ_0, θ and θ_1 are, respectively, the second-order moments of f_0, f and f_1. Then, because it is easy to check (see appendix A for details) that μ_1 is the zero of a two-order polynomial, we obtain that either $(p_1,\mu_1,\theta_1)=(p,\mu,\theta)$ or

$$\begin{cases} p_1=p\left(\dfrac{2\mu^2}{3\theta+\mu^2-3\theta_0}\right) \\ \mu_1=\mu+\dfrac{3\theta-\mu^2-3\theta_0}{2\mu} \\ \theta_1=\theta+\dfrac{(\theta+\mu^2-\theta_0)(3\theta_0+\mu^2-3\theta)}{4\mu^2}. \end{cases} \tag{6}$$

Note that if f_0 is not symmetric with third-order moment equal to γ_0, the first two equations in (5) are unchanged, whereas the third equation becomes

$$(1-p)\gamma_0+p(3\mu\theta+\mu^3)=(1-p_1)\gamma_0+p_1(3\mu_1\theta_1+\mu_1^3).$$

Then, with this new system of equations we obtain that either $\mu=\mu_1$ or

$$-2\mu\mu_1^2+(3\theta-3\theta_0+\mu^2)\mu_1+\gamma_0=0.$$

It follows that there are at most three solutions for (p_1,μ_1,θ_1). As, from (1), we have

$$f(x)=\frac{g(x+\mu)-(1-p)f_0(x+\mu)}{p},\quad x\in\mathbb{R}. \tag{7}$$

It can be seen that the pdf f is uniquely determined by g, f_0, p and μ. The proposition is proved.

The above proposition proves that in examples (2) and (3) there is no other way to write the mixture, because in both examples f_0 is an even pdf. Note also that this proposition leads to a local identifiability result. Indeed, as $(p_1,\mu_1,f_1)=(p,\mu,f)$ is a solution of (4) there exists a neighbourhood of (p,μ) where (p,μ,f) is the unique solution of (4). Note also that if $(p,\mu,\theta)=(1/3,3,16/3)$ then by (6) we obtain $(p_1,\mu_1,\theta_1)=(1/4,4,3)$, which corresponds to the non-identifiability example given in (2).

2.3. Identifiability and characteristic functions

In this section, we investigate identifiability for model (1) when f_0 is a symmetric pdf having a third-order moment, or equivalently, when $f_0 \in \mathcal{F}_3$ is an even function (if $\mu^{(0)}$ is the known symmetry point of f_0, then consider $g(\cdot + \mu^{(0)})$). Let us look at (4) for (p, μ, f) and (p_1, μ_1, f_1) in $(0, 1) \times \mathbb{R}^* \times \mathcal{F}_3$. Denoting by $\hat{f}$ the Fourier transform (or characteristic function) of a pdf f, we obtain the following equations by identifying the real and imaginary parts of the Fourier transform of (4):

$$0 = \det \begin{pmatrix} \hat{f}(t) & p_1 \sin(\mu_1 t) \\ \hat{f}_1(t) & p \sin(\mu t) \end{pmatrix}, \quad \forall t \in \mathbb{R} \tag{8}$$

and

$$(p_1 - p)\hat{f}_0(t) = \det \begin{pmatrix} \hat{f}(t) & p_1 \cos(\mu_1 t) \\ \hat{f}_1(t) & p \cos(\mu t) \end{pmatrix}, \quad \forall t \in \mathbb{R}. \tag{9}$$

The following proposition gives an identifiability result when $\hat{f}_0 > 0$, which is true, e.g. for Gaussian or Student centred distributions.

Proposition 2

The mixture model (1), with $f_0 \in \mathcal{F}_3$ and $\hat{f}_0 > 0$, is identifiable if

$$(p, \mu, f) \in (0, 1) \times \mathbb{R}^* \times \mathcal{F}_3 \quad \text{and} \quad \theta \neq \theta_0 + \frac{k \pm 2}{3k}\mu^2, \quad \forall k \in \mathbb{N}^*,$$

where θ and θ_0 are the second-order moments of f and f_0 respectively.

Proof. Multiplying (9) by $\sin(t\mu)$ and using (8) we get the following equation

$$(p_1 - p)\sin(\mu t)\hat{f}_0(t) = p_1 \hat{f}_1(t) \sin(t(\mu - \mu_1)), \quad \forall t \in \mathbb{R}.$$

Because $\hat{f}_0 > 0$, the above equation implies that $\sin(\mu t) = 0$ whenever $\sin(t(\mu - \mu_1)) = 0$. By considering the particular argument value $t^* = \pi/(\mu - \mu_1)$ we obtain that:

$$\sin(t^* \mu) = \sin\left(\left[\frac{\mu}{\mu - \mu_1}\right]\pi\right) = 0 \Rightarrow \frac{\mu}{\mu - \mu_1} \in \mathbb{N}.$$

But according to proposition 1 there exists at most one other solution $\mu_1 \neq \mu$ to problem (4), which in turn implies that there exists at most one positive integer k_0 such that $|\mu| = k_0|\mu - \mu_1|$. From this last equality it follows that

$$\mu_1 = \frac{k_0 \pm 1}{k_0}\mu.$$

The above equality, together with the second equality in (6), entails:

$$\theta = \theta_0 + \mu^2\left(\frac{k_0 \pm 2}{3k_0}\right). \tag{10}$$

Finally, if f belongs to the set of densities that do not satisfy (10), we have $\mu_1 = \mu$, and then, from the first moment equation we obtain $p_1 = p$, and from (8) we obtain $\hat{f}_1 = \hat{f}$ or equivalently $f_1 = f$ almost everywhere on $\mathbb{R}$ (with respect to the Lebesgue measure).

Using the first-order moment and (8) we show that $f = f_1$ almost everywhere (with respect to the Lebesgue measure) whenever $p = p_1$ or $\mu = \mu_1$. It follows that model (1) is identifiable whenever $(\mu, \theta) = (\mu_1, \theta_1)$, i.e. if

$$(\mu, \theta) \in \Phi = \{(\mu, \theta) \in \mathbb{R}^* \times (0, +\infty)\} \setminus \bigcup_{k \in \mathbb{N}^*} \left\{\left(\frac{\mu, \theta_0 + \mu^2(k \pm 2)}{3k}\right); \mu \in \mathbb{R}^*\right\}.$$

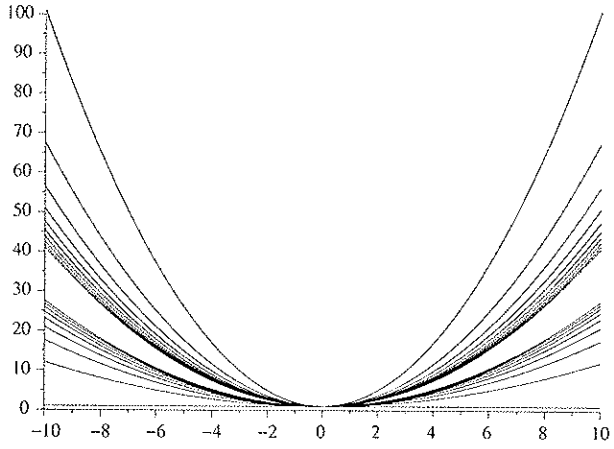

Fig. 1. Model (1) is identifiable if $(\mu, 0)$ does not belong to one of the curves.

Identifiability is therefore obtained on $\mathbb{R}^* \times (0, +\infty)$ except on a set (of uncertainty) with Lebesgue measure equal to 0. We can see in Fig. 1 below the domain of identifiability for $(\mu, 0) \in (0, +\infty)^2$. Notice that the second non-identifiable case given in (3) (with φ and f symmetric) satisfies $\theta = \theta_0$, which is a particular condition of uncertainty as $\theta = \theta_0 + \mu^2\left((2-2)/6\right)$, corresponding to $k_0 = 2$ in (10).

To conclude this section we now remark that there are at least two other cases where identifiability of model (1) holds.

Proposition 3

(i) The mixture model (1) is identifiable if $f_0 > 0$, f is an even function and both have first-order moments and satisfy

$$\forall \beta \in \mathbb{R}, \quad \lim_{x \to +\infty} \frac{f(x-\beta)}{f_0(x)} = 0, \quad \text{or} \quad \lim_{x \to -\infty} \frac{f(x-\beta)}{f_0(x)} = 0. \tag{11}$$

(ii) The mixture model (1) is identifiable if $f > 0$ has a first-order moment, and there exists a real number $a > 0$ such that for all $|x| > a$ we have $f_0(x) = 0$ and $f(x) = f(-x)$.

