ESTADISTICA ESPAÑOLA
Vol. 31, Núm. 121, 1989, págs. 165 a 206

Modelos lineales dinámicos y mixturas de distribuciones

por
F. J. GIRON, M. L. MARTINEZ y J. C. ROJANO
Departamento de Matemáticas (U. D. de Estadística)
Facultad de Ciencias
Campus de Teatinos s/n
Universidad de Málaga
29071 - MALAGA

RESUMEN

Uno de los problemas prácticos a la hora de modelizar una serie temporal mediante un modelo lineal dinámico es la especificación de las fuentes de error asociadas a las ecuciones de observación y del sistema. El modelo de Snyder (1985) solamente incluye una fuente de variabilidad, reduciendo así el problema de especificación al del vector de efectos permanentes. Tomando este modelo como punto de partida, en este artículo consideramos dos extensiones importantes del modelo, una que permite estimar la varianza común de los términos de error y otra, más importante, que permite considerar errores no normales. De la infinidad de posibles modelos para los términos de error, consideramos el caso importante del modelo multiplicativo, siendo inmediata la extensión de estos resultados a una amplia clase de modelos. Mediante técnicas de aproximación de las distribuciones asociadas al modelo, mixturas de ciertas distribuciones, se generaliza el filtro de Kalman, manteniendo su estructura básica sin aumentar sustancialmente el esfuerzo computacional.

Palabras clave: aproximación de mixturas: filtro de Kalman robusto; modelos lineales dinámicos; mixturas de distribuciones; observaciones anómalas.

Clasificación AMS: 62M20, 62F15.

1. INTRODUCCION

El modelo lineal dinámico de Harrison y Stevens (1976) dado por las ecuaciones

$$\text{Ecuación de observación:} \quad y_t = F_t\theta_t + v_t, \quad \{v_t \sim N(0, V_t)\}$$

$$\text{Ecuación del sistema:} \quad \theta_t = G\theta_{t-1} + w_t, \quad \{w_t \sim N(0, W_t)\},$$

aunque muy general en su formulación y que constituye una alternativa a otros modelos de series temporales, presenta la desventaja, como han señalado Harrison y Ameen (1985) y Snyder (1985), de la necesidad de especificar en cada instante de tiempo las matrices de covarianzas asociadas a las *ecuaciones de observación y del sistema,* V_t y W_t, respectivamente.

En el mencionado artículo de Snyder, éste argumenta que, al menos en el tratamiento de las series temporales univariantes, es suficiente el considerar desde la perspectiva de los MLD (modelos lineales dinámicos), una única fuente de variabilidad que sirve a la vez para describir los errores en la ecuación de observación así como en la ecuación del sistema a través de lo que denomina *vector de efectos permanentes*. Esta formulación, aunque aparentemente más restrictiva, contiene como caso particular a muchos de los modelos clásicos de series temporales, en particular a los modelos *ARMA* de Box y Jenkins (1970), y su tratamiento se hace a través de una versión del filtro de Kalman para el caso en que los términos de error de la ecuación de observación sean normales, independientes, idénticamente distribuidos de media 0 y varianza conocida. Así el modelo de Snyder constituye el punto de partida para el presente artículo, que generalizaremos en las secciones siguientes, al menos en dos direcciones.

En la sección 2 tratamos el caso más general de suponer la varianza común desconocida, generalizando así el filtro de Kalman mediante la inclusión de dos nuevas ecuaciones recurrentes que permiten actualizar los parámetros que describen la distribución a posteriori de la varianza. La deducción del filtro se realiza simplemente calculando las distribuciones a posteriori correspondientes, a diferencia de la de Snyder que se basa en minimizar el error cuadrático medio del error de la predicción de una observación futura.

A continuación, en la sección 3, y siguiendo una sugerencia de Guttman y Peña (1985) a un artículo de West, Harrison y Migon (1985), generalizamos el modelo de Snyder al caso en que los errores no sigan necesariamente una ley normal. De los muchos modelos probabilísticos posibles

consideramos el caso más interesante en que el error sigue un modelo multiplicativo, ésto es, una mixtura de una normal $N(0, \sigma^2)$ y una $N(0, k^2\sigma^2)$ con probabilidades, bien conocidas o desconocidas, $1 - \lambda$ y λ, respectivamente. Este modelo, debido a Box y Tiao (1968), es bien conocido para el tratamiento de observaciones anómalas (*outliers*) en regresión y, desde nuestra perspectiva, podría ser un enfoque interesante al difícil problema de determinación de observaciones anómalas en series temporales. De hecho, mediante una técnica de aproximación que se trata en la sección siguiente, se extiende el filtro de Kalman a esta situación preservando su carácter recurrente y su facilidad de cálculo. El método proporciona además probabilidades a posteriori aproximadas *on line* de que la observación correspondiente sea anómala. Si se dispone de toda la serie temporal, éstas probabilidades pueden revisarse mediante la técnica del suavizado (*smoothing*).

En estas dos secciones, y con la finalidad de mantener las demostraciones dentro del contexto general del artículo, en el sentido de no utilizar la función de verosimilitud de las mixturas que aparecen, que son complejas debido a la naturaleza de las mismas, se dan en la sección 3 tres lemas que simplifican considerablemente la demostración y presentación de los resultados.

Nuestro modelo es susceptible de incorporar otros modelos para los errores, como pudiera ser el modelo aditivo de Abraham y Box (1978) para el tratamiento de observaciones anómalas en modelos dinámicos, así como a modelos más complejos como pudieran ser las mixturas de normales respecto del parámetro de escala que incluiría distribuciones como la t de Student (véase, p.e., West (1981)) y a la familia exponencial de potencias de Box y Tiao (1973) (véase, p.e., West (1984)), siguiendo una metodología análoga a la que presentamos. Este presenta además la ventaja de la economía, no sólo en la descripción de modelos para incorporar observaciones anómalas sino en el aspecto computacional, sobre los modelos multiproceso de Harrison y Stevens (loc. cit.). Estas extensiones, así como el caso de datos multivariantes, serán objeto de un posterior trabajo.

En la sección 4, y como consecuencia de los resultados de la anterior, se analiza el problema de la aproximación de ciertas mixturas de distribuciones y se dan resultados específicos y se comparan con otros métodos de aproximación.

Por último, en la sección 5, se comentan y discuten ciertos aspectos del modelo de Snyder (sus ventajas, sus limitaciones, sus generalizaciones y otras posibles alternativas).

2. ANALISIS BAYESIANO DEL MODELO DE SNYDER: EXTENSIONES

El modelo original de Snyder viene dado por las ecuaciones siguientes

$$y_t = \mathbf{x}_t'\theta_t + u_t \tag{2.1a}$$

$$\theta_{t+1} = \mathbf{T}\theta_t + \alpha u_t; \tag{2.1b}$$

donde $y_1, \ldots, y_t$ es una serie de variables aleatorias observables, $u_1, \ldots, u_t$ son errores aleatorios independientes normalmente distribuidos con media 0 y varianza σ^2, θ_t es el vector $(k \times 1)$ de parámetros del proceso en el tiempo t, $\mathbf{x}_t$ es un vector $(k \times 1)$ de variables independientes, *conocido en el instante t*, $\mathbf{T}$ es una matriz de transición $(k \times k)$ fija (aunque puede hacerse que dependa del tiempo sin modificaciones sustanciales en los resultados) y α es un vector $(k \times 1)$ fijo que, siguiendo a Snyder, denominaremos *vector de efectos permanentes.*

El resultado siguiente establece el filtro de Kalman para los parámetros del proceso. Es importante hacer notar que, con las hipótesis del teorema, el filtro es independiente del valor de la varianza σ^2.

Teorema 2.1. *Si* $\theta_1 \mid \sigma^2 \sim N(\mathbf{m}_1, \sigma^2\mathbf{C}_1)$ *independientemente de* $u_1, \ldots, u_n, \ldots$, *entonces para todo t la distribución a posteriori de*

$$\theta_{t+1} \mid \sigma^2, y_1, \ldots, y_t \sim N(\mathbf{m}_{t+1}; \sigma^2\mathbf{C}_{t+1}),$$

donde los parámetros se calculan a partir de las relaciones siguientes: si

$$\begin{aligned} \hat{y}_t &= \mathbf{x}_t'\mathbf{m}_t \\ e_t &= y_t - \hat{y}_t \\ \mathbf{B}_t &= \mathbf{T}\mathbf{C}_t\mathbf{T}' + \alpha\alpha' \\ v_t &= \mathbf{x}_t'\mathbf{C}_t\mathbf{x}_t + 1 \\ \mathbf{a}_t &= (\mathbf{T}\mathbf{C}_t\mathbf{x}_t + \alpha) / v_t, \end{aligned} \tag{2.2}$$

entonces

$$\begin{aligned} \mathbf{m}_{t+1} &= \mathbf{T}\mathbf{m}_t + \mathbf{a}_t e_t \\ \mathbf{C}_{t+1} &= \mathbf{B}_t - v_t\mathbf{a}_t\mathbf{a}_t'. \end{aligned} \tag{2.3}$$

La demostración de este resultado se realiza por inducción. Supongamos que el teorema fuese cierto hasta el instante t, es decir

$$\theta_t \mid \sigma^2, y_1, \ldots, y_{t-1} \sim N(\mathbf{m}_t; \sigma^2\mathbf{C}_t).$$

Si consideramos ahora las dos ecuaciones del modelo como una sola ecuación, éste puede escribirse como

$$\begin{pmatrix} y_t \\ \theta_{t+1} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_t' & 1 \\ \mathbf{T} & \alpha \end{pmatrix} \begin{pmatrix} \theta_t \\ u_t \end{pmatrix}. \tag{2.4}$$

Ahora, por la hipótesis de inducción, la independencia condicional de θ_t y u_t dado $y_1, \ldots, y_{t-1}$ y las propiedades de la distribución normal multivariante, se tiene que

$$\begin{pmatrix} y_t \\ \theta_{t+1} \end{pmatrix} \mid \sigma^2, y_1, \ldots, y_{t-1} \sim N\left[\begin{pmatrix} \mathbf{x}_t' \mathbf{m}_t \\ \mathbf{T}\,\mathbf{m}_t \end{pmatrix} ; \sigma^2 \begin{pmatrix} \mathbf{x}_t' \mathbf{C}_t \mathbf{x}_t + 1 & \mathbf{x}_t' \mathbf{C}_t \mathbf{T}' + \alpha' \\ \mathbf{T}\,\mathbf{C}_t \mathbf{x}_t + \alpha & \mathbf{T}\,\mathbf{C}_t \mathbf{T}' + \alpha\alpha' \end{pmatrix} \right].$$

De aquí se sigue que la distribución de θ_{t+1} condicionada a $\sigma^2, y_1, \ldots, y_{t-1}, y_t$ es, por las propiedades de la normal multivariante,

$$\theta_{t+1} \mid \sigma^2, y_1, \ldots, y_{t-1}, y_t \sim N(\mathbf{m}_{t+1}; \sigma^2 \mathbf{C}_{t+1}),$$

donde

$$\mathbf{m}_{t+1} = \mathbf{T}\mathbf{m}_t + [\mathbf{T}\mathbf{C}_t\mathbf{x}_t + \alpha] (\mathbf{x}_t' \mathbf{C}_t \mathbf{x}_t + 1)^{-1} (y_t - \mathbf{x}_t' \mathbf{m}_t)$$

$$\mathbf{C}_{t+1} = [\mathbf{T}\mathbf{C}_t\mathbf{T}' + \alpha\alpha'] - [\mathbf{T}\mathbf{C}_t\mathbf{x}_t + \alpha] (\mathbf{x}_t' \mathbf{C}_t \mathbf{x}_t + 1)^{-1} [\mathbf{x}_t' \mathbf{C}_t \mathbf{T}' + \alpha'],$$

que son equivalentes a las ecuaciones del filtro. Al vector $\mathbf{a}_t$ se le conoce con el nombre de *tasa de respuesta* o *constante suavizadora*.

A partir del teorema anterior pueden calcularse fácilmente predicciones de observaciones futuras, siguiendo un procedimiento análogo al de Harrison y Stevens. En particular, la distribución de una observación futura y_{t+1} dadas las observaciones $y_1, \ldots, y_t$ viene dada por la distribución predictiva de $y_{t+1} \mid \sigma^2, y_1, \ldots, y_t$ que se obtiene a partir de la ecuación de observación (2.1a) para el instante $t+1$

$$y_{t+1} = \mathbf{x}_{t+1}' \theta_{t+1} + u_{t+1}.$$

De aquí se sigue, por la independencia de θ_{t+1} y u_{t+1} dados $\sigma^2, y_1, \ldots, y_t$, que

$$y_{t+1} \mid \sigma^2, y_1, \ldots, y_t \sim N(\mathbf{x}_{t+1}' \mathbf{m}_{t+1}; \sigma^2 \{\mathbf{x}_{t+1}' \mathbf{C}_{t+1} \mathbf{x}_{t+1} + 1\}).$$

El caso en que σ^2 sea desconocida, requiere la especificación de una distribución a priori sobre el par (θ_1, σ^2). Si se elige convenientemente la distribución a priori, ésta se va propagando a lo largo del tiempo tal como sucede con las familias conjugadas en los modelos estadísticos estáticos. La estructura del modelo sugiere, y de hecho así sucede como demuestra el teorema que a continuación sigue, elegir una distribución normal-gamma invertida.

