Mixturas de distribuciones normales con aplicaciones a problemas estadísticos complejos

Por F. J. GIRON

Recibido: 1 marzo 1989

Académico Correspondiente

RESUMEN

El presente artículo trata de ilustrar el cómo propiedades sencillas de los vectores normales, que tienen inmediata aplicación a problemas estadísticos referentes a datos normales, pueden ser generalizados a modelos más complejos cuando éstos vienen descritos por mixturas de distribuciones normales. En particular veremos cómo el filtro de Kalman, que desde nuestro enfoque no es sino una simple consecuencia de una propiedad elemental de la distribución normal multivariante, se extiende a situaciones o problemas complejos regulados por modelos probabilísticos representados por mixturas de distribuciones normales.

ABSTRACT

The present paper illustrates how simple properties of the normal random vectors, which have immediate applications to statistical problems involving normal data, can be generalized to more complex models when these are described by means of normal mixture models. In particular, we shall see how the Kalman filter, which is but a simple consequence of an elementary property of the multivariate normal distribution, is generalized to more complex situations or problems which may be described by models based on mixtures of normal distributions.

1. INTRODUCCION

En esta sección vamos a considerar una propiedad de la distribución normal multivariante que tiene aplicación inmediata a problemas de regresión con errores normales y permite deducir de una manera sencilla el llamado filtro de Plackett-Kalman, del que analizaremos con cierto detalle su relación con el enfoque bayesiano clásico y con otros conceptos básicos de la inferencia. Como el desarrollo del trabajo es mayormente ilustrativo de la potencialidad de los resultados de las proposiciones 1 y 2, a lo largo del mismo supondremos que la varianza σ^2, que aparece en los modelos considerados, es conocida. Esto no supone, en principio, una merma en la aplicabilidad de los resultados obtenidos, toda vez que las proposiciones 1 y 2 pueden generalizarse a distribuciones normales-gamma o normales-gamma invertidas, como se puede ver, p.e., en Girón et al. (1989) y en trabajos posteriores.

El resultado siguiente, cuya demostración es trivial, relaciona las distribuciones condicionadas de dos subvectores de una distribución normal multivariante. Obsérvese la (posible) dependencia lineal respecto de la variable condicionante de la media de las distribuciones condicionadas, así como la independencia de las matrices de covarianza. Sin estas restricciones la distribución conjunta no sería, obviamente, normal.

Proposición 1.— Si $X_1 | X_2 = x_2 \sim N(x_1 | Ax_2 + b; S)$ y $X_2 \sim N(x_2 | m; V)$, entonces la distribución de $X_2 | X_1 = x_1$ viene dada por

$$N(x_2 | m + VA'(S + AVA')^{-1}(x_1 - Am - b); V - VA'(S + AVA')^{-1} AV).$$

1.1. El modelo de regresión con errores normales

El resultado anterior está pensado para que se pueda aplicar de forma inmediata a problemas de inferencia bayesiana en poblaciones normales cuando la información a priori viene representada por una distribución normal.

En efecto, consideremos el modelo de regresión con errores normales, independientes e idénticamente distribuidos, dado por

$$y_i = x_i'\theta + u_i \quad \text{donde} \quad u_i | \sigma^2 \sim N(u_i | 0, \sigma^2); \quad i = 1, ..., n. \qquad (1.1)$$

Supongamos que la distribución a priori de θ es una normal $N(\theta | m_0; \sigma^2 C_0)$. Entonces si consideramos las dos distribuciones

$$y_1 | \theta \sim N(y_1 | x_1'\theta; \sigma^2) \quad \text{y} \quad \theta \sim N(\theta | m_0; \sigma^2 C_0),$$

estamos en las condiciones de la proposición 1, de modo que la distribución de θ condicionada por y_1 es una $N(\theta | m_1; \sigma^2 C_1)$, donde

$$m_1 = m_0 + \frac{1}{1 + x_1' C_0 x_1} C_0 x_1 (y_1 - x_1' m_0)$$

$$C_1 = C_0 - \frac{1}{1 + x_1' C_0 x_1} C_0 x_1 x_1' C_0.$$

De aquí podríamos conjeturar que la distribución a posteriori de θ condicionada por $y_1, ..., y_n$ es también una normal. En efecto, procediendo por inducción sobre n, si suponemos que la distribución a posteriori de θ condicionada por $y_1, ..., y_{n-1}$ es una $N(\theta | m_{n-1}; \sigma^2 C_{n-1})$, entonces como la distribución de $y_n | \theta, y_1, ..., y_{n-1} \equiv y_n | \theta$, por la hipótesis de independencia condicional, resulta que aplicando de nuevo la proposición 1, la distribución a posteriori de $\theta | y_1, ..., y_{n-1}, y_n$ es una normal $N(\theta | m_n; \sigma^2 C_n)$, donde los parámetros m_n y C_n vienen dados por las ecuaciones

$$m_n = m_{n-1} + \frac{1}{1 + x_n' C_{n-1} x_n} C_{n-1} x_n (y_n - x_n' m_{n-1})$$

$$C_n = C_{n-1} - \frac{1}{1 + x_n' C_{n-1} x_n} C_{n-1} x_n x_n' C_{n-1} . \tag{1.2}$$

A las ecuaciones anteriores se las conoce con el nombre de filtro de Plackett-Kalman y proporcionan un método computacionalmente eficiente de cálculo del vector de medias y de la matriz de covarianzas de los coeficientes de regresión, que no dependen del valor de σ^2, y que presenta además, dado su carácter secuencial, la ventaja de incorporar nueva información muestral sin necesidad de rehacer todos los cálculos.

