ALGORITMOS ADAPTATIVOS DE GIBBS SAMPLING PARA LA IDENTIFICACIÓN DE HETEROGENEIDAD EN REGRESIÓN Y SERIES TEMPORALES

TESIS DOCTORAL

Autora: Ana Justel Eusebio

Director: Daniel Peña Sánchez de Rivera

UNIVERSIDAD CARLOS III DE MADRID

Departamento de Estadística y Econometría

Getafe, septiembre de 1995

A mis padres.

Índice

Resumen

El objetivo principal de esta tesis doctoral es desarrollar nuevos procedimientos para la identificación de observaciones atípicas que introducen heterogeneidad en muestras con datos independientes y dependientes. Se proponen dos algoritmos diferentes para los problemas de regresión y series temporales basados en el algoritmo de Gibbs Sampling.

Al igual que sucede con los métodos clásicos de identificación de valores atípicos, se demuestra que la aplicación estándar del Gibbs Sampling no proporciona una identificación correcta de estos valores atípicos en problemas que presentan grupos de observaciones atípicas enmascaradas. Dado un vector cualquiera de valores iniciales, teóricamente el algoritmo converge a la verdadera distribución a posteriori de los parámetros, sin embargo, la velocidad de convergencia puede ser extremadamente lenta cuando el espacio paramétrico tiene dimensión alta y los parámetros están muy correlacionados. Los nuevos algoritmos que se discuten en este trabajo permiten mediante un proceso de aprendizaje adaptar las condiciones iniciales del Gibbs Sampling y mejorar su convergencia a la distribución a posteriori de los parámetros del modelo.

En el primer capítulo se presenta la situación actual del problema de la identificación de observaciones atípicas en modelos de regresión y series temporales, así como una descripción del Gibbs Sampling y sus principales propiedades. Las contribuciones originales que se desarrollan en esta tesis doctoral se exponen en los capítulos 2, 3 y 4.

En el capítulo 2 se extiende la aplicación del Gibbs Sampling a la identificación de observaciones atípicas en regresión con un modelo lineal de contaminación de escala. Se demuestra que el efecto del *potencial* en los modelos de regresión puede provocar una convergencia extremadamente lenta del algoritmo en muestras que contienen grupos de atípicos influyentes. Si estos datos son considerados inicialmente como buenos, la solución que proporciona el algoritmo a lo largo de miles de iteraciones es errónea, indicando que aparentemente se ha alcanzado la convergencia. Los estimadores de los parámetros que se obtienen son sesgados y la identificación de los valores atípicos ignora los grupos de atípicos, identificándose únicamente los aislados. Como generalización del modelo de contaminación se propone un modelo bayesiano no parámetrico de contaminación de escala y nivel, y se obtienen las distribuciones condicionadas de los parámetros necesarias para la aplicación del Gibbs Sampling. Se demuestra que este modelo no mejora la convergencia del algoritmo cuando existen grupos de atípicos influyentes.

En el capítulo 3 se propone un algoritmo adaptativo de Gibbs Sampling que supera los problemas de convergencia detectados en el capítulo 2 en la identificación de grupos de observaciones atípicas. Este nuevo algoritmo consta de dos etapas en las que se adaptan las condiciones iniciales del Gibbs Sampling haciendo uso de la información que proporciona la matriz de covarianzas de ciertas variables de clasificación. Se ilustra con varios ejemplos como con este algoritmo se obtienen resultados buenos en situaciones extremas en las que fallan algunos de los procedimientos para la identificación de atípicos que se han propuesto más recientemente en la literatura.

En el capítulo 4 se analiza la aplicación del Gibbs Sampling estándar a la identificación de rachas de valores atípicos aditivos en procesos autorregresivos. Se demuestra que su convergencia es muy lenta y que únicamente se identifican los extremos de la secuencia de atípicos. Se propone un nuevo algoritmo adaptativo que permite detec-

tar en la primera etapa la posición de las rachas de atípicos mediante la ejecución del Gibbs Sampling y, en la segunda etapa, se adaptan las distribuciones a priori del modelo y las condiciones iniciales para incorporar esta información. La ejecución del Gibbs Sampling en la segunda etapa se realiza sobre un espacio paramétrico reducido que requiere el cálculo de las distribuciones condicionadas correspondientes a bloques de observaciones. El comportamiento del nuevo algoritmo se ilustra en un ejemplo.

Finalmente, en el capítulo 5 se presentan algunas conclusiones y se indican las líneas de investigación futuras a partir del trabajo desarrollado en esta tesis doctoral.

Las principales contribuciones que aporta este trabajo se pueden resumir en los siguientes puntos:

1. Extender la aplicación del Gibbs Sampling a la detección de observaciones atípicas con un modelo lineal de contaminación de escala.
2. Demostrar que el Gibbs Sampling estándar falla en la identificación de grupos de atípicos enmascarados en problemas de regresión, y que se pueden identificar como atípicas observaciones que no lo son.
3. Proponer un modelo bayesiano semiparamétrico de generación de datos atípicos en regresión y desarrollar la aplicación del Gibbs Sampling para obtener las distribuciones a posteriori de los parámetros del modelo.
4. Desarrollar un algoritmo adaptativo de Gibbs Sampling para identificar grupos de observaciones atípicas en situaciones generales de enmascaramiento.
5. Demostrar que el Gibbs Sampling estándar no permite identificar rachas de valores atípicos aditivos en procesos autorregresivos.
6. Obtener las distribuciones condicionadas del vector que clasifica los valores atípicos y del vector de los tamaños de un bloque de observaciones consecutivas en una serie temporal.

7. Desarrollar un algoritmo adaptativo de Gibbs Sampling para identificar rachas de valores atípicos aditivos en procesos autorregresivos.

Algunas de las principales aportaciones de esta tesis se pueden encontrar en los artículos redactados en inglés de Justel y Peña (1995a, 1995b).

AGRADECIMIENTOS

Quiero agradecer en primer lugar al profesor Daniel Peña la gran ayuda que me ha prestado en todo momento durante la elaboración de esta tesis. Su apoyo y confianza constantes me han impulsado a realizar este trabajo con interés e ilusión.

También quiero expresar mi agradecimiento a los miembros del Departamento de Estadística y Econometría de la Universidad Carlos III de Madrid que han puesto a mi disposición todos los medios necesarios para realizar este trabajo. En particular, a Pedro Delicado que me ha ayudado a resolver muchos problemas informáticos. Las discusiones mantenidas con Ruey Tsay, Christian Robert y Mike West han servido para mejorar los resultados que se recogen en esta tesis; a ellos también estoy agradecida, especialmente a los dos últimos por sus invitaciones y hospitalidad cuando visité el Institute of Statistics and Decision Sciences de la Universidad de Duke y el Center of Researchs in Economics and Statistics del INSEE en París.

Getafe, septiembre de 1995.

Capítulo 1

Introducción

1.1 La heterogeneidad en los datos económicos

El tipo de datos que se analizan en Economía están frecuentemente contaminados por valores atípicos ya que proceden de la observación de fenómenos que no son repetibles ni controlables, o están construidos en condiciones muy diversas de agregación. Por tanto, la identificación y tratamiento de observaciones atípicas es una condición indispensable para que el análisis econométrico conduzca a resultados fiables, ya que unos pocos datos atípicos pueden alterar profundamente las conclusiones del análisis (véase Peña y Sánchez Albornoz, 1984, o Peña y Ruiz Castillo, 1984). La heterogeneidad suele considerarse en Econometría como un fenómeno conocido que se modeliza explícitamente, y no como un problema potencial que puede ocurrir de manera imprevista y que se debe identificar (con métodos de diagnóstico) o cubrirse ante su posible presencia (con métodos robustos). Por ejemplo, la heterogeneidad se modeliza con parámetros que varían con el tiempo o con estudios de cambio estructural, pero la sensibilidad de estos procedimientos a un grupo pequeño de valores atípicos no detectados es, en general, desconocida.

