TRABAJOS DE ESTADISTICA
Vol. 6, núm. 1, 1991, pp. 89 a 106

INFERENCIA BAYESIANA EN MIXTURAS: METODOS APROXIMADOS

E. CARO, J. I. DOMÍNGUEZ y
F. J. GIRÓN
Departamento de Estadística
Facultad de Ciencias
Universidad de Málaga

SUMMARY

The problem of approximating mixtures of distributions has received considerable attention recently. In this paper we consider problems of estimating the mixing proportions of a finite mixture from a Bayesian perspective. The problems which arise from this methodology are basically approximations of finite mixtures of distributions. We propose two approximating methods and prove that under certain conditions both methods are asymptotically equivalent to a third method, which turns out to be simpler and computationally more efficient than the others. The paper concludes with a simulation study which analyses the goodness of the three methods with respect to the exact solution.

Key words: approximations; finite mixtures; Kullback-Leibler divergence; method of moments; simulation.

AMS Classification: 62J05; 62E15; 62F10.

RESUMEN

Últimamente se ha dedicado una gran atención a las técnicas de aproximación de mixturas de distribuciones. En este trabajo se consideran problemas de estimación de los parámetros de mezcla en una mixtura finita, desde una metodología bayesiana, que conducen a problemas de aproximación de mixtu-

Recibido: Mayo 1990.
Revisado: Octubre 1990.

ras finitas y se proponen dos nuevos métodos de aproximación. Bajo ciertas condiciones se demuestra que ambos métodos son asintóticamente equivalentes a un tercer método, de aplicación mucho más sencilla. El trabajo se concluye con un estudio de simulación en el que se analiza la bondad de los métodos de aproximación que aquí se exponen.

Palabras y frases clave: aproximaciones; divergencia de Kullback-Leibler; método de los momentos; mixturas finitas de distribuciones; simulación.

Clasificación AMS: 62J05; 62E15; 62F10.

1. INTRODUCCION

Como es bien sabido, los problemas de estimación y clasificación simultánea de observaciones, cuando se plantean desde el punto de vista bayesiano, conducen a una distribución a posteriori que tiene forma de una mixtura finita con un número de términos que crece exponencialmente con el tamaño muestral, lo que hace que la solución exacta del problema sea generalmente intratable desde un punto de vista analítico debido al carácter expansivo del número de términos de la mixtura.

En este artículo se considera un caso particular de estimación de mixturas cuya solución se puede generalizar a problemas más complejos que pueden englobar al análisis de conglomerados, clasificación probabilística, detección de outliers, etc.

Una característica importante de estos modelos es que pueden considerarse como modelos jerárquicos en varias etapas, lo que contribuye a clarificar mejor su estructura y permite un tratamiento homogéneo de problemas, en apariencia, diferentes.

En el apartado dos se plantea el problema objeto del trabajo, se le reinterpreta como modelo jerárquico y se analiza su solución bayesiana y se comentan los problemas de cálculo que se originan. En el apartado tres se presentan dos aproximaciones de tipo secuencial para el problema que, siguiendo una nomenclatura usual llamaremos *cuasi-bayesianas*, aproximaciones que retienen muchos aspectos de la solución bayesiana exacta y que son computacionalmente mucho más tratables. Por último, en el apartado cuatro se presenta un estudio de simulación que permite comparar la solución exacta con las que se proponen en este artículo.

2. PLANTEAMIENTO DEL PROBLEMA

Supongamos una muestra aleatoria formada por n observaciones $\boldsymbol{x} = (x_1, ..., x_n)$, que han sido generadas de la siguiente forma:

i) Cada observación procede de una y sólo una de las poblaciones $1, ..., j, ..., k$.

ii) La probabilidad de que una observación proceda de la población j-ésima es λ_j con $(0 < \lambda_j < 1)$.

iii) La distribución de las observaciones dentro de la población j-ésima es perfectamente conocida (es decir, no depende de parámetros desconocidos) y está caracterizada por una función de densidad $f_j(x)$. Aunque para formular el modelo teórico suponemos que la distribución de las observaciones en cada población es perfectamente conocida, el método que aquí proponemos se puede usar también para el caso en el que la distribución de las observaciones dentro de cada población dependa de parámetros desconocidos sobre los que se desea aprender.

iv) No se conoce para ningún elemento de la muestra de qué población procede.

En el caso de que hubiese algunas observaciones confirmadas o se dispusiese de un banco de datos confirmados previamente, obvias modificaciones de nuestro modelo permitirían su tratamiento.

Como consecuencia del planteamiento que hemos formulado para nuestro problema, el modelo probabilístico de obtención de la muestra viene dado por:

$$f(x; \boldsymbol{\lambda}) = \sum_{j=1}^{k} \lambda_j f_j(x); \quad \boldsymbol{\lambda} = (\lambda_1. ..., \lambda_k); \quad \sum_{j=1}^{k} \lambda_j = 1. \tag{2.1}$$

Nuestro objetivo ante problemas de este tipo suele ser doble: Por una parte, tratamos de determinar las probabilidades de clasificación de cada observación en cada una de las k poblaciones y, por otra parte, queremos estimar las proporciones λ_j.

Una manera útil de mirar al modelo (2.1), sobre todo considerando el primer objetivo señalado, es considerarlo como un modelo jerárquico, en el que la introducción de los hiperparámetros de la primera etapa permite de modo inmediato resolver este primer punto.

Si consideramos que la muestra procede de una colección potencialmente infinita de variables aleatorias intercambiables tendremos que

$$X_i \sim f(x_i \mid \pi_i) \quad i = 1, 2, ..., n.$$

Si a su vez suponemos que los hiperparámetros $\pi_1, ..., \pi_n$ son intercambiables y procedentes de una distribución discreta Π, tendremos el siguiente modelo jerárquico.