Si, con las mismas hipótesis sobre el modelo de regresión y sobre la distribución a priori de θ dado σ^2, la varianza fuese desconocida y su distribución a priori fuese una gamma-invertida, $\sigma^2 \sim GaI(a_0, p_0)$, entonces el filtro anterior se puede ampliar con las ecuaciones siguientes

$$a_n = a_{n-1} + \frac{1}{2} \frac{(y_n - x_n' m_{n-1})^2}{1 + x_n' C_{n-1} x_n}$$

$$p_n = p_{n-1} + \frac{1}{2}, \tag{1.3}$$

de modo que la distribución a posteriori de $\sigma^2 | y_1, ..., y_n \sim GaI(a_n, p_n)$.

Otra alternativa al proceso anterior, que de nuevo resultaría de una simple aplicación de la proposición 1, es la de expresar el modelo (1.1) en la forma matricial usual, siendo $y = (y_1, ..., y_n)'$ y X la matriz de diseño formada por los vectores fila x_i',

$$y | \theta \sim N(y | X\theta; \sigma^2 I).$$

Entonces, si la distribución a priori de θ es una normal $N(\theta | m_0; \sigma^2 C_0)$, por la proposición 1 se tiene que la distribución a posteriori de $\theta | y$ es una normal

$$N(\theta | m_0 + C_0 X' (I + X C_0 X')^{-1} (y - X m_0);$$

$$\sigma^2 (C_0 - C_0 X' (I + X C_0 X')^{-1} X C_0).$$

Por el teorema de Bayes sabemos que esta distribución a posteriori coincide con la dada por las ecuaciones (1.2). Sin embargo, desde un punto de vista práctico, conviene señalar la diferencia entre ambas. En el caso de emplear la proposición 1 con toda la información muestral hay que invertir una matriz de dimensión $n \times n$, mientras que si se utiliza el filtro *no* es necesaria dicha inversión matricial. Además, de nuevo el carácter secuencial del filtro al estar basado de forma implícita en el teorema de Bayes, le hace independiente del orden en que aparecen los elementos de la muestra.

Otra posibilidad de analizar el modelo, que sería intermedia entre la de utilizar la muestra de una vez y de elemento en elemento (el filtro de Kalman), dentro de la estricta utilización de la proposición 1, es la de considerar la idea de suficiencia. En efecto, es bien conocido que la distribución a posteriori de $\boldsymbol{\theta}$ condicionada por toda la muestra coincide con la distribución condicionada por el estadístico suficiente.

El estadístico minimal suficiente es el estimador de máxima verosimilitud $\tilde{\boldsymbol{\theta}} = (X'X)^{-1} X'y$. Además su distribución dado $\boldsymbol{\theta}$ es una normal $N(\tilde{\boldsymbol{\theta}} | \boldsymbol{\theta}; \sigma^2 (X'X)^{-1})$, de modo que si a priori $\boldsymbol{\theta} \sim N(\boldsymbol{\theta} | \boldsymbol{m}_0; \sigma^2 \boldsymbol{C}_0)$, por la proposición 1 se tiene que

$$\boldsymbol{\theta} | \tilde{\boldsymbol{\theta}} \sim N(\boldsymbol{m}_0 + \boldsymbol{C}_0 ((X'X)^{-1} + \boldsymbol{C}_0)^{-1} (\tilde{\boldsymbol{\theta}} - \boldsymbol{m}_0);$$

$$\sigma^2 (\boldsymbol{C}_0 - \boldsymbol{C}_0 ((X'X)^{-1} + \boldsymbol{C}_0)^{-1} \boldsymbol{C}_0)).$$

Obsérvese cómo la utilización del estadístico suficiente ha reducido el cálculo de una inversa de orden $n \times n$ a una de dimensión fija (independiente del tamaño muestral) $k \times k$, siendo k la dimensión del vector $\boldsymbol{\theta}$.

Así pues conviene destacar los puntos de interés del desarrollo anterior, en particular lo que se refiere al filtro de Kalman y relacionarlo con otros conceptos de importancia en inferencia:

i) No se ha utilizado, de forma explícita, el teorema de Bayes para el cálculo de las distribuciones a posteriori. *Solamente la proposición 1.*
ii) No ha sido necesario de forma explícita mencionar, ni calcular ni reducir por suficiencia, la función de verosimilitud.
iii) La idea de familia conjugada, aunque está implícita en las propiedades de la normal multivariante, no se ha empleado en el desarrollo.
iv) La existencia de un estadístico suficiente para este modelo *no* se utiliza en la demostración del filtro (el procedimiento es secuencial y se utiliza un elemento de la muestra en cada etapa; no se reduce la información muestral por suficiencia).
v) El carácter secuencial del teorema de Bayes garantiza que el resultado final del filtro es *independiente* del orden de los elementos de la muestra y que coincide con los otros dos procedimientos basados en toda la muestra y en el estadístico suficiente.
vi) La deducción del filtro depende *exclusivamente* de la propiedad elemental 1, y de la hipóstesis de *independencia condicional* de los $\{u_i\}$ o los $\{y_i\}$ dado $\boldsymbol{\theta}$.
vii) Como valor añadido del filtro, comparado con los otros procedimientos, se tiene el que no es necesaria inversión matricial alguna y que la incorporación de nueva información muestral no requiere rehacer de nuevo los cálculos; basta aplicar de nuevo las ecuaciones del filtro.