Desde un punto de vista bayesiano se han desarrollado métodos para el diagnóstico y tratamiento de observaciones atípicas e influyentes aisladas, fundamentalmente, para modelos uniecuacionales estáticos. La inevitable complejidad que supone la obtención de las distribuciones a posteriori en modelos más avanzados (véase Steel, 1991) provoca que el desarrollo de métodos para la identificación de grupos de datos atípicos en series temporales o modelos econométricos dinámicos sea actualmente escaso. Sin embargo, la aparición en los últimos años de nuevos algoritmos de remuestreo de tipo Markoviano, entre los que destacan el algoritmo de Metropolis *et al.* (1953) —adoptado por Hastings (1970) para su aplicación en problemas estadísticos—, el algoritmo de Sustitución de Tanner y Wong (1987) y, principalmente, el Gibbs Sampling de Geman y Geman (1984), ha supuesto una revisión del problema con prometedoras perspectivas. Un ejemplo ilustrativo de esta nueva línea de trabajo es el estudio de modelos Tobit de regresión censurada de Chib (1992) y Geweke (1992) llevado a cabo mediante Gibbs Sampling, o las aplicaciones de este mismo algoritmo a modelos dinámicos en el espacio de los estados de Carlin et al. (1992), Carter y Kohn (1994) y Frühwirth-Schnatter (1994). Para la identificación de observaciones atípicas, Verdinelli y Wasserman (1991) y McCulloch y Tsay (1994a) desarrollan nuevos procedimientos basados en el Gibbs Sampling.

En los últimos años se ha puesto de manifiesto que los métodos de identificación de valores atípicos e influyentes individuales son insuficientes y pueden ser muy engañosos con determinadas estructuras de heterogeneidad muestral. Cuando existen grupos de valores atípicos estos procedimientos no son fiables ya que: (1) pueden no identificar conjuntos de atípicos (este fenómeno se conoce como enmascaramiento, *masking* en inglés); y (2) pueden señalar como atípicos a datos que no lo son (este fenómeno se conoce en inglés como *swamping*).

En las siguientes secciones se presentan las propuestas más relevantes que han

aparecido en la literatura bayesiana para la identificación de valores atípicos en modelos de regresión y series temporales, así como una descripción del algoritmo de Gibbs Sampling y sus principales propiedades. Una revisión más completa sobre los métodos clásicos y bayesianos se puede encontrar en los libros de Belsley *et al.* (1980), Cook y Weisberg (1982), Barnett y Lewis (1984), Chaterjee y Hadi (1988) y Spall (1988); o en los artículos de Beckman y Cook (1983), Chaterjee y Hadi (1986), Peña (1987a), Hotta y Neves (1992) y Justel, Peña y Sánchez (1993).

1.2 Datos atípicos en modelos de regresión

1.2.1 Detección y tratamiento de datos atípicos e influyentes aislados

Se considera el modelo de regresión lineal

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i \qquad i = 1, \ldots, n, \tag{1.1}$$

donde $\boldsymbol{y} = (y_1, \ldots, y_n)'$ es un vector de variables endógenas, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ una matriz $n \times p$ de variables exógenas y rango $p < n$, $\boldsymbol{\beta}$ es un vector de p parámetros desconocidos y $\boldsymbol{u} = (u_1, \ldots, u_n)'$ un vector de perturbaciones no observadas.

Los distintos métodos para el tratamiento y la identificación de datos atípicos e influyentes en modelos de regresión dentro del enfoque bayesiano se pueden clasificar en dos categorías:

i) Métodos de diagnóstico en los que no se especifica la distribución de los datos atípicos.

ii) Métodos robustos en los que se especifica la distribución de los datos atípicos.

Los métodos de diagnóstico bayesianos no suponen un modelo particular para explicar el mecanismo que genera los datos atípicos. Las observaciones atípicas se iden-

tifican analizando las densidades predictivas $p(y_i \mid \boldsymbol{y}_{(i)})$, donde $\boldsymbol{y}_{(i)}$ significa que se ha eliminado el dato i-ésimo de la muestra $\boldsymbol{y}$. Si la observación i-ésima es atípica y las restantes son buenas, la probabilidad de predecir y_i observando toda la muestra menos el dato i-ésimo será muy baja. Este procedimiento se conoce con el nombre de *método de la ordenada de la distribución predictiva condicionada* y fue introducido por Geisser (1980). Como

$$p(y_i \mid \boldsymbol{y}_{(i)}) = \frac{p(\boldsymbol{y})}{p(\boldsymbol{y}_{(i)})},$$

la densidad predictiva condicionada puede verse como el cociente entre dos formulaciones de la distribución predictiva $p(\boldsymbol{y})$, recomendada por Box (1980) como herramienta general de diagnóstico de un modelo estadístico. Esta medida ha sido también utilizada entre otros por Pettit y Smith (1985) y Pettit (1990).

La ordenada predictiva está muy relacionada con el contraste clásico del residuo estudentizado. Puede demostrarse (Pettit, 1990) que para modelos de regresión con distribuciones a priori no informativas

$$p(y_i \mid \boldsymbol{y}_{(i)}) = c\, s_{(i)}^{-1}(1-h_i)^{1/2}\left(1+\frac{t_i^2}{n-p-1}\right)^{-\frac{n-p}{2}} \tag{1.2}$$

donde t_i es el residuo estudentizado (véase, por ejemplo, Cook y Weisberg, 1982 o Peña, 1993)

$$t_i = \frac{y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}}{s_{(i)}(1-h_i)^{1/2}}, \tag{1.3}$$

$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ es el estimador de mínimos cuadrados, $s_{(i)}^2 = \sum(y_j - \boldsymbol{x}_j'\hat{\boldsymbol{\beta}}_{(i)})/(n-p-1)$ es el estimador insesgado de la varianza residual cuando se elimina el dato i-ésimo, y h_i es el elemento i-ésimo de la diagonal principal de la matriz $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, que se conoce como el potencial (*leverage*, en inglés) de la observación i-ésima

$$h_i = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i \qquad i = 1, \ldots, n. \tag{1.4}$$

En consecuencia, observaciones con residuo estudentizado alto tendrán un valor pequeño de la ordenada predictiva (1.2) y serán consideradas como atípicas.

El problema que presenta el método de la ordenada predictiva es que observaciones con alto potencial (h_i está acotado por 1) tendrán un valor grande de la ordenada predictiva (1.2) independientemente de que sean atípicas. Esto se deduce de expresar el residuo estudentizado como

$$t_i = \frac{(1-h_i)^{1/2} e_{i(i)}}{s_{(i)}},$$

donde $e_{i(i)} = y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_{(i)}$ es el residuo mínimo cuadrático que se obtiene al estimar la recta de regresión eliminando el dato i-ésimo. Cuando h_i tiende a 1, el residuo estudentizado tiende a cero, independientemente de que la observación i-ésima sea atípica ($e_{i(i)}$ es grande) o sea una observación buena ($e_{i(i)}$ es pequeña). En este caso se dice que la observación i-ésima está muy alejada del resto en el espacio de las variables independientes y es un *dato influyente*. Por tanto, se distinguen tres tipos entre datos atípicos e influyentes:

1. Los *datos atípicos no influyentes* tienen potencial pequeño y su residuo estudentizado es grande: se identifican sin problemas.
2. Los *datos atípicos influyentes* tienen potencial grande y su residuo estudentizado es pequeño: no se identifican con el método de la ordenada predictiva.
3. Los *datos buenos potencialmente influyentes* tienen potencial grande y su residuo estudentizado es pequeño.