Definición 2.1. Si $\mathbf{x}$ es un vector aleatorio k dimensional y v es una variable aleatoria no negativa, se dice que el par $(\mathbf{x}, v)$ sigue una distribución normal-gamma invertida de parámetros $\mathbf{m}$, $\mathbf{C}$, a y p, que representaremos abreviadamente por

$$(\mathbf{x}, v) \sim NGal(\mathbf{m}, \mathbf{C}; a, p)$$

si $\mathbf{x} \mid v \sim N(\mathbf{m}, v\mathbf{C})$ y $v \sim Gal(a, p)$, donde la densidad de la distribución gamma invertida de parámetros a y p, viene dada por

$$g(v; a, p) = \frac{a^p}{\Gamma(p)} v^{-(p+1)} e^{-a/v}.$$

Teorema 2.2. *Si* $(\theta_1, \sigma^2) \sim NGal(\mathbf{m}_1, \mathbf{C}_1; a_1, p_1)$ y dado σ^2, θ_1 es *independiente de los errores* $u_1, \ldots, u_n, \ldots$, entonces para todo t la distribución a posteriori de

$$\theta_{t+1}, \sigma^2 \mid y_1, \ldots, y_t \sim NGal(\mathbf{m}_{t+1}, \mathbf{C}_{t+1}; a_{t+1}, p_{t+1}),$$

donde los parámetros $\mathbf{m}_{t+1}$ y $\mathbf{C}_{t+1}$ *vienen dados por el teorema anterior y los nuevos parámetros* a_{t+1} *y* p_{t+1} *se obtienen recursivamente de las fórmulas*

$$\begin{aligned} p_{t+1} &= p_t + \frac{1}{2} \\ a_{t+1} &= a_t + \frac{1}{2} e_t^2 / v_t. \end{aligned} \tag{2.5}$$

Obsérvese que la solución de las ecuaciones (2.5) es inmediata y viene dada por

$$\begin{aligned} p_{t+1} &= p_1 + \frac{t}{2} \\ a_{t+1} &= a_1 + \frac{1}{2} \sum_{i=1}^{t} e_i^2 / v_i. \end{aligned} \tag{2.5'}$$

La demostración se hace como en el teorema anterior por inducción. Si la distribución de

$$\theta_t, \sigma^2 \mid y_1, \ldots, y_{t-1} \sim NGal(\mathbf{m}_t, \mathbf{C}_t; a_t, p_t),$$

entonces la distribución de

$$\theta_t \mid \sigma^2, y_1, \ldots, y_{t-1} \sim N(\mathbf{m}_t; \sigma^2 \mathbf{C}_t).$$

Por el teorema anterior, la distribución de $\theta_{t+1} \mid \sigma^2$ sería

$$\theta_{t+1} \mid \sigma^2, y_1, \ldots, y_t \sim N(\mathbf{m}_{t+1}; \sigma^2 \mathbf{C}_{t+1}),$$

donde los parámetros $\mathbf{m}_{t+1}$ y $\mathbf{C}_{t+1}$ vendrían dados por el teorema anterior. Sólo queda por demostrar que la distribución de σ^2 condicionada por

$y_1, \ldots, y_t$ es una gamma invertida de parámetros a_{t+1} y p_{t+1} dados por las ecuaciones (2.5). En efecto, se tiene, por el teorema de Bayes, que

$$p(\sigma^2 \mid y_1, \ldots, y_t) \propto p(\sigma^2 \mid y_1, \ldots, y_{t-1}) \cdot f(y_t \mid \sigma^2, y_1, \ldots, y_{t-1}),$$

es decir

$$p(\sigma^2 \mid y_1, \ldots, y_t) \propto (\sigma^2)^{-1/2} \exp\left\{-\frac{e_t^2}{2\sigma^2 v_t}\right\} \cdot (\sigma^2)^{-(p_t+1)} \exp\left\{-\frac{a_t}{\sigma^2}\right\}$$

$$\propto (\sigma^2)^{-(p_t+1/2+1)} \exp\left\{-\frac{1}{\sigma^2}\left[a_t + \frac{1}{2} e_t^2 / v_t\right]\right\},$$

que es proporcional a la densidad de una gamma invertida de parámetros a_{t+1} y p_{t+1}. Esto es equivalente a que a posteriori, es decir condicionado a $y_1, \ldots, y_t$, $a_{t+1} / \sigma^2 \sim \chi^2(p_{t+1})$.

Del teorema precedente es fácil obtener, *p. e., la distribución marginal de* θ_{t+1} dados los datos $y_1, \ldots, y_t$, que es una *t* de Student *k*-variante

$$\theta_{t+1} \mid y_1, \ldots, y_t \sim St(2p_{t+1}; \mathbf{m}_{t+1}; (a_{t+1} / p_{t+1}) \mathbf{C}_{t+1}),$$

con $2p_{t+1}$ grados de libertad, vector de centralización $\mathbf{m}_{t+1}$ y matriz de precisión $(p_{t+1} / a_{t+1}) \mathbf{C}_{t+1}^{-1}$.

De modo análogo se puede obtener la distribución predictiva de una observación futura. De ser $y_{t+1} \mid \sigma^2, y_1, \ldots, y_t \sim N(\hat{y}_{t+1}; \sigma^2 v_{t+1})$, por el teorema anterior y $\sigma^2 \mid y_1, \ldots, y_t \sim Gal(a_{t+1}, p_{t+1})$, por el resultado que acabamos de probar, se deduce que la distribución marginal de $y_{t+1} \mid y_1, \ldots, y_t$ es una *t* de Student univariante

$$y_{t+1} \mid y_1, \ldots, y_t \sim St(2p_{t+1}; \hat{y}_{t+1}, (a_{t+1} / p_{t+1}) v_{t+1}).$$

Está claro, dada la naturaleza recursiva del filtro de Kalman, que éste proporciona la distribución del parámetro θ_{t+1} dada la información $y_1, \ldots, y_t$ hasta el tiempo *t*. Sin embargo, también está claro que datos posteriores en el tiempo $y_{t+1}, \ldots$ proporcionan información sobre $\theta_1, \ldots, \theta_{t+1}$. Esta técnica, la de obtener la distribución de cualquier parámetro del modelo utilizando todos los datos, conocida con el nombre de suavizamiento, puede realizarse definiendo un nuevo MLD (aunque de dimensión creciente linealmente con el tiempo) en función de la sucesión de todos los parámetros del modo siguiente:

Si definimos $\phi_{t+1} = (\theta_{t+1}, \theta_t, \ldots, \theta_1)$, basta considerar el MLD dado por

$$y_t = (\mathbf{x}_t', 0, \ldots, 0)\ \phi_t + u_t$$

$$\theta_{t+1} = \begin{pmatrix} \mathbf{T} & \mathbf{0} \dots \mathbf{0} \\ \hline \mathbf{I} & \end{pmatrix} \theta_t + \begin{pmatrix} \alpha \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{pmatrix} u_t; \qquad (2.6)$$

que es también un modelo del tipo considerado por Snyder, donde la fuente de variación es la del modelo original, es decir u_t y el vector de efectos permanentes es el original aumentado con la inclusión de un número creciente de vectores nulos.

3. EXTENSIONES DEL MODELO: ERRORES NO NORMALES

Como comentabamos en la introducción, la importancia del modelo de Snyder es la de incluir sólamente una única fuente de variabilidad tanto en la ecuación de observación como en la del sistema, ésta última modificada por el vector de efectos permanentes α. De este modo es fácil generalizar el filtro propuesto por Guttman y Peña (1985) en las dos direcciones siguientes: considerar errores no normales en ambas ecuaciones y considerar el caso en que la varianza de la fuente principal sea desconocida.

Consideremos el modelo dado por las ecuaciones (2.1a) y (2.1b), donde ahora la sucesión $u_1, \dots, u_t$ es de variables aleatorias independientes igualmente distribuidas según la mixtura

$$u_t \sim (1 - \lambda_0) N(0, \sigma^2) + \lambda_0 N(0, k_0^2 \sigma^2),$$

con λ_0 y $k_0^2 > 1$ ambos conocidos. El caso en que λ_0 fuese desconocido podría tratarse de manera análoga a como se hace en inferencia bayesiana, es decir especificando una distribución conjunta para todos los parámetros del modelo incluido λ, tal como hacen por ejemplo Bernardo y Girón (1988a, 1988b) en el análisis de ciertos problemas de mixturas. Sin embargo, el caso k_0^2 desconocido es más complicado, recomendándose en este caso un análisis de sensibilidad del modelo respecto de este parámetro.

Supongamos, en primer lugar que σ^2 es conocido. Como ocurre en el teorema 2.1. las distribuciones que se van obteniendo a lo largo del tiempo son todas del mismo tipo, concretamente mixturas de distribuciones normales, con un número creciente de términos. Entonces el equivalente del teorema 2.1 es el siguiente

Teorema 3.1. *Si*

$$\theta_1 \mid \sigma^2 \sim \sum_{i_0=1}^{t_0} \mu_{i_0}^1 \, N(\mathbf{m}_1^{i_0}, \sigma^2 \, \mathbf{C}_1^{i_0})$$

independientemente de $u_1, \ldots, u_n, \ldots$, *entonces para todo t la distribución a posteriori de* $\theta_{t+1} \mid \sigma^2$, $y_1, \ldots, y_t$ *es una mixtura de* $t_0 2^t$ *términos de la forma*

$$\sum_{i_t=1}^{2} \sum_{i_{t-1}=1}^{2} \cdots \sum_{i_1=1}^{2} \sum_{i_0=1}^{t_0} \mu_{i_0 i_1 \ldots i_{t-1} i_t}^{t+1} \, N(\mathbf{m}_{t+1}^{i_0 i_1 \ldots i_{t-1} i_t}, \sigma^2 \, \mathbf{C}_{t+1}^{i_0 i_1 \ldots i_{t-1} i_t})$$

donde los nuevos parámetros se calculan a partir de las relaciones siguientes (para simplificar la notación, los superíndices y subíndices $i_0, \ldots, i_{t-1}$, 1 *y* $i_0, \ldots, i_{t-1}$, 2 *se representarán por* **1** *y* **2**, *respectivamente): si*

$$\begin{aligned} \hat{y}_t^1 &= \mathbf{x}_t' \, \mathbf{m}_t^1 \\ e_t^1 &= y_t^1 - \hat{y}_t^1 \\ \mathbf{B}_t^1 &= \mathbf{T}\mathbf{C}_t^1\mathbf{T}' + \alpha\alpha' \\ v_t^1 &= \mathbf{x}_t' \, \mathbf{C}_t^1 \mathbf{x}_t + 1 \\ \mathbf{a}_t^1 &= (\mathbf{T}\mathbf{C}_t^1 \mathbf{x}_t + \alpha) \, / \, v_t^1, \end{aligned} \tag{3.1a}$$

y

$$\begin{aligned} \hat{y}_t^2 &= \mathbf{x}_t' \, \mathbf{m}_t^2 \\ e_t^2 &= y_t^2 - \hat{y}_t^2 \\ \mathbf{B}_t^2 &= \mathbf{T}\mathbf{C}_t^2\mathbf{T}' + k^2 \, \alpha\alpha' \\ v_t^2 &= \mathbf{x}_t' \, \mathbf{C}_t^2 \mathbf{x}_t + k^2 \\ \mathbf{a}_t^2 &= (\mathbf{T}\mathbf{C}_t^2 \mathbf{x}_t + k^2 \, \alpha) \, / \, v_t^2, \end{aligned} \tag{3.1b}$$

entonces

$$\begin{aligned} \mathbf{m}_{t+1}^1 &= \mathbf{T}\mathbf{m}_t^1 + \mathbf{a}_t^1 e_t^1 \\ \mathbf{C}_{t+1}^1 &= \mathbf{B}_t^1 - v_t^1 \mathbf{a}_t^1 \mathbf{a}_t^{1\prime}, \end{aligned} \tag{3.2a}$$

y

$$\begin{aligned} \mathbf{m}_{t+1}^2 &= \mathbf{T}\mathbf{m}_t^2 + \mathbf{a}_t^2 + \mathbf{a}_t^2 e_t^2 \\ \mathbf{C}_{t+1}^2 &= \mathbf{B}_t^2 - v_t^2 \mathbf{a}_t^2 \mathbf{a}_t^{2\prime}, \end{aligned} \tag{3.2b}$$

Los coeficientes de la mixtura

$$\mu_{i_0 i_1 \ldots i_{t-1} i_t}^{t+1}$$

se calculan a partir de las relaciones siguientes:

$$\begin{aligned} \mu_1^{t+1} &\propto (1 - \lambda_0)\, \mu_{i_0 \ldots i_{t-1}}\, f_1(y_t \mid \sigma^2, y_1, \ldots, y_{t-1}) \\ \mu_2^{t+1} &\propto \lambda_0\, \mu_{i_0 \ldots i_{t-1}}\, f_2(y_t \mid \sigma^2, y_1, \ldots, y_{t-1}) \end{aligned} \tag{3.3}$$

donde $f_{i_0 \ldots i_{t-1} i_t}(y_t \mid \sigma^2, y_1, \ldots, y_{t-1})$ *es la densidad predictiva de* y_t *condicionada a* $\sigma^2, y_1, \ldots, y_{t-1}$ *y a que la distribución inicial de* θ_1 *sea normal de índice* i_0 *y los errores de observación* $u_1, \ldots, u_t$ *procedan, respectivamente, de las poblaciones* $i_1, \ldots, i_t$. *Por el teorema 3.1, ésta viene dada por una densidad normal de parámetros*

$$N(\hat{y}_t^{i_0 \ldots i_{t-1} i_t};\ \sigma^2\, v_t^{i_0 \ldots i_{t-1} i_t}).$$

La demostración sigue los pasos del teorma 2.1. Sin embargo es necesario utilizar los dos resultados siguientes, cuya demostración se da en el Apéndice.