Proof. (i) If condition (11) is satisfied we have $1 - p = \lim_{x \to +\infty} g(x)/f_0(x)$ (or $1 - p = \lim_{x \to -\infty} g(x)/f_0(x)$ for convenience) and then by (4) we have $p = p_1$. From the first moment equation, denoting by m and m_0 the first-order moments of g and m respectively, we have $\mu = (m - (1-p)m_0)/p$, and thus $\mu = \mu_1$ and by (7) f is unique.

(ii) For all $x \in \mathbb{R}$ such that $|x| > a$ we have by (4): $pf(x-\mu) = p_1 f(x-\mu_1)$. Then, for large values of $|x|$ we obtain

$$pf(x + \mu_1 - \mu) = p_1 f_1(x) = p_1 f_1(-x) = pf(\mu_1 - \mu - x),$$

which is possible only if $\mu = \mu_1$. The end of the proof is the same as for case (i).

We note that cases (i) and (ii) of proposition 3 do not require the symmetry of f_0, which can be useful in microarray data analysis with more than two conditions (see section 6.1).

3. Estimating the Euclidean parameter by symmetrization

Suppose that we observe n independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ with cumulative distribution function (cdf) G defined by model (1), i.e.

$$G(x) = (1-p)F_0(x) + pF(x-\mu), \quad \forall x \in \mathbb{R},$$

where G, F_0 and F are cdfs corresponding to pdfs g, f_0 and f respectively. From now on, we assume that f_0 is the density of a centred distribution. If it is not, we have simply to change the X_is into $X_i - m_0$, where $m_0 = \int_{\mathbb{R}} x f_0(x)\,\mathrm{d}x$. Assuming that there exists a unique triple (p, μ, F) defining G in the previous equation, then we get

$$F(x) = \frac{1}{p}\left(G(x+\mu) - (1-p)F_0(x+\mu)\right), \quad \forall x \in \mathbb{R}. \tag{12}$$

As F is the cdf of a symmetric distribution with respect to 0, we have $F(x) = 1 - F(-x)$, for all $x \in \mathbb{R}$. We denote by p_0 and μ_0, respectively, the unknown values of p and μ. Defining for all $x \in \mathbb{R}$ the functions

$$H_1(x; \mu, m, G, F_0) = \frac{\mu}{m} G(x+\mu) + \frac{m-\mu}{m} F_0(x+\mu)$$

and

$$H_2(x; \mu, m, G, F_0) = 1 - \frac{\mu}{m} G(\mu - x) + \frac{\mu - m}{m} F_0(\mu - x),$$

where $m = p_0\mu_0$ is the first-order moment of G, we have, using (12) and the symmetry of F, $H_1(\cdot; \mu_0, m, G, F_0) = H_2(\cdot; \mu_0, m, G, F_0)$. Consequently, if d is a distance measure between two functions, we have $d(H_1(\cdot; \mu_0, m, G, F_0), H_2(\cdot; \mu_0, m, G, F_0)) = 0$.

Now, since G and m are unknown, it is natural to replace G and m by their estimators, i.e. $\hat{G}_n$ and $\hat{m}_n$, respectively, defined by

$$\hat{G}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \leq x) \quad \forall x \in \mathbb{R} \quad \text{and} \quad \hat{m}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

where $\mathbf{1}(\cdot)$ is the indicator function. Therefore, we get an empirical version d_n of d defined by

$$d_n(\mu) = d(H_1(\cdot; \mu, \hat{m}_n, \hat{G}_n, F_0), H_2(\cdot; \mu, \hat{m}_n, \hat{G}_n, F_0)), \quad \mu \in v, \tag{13}$$

where v is a compact subset of $\mathbb{R}^*$ on which model (1) is identifiable. It is natural to estimate the unknown parameters μ_0 and p_0 by

$$\hat{\mu}_n = \arg\min_{\mu \in v} d_n(\mu) \quad \text{and} \quad \hat{p}_n = \frac{\hat{m}_n}{\hat{\mu}_n}. \tag{14}$$

For d we can choose the $L^q(\mathbb{R})$-norm, defined for $1 \leq q < +\infty$ and $\mu \in v$ by

$$d(\mu) \equiv \|H_1 - H_2\|_q = \left(\int_{\mathbb{R}} |H_1(x; \mu, m, G, F_0) - H_2(x; \mu, m, G, F_0)|^q \,\mathrm{d}x\right)^{1/q}.$$

Remark 1. Replacing μ/m by $1/p$ in H_1 and H_2 we obtain a new contrast function $d(p, \mu; G)$ depending on G and on the unknown parameters p and μ. Replacing G by $\hat{G}_n$ we are led to an empirical contrast $d_n(p, \mu) = d(p, \mu; \hat{G}_n)$ whose minimizer $(\hat{p}_n, \hat{\mu}_n)$ is an estimator of (p, μ). This approach should be used when g does not have a first-order moment. Note also that when f is not exactly an even function, simulation results show robustness in estimating (p, μ) by using $d_n(p, \mu)$ instead of $d_n(\mu)$.

Using relation (12) we can estimate F by

$$\hat{F}_n(x) = \frac{\hat{\mu}_n}{\hat{m}_n}\hat{G}_n(x+\hat{\mu}_n) + \frac{\hat{m}_n - \hat{\mu}_n}{\hat{m}_n}F_0(x+\hat{\mu}_n), \quad \forall x \in \mathbb{R}. \tag{15}$$

Note that generally $\hat{F}_n$ will not be a legitimate cdf, as it is not non-decreasing in general. However, the Glivenko–Cantelli strong consistency result obtained in section 5 shows that this is not a serious drawback provided that the sample size is large enough.

Again by formula (7) a natural estimator of the pdf f is defined by

$$\tilde{f}_n(x) = \frac{1}{\hat{p}_n}\left(\hat{g}_n(x+\hat{\mu}_n) - (1-\hat{p}_n)f_0(x+\hat{\mu}_n)\right), \quad \forall x \in \mathbb{R},$$

where

$$\hat{g}_n(x) = \frac{1}{nb_n}\sum_{i=1}^{n} q\left(\frac{x-X_i}{b_n}\right), \quad \forall x \in \mathbb{R},$$

with $b_n \to 0$, $nb_n \to +\infty$ and q is a symmetric kernel pdf with finite second-order moment. For example, we may choose $q(x) = (1-|x|)\mathbf{1}(-1 \le x \le 1)$. Because $\tilde{f}_n$ is generally not a pdf it can be modified into the estimator $\hat{f}_n$ which is itself a pdf

$$\hat{f}_n = \frac{1}{s_n}\tilde{f}_n\mathbf{1}(\tilde{f}_n \ge 0), \tag{16}$$

where

$$s_n = \int_{\mathbb{R}} \tilde{f}_n(x)\mathbf{1}(\tilde{f}_n(x) \ge 0)\,\mathrm{d}x.$$

However, there are other ways to modify kernel estimators to make them non-negative (see Glad *et al.*, 2003).

4. Moments method for estimating the Euclidean parameters

Let $\hat{G}_n$ be the empirical cdf obtained from n i.i.d. random variables with common cdf G. Let us denote by μ_0, 0_0 and γ_0 the first three moments of f_0. Define $\tilde{g}(\cdot) = g(\cdot + \mu_0)$, where g is defined by (1). We then have

$$\tilde{g}(x) = (1-p)\tilde{f}_0(x) + pf(x-\tilde{\mu}), \quad \forall x \in \mathbb{R},$$

where $\tilde{f}_0(\cdot) = f_0(\cdot + \mu_0)$ and $\tilde{\mu} = \mu - \mu_0$. Now we write $\tilde{m}_i = \int_{\mathbb{R}} x^i \tilde{g}(x)\,\mathrm{d}x$ for $i = 1, 2, 3$, and we get

$$\begin{cases} p\tilde{\mu} = \tilde{m}_1, \\ (1-p)\tilde{\theta}_0 + p(\theta + \tilde{\mu}^2) = \tilde{m}_2, \\ (1-p)\tilde{\gamma}_0 + p(3\theta\tilde{\mu} + \tilde{\mu}^3) = \tilde{m}_3, \end{cases} \tag{17}$$

where θ is the second-order moment of f and $\tilde{\theta}_0$ and $\tilde{\gamma}_0$ are moments of order 2 and 3 of $\tilde{f}_0$. If $\tilde{m}_1 = 0$ and $\tilde{\gamma}_0 \ne 0$, then we have

$$\begin{cases} \mu = \mu_0, \\ p = \dfrac{\tilde{\gamma}_0 - \tilde{m}_3}{\tilde{\gamma}_0}, \end{cases}$$

whereas if $\tilde{m}_1 = 0$ and $\tilde{\gamma}_0 = 0$ there are infinitely many solutions. Otherwise, if $\tilde{m}_1 \ne 0$ we show that $\tilde{\mu}$ is a zero of the following polynomial

$$\tilde{\mu}^3 + \frac{3(\tilde{\theta}_0 - \tilde{m}_2)}{2\tilde{m}_1}\tilde{\mu}^2 + \frac{\tilde{m}_3 - \tilde{\gamma}_0 - 3\tilde{m}_1\tilde{\theta}_0}{2\tilde{m}_1}\tilde{\mu} + \frac{\tilde{\gamma}_0}{2} = 0.$$