La identificación de observaciones atípicas influyentes se obtiene examinando los cambios que se producen en la distribución a posteriori de los parámetros o en la distribución predictiva cuando se elimina una observación. Johnson y Geisser (1983, 1985) y Geisser (1985) proponen utilizar la divergencia de Kullback-Leibler (Kullback y Leibler, 1951)

$$J(f_1, f_2) = \int f_1 \ln \frac{f_1}{f_2} dx + \int f_2 \ln \frac{f_2}{f_1} dx \tag{1.5}$$

para medir la distancia entre las distribuciones predictivas cuando la observación i-ésima es eliminada, $p(\boldsymbol{y}_{(i)})$, y cuando se considera toda la muestra, $p(\boldsymbol{y})$. Por otro lado, Pettit y Smith (1985) y Guttman y Peña (1988a, 1993) sugieren el uso de la "distancia" (1.5) para medir los cambios en la distribución a posteriori de los parámetros.

Al igual que sucede con el procedimiento para identificar observaciones atípicas, existe una estrecha relación entre las medidas de influencia basadas en la divergencia de Kullback-Leibler y los estadísticos clásicos para detectar influencia. Johnson y Geisser (1983) prueban que su medida es asintóticamente equivalente a la suma del estadístico de Cook (Cook, 1977) y una función convexa de los residuos estudentizados. El estadístico de Cook mide la influencia en la estimación de los parámetros y se puede expresar como

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})' \boldsymbol{X}' \boldsymbol{X} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{p s^2},$$

donde s^2 es el estimador insesgado de la varianza residual y $\hat{\boldsymbol{\beta}}_{(i)}$ es el estimador de mínimos cuadrados eliminando la observación i-ésima. Este estadístico se relaciona con el residuo estudentizado (1.3) mediante

$$D_i = \frac{(n-p)}{p} \frac{t_i^2}{(n-p-1+t_i^2)} \frac{h_i}{1-h_i}.$$

El estadístico de Cook es grande para las observaciones atípicas influyentes y pequeño para las observaciones buenas.

Guttman y Peña (1988a) demuestran que el cambio en la distribución a posteriori del parámetro $\boldsymbol{\beta}$ es también una función del estadístico de Cook y se puede escribir como

$$J(p(\boldsymbol{\beta} \mid \boldsymbol{y}_{(i)}), p(\boldsymbol{\beta} \mid \boldsymbol{y})) = \frac{p D_{(i)}^2}{2} + \frac{s_{(i)}^2}{2s^2} \left(p + \frac{h_{ii}}{1-h_{ii}} \right) + \frac{s^2}{2s_{(i)}^2} (p - h_{ii}) - p,$$

donde $D_{(i)}$ es

$$pD_{(i)} = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})' \boldsymbol{X}'_{(i)} \boldsymbol{X}_{(i)} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}) / s^2_{(i)}.$$

También prueban que el cambio en la distribución a posteriori del parámetro σ^2 se puede interpretar como una medida de atipicidad dependiente de los residuos estudentizados t_i y estandarizados r_i, según la relación

$$J(p(\sigma^2 \mid \boldsymbol{y}_{(i)}), p(\sigma^2 \mid \boldsymbol{y})) = \frac{t_i^2 - r_i^2}{2},$$

donde

$$r_i = \frac{e_i}{s(1-h_i)^{1/2}}.$$

Por último, el cambio en la distribución a posteriori conjunta de los dos parámetros es una combinación de las medidas de influencia del dato i-ésimo en la distribución de $\boldsymbol{\beta}$ y de su atipicidad.

Otras medidas de influencia son descritas por Kass, Tierney y Kadane (1985), que emplean métodos asintóticos para estudiar el cambio en determinadas funciones de interés cuando se elimina una observación, y por Kempthorne (1986) y Carlin y Polson (1991a) que analizan cambios en el riesgo de Bayes en un contexto de teoría de la decisión.

Para concluir con los métodos de diagnóstico bayesianos conviene destacar que todas las propuestas que se han mencionado para identificar datos atípicos e influyentes aislados se pueden extender al problema de identificación de un grupo de tamaño m. Sin embargo, los cálculos necesarios en este caso aumentan hasta cubrir todas las $\binom{n}{m}$ combinaciones posibles dentro de la muestra.

Los métodos robustos bayesianos suponen un modelo que acomoda los datos atípicos de manera que la estimación de los parámetros se lleva a cabo con todas las observaciones, incluidas las anómalas. Box y Tiao (1973) proponen la familia exponencial de

potencias para la distribución de los errores, que viene dada por

$$f(u) = k_1(\alpha)\sigma^{-1}e^{-\frac{1}{2}\left|\frac{u}{\sigma}\right|^{\frac{2}{1+\alpha}}} \qquad -1 < \alpha \leq 1.$$

Cuando $\alpha = 0$, esta distribución coincide con la normal, mientras que cuando $0 < \alpha \leq 1$ se obtienen distribuciones de colas más pesadas. West (1984) sugiere utilizar modelos con distribuciones de colas pesadas para los errores que pueden ser descompuestas en mixturas de normales con distinta varianza, entre las que se incluyen familias de distribuciones como la t de Student, las estables, la logística y la doble exponencial. Sobre estas distribuciones se puede llevar a cabo un análisis de las distribuciones a posteriori de los parámetros explotando ciertas propiedades de los errores, objetivo que no siempre es alcanzable cuando se supone una distribución cualquiera de colas pesadas.

Entre los métodos robustos destacan también los que suponen una distribución alternativa para los datos atípicos. El caso más estudiado es el de los modelos introducidos por Tukey (1960) de mixturas de distribuciones para los errores; mezcla de una distribución central y una distribución contaminante. Por un lado, se sugiere un modelo de distribución de los errores que acomoda las observaciones anómalas, generadas según una distribución alternativa; por otro lado, cuando la muestra no contiene datos atípicos la estimación de los parámetros es la adecuada para el modelo central que trata de reproducir el mecanismo generador de datos. La identificación se realiza a través del análisis de las distribuciones a posteriori de que cada observación sea generada por el modelo alternativo o comparando distribuciones predictivas.

En el modelo de Box y Tiao (1968), *modelo normal de contaminación de escala* y que denotaremos por BT, se establece la regresión estándar de la ecuación (1.1), pero suponiendo que los errores u_i se distribuyen según una mixtura de normales

$$u_i \sim (1-\alpha)\, N(0,\sigma^2) + \alpha\, N(0,k^2\sigma^2) \qquad i = 1,\ldots,n. \tag{1.6}$$

Alternativamente, Abraham y Box (1978) proponen el *modelo normal de cambio de nivel* (AB) en el que la heterogeneidad se refleja en la media. En este caso, los errores se distribuyen

$$u_i \sim (1-\alpha)\, N(0,\sigma^2) + \alpha\, N(\lambda,\sigma^2) \qquad i = 1,\ldots,n.$$

El modelo aditivo de Guttman, Dutter y Freeman (1978) (GDF) supone la presencia de m datos atípicos (m se fija analizando el modelo para $m = 0,1,\ldots$) y, en consecuencia, la distribución de los errores u_{i_j} es $\mathrm{N}(\lambda_j,\sigma^2)$ para $j = 1,\ldots,m$ y $\mathrm{N}(0,\sigma^2)$ para los restantes. Finalmente, Eddy (1980) propone un modelo mixto que combina las formulaciones de Box y Tiao (1968) y Abraham y Box (1978).