1.ª etapa: $\Pi = j$ con probabilidad $\lambda_j, (1 \leqslant j \leqslant k)$;

2.ª etapa: $\boldsymbol{\lambda} \sim p(\boldsymbol{\lambda})$ es la distribución a priori sobre el simplex S_k.

El modelo (2.1), junto con la distribución a priori $p(\boldsymbol{\lambda})$ es equivalente al modelo jerárquico propuesto si convenimos en que

$$f(x_i \mid \pi_i = j) = f_j(x_i) \quad i = 1, 2, ..., n; \quad j = 1, 2, ..., k.$$

La verosimilitud del vector $\boldsymbol{\lambda}$ viene dada directamente por (2.1)

$$l(\boldsymbol{\lambda}; \boldsymbol{x}) = \prod_{i=1}^{n} \left(\sum_{j=1}^{k} \lambda_j f_j(x_i) \right). \tag{2.2a}$$

Otra forma más interesante de obtener un resultado equivalente al anterior es, dada la equivalencia con el modelo jerárquico, calcular la distribución predictiva de $\boldsymbol{x}$ condicionada a $\boldsymbol{\lambda}$, a veces denominada verosimilitud integrada, que se obtiene eliminando los hiperparámetros $\pi_1, ..., \pi_n$.

$$\begin{aligned} p(\boldsymbol{x} \mid \boldsymbol{\lambda}) &= \sum_{j_n=1}^{k} \cdots \sum_{j_1=1}^{k} \prod_{i=1}^{n} f(x_i \mid \pi_i = k_i) \Pr(\pi_i = j_i \mid \boldsymbol{\lambda}) \\ &= \sum_{j_1=1}^{k} \cdots \sum_{j_n=1}^{k} \lambda_{j_1} \cdots \lambda_{j_n} \prod_{i=1}^{n} f_{j_i}(x_i); \end{aligned} \tag{2.2b}$$

ya que, por ii), $\Pr \pi_i = j_i \mid \boldsymbol{\lambda}) = \lambda_{j_i}$.

De (2.2a), o mejor de (2.2b), se obtiene que la distribución a posteriori de $\boldsymbol{\lambda}$ dado $\boldsymbol{x}$ viene dada por

$$p(\boldsymbol{\lambda} \mid \boldsymbol{x}) \propto p(\boldsymbol{\lambda}) \sum_{j_1=1}^{k} \cdots \sum_{j_n=1}^{k} \prod_{i=1}^{n} f_{j_i}(x_i) \lambda_{j_1} \cdots \lambda_{j_n}; \tag{2.3}$$

que es la solución bayesiana al problema de estimación de $\boldsymbol{\lambda}$.

Obsérvese que dada la forma especial de la verosimilitud integrada, si a priori $p(\lambda) \sim \mathrm{Di}(\lambda; \alpha_1^{(0)}, ..., \alpha_k^{(0)})$, entonces la distribución a posteriori es siempre una mixtura finita de distribuciones Dirichlet.

El problema de clasificación queda reducido a calcular $\Pr(\pi_1 = j_1, ..., \pi_n = j_n \mid \boldsymbol{x})$, que representa la probabilidad de clasificación conjunta. A partir de aquí, la probabilidad individual de clasificación a posteriori viene dada por la distribución marginal $\Pr(\pi_i = j_i \mid \boldsymbol{x})$.

Las probabilidades a priori de los hiperparámetros $\pi_1, ..., \pi_n$ se obtienen eliminando los hiperparámetros $\lambda_1, ..., \lambda_k$ de la última etapa del modelo jerárquico (determinación del baricentro de la densidad a priori $p(\lambda)$), del modo siguiente

$$\Pr(\pi_1 = j_1, ..., \pi_n = j_n) = \int_{S_k} \lambda_{j_1} \cdots \lambda_{j_n} p(\lambda)\, d\lambda.$$

Como por otro lado $\Pr(\boldsymbol{x} \mid \pi_1 = j_1, ..., \pi_n = j_n) = f_{j_1}(x_1) \cdots f_{j_n}(x_n)$, del teorema de Bayes se tiene que

$$\Pr(\pi_1 = j_1, ..., \pi_n = j_n \mid \boldsymbol{x}) \propto \Pr(\pi_1 = j_1, ..., \pi_n = j_n) \prod_{i=1}^{n} f_{j_i}(x_i). \qquad (2.4)$$

El problema fundamental con las fórmulas (2.3) y (2.4) es el elevado número de sumandos, k^n, en la mixtura de la distribución a posteriori de λ condicionada a $\boldsymbol{x}$ y el cálculo de k^n probabilidades de clasificación conjunto para obtener éstas y también las probabilidades de clasificación individual.

La solución que se ofrece en la sección siguiente, similar a otras propuestas en las referencias, consiste en aplicar secuencialmente el teorema de Bayes, e ir aproximando en cada etapa la mixtura (2.3) de distribuciones de Dirichlet por una distribución también de la familia de las Dirichlet, de modo que la distribución a posteriori se mantenga en cada etapa dentro de esta familia, y estimar los nuevos parámetros mediante una especie de filtro recursivo que se obtiene para este modelo y puede ser generalizado a situaciones más complejas.

3. METODO SECUENCIAL DE APRENDIZAJE CUASI-BAYESIANO

Como hemos comentado en el apartado anterior, nuestro objetivo es aprender sobre $\lambda = (\lambda_1, ..., \lambda_n)$. Por los resultados del apartado anterior sabemos que si la distribución a priori de λ es una Dirichlet, es decir $p(\lambda) \sim \text{Di}(\lambda; \alpha_1^{(0)}, ..., \alpha_k^{(0)})$, entonces la distribución a posteriori, para cualquier tamaño muestral, es una mixtura finita de distribuciones Dirichlet. En esta sección y en las siguientes supondremos que la información a priori viene dada por una distribución de este tipo.