Now, replacing unknown moments $\tilde{m}_k$ by their empirical counterpart $\tilde{m}_k^{(n)}$ defined for $k=1,2,3$ by

$$\tilde{m}_k^{(n)} = \int_{\mathbb{R}} x^k \, \mathrm{d}\hat{G}_n(x+\mu_0) = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)^k,$$

we solve the following random polynomial equation

$$\tilde{\mu}^3 + \frac{3(\tilde{\theta}_0 - \tilde{m}_2^{(n)})}{2\tilde{m}_1^{(n)}}\tilde{\mu}^2 + \frac{\tilde{m}_3^{(n)} - \tilde{\gamma}_0 - 3\tilde{m}_1^{(n)}\tilde{\theta}_0}{2\tilde{m}_1^{(n)}}\tilde{\mu} + \frac{\tilde{\gamma}_0}{2} = 0. \tag{18}$$

This leads to at most three solutions written $\tilde{\tilde{\mu}}_n^{(i)}$ which in turn lead to three possible estimators for μ. Let us write $\hat{\mu}_n^{(i)} = \mu_0 + \tilde{\tilde{\mu}}_n^{(i)}$ $(i=1,2,3)$ these three possible estimators of μ. We finally estimate μ by the value from among $\{\hat{\mu}_n^{(1)}, \hat{\mu}_n^{(2)}, \hat{\mu}_n^{(3)}\}$ that minimizes the empirical discrepancy measure d_n defined by (13).

Remark 2. Solving (18) can be done, e.g. by using the function *polyroot* of the statistical program *R*. Note that (18) can have two conjugate complex roots, because of errors in coefficients of the polynomial. In this case we have to take the real part of the roots. Typically, when $\tilde{\gamma}_0 = 0$ and the model is identifiable, the polynomial (18) reduces to a polynomial of degree 2, the discriminant of which can be null (if the moment equations give the identifiability). In this case the estimator of the discriminant can be positive or negative, and then we can obtain complex roots of (18). However this is not a serious drawback because the more precise the estimation of moments, the smaller (and more negligible) the estimated value of the discriminant.

Remark 3. It is worth noting that the moments method can provide an interesting initial guess value for minimizing the discrepancy measure d_n.

5. Consistency

We denote by (p_0, μ_0) the true value of the unknown Euclidean part (p, μ) of model (1) and by θ_0 and θ the moments of order 2 of f_0 and f respectively. Let us introduce the set

$$\Phi = \mathbb{R}^* \times (0, +\infty) \backslash \cup_{k \in \mathbb{N}^*} \Phi_k,$$

where

$$\Phi_k = \left\{ (\mu, \theta) \in \mathbb{R}^* \times (0, +\infty); \theta = \theta_0 + \frac{k \pm 2}{3k}\mu^2 \right\}.$$

We consider the following assumptions.

A1. $(f_0, f) \in \mathcal{F}_3^2$, $\hat{f}_0 > 0$ and $(\mu_0, \theta) \in \Phi_c \subset \Phi$, where Φ_c is a compact subset of Φ.

A2. (f_0, f) satisfies the identifiability condition of proposition 3 (i), and in addition f_0 satisfies the following tail condition:

$$\forall \alpha \in \mathbb{R}, \lim_{x \to +\infty} \frac{f_0(-x+\alpha)}{f_0(x)} = 0, \quad \text{or} \lim_{x \to +\infty} \frac{f_0(x)}{f_0(-x+\alpha)} = 0. \tag{19}$$

A3. (f_0, f) satisfies the identifiability condition of proposition 3 (ii).

The set $\mathcal{F}_3$ may include, for example, centred Gaussian distributions, centred Laplace distributions or Student's t-distributions (if the number of degrees of freedom is large enough). However, even if the centred Cauchy distributions have positive characteristic functions they

do not satisfy the third-order moment condition that is required. Consequently, many classical parametric distributions satisfy assumption A1. Moreover under A1, it is possible to have one heavy-tailed component distribution while the other is light-tailed. Conditions A2 and A3 are more specific, but they allow a non-symmetric known component to be included. For example (19) is satisfied if f_0 has exponential tails with different rates. This assumption can be also fulfilled if f_0 has polynomial tails with non-equal degrees. Assumption A3 is of interest because it includes some cases where f_0 has compact support, for example, when it is the pdf of a uniform distribution on a finite interval.

Let us consider $v=\xi(\Phi_c)$ under A1, and $v=\xi(\Phi_K)$ under A2 or A3, with $\xi(x,y)=x$ and Φ_K denotes any compact subset of $\mathbb{R}^*\times(0,+\infty)$. Therefore, we assume that estimators $\hat{\mu}_n$, $\hat{p}_n$, $\hat{F}_n$ and $\hat{f}_n$ are defined by (14)–(16) in section 3. We denote by $\|\cdot\|_\infty$ the supremum norm.

Theorem 1

Assume that one of assumptions A1–A3 is satisfied. As n tends to infinity we have:

(i) $(\hat{p}_n,\hat{\mu}_n)$ *converge almost surely to* (p_0,μ_0).

(ii) $\|\hat{F}_n-F\|_\infty$ *converges almost surely to 0.*

(iii) $\|\hat{f}_n-f\|_1$ *converges almost surely to 0 if* $b_n\to 0$, $nb_n\to+\infty$ *and if both* f_0 *and* f *belong to the Besov space*

$$\mathcal{B}^1_{1,\infty}=\left\{f\in\mathcal{B}(\mathbb{R}):\sup_{h\neq 0}\frac{1}{|h|}\int_{\mathbb{R}}|f(x+h)-f(x)|\,dx<\infty\right\},$$

where $\mathcal{B}(\mathbb{R})$ *denotes the class of Borel measurable functions defined on* $\mathbb{R}$.

(iv) Assume that q is a symmetric pdf with finite second-order moment satisfying the Geffroy properties (see Bosq & Lecoutre, 1987, p. 65):

(a) the set of discontinuities of q has null Lebesgue measure.

(b) $x\mapsto\sup\{|q(u)|;|u-x|<1\}$ *is integrable on* $\mathbb{R}$.

In addition we assume that $b_n\to 0$, $nb_n/\log n\to+\infty$, *and that both* f_0 *and* f *are uniformly continuous on* $\mathbb{R}$. *Then* $\|\hat{f}_n-f\|_\infty$ *converges almost surely to 0.*

Remark 4. The results of theorem 1 also hold if $\hat{p}_n$ and $\hat{\mu}_n$ are the estimators of remark 1.

Proof. (i) Let $\varepsilon>0$ be a real number. From lemma 4 there exists $\delta>0$ such that

$$\limsup_{n>0}\{|\hat{\mu}_n-\mu_0|>\varepsilon\}\subseteq\limsup_{n>0}\{d(\hat{\mu}_n)>\delta\}.$$

By lemma 3, the last set has probability zero. The almost sure convergence of $\hat{\mu}_n$ follows, which implies the almost sure convergence of $\hat{p}_n$ as $\hat{p}_n=\hat{m}_n/\hat{\mu}_n$.