Mientras que con un enfoque clásico la estimación mínimo cuadrática de estos modelos requiere el conocimiento previo de qué observaciones son atípicas, la ventaja que ofrece el planteamiento del problema desde un punto de vista bayesiano es que la distribución a posteriori del vector de parámetros $\boldsymbol{\beta}$ se puede expresar como una media ponderada de las 2^n distribuciones de $\boldsymbol{\beta}$ condicionadas a cada configuración de datos buenos y atípicos que se pueda dar en la muestra. Si $A(r)$ representa el suceso "r observaciones son atípicas y $n-r$ proceden del modelo central" la distribución a posteriori de $\boldsymbol{\beta}$ es

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) = \sum_r p(A(r) \mid \boldsymbol{y})\; p(\boldsymbol{\beta} \mid A(r), \boldsymbol{y}),$$

donde las ponderaciones $p(A(r) \mid \boldsymbol{y})$ son la probabilidad a posteriori de las distintas configuraciones posibles. La expresión de estas probabilidades cuando se supone k y α (en BT) y α y λ (en AB) conocidos, y la distribución a priori para $\boldsymbol{\beta}$ y σ^2 de referencia habitual $p(\boldsymbol{\beta},\sigma^2) \propto \sigma^{-2}$, se presentan de forma compacta en Freeman (1980). En la discusión que sigue a este artículo, Eddy (1980) pone de manifiesto que la estimación de los parámetros de la distribución $p(\boldsymbol{\beta} \mid A(r), \boldsymbol{y})$ en los tres modelos puede verse como una estimación de mínimos cuadrados ponderados, donde los pesos varían

según el modelo que se especifique. Si el modelo especificado es el BT, la matriz de ponderaciones es

$$\boldsymbol{V}_{(r)\perp} = \begin{pmatrix} \boldsymbol{I}_{n-r} & \mathbf{0} \\ \mathbf{0} & k^{-2}\boldsymbol{I}_r \end{pmatrix},$$

si el modelo es el AB

$$\boldsymbol{V}_{(r)\perp} = \begin{pmatrix} \boldsymbol{I}_{n-r} & \mathbf{0} \\ \mathbf{0} & r^{-1}\mathbf{1}\mathbf{1}' \end{pmatrix},$$

y si es el GDF

$$\boldsymbol{V}_{(r)\perp} = \begin{pmatrix} \boldsymbol{I}_{n-r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Como se dijo anteriormente, uno de los posibles mecanismos de identificación de atípicos consiste en utilizar los pesos $p(A(r) \mid \boldsymbol{y})$. En el caso particular de observaciones atípicas aisladas, donde se supone que hay un único dato anómalo y los restantes son buenos, las probabilidades vienen dadas por la expresión

$$p(A_i(1) \mid \boldsymbol{y}) \propto \omega |\boldsymbol{X}'\boldsymbol{V}_{(1)}\boldsymbol{X}|^{-1/2} s_{(i)}^{-\upsilon},$$

donde, en general, el subíndice en $A_i(r)$ indica que entre las r observaciones que contaminan la muestra se encuentra y_i. Según el modelo, los valores de ω y υ son:

$$\begin{array}{lllll} \text{BT:} & \upsilon = n-p & \text{y} & \omega = \alpha/k(1-\alpha) \\ \text{AB:} & \upsilon = n-p-1 & \text{y} & \omega = \alpha/(1-\alpha) \\ \text{GDF:} & \upsilon = n-p-1 & \text{y} & \omega = 1. \end{array}$$

La extensión al caso general, en el que se plantea si una observación es atípica independientemente de que lo sean las demás, es inmediata teóricamente, ya que $p_i = \sum_r p(A_i(r) \mid \boldsymbol{y})$, pero supone calcular la probabilidad de todas las 2^n combinaciones posibles.

La segunda vía para identificar las observaciones atípicas consiste en comparar las distribuciones predictivas. Siguiendo a Jeffreys (1961), podemos comparar mediante el factor de Bayes el modelo que supone que únicamente la observación i-ésima es atípica con el modelo de regresión estándar. Para estos modelos el factor de Bayes se puede expresar como

$$F_{10}(i) = \frac{p(\boldsymbol{y} \mid A_i(1))}{p(\boldsymbol{y} \mid A(0))}.$$

Pettit (1992) extiende la idea del factor de Bayes para distribuciones a priori impropias, recogiendo el método de Spiegelhalter y Smith (1982) de observaciones imaginarias de tamaño mínimo.

Peña y Guttman (1993) estudian la relación entre los métodos de identificación de atípicos basados en mixturas de distribuciones y el método de la ordenada de la distribución predictiva condicionada para modelos que no suponen distribución alternativa. Prueban que la probabilidad a posteriori de cualquier combinación de datos atípicos y buenos para los modelos BT y GDF es inversamente proporcional a la ordenada de la distribución predictiva condicionada.

Como conclusión a esta discusión, podemos resumir las ideas que aportan las dos vías de aproximación al problema de la identificación y tratamiento de observaciones atípicas e influyentes en modelos de regresión, en los siguientes puntos:

(1) Los métodos robustos se ocupan del tratamiento de las observaciones anómalas proponiendo modelos que explican la generación de toda la muestra, tanto de los datos buenos como de los atípicos.

(2) Las medidas de atipicidad e influencia en modelos bayesianos que no suponen distribución alternativa para las observaciones atípicas y los métodos de diagnóstico clásicos pueden ser vistos como equivalentes: ambos miran al residuo estudentizado y miden la influencia por los cambios que se producen en la estimación de

los parámetros (o de su distribución) al eliminar las observaciones sospechosas de no haber sido generadas por el modelo especificado.

(3) Las propuestas para identificar datos atípicos con modelos sin distribución alternativa o especificando ésta pueden expresarse de manera única. En ambos casos la identificación puede verse como el análisis de las distribuciones predictivas eliminando el dato.

1.2.2 Detección y tratamiento de grupos de atípicos

Cuando se utilizan los procedimientos de identificación de observaciones atípicas individuales en muestras que contienen grupos de atípicos es posible que se produzca el fenómeno conocido como *enmascaramiento*: una observación atípica no se identifica por la presencia de otra atípica. Otro fenómeno que tiende a producirse con grupos de atípicos es el *señalamiento* de observaciones buenas como atípicas.

El enmascaramiento se produce generalmente cuando existen varias observaciones que son atípicas y todas ellas alejadas del resto en el espacio de las variables independientes. Los métodos de identificación de atípicos basados en eliminar una observación no funcionan en este caso ya que los residuos (calculados con y sin el dato) tienden a ser pequeños mientras no se eliminan los otros atípicos. Además cuando el tamaño del grupo es grande el potencial de estos datos tiende a ser pequeño aunque estén muy alejados. En este caso, tanto el residuo estudentizado como el estadístico de Cook serán pequeños. Peña y Yohai (1995) prueban este hecho en el caso límite en el que existe un grupo I de n_{I} observaciones atípicas $(y_a, \boldsymbol{x}'_a)$ todas ellas a una "distancia" $h_a = \boldsymbol{x}'_a(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}'_a$. Los residuos de la regresión estimada con toda la muestra son

$$e_a = \frac{y_a - \boldsymbol{x}'_a\hat{\boldsymbol{\beta}}_{(\mathrm{I})}}{1 + n_{\mathrm{I}}\, h_a},$$

donde $\hat{\boldsymbol{\beta}}_{(\mathrm{I})}$ es el estimador de mínimos cuadrados cuando la muestra no contiene el