De (2.3) se sigue, tras algunos cálculos, que la distribución a posteriori de λ condicionada a x_1 viene dada por la mixtura

$$(p\lambda \mid x_1) = \sum_{i=1}^{k} p_i(x_1) \, \text{Di}(\lambda; \alpha_1^{(0)} + \delta_{i1}, ..., \alpha_k^{(0)} + \delta_{ik}), \tag{3.1}$$

donde δ_{ij} es la δ de Kronecker y los coeficientes de la mixtura, $p_i(x_1)$, vienen dados por

$$p_i(x_1) = \frac{\alpha_i^{(0)} f_i(x_1)}{\sum_{j=1}^{k} \alpha_j^{(0)} f_j(x_1)}; \quad i = 1, ..., k.$$

Tenemos así que la distribución a posteriori, después de observar el valor x_1, viene dada por una mixtura de distribuciones Dirichlet en donde el vector de coeficientes o pesos de la mixtura $\boldsymbol{p}(x_1) = (p_1(x_1), ..., p_k(x_1))$, como resulta fácil comprobar, representa las probabilidades de clasificación de esta primera observación en cada una de las poblaciones $1, ..., j, ..., k$.

La aplicación secuencial del teorema de Bayes conduce a la ecuación (2.3). Para evitar esta explosión combinatoria y a la vez permanecer dentro de la familia conjugada de las mixturas finitas de distribuciones Dirichlet, que nos van a permitir la construcción de un filtro aproximado, se recurre a aproximaciones de estas mixturas por un miembro de la familia de las Dirichlet de la forma $\text{Di}(\lambda; \alpha_1^{(1)}, ..., \alpha_k^{(1)})$; y así se procedería sucesivamente en cada etapa.

De donde resulta que si en la etapa h-ésima la distribución a posteriori aproximada es una $\text{Di}(\lambda; \alpha_1^{(h)}, ..., \alpha_k^{(h)})$, entonces la distribución a posteriori después de observar x_{h+1} es

$$p(\lambda \mid x_1, ..., x_{h+1}) = \sum_{i=1}^{k} p_i(x_1, ..., x_{h+1}) \, \mathrm{Di}\,(\lambda; \alpha_1^{(h)} + \delta_{i1}, ..., \alpha_k^{(h)} + \delta_{ik}), \quad (3.2)$$

y el vector de coeficientes de la mixtura, que representa las probabilidades aproximadas de clasificación a posteriori, $\boldsymbol{p}(\boldsymbol{x}_{h+1}) = \boldsymbol{p}(x_1, ..., x_{h+1})$, viene dado por sus coordenadas

$$p_i(x_1, ..., x_{h+1}) = \frac{\alpha_i^{(h)} f_i(x_{h+1})}{\sum_{j=1}^{k} \alpha_j^{(h)} f_j(x_{h+1})}; \quad i = 1, ..., k.$$

Este problema de aproximación ha sido tratado ampliamente en la literatura. Entre los autores que han tratado este tema podemos citar a Owen (1975) quien basa su método de aproximación en lo que denomina un *editor probabilístico* que no es sino una variante del método de los momentos; en Makov y Smith (1977) y Smith y Makov (1978) se propone un método que denominan *cuasi-bayes*, en el que en cada etapa del proceso de aprendizaje $\alpha_i^{(h+1)} = \alpha_i^{(h)} + p_i(x_1, ..., x_{h+1})$ para $i = 1, ..., k$. Makov (1981) y Caro, Domínguez y Girón (1985, 1986a, 1986b) consideran éste y otros problemas más complejos como la aproximación de mixturas de normales desde diferentes puntos de vista; Bernardo y Girón (1988a, 1988b) estudian la aproximación que resulta de minimizar la divergencia logarítmica de Kullback-Leibler y Bermúdez y Sendra (1989) consideran un método híbrido que denominan *minimización restringida de la divergencia.*

Nuestro planteamiento parte de la base de que cualquier método de aproximación que se elija debe ser compatible con las tres condiciones siguientes:

a) En cada etapa $h = 1, ..., n$, el parámetro $\alpha_+^{(h)} = \sum \alpha_i^{(h)}$ debe aumentar como máximo en una unidad respecto de la etapa anterior, es decir, $\alpha_+^{(h+1)} \leqslant \alpha_+^{(h)} + 1$. Esta condición sí la cumple, por ejemplo, la solución *cuasi-bayesiana* dada por Smith y Makov, aunque debemos observar que en cada etapa $\alpha_+^{(h)}$ aumenta exactamente una unidad, lo que no es muy razonable cuando las poblaciones $f_j(x)$ están muy próximas.

b) En caso de clasificación perfecta, es decir cuando $\boldsymbol{p}(x_1, ..., x_{h+1}) = (0, ..., 1, ..., 0)$, se debe dar la igualdad del requisito a), es decir, $\alpha_+^{(h+1)} = \alpha_+^{(h)} + 1$.

c) Si las probabilidades de clasificación en una etapa son proporciona-

les a los parámetros de la Dirichlet de la etapa anterior, es decir, $p(x_1, ..., x_{h+1}) \propto (\alpha_1^{(h)}, ..., \alpha_k^{(h)})$, estos parámetros no deben cambiar, es decir, $\alpha_i^{(h+1)} = \alpha_i^{(h)}$ para todo $i = 1, ..., k$. Condición que no se cumple, en general, para la solución *cuasi-bayes.*

Los requisitos anteriores parecen intuitivamente obvios: a) y b) afirman que en cada etapa del proceso de aprendizaje el *tamaño muestral equivalente* $\alpha_+^{(h)}$ nunca debe aumentar más de una unidad, salvo en el caso de clasificación perfecta en que el *tamaño muestral equivalente* aumenta exactamente en una unidad. c) recoge la idea de que cuando las probabilidades de clasificación observadas coinciden con las esperadas, no hay aprendizaje. Obsérvese que no se exige la condición de monotonía creciente de $\alpha_+^{(h)}$ como función de h para todo h, es decir, que ciertas observaciones mal clasificadas pueden incluso disminuir el *tamaño muestral equivalente* en alguna etapa del proceso de aprendizaje.