(ii) Let us consider, for all $x\in\mathbb{R}$, the inequality:

$$|\hat{F}_n(x)-F(x)|\leq T_1(x)+T_2(x),$$

where

$$T_1(x)=\left|\frac{\hat{\mu}_n}{\hat{m}_n}\hat{G}_n(x+\hat{\mu}_n)-\frac{\mu_0}{m}G(x+\mu_0)\right|,$$

$$T_2(x)=\left|\frac{\hat{m}_n-\hat{\mu}_n}{\hat{m}_n}F_0(x+\hat{\mu}_n)-\frac{m-\mu_0}{m}F_0(x+\mu_0)\right|.$$

For the treatment of T_1 let us remark that

$$T_1(x) \leq \left| \frac{\hat{\mu}_n}{\hat{m}_n} \hat{G}_n(x+\hat{\mu}_n) - \frac{\hat{\mu}_n}{\hat{m}_n} G(x+\hat{\mu}_n) \right| + \left| \frac{\hat{\mu}_n}{\hat{m}_n} G(x+\hat{\mu}_n) - \frac{\mu_0}{m} G(x+\mu_0) \right|.$$

According to the Glivenko–Cantelli theorem the first term of the right-hand side converges almost surely and uniformly in x to 0 as n tends to infinity. Remarking that the second term of the right-hand side of the above inequality is very similar to T_2, we consider the following inequality

$$\left| \frac{\hat{\mu}_n}{\hat{m}_n} G(x+\hat{\mu}_n) - \frac{\mu_0}{m} G(x+\mu_0) \right| \leq \left| \frac{\hat{\mu}_n}{\hat{m}_n} - \frac{\mu_0}{m} \right| + \left| \frac{\mu_0}{m} \right| |G(x+\hat{\mu}_n) - G(x-\mu_0)|.$$

The left-hand side of the above inequality tends clearly to 0 as n tends to infinity since, as was shown in (i), $\hat{\mu}_n \xrightarrow{\text{a.s.}} \mu_0$ and $\hat{m}_n \xrightarrow{\text{a.s.}} m$, and since G is uniformly continuous. We thus obtain that $\|T_1\|_\infty \xrightarrow{\text{a.s.}} 0$ and $\|T_2\|_\infty \xrightarrow{\text{a.s.}} 0$ (this result is straightforward in virtue of the analogy between T_2 and the last term we discussed), which concludes the proof for (ii).

(iii) Let us notice that

$$\|\tilde{f}_n - f\|_1 \leq \left\| \frac{\hat{\mu}_n}{\hat{m}_n} \hat{g}_n(\cdot - \hat{\mu}_n) - \frac{\mu_0}{m} g(\cdot - \mu) \right\|_1 + \left\| \frac{\hat{m}_n - \hat{\mu}_n}{\hat{m}_n} f_0(\cdot - \hat{\mu}_n) - \frac{m - \mu_0}{m} f_0(\cdot - \mu) \right\|_1.$$

For simplicity we restrict ourselves to the first term on the right-hand side (a similar but simpler proof holds for the second term):

$$\begin{aligned}
&\left\| \frac{\hat{\mu}_n}{\hat{m}_n} \hat{g}_n(\cdot - \hat{\mu}_n) - \frac{\mu_0}{m} g(\cdot - \mu) \right\|_1 \\
&\quad \leq \left\| \frac{\hat{\mu}_n}{\hat{m}_n} \hat{g}_n(\cdot - \hat{\mu}_n) - \frac{\hat{\mu}_n}{\hat{m}_n} g(\cdot - \hat{\mu}_n) \right\|_1 + \left\| \frac{\hat{\mu}_n}{\hat{m}_n} g(\cdot - \hat{\mu}_n) - \frac{\mu_0}{m} g(\cdot - \hat{\mu}_n) \right\|_1 \\
&\qquad + \left\| \frac{\mu_0}{m} g(\cdot - \hat{\mu}_n) - \frac{\mu_0}{m} g(\cdot - \mu_0) \right\|_1 \\
&\quad \leq \left| \frac{\hat{\mu}_n}{\hat{m}_n} \right| \|\hat{g}_n - g\|_1 + \left| \frac{\hat{\mu}_n}{\hat{m}_n} - \frac{\mu_0}{m} \right| + \left| \frac{\mu_0}{m} \right| \|g(\cdot - \hat{\mu}_n) - g(\cdot - \mu_0)\|_1.
\end{aligned}$$

From (i) and the Devroye (1983) L^1-consistency result, which establishes that $\|\hat{g}_n - g\|_1 \xrightarrow{\text{a.s.}} 0$, we obtain the almost sure convergence to 0 of the two first terms on the right-hand side of the previous inequality. For the third term we use the fact that g belongs to the Besov space $\mathcal{B}^1_{1,\infty}$, which implies that

$$\begin{aligned}
&\|g(\cdot + \hat{\mu}_n) - g(\cdot + \mu_0)\|_1 \leq |\hat{\mu}_n - \mu_0| \\
&\times \left((1-p_0) \sup_{h \neq 0} \frac{1}{|h|} \int_{\mathbb{R}} |f_0(x+h) - f_0(x)| \, \mathrm{d}x + p_0 \sup_{h \neq 0} \frac{1}{|h|} \int_{\mathbb{R}} |f(x+h) - f(x)| \, \mathrm{d}x \right), \qquad (20)
\end{aligned}$$

and proves that $\|\tilde{f}_n - f\|_1 \xrightarrow{\text{a.s.}} 0$ from (i). In addition, it is straightforward to show that $\|\hat{f}_n - f\|_1 \leq \|\tilde{f}_n - f\|_1$, and then we have the almost sure convergence of $\|\hat{f}_n - f\|_1$ to 0. Moreover, under the same assumptions and with $s_n = \int_{\mathbb{R}} \tilde{f}_n(x) \, \mathrm{d}x$, we have

$$|s_n - 1| = \left| \int_{\mathbb{R}} (\tilde{f}_n(x) - f(x)) \, \mathrm{d}x \right| \leq \|\tilde{f}_n - f\|_1 \to 0, \quad \text{a.s.}$$

Therefore, $\hat{f}_n = s_n^{-1} \tilde{f}_n \mathbf{1}(\tilde{f}_n > 0)$ are density functions that satisfy $\|\hat{f}_n - f\|_1 \xrightarrow{\text{a.s.}} 0$.

(iv) Regarding this last point, it is sufficient to remark that as q is a Geffroy kernel, then $\|\hat{g}_n - g\|_\infty \xrightarrow{\text{a.s.}} 0$, under the conditions specified in (iv) (see Bosq & Lecoutre, 1987, p. 65). In fact, using the analogy of the triangular inequalities of the previous proof in supremum norm, the required result holds by replacing the argument in (20) by a uniform continuity argument.

6. Applications

6.1. Microarray data analysis

Microarrays are a technique for revealing the simultaneous expression levels of a large number of genes in a biological sample. More precisely, a large number (up to several thousands) of gene probes, consisting of cDNA, are localized on a membrane. The gene targets, in the form of a solution of mRNA, are then extracted from a biological tissue and hybridized on this support. The expression level of each gene in the biological tissue is given by the concentration of mRNA hybridized on each probe. The experiments are repeated to assess the experimental variability and conducted under different conditions (different processes, stages, tissues, etc.). The conditions are compared in order to detect differences in expression.

We denote by R the number of repetitions, n the number of genes and J the number of conditions. The data are the random variables (A_{ijr}) corresponding to the rth repetition of the expression level of gene i in condition j. We divide A_{ijr} by the sum of all the expression levels on the membrane to obtain the concentrations:

$$P_{ijr} = \frac{A_{ijr}}{\sum_{i=1}^{n} A_{ijr}}.$$

We consider the following transformation of the P_{ijr}s:

$$X_{ijr} = \ln\left(\frac{P_{ijr}}{1 - P_{ijr}}\right).$$

We write

$$X_{ij\cdot} = \frac{1}{R}\sum_{r=1}^{R} X_{ijr}, \quad X_{i\cdot\cdot} = \frac{1}{JR}\sum_{j=1}^{J}\sum_{r=1}^{R} X_{ijr}$$

and we assume that the data are of good quality and have been correctly normalized to ensure that experimental biases have been removed. We assume that for r in $\{1,\ldots,R\}$, X_{ijr} is normally distributed with mean m_{ij} and variance σ_{ij}^2. Therefore, for i in $\{1,\ldots,n\}$, the null hypothesis $\mathcal{H}_{0,i}$:

'There is no expression difference between the J conditions for gene i', is equivalent to

$$\left\{ m_{ij} = m_{i\cdot}, \sigma_{ij} = \sigma_{i\cdot}, \forall j = 1,\ldots,J \quad \text{where } m_{i\cdot} = \frac{1}{J}\sum_{j=1}^{J} m_{ij}, \quad \text{and } \sigma_{i\cdot} = \frac{1}{J}\sum_{j=1}^{J} \sigma_{ij} \right\}.$$

In order to compare two conditions ($J = 2$) we can use, for the ith gene, the test statistic S_i defined by:

$$S_i = \frac{X_{i1\cdot} - X_{i2\cdot}}{\sqrt{\frac{\sum_{r=1}^{R}(X_{i1r} - X_{i1\cdot})^2 + \sum_{r=1}^{R}(X_{i2r} - X_{i2\cdot})^2}{R(R-1)}}}. \tag{21}$$

For each i in $\{1,\ldots,n\}$ under the null hypothesis $\mathcal{H}_{0,i} = \{m_{i1} = m_{i2}, \sigma_{i1} = \sigma_{i2}\}$ (that there is no expression difference between the two conditions for gene i) the statistic S_i is Student's distributed with $2R - 2$ degrees of freedom. Generally speaking, when we compare J conditions, we can use for the ith gene the test statistic S_i:

$$S_i = \frac{RJ(R-1)}{(J-1)} \times \frac{\sum_{j=1}^{J}(X_{ij\cdot} - X_{i\cdot\cdot})^2}{\sum_{j=1}^{J}\sum_{r=1}^{R}(X_{ijr} - X_{ij\cdot})^2}.$$

Under the null hypothesis $\mathcal{H}_{0,i} = \{m_{ij} = m_{i\cdot}, \sigma_{ij} = \sigma_{i\cdot}, \forall j = 1,\ldots,J\}$ that there is no expression difference between the J conditions for gene i, the test statistic S_i is Fisher distributed with $(J-1, JR-J)$ degrees of freedom.