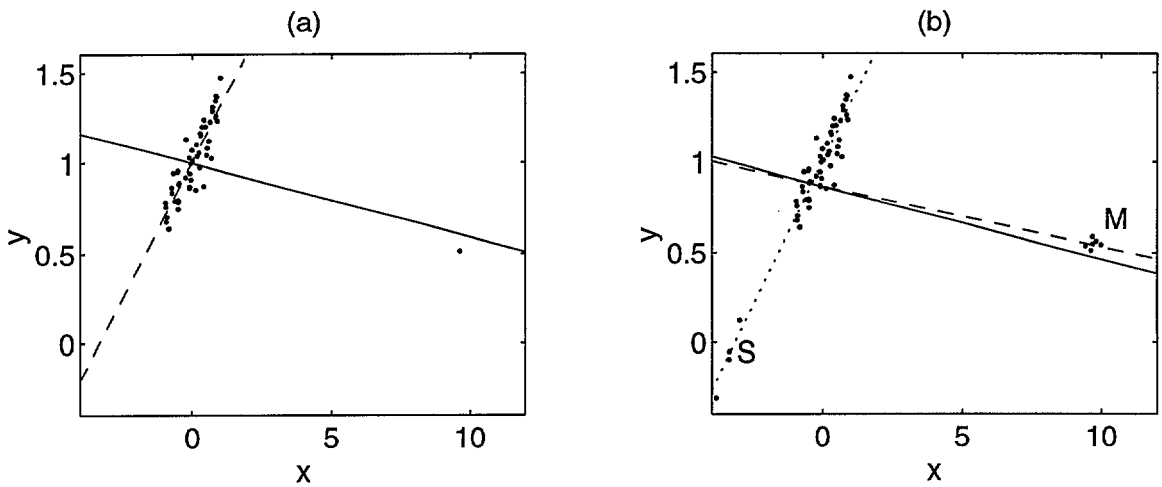

Figura 1.1: Ejemplos de datos atípicos: (a) aislados y (b) enmascarados (M) y señalados (S). La línea continua es el estimador de la regresión con todos los datos, la discontinua eliminando un dato atípico y la de puntos eliminando todos los atípicos.

grupo I. Si h_a es grande el residuo será pequeño y no cambiará mucho si se elimina la observación y se sustituye n_{I} por $n_{\mathrm{I}} - 1$. Los términos diagonales de la matriz $\boldsymbol{H}$ para los datos de I son $h_a/(1 + n_{\mathrm{I}} h_a)$, que tienden a ser pequeños cuando aumenta n_{I}.

Para entender mejor el fenómeno de enmascaramiento, en la figura 1.1(a) se muestra un conjunto de datos en el que existe un dato atípico aislado, mientras que en la figura 1.1(b) existe un grupo (M) de atípicos que se enmascaran y un grupo (S) de datos buenos que son señalados por los atípicos de M. Se puede observar que al eliminar el dato atípico aislado en la figura 1.1(a) la estimación de la recta (la línea discontinua) cambia drásticamente, de manera que cualquier medida de influencia permitirá identificar este dato. En el caso del grupo de atípicos de la figura 1.1(b) el cambio es mínimo y cualquier método de los descritos para detectar atípicos aislados identificará como atípicas las observaciones del grupo S, que no lo son. Únicamente se identificarán los atípicos cuando se elimine el grupo completo (línea de puntos)

Para superar este problema se han desarrollado en la literatura diversos procedimientos que desde un punto de vista clásico tratan la identificación de grupos de atípicos mediante el análisis recursivo de los residuos, la estimación robusta o el análisis multivariante. Aunque existen muchos otros métodos, destacamos los de Hadi y Simonoff (1993) y Peña y Yohai (1995) (estos serán comparados con el nuevo procedimiento que se propone en el capítulo 3). Hadi y Simonoff (1993) proponen un método con dos variantes en el que se analizan los residuos estudentizados iterativamente. El primer objetivo es separar los datos en un conjunto que no contenga datos atípicos y otro en el que estén todos los posibles atípicos. En la segunda fase se contrasta si los posibles atípicos están suficientemente alejados del resto de las observaciones mediante una versión adecuadamente escalada del error de predicción. Peña y Yohai (1995) proponen un método basado en un enfoque multivariante. Su idea es construir una matriz de influencia que describa el cambio en la predicción de cada dato cuando se elimina cada una de las observaciones. Como los grupos de observaciones atípicas deben aparecer con el mismo signo y estructura dentro de los vectores propios de la matriz de influencia, desarrollan un procedimiento para detectar los valores con mayor peso en los vectores propios y contrastar si son atípicos.

Los métodos bayesianos para la identificación de observaciones atípicas e influyentes requieren un número de cálculos no muy elevado cuando se trata de la identificación de un dato atípico aislado, pero la generalización al caso de grupos de datos atípicos conduce inmediatamente a un problema computacional difícil de resolver. Desde un punto de vista bayesiano son pocos los procedimientos que se han desarrollado para el caso de grupos de atípicos. Peña y Tiao (1992) proponen dos herramientas nuevas de diagnóstico para identificar grupos de datos atípicos con el modelo de BT: la *Curva de Robustez Bayesiana* (BROC) y la *Curva de Robustez Bayesiana Secuencial* (SEBROC). Estas curvas comparan el modelo que contiene h datos atípicos (M_h) con el modelo que

no contiene observaciones atípicas (M_0). La Curva de Robustez Bayesiana se define como el cociente de las probabilidades a posteriori de los modelos M_h y M_0, frente a la cantidad de datos atípicos h, y se puede expresar en términos del factor de Bayes $F_{h,0}$ como:

$$P_{h0} = \frac{P(M_h \mid \boldsymbol{y})}{P(M_0 \mid \boldsymbol{y})} = \binom{n}{h}\left(\frac{\alpha}{1-\alpha}\right)^h F_{h,0},$$

donde $\boldsymbol{X}_{(h)}$ denota la matriz que contiene todas las columnas de $\boldsymbol{X}$ correspondientes a las h observaciones atípicas. El factor de Bayes se puede escribir como

$$F_{h0} = \binom{n}{h}^{-1} k^{-h} \sum_r \frac{|\boldsymbol{X}'\boldsymbol{X}|^{\frac{1}{2}}}{|\boldsymbol{X}'\boldsymbol{X} - \phi \boldsymbol{X}'_{(r)}\boldsymbol{X}_{(r)}|^{\frac{1}{2}}} \left(\frac{s^2}{s^2_{(r)}}\right)^{\frac{n-p}{2}}. \tag{1.7}$$

El sumatorio en (1.7) se extiende a todas las $\binom{n}{h}$ combinaciones posibles de que la muestra contenga exactamente h observaciones atípicas y $s^2_{(r)}$ es el término equivalente a la suma de cuadrados residual teniendo en cuenta el modelo que genera cada observación y cuya expresión exacta se puede encontrar en Box y Tiao (1968). Así definida, la BROC da información del número de datos atípicos en la muestra. Sin embargo, cuando existen varios datos atípicos esta curva puede no identificar alguno de ellos si se produce un efecto de enmascaramiento. Para evitar este problema se propone la Curva de Robustez Bayesiana Secuencial (SEBROC), que se construye para cada h comparando secuencialmente los cocientes

$$S_{h,h-1} = \frac{P_{h,0}}{P_{h-1,0}}.$$

Una parte importante de la propuesta de Peña y Tiao (1992) es el procedimiento de muestreo estratificado para reducir los $\binom{n}{h}$ cálculos necesarios para obtener $P_{h,0}$ o los $\binom{n}{h-1} + \binom{n}{h}$ para $S_{h,h-1}$, sin que por ello disminuya la eficiencia del método. El procedimiento consiste en: (1) dividir la muestra en dos partes, una que contiene los potencialmente atípicos y otra que contiene las observaciones presumiblemente correctas. Esta división se efectúa a partir de la matriz que proponen Peña y Tiao (1992),