1.2. El modelo de regresión con errores no normales

Consideremos ahora en el modelo de regresión (1.1) una estructura más compleja en la especificación de los errores. En particular consideremos los cuatro casos siguientes, recordando que la varianza es conocida y por lo tanto todas las distribuciones que consideramos se suponen condicionadas a σ^2.

a) $\{u_i\}$ son i.i.d. y siguen una distribución de Laplace o exponencial doble $u_i \sim La\,(u_i|0, \sigma^2)$; es decir, la densidad común de los errores viene dada por

$$f_L\,(u_i|\sigma^2) = \frac{1}{\sigma\sqrt{2}} \exp\left[-\frac{\sqrt{2}}{\sigma}\,|u_i|\right].$$

b) $\{u_i\}$ son i.i.d. y siguen una distribución t de Student con ν grados de libertad $u_i \sim St\,(u_i|0, \sigma^2; \nu)$; es decir, la densidad común de los errores viene dada por

$$f_t\,(u_i|0, \sigma^2; \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\sqrt{2\pi}\,\Gamma\,(\nu/2)} \left\{1 + \frac{1}{\nu\sigma^2}\,u_i^2\right\}^{-((\nu+1)/2)}$$

c) (Modelo conjunto de Box, Tiao y Abraham). $\{u_i\}$ son i.i.d. y siguen una distribución dada por la mixtura de normales siguiente

$$u_i \sim (1 - \lambda_0)\,n\,(u_i|0, \sigma^2) + \lambda_0 n\,(u_i|\delta, k^2\sigma^2).$$

d) $\{u_i\}$ son intercambiables y, condicionalmente en λ con $(\lambda > 0)$, se distribuyen i.i.d. como $N(u_i|0, \lambda\sigma^2)$, donde $\lambda \sim F(\lambda)$ y $F(\cdot)$ es una función de distribución arbitraria sobre $(0, \infty)$.

Los cuatro ejemplos anteriores tienen en común el que las distribuciones de los errores son mixturas de distribuciones normales. En los casos a), b) y d) la mixtura es respecto del parámetro de escala, mientras que en el caso c) tenemos una mixtura respecto de los parámetros de localizacion y escala.

El caso d) tiene un particular interés puesto que aunque los errores presentan simetría esférica no son independientes, salvo que $F(\cdot)$ sea degenerada; además en el caso de que existan los momentos de orden 2, los errores son incorrelados como es fácil comprobar.

El modelo de regresión con estructura de errores dada por a), b) y d) se puede escribir en forma de mixtura del modo siguiente:

a) $$y_i|\boldsymbol{\theta} \sim \int N\,(y_i|\mathbf{x}_i'\boldsymbol{\theta}; \lambda\sigma^2)\,dGa\,(\lambda; 1, 1);$$

b) $$y_i \mid \theta \sim \int N(y_i \mid x_i'\theta;\ \lambda\sigma^2)\, dGaI(\lambda;\ \frac{\nu}{2},\ \frac{\nu}{2});$$

d) $$y \mid \theta \sim \int N(y \mid X\theta;\ \lambda\sigma^2 I)\, dF(\lambda).$$

2. GENERALIZACION DE LA PROPOSICION 1 AL CASO DE MIXTURAS DE DISTRIBUCIONES NORMALES

En este apartado analizamos con cierto detalle la extensión de la proposición 1 a mixturas arbitrarias de distribuciones normales, que permitirá analizar modelos de regresión de los considerados al final de la sección anterior y otros más generales que se tratarán en la sección siguiente. Debido a la generalidad del resultado, el análisis exacto del filtro es complicado, dependiendo de la estructura del problema específico al que se aplique, de modo que proponemos en cada etapa del mismo una simplificación que conserva su estructura básica y lo hace particularmente atractivo desde el punto de vista computacional. En particular, uno de los modelos considerados en la sección anterior permite generalizar el teorema de Gauss-Markov a situaciones donde el criterio de los mínimos cuadrados no es, en principio directamente aplicable, obteniéndose además la distribución a posteriori exacta de los coeficientes de regresión.

El teorema siguiente es una generalización de la proposición 1.

Proposición 2.— Sea Λ un boreliano de $\mathbb{R}^m$, sean $A(\lambda)$, $S(\lambda)$, $b(\lambda)$ matrices medibles de dimensiones $k \times k$, $k \times k$ y $k \times 1$, respectivamente y sea $F(\lambda)$ una función de distribución sobre Λ. Si $X_1 \mid X_2 = x_2 \sim \sim \int N[x_1 \mid A(\lambda)\, x_2 + b(\lambda);\ S(\lambda)]\, dF(\lambda)$ y $X_2 \sim N(x_2 \mid m;\ V)$, entonces la distribución de $X_2 \mid X_1 = x_1$ viene dada por

$$\int N[x_2 \mid m(x_1, \lambda);\ V(\lambda)]\, dF(\lambda \mid x_1);$$

donde

$$m(x_1, \lambda) = m + VA'(\lambda)\,(S(\lambda) + A(\lambda)\, VA'(\lambda))^{-1}\,(x_1 - A(\lambda)\, m - b(\lambda))$$

$$V(\lambda) = V - VA'(\lambda)\,(S(\lambda) + A(\lambda)\, VA'(\lambda))^{-1}\, A(\lambda)\, V)$$

$$dF(\lambda \mid x_1) \propto n[x_1 \mid A(\lambda)\, m + b(\lambda);\ S(\lambda) + A(\lambda)\, VA'(\lambda)]\, dF(\lambda);$$

y $n[x_1 \mid \cdot;\ \cdot]$ representa la función de densidad de una normal multivariante.

Demostración.— En primer lugar demostraremos que la distribución conjunta del vector $X = (X_1, X_2)$ es también una mixtura de distribuciones normales con la misma distribución de mezcla $F(\lambda)$.