Under the alternative hypothesis $\mathcal{H}_{1,i} = \{\exists j \in \{1,\ldots,J\} : m_{ij} \neq m_{i.} \text{ or } \sigma_{ij} \neq \sigma_{i.}\}$ that there is at least one expression difference between the J conditions, the distribution of S_i is unknown. Therefore, the S_is distribution can be modelled by:

$$g(x) = (1-p)f_0(x) + pf(x), \tag{22}$$

where p is the proportion of non-null statistics, f_0 is the null pdf of S_i (Student's or Fisher) and f is the unknown non-null pdf.

The estimation of the unknown parameter p and the pdf f enables us to estimate the probability $\alpha^{(i)}$ that gene i is differentially expressed given $\{S_i = s_i\}$:

$$\alpha^{(i)} = P(\text{gene } i \text{ is differentially expressed } | S_i = s_i) = \frac{pf(s_i)}{(1-p)f_0(s_i) + pf(s_i)}.$$

Under the hypothesis that f is a symmetric pdf on $\mathbb{R}$, model (22) reduces to model (1) and we can estimate p and μ by symmetrization or by the moments method as indicated in sections 3 and 4 respectively. Then we define natural consistent estimators of f and $\alpha^{(i)}$ given $\{S_i = s_i\}$ by:

$$\hat{f}_n(x) = \frac{\tilde{f}_n(x)}{\int_{\mathbb{R}} \tilde{f}_n(y)\mathbb{1}(\tilde{f}_n(y) \geq 0)\,\mathrm{d}y},$$

$$\hat{\alpha}_n^{(i)} = \frac{\hat{p}_n \hat{f}_n(s_i - \hat{\mu}_n)}{(1-\hat{p}_n)f_0(s_i) + \hat{p}_n \hat{f}_n(s_i - \hat{\mu}_n)},$$

where

$$\tilde{f}_n(x) = \frac{1}{\hat{p}_n}(\hat{g}_n(x + \hat{\mu}_n) - (1-\hat{p}_n)f_0(x + \hat{\mu}_n)).$$

Let us remark that the strong consistency of the $\hat{\alpha}_n^{(i)}$s to the $\alpha^{(i)}$s, is insured by theorem 1. As a consequence, a heuristic means of identifying differentially expressed genes consists in selecting genes i for which the $\hat{\alpha}_n^{(i)}$s are among the $[n\hat{p}_n]$ greatest values of $\{\hat{\alpha}_n^{(j)}; j = 1,\ldots,n\}$. Identification of differentially expressed genes may also be carried out using standard classification procedures (e.g. see Benjamini & Hochberg, 1995).

6.2. Mixture model with training data

We are still considering model (1), in which both f_0 and f, together with parameters p and μ, are unknown but for which training data are available for the first component f_0, that is, we still have an n-sample from g and, in addition, an n'-sample from f_0 is given. In classical finite mixture models involving training data, samples from each component are given and then the inference reduces to estimating the mixture proportions (see Titterington, 1983). Following the methodology of section 3 and replacing the unknown cdf F_0 by the empirical cdf $\hat{F}_{0,n'}$ obtained from the training sample, we are able to propose an estimation function similar to (13), defined by

$$d_{n,n'}(\mu) = d(H_1(\cdot;\mu,\hat{m}_n,\hat{G}_n,\hat{F}_{0,n'}), H_2(\cdot;\mu,\hat{m}_n,\hat{G}_n,\hat{F}_{0,n'})), \quad \mu \in \nu, \tag{23}$$

where ν is a compact subset of $\mathbb{R}^*$ on which model (1) is identifiable. Therefore, it is natural to estimate the unknown parameter μ_0 by

$$\hat{\mu}_{n,n'} = \arg\min_{\mu \in \nu} d_{n,n'}(\mu).$$

Therefore, as in section 3, we can derive natural estimators of the proportion p and the unknown pdf f. On the one hand, we have $p = (m_0 - m)/(m_0 - \mu)$ where m and m_0, respectively, are the expectation of g and f_0 (m can be estimated from the g-sample and m_0 can be

estimated from the f_0-sample). On the other hand, as F can be explicitly expressed as a function of G, F_0, p and μ, it can be estimated by plugging estimators of these four quantities into formula (12). Therefore, smoothing this estimator we can estimate the pdf f as in section 3.

Note also that the moments method of section 4 can be used by simply replacing the unknown quantities μ_0, θ_0 and γ_0 by their estimators obtained from the f_0-sample.

Finally, consistency results analogous to those of theorem 1 can be established by assuming that $\min(n, n')$ tends to infinity.

7. Simulations and example

7.1. Simulations

In this section, we simulate K samples of n i.i.d. random variables whose the common distribution is given by the following two-component mixture model:

$$(1-p)\mathcal{N}(0,4)+p\mathcal{N}(\mu,1), \tag{24}$$

where $\mathcal{N}(\alpha,\beta)$ denotes the Gaussian distribution with mean α and variance equal to β. For each sample we estimate (p,μ), given that the known component is $\mathcal{N}(0,4)$-distributed. Finally, for different values of n and p, we provide the mean and the standard deviation of the estimates obtained both by the symmetrization method and by the parametric maximum likelihood method.

From a semiparametric point of view the estimator of μ is given by (14) where we choose the L^2-norm for d; thus, using the first moment equation, we derive an estimator of p (see section 3). Note that the computation of $d_n(\mu)$ requires an integration step that is performed numerically. Because numerical estimation of the derivatives of $d_n(\mu)$ are quite unstable we do not look for the minimum argument $\hat{\mu}_n$ of d_n by using a standard optimization routine, but we simply look for the minimizer of d_n over a parameter space discretization.

We should mention a weak point when using $d_n(\mu)$. Where f_0 is an even function it is easy to check that $\mu=0$ is a (non-admissible) zero of both d and d_n. Therefore, if the model is identifiable d has two roots (0 and μ) corresponding to two minima. Thus, the approximate d_n should have two minima too. But in practice, if the first moment m is not well estimated (typically when both n and p are small) it may happen that the only minimum of d_n is the non-admissible value 0, and in any case we have $d_n(\mu)\geq d_n(0)=0$. In this case we recommend estimating (p,μ) using the two-parameter function d_n given in remark 1. As we have mentioned, when both p and n are small the first moment equation can lead to a constraint that is not well satisfied by the data. This fact is illustrated by Fig. 2: in Fig. 2A d_n has only one minimum whereas in Fig. 2B d_n has two minima.

We can see from Tables 1 and 2 that for the standard deviation criterion the parametric estimates outperform the semiparametric estimates. Although for the smallest sample size performances of semiparametric and parametric estimators are quite close, it should be noted that in the semiparametric symmetrization method, samples for which the empirical contrast function was monotonous (as in Fig. 2A) were rejected (for $n=250$ about 10% and 20% for $p=0.3$ and $p=0.15$ respectively; very few cases for $n=1000$). This drawback does not occur in the parametric setup.