Lema 3.1. *Si* $\mathbf{X}$, $\mathbf{Y}$ *son dos vectores aleatorios independientes* k *y* l *variantes, respectivamente;* $\mathbf{X} \sim \sum_i \lambda_i N(\mu_i, \Sigma_i)$ *e* $\mathbf{Y} \sim \sum_j \lambda_j^* N(\mu_j^*, \Sigma_j^*)$, *entonces para cualquier matriz* $\mathbf{A}$ *de dimensiones* $m \times (k+l)$, *se tiene que*

$$\mathbf{Z} = \mathbf{A}\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$$

se distribuye también como una mixtura de normales dada por

$$\mathbf{Z} \sim \sum_{i,j} \lambda_i \lambda_j^* N\left[\mathbf{A}\begin{pmatrix} \mu_i \\ \mu_j^* \end{pmatrix};\ \mathbf{A}\begin{pmatrix} \Sigma_i & \mathbf{O} \\ \mathbf{O} & \Sigma_j^* \end{pmatrix}\mathbf{A}'\right].$$

Lema 3.2. *Si el vector aleatorio* $\mathbf{X}$, *que consta de dos subvectores* $\mathbf{X}' = (\mathbf{X}_1', \mathbf{X}_2')$, *tiene como función de densidad* $f(\mathbf{x}) = \sum_i \lambda_i f_i(\mathbf{x})$, *entonces la distribución de* $\mathbf{X}_2 \mid \mathbf{X}_1 = \mathbf{x}_1$ *tiene como densidad*

$$f(\mathbf{x}_2 \mid \mathbf{x}_1) = \sum_i \lambda_i' f_i(\mathbf{x}_2 \mid \mathbf{x}_1)$$

donde $\lambda_i' \propto \lambda_i f_i(\mathbf{x}_1)$, *con la condición* $\sum_i \lambda_i' = 1$, *y* $f_i(\mathbf{x}_2 \mid \mathbf{x}_1)$ *es la densidad condicionada de cada término de la mixtura.*

Si suponemos el teorema cierto hasta el instante t, sería

$$\theta_t \mid \sigma^2, y_1, \ldots, y_{t-1} \sim \sum_{i_{t-1}=1}^{2} \cdots \sum_{i_0=1}^{i_0} \mu_{i_0 \ldots i_{t-1}}^t\, N(\mathbf{m}_t^{i_0 \ldots i_{t-1}}, \sigma^2 \mathbf{C}_t^{i_0 \ldots i_{t-1}}).$$

De la independencia condicional de θ_t y u_t dado $y_1, \ldots, y_{t-1}$ y del lema 3.1 aplicado al sistema (2.4) se deduce, tras algunas manipulaciones algebraicas, que la distribución de

$$\begin{pmatrix} Y_t \\ \theta_{t+1} \end{pmatrix} \mid \sigma^2, y_1, \ldots, y_{t-1}$$

es la mixtura

$$\sum_{i_{t-1}=1}^{2} \cdots \sum_{i_0=1}^{i_0} (1 - \lambda_0)\, \mu^t_{i_0 \ldots i_{t-1}}\, N\left[\begin{pmatrix} \mathbf{x}_t' \mathbf{m}_t^1 \\ \mathbf{T}\, \mathbf{m}_t^1 \end{pmatrix}; \sigma^2 \begin{pmatrix} \mathbf{x}_t' \mathbf{C}_t^1 \mathbf{x}_t + 1 & \mathbf{x}_t' \mathbf{C}_t^1 \mathbf{T}' + \alpha' \\ \mathbf{T}\, \mathbf{C}_t^1 \mathbf{x}_t + \alpha & \mathbf{T}\, \mathbf{C}_t^1 \mathbf{T}' + \alpha\alpha' \end{pmatrix}\right]$$

+

$$\sum_{i_{t-1}=1}^{2} \cdots \sum_{i_0=1}^{i_0} \lambda_0\, \mu^t_{i_0 \ldots i_{t-1}}\, N\left[\begin{pmatrix} \mathbf{x}_t' \mathbf{m}_t^2 \\ \mathbf{T}\, \mathbf{m}_t^1 \end{pmatrix}; \sigma^2 \begin{pmatrix} \mathbf{x}_t' \mathbf{C}_t^2 \mathbf{x}_t + k^2 & \mathbf{x}_t' \mathbf{C}_t^2 \mathbf{T}' + k^2 \alpha' \\ \mathbf{T}\, \mathbf{C}_t^2 \mathbf{x}_t + k^2 \alpha & \mathbf{T}\, \mathbf{C}_t^2 \mathbf{T}' + k^2 \alpha\alpha' \end{pmatrix}\right].$$

De la expresión anterior y del lema 3.2, se deduce que la distribución condicionada de $\theta_{t+1} \mid \sigma^2, y_1, \ldots, y_t$ es la mixtura dada por el teorema.

Es también fácil comprobar que los coeficientes de la mixtura

$$\mu^{t+1}_{i_0, i_1, \ldots, i_{t-1}, i_t}$$

son las probabilidades de que los términos de error del modelo (3.1) $u_1, \ldots, u_t$ provengan del modelo $i_1, \ldots, i_{t-1}, i_t$, respectivamente, donde $i_t = 1$ corresponde al modelo $N(0, \sigma^2)$ y $i_t = 2$ al modelo $N(0, k^2 \sigma^2)$, condicionado además a que el vector de parámetros inicial θ_1 se distribuya según una $N(\mathbf{m}_1^{i_0}, \sigma^2 \mathbf{C}_1^{i_0})$.

Como en el teorema 2.1, el resultado anterior puede utilizarse para calcular, p.e., la distribución de una observación futura. Es fácil demostrar, utilizando de nuevo el lema 3.1, que la distribución de $y_{t+1} \mid \sigma^2, y_1, \ldots, y_t$ es también una mixtura de $t_0 2^{t+1}$ términos normales.

Aunque los resultados del teorema anterior son exactos, la complejidad de los cálculos que supone programar éstos (el número de términos de la mixtura en el instante t es de $t_0 2^t$) los hace poco atractivos desde el punto de vista práctico. Se impone pues, la necesidad de emplear aproximaciones que mantengan la simplicidad del filtro del teorema 2.1. Como todas las distribuciones que aparecen como consecuencia de la demostración del teorema 3.1 son mixturas de normales, es necesario aproximar tales mixturas por distribuciones normales en cada etapa. Esto lo podemos hacer de dos maneras:

a) aproximar la distribución conjunta para cada t del par

$$\begin{pmatrix} Y_t \\ \theta_{t+1} \end{pmatrix} \mid \sigma^2, y_1, \ldots, y_{t-1}$$

por una distribución normal y, a continuación, calcular la distribución condicionada de $\theta_{t+1} \mid y_1, \ldots, y_t$.

b) utilizando el lema 3.2, calcular a partir de la distribución conjunta anterior, la distribución condicional $\theta_{t+1} \mid y_1, \ldots, y_t$, que sería una mixtura y aproximar ésta por una distribución normal.

La segunda alternativa parece preferible ya que el proceso de aproximación se realiza después de calcular la distribución *exacta* y, además, permite calcular las probabilidades *aproximadas* de que la observación provenga de cada una de las dos poblaciones.

La manera de aproximar mixturas de distribuciones normales por una única distribución normal depende del criterio de aproximación elegido. Sin embargo en este caso, todos los criterios razonables de aproximación equivalen a sustituir la mixtura por una distribución normal con vector de medias y matriz de covarianzas idénticos a los de la mixtura. Si en cada etapa del proceso de incorporación de nueva información, se aproxima la mixtura correspondiente por la distribución normal calculada por el criterio anterior, se obtiene la versión aproximada del teorema 3.1 que tiene cierto parecido al *filtro robusto* de Guttman y Peña (loc. cit.).

Teorema 3.2. *Si, bajo las condiciones del teorema anterior, en cada instante de tiempo t la distribución de* $\theta_t \mid \sigma^2, y_1, \ldots, y_{t-1}$ *se aproxima por una distribución normal k-variante* $N(\mathbf{m}_t, \sigma^2 \mathbf{C}_t)$, *entonces la distribución aproximada de*

$$\theta_{t+1} \mid \sigma^2, y_1, \ldots, y_t$$

es la mixtura

$$\mu_{t+1}^{(1)} N(\mathbf{m}_{t+1}^{(1)}, \sigma^2 \mathbf{C}_{t+1}^{(1)}) + \mu_{t+1}^{(2)} N(\mathbf{m}_{t+1}^{(2)}, \sigma^2 \mathbf{C}_{t+1}^{(2)}) \qquad (3.4)$$

donde los nuevos parámetros se calculan a partir de las relaciones siguientes: si

$$\hat{y}_t = \mathbf{x}_t' \mathbf{m}_t$$

$$e_t = y_t - \hat{y}_t$$

$$\mathbf{B}_t^{(1)} = \mathbf{T}\mathbf{C}_t\mathbf{T}' + \alpha\alpha' \qquad \mathbf{B}_t^{(2)} = \mathbf{T}\mathbf{C}_t\mathbf{T}' + k^2\,\alpha\alpha'$$

$$v_t^{(1)} = \mathbf{x}_t'\mathbf{C}_t\mathbf{x}_t + 1 \qquad v_t^{(2)} = \mathbf{x}_t'\mathbf{C}_t\mathbf{x}_t + k^2$$

$$\mathbf{a}_t^{(1)} = (\mathbf{T}\mathbf{C}_t\mathbf{x}_t + \alpha) / v_t^{(1)}, \qquad \mathbf{a}_t^{(2)} = (\mathbf{T}\mathbf{C}_t\mathbf{x}_t + k^2\alpha) / v_t^{(2)};$$

entonces

$$\begin{aligned} \mathbf{m}_{t+1}^{(1)} &= \mathbf{T}\mathbf{m}_t + \mathbf{a}_t^{(1)} e_t & \mathbf{m}_{t+1}^{(2)} &= \mathbf{T}\mathbf{m}_t + \mathbf{a}_t^{(2)} e_t \\ \mathbf{C}_{t+1}^{(1)} &= \mathbf{B}_t^{(1)} - v_t^{(1)} \mathbf{a}_t^{(1)} \mathbf{a}_t^{\prime(1)}, & \mathbf{C}_{t+1}^{(2)} &= \mathbf{B}_t^{(1)} - v_t^{(2)} \mathbf{a}_t^{(2)} \mathbf{a}_t^{\prime(2)}, \end{aligned} \tag{3.5}$$

Los coeficientes de la mixtura se calculan por las fórmulas

$$\begin{aligned} \mu_{t+1}^{(1)} &\propto (1 - \lambda_0) v_t^{(1)-1/2} \exp\left\{ - \frac{e_t^2}{2\sigma^2 v_t^{(1)}} \right\} \\ \mu_{t+1}^{(2)} &\propto \lambda_0 v_t^{(2)-1/2} \exp\left\{ - \frac{e_t^2}{2\sigma^2 v_t^{(2)}} \right\} \end{aligned} \tag{3.6}$$

sujetos a la condición de ser $\mu_{t+1}^{(1)} + \mu_{t+1}^{(2)} = 1$.

A diferencia del teorema anterior donde los coeficientes de la mixtura son las probabilidades exactas, ahora los coeficientes de la mixtura $\mu_{t+1}^{(1)}$ y $\mu_{t+1}^{(2)}$ son las probabilidades aproximadas, condicionadas a la información muestral $y_1, \ldots, y_t$ y a la información a priori, de que el término de error u_t provenga de las poblaciones $N(0,\sigma^2)$ o $N(0, k^2\sigma^2)$, respectivamente.

La demostración se basa en la del teorema anterior. La distribución *aproximada* de

$$\begin{pmatrix} y_t \\ \theta_{t+1} \end{pmatrix} \mid \sigma^2, y_1, \ldots, y_{t-1}$$

sería la mixtura

$$\begin{aligned} &(1 - \lambda_0) N\left[\begin{pmatrix} \mathbf{x}_t' \mathbf{m}_t \\ \mathbf{T}\mathbf{m}_t \end{pmatrix}; \sigma^2 \begin{pmatrix} \mathbf{x}_t' \mathbf{C}_t \mathbf{x}_t + 1 & \mathbf{x}_t' \mathbf{C}_t \mathbf{T}' + \alpha' \\ \mathbf{T}\mathbf{C}_t \mathbf{x}_t + \alpha & \mathbf{T}\mathbf{C}_t \mathbf{T}' + \alpha\alpha' \end{pmatrix} \right] \\ &+ \lambda_0 N\left[\begin{pmatrix} \mathbf{x}_t' \mathbf{m}_t \\ \mathbf{T}\mathbf{m}_t \end{pmatrix}; \sigma^2 \begin{pmatrix} \mathbf{x}_t' \mathbf{C}_t \mathbf{x}_t + k^2 & \mathbf{x}_t' \mathbf{C}_t \mathbf{T}' + k^2 \alpha' \\ \mathbf{T}\mathbf{C}_t \mathbf{x}_t + k^2 \alpha & \mathbf{T}\mathbf{C}_t \mathbf{T}' + k^2 \alpha\alpha' \end{pmatrix} \right]; \end{aligned} \tag{3.7}$$

y de aquí, por el lema 3.2, se deducen las ecuaciones precedentes.

La nueva distribución (3.4) ha de aproximarse por una distribución normal $N(\mathbf{m}_{t+1}, \sigma^2 \mathbf{C}_{t+1})$, donde los nuevos parámetros vienen dados por las ecuaciones siguientes:

$$\mathbf{m}_{t+1} = \mu_{t+1}^{(1)}\, \mathbf{m}_{t+1}^{(1)} + \mu_{t+1}^{(2)}\, \mathbf{m}_{t+1}^{(2)}$$

$$\mathbf{C}_{t+1} = \mu_{t+1}^{(1)}\, \mathbf{C}_{t+1}^{(1)}$$

$$+ \mu_{t+1}^{(2)}\, \mathbf{C}_{t+1}^{(2)} + \mu_{t+1}^{(1)}\, (\mathbf{m}_{t+1}^{(1)} - \mathbf{m}_{t+1})\, (\mathbf{m}_{t+1}^{(1)} - \mathbf{m}_{t+1})' \qquad (3.8)$$

$$+ \mu_{t+1}^{(2)}\, (\mathbf{m}_{t+1}^{(2)} - \mathbf{m}_{t+1})\, (\mathbf{m}_{t+1}^{(2)} - \mathbf{m}_{t+1})'.$$

Estas ecuaciones pueden escribirse de la siguiente manera, cuya interpretación resulta más intuitiva:

$$\mathbf{m}_{t+1} = \mathbf{T}\mathbf{m}_t + [\, \mu_{t+1}^{(1)}\, \mathbf{a}_t^{(1)} + \mu_{t+1}^{(2)}\, \mathbf{a}_t^{(2)} \,]\, e_t$$

$$\mathbf{C}_{t+1} = \mu_{t+1}^{(1)}\, \mathbf{C}_{t+1}^{(1)} + \mu_{t+1}^{(2)}\, \mathbf{C}_{t+1}^{(2)} + \mu_{t+1}^{(1)}\, \mu_{t+1}^{(2)}\, [\, \mathbf{a}_t^{(1)} - \mathbf{a}_t^{(2)} \,]\, [\, \mathbf{a}_t^{(1)} - \mathbf{a}_t^{(2)} \,]'\, e_t^2.$$

La primera ecuación es análoga a la primera de (2.3) con la diferencia de que ahora el vector tasa de respuesta en cada instante t es la mixtura correspondiente de los vectores de tasa de respuesta de ambos modelos ponderados por las probabilidades correspondientes, es decir, $\bar{\mathbf{a}}_t = [\, \mu_{t+1}^{(1)}\, \mathbf{a}_t^{(1)} + \mu_{t+1}^{(2)}\, \mathbf{a}_t^{(2)} \,]$. La segunda ecuación contiene un término adicional, aparte de la mixtura de las matrices de covarianza correspondientes, que representa la incertidumbre adicional debida al desconocimiento del modelo del cual proviene la observación.