Los métodos de aproximación que aquí se proponen son dos: uno es una variante simplificada del método de los momentos y el otro es el método de minimización de la divergencia logarítmica de Kullback-Leibler, que ya ha sido estudiado con cierto detalle por Bernardo y Girón (1988a, 1988b). Ambos métodos cumplen las tres condiciones dadas anteriormente.

La variante del método de los momentos que proponemos utiliza, aparte de los momentos de primer orden, la traza de la matriz de covarianzas, lo que permite dar expresiones explícitas de los nuevos parámetros en función de los anteriores y de las probabilidades de clasificación en cada etapa. De modo que nuestro procedimiento consiste en igualar el vector de medias y la traza de la matriz de covarianzas de la mixtura (3.2) a los de la nueva distribución $\mathrm{Di}(\lambda; \alpha_1^{(h+1)}, ..., \alpha_k^{(h+1)}$ y resolver el sistema de ecuaciones resultante en las variables $\boldsymbol{\alpha}^{(h+1)} = (\alpha_1^{(h+1)}, ..., \alpha_k^{(h+1)})$. Otra posibilidad que no contemplamos aquí, computacionalmente más complicada salvo para el caso de dos poblaciones, sería utilizar la varianza generalizada.

El vector de medias de la $\mathrm{Di}(\lambda; \boldsymbol{\alpha}^{(h+1)})$ es $\boldsymbol{\alpha}^{(h+1)}/\alpha_+^{(h+1)}$, mientras que el de la mixtura (3.2) es $(\boldsymbol{\alpha}^{(h)} + \boldsymbol{p}(\boldsymbol{x}_{h+1}))/(\alpha_+^{(h)} + 1)$, como puede verse fácilmente. De modo que la ecuación vectorial resultante de igualarlos es

$$\frac{\boldsymbol{\alpha}^{(h+1)}}{\alpha_+^{(h+1)}} = \frac{\boldsymbol{\alpha}^{(h)} + \boldsymbol{p}(\boldsymbol{x}_{h+1})}{\alpha_+^{(h)} + 1}. \tag{3.3}$$

La otra ecuación, resultante de igualar las trazas de las matrices de momentos de segundo orden respecto del origen de la Di $(\lambda; \alpha^{(h+1)})$ y de la mixtura (3.2), es la siguiente,

$$\frac{\sum_{i=1}^{k} \alpha_i^{(h+1)}(\alpha_i^{(h+1)}+1)}{\alpha_+^{(h+1)}(\alpha_+^{(h+1)}+1)} = \frac{\sum_{i=1}^{k} (\alpha_i^{(h)}+1)(\alpha_i^{(h)}+2p_i(x_{h+1}))}{(\alpha_+^{(h)}+1)(\alpha_+^{(h)}+2)}. \tag{3.4}$$

Hay que hacer constar que si en lugar de utilizar la traza de la matriz de momentos de segundo orden se hubiese utilizado la traza de la matriz de varianzas y covarianzas, como hemos sugerido en el procedimiento de aproximación, las ecuaciones resultantes serían equivalentes y el resultado final sería el mismo. El hacerlo así simplifica notablemente los cálculos.

De la ecuación vectorial (3.3) se deduce que

$$\alpha_i^{(h+1)} = (\alpha_i^{(h)} + p_i(x_{h+1}))\cdot C(h); \quad i = 1, ..., k; \quad h = 1, ..., n; \tag{3.5}$$

donde la constante de proporcionalidad en cada etapa de la recurrencia $C(h)$, que se determina a partir de la ecuación (3.4), resulta ser

$$C(h) = \frac{A(h)}{A(h) + B(h)}, \quad h = 1, ..., n; \tag{3.6a}$$

donde

$$\begin{aligned} A(h) &= \alpha_+^{(h)}(\alpha_+^{(h)} + 2) - \sum_{i=1}^{k} \alpha_i^{(h)}(\alpha_i^{(h)} + 2p_i(x_{h+1})); \\ B(h) &= (\alpha_+^{(h)} + 2)\left(1 - \sum_{i=1}^{k} p_i^2(x_{h+1})\right). \end{aligned} \tag{3.6b}$$

Es evidente que ambas $A(h)$ y $B(h)$ son no negativas, por lo cual $C(h) \leqslant 1$, que implica la condición a), verificándose la igualdad solamente en el caso de clasificación perfecta, en cuyo caso $B(h) = 0$, es decir se cumple la condición b). Para demostrar c) basta probar que, bajo esta condición, la constante de proporcionalidad $C(h) = \alpha_+^{(h)}/(\alpha_+^{(h)} + 1)$, ya que esto implicaría, por la ecuación (3.3), que $\alpha^{(h+1)} = \alpha^{(h)}$.

El método de aproximación de la mixtura (3.2) por una Dirichlet Di $(\lambda; \alpha^{(h+1)})$, basado en la minimización de la divergencia de Kullback-

Leibler, como puede verse, por ejemplo, en Bernardo y Girón (1988a), conduce a las ecuaciones siguientes

$$\begin{aligned}
&\psi(\alpha_+^{(h+1)}) - \psi(\alpha_1^{(h+1)}) = \psi(\alpha_+^{(h)} + 1) - \psi(\alpha_1^{(h)}) - \frac{p_1(x_{h+1})}{\alpha_1^{(h)}} \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \qquad (3.7) \\
&\psi(\alpha_+^{(h+1)}) - \psi(\alpha_k^{(h+1)}) = \psi(\alpha_+^{(h)} + 1) - \psi(\alpha_k^{(h)}) - \frac{p_k(x_{h+1})}{\alpha_k^{(h)}}
\end{aligned}$$

donde $\psi(\cdot)$ representa la función *digamma*, que es la derivada logarítmica de la función $\Gamma(\cdot)$. En Bernardo y Girón (1988b) se demuestra que el procedimiento anterior cumple las condiciones a), b) y c).