7.2. Real data: bovine gestation mode comparison

We examine data used to detect genes that are statistically differentially expressed in bovine trophoblast between artificial insemination (AI) and *in vitro* fertilization (IVF) gestation

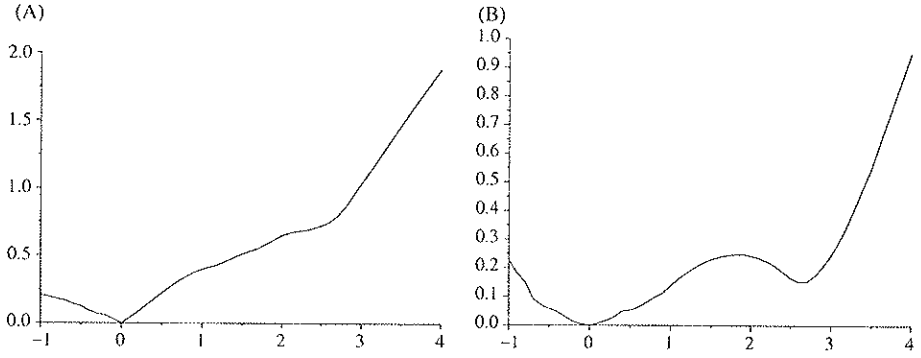

Fig. 2. Two examples of the behaviour of the empirical contrast function d_n. (A) An example where μ cannot be estimated from $d_n(\mu)$. (B) An example where μ is estimated from $d_n(\mu)$.

Table 1. *Mean (SD) of 200 semiparametric estimates of $p|\mu$ (obtained by the symmetrization method)*

$K=200$	$n=250$	$n=1000$
$p=0.3/\mu=3$	0.303 (0.057)/2.963 (0.226)	0.301 (0.029)/2.976 (0.131)
$p=0.15/\mu=3$	0.165 (0.055)/2.878 (0.418)	0.154 (0.031)/2.944 (0.272)

Table 2. *Mean (SD) of 200 parametric estimates of $p|\mu$ (obtained by the maximum likelihood method)*

$K=200$	$n=250$	$n=1000$
$p=0.3/\mu=3$	0.293 (0.045)/2.989 (0.058)	0.300 (0.022)/2.991 (0.041)
$p=0.15/\mu=3$	0.156 (0.051)/2.993 (0.101)	0.152 (0.026)/2.990 (0.051)

modes. AI mode is the reference gestation mode in animal sciences. This statistical analysis helps the biologist in understanding the biological differences between the two gestation modes and in improving IVF techniques to reduce the mortality rate associated with this gestation mode. Ten microarrays were obtained, each with $n=10,214$ genes, for each condition (AI and IVF). Let A_{ijr} denote the mean intensity of the signal for the rth repetition of gene i in condition j, where, using the notation of section 6.1, we have $(i,j,r)\in\{1,\ldots,n\}\times\{1,\ldots,J\}\times\{1,\ldots,R\}$ with $J=2$ and $R=10$. Each S_i is therefore computed using formula (21) and, under the null hypothesis, it follows a Student's distribution with 18 degrees of freedom (denoted by T_{18}).

We assume that the S_is are i.i.d. with common distribution defined by (1) where f_0 is the pdf of a T_{18}, and p, μ and f are unknown. We estimate the unknown Euclidean parameters p and μ using the two-parameter contrast function defined in remark 1, rather than the method involving the first moment equation, which did not prove to be suitable (the contrast function with only one parameter appears to be more sensitive to the f symmetry).

We obtain $\hat{p}=0.037$ and $\hat{\mu}=1.05$ by discretizing the Euclidean parameter space $([0,1]\times[\min S_i, \max S_i])$ and taking for $(\hat{p},\hat{\mu})$ the value that makes the empirical contrast d_n minimum on the discretized space. The graph of the empirical contrast function d_n is given in Fig. 3 (with p and μ restrained to $[0.01,0.1]\times[0.5,1.5]$) where d_n reaches its minimum value 0.2257 at $(\hat{p},\hat{\mu})$.

Figure 4A,B show, respectively, the reconstruction of the mixture density g [defined by model (1)] and the estimate of the unknown density f (not symmetrized). Even though the estimator of f is not really an even function, it nevertheless reveals the deviations of g with respect to the T_{18} pdf, thus ensuring a good reconstruction of g.

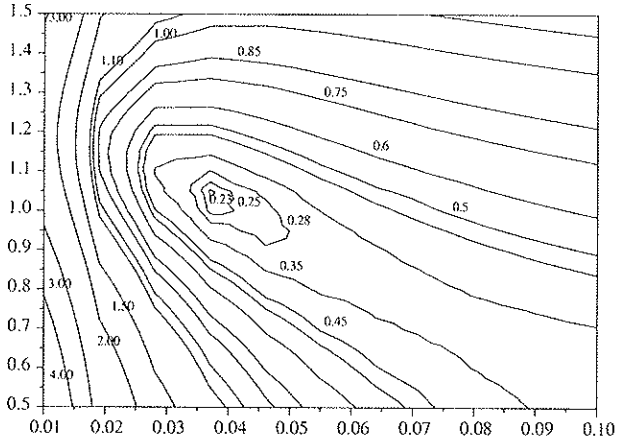

Fig. 3. Level curves of $(p, \mu) \mapsto d_n(p, \mu)$ for the real data set (with $(p, \mu) \in [0.01, 0.1] \times [0.5, 1.5]$).

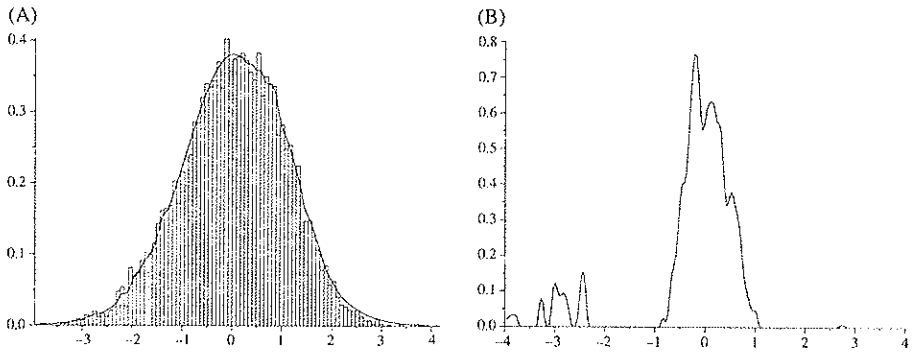

Fig. 4. Reconstruction of the mixture distribution using estimate of (p, μ, f) and estimator of the density f. (A) Histogram of the real data set compared with $(1-\hat{p})f_0(\cdot)+\hat{p}\hat{f}(\cdot-\hat{\mu})$ for $\hat{p}=0.037$ and $\hat{\mu}=1.05$. (B) Estimate of the unknown density f.

Finally, the above identification of model parameters and the heuristic classification method of section 6.1, allows around 370 genes to be detected as possibly differentially expressed.

8. Discussion and concluding remarks

We introduced a new semiparametric finite mixture model that completes the recent semiparametric finite mixture models introduced by Hall & Zhou (2003), Bordes *et al.* (2006) and Hunter *et al.* (2004). We studied the identifiability of our model but we observed that even if one component is completely specified, identifiability is not guaranteed in general. We proposed two types of estimator for the Euclidean part of the model. One is a minimum contrast estimator, while the other is based on the moments method. These two methods rely heavily on the fact that the pdf of the unknown component is symmetric. In our opinion a challenging problem would be to consider model (1) without the symmetry assumption on the unknown component. We obtained the consistency of our estimators for several classes of identifiable models. Convergence rates and the efficiency of our estimators remain open

problems (very little is known about these aspects: see Hall & Zhou, 2003; Bordes *et al.*, 2006).

We indicated two fields of application for our model: first, microarray data analysis [which was the initial motivation for the introduction of model (1)] (see e.g. McLachlan *et al.*, 2004; and Dudoit *et al.*, 2002); and, secondly, finite mixture models with training data (where our approach provides more flexibility in the sense that it is not necessary to have training data from each component of the model) (see e.g. Murray & Titterington, 1978; Hall, 1981; Titterington, 1983; Qin, 1998, 1999).

Another important issue will be to provide efficient algorithms to estimate both the Euclidean and the functional parts of these kinds of semiparametric models. L. Bordes, D. Chauveau and P. Vandekerkhove develop a promising approach based on expectation maximization type algorithms.

References

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289–300.

Bordes, L., Mottelet, S. & Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model. *Ann. Statist.* **34** (in press).

Bosq, D. & Lecoutre, J.-P. (1987). *Théorie de l'estimation fonctionnelle.* Economica, Paris.

Cruz-Medina, I. R. & Hettmansperger, T. P. (2004). Nonparametric estimation in semiparametric univariate mixture models. *J. Statist. Comp. Sim.* **74**, 513–524.

Devroye, L. (1983). The equivalence of weak, strong and complete convergence in L_1 for kernel density estimates. *Ann. Statist.* **11**, 896–904.

Dudoit, S., Yang, Y. H., Callow, M. J. & Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statist. Sin.* **12**, 111–139.