Consideremos ahora el caso en que σ^2 sea desconocida. El teorema 3.1 tiene su equivalente sustituyendo mixturas de normales por mixturas de distribuciones normales-gamma invertidas. En vez de dar los resultados exactos, que son similares aunque algo más complicados que los del teorema 3.1, consideraremos en cada etapa el problema de aproximar una mixtura de distribuciones normales-gamma invertidas por una distribución del mismo tipo. Para no interferir con el planteamiento del problema objeto de esta sección, trataremos con cierto detalle éste problema en la sección siguiente.

Para demostrar el teorema que sigue, será necesario el lema siguiente, cuya demostración también se incluye en el Apéndice.

Lema 3.3. *Si el vector aleatorio*

$$\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \sigma^2 \end{pmatrix} \sim N\,Gal(\mathbf{m}, \mathbf{C}; a.p)$$

entonces, la distribución de

$$\begin{pmatrix} \mathbf{x}_2 \\ \sigma^2 \end{pmatrix} \Big| \mathbf{x}_1 \sim N\,Gal(\mathbf{m}_{2|1}, \mathbf{C}_{2|1}; a_{2|1}, p_{2|1})$$

donde, con notación obvia, los nuevos parámetros son:

$$\mathbf{m}_{2|1} = \mathbf{m}_2 + \mathbf{C}_{21}\mathbf{C}_{11}^{-1}(\mathbf{x}_1 - \mathbf{m}_1)$$
$$\mathbf{C}_{2|1} = \mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}$$
$$a_{2|1} = a + \frac{1}{2}(\mathbf{x}_1 - \mathbf{m}_1)'\,\mathbf{C}_{11}^{-1}(\mathbf{x}_1 - \mathbf{m}_1)$$
$$p_{2|1} = p + \frac{k}{2},$$

siendo k la dimensión de $\mathbf{x}_1$.

Teorema 3.3. *Si a priori* $(\theta_1, \sigma^2) \sim N\,Gal(\mathbf{m}_1, \mathbf{C}_1; a_1, p_1)$ y dado σ^2, θ_1 es *independiente de los errores* $u_1, \ldots, u_n, \ldots$, *y si en cada instante de tiempo t* la distribución de $\theta_t, \sigma^2 \mid y_1, \ldots, y_{t-1}$ *se aproxima por una distribución normal-gamma invertida* $N\,Gal(m_t, \mathbf{C}_t; a_t, p_t)$, *entonces la distribución a posteriori aproximada de*

$$\theta_{t+1}, \sigma^2 \mid y_1, \ldots, y_t$$

es la siguiente mixtura

$$\omega_{t+1}^{(1)}\,N\,Gal(\mathbf{m}_{t+1}^{(1)}, \mathbf{C}_{t+1}^{(1)}; a_{t+1}^{(1)}, p_{t+1}^{(1)}) + \omega_{t+1}^{(2)}\,N\,Gal(\mathbf{m}_{t+1}^{(2)}, \mathbf{C}_{t+1}^{(2)}; a_{t+1}^{(2)}, p_{t+1}^{(2)}); \qquad (3.9)$$

donde los parámetros $\mathbf{m}_{t+1}^{(1)}, \mathbf{C}_{t+1}^{(1)}, \mathbf{m}_{t+1}^{(2)}, C_{t+1}^{(2)}$ *son los dados por las ecuaciones* (3.5): los nuevos parámetros $a_{t+1}^{(1)}, p_{t+1}^{(1)}, a_{t+1}^{(2)}, p_{t+1}^{(2)}$ *vienen dados por*

$$a_{t+1}^{(1)} = a_t + \frac{1}{2}\,e_t^2 / v_t^{(1)};\quad a_{t+1}^{(2)} = a_t + \frac{1}{2}\,e_t^2 / v_t^{(2)};\quad p_{t+1}^{(1)} = p_{t+1}^{(2)} = p_t + \frac{1}{2};$$

y los coeficientes de la mixtura se calculan mediante las fórmulas

$$\omega_{t+1}^{(1)} \propto (1 - \lambda_0)\,v_t^{(1)-1/2}\left[1 + \frac{e_t^2}{2\,a_t v_t^{(1)}}\right]^{-(p_t+1/2)}$$

$$\omega_{t+1}^{(2)} \propto \lambda_0\,v_t^{(2)-1/2}\left[1 + \frac{e_t^2}{2\,a_t v_t^{(2)}}\right]^{-(p_t+1/2)}$$

sujetos a la condición de ser $\omega_{t+1}^{(1)} + \omega_{t+1}^{(2)} = 1$.

Como en el teorema anterior, los coeficientes de la mixtura $\omega_{t+1}^{(1)}$ y $\omega_{t+1}^{(2)}$ son las probabilidades aproximadas, condicionadas a las observaciones

$y_1, \ldots, y_t$ y a la información a priori, de que el término de error u_t provenga de las poblaciones $N(0,\sigma^2)$ o $N(0,k^2\sigma^2)$, respectivamente.

La demostración del teorema es como sigue: si en cada instante t la distribución aproximada de $\theta_t, \sigma^2 \mid y_1, \ldots, y_{t-1}$, que es una $NGal(\mathbf{m}_t, \mathbf{C}_t; a_t, p_t)$, es independiente de

$$u_t - (1 - \lambda_0) N(0,\sigma^2) + \lambda_0 N(0,k^2\sigma^2),$$

entonces por el teorema 3.2, la distribución de $(y_t, \theta_{t+1}) \mid \sigma^2, y_1, \ldots, y_{t-1}$ es la mixtura de normales dada por las ecuaciones (3.7). Por otra parte, la distribución de $\sigma^2 \mid y_1, \ldots, y_{t-1}$ es, por hipótesis, una $Gal(a_t, p_t)$. Como la marginal y las condicionadas caracterizan unívocamente la distribución conjunta, es fácil comprobar que la distribución conjunta

$$\begin{pmatrix} y_t \\ \theta_{t+1} \\ \sigma^2 \end{pmatrix} \Bigg| \; y_1, \ldots, y_{t-1}$$

viene dada por la mixtura

$$(1 - \lambda_0) NGal\left[\begin{pmatrix} \mathbf{x}_t'\mathbf{m}_t \\ \mathbf{T}\,\mathbf{m}_t \end{pmatrix}, \begin{pmatrix} \mathbf{x}_t'\mathbf{C}_t\mathbf{x}_t + 1 & \mathbf{x}_t'\mathbf{C}_t\mathbf{T}' + \alpha' \\ \mathbf{T}\,\mathbf{C}_t\mathbf{x}_t + \alpha & \mathbf{T}\,\mathbf{C}_t\mathbf{T}' + \alpha\alpha' \end{pmatrix}; a_t, p_t\right]$$

$$+\lambda_0 NGal\left[\begin{pmatrix} \mathbf{x}_t'\mathbf{m}_t \\ \mathbf{T}\,\mathbf{m}_t \end{pmatrix}, \begin{pmatrix} \mathbf{x}_t'\mathbf{C}_t\mathbf{x}_t + k^2 & \mathbf{x}_t'\mathbf{C}_t\mathbf{T}' + k^2\alpha' \\ \mathbf{T}\,\mathbf{C}_t\mathbf{x}_t + k^2\alpha & \mathbf{T}\,\mathbf{C}_t\mathbf{T}' + k^2\alpha\alpha' \end{pmatrix}; a_t, p_t\right].$$

Si a esta mixtura le aplicamos, primero el lema 3.2, a continuación, el lema 3.3 y tenemos en cuenta el teorema 3.2, tras algunos cálculos resulta que la distribución de

$$\begin{pmatrix} \theta_{t+1} \\ \sigma^2 \end{pmatrix} \Bigg| \; y_1, \ldots, y_t$$

es la mixtura

$$\omega_{t+1}^{(1)} NGal(\mathbf{m}_{t+1}^{(1)}, \mathbf{C}_{t+1}^{(1)}; a_{t+1}^{(1)}, p_{t+1}^{(1)}) + \omega_{t+1}^{(2)} NGal(\mathbf{m}_{t+1}^{(2)}, \mathbf{C}_{t+1}^{(2)}; a_{t+1}^{(2)}, p_{t+1}^{(2)});$$

donde los parámetros $\mathbf{m}_{t+1}^{(i)}$, $\mathbf{C}_{t+1}^{(i)}$ y $p_{t+1}^{(i)}$ son los dados por el teorema y

$$a_{t+1}^{(i)} = a_t + \frac{1}{2}(y_t - \mathbf{x}_t'\mathbf{m}_t)' v_t^{(i)-1} (y_t - \mathbf{x}_t'\mathbf{m}_t) = a_t + \frac{1}{2} e_t^2 / v_t^{(i)}.$$

Finalmente, por el lema 3.2, los coeficientes de la mixtura vienen dados por

$$\omega_{t+1}^{(1)} \propto (1 - \lambda_0)\, f_1(y_t \mid y_1, \ldots, y_{t-1})$$

$$\omega_{t+1}^{(2)} \propto \lambda_0\, f_2(y_t \mid y_1, \ldots, y_{t-1}),$$

donde $f_i(y_t \mid y_1, \ldots, y_{t-1})$ es la densidad predictiva condicionada a la información $y_1, \ldots, y_{t-1}$ y a que el error u_t proviene de la población i-ésima, que es una t de Student $St(2p_t; \hat{y}_t, (a_t/p_t)v_t^{(i)})$. Y de aquí se sigue el teorema.

De este teorema es fácil obtener, p.e., la distribución marginal de θ_{t+1} dados los datos $y_1, \ldots, y_t$ que, por las propiedades de las mixturas de normales-gamma invertidas, es una mixtura de distribuciones t de Student k-variantes

$$\omega_{t+1}^{(1)}\, St(2p_{t+1}^{(1)}; \mathbf{m}_{t+1}^{(1)}; (a_{t+1}^{(1)} / p_{t+1}^{(1)})\, \mathbf{C}_{t+1}^{(1)}) + \omega_{t+1}^{(2)}\, St(2p_{t+1}^{(2)}; \mathbf{m}_{t+1}^{(2)}; (a_{t+1}^{(2)} / p_{t+1}^{(2)})\, \mathbf{C}_{t+1}^{(2)})$$

con los mismos grados de libertad $2p_t + 1$, vectores de centralización $\mathbf{m}_{t+1}^{(1)}$, $\mathbf{m}_{t+1}^{(2)}$ y matrices de escala $(a_{t+1}^{(1)} / p_{t+1}^{(1)})\, \mathbf{C}_{t+1}^{(1)}$ y $(a_{t+1}^{(2)} / p_{t+1}^{(2)})\, \mathbf{C}_{t+1}^{(2)}$, respectivamente.

Análogamente, la distribución marginal aproximada de $\sigma^2 \mid y_t \mid y_1, \ldots, y_t$ es la mixtura

$$\omega_{t+1}^{(1)} Gal(a_{t+1}^{(1)}, p_{t+1}^{(1)}) + \omega_{t+1}^{(2)} Gal(a_{t+1}^{(2)}, p_{t+1}^{(2)}).$$

De modo análogo, aunque los resultados son más complicados, se podría obtener, utilizando el lema 3.1 y el resultado anterior, la distribución predictiva aproximada de una observación futura y_{t+1} $y_1, \ldots, y_t$, que sería una mixtura de, en principio ocho términos, cada uno de los cuales es una t univariante.

Queda patente de estos resultados, la necesidad de aproximar mixturas de distribuciones como las normales-gamma invertidas, que son las básicas en el proceso de aprendizaje dado por el teorema, y también de otro tipo de distribuciones como las mixturas de gammas-invertidas y t de Student multivariantes.

Estos problemas se tratan con cierto detalle en la sección siguiente. En ésta, sólamente aplicaremos el teorema 4.1 a la aproximación de la mixtura (3.8), que es la necesaria para mantener simplificada la estructura del filtro.

Teorema 3.4. *La mejor aproximación de la mixtura (3.8) por una distribución normal-gamma invertida* $NGaI(\mathbf{m}_{t+1}, \mathbf{C}_{t+1}; a_{t+1}, p_{t+1})$ *es aquella cuyos parámetros satisfacen las ecuaciones siguientes:*

$$\mathbf{m}_{t+1} = \omega_{t+1}^{(1)}\mathbf{m}_{t+1}^{(1)} + \omega_{t+1}^{(2)}\mathbf{m}_{t+1}^{(2)}$$

$$\mathbf{C}_{t+1} = \omega_{t+1}^{(1)}\mathbf{C}_{t+1}^{(1)} + \omega_{t+1}^{(2)}\mathbf{C}_{t+1}^{(2)} + \omega_{t+1}^{(1)}(\mathbf{m}_{t+1}^{(1)} - \mathbf{m}_{t+1})\,(p_{t+1}^{(1)} / a_{t+1}^{(1)})\,(\mathbf{m}_{t+1}^{(1)} - \mathbf{m}_{t+1})' + \omega_{t+1}^{(2)}(\mathbf{m}_{t+1}^{(2)} - \mathbf{m}_{t+1})\,(p_{t+1}^{(2)} / a_{t+1}^{(2)})\,(\mathbf{m}_{t+1}^{(2)} - \mathbf{m}_{t+1})'$$

$$\frac{p_{t+1}}{a_{t+1}} = \omega_{t+1}^{(1)}\frac{p_{t+1}^{(1)}}{a_{t+1}^{(1)}} + \omega_{t+1}^{(2)}\frac{p_{t+1}^{(2)}}{a_{t+1}^{(2)}}$$

$$-\log a_{t+1} + \psi(p_{t+1}) = \omega_{t+1}^{(1)}\,[\,-\log a_{t+1}^{(1)} + \psi(p_{t+1}^{(1)})\,] + \omega_{t+1}^{(2)}\,[\,-\log a_{t+1}^{(2)} + \psi(p_{t+1}^{(2)})\,].$$

Las dos primeras ecuaciones son explícitas y similares a las (3.8) salvo que ahora los sumandos segundo y tercero de la segunda ecuación están *modulados* respectivamente por los coeficientes $p_{t+1}^{(1)} / a_{t+1}^{(1)}$ y $p_{t+1}^{(2)} / a_{t+1}^{(2)}$, que son debidos al desconocimiento de la varianza σ^2. Las dos ecuaciones restantes se pueden resolver, de modo simple mediante técnicas de cálculo numérico (véase, p.e., Rojano (no publicado)). Si, además, tenemos en cuenta que los grados de libertad de los dos términos de la mixtura son idénticos $(2p_t + 1)$ entonces, por el teorema 4.2, el número de grados de libertad de la aproximación $2p_{t+1}$ es menor que el de los términos. De hecho, se puede demostrar fácilmente que, a medida que t aumenta $a_{t+1}^{(1)} \simeq a_{t+1}^{(2)}$, con lo que $p_{t+1} \simeq p_t + 1/2$, es decir los grados de libertad de la aproximación en cada etapa aumentan en una unidad, como ocurre con los procedimientos tradicionales de aproximación. Sin embargo, sobre todo en las primeras etapas de aproximación del filtro que son cruciales, creemos que el método que proponemos es más adecuado y efectivamente refleja la incertidumbre adicional que entraña el proceso de aproximación. Queda claro también, teniendo en cuenta los resultados de la sección siguiente, que para t grande los dos métodos de aproximación son equivalentes.