cuyos términos comparan la probabilidad de que haya dos observaciones atípicas en la muestra con el producto de las probabilidades de que cada una de ellas sea la única atípica. Esta matriz mide la interacción entre pares de observaciones que no se manifiesta en las probabilidades para datos atípicos aislados. Si la observación i es atípica, todos los elementos

$$d(i,j) = \frac{p(A_{i,j}(2) \mid \boldsymbol{y})}{p(A(0) \mid \boldsymbol{y})} - \frac{p(A_i(1) \mid \boldsymbol{y})}{p(A(0) \mid \boldsymbol{y})} \frac{p(A_j(1) \mid \boldsymbol{y})}{p(A(0) \mid \boldsymbol{y})}$$

de la columna correspondiente tomarán valores relativamente altos. Sea n_1 el tamaño del conjunto que contiene los datos atípicos y $n - n_1$ el del complementario. (2) Utilizando que

$$\binom{n}{h} = \sum_{r=0}^{h} \binom{n_1}{r} \binom{n-n_1}{h-r}$$

podemos calcular las $\binom{n}{h}$ combinaciones resultantes de eliminar h elementos de los n a partir de los grupos n_1 y $n - n_1$ calculando todas las combinaciones en el grupo n_1, pero sólo una muestra pequeña del conjunto $n - n_1$. Por ejemplo, si $n_1 = 10$, $n_2 = 20$ y $h = 3$, calcularemos las $\binom{10}{3}$ combinaciones de eliminar tres elementos de n_1, las $\binom{10}{2}$ combinaciones de eliminar dos cruzadas con la eliminación al azar de uno cualquiera de los 20 elementos de n_2, las $\binom{10}{1}$ combinaciones de eliminar uno de n_1 combinada con una muestra aleatoria de las $\binom{20}{2}$ combinaciones posibles de observaciones buenas y, finalmente, una muestra aleatoria pequeña de las $\binom{20}{3}$ posibilidades de eliminar observaciones buenas.

1.3 Datos atípicos en modelos de series temporales

El análisis bayesiano de datos atípicos en series temporales se inicia con el trabajo de Abraham y Box (1979) para procesos autorregresivos. Debido a la complejidad analítica que supone obtener las distribuciones a posteriori su desarrollo se ha limitado

al tratamiento de valores atípicos aislados; siendo el problema de enmascaramiento un campo de investigación en el que la utilización de algoritmos como el Gibbs Sampling pueden facilitar la identificación de heterogeneidad en una serie temporal.

Abraham y Box (1979) proponen un modelo de generación de atípicos alternativo al modelo central que sirve para generar el resto de la serie. Las distribuciones a posteriori de los parámetros de un AR(p) con valores atípicos innovativos son obtenidas para los casos en que: (1) se conoce la posición de las innovaciones que contaminan la serie y (2) únicamente se sabe que con cierta probabilidad α una innovación es atípica. Este segundo caso es el más complejo y se convierte en el problema de calcular 2^n probabilidades. Cuando los valores atípicos en el proceso autorregresivo son aditivos, la obtención de las distribuciones marginales a posteriori de los parámetros y de la magnitud del dato atípico son difíciles de tratar analíticamente. Haciendo uso del algoritmo de Gibbs Sampling, McCulloch y Tsay (1994a) estiman la distribución a posteriori de estos parámetros y Marriott *et al.* (1992) consideran cada dato atípico (cuando la posición es conocida) como un parámetro más en la ejecución del Gibbs Sampling. El mismo tratamiento lo proponen Marriott *et al.* (1994) para procesos ARMA, utilizando un algoritmo descrito por Müeller (1991) que combina el Gibbs Sampling y el algoritmo de Metropolis.

Las ventajas computacionales que ofrece el Gibbs Sampling también han sido utilizadas para analizar modelos de cambio de nivel o varianza. Los modelos que estudian McCulloch y Tsay (1993, 1994a) suponen estos cambios en procesos autorregresivos, cambios que pueden ocurrir en cualquier instante t con cierta probabilidad α. Las distribuciones condicionadas necesarias para la ejecución del Gibbs Sampling no son difíciles de obtener e incluyendo un conjunto de variables ficticias δ_t (con valores 1 ó 0 dependiendo de si hay o no cambio en la serie en el instante t) se obtiene la probabilidad de que haya un cambio en t. Para predecir los cambios McCulloch y Tsay (1993)

proponen relacionarlos con otras variables explicativas mediante la formulación de un modelo probit. Así, la probabilidad $p_t = P(\delta_t = 1)$ depende de un conjunto de variables $\boldsymbol{x}_t$ mediante la ecuación

$$p_t = \Phi(\boldsymbol{x}'\gamma),$$

donde Φ es la función de distribución de una normal de media cero y varianza uno y γ es un vector de parámetros desconocidos. Cuando los cambios dependen de un proceso de Markov con probabilidades de transición desconocidas, Albert y Chib (1993) y McCulloch y Tsay (1994b) estiman las distribuciones a posteriori y las probabilidades de transición; estos últimos con aplicación al análisis de series macroeconómicas.

Otra forma de analizar observaciones atípicas en series temporales es considerar la serie como un caso particular del modelo lineal dinámico (Harrison y Stevens, 1976). En síntesis, estos modelos se formulan como

$$y_t = \boldsymbol{F}_t\boldsymbol{\theta}_t + v_t \tag{1.8}$$

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{a}_t, \tag{1.9}$$

donde $\boldsymbol{\theta}_t$ es un vector de parámetros desconocidos, las matrices $\boldsymbol{F}_t$ y $\boldsymbol{G}_t$ se suponen conocidas para todo t y v_t y $\boldsymbol{a}_t$ son procesos no observables de ruido blanco. La estimación de estos modelos se realiza mediante el filtro de Kalman (Kalman, 1960) y el tratamiento de valores atípicos, al igual que en regresión, se ha realizado desde tres puntos de vista. Dentro de los métodos que proponen un modelo robusto frente a las observaciones atípicas, Zellner (1976) y West (1981) estudian como robustificar el procedimiento de estimación suponiendo que las distribuciones son t de Student. La segunda vía trata de identificar valores atípicos sin un modelo particular de generación de atípicos. En este contexto, West y Harrison (1986) propusieron inicialmente estudiar secuencialmente el cociente

$$W_t(k) = \frac{p(y_t, \ldots, y_{t-k+1} \mid D_{t-k})}{p_a(y_t, \ldots, y_{t-k+1} \mid D_{t-k})},$$

donde D_t es la información hasta el instante t, el numerador es la distribución predictiva de las últimas k observaciones dada la información hasta el instante $t-k$ y el denominador es un valor de referencia. Este cociente puede escribirse como

$$W_t(k) = W_{t-1}(k-1)\, H_t,$$

donde H_t es el factor de Bayes

$$H_t = \frac{p(y_t \mid D_{t-1})}{p_a(y_t \mid D_{t-1})}.$$

El principal problema que presenta este método es decidir la distribución p_a que servirá como valor de referencia para juzgar cuándo un dato, o un grupo de datos, es atípico. Posteriormente, Harrison y West (1991) optan por seguir un enfoque similar a Peña (1987b) y construir diagnósticos individuales suponiendo que cada observación es un valor ausente; recogiendo también el enfoque de Johnson y Geisser (1983) de medir la influencia de cada observación por el cambio en una distribución (la predictiva o la a posteriori) medido por la divergencia de Kullback-Leibler.