Glad, I. K., Hjort, N. L. & Ushakov, N. G. (2003). Correction of density estimators that are not densities. *Scand. J. Statist.* **30**, 415–427.

Hall, P. (1981). On the nonparametric estimation of mixture proportions. *J. Roy. Statist. Soc. Ser. B* **43**, 147–156.

Hall, P. & Zhou, X. H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.* **31**, 201–224.

Hunter, D. R., Wang, S. & Hettmansperger, T. P. (2004). Inference for mixtures of symmetric distributions. *Technical Report* 04-01, Penn State University, Philadelphia, PA.

McLachlan, G. J. & Peel, D. (2000). *Finite mixture models.* Wiley, New York.

McLachlan, G. J., Do, K. A. & Ambroise, C. (2004). *Analyzing microarray gene expression data.* Wiley, Hobokin, NJ.

Murray, G. D. & Titterington, D. M. (1978). Estimation problems with data from a mixture. *Appl. Statist.* **27**, 325–334.

Qin, J. (1998). Semiparametric likelihood based method for goodness of fit tests and estimation in upgraded mixture models. *Scand. J. Statist.* **25**, 681–691.

Qin, J. (1999). Empirical likelihood ratio based confidence intervals for mixture proportions. *Ann. Statist.* **27**, 1368–1384.

Titterington, D. M. (1983). Minimum-distance non-parametric estimation of mixture proportions. *J. Roy. Statist. Soc. Ser. B* **45**, 37–46.

Titterington, D. M., Smith, A. F. M. & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions.* Wiley, Chichester.

Received June 2005, in final form February 2006

P. Vandekerkhove, Université de Marne-la-Vallée, Citée Descartes - 5 Bd Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée Cedex 2, France.
E-mail: pierre.vandek@univ-mlv.fr

Appendix

A. Inverting the moment equations

Proof of (6). Let us consider the system of moments equations

$$\begin{cases} p\mu = p_1\mu_1 & \text{(a)} \\ (1-p)\theta_0 + p(\mu^2+\theta) = (1-p_1)\theta_0 + p_1(\mu_1^2+\theta_1) & \text{(b)} \\ p(3\mu\theta+\mu^3) = p_1(3\mu_1\theta_1+\mu_1^3). & \text{(c)} \end{cases}$$

From relations (b) and (a) we obtain:

$$\begin{aligned} p_1\theta_1 &= (p_1-p)\theta_0 + p(\theta+\mu^2) - p_1\mu_1^2 \\ &= (p_1-p)\theta_0 + p(\theta+\mu^2) - p\mu\mu_1. \end{aligned} \quad \text{(d)}$$

From (c) and (d) we write

$$\begin{aligned} & 3\mu_1 p_1\theta_1 = 3p\mu\theta + p\mu^3 - p\mu\mu_1^2 \\ \Longleftrightarrow\ & 3p\mu(\theta_0-\theta) - p\mu_3 + \mu_1[3p(\theta+\mu_2-\theta_0)] - 2p\mu\mu_1^2 = 0. \end{aligned} \quad \text{(e)}$$

Equation (e) gives us a polynomial of degree two (in μ_1) which admits $\mu_1=\mu$ as a trivial zero, hence (e) is equivalent to $(\mu_1-\mu)(a\mu_1+b)=0$, with

$$a=-2p\mu \quad \text{and} \quad b=p\mu^2-3p(\theta_0-\theta)=p(\mu^2-3\theta_0+3\theta).$$

Hence the second zero of (e) is

$$\mu_1=\frac{\mu^2-3\theta_0+3\theta}{2\mu}=\mu+\frac{+3\theta-3\theta_0-\mu^2}{2\mu},$$

which concludes the proof.

B. Technical results

Lemma 1

Let H be a cdf such that $\int_{\mathbb{R}}|x|\,\mathrm{d}H(x)<+\infty$. Then, for all $(\alpha,\beta)\in\mathbb{R}^2$ we have:

$$\int_{\mathbb{R}}|H(x+\alpha)-H(x+\beta)|\,\mathrm{d}x=|\alpha-\beta|.$$

Proof. Obviously, it is sufficient to prove the result for $\alpha\geq 0$ and $\beta=0$. Because

$$\int_{-\infty}^{0}H(x)\,\mathrm{d}x+\int_{0}^{+\infty}(1-H(x))\,\mathrm{d}x=\int_{\mathbb{R}}|x|\,\mathrm{d}H(x)<+\infty,$$

we have

$$\begin{aligned} &\int_{\mathbb{R}}|H(x+\alpha)-H(x)|\,\mathrm{d}x=\int_{\mathbb{R}}(H(x+\alpha)-H(x))\,\mathrm{d}x \\ &=\int_{-\alpha}^{0}H(x)\,\mathrm{d}x+\int_{-\alpha}^{+\alpha}H(x)\,\mathrm{d}x+\int_{\alpha}^{2\alpha}\bar{H}(x)\,\mathrm{d}x=\alpha, \end{aligned}$$

where $\bar{H}(x)=1-H(x)$.

Lemma 2

Let H_i $(i=1,2)$ be the cdf of two distributions having first-order moments. Then for all $(\alpha,\beta)\in\mathbb{R}^2$ we have

$$\int_{\mathbb{R}} |H_1(x+\alpha) - H_2(x+\beta)|\,\mathrm{d}x \le m_1 + m_2 + |\alpha-\beta|,$$

where $m_i = \int_{\mathbb{R}} |x|\,\mathrm{d}H_i(x)$ for $i=1,2$.

Proof. We can consider without loss of generality that $\beta=0$. Then we have

$$\begin{aligned}\int_{\mathbb{R}} |H_1(x+\alpha) - H_2(x)|\,\mathrm{d}x &\le \int_{\mathbb{R}} |H_1(x+\alpha) - H_1(x)|\,\mathrm{d}x + \int_{\mathbb{R}} |H_1(x) - H_2(x)|\,\mathrm{d}x \\ &= |\alpha| + \int_{\mathbb{R}} |H_1(x) - H_2(x)|\,\mathrm{d}x,\end{aligned}$$

by lemma 1. Therefore, we have

$$\begin{aligned}&\int_{\mathbb{R}} |H_1(x) - H_2(x)|\,\mathrm{d}x \\ &\le \int_{-\infty}^{0} H_1(x)\,\mathrm{d}x + \int_{-\infty}^{0} H_2(x)\,\mathrm{d}x + \int_{0}^{+\infty} |1 - H_1(x) - (1 - H_2(x))|\,\mathrm{d}x \\ &\le \int_{-\infty}^{0} H_1(x)\,\mathrm{d}x + \int_{-\infty}^{0} H_2(x)\,\mathrm{d}x + \int_{0}^{+\infty} (1 - H_1(x))\,\mathrm{d}x + \int_{0}^{+\infty} (1 - H_2(x))\,\mathrm{d}x \\ &= m_1 + m_2.\end{aligned}$$

This concludes the proof.

Lemma 3

Assume that both F and F_0 have first-order moment. Then, as $n\to+\infty$, we have $d(\hat{\mu}_n)\to 0$ a.s.

Proof. First remark that

$$d^q(\hat{\mu}_n) \le d^q(\hat{\mu}_n) - d_n^q(\hat{\mu}_n) + d_n^q(\mu_0) - d^q(\mu_0) \le 2\sup_{\mu\in\nu} |d^q(\mu) - d_n^q(\mu)|, \tag{25}$$

because $d_n^q(\hat{\mu}_n) \le d_n^q(\mu_0)$ and $d^q(\mu_0)=0$. To simplify our notations we write $H_i(x) = H_i(x;\mu,m,G,F_0)$ and $\hat{H}_i(x) = H_i(x;\mu,\hat{m}_n,\hat{G}_n,F_0)$ for $i=1,2$. Let $\hat{c}_n$ be defined by

$$\hat{c}_n = \max\left\{\|H_1\|_\infty + \|H_2\|_\infty, \|\hat{H}_1\|_\infty + \|\hat{H}_2\|_\infty\right\} \le \frac{c_1}{\min(|m|,|\hat{m}_n|)} + 2,$$

where c_1 is finite and does not depend on μ. Because $||a|^q - |b|^q| \le q|a-b|$ for $|a|\le 1$ and $|b|\le 1$, we have:

$$\begin{aligned}|d^q(\mu) - d_n^q(\mu)| &\le \left| \int_{\mathbb{R}} \left(|H_1(x) - H_2(x)|^q - |\hat{H}_1(x) - \hat{H}_2(x)|^q \right) \mathrm{d}x \right| \\ &\le q\hat{c}_n^{q-1} \int_{\mathbb{R}} |H_1(x) - H_2(x) - \hat{H}_1(x) + \hat{H}_2(x)|\,\mathrm{d}x. \end{aligned} \tag{26}$$