La función característica, $\varphi(t)$, del vector X se puede expresar como:

$$\varphi(t) = \varphi(t_1, t_2) = \int \exp\{it_2' x_2\}\, \varphi(t_1 | x_2)\, dF(x_2),$$

donde $\varphi(t_1 | x_2)$ es la función característica del vector $X_1 | X_2 = x_2$, es decir

$$\varphi(t_1 | x_2) = \int \exp\{i(A(\lambda) x_2 + b'(\lambda) t_1) - \frac{1}{2} t_1' S(\lambda) t_1\}\, dF(\lambda).$$

Sustituyendo en la expresión anterior, haciendo operaciones en el integrando y aplicando el teorema de Fubini, se tiene que

$$\varphi(t) = \int \left[\int \exp\{i[t_2' + t_1' A(\lambda) x_2]\}\, dF(x_2)\right] \exp\{ib'(\lambda) t_1 - \frac{1}{2} t_1' S(\lambda) t_1\}\, dF(\lambda).$$

Ahora bien, el término entre corchetes es la función característica del vector X_2 evaluada en $t_2' + t_1' A(\lambda)$. Sustituyendo y haciendo operaciones, finalmente se obtiene que la función característica $\varphi(t)$ es igual a

$$\int \exp\left[it' \begin{pmatrix} A(\lambda) m + b(\lambda) \\ m \end{pmatrix} - \frac{1}{2} t' \begin{pmatrix} S(\lambda) + A(\lambda) V A'(\lambda) & A(\lambda) V \\ V A(\lambda) & V \end{pmatrix} t\right] dF(\lambda).$$

De aquí se deduce, por la unicidad de la función característica, que la distribución conjunta de (X_1, X_2) es la siguiente mixtura

$$\int N\left[x \mid \begin{pmatrix} A(\lambda) m + b(\lambda) \\ m \end{pmatrix}; \begin{pmatrix} S(\lambda) + A(\lambda) V A'(\lambda) & A(\lambda) V \\ V A(\lambda) & V \end{pmatrix}\right] dF(\lambda);$$

y de aquí, por las propiedades de las distribuciones condicionadas y de la distribución normal multivariante, se sigue el teorema.

Como una primera aplicación de este resultado consideremos el caso d) de la sección anterior. La aplicación directa del teorema anterior, suponiendo que la distribución a priori de θ es una normal $N(\theta | m_0; \sigma^2 C_0)$, nos dice que la distribución de $\theta | y$ es la mixtura

$$\int N(\theta | m_0 + C_0 X' (\lambda I + X C_0 X')^{-1} (y - X m_0);$$
$$\sigma^2 (C_0 - C_0 X' (\lambda I + X C_0 X')^{-1} X C_0)\, dF(\lambda | y);$$

donde

$$dF(\lambda|y) \propto n\,[y|Xm_0;\ \sigma^2\,(\lambda I + XC_0X')]\,dF(\lambda);$$

que presenta la desventaja de tener que invertir, condicionalmente en λ, una matriz de dimensión $n \times n$.

Otra alternativa sería reducir el problema mediante la utilización del estadístico suficiente $\tilde{\theta}$, toda vez que su distribución condicionada a θ es también una mixtura de normales. En efecto, se puede demostrar (véase el lema 3.1 de Girón et al. (1989)), que

$$\tilde{\theta}|\theta \sim \int N(\tilde{\theta}|\theta;\ \lambda\sigma^2\,(X'X)^{-1})\,dF(\lambda);$$

y por la proposición 2, la distribución de $\theta|\tilde{\theta}$ es la mixtura

$$\int N(m_0 + C_0\,(\lambda\,(X'X)^{-1} + C_0)^{-1}\,(\tilde{\theta} - m_0);$$

$$C_0 - C_0\,(\lambda\,(X'X)^{-1} + C_0)^{-1}\,C_0)\,dF(\lambda|\tilde{\theta}); \qquad (2.1)$$

donde

$$dF(\lambda|\tilde{\theta}) \propto n\,(\tilde{\theta}|\theta;\ \sigma^2\,(\lambda\,(X'X)^{-1} + C_0)\,dF(\lambda).$$

O bien, por último, se podría trabajar condicionalmente en λ, aplicar el filtro de la sección anterior, calculando simbólicamente $m_n(\lambda)$ y $C_n(\lambda)$ mediante las ecuaciones

$$m_n(\lambda) = m_{n-1}(\lambda) + \frac{1}{\lambda + x_n'C_{n-1}(\lambda)\,x_n}\,C_{n-1}(\lambda)\,x_n\,(y_n - x_n'm_{n-1}(\lambda))$$

$$C_n(\lambda) = C_{n-1}(\lambda) - \frac{1}{\lambda + x_n'C_{n-1}(\lambda)\,x_n}\,C_{n-1}(\lambda)\,x_nx_n'C_{n-1}(\lambda);$$

con las condiciones iniciales $m_0(\lambda) = m_0$ y $C_0(\lambda) = C_0$, de manera que

$$\theta|y \sim \int N[\theta|m_n(\lambda);\ \sigma^2C_n(\lambda)]\,dF(\lambda|y).$$

Como en la sección 1, los tres procedimientos conducen a la misma distribución a posteriori.

Un caso particularmente interesante de lo anterior se presenta cuando no se tiene información a priori sobre θ. Entonces si tomamos como distribución a priori la no informativa, es decir, hacemos tender $C_0^{-1} \to O$, es fácil comprobar utilizando la ecuación (2.1) que la distribución a posteriori de referencia, condicionada a σ^2, es

$$\theta|y \sim \int N[\theta|(X'X)^{-1}\,X'y;\ \lambda\sigma^2\,(X'X)^{-1})\,dF(\lambda),$$

toda vez que $dF(\lambda|y) \propto dF(\lambda)$, como es fácil comprobar.