Por último, los modelos de mixturas para el tratamiento de los valores atípicos suponen un modelo explícito de generación de datos atípicos que, de esta manera, son automáticamente incorporados con menos peso en el proceso de estimación de los parámetros, obteniéndose simultáneamente un método robusto de estimación y un método de detección de valores atípicos. Harrison y Stevens (1976) sugieren suponer que cada observación del modelo (1.8) y (1.9) puede estar generada con cierta probabilidad p_i (que suponemos conocida) por un modelo M_i entre varios posibles. Por tanto, si la distribución a priori y la verosimilitud son mezcla de h componentes, la distribución a posteriori

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = c\, p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

es una mezcla de h^2 distribuciones y cuando se estima el filtro de Kalman la dimensión del problema crece en cada etapa. Con $h = 2$, el número de componentes de la dis-

tribución a posteriori final es 2^n. Estos autores sugieren que se emplee un procedimiento para colapsar por momentos en cada etapa la distribución resultante. En esta línea, Guttman y Peña (1985, 1988b) proponen un algoritmo simple de estimación robusta para el modelo donde el ruido v_t en la ecuación de observación (1.8) proviene de una normal con contaminación de escala como en (1.6). El algoritmo está basado en la idea de colapsar en cada etapa la distribución a posteriori a una única distribución normal, ajustada por momentos. Colapsar por momentos es óptimo en el sentido de minimizar la distancia de Kullback-Leibler entre la verdadera distribución *a posteriori* y la colapsada. Meinhold y Singpurwalla (1989) trabajan con distribuciones t de Student y proponen representar la distribución *a priori* en la etapa t como una mezcla de h_t distribuciones t. Cuando se pasa a la siguiente etapa, el teorema de Bayes se aplica a cada uno de los h_t componentes de la mezcla resultando una distribución que puede ser unimodal o bimodal. Si es unimodal, la distribución resultante se aproxima por momentos por una distribución t. Si es bimodal, se utiliza una mezcla de dos distribuciones t. Además, en cada etapa se estudia la distribución *a posteriori* para ver si algún componente de la mezcla tiene un peso menor que un determinado umbral, en cuyo caso es eliminado. La distribución *a posteriori* resultante en cada etapa se convierte en distribución *a priori* para la siguiente etapa y se repite el mismo procedimiento.

1.4 Gibbs Sampling

1.4.1 Algoritmos MCMC

Desde que Tanner y Wong (1987) y Gelfand y Smith (1990) proponen los algoritmos markovianos conocidos como MCMC (del inglés, *Monte Carlo Markov Chain*) para estimar las distribuciones a posteriori de los parámetros de un modelo bayesiano, se

ha desarrollado una extensa literatura sobre estos métodos de simulación. El origen de los MCMC se remonta al algoritmo de Metropolis *et al.* (1953) para simular sistemas complejos con aplicaciones en física del estado sólido. Otro antecedente se encuentra en Hastings (1970) para generar variables aleatorias. Posteriormente, Geman y Geman (1984) introducen el algoritmo de Gibbs Sampling para reconstruir imágenes y simular campos aleatorios markovianos, un caso particular de una distribución de Gibbs, siendo éste el origen del nombre con el que es conocido.

La aplicación del Gibbs Sampling a la resolución de problemas bayesianos se inicia a partir de Gelfand y Smith (1990). El algoritmo permite estimar densidades marginales, no obtenibles analíticamente, a partir de muestras aleatorias de distribuciones conocidas. El requerimiento básico consiste en ser capaces de generar muestras de las distribuciones condicionadas a posteriori de todos los parámetros del modelo. A partir de unos valores iniciales de los parámetros se genera iterativamente una secuencia de muestras de las distribuciones condicionadas que converge en distribución a la conjunta de los parámetros, sea cual sea la selección de los valores iniciales. El Gibbs Sampling es un método menos eficiente para realizar inferencia que la simulación *directa* consistente en extraer muestras de la distribución verdadera; no obstante, el número de problemas en los que conocemos la distribución exacta es bastante reducido en comparación con los problemas en que se puede aplicar Gibbs Sampling. Además, en trabajos aplicados el Gibbs Sampling presenta dos ventajas importantes frente a otros procedimientos: (1) los requerimientos son escasos y (2) es fácil de programar.

En general, la literatura existente sobre algoritmos MCMC se puede dividir en dos grandes bloques. Por un lado, se estudian las propiedades teóricas del algoritmo, prestando una atención especial a la velocidad de convergencia y al estudio de los criterios de parada (véase, por ejemplo, Smith y Robert, 1993, y Tierney, 1994). Por otro lado, se emplea el Gibbs Sampling en la resolución de modelos con solución analítica

compleja en campos tan diversos como la economía, la genética o la paleontología.

1.4.2 Descripción del algoritmo

Sean $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ variables aleatorias vectoriales con función de densidad conjunta estrictamente positiva en el espacio muestral. Supongamos que se pueden obtener todas las distribuciones de $\boldsymbol{Z}_i$ condicionadas a $\boldsymbol{Z}_{(i)} = \boldsymbol{z}_{(i)}$, donde $\boldsymbol{Z}_{(i)}$ denota que se ha eliminado $\boldsymbol{Z}_i$, $\boldsymbol{Z}_{(i)} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{i-1}, \boldsymbol{Z}_{i+1}, \ldots, \boldsymbol{Z}_n)$, bajo ciertas condiciones débiles estas distribuciones condicionadas determinan la distribución conjunta de $(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ y, por tanto, todas las distribuciones marginales (véase Besag, 1974). A menudo es suficiente conocer para algún i la distribución de $\boldsymbol{Z}_i$ condicionada a $\boldsymbol{Z}_T = \boldsymbol{z}_T$ (donde $T \subset \{1, \ldots, n\} - \{i\}$), siempre y cuando éstas condicionadas sirvan para determinar unívocamente la distribución conjunta. Muestrear a partir de un conjunto diferente de distribuciones condicionadas puede significar una convergencia más rápida como muestran Diebolt y Robert (1994). Aunque no siempre se da el caso de que variables con distribuciones condicionadas propias tengan marginales propias (en situaciones de este tipo no está garantizada la convergencia del algoritmo), en adelante supondremos la existencia de función de densidad (respecto de la medida de Lebesgue o la medida contable) para todas las marginales y condicionadas.

Considerando un vector de valores iniciales $(\boldsymbol{z}_1^{(0)}, \ldots, \boldsymbol{z}_n^{(0)})$ extraído aleatoriamente de una distribución inicial arbitraria P_0, se generan iterativamente valores de las distribuciones condicionadas. La primera iteración se realiza del siguiente modo:

$$\begin{aligned}
\boldsymbol{z}_1^{(1)} &\sim f(\boldsymbol{z}_1 \mid \boldsymbol{z}_2^{(0)}, \boldsymbol{z}_3^{(0)}, \ldots, \boldsymbol{z}_n^{(0)}) \\
\boldsymbol{z}_2^{(1)} &\sim f(\boldsymbol{z}_2 \mid \boldsymbol{z}_1^{(1)}, \boldsymbol{z}_3^{(0)}, \ldots, \boldsymbol{z}_n^{(0)}) \\
&\vdots \\
\boldsymbol{z}_n^{(1)} &\sim f(\boldsymbol{z}_n \mid \boldsymbol{z}_1^{(1)}, \boldsymbol{z}_2^{(1)}, \ldots, \boldsymbol{z}_{n-1}^{(1)}).
\end{aligned}$$