Straightforward calculations lead to

$$\int_{\mathbb{R}} |H_1(x) - H_2(x) - \hat{H}_1(x) + \hat{H}_2(x)|\,\mathrm{d}x \le \frac{2\mu}{\hat{m}_n}\|\hat{G}_n - G\|_1 + \frac{\mu|\hat{m}_n - m|}{\hat{m}_n m} I(\mu), \tag{27}$$

where

$$I(\mu)=\int_{\mathbb{R}}|G(x+\mu)+G(\mu-x)-F_0(x+\mu)-F_0(\mu-x)|\,\mathrm{d}x.$$

Using the fact that

$$G(x)=(1-p_0)F_0(x)+p_0F(x-\mu_0),\quad \forall x\in\mathbb{R},$$

and that both F_0 and F have first-order moment, we show that

$$I(\mu)\le 2p_0\int_{\mathbb{R}}|F(x-\mu_0)-F_0(x)|\,\mathrm{d}x\le c_2,$$

where c_2 arises from lemma 2. Finally, this last result, with inequalities (26) and (27), gives:

$$|d^q(\mu)-d_n^q(\mu)|\le\hat{K}_n^{(1)}\|\hat{G}_n-G\|_1+\hat{K}_n^{(2)}|\hat{m}_n-m|,$$

where $\hat{K}_n^{(1)}$ and $\hat{K}_n^{(2)}$ do not depend on μ and converge almost surely to finite constants as n tends to infinity, $|\hat{m}_n-m|$ converges almost surely to 0 by the strong law of large numbers and, from Hunter *et al.* (2004), $\|\hat{G}_n-G\|_1$ converges almost surely to 0. We conclude the proof using (25).

Lemma 4

Assume that one of assumptions A1–A3 is satisfied. Then for all $\varepsilon>0$, there exists $\delta_\varepsilon>0$, such that:

$$\forall\mu\in v,\quad |\mu-\mu_0|>\varepsilon\Rightarrow d(\mu)>\delta_\varepsilon.$$

Proof. Step 1. Let us show that $\mu\mapsto d(\mu)$ is continuous on v. Using the beginning of the proof of lemma 3 we show that there exists a finite constant c such that

$$\begin{aligned}
|d^q(\mu)-d^q(\mu')|\le c\int_{\mathbb{R}}\Big|&\frac{\mu}{m}G(x+\mu)-\frac{\mu'}{m}G(x+\mu')+\left(1-\frac{\mu}{m}\right)F_0(x+\mu)-\left(1-\frac{\mu'}{m}\right)F_0(x+\mu')\\
&+\frac{\mu}{m}G(\mu-x)-\frac{\mu'}{m}G(\mu'-x)+\left(1-\frac{\mu'}{m}\right)F_0(\mu'-x)-\left(1-\frac{\mu}{m}\right)F_0(\mu-x)\Big|\,\mathrm{d}x\\
\le c\int_{\mathbb{R}}\Big|&\frac{\mu}{m}(G(x+\mu)-G(x+\mu'))+\frac{\mu-\mu'}{m}G(x+\mu)\\
&+\frac{m-\mu}{m}(F_0(x+\mu)-F_0(x+\mu'))+\frac{\mu'-\mu}{m}F_0(x+\mu')\\
&+\frac{\mu}{m}(G(\mu-x)-G(\mu'-x))+\frac{\mu-\mu'}{m}G(\mu'-x)\\
&+\frac{\mu'-m}{m}(F_0(\mu'-x)-F_0(\mu-x))+\frac{\mu'-\mu}{m}F_0(\mu-x)\Big|\,\mathrm{d}x\\
\le c_1|\mu-\mu'|+&\frac{c}{m}|\mu-\mu'|\int_{\mathbb{R}}|G(x+\mu')-F_0(x+\mu')+G(\mu'-x)-F_0(\mu-x)|\,\mathrm{d}x,
\end{aligned}$$

where the finite constant c_1 arises from lemma 1 and from the fact that v is compact. It remains to show that

$$\int_{\mathbb{R}}|G(x+\mu')-F_0(x+\mu')+G(\mu'-x)-F_0(\mu-x)|\,\mathrm{d}x$$

is bounded uniformly with respect to μ. Using (1) for the cdf G we have

$$\begin{aligned}
&\int_{\mathbb{R}} |G(x+\mu') - F_0(x+\mu') + G(\mu'-x) - F_0(\mu-x)|\,\mathrm{d}x \\
&\le \int_{\mathbb{R}} |(1-p_0)(F_0(x+\mu) - F_0(x+\mu')) + p_0(F(\mu-\mu_0+x) - F_0(x-\mu') \\
&\quad + (1-p_0)(F_0(\mu'-x) - F_0(\mu-x)) + p_0(F(\mu'-\mu_0-x) - F_0(\mu-x)|\,\mathrm{d}x \\
&\le 2(1-p_0)|\mu-\mu'| + 2p_0\left(|\mu| + |\mu'| + |\mu_0| + \int_{\mathbb{R}} |x|\,\mathrm{d}F_0(x) + \int_{\mathbb{R}} |x|\,\mathrm{d}F(x)\right) \\
&\le c_2 < +\infty,
\end{aligned}$$

where we used lemmas 1 and 2 and the compactness of v. The continuity of d on v follows.

Step 2. Clearly, if $\mu=\mu_0$ then $d(\mu)=0$. Let us prove the converse. If $d(\mu)=0$, then we have $H_1=H_2$, i.e.:

$$\mu g(\mu+x) - (\mu-m)f_0(\mu+x) = \mu g(\mu-x) - (\mu-m)f_0(\mu-x), \tag{28}$$

almost everywhere on $\mathbb{R}$.

Under A1. Let us consider the Fourier transform of the above equality. Using the fact that

$$g(x) = (1-p_0)f_0(x) + p_0 f(x-\mu_0), \quad \forall x \in \mathbb{R}, \tag{29}$$

we get

$$(p-p_0)\sin(t\mu)\hat{f}_0(t) = p_0 \sin(t(\mu_0-\mu))\hat{f}(t), \quad \forall t \in \mathbb{R}. \tag{30}$$

Assume that $\mu \neq \mu_0$. On the one hand, because $\hat{f}_0 > 0$, the above equality implies that there exists $k_0 \in \mathbb{N}^*$ such that $|\mu| = k_0|\mu_0-\mu|$. On the other hand, taking the derivatives of order 1 and 3 of (30) at $t=0$ we get

$$p_0\mu_0 = p\mu \quad \text{and} \quad (p-p_0)(\mu^3+3\theta_0\mu) = p_0((\mu_0-\mu)^3 + 3\theta(\mu_0-\mu)),$$

where θ_0 and θ denote the moments of order 2 of f_0 and f respectively. These last two equalities with $|\mu| = k_0|\mu_0-\mu|$ imply that

$$\theta = \theta_0 + \mu_0^2 \frac{k_0 \pm 2}{3k_0},$$

and then, $(\mu,\theta) \in \cup_{k\in\mathbb{N}^*}\Phi_k$, which in turn implies that $(\mu,\theta) \notin \Phi_c$. It follows that $\mu=\mu_0$.

Under A2. From (28), (29), and from the fact that f is an even pdf, we obtain for $x \in \mathbb{R}$:

$$f_0(x+\mu)(m-p_0\mu) + p_0\mu f(x+\mu-\mu_0) = f_0(\mu-x)(m-p_0\mu) + p_0\mu f(x+\mu_0-\mu).$$

Dividing the above equality by $f_0(x+\mu)$ (or by $f_0(\mu-x)$ depending on the tail property of f_0) and making x tend to infinity, we obtain $m=p_0\mu$, which leads to $\mu=\mu_0$.

Under A3. Considering relation (31) for large values of $|x|$, we obtain $f(\mu-\mu_0+x) = f(\mu-\mu_0-x)$ which, according to assumption A3, is possible only if $\mu=\mu_0$.

Step 3. By step 1 d is continuous on the compact subset $v_\varepsilon = \{\mu \in v; |\mu-\mu_0| \ge \varepsilon\}$ of v. Therefore, there exists $\mu^* \in v_\varepsilon$ such that $d(\mu) \ge d(\mu^*)$ on v_ε. By step 2 we have $\delta_\varepsilon = d(\mu^*) > 0$. It follows that the expected result also holds with strict inequalities.