Este resultado es interesante, puesto que la distribución a posteriori resultante es una distribución elipsoidal (véase, p.e., Dickey and Chen (1985)), cuya moda es precisamente $\tilde{\theta}$, es decir, el estimador usual de mínimos cuadrados, que *no depende* de $F(\lambda)$, y que coincide con el vector de medias, caso de que éste exista, con lo cual se tiene una generalización del estimador de mínimos cuadrados, que es además robusto respecto de la forma de los errores u_t, siempre que éstos sean intercambiables y generados por una mixtura arbitraria $F(\lambda)$ respecto del parámetro de escala. Se tiene también que las M.D.P. de contenido probabilístico $1-\alpha$ son elipsoides de la forma $\{\theta \,|(\theta-\tilde{\theta})'\,(\sigma^2\,(X'X)^{-1})\,(\theta-\tilde{\theta}) \leqslant K_\alpha\}$, donde K_α se calcula a partir de la distribución $F(\lambda)$. Así, p.e., si $F(\lambda)$ fuese una $GaI(\lambda, \nu/2, \nu/2)$, es decir, los errores $\{u_t\}$ siguieran una t de Student multivariante (que serían incorrelados si $\nu > 2$) se tendría que la distribución a posteriori de $\theta|y$ sería una t de Student k-variante

$$St\,(\theta\,|\tilde{\theta},\, \sigma^2\,(X'X)^{-1};\, \nu)$$

y K_α se calcularía a partir de la $F(k, \nu)$ de Snedecor.

El caso c) se tratará en la sección 3; mientras, el análisis de los casos a) y b) es más complicado, ya que la forma exacta del filtro es generalmente intratable, debido entre otras causas a la no existencia de estadísticos suficientes. Lo que proponemos es una aproximación del filtro que sea computacionalmente sencilla y, a la vez, conserve aquellas propiedades que lo hacen tan atractivo, basado todo ello en la proposición 2.

2.1. El filtro continuo de Kalman

Los casos a) y b) se pueden considerar como casos particulares del modelo siguiente

$$y_t\,|\theta \sim \int N\,(y_t\,|x_t'\theta;\, \lambda\sigma^2)\, dF\,(\lambda).$$

Si, a priori θ se distribuye según una normal $N\,(\theta\,|\,\boldsymbol{m}_0;\, \sigma^2 \boldsymbol{C}_0)$, entonces, por la proposición 2, se tiene que

$$\theta\,|y_1 \sim \int N\,[\theta\,|\boldsymbol{m}_1\,(\lambda);\, \sigma^2 \boldsymbol{C}_1\,(\lambda)]\, dF\,(\lambda\,|y_1); \tag{2.2}$$

donde si

$$e_1 = y_1 - \boldsymbol{x}_1' \boldsymbol{m}_0$$

$$v_1\,(\lambda) = \lambda + \boldsymbol{x}_1' \boldsymbol{C}_0 \boldsymbol{x}_1$$

$$\boldsymbol{a}_1\,(\lambda) = \frac{\boldsymbol{C}_0 \boldsymbol{x}_1}{v_1\,(\lambda)}$$

entonces

$$\boldsymbol{m}_1(\lambda) = \boldsymbol{m}_0 + \boldsymbol{a}_1(\lambda)\, e_1$$

$$\boldsymbol{C}_1(\lambda) = \boldsymbol{C}_0 - \boldsymbol{C}_0 v_1(\lambda)\, \boldsymbol{a}_1(\lambda)\, \boldsymbol{a}_1'(\lambda)$$

$$dF(\lambda | y_1) \propto v_1(\lambda)^{-1/2} \exp\left[-\frac{e_1^2}{2\sigma^2 v_1(\lambda)}\right] dF(\lambda).$$

Como ya hemos comentado, el cálculo de la distribución exacta de $\boldsymbol{\theta} | y_1, y_2$ y subsiguientes distribuciones a posteriori, requeriría una generalización de la proposición 2, demasiado complicada para ser de utilización práctica. La idea que proponemos es aproximar la distribución exacta de $\boldsymbol{\theta} | y_1$ por una distribución normal con los mismos parámetros y proceder secuencialmente por aplicación repetida de la proposición 2, seguida de la aproximación correspondiente. Esta idea, aplicada dentro de un contexto similar al nuestro, puede encontrarse también en los trabajos de Girón et al. (1989) y Guttman y Peña (1985).

La aproximación normal, $N(\boldsymbol{\theta} | \boldsymbol{m}_1; \sigma^2 \boldsymbol{C}_1)$, a (2.2) viene dada por las ecuaciones

$$\boldsymbol{m}_1 = \int \boldsymbol{m}_1(\lambda)\, dF(\lambda | y_1)$$

$$\boldsymbol{C}_1 = \int \boldsymbol{C}_1(\lambda)\, dF(\lambda | y_1) + \frac{1}{\sigma^2} \int (\boldsymbol{m}_1(\lambda) - \boldsymbol{m}_1)(\boldsymbol{m}_1(\lambda) - \boldsymbol{m}_1)'\, dF(\lambda | y_1).$$

Si procedemos por inducción como en la sección 1, mediante la aplicación de la proposición 1, junto con la hipótesis de independencia condicionada y la aproximación de la mixtura resultante, se tiene el siguiente filtro aproximado:

Si

$$\begin{cases} e_n = y_n - \boldsymbol{x}_n' \boldsymbol{m}_{n-1} \\ v_n(\lambda) = \lambda + \boldsymbol{x}_n' \boldsymbol{C}_{n-1} \boldsymbol{x}_n \\ \boldsymbol{a}_n(\lambda) = \dfrac{\boldsymbol{C}_{n-1} \boldsymbol{x}_n}{v_n(\lambda)} \end{cases} \tag{2.3a}$$

$$\begin{cases} \boldsymbol{m}_n(\lambda) = \boldsymbol{m}_{n-1} + \boldsymbol{a}_n(\lambda)\, e_n \\ \boldsymbol{C}_n(\lambda) = \boldsymbol{C}_{n-1} - \boldsymbol{C}_{n-1} v_n(\lambda)\, \boldsymbol{a}_n(\lambda)\, \boldsymbol{a}_n'(\lambda), \end{cases} \tag{2.3b}$$

y

$$d\tilde{F}(\lambda | y_1, \ldots, y_n) \propto v_n(\lambda)^{-1/2} \exp\left[-\frac{e_n^2}{2\sigma^2 v_n(\lambda)}\right] dF(\lambda), \tag{2.3c}$$

entonces

$$\boldsymbol{\theta}\,|y_1, \ldots, y_n \approx \int N\,[\boldsymbol{\theta}\,|\boldsymbol{m}_n\,(\lambda);\ \sigma^2 \boldsymbol{C}_n\,(\lambda)]\ d\tilde{F}\,(\lambda|y_1, \ldots, y_n)$$

$$\approx N\,(\boldsymbol{\theta}\,|\boldsymbol{m}_n;\ \sigma^2 \boldsymbol{C}_n);$$

donde $\boldsymbol{m}_n$ y $\boldsymbol{C}_n$ vienen dados por las fórmulas

$$\boldsymbol{m}_n = \int \boldsymbol{m}_n\,(\lambda)\ d\tilde{F}\,(\lambda|y_1, \ldots, y_n)$$

$$\boldsymbol{C}_n = \int \boldsymbol{C}_n\,(\lambda)\ d\tilde{F}\,(\lambda|y_1, \ldots, y_n) + \tag{2.3d}$$

$$+ \frac{1}{\sigma^2} \int (\boldsymbol{m}_n\,(\lambda) - \boldsymbol{m}_n)\,(\boldsymbol{m}_n\,(\lambda) - \boldsymbol{m}_n)'\ d\tilde{F}\,(\lambda|y_1, \ldots, y_n);$$

y $\tilde{F}\,(\lambda|y_1, \ldots, y_n)$ representa la distribución a posteriori *aproximada* de λ dada la información muestral $\boldsymbol{y}$.

Obsérvese que la aplicación del filtro requiere en cada etapa el cálculo numérico de varias integrales: una para determinar la constante de proporcionalidad de la ecuación (2.3c) y varias para calcular la aproximación dada por (2.3d) que de hecho, como se puede demostrar con facilidad, se reducen al cálculo de dos tipos de integrales distintas. Caso de que la distribución $F\,(\lambda)$ tuviese un número finito de saltos, las integraciones anteriores se reducirían a simples sumas.

3. OTRAS APLICACIONES

En esta sección consideramos un caso particularmente importante de la proposición 2 que tiene aplicaciones a problemas estadísticos que ocurren con frecuencia en situaciones tan aparentemente dispersas como métodos robustos de inferencia, aprendizaje secuencial, clasificación de observaciones, diagnóstico automático, análisis de conglomerados y a problemas que hemos denominado de regresión-cluster. Debido también al carácter ilustrativo de este apartado, solamente consideraremos el caso de observaciones unidimensionales.

Consideremos la siguiente especialización de la proposición 2 al caso de observaciones $y_1, \ldots, y_n$ i.i.d. generadas por el modelo

$$y_i \sim \sum_{j=1}^{k} \lambda_j N\,(y_i|\boldsymbol{x}_i'\,(j)\ \boldsymbol{\theta};\ k_j^2\,\sigma^2) \tag{3.1}$$

en el cual se suponen conocidos los parámetros λ_j, los vectores de diseño $\boldsymbol{x}_i'\,(j)$ y los factores de heterocedasticidad k_j^2.

El modelo anterior se puede considerar como una generalización del modelo de regresión clásico en el sentido de que hay k posibles modelos de regresión heterocedásticos, con matrices de diseño distintas y cada uno de ellos tiene una probabilidad λ_j de generar los datos y_i.

Si la distribución a priori de $\boldsymbol{\theta}$ es una normal $N(\boldsymbol{\theta}|\boldsymbol{m}_0; \sigma^2\boldsymbol{C}_0)$ y en cada etapa del filtro se aproxima la mixtura resultante por una distribución normal tal como se hizo en la sección anterior, de la proposición 2 se tiene que si definimos recursivamente

$$\begin{aligned} e_n(j) &= y_n - \boldsymbol{x}_i'(j)\,\boldsymbol{m}_{n-1} \\ v_n(j) &= k_j^2 + \boldsymbol{x}_i'(j)\,\boldsymbol{C}_{n-1}\boldsymbol{x}_i(j) \\ \boldsymbol{a}_n(j) &= \frac{\boldsymbol{C}_{n-1}\boldsymbol{x}_i(j)}{v_n(j)} \end{aligned} \tag{3.2a}$$

y

$$\begin{aligned} \boldsymbol{m}_n(j) &= \boldsymbol{m}_{n-1} + \boldsymbol{a}_n(j)\,e_n(j) \\ \boldsymbol{C}_n(j) &= \boldsymbol{C}_{n-1} - \boldsymbol{C}_{n-1}v_n(j)\,\boldsymbol{a}_n(j)\,\boldsymbol{a}_n'(j) \\ \lambda_n(j) &\propto \lambda_j v_n(j)^{-1/2} \exp\left[-\frac{e_n(j)^2}{2\sigma^2 v_n(j)}\right] \end{aligned} \tag{3.2b}$$

entonces la distribución a posteriori aproximada de $\boldsymbol{\theta}$ es

$$\boldsymbol{\theta}|y_1, \ldots, y_n \approx \sum_{j=1}^{k} \lambda_n(j)\,N(\boldsymbol{m}_n(j); \sigma^2\boldsymbol{C}_n(j)).$$