4. APROXIMACION DE CIERTAS MIXTURAS DE DISTRIBUCIONES

Es bien sabido que la mayoría de los modelos propuestos para describir datos que incluyen observaciones espurias, observaciones aberrantes o que simplemente no pueden considerarse como datos normales, en los que el problema central es estimar el parámetro de localización de los mismos y/o detectar la presencia de estas observaciones, suelen venir descritos por modelos basados en mixturas de distribuciones y típicamente los elementos de éstas suelen ser distribuciones normales (como, p.e., Box y Tiao

(1968), Abraham y Box (1978), Guttman, Dutter y Freeman (1978). Freeman (1980) y Pettit y Smith (1984)). Como hemos visto en la sección anterior, las generalizaciones de estos modelos al caso dinámico se centran también en la estimación dinámica de los parámetros en cada instante de tiempo y de la varianza de la fuente de variación común.

Desde el punto de vista bayesiano, los parámetros de interés siguen generalmente una distribución normal, mientras que el parámetro de escala, el grado de contaminación, el desplazamiento de los datos espurios y la proporción de éstos se consideran o bien parámetros marginales o, simplemente, se realiza un análisis condicional y posteriormente se estudia la sensibilidad de la distribución a posteriori respecto a cambios en estos parámetros.

Básicamente todos los modelos propuestos dan como resultado el que la distribución a posteriori del parámetro de interés, que denominaremos θ, sea una mixtura de distribuciones t de Student con el mismo número de grados de libertad: distribución que se obtiene calculando la marginal a partir de la distribución conjunta de (θ,σ^2), que típicamente es una mixtura de normales-gamma invertidas, donde σ^2 es un parámetro de escala (en el modelo de las secciones anteriores, la varianza común de la mayoría de los términos de error).

Señalábamos que desde el punto de vista computacional es interesante, sobre todo si el número de términos de la mixtura es elevado o se quiere mantener la sencilla estructura de filtro lineal en cada etapa, aproximar la verdadera distribución a posteriori de (θ,σ^2) o la marginal de θ por otras más manejables. Los candidatos obvios son una normal-gamma invertida para el primer caso o bien una distribución t o una normal para el segundo.

Para tratar este problema se han propuesto varios métodos (véanse, p.e., Murray (1979), Johnson y Geisser (1982, 1983) y Johnson y Utts (1986)). Concretamente Johnson y Geisser (1983) señalan que el problema de aproximar una distribución por una t de Student, minimizando la divergencia de Kullback-Leibler no es factible analíticamente por lo que sugieren aproximar la mixtura de distribuciones t por una distribución normal.

Sin embargo, teniendo en cuenta que la distribución a posteriori del parámetro de interés, se obtiene marginalizando la distribución conjunta de (θ,σ^2), parece más natural aproximar esta distribución a posteriori por una normal-gamma que minimice la divergencia a la conjunta y, a partir de ella, calcular la marginal. Esto presenta, además, la ventaja de que la aproximación se puede calcular de forma relativamente simple como se demuestra a continuación.

Este procedimiento conduce además a otros resultados interesantes (véase el teorema 4.2 y las aproximaciones subsiguientes) sobre todo si se le compara con otros métodos de aproximación (como el de los momentos) y es el de que la incertidumbre resultante de aproximar una mixtura de distribuciones t, con los mismos grados de libertad, por una t se refleja en un número ligeramente menor de grados de libertad (es decir, colas algo más altas).

Así pues, el problema que abordamos en esta sección lo podemos plantear como sigue: supongamos que la variable (θ,σ^2) se distribuye como una mixtura de normales-gamma invertidas

$$(\theta,\sigma^2) \sim \sum_{i=1}^{k} \lambda_i NGal(\theta,\sigma^2 \mid \mathbf{m}_i, \mathbf{C}_i; a_i, p_i) \tag{4.1}$$

que queremos aproximar por una normal-gamma invertida

$$NGal(\theta,\sigma^2 \mid \tilde{\mathbf{m}}, \tilde{\mathbf{C}}; \tilde{a}, \tilde{p})$$

de modo que la divergencia de Kullback-Leibler a la mixtura sea mínima. Entonces se tiene el resultado siguiente

Teorema 4.1. *Los parámetros* $\tilde{\mathbf{m}}, \tilde{\mathbf{C}} : \tilde{a}, \tilde{p}$ *satisfacen las siguientes ecuaciones*

$$\tilde{\mathbf{m}} = \sum_{i=1}^{k} \lambda_i \mathbf{m}_i$$

$$\frac{\tilde{p}}{\tilde{a}} = \sum_{i=1}^{k} \lambda_i \frac{p_i}{a_i}$$

$$-\log \tilde{a} + \psi(\tilde{p}) = \sum_{i=1}^{k} \lambda_i [\, -\log a_i + \psi(p_i) \,]$$

$$\tilde{\mathbf{C}} = \sum_{i=1}^{k} \lambda_i \mathbf{C}_i + \sum_{i=1}^{k} \lambda_i (\mathbf{m}_i - \tilde{\mathbf{m}})\,(p_i/a_i)\,(\mathbf{m}_i - \tilde{\mathbf{m}})';$$

donde $\psi(\cdot)$ *es la función digamma.*

Para el caso más frecuente e importante en que los p_i sean todos iguales se tiene el siguiente

Teorema 4.2. *Si todos los* p_i *son iguales a* p_0, *entonces se verifica que* $\tilde{p} \leq p_0$ *con igualdad si y sólo si todos los* a_i *son idénticos.*

Esquema de la demostración. Llamando a^h y a^g a las medias ponderadas armónica y geométrica de los a_i, respectivamente, de las ecuaciones segunda y tercera del teorema 2.1 se deduce que

$$-\log \tilde{p} + \psi(\tilde{p}) = -\log p_0 + \psi(p_0) + \log\left(\frac{a^h}{a^g}\right)$$

y como la función $-\log p + \psi(p)$ es una función creciente de p y la media geométrica es mayor que la armónica, salvo si todos los a_i son iguales en cuyo caso coinciden, de aquí se sigue el teorema.

Si consideramos ahora el problema de aproximar una mixtura de distribuciones t de Student por una t aplicando el procedimiento anterior, es decir considerando las t como marginales de normales-gamma invertidas, nos encontramos con el nuevo problema de la falta de identificación de uno de los parámetros. Existen, por consiguiente, una infinidad de soluciones. Como caso límite de éstas se puede obtener una solución similar a la proporcionada por el método de los momentos. Más detalles, así como las demostraciones y procedimientos de cálculo, pueden encontrarse en Rojano (no publicado). Aquí nos limitaremos a dar aproximaciones a las distribuciones marginales de $\theta_{t+1} \mid y_1, \ldots, y_t$ y $\sigma^2 \mid y_1, \ldots, y_t$ de la sección precedente y compararlas con las dadas por el método de los momentos.

Si (θ,σ^2) se distribuye según (4.1) y además se verifica la condición del teorema 4.2, entonces la distribución marginal de θ es

$$\theta \sim \sum_{i=1}^{k} \lambda_i St(2p_0; \mathbf{m}_i, \mathbf{S}_i); \tag{4.2}$$

donde

$$\mathbf{S}_i = \frac{a_i}{p_i}\, \mathbf{C}_i$$

es la matriz de dispersión de cada término. Entonces, de los teoremas 4.1 y 4.2, la mejor aproximación a (4.2) es la t de Student $St(2\tilde{p}; \tilde{\mathbf{m}}, \mathbf{S})$, donde $\tilde{p}$ y $\tilde{\mathbf{m}}$ son los dados por los teoremas 4.1 y 4.2 y $\mathbf{S}$ viene dado por

$$\mathbf{S} = \sum_{i=1}^{k} \lambda_i' \{ \mathbf{S}_i + (\mathbf{m}_i - \tilde{\mathbf{m}})(\mathbf{m}_i - \tilde{\mathbf{m}})' \},$$

donde

$$\lambda_i' = \frac{\lambda_i}{a_i} \Big/ \sum_{i=1}^{k} \lambda_i \frac{\lambda_i}{a_i}.$$

Si ahora aplicamos el método de los momentos para aproximar la mixtura (4.2) por una t con los mismos grados de libertad que los de cada uno de los términos, suponiendo que $p_0 > 1$, resulta que la aproximación es una $St(2p_0; \tilde{\mathbf{m}}, \mathbf{S}')$, donde $\mathbf{S}'$ viene dada por

$$\mathbf{S}' = \sum_{i=1}^{k} \lambda_i \left\{ \mathbf{S}_i + \left(1 - \frac{1}{p_0}\right) (\mathbf{m}_i - \tilde{\mathbf{m}})\,(\mathbf{m}_i - \tilde{\mathbf{m}})' \right\}.$$

Si comparamos ambas aproximaciones se tiene que únicamente el vector de localización es igual en ambos casos. El número de grados de libertad utilizando el método de minimizar la divergencia de Kullback-Leibler es ligeramente menor que el generalmente utilizado por el método de los momentos, mientras que las matrices de dispersión tienen una apariencia similar. Ambas matrices convergen cuando el número de grados de libertad crece y los valores de los a_i son aproximadamente iguales, como ocurre con las aproximaciones de la sección 3.

La aplicación del método de minimizar la divergencia de Kullback-Leibler para aproximar la distribución marginal de σ^2 de la mixtura (4.1), conduce a las ecuaciones segunda y tercera del teorema 4.1, como era lógico esperar. Los mismos resultados se obtendrían si en vez de haber considerado la varianza como parámetro se hubiese considerado la precisión. Sin embargo, los resultados de la aplicación del método de los momentos dependería, en gran manera, de la parametrización elegida.

5. COMENTARIOS Y DISCUSION

Ya hemos comentado en las secciones anteriores que la principal ventaja del modelo de Snyder y sus generalizaciones es, precisamente la de incluir una única fuente de aleatoriedad en el modelo. Esta ventaja podría, por otra parte, constituir un inconveniente sobre todo si lo comparamos con el modelo general de Harrison y Stevens en el que, en principio las dos fuentes de error $\mathbf{v}_t$ y $\mathbf{w}_t$ pueden ser independientes o correladas, mientras que ahora las dos fuentes de aleatoriedad serían siempre correladas, salvo que el vector de efectos permanentes fuese nulo. De hecho, uno de los problemas cruciales a la hora de construir un modelo de este tipo es el de especificar el vector de efectos permanentes. Este problema es similar al de estimar los parámetros de un proceso de medias móviles (véase, p.e., Broemeling (1985)) en el que la dificultad mayor está en la no linealidad del modelo (en el análisis del modelo aparece la distribución de un producto de dos distribuciones normales independientes que no tiene expresión analítica conocida). Este es un problema en el que sería importante encontrar buenas aproximaciones a dicha distribución, p.e., en términos de mixturas, que permitieran un tratamiento análogo al de la sección 3.

Otra observación importante se refiere al modelo considerado en la sección 3; en esta sección se ha hecho hincapié en que las probabilidades de

clasificación, tanto exactas como aproximadas, se refieren a las componentes de los términos de error u_t, *no* a que las observaciones y_t sean o no outliers. La razón es que, debido a la forma del modelo (2.1), una realización anómala en el término de error u_t no sólo influye en que lo sea la observación y_t sino también la inmediatamente siguiente y_{t+1}, dependiendo su magnitud o influencia del vector de efectos permanentes. Por lo tanto, desde esta perspectiva es imposible atribuir la presencia de una observación anómala a (2.1a) o a (2.1b) a no ser, obviamente, que el vector de efectos permanentes fuese nulo.

Una manera de proceder podría ser el considerar la siguiente modificación del modelo de Snyder:

$$y_t = \mathbf{x}_t' \theta_t + u_t$$

$$\theta_t = \mathbf{T}\, \theta_{t-1} + \alpha u_t;$$

en el que la presencia de una observación anómala en u_t influiría directamente en y_t y no en la futura observación y_{t+1}. El estudio de este modelo sería similar al de las secciones 2 y 3.

6. AGRADECIMIENTOS

Este trabajo ha sido subvencionado por la *Consejería de Educación de la Junta de Andalucía.*

APENDICE

En este apéndice damos las demostraciones de los tres lemas de la sección 3.