4. COMPARACION ENTRE LOS METODOS DE APRENDIZAJE: APROXIMACIONES Y SIMULACION

En este apartado resumimos algunos de los resultados obtenidos en un extenso estudio de simulación. Nos restringiremos al caso de $k = 2$ poblaciones normales con diversas proporciones de mezcla λ y un tamaño muestral moderado $n = 50$.

Para el caso de dos poblaciones el modelo probabilístico sería

$$f(x; \lambda) = \lambda f_1(x) + (1 - \lambda) f_2(x);$$

de modo que si a priori $\lambda \sim \text{Be}(\alpha^{(0)}, \beta^{(0)})$, la distribución a posteriori aproximada de λ correspondiente a la etapa $h + 1$ sería una $\text{Be}(\alpha^{(h+1)}, \beta^{(h+1)})$, donde $\alpha^{(h+1)}$ y $\beta^{(h+1)}$ son funciones de los $\alpha^{(h)}$ y $\beta^{(h)}$ y la probabilidad aproximada de clasificación $p_{h+1} = p(x_{h+1})$ de x_{h+1} en la primera población, cuya forma depende del método de aproximación elegido.

Para este caso de dos poblaciones presentamos una nueva aproximación, más sencilla de calcular, que demostramos es asintóticamente equivalente a los dos métodos propuestos en el apartado anterior.

La idea de la nueva aproximación es la siguiente: buscar en cada etapa la aproximación más simple a $\alpha^{(h+1)}$ y $\beta^{(h+1)}$ que cumpla las condiciones a), b) y c).

La aproximación lineal en p_{h+1}, que cumple los requisitos a) y b), es

precisamente la solución cuasi-bayesiana de Smith y Makov, dada por las ecuaciones

$$\alpha_{SM}^{(h+1)} = \alpha_{SM}^{(h)} + p_{h+1}$$
$$\beta_{SM}^{(h+1)} = \beta_{SM}^{(h)} + 1 - p_{h+1};$$

que en general no cumple la condición c); además presenta el inconveniente, como ya se ha señalado, de que el *tamaño muestral equivalente* aumenta siempre exactamente una unidad de en cada etapa.

La mejor aproximación cuadrática en p_{h+1} que cumple los requisitos anteriores, y que además es única como es fácil comprobar, es la dada por las ecuaciones siguientes

$$\begin{aligned} \alpha_a^{(h+1)} &= \alpha_a^{(h)} - \frac{\alpha_a^{(h)}}{\beta_a^{(h)}} p_{h+1} + \left(\frac{\alpha_a^{(h)} + \beta_a^{(h)}}{\beta_a^{(h)}}\right) p_{h+1}^2 \\ \beta_a^{(h+1)} &= \beta_a^{(h)} + 1 - \left(\frac{2\alpha_a^{(h)} + \beta_a^{(h)}}{\alpha_a^{(h)}}\right) p_{h+1} + \left(\frac{\alpha_a^{(h)} + \beta_a^{(h)}}{\alpha_a^{(h)}}\right) p_{h+1}^2; \end{aligned} \tag{4.1}$$

Para el caso de dos poblaciones es bien sabido que asintóticamente, es decir cuando $\alpha^{(h)} \to \infty$, $\beta^{(h)} \to \infty$ y $\alpha^{(h)}/\beta^{(h)}$ no converge ni hacia 0 ni a ∞, el método de Kullback-Leibler y el de los momentos son equivalentes toda vez que, en estas condiciones, las distribuciones Beta son asintóticamente normales y para las distribuciones normales minimizar la divergencia de Kullback-Leibler es equivalente al método de los momentos.

Vamos a probar aquí que también la aproximación dada por las ecuaciones (4.1) es asintóticamente equivalente al método de los momentos, de modo que los tres procedimientos serían numéricamente equivalentes para valores grandes de los parámetros.

En efecto, las ecuaciones (3.5) y (3.6), para el caso de dos poblaciones se reducen a

$$\alpha_m^{(h+1)} = \frac{(\alpha^{(h)} + p_{h+1})(\alpha^{(h)}\beta^{(h)} + \alpha^{(h)}(1 - p_{h+1}) + \beta^{(h)} p_{h+1})}{\alpha^{(h)}\beta^{(h)} + \alpha^{(h)}(1 - p_{h+1}) + \beta^{(h)} p_{h+1} + (\alpha^{(h)} + \beta^{(h)} + 2) p_{h+1}(1 - p_{h+1})},$$

$$\beta_m^{(h+1)} = \frac{(\beta^{(h)} + 1 - p_{h+1})(\alpha^{(h)}\beta^{(h)} + \alpha^{(h)}(1 - p_{h+1}) + \beta^{(h)} p_{h+1})}{\alpha^{(h)}\beta^{(h)} + \alpha^{(h)}(1 - p_{h+1}) + \beta^{(h)} p_{h+1} + (\alpha^{(h)} + \beta^{(h)} + 2) p_{h+1}(1 - p_{h+1})};$$

donde el subíndice m hace referencia a los parámetros calculados por el método de los momentos.

La idea de la demostración es acotar en valor absoluto las diferencias $\alpha_a^{(h+1)} - \alpha_m^{(h+1)}$ y $\beta_a^{(h+1)} - \beta_m^{(h+1)}$, uniformemente en p_{h+1}.