Esta distribución a posteriori puede ser de nuevo aproximada por una $N(\boldsymbol{m}_n; \sigma^2\boldsymbol{C}_n)$, donde los nuevos parámetros se obtienen de las ecuaciones siguientes

$$\begin{aligned} \boldsymbol{m}_n &= \sum_{j=1}^{k} \lambda_n(j)\,\boldsymbol{m}_n(j) \\ \boldsymbol{C}_n &= \sum_{j=1}^{k} \lambda_n(j)\,\boldsymbol{C}_n(j) + \frac{1}{\sigma^2}\sum_{j=1}^{k} \lambda_n(j)\,(\boldsymbol{m}_n(j) - \boldsymbol{m}_n)(\boldsymbol{m}_n(j) - \boldsymbol{m}_n)'. \end{aligned} \tag{3.2c}$$

El conjunto de las ecuaciones (3.2), (3.2a), (3.2b) y (3.2c) constituye la base del filtro aproximado para el modelo (3.1), del que vamos a considerar brevemente algunos casos particulares.

1. El modelo de regresión lineal homocedástico, considerado en la introducción, se obtiene como caso particular tomando $k = 1$ y en este caso el filtro es exacto ya que $F(\lambda)$ es degenerada y no hay necesidad de aproximaciones.

2. El modelo de regresión contaminado con posibilidad de desplazamientos (parámetro δ) y/o heterocedasticidad (parámetro ω^2), caso c) de la sección 1, que es una generalización de los modelos propuestos por Box and Tiao (1968) y Abraham and Box (1978).

Condicionado a β, δ y σ^2, las observaciones $\{y_i\}$ son i.i.d. y se distribuyen según

$$y_i \sim (1 - \lambda_0)\, N(y_i | x'\beta;\, \sigma^2) + \lambda_0 N(y_i | x'\beta + \delta;\, \omega^2 \sigma^2).$$

Basta tomar $k = 2$, $\theta = (\beta, \delta)$; $k_1^2 = 1$, $k_1^2 = \omega^2$ y

$$x_i'(1) = (x_i', 0) \quad x_i'(2) = (x_i', 1) \quad \text{para todo } i = 1, ..., n.$$

3. El análisis de conglomerados o "clusters" unidimensionales basado en mixturas de distribuciones normales (Fig. 1) es un modelo probabilístico para la descripción de datos que presentan una estructura de subgrupos más o menos dispersos (véase, p.e., el reciente libro de McLachan and Basford (1988)). Lo podemos considerar un caso particular del modelo (3.1), simplemente tomando $x_i'(j) = e_j'$, para todo $i = 1, ..., n$, $j = 1, ..., k$, donde e_j es el vector unitario con la j-ésima coordenada igual a 1.

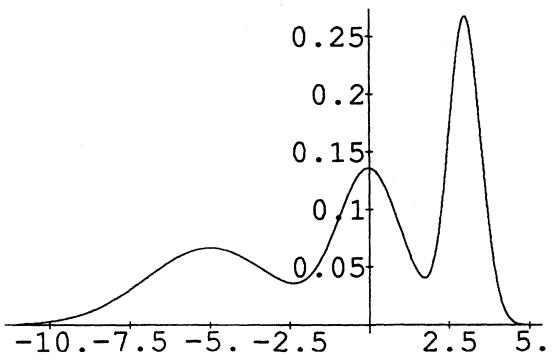

Fig. 1. *Una mixtura de tres distribuciones normales.*

Este modelo no es del todo general puesto que las probabilidades de pertenencia a cada conglomerado λ_j y los factores de heterocedasticidad k_j^2 se suponen conocidos. Para un desarrollo más general de este problema, que tiene en cuenta el desconocimiento de estos parámetros, dentro de un contexto distinto aunque similar al presente, véase Bernardo y Girón (1989). Un problema importante en la especificación de este modelo es el de la distribución a priori sobre los parámetros θ que parece natural considerar intercambiables, con lo que introduciríamos un submodelo jerárquico en los parámetros del modelo que se trataría de manera similar a la del filtro considerado.

4. Detección y clasificación de señales (problemas de aprendizaje secuencial). Estos problemas se pueden modelar dentro del contexto de mixturas de distribuciones normales que estamos considerando (véase, p.e., Makov (1980)) y se centran, básicamente, en los problemas de clasificación probabilística de las señales y en los de aprendizaje sobre el o los parámetros del modelo. A modo de ilustración vamos a considerar dos casos particulares:

4.1. Modelo de señal-ruido: las observaciones $\{y_i\}$ son i.i.d. y se distribuyen según

$$y_i \sim (1 - \lambda_0) N(y_i|\theta; \sigma^2) + \lambda_0 N(y_i|0; \sigma^2).$$

Basta tomar $k = 2$, $x_i'(1) = 1$, $x_i'(2) = 0$, para todo $i = 1, ..., n$, y $k_1^2 = k_2^2 = 1$.

4.2. Modelo bipolar: las observaciones $\{y_i\}$ son i.i.d. y se distribuyen según

$$y_i \sim (1 - \lambda_0) N(y_i|\theta; \sigma^2) + \lambda_0 N(y_i|-\theta; \sigma^2).$$

Basta tomar $k = 2$, $x_i'(1) = 1$, $x_i'(2) = -1$, para todo $i = 1, ..., n$, y $k_1^2 = k_2^2 = 1$.