Lema 3.1. La demostración se basa en el cálculo de la función característica del vector **Z** y en la independencia de **X** e **Y**. Si particionamos el vector **Z** y la matriz **A** de manera obvia, se tiene que

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11}\mathbf{X} + \mathbf{A}_{12}\mathbf{Y} \\ \mathbf{A}_{21}\mathbf{Y} + \mathbf{A}_{22}\mathbf{X} \end{pmatrix}.$$

Si $\mathbf{t}' = (\mathbf{t}_1', \mathbf{t}_2')$, entonces la función característica del vector **Z** es

$$\psi_{\mathbf{Z}}(\mathbf{t}) = E\left\{ e^{i\mathbf{t}'\mathbf{Z}} \right\} = E\left\{ e^{i\mathbf{t}_1'\mathbf{Z}_1 + i\mathbf{t}_2'\mathbf{Z}_2} \right\}$$

$$= E\left\{ e^{i\mathbf{t}_1'(\mathbf{A}_{11}\mathbf{X} + \mathbf{A}_{12}\mathbf{Y}) + i\mathbf{t}_2'(\mathbf{A}_{21}\mathbf{X} + \mathbf{A}_{22}\mathbf{Y})} \right\}$$

$$= E\left\{ e^{i(t_1'A_{11}+t_2'A_{21})X + i(t_1'A_{12}+t_2'A_{22})Y} \right\}$$

$$= \psi_X(t_1'A_{11} + t_2'A_{21}) \cdot \psi_Y(t_1'A_{12} + t_2'A_{22})$$

$$= \left[\sum_i \lambda_i \exp\left\{ i(t_1'A_{11} + t_2'A_{21})\mu_i - \frac{1}{2}(t_1'A_{11} + t_2'A_{21})\Sigma_i(A_{11}'t_1 + A_{21}'t_2) \right\} \right] \cdot$$

$$\cdot \left[\sum_j \lambda_j^* \exp\left\{ i(t_1'A_{12} + t_2'A_{22})\mu_j^* - \frac{1}{2}(t_1'A_{12} + t_2'A_{22})\Sigma_j^*(A_{12}'t_1 + A_{22}'t_2) \right\} \right] \cdot$$

$$= \sum_{ij} \lambda_i \lambda_j^* \exp\left\{ i(t_1', t_2') \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} - \frac{1}{2}(t_1', t_2') \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \right\},$$

donde, tras algunos cálculos, se demuestra que

$$B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} = \begin{pmatrix} A_{11}X + A_{12}Y \\ A_{21}Y + A_{22}X \end{pmatrix} \begin{pmatrix} \mu_i \\ \mu_j^* \end{pmatrix}$$

y

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} \Sigma_i & O \\ O & \Sigma_j^* \end{pmatrix} \begin{pmatrix} A_{11}' & A_{21}' \\ A_{12}' & A_{22}' \end{pmatrix}.$$

Como la función característica determina unívocamente la distribución y ésta corresponde a la mixtura de normales multivariantes, de aquí se sigue el lema.

Lema 3.2. La demostración se sigue inmediatamente de la definición de densidad condicionada.

Lema 3.3. Hay que probar primero que $\mathbf{x}_2 \mid \mathbf{x}_1, \sigma^2 \sim N(\mathbf{m}_{2\,1}, \sigma^2 \mathbf{C}_{2\,1})$, lo que es evidente de la hipótesis del lema y las propiedades de la normal multivariante. Y a continuación, que la distribución de $\sigma^2 \mid \mathbf{x}_1$ es una gamma invertida de parámetros los especificados. En efecto, por el teorema de Bayes, se tiene que

$$p(\sigma^2 \mid \mathbf{x}_1) \propto p(\mathbf{x}_1, \sigma^2) = p(\mathbf{x}_1 \mid \sigma^2) \cdot p(\sigma^2)$$

y como $\mathbf{x}_1 \mid \sigma^2 \sim N(\mathbf{m}_1, \sigma^2 \mathbf{C}_{11})$ y $\sigma^2 \sim Gal(a,p)$, operando se concluye el lema.

REFERENCIAS

ABRAHAM, B. and BOX, G. E. P. (1978). Linear models and spurious observations. *Appl. Statist.*, **27**, 120-130.

BERNARDO, J. M. and GIRON, F. J. (1988). A Bayesian approach to cluster analysis. *Qüestió.*, **12**, n.º 1, pp. 97-112.

BERNARDO, J. M. and GIRON, F. J. (1988). A Bayesian analysis of simple mixture problems. In *Bayesian Statistics* **3**. (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds.), (with discussion). pp. 67-78.

BOX, G. E. P. and JENKINS, G. M. (1970). *Time Series Analysis, Forecasting and Control.* Holden Day, San Francisco.

BOX, G. E. P. and TIAO, G. C. (1968). A Bayesian approach to some outlier problems. *Biometrika,* **55,** 119-129.

BOX, G. E. P. and TIAO, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading.

BROEMELING, L. D. (1985). *Bayesian Analysis of Linear Models.* Marcel Dekker, New York.

FREEMAN, P. R. (1980). On the number of outliers in data from a linear model. In *Bayesian Statistics* **1** (J. M. Bernardo, D. V. Lindley, M. H. DeGroot and A. F. M. Smith, eds.) pp. 349-365. Valencia University Press.

GUTTMAN, I., DUTTER, R. and FREEMAN, P. R. (1978). Care and handling of univariate outliers in the general linear model to detect spuriosity-A Bayesian approach. *Technometrics,* **20,** 187-193.

GUTTMAN, I. and PEÑA, D. (1985). Robust filtering. *J. Amer. Statist. Ass.* **80,** 91-92.

HARRISON, P. J. and AMEEN, J. R. M. (1985). Normal discount Bayes models. In *Bayesian Statistics* **2** (J. M. Bernardo, D. V. Lindley, M. H. DeGroot and A. F. M. Smith, eds.) pp. 271-298. North Holland: Amsterdam.

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

JOHNSON, W. and GEISSER, S. (1982). Assesing the predictive influence of observations. In *Statistics and Probability: Essays in Honor of C. R. Rao,* (G. Kallianpur, P. R. Krishnaiah and J. K. Ghosh, eds.) pp. 343-358. New York: North Holland.

JOHNSON, W. and GEISSER, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis. *J. Amer. Statist. Ass.,* **78,** 137-144.

JOHNSON, W. and UTTS, J. (1986). Bayesian robust estimation of the mean. *Appl. Statist.,* **35**, 63-72.

MURRAY, G. D. (1979). The estimation of multivariate normal density functions using incomplete data. *Biometrika,* **66**, 375-380.

PETTIT, L. I. and SMITH, A. M. F. (1984). Outliers and influential observations in linear models. In *Bayesian Statistics* **2** (J. M. Bernardo, D. V. Lindley, M. H. DeGroot and A. F. M. Smith, eds.) pp. 473-494. North Holland: Amsterdam.

ROJANO, J. C. (sin publicar). 'Métodos bayesianos aproximados para mixturas de distribuciones. (Tesis doctoral).

SNYDER, R. D. (1985). Recursive estimation of dynamic linear models. *J. Roy. Statist. Soc. B,* **47,** 272-276.

WEST, M. (1981). Robust sequential approximate Bayesian estimation. *J. Roy. Statist. Soc. B,* **43,** 157-166.

WEST, M. (1984). Outlier models and prior distributions in Bayesian linear regression. *J. Roy. Statist. Soc. B,* **46,** 431-439.

WEST, M., HARRISON, P. J. and MIGON, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting (with discussion). *J. Amer. Statist. Ass.,* **80,** 73-97.

LINEAR DYNAMIC MODELS AND MIXTURES OF DISTRIBUTIONS

SUMMARY

When modelling a time series by means of a DLM, one of the most important aspects of it is the specification of the error sources associated to the observation and system equations. Snyder's model includes only a source of randomness, thus reducing the problem of specification to that of the permanent effects vector. Starting from this model, we consider two important extensions of it: the first allows for the computation of the common variance of the error terms and the second, and more important, makes provision for non normal errors. From the many possible models for the error terms, we consider the multiplicative model, the extension to other models being straightforward. By means of approximation techniques to the distributions generated from the model, namely mixtures of some distributions, we generalize the Kalman filter, keeping its basic structure, without greatly increasing the computational effort.

Key words: approximation of mixtures; dynamic linear models; mixtures of distributions; outliers; robust Kalman filter.

AMS Subject Classification: 62M20, 62F15.

COMENTARIOS

JOSE M. BERNARDO
Departamento de Estadística, Presidencia de la Generalidad Valenciana (*)

El brillante trabajo que Girón, Martínez y Rojano nos han ofrecido aporta, en el campo concreto de los modelos dinámicos, avances concretos en las líneas de investigación centrales a la estadística matemática de los años 80; específicamente, (i) extensiones del modelo teórico capaces de incorporar observaciones atípicas y (ii) aproximaciones analíticas que reducen a expresiones calculables resultados teóricos intratables. Por otra parte, la comparación de estos resultados con los de Snyder (1985) ejemplifica, una vez más, la potencia de la metodología Bayesiana, proporcionando elegantes generalizaciones de los resultados tradicionales.

Para muchos de sus lectores, sin embargo, el trabajo puede haber quedado a un nivel excesivamente teórico; hubiera resultado atractiva la inclusión de un ejemplo específico, que incluyera una discusión detallada de la verificación de la hipótesis del modelo, de la asignación de los distintos parámetros que contienen, (en particular de los que describen la distribución inicial), y de la bondad de las aproximaciones propuestas, cuestiones todas ellas no triviales.

La consideración del trabajo desde el punto de vista de su aplicabilidad sugiere varios interrogantes.

(i) *Series multivariantes.* El objeto fundamental de la mayor parte de los análisis de series temporales con modelos dinámicos es disponer de un

(*) En Servicios Especiales procedente del Departamento de Estadística de la Universidad de Valencia.

instrumento fiable de predicción capaz de detectar rápidamente un cambio de tendencia. Esto suele implicar que únicamente las últimas observaciones de la serie estudiada y *las de otras series correlacionadas con ella* contienen información útil. Esto significa que la generalización del trabajo presentado a series multivariantes no es una cuestión puramente académica, sino mas bien una exigencia para su aplicación realista. Así, no parece razonable excluir del análisis de la evolución de la tasa de paro el de la tasa de población activa, estudiar el desarrollo de las transaminasas de un paciente sin considerar el de su colesterol, o predecir la evolución de la tendencia de voto hacia un partido político sin considerar la evolución de la tendencia de voto hacia los demás.

año	1983			1984				
tasa de paro	18.19	16.37	17.28	17.91	18.36	18.48	19.74	21.14
tasa de actividad	50.00	48.75	49.21	49.20	49.30	48.79	48.62	49.14

año	1985			1986				
tasa de paro	21.36	20.37	21.55	19.79	19.81	19.48	20.12	18.86
tasa de actividad	48.85	48.55	48.45	48.33	48.67	48.57	48.41	48.67

año	1987			1988			
tasa de paro	19.61	19.83	18.32	18.31	17.97	18.10	16.92
tasa de actividad	49.34	50.46	50.44	50.83	50.93	50.97	50.80

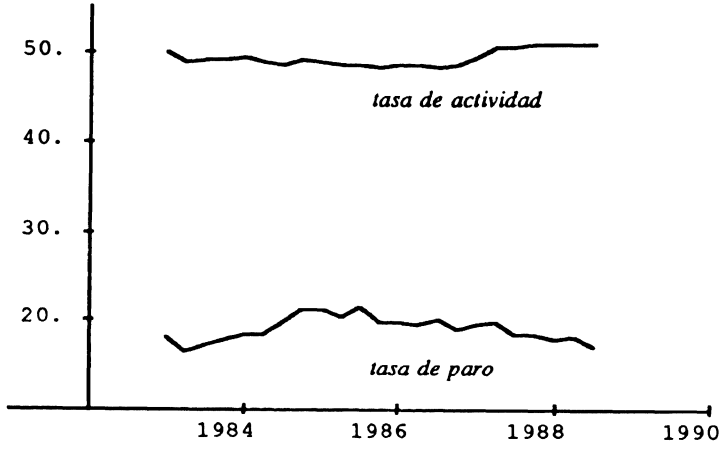


Figura 1. Evolución de las tasas de desempleo y de población activa en la Comunidad Valenciana (1983–1988). Fuente: INE, *Encuesta de la población activa.*

Por ejemplo, en la Figura 1 pueden observarse las evoluciones trimestrales desde 1983, correspondientes a la Comunidad Valenciana, de las tasas de desempleo y población activa. La Figura 2(a), que describe una tasa en función de la otra, pone de manifiesto su fuerte interdependencia: exceptuando el dato correspondiente al segundo trimestre de 1983, que claramente constituye una observación atípica, existe una clara relación probabilística inversa, aproximadamente lineal, entre ambas tasa. Resulta interesante observar (Figura 2(b)) la sucesión cronológica correspondiente a la nube de puntos descrita en la Figura 2: entre 1983 y 1985, hay una fuerte

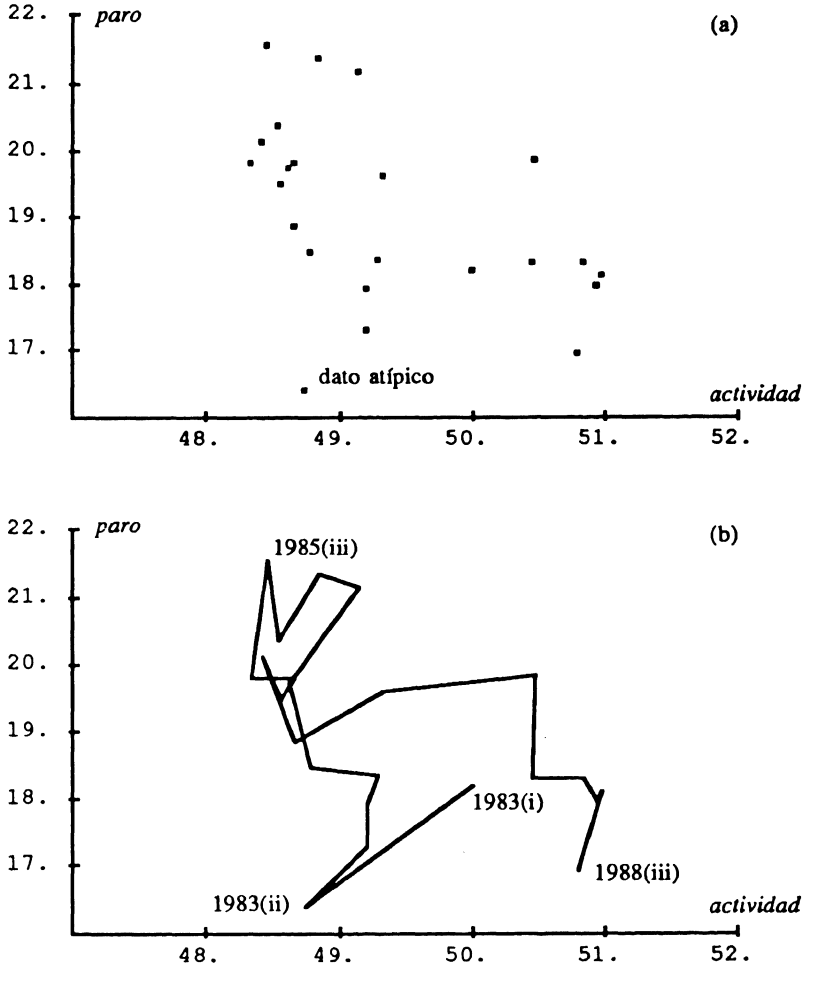


Figura 2. Relación entre las tasas de paro y de población activa en la Comunidad Valenciana (1983–1988).

tendencia a aumentar el paro con una pequeña reducción de la tasa de actividad; a partir del tercer trimestre de 1985, *la tendencia se invierte* radicalmente, observándose una disminución progresiva del paro, acompañada de un preceptible aumento de la tasa de actividad.