Así tenemos que, tras algunas simplificaciones, $|\alpha_a^{(h+1)} - \alpha_m^{(h+1)}|$ es igual a

$$\left| \frac{p_{h+1}(1-p_{h+1})(-\alpha^{(h)}+(\alpha^{(h)}+\beta^{(h)})p_{h+1})(\alpha^{(h)}-\beta^{(h)}+(2+\alpha^{(h)}+\beta^{(h)}p_{h+1})}{\beta^{(h)}(-\alpha^{(h)}-\alpha^{(h)}\beta^{(h)}-(2+2\beta^{(h)})p_{h+1}+(2-\alpha^{(h)}+\beta^{(h)})p_{h+1}^2} \right|.$$

Acotando superiormente cada uno de los tres factores del numerador y acotando inferiormente el denominador, se obtiene finalmente la siguiente acotación

$$|\alpha_a^{(h+1)} - \alpha_m^{(h+1)}| \leqslant \begin{cases} \dfrac{\alpha^{(h)}}{2\beta^{(h)2}}, & \text{si} \quad 0 < \beta^{(h)} \leqslant \alpha^{(h)}; \\ \dfrac{\alpha^{(h)}+1}{2\alpha^{(h)}(\beta^{(h)}+1)}, & \text{si} \quad \alpha^{(h)} < \beta^{(h)} < 3\alpha^{(h)}+2; \\ \dfrac{\beta^{(h)}-\alpha^{(h)}}{4\alpha^{(h)}(\beta^{(h)}+1)}, & \text{si} \quad 3\alpha^{(h)}+2 \leqslant \beta^{(h)} < \infty; \end{cases}$$

de donde se deduce inmediatamente que cuando $\alpha^{(h)} \to \infty$, $\beta^{(h)} \to \infty$ y $\alpha^{(h)}/\beta^{(h)}$ no converge hacia ∞, la diferencia $|\alpha_a^{(h+1)} - \alpha_m^{(h+1)}| \to 0$.

Análogamente, intercambiando los papeles de $\alpha^{(h)}$ y $\beta^{(h)}$, se obtendría una acotación análoga para el valor absoluto de la diferencia $|\beta_a^{(h+1)} - \beta_m^{(h+1)}|$, que convergiría hacia 0 cuando $\alpha^{(h)} \to \infty$, $\beta^{(h)} \to \infty$ y $\alpha^{(h)}/\beta^{(h)}$ no converge hacia 0, con lo que queda demostrada la equivalencia asintótica.

Este método para calcular valores aproximados se ha comprobado que da resultados muy satisfactorios si se comienza a utilizar cuando los dos los parámetros de la distribución Beta aproximada son mayores que cinco, lo que generalmente se consigue cuando las probabilidades de pertenencia a cada población no son demasiado pequeñas, aún en el caso de que se haya analizado un número no muy grande de observaciones.

Los resultados de este artículo están dirigidos a estudiar el comportamiento de los métodos de aproximación propuestos para muestras de

tamaño moderado, de ahí la justificación de incluir los resultados de nuestra simulación. Quedan pendientes para un próximo trabajo, como acertadamente nos han señalado los dos evaluadores, el estudio del comportamiento asintótico de estos procedimientos, utilizando técnicas de aproximación estocástica, para procedimientos recursivos, análogas a las de Makov y Smith (1977).

Las figuras que a continuación siguen ilustran algunos aspectos de los métodos de aprendizaje que hemos considerado y su comparación con los resultados exactos, tales como las distribuciones a posteriori aproximadas del parámetro de mezcla λ y las probabilidades aproximadas de clasificación de los elementos de la muestra en la primera de las poblaciones, calculadas por uno de los tres métodos de aproximación, para tres supuestos distintos. El no incluir en las figuras 1b, 2b y 3b comparaciones entre las probabilidades de clasificación aproximadas calculadas por los tres métodos se debe a que en ninguno de los casos simulados se obtuvieron diferencias apreciables.

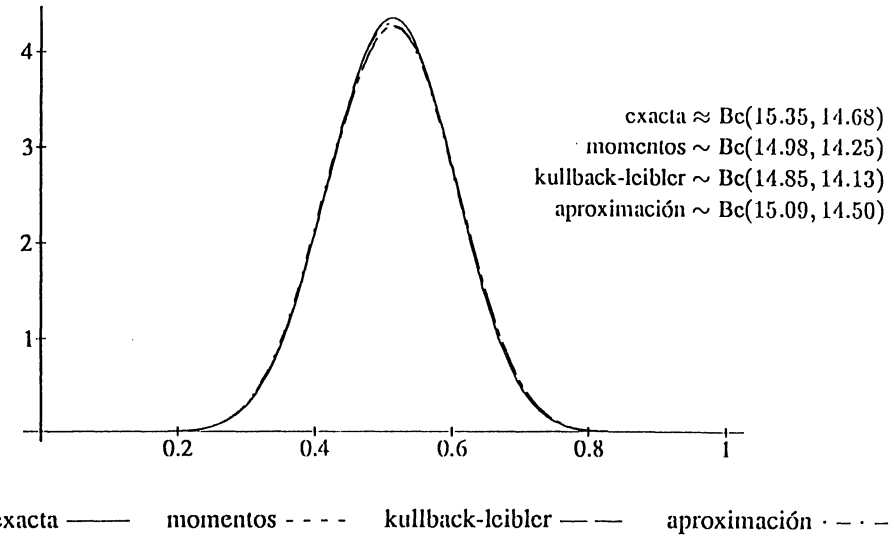


Figura 1a *Comparación entre la función de verosimilitud exacta de* λ *y las aproximaciones dadas por los otros tres métodos para una muestra de tamaño* 50 *del modelo* $.5N(x;-1,1)+.5N(x;1,1)$.