5. El modelo de regresión-cluster es básicamente el modelo general (3.1) sin suponer restricciones en las matrices de diseño X_j. Un ejemplo ilustrativo de este modelo es la situación descrita por los datos de la figura 2.

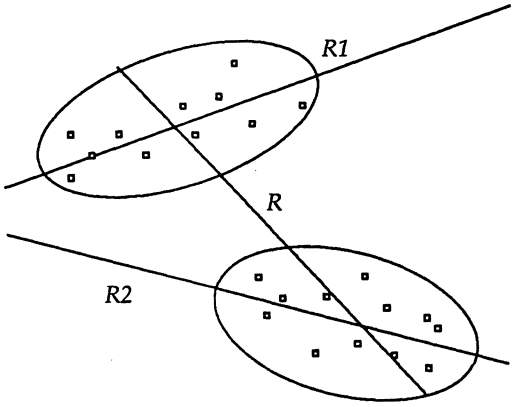

Fig. 2. *Datos generados por un modelo de "regresión-cluster".*

Un análisis de regresión estándar de estos datos conduciría a la recta de regresión R y a una sobreestimación de la varianza del modelo, con la consecuente falta de capacidad estimativa y predictiva del mismo. Está claro, de la estructura de los datos, que hay dos regresiones subyacentes distintas R_1 y R_2, y que un modelo como el (3.1) proporcionaría una mejor descripción de los datos que, a su vez, tendría un mayor poder predictivo que el modelo de regresión simple.

Así pues, el modelo sería el siguiente:

$$y_i \sim (1 - \lambda_0) N(y_i|\alpha_1 + \beta_1 x_i; \sigma^2) + \lambda_0 N(y_i|\alpha_2 + \beta_2 x_i; \omega^2 \sigma^2)$$

para lo cual basta tomar $k = 2$; $k_1^2 = 1$, $k_2^2 = \omega^2$, $\theta' = (\alpha_1, \beta_1, \alpha_2, \beta_2)$ y

$$x_i'(1) = (1, x_i, 0, 0)$$

$$x_i'(2) = (0, 0, 1, x_i).$$

4. COMENTARIOS

En las secciones anteriores hemos considerado, desde un punto de vista estadístico, el filtro de Kalman y sus generalizaciones todo dentro del contexto de los modelos estáticos a los que es directamente aplicable la proposición 2 y que generalizan los modelos clásicos de regresión en varias direcciones. Este enfoque pone de manifiesto la simplicidad y elasticidad de las herramientas utilizadas en la metodología bayesiana y su capacidad de abordar, aunque a veces sea utilizando métodos aproximados, problemas de inferencia complejos mediante la generalización de ciertas propiedades de la distribución normal a familias generadas a partir de ella por el procedimiento de mixturas.

Hay sin embargo situaciones donde los parámetros del modelo evolucionan a lo largo del tiempo, como son en particular los denominados *modelos lineales dinámicos*, para cuyo análisis se necesitan otras propiedades de la normal multivariante, que sean a su vez susceptibles de ser generalizadas a mixturas de distribuciones normales.

En particular, la deducción del filtro de Kalman en los modelos de Harrison and Stevens (1976) y de Snyder (1985) se basa también en una sencilla propiedad de los vectores normales, como es la de que transformaciones lineales de vectores normales independientes son tambien normales. Este resultado puede extenderse al caso de mixturas de vectores normales, como puede verse en el lema 3.1 de Girón et al. (1989), para un caso particular, y en Rojano (1989) para mixturas arbitrarias. El estudio de estos problemas y otros relacionados con la estimación de los parámetros de estos modelos (que incluyan, p.e., la varianza y el vector de efectos fijos), serán objeto de posteriores estudios.

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.

BIBLIOGRAFIA

[1] ABRAHAM, B. AND BOX, G. E. P. *Linear models and spurious observations.* Appl. Statist., 27, 120-130, 1978.

[2] BERNARDO, J. M. AND GIRON, F. J.: *A Bayesian approach to cluster analysis.* Qüestiió, vol 12, 1, 97-112, 1988.

[3] BOX, G. E. P. AND TIAO, G. C.: *A Bayesian approach to some outlier problems.* Biometrika, 55, 119-129, 1968.

[4] DICKEY, J. M. AND CHEN, C-H.: *Direct subjective-probability modelling using ellipsoidal distributions* (with discussion). In Bayesian Statistics 2 (J. M. Bernardo, D. V. Lindley, M. H. De Groot and A. F. M. Smith, eds.), pp. 157-182. North Holland: Amsterdam, 1985.

[5] GIRON, F. J., MARTINEZ, M. L. Y ROJANO, J. C.: *Modelos lineales dinámicos y mixturas de distribuciones* (artículo invitado con discusión). Estadist. Española 31, 1989.

[6] GUTTMAN, I. AND PEÑA, D.: *Robust filtering.* J. Amer. Statist. Ass., 80, 91-92, 1985.

[7] HARRISON, P. J. AND STEVENS, C. F.: *Bayesian forecasting* (with discussion). J. Roy. Statist. Soc. B, 38, 205-247, 1976.

[8] MCLACHLAN, G. J. AND BASFORD, K. E.: *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker: New York, 1988.

[9] MAKOV, U. E.: *Approximations of unsupervised Bayes learning procedures* (with discussion). In Bayesian Statistics 1 (J. M. Bernardo, D. V. Lindley, M. H. DeGroot and A. F. M. Smith, eds.), pp. 69-81. Valencia: University Press, 1980.

[10] ROJANO, J. C. (sin publicar): *Métodos bayesianos aproximados para mixturas de distribuciones.* (Tesis doctoral.)

[11] SNYDER, R. D.: *Recursive estimation of dynamic linear models.* J. Roy Statist. Soc. B, 47, 272-276, 1985.