Siguiendo el mismo esquema se realiza la iteración s del siguiente modo:

$$\begin{aligned} z_1^{(s)} &\sim f(z_1 \mid z_2^{(s-1)}, z_3^{(s-1)}, \ldots, z_n^{(s-1)}) \\ z_2^{(s)} &\sim f(z_2 \mid z_1^{(s)}, z_3^{(s-1)}, \ldots, z_n^{(s-1)}) \\ &\vdots \\ z_n^{(s)} &\sim f(z_n \mid z_1^{(s)}, z_2^{(s)}, \ldots, z_{n-1}^{(s)}). \end{aligned}$$

Después de replicar este mismo esquema un número S de veces, se obtiene una secuencia de S vectores n-dimensionales $(z_1^{(1)}, \ldots, z_n^{(1)}), \ldots, (z_1^{(S)}, \ldots, z_n^{(S)})$. Geman y Geman (1984) prueban que bajo condiciones de regularidad esta secuencia converge en distribución a $(Z_1, \ldots, Z_n)$ y, por tanto, $\mathcal{L}(Z_k^{(j)}) \to P(z_k)$ cuando $j \to \infty$, donde P es la distribución marginal de Z_k. La demostración está basada en la naturaleza markoviana de las iteraciones y se realiza en dos etapas: (1) se demuestra que la secuencia $(Z^{(1)}, Z^{(2)}, \ldots)$ es una cadena de Markov con una única distribución estacionaria; y (2) se demuestra que la distribución estacionaria coincide con la distribución objetivo. Geman y Geman (1984), Schervish y Carlin (1992) y Liu *et al.* (1992) prueban que la tasa de convergencia es geométrica (con la norma del supremo), y Tierney (1994) estudia la convergencia con la norma de la variación total. Además, la secuencia que se genera con el Gibbs Sampling es ergódica: para cualquier función medible T de $(Z_1, \ldots, Z_n)$, cuya esperanza existe, se verifica que

$$\lim_{j \to \infty} \frac{1}{j} \sum_{l=1}^{j} T(Z_1^{(l)}, \ldots, Z_n^{(l)}) = E[T(Z_1, \ldots, Z_n)] \quad \text{c.s.}$$

Para un número de iteraciones S suficientemente grande podemos considerar cada $z_i^{(S)}$ como una muestra aleatoria de la distribución marginal de la variable. Si se repite R veces este proceso obtenemos una muestra de variables aleatorias independientes $z_{i_1}^{(S)}, z_{i_2}^{(S)}, \ldots, z_{i_R}^{(S)}$ con la que podemos estimar la densidad de la variable. Existen diversos métodos para estimar la densidad, entre los que destacamos los dos tipos de

estimadores que se han empleado más frecuentemente en el contexto del Gibbs Sampling: estimadores tipo núcleo y estimadores tipo Rao-Blackwell.

Un estimador tipo núcleo y con parámetro de suavizado h es

$$\hat{f}_{\mathrm{S,R}}(\boldsymbol{z}_i) = \frac{1}{R\,h}\sum_{r=1}^{\mathrm{R}} \mathrm{K}\left(\frac{\boldsymbol{z}_i - \boldsymbol{z}_{ir}^{(\mathrm{S})}}{h}\right),$$

donde $\mathbf{K}$ es una función de densidad, generalmente simétrica.

El estimador tipo Rao-Blackwell surge de manera natural al expresar la densidad marginal como la esperanza de la densidad condicionada

$$\begin{aligned} f(\boldsymbol{z}_i) &= \int f(\boldsymbol{z}_i \mid \boldsymbol{z}_{(i)})\, f(\boldsymbol{z}_{(i)})\, d\boldsymbol{z}_{(i)} \\ &= E_{\boldsymbol{z}_{(i)}}\left(f(\boldsymbol{z}_i \mid \boldsymbol{z}_{(i)})\right), \end{aligned}$$

y empleando como estimador el análogo muestral, de la forma

$$\hat{f}_{\mathrm{S,R}}(\boldsymbol{z}_i) = \frac{1}{R}\sum_{r=1}^{\mathrm{R}} f(\boldsymbol{z}_i \mid \boldsymbol{z}_{(i)r}^{(\mathrm{S})}).$$

Este estimador es más eficiente ya que recoge toda la información conocida sobre la distribución e incorpora la información que aporta la obtención de una muestra equivalente para el resto de variables. Este resultado lo prueban Gelfand y Smith (1990) para muestras independientes, y Liu *et al.* (1994) en el caso general. Del mismo modo el estimador de la media $\mu = E(T(\boldsymbol{Z}_i))$ dado por

$$\hat{\mu}_{\mathrm{S,R}} = \frac{1}{R}\sum_{r=1}^{\mathrm{R}} E\left(T(\boldsymbol{Z}_i) \mid \boldsymbol{z}_{(i)r}^{(\mathrm{S})}\right)$$

es más eficiente que la media muestral de $T(\boldsymbol{z}_{i_1}^{(\mathrm{S})}), \ldots, T(\boldsymbol{z}_{i_{\mathrm{R}}}^{(\mathrm{S})})$.

La elección del número R de veces que repetimos la ejecución del algoritmo dependerá, por tanto, de las propiedades asintóticas de los estimadores propuestos (Silverman 1986). Conviene advertir que, con frecuencia, estos estimadores se emplean cuando se

considera una muestra formada por las R últimas iteraciones de una única secuencia. En este caso se debe tener en cuenta que, en general, las observaciones no son independientes.

Así pues, para implementar el algoritmo es suficiente disponer de las distribuciones condicionadas y de técnicas eficientes de simulación de variables aleatorias. Para esto son útiles los métodos propuestos por Devroye (1986) y Ripley (1987). En un contexto bayesiano —donde las distribuciones necesarias para implementar el Gibbs Sampling son las correspondientes a cada uno de los parámetros condicionados a la muestra observada y al resto de los parámetros—, se emplean frecuentemente distribuciones a priori conjugadas con objeto de simplificar los cálculos. En general, esta restricción no refleja la situación real (ver, por ejemplo, Carlin y Polson, 1991b) y cuando se elimina es muy frecuente que la expresión analítica de alguna de las distribuciones condicionadas no se pueda calcular. Existen varias soluciones a este problema, la más sencilla consiste en evaluar el producto de la verosimilitud y la distribución a priori en una red de valores del parámetro para valores fijos del resto de las variables y de los datos de la muestra. De esta forma, la distribución puede ser aproximada salvo por una constante que se obtiene con técnicas de integración numérica. Incluso sin necesidad de normalizar es posible, a partir de muestras generadas de otra distribución, obtener muestras aleatorias de las distribuciones condicionadas aplicando la técnica de *muestreo de rechazo* (Smith y Gelfand, 1992). Para aplicar este método, el cociente entre el producto de la verosimilitud y la distribución a priori y la densidad de la distribución que se muestrea debe estar acotado. Se pueden emplear, alternativamente, modificaciones como el método SIR (o "bootstrap ponderado") de Rubin (1988) o el método adaptativo de muestreo de rechazo para funciones de densidad log-cóncava (el logaritmo de la densidad es una función cóncava) de Gilks y Wild (1992). La naturaleza adaptativa de esta técnica permite extraer muestras a partir de un pequeño número

de evaluaciones del producto de la verosimilitud y la distribución a priori, mejorando la eficiencia en comparación con el método no adaptativo. En aplicaciones del Gibbs Sampling también se emplea el Griddy-Gibbs de Ritter y Tanner (1992), más simple y fácil de aplicar que los mencionados anteriormente aunque no permite extraer muestras de la distribución exacta.

1.4.3 Control de la convergencia

Ante las buenas propiedades asintóticas del algoritmo no se debe olvidar que únicamente es posible realizar un número finito de iteraciones y que, en general, las muestras obtenidas en cada iteración no son independientes y sólo tienen la misma distribución en el límite. Gelman