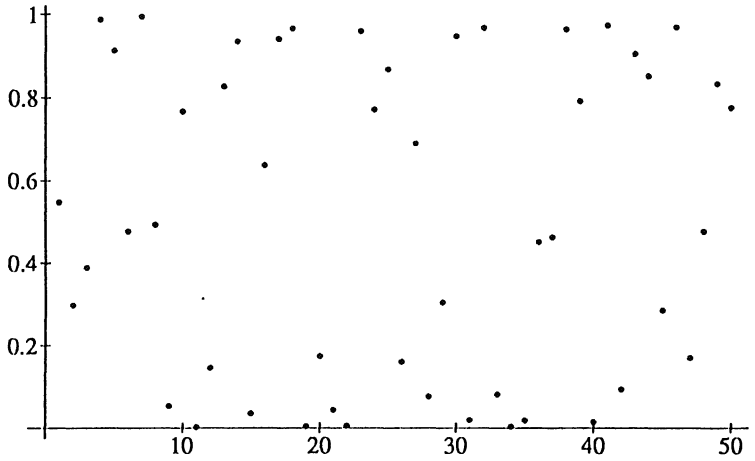
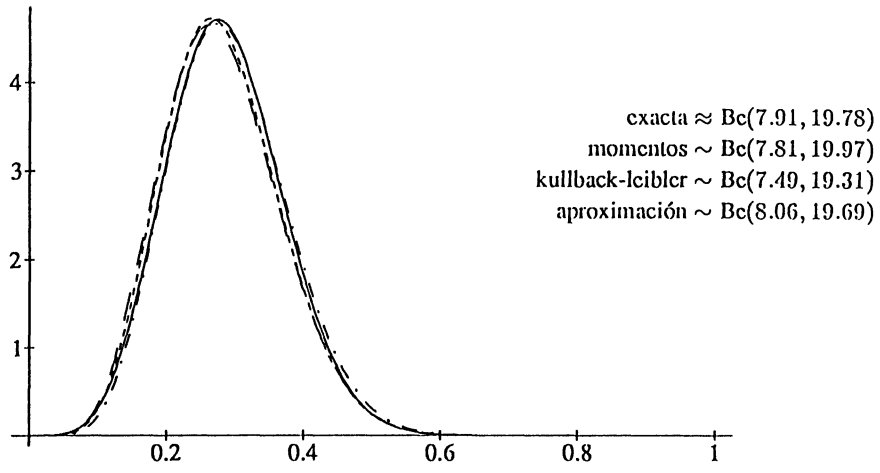


Figura 1b *Probabilidades aproximadas de clasificación de los elementos de una muestra de tamaño* 50 *del modelo* $.5N(x;-1,1)+.5N(x;1,1)$ *en la primera población.*



exacta —— momentos - - - - kullback-leibler — — aproximación · — · —

Figura 2a *Comparación entre la función de verosimilitud exacta de* λ *y las aproximaciones dadas por los otros tres métodos para una muestra de tamaño* 50 *del modelo* $.2N(x;-1,1)+.8N(x;1,1)$.

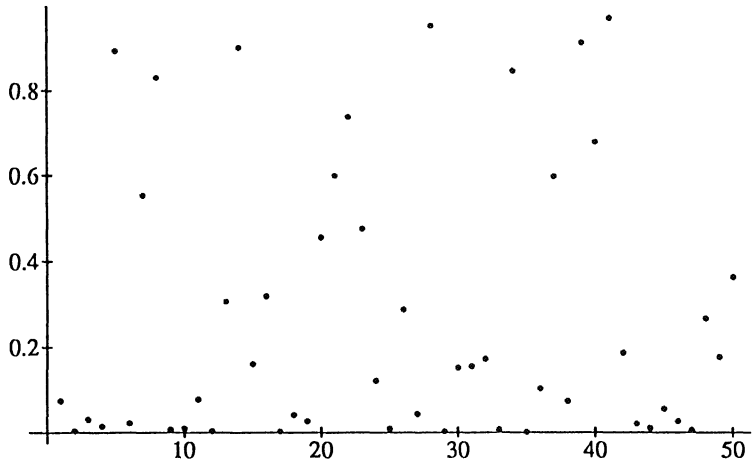


Figura 2b *Probabilidades aproximadas de clasificación de los elementos de una muestra de tamaño* 50 *del modelo* $.2N(x;-1,1)+.8N(x;1,1)$ *en la primera población.*

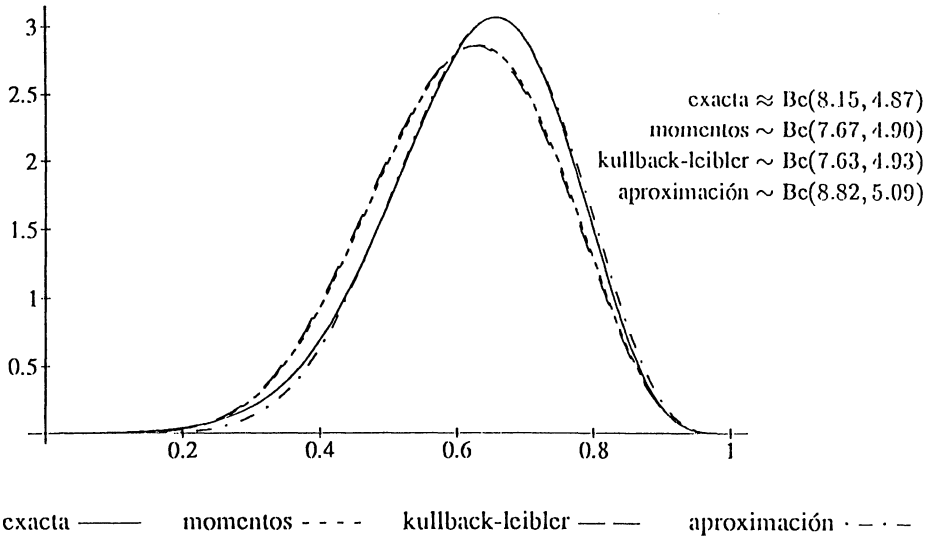


Figura 3a *Comparación entre la función de verosimilitud exacta de* λ *y las aproximaciones dadas por los otros tres métodos para una muestra de tamaño* 50 *del modelo* $.5N(x;0,1)+.5N(x;0,\sqrt{6})$.