(ii) *Estructura de la matriz de covarianzas*. El modelo de mixtura propuesto para el análisis de series temporales univariantes es atractivo, pero parece necesario desarrollar técnicas que permitan verificar su adecuación a una serie concreta. En el caso multivariante, la situación es notablemente más compleja. Para empezar a entender la estructura de la matriz de covarianzas en situaciones como las descritas en el ejemplo anterior, resulta interesante analizar los residuos correspondientes a funciones ajustadas a las series temporales analizadas, como función de las magnitudes consideradas; los resultados obtenidos típicamente sugieren, si se quiere ser realista, la necesidad de modelizar estructuras sofisticadas para las covarianzas.

(iii) *Distribuciones Iniciales*. El uso de distribuciones iniciales conjugadas facilita enormemente el análisis, pero puede condicionar seriamente los resultados, especialmente en el caso de considerar simultáneamente muchos parámetros. Puesto que, como ya hemos apuntado, es frecuente que sólo tenga sentido trabajar con series muy cortas, la importancia de la distribución inicial se magnifica; si se consideran además las dificultades que lleva consigo la estimación subjetiva de una distribución inicial multivariante, es fácil concluir que un análisis realista exige la identificación de una distribución de referencia adecuada. El problema no es sencillo, pero está emergiendo una metodología general capaz de resolverlo. Véase Berger y Bernardo (1988, 1989).

Quisiera terminar con un ruego; el impacto, tanto científico como social, de la estadística matemática, especialmente el de sus métodos más sofisticados, depende radicalmente de la existencia de programas, preferentemente para microordenadores, que sean capaces de implementarlos. Espero que los autores de este trabajo nos ofrezcan pronto los programas correspondientes a los resultados que han obtenido.

ADDENDUM

Ante el extraño comportamiento de la tasa de desempleo en la Comunidad Valenciana en el segundo trimestre de 1983, dirigimos una consulta al Instituto Valenciano de Estadística (IVE). Como el análisis sugería, el dato que aparece en el anuario del IVE de 1987, del que se tomaron los datos, es incorrecto; el verdadero valor es 17.36, no 16.37, con lo que 1983 (ii)

deja de ser una observación atípica. Esta anécdota, constituye un buen ejemplo de la potencia del análisis de datos para el filtrado de los inevitables errores contenidos en las bases de datos reales.

REFERENCIAS ADICIONALES

BERGER, J. O. and BERNARDO, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* 84, pp. 200 - 207.

BERGER, J. O. and BERNARDO, J. M. (1989). Ordered group reference priors with applications to multinomial and variance components problems. *Tech. Rep.* 01/89, *Dept. Estadística.* Presidencia de la Generalidad Valenciana, Spain.

ELIAS MORENO

Departamento de Estadística e I.O.
Universidad de Granada.
Granada. España

LUIS RAUL PERICCHI

Departamento de Estadística
Universidad Simón Bolivar
Caracas. Venezuela

Vaya de antemano nuestra felicitación a los autores de este excelente artículo, al que a buen seguro le seguirán otros en la misma linea, por el punto de vista adoptado y por la claridad de exposición tanto de motivos como de resultados.

Por lo que respecta al modelo que sirve de base al artículo solo tenemos que decir que toda la inferencia que se lleva a cabo es condicional al vector α de efectos permanentes. De hecho los autores en sus Comentarios y Discusión ya anuncian "... uno de los problemas cruciales a la hora de construir un modelo de este tipo es la especificación del vector de efectos permanentes". A pesar de ello no se aporta una solución a tal problema.

El hecho observado por los autores de que ni la normal ni la normal-gamma pierdan su caracter a lo largo del tiempo es un resultado importante sobre todo cara a la tratabilidad del análisis. Sin embargo, esta familia no admite elementos con colas "pesadas" lo cual es un serio inconveniente si se tiene en cuenta que la parte más insegura en el proceso de especificación de una distribución a priori es esa referida a la tasa con que la densidad tiende a cero. Esto obliga a considerar familias capaces de absorver esta incertidumbre a priori. En este orden de ideas, la consideración en

el artículo de la mixtura de normales $(1-\lambda)N(0,\sigma^2) + \lambda N(0,k^2\sigma^2)$ es interesante, sobre todo desde el momento en que obtienen una aproximación razonable y sencilla de tratar bajo el punto de vista de la computabilidad.

En general, mixturas de escala de distribuciones normales resuelven el problema dando lugar a t-distribuciones. La distribución o familia de distribuciones que efectúa la mixtura, (en el trabajo $(1-\lambda)\,\delta_{\sigma 2} + \lambda\,\delta_{k2\,\sigma 2}$, con $\delta_{\sigma 2}$ la masa unidad concentrada en σ^2 y k^2 conocido), estará siempre sujeta a arbitrariedad y condicionada a su manejabilidad.

Dos tipos de familias que son razonables, tanto por el conocimiento a priori que modelizan como por su tratabilidad analítica y numérica son las siguientes:

1) La clase de todas las distribuciones con algunos cuantiles fijados; o equivalentemente la clase de todas las distribuciones tales que algunos conjuntos mantienen constante su medida en todos los elementos de la familia. Sin excesivas complicaciones pueden añadirse resticciones de forma tales como unimodalidad y esfericidad.

2) La clase de todas las distribuciones que están comprendidas en un "intervalo" prefijado de medidas.

En ambos casos pueden calcularse rangos de esperanzas a posteriori sin gran esfuerzo de cálculo. (véase Moreno y Pericchi (1988) y DasGupta y Studden (1988a, 1988b), respectivamente). Sería interesante obtener rangos de $E(\theta_i|\text{datos})$, lo que supondría un diagnóstico natural de la robustez del modelo, dados los datos.

Creemos, por último, que sería deseable que este artículo contara con una ilustración práctica que nos permitiera observar el comportamiento del modelo toda vez que exigiría de los autores una propuesta de solución en torno a α.

REFERENCIAS

DASGUPTA A and STUDDEN W. J. (1988a) "Robust Bayesian Analysis and Optimal Experimental Designs in Normal Linear Models with Many Parameters I". Technical Report. Dept. of Statistics, Purdue University.

DASGUPTA A and STUDDEN W. J. (1988b) "Robust Bayesian Analysis and Optimal Experimental Designs in Normal Linear Models with Many Parameters II". Technical Report. Dept. of Statistics, Purdue University.

MORENO E. y PERICCHI L. R. (1988) "A Robust Bayesian Elimination of Hyperparameters". Informe Técnico. Dpto. Estadística. Universidad de Granada.

DANIEL PEÑA
Universidad Politécnica de Madrid

En primer lugar, quiero felicitar a los autores por la excelente presentación de sus contribuciones a un tema importante y actual. Mi comentario va a centrarse, por razones de brevedad, en dos puntos: el modelo utilizado y los métodos para aproximar mezclas de distribuciones.

Comenzando con el primer aspecto, la formulación del modelo de espacio de los estados en términos de un sólo vector de innovaciones ha sido utilizada previamente por Akaike (1976) y Aoki (1976), entre otros. Sneyder introduce esta formulación como una restricción en el modelo general, lo que es incorrecto: todo modelo dinámico puede representarse indistintamente en ambas formulaciones. Este resultado, implícito en la presentación de Aoki (1987), puede demostrarse geométricamente como sigue: supongamos el modelo general:

$$y_t = x'_t \beta_t + v_t \tag{1}$$
$$\beta_t = T \beta_{t-1} + w_t$$

donde β_t es el vector de variables de estado y v_t y w_t son procesos incorrelados de media cero. Definamos:

$$e_t = y_t - \hat{y}_{t/t-1} = y_t - x'_t \beta_{t/t-1} \tag{2}$$

la innovación o error de predicción a un paso del modelo, es decir, el error al prever y_t con la información disponible hasta t-1 (el conjunto de observaciones $y_1,..., y_{t-1}$). Sea

$$\Theta_t = \beta_{t/t-1} = E[\beta_t | y_{t-1},..., y_1] \tag{3}$$

un nuevo vector de variables de estado obtenido mediante la proyección ortogonal de β_t sobre el espacio definido por las observaciones hasta t-1. Entonces:

$$\Theta_{t+1} = E[\beta_{t+1} | y_t,..., y_1] = E[\beta_{t+1} | e_t; y_{t-1},..., y_1]$$

ya que, por (2), y_t es la suma de e_t y un término que, por (3), está en el espacio generado por $y_{t-1},..., y_1$. Por tanto $(y_t,... y_1)$ y $(e_t; y_{t-1},..., y_1)$ definen el mismo espacio. Además, e_t es ortogonal al espacio definido por $y_{t-1},..., y_1$ ya que:

$$E[y_t | y_{t-1},..., y_1] = x'_t E[\beta_t | y_{t-1},..., y_1] = x'_t \beta_{t/t-1}$$

y, por tanto, e_t es la diferencia entre y_t y su proyección ortogonal sobre $y_{t-1},\ldots, y_1$. Entonces:

$$E[\beta_{t+1} \mid e_t; y_{t-1},\ldots, y_1] = E[\beta_{t+1} \mid e_t] + E[\beta_{t+1} \mid y_{t-1},\ldots, y_1] \tag{4}$$

ya que los coeficientes de la regresión múltiple son iguales a los de las regresiones simples con variables independientes, (véase una demostración formal en Aoki (1987), pág. 80); además:

$$E[\beta_{t+1} \mid e_t] = \mathrm{Var}(e_t)^{-1}\, \mathrm{Cov}(e_t \beta_{t+1})\, e_t = \alpha\, e_t$$

donde α representa los coeficientes de regresión entre los componentes del vector de estado y la innovación; el otro término es

$$\begin{aligned} E[\beta_{t+1} \mid y_{t-1},\ldots, y_1] &= T\, E[\beta_t \mid y_{t-1},\ldots, y_1] + E[w_{t+1}] \\ &= T\, \beta_{t \mid t-1} = T\, \Theta_t \end{aligned}$$

con lo que, sustituyendo en (4), se obtiene la nueva ecuación de estado:

$$\Theta_{t+1} = T\, \Theta_t + \alpha\, e_t \tag{5}$$

mientras que (2) proporciona la ecuación de predicción:

$$y_t = x'_t\, \Theta_t + e_t \tag{6}$$

En resumen, cualquier modelo del tipo (1) puede escribirse de forma equivalente de acuerdo con (5) y (6), donde: (a) aparece únicamente un término de error que es, además el error de predicción a un paso; (b) las variables de estado son proyecciones ortogonales sobre el espacio de datos observados de las variables primitivas.

Respecto al segundo problema, cómo evitar la multiplicación de la distribución a posteriori cuando se combinan distribuciones mezcladas, la sugerencia heurística propuesta por Harrison y Stevens de ajustar una distribución normal igualando momentos ha sido justificada por Peña y Guttman (1989) que han demostrado que, trabajando con distribuciones normales, este procedimiento es óptimo con la medida de distancia de Kullblack-Leibler. El enfoque de este trabajo es distinto e interesante en dos aspectos: en primer lugar suponen σ^2 desconocida y trabajan con la distribución t en lugar de hacerlo con la normal; en segundo, sugieren hacer primero la aproximación sobre la distribución a posteriori conjunta y obtener después la marginal. El enfoque resulta atractivo y sería interesante investigar desde un punto de vista práctico en qué situaciones se obtienen ventajas apreciables respecto al método anterior.

Por último, en Kitagawa (1987), Spall (1988), Peña y Guttman (1988) y Meinhold y Singpurwalla (1989), se presentan resultados alternativos o complementarios a los propuestos en este interesante trabajo.

REFERENCIAS

AKAIKE, H. (1976). "Canonical Correlation analysis of Time Series and the use of an Information criterion", en Mehra y D. Lainiotis eds. *System Identification: Advances and case studies*. Academic Press.

AOKI, M. (1976). *Optimal Control and System Theory in Dynamic Analysis*, N. H. Amsterdam.

AOKI, M. (1987). *State Space Modeling of Time Series*. Springer-Verlag.

PEÑA, D. y GUTTMAN, I. (1988). "Robust Recursive estimation using Kalman Filtering and its applications". en *Bayesian Analysis of Time Series and Dynamic Models*, J. Spall editor, Marcel Dekker, Cap. 9, pp: 227-255.

PEÑA, D. y GUTTMAN, I. (1989). "Optimal collapsing of mixture distributions in Robust Recursive Estimation". *Communications in statistics, (Theory and Methods)* 18, 3, 817-834.

KITAGAWA, G. (1987). "Non Gaussian State-Space Modeling of Nonstationary Time Series" (con discusión). *Journal of American Statistical Association*, 82, 400, 1032-1064.

MEINHOLD, R. J. y SINGPURWALLA, N. (1989). "Robustification of Kalman Filter Models". *Journal of American Statistical Association*, 84, 406, 479-486.

SPALL, J. (1988). *Bayesian Analysis of Time Series and Dynamic Models*. Marcel Dekker.

FERNANDO TUSELL
Universidad del País Vasco

Es una satisfacción poder comentar un trabajo de espléndida factura, como el de GIRON et al. (1988). Vaya por delante mi felicitación a los autores que han conseguido un tratamiento notablemente diáfano en un desarrollo en que la notación tiende inevitablemente a hacerse compleja.

Quiero solo hacer algunos comentarios sobre una cuestión que me parece necesitada todavía de investigación. Me referiré, por simplicidad, al modelo presentado por los autores en el caso en que σ^2 es conocida. En este caso, el resultado básico es el proporcionado por el Teorema 3.1, que proporciona la densidad *a posteriori* de $\theta | \sigma^2, y_1, \ldots, y_t$ como una mixtura de $t_0 2^t$ distribuciones normales.