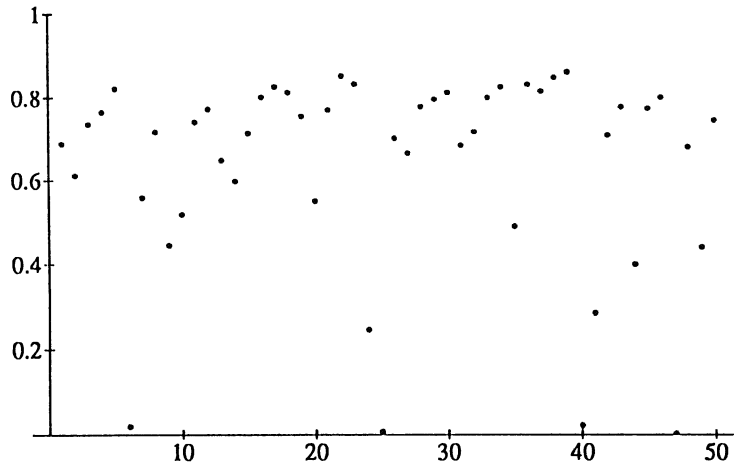


Figura 3b *Probabilidades aproximadas de clasificación de los elementos de una muestra de tamaño* 50 *del modelo* $.5N(x;0,1) + .5N(x;0,\sqrt{6})$ *en la primera población.*

En todos los casos hemos considerado muestras del mismo tamaño, $n = 50$, que se puede considerar moderado y la misma distribución inicial para $\lambda \sim \text{Be}(1,1)$, es decir una distribución uniforme en (0, 1). La aproximación a la distribución exacta por una distribución Beta, que aparece en las gráficas, se ha obtenido minimizando la discrepancia de Kullback-Leibler a la verdadera distribución.

5. CONCLUSIONES

Del estudio de simulación para el caso de una mixtura de dos poblaciones, podemos deducir, entre otros, los siguientes resultados:

i) Cuando, para α típicamente pequeño, los $1 - \alpha$ soportes de las distribuciones —regiones de máxima densidad de contenido probabilístico $1 - \alpha$— son relativamente disjuntos, tanto el proceso de aprendizaje sobre λ, como las probabilidades de clasificación por cualquiera de los tres métodos dan resultados muy similares entre sí y también a los de la solución exacta. Además el orden en el que se examina el banco de datos es, en este caso, generalmente poco relevante, incluso cuando el tamaño del banco de datos es pequeño.

ii) Cuando los $1 - \alpha$ soportes de una distribución están contenidos en los de la otra, incluso en el caso en que la distancia de Mahalanobis entre ellas sea relativamente grande, los resultados son perores: se aprende lentamente y se clasifican mal aquellas observaciones pertenecientes al soporte común. Incluso la distribución exacta no se aproxima bien por una distribución Beta. En estos casos, los resultados de la simulación nos sugieren que una mixtura, aunque sea de dos términos únicamente, resulta ser una aproximación mucho mejor a la verdadera distribución a posteriori de λ. También en estos casos, y para muestras de tamaño moderado, el orden en que se procesan los datos en los procedimientos de aprendizaje secuencial influye en la distribución a posteriori.

En un próximo artículo daremos resultados más extensos y precisos sobre los resultados y conclusiones del estudio de simulación llevado a cabo, que incluye también el caso de más de dos poblaciones y el análisis de datos multivariantes.

AGRADECIMIENTOS

Este trabajo ha sido realizado con ayuda de la *Consejería de Educación de la Junta de Andalucía* y de la *Dirección General de Investigación Científica y Técnica* (DGICYT) como parte del Proyecto de Referencia PB87-0607-C02-02.

También queremos dar las gracias a dos evaluadores anónimos del trabajo, cuyos comentarios y sugerencias han mejorado la presentación del mismo.

REFERENCIAS

BERMUDEZ, J. D., y SENDRA, M. (1989): «Aproximaciones en la inferencia bayesiana sobre mixturas». (Unpublished technical report).

BERNARDO, J. M., y GIRON, F. J. (1988a): «A Bayesian approach to cluster analysis», *Qüestió*, 12, n. 1, pp. 97-112.

BERNARDO, J. M., y GIRON, F. J. (1988b): «A Bayesian analysis of simple

mixture problems», in *Bayesian Statistics,* 3. (J. M. Bernardo, M. H. DeGroot, D. V. Lindley y A. F. M. Smith, eds.), (with discussion), pp. 67-78, Oxford University Press, Oxford.

CARO, E.; DOMINGUEZ, J. I., y GIRON, F. J. (1985): «Métodos de aprendizaje secuencial cuasi-bayesiano», *Actas XV Reunión Nacional de Estadística, Investigación Operativa e Informática,* vol. 1, pp. 78-85, Univ. Oviedo.

CARO, E.; DOMINGUEZ, J. I., hy GIRON, F. J. (1986a): «Métodos bayesianos aproximados para mixturas de normales», *Actas XVI Reunión Nacional de Estadística, Investigación Operativa e Informática,* vol. 1, pp. 213-220. Univ. de Málaga.

CARO, E.; DOMINGUEZ, J. I., 6 GIRON, F. J. (1986b): «Métodos bayesianos para mixturas de distribuciones», *Actas XVI Reunión Nacional de Estadística, Investigación Operativa e Informática,* vol. 1, pp. 221-229. Univ. de Málaga.

MAKOV, U. E. (1981): «Approximations of unsupervised Bayes learning procedures», *Bayesian Statistics,* 1 (Bernardo, J. M., DeGroot, M. H., Lindley, D. V. and Smith, A. F. M., eds.), pp. 69-82 (with discussion), Valencia University Press: Valencia.

MAKOV, U. E., and SMITH, A. F. M. (1977): «A quasi-Bayes unsupervised learning procedure for priors», *IEEE Trans. Inform. Th.,* IT-23, pp. 761-764.

OWEN, J. R. (1975): «A Bayesian sequential procedure for quantal response in the context of adaptive mental testing», *J. Amer. Statist. Associ.,* 70, pp. 351-356.

SMITH, A. F. M., y MAKOV, U. E. (1978): «A quasi-Bayes sequential procedure for mixtures», *J. Roy. Statist. Soc. B,* 40, pp. 106-111.