Como ya indican los autores, este resultado es de inviable aplicación en la práctica tal y como se propone, pues requeriría tiempo de cálculo y

memoria creciendo exponencialmente con t. Se impone pues algún tipo de simplificación, y una que se sugiere es la de aproximar cada mixtura por una normal en cada momento t (Teorema 3.2).

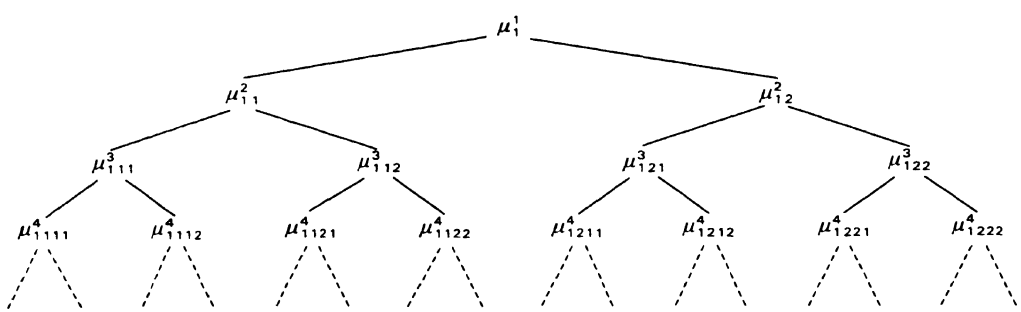
La cuestión es entonces: ¿es buena esta aproximación?. ¿Bajo que condiciones lo es?. Parece, por ejemplo, que debiera ser una buena solución si k_0^2 no es mucho mayor que 1, y/o λ_0 es muy pequeña. Por el contrario, si la distribución generadora de *outliers* es considerablemente diferente de la que genera las u_t bajo regimen normal, y la probabilidad de aparición λ_0 de los mismos es apreciable, la aproximación podría resentirse notablemente.

Desde el punto de vista de un posible usuario, las siguientes preguntas tendrían repuestas de interés:

1. La aproximación reiterada de una mixtura en cada momento t por una única distribución normal, ¿es el mejor procedimiento?. ¿Cabría la posibilidad, por ejemplo, de que propagando las mixturas a lo largo de s unidades de tiempo y efectuando la aproximación en los momentos $t+ns$, $(n = 1,2,...)$ los resultados mejoraran?. El hacerlo solo requeriría $O(2^s)$ memoria adicional.

2. ¿Cabría la posibilidad de reducir la complejidad del algoritmo propagando solo *una porción* de la mixtura que en cada momento t se genera?.

En relación a esta segunda cuestión, la idea directriz podría bosquejarse como sigue. Si comenzamos, por simplicidad, con un único μ_1^1 ($t_0 = 1$ en el Teorema 3.1), sucesivas iteraciones van produciendo coeficientes $\mu^t_{i_0 i_1 \ldots i_t}$ de acuerdo con un proceso que puede esquematizarse en el siguiente árbol binario:



Se puede imaginar como posibilidad el recurrir a la *poda* de dicho árbol binario conforme se va profundizando, despreciando algunas de sus ramas y propagando las demás. Esta idea viene sugerida por algo que es habitual en Inteligencia Artificial; la exploración de un árbol de decisiones en profundidad abandonando aquellas alternativas que, examinadas en unos cuantos niveles, no parecen prometedoras. Los programas que juegan a ajedrez, entre otros, recurren de ordinario a este tipo de estrategia *(selective pruning).*

No está claro el criterio que debería seguirse en esta poda. Parece que coeficientes $\mu^t_{ij,i_1,\ldots,i_l} \simeq 0$ son candidatos idóneos a ser eliminados; pero pueden no serlo si están asociados a términos de la mixtura lo suficientemente "diferentes" (en términos de distancia de Kullback-Leibler, quizás) de los restantes.

Si la poda es lo suficientemente intensa, la mixtura a propagar podría mantenerse con un número manejable de términos. Cuántos y cuáles hayan de ser éstos es, no obstante, una cuestión que parece ardua de solventar; y aún más difícil si la solución que se adopte ha de ser susceptible de implementarse en el contexto de un algoritmo capaz de hacer predicción en tiempo real.

REFERENCIAS

GIRON, F. J. et al. (1988) "Modelos lineales dinámicos y mixturas de distribuciones", en este número.

CONTESTACION

En primer lugar queremos agradecer globalmente a todos los autores que han comentado nuestro artículo por la favorable acogida que ha tenido, a pesar de nuestro temor inicial de que la notación, necesariamente farragosa, ocultara las ideas y líneas esenciales del trabajo. En segundo lugar, muchos de sus comentarios nos brindan nuevas ideas sobre posibles extensiones del modelo que lo hagan más flexible y de mayor aplicabilidad. Algunas de estas ideas ya estaban en fase de desarrollo, mientras que otras que nos han sugerido sus comentarios nos parecen interesantes y dignas de tener en cuenta en sucesivos trabajos.

Como los comentarios se centran en aspectos diferentes del artículo, vamos a contestar por separado a cada uno de ellos.

Agradecemos al Profesor Daniel Peña su amabilidad al aceptar comentar nuestro trabajo sobre un tema en el que es un reconocido experto y al que ha contribuido con varios artículos. El haber tomado como modelo de partida el de Snyder (1985) tuvo su origen, precisamente, en el problema de estimar secuencialmente la varianza de la única fuente de error σ^2, de modo análogo a como se estiman recursivamente los parámetros del modelo mediante el filtro de Kalman. Es interesante comprobar, como nos demuestra Daniel Peña en sus comentarios, que todo modelo lineal dinámico se puede representar de esa manera. Este resultado es revelador pues nos da una interpretación del vector de efectos permanentes y de los términos de error de la única fuente de variación. Nos gustaría conocer si ese resultado, o uno similar, es también válido para series o datos multivariantes.

Respecto del problema de aproximación de mixturas normales-gamma invertidas, que tiene implicaciones obvias en la aproximación de mixturas de distribuciones *t* de Student, nos congratula saber que empleamos una

técnica análoga a la de Peña y Guttman (1989) para el caso normal. La idea de utilizar la divergencia de Kullback-Leibler en nuestro artículo se debe, aparte de otras razones bien conocidas, a la de que es invariante respecto de posibles reparametrizaciones. Concretamente, si en vez de utilizar como parámetro σ^2, hubiésemos estado interesados en la precisión $\tau = \sigma^{-2}$, o en la desviación típica σ, la minimización de la divergencia de Kullback-Leibler nos conduciría a los mismos resultados, a diferencia de otros métodos de aproximación de mixturas como pudiera ser el de los momentos, que es muy sensible a la reparametrización.

José Miguel Bernardo, a quien agradecemos sus elogios, más que comentar sobre nuestro trabajo, lo que sugiere es que nos pongamos a trabajar en un ambicioso proyecto en el cual hay una gran cantidad de problemas abiertos no triviales, como son la modelización de series multivariantes y la especificación de distribuciones iniciales para modelos dinámicos, un tema muy del agrado del Profesor Bernardo al que, desde estas líneas, invitamos a colaborar.

Respecto de la extensión del modelo a series multivariantes, se han conseguido algunos resultados en la misma línea que los de nuestro artículo (véase el reciente trabajo de Serrano (1989)), mediante un modelo de la forma

$$\mathbf{y}_t' = \mathbf{x}_t' \Theta_t + \mathbf{u}_t'$$

$$\Theta_{t+1} = \mathbf{T} \Theta_t + \alpha \mathbf{u}_t'$$

en el que $\mathbf{y}_t$ es un vector $p \times 1$ de variables observables, $\mathbf{x}_t$ es un vector $m \times 1$, $\mathbf{T}$ es una matriz $m \times m$, α es un vector $m \times 1$, Θ_t es una matriz $m \times p$ y $\mathbf{u}_t$ es un vector aleatorio $p \times 1$.

Este modelo, que aunque implica una estructura ciertamente algo rígida de la matriz de covarianzas de la serie temporal, tiene la ventaja de permitir un análisis similar al del caso univariante (las ecuaciones del filtro de Kalman son análogas), y contiene como casos particulares a los modelos de regresión y a los procesos autorregresivos multivariantes, entre otros.

Para finalizar nuestra respuesta a los comentarios del Profesor Bernardo, queremos indicarle que estamos poniendo a punto los programas correspondientes al modelo propuesto y su extensión multivariante; además de generalizaciones que incluyen el desconocimiento de la proporción de observaciones anómalas λ_0, la posibilidad de *outliers* aditivos, en (y a la) vez de multiplicativos. El modelo, como comentamos a continuación en respuesta a los Profs. E. Moreno y L. R. Perichi, es fácilmente generalizable al

caso de que la estructura de los errores venga dadà por distribuciones generadas por mixturas arbitrarias de distribuciones normales respecto del parámetro de escala, lo que permitiría suponer en los errores distribuciones con colas *pesadas*, como las *t*, las *estables simétricas* o la familia *exponencial de potencias*.

Como nos recuerdan los Profesores Moreno y Pericchi las inferencias que se llevan a cabo en el artículo son condicionadas, no sólo al vector de efectos fijos α sino también a λ_0. Como hemos comentado en el parágrafo, anterior, el filtro de Kalman también puede generalizarse y adaptarse para poder aprender secuencialmente sobre el parámetro de mezcla λ_0 (ver, p.e., Serrano (1989)). Sin embargo, la generalización del filtro al parámetro α no puede realizarse ni siquiera dentro del contexto de familias conjugadas extendidas al caso de mixturas, por lo que habría que recurrir a otros métodos de aproximación. Este es un problema abierto que merecería la pena estudiar. La interpretación dada al vector de efectos fijos por Daniel Peña en su comentario, podría ser útil. En nuestra opinión la dificultad reside, básicamente, en la complejidad de la distribución del producto de dos variables normales independientes y su posible aproximación por una distribución normal.

Desde el punto de vista de la robustez del modelo respecto de la distribución de los errores, su estudio se puede realizar siguiendo las líneas de nuestro trabajo, como puede verse en Girón (1989) para problemas de regresión y en Rojano (1990) para modelos lineales dinámicos, al caso de errores generados por mixturas de normales respecto del parámetro de escala, de la forma

i) Los $\{u_i\}$ son i.i.d. y siguen una distribución

$$\int N(u_i \mid 0, \lambda\sigma^2)\, dF(\lambda)$$

donde $F(\lambda)$ es una función de distribución arbitraria sobre $(0, \infty)$.

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

La sugerencia de considerar familias más generales que las mixturas de normales respecto del parámetro de escala en los modelos dinámicos, como las clases 1) y 2) de sus comentarios que se han utilizado ampliamente en estudios de robustez bayesiana para modelos estáticos, nos parece extraordinariamente interesante, siempre que la obtención de, p.e., rangos del vector de medias o moda a posteriori de $\theta_t \mid y_1, \ldots, y_t$ fuese computacionalmente sencilla como en los casos de inferencia estática.

Los comentarios del Profesor F. Tusell se centran en un punto especialmente importante y delicado de nuestro trabajo como es la cuestión de simplificar la distribución exacta de $\theta_t | \sigma^2 y_1, \ldots, y_t$.

La idea de aproximar en cada etapa la mixtura de normales por una normal, se podría extender, como acertadamente señala el Profesor Tusell, a s etapas. De hecho no sería necesario efectuar la aproximación en los instantes $t+ns$, sino únicamente al cabo de las s primeras etapas, que son las cruciales en el proceso de aproximación, como se ha señalado previamente en Bernardo y Girón (1989a,b) y se podría justificar de un modo teórico utilizando, por ejemplo, los resultados de Heyde and Johnstone (1979) que garantizan la normalidad asintótica de la distribución de $\theta_t | \sigma^2 y_1, \ldots, y_t$, cuando $t \to \infty$. Si tenemos en cuenta que la verdadera distribución es siempre una mixtura de normales, la lectura del resultado anterior nos dice que cuando t crece los términos de la mixtura con pesos mayores tienden a converger hacia una distribución (la asintótica) mientras que el resto de los términos de la mixtura con distribuciones alejadas de la asintótica tendrían pesos cada vez menores, es decir $\alpha^t_{i_0 i_1 \ldots i_t} \to 0$, con lo que se conseguiría lo que Tusell denomina la *poda* del árbol.

Todo lo anterior parece reforzar nuestra conjetura de que el problema de aproximación es básico en las s primeras etapas, donde s dependería de varios parámetros: λ_0 y k_0^2 tal como señala el Profesor Tusell, y de la dimensión del vector θ_t. En estas etapas podría ser interesante aplicar la idea de Tusell de examinar selectivamente parte del árbol y desechar aquellas ramas que tuvieran probabilidad pequeña. Otra idea que aquí queremos apuntar, y que está relacionada con la parte final de sus comentarios es la de ir aproximando dinámicamente la mixtura por otra de menor número de términos, combinando aquellos términos que o bien son próximos entre si (en términos de alguna divergencia o métrica) o tienen un peso muy pequeño (menor que un cierto umbral) y no son prometedoras en el sentido apuntado por Tusell. De acuerdo con los razonamientos del punto anterior, a medida que t creciera el número de términos de la mixtura iría decreciendo hasta llegar a uno.

Por último, queremos mencionar que en el reciente libro de West and Harrison (1989), también se presentan resultados alternativos a los de nuestro artículo.

REFERENCIAS ADICIONALES

BERNARDO, J. M. and GIRÓN, F. J. (1988a). A Bayesian approach to cluster analysis. *Qüestiió.*, 12, n. 1, pp. 97-112.

BERNARDO, J. M. and GIRÓN, F. J. (1988b). A Bayesian analysis of simple mixture problems. In *Bayesian Statistics* 3. (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), (with discussion), pp. 67-78. Oxford University Press: Oxford.

GIRÓN, F. J. (1989). Mixturas de distribuciones normales con aplicaciones a problemas estadísticos complejos. *Real Acad. Cien. Exac. Fis. Nat.*, (pendiente de publicación).

HEYDE, C. C. and JOHNSTONE, I. M. (1979). On asymptotic posterior normality for stochastic procesess. *J. Roy. Statist. Soc. B,* 41, 184-189.

SERRANO, J. (1989). Contribuciones a la teoría de los modelos lineales dinámicos con errores no normales. *Tesis doctoral.* Universidad de Málaga.

WEST, M. and HARRISON, J. (1989). *Bayesian Forecasting and Dynamic Models.* Springer-Verlag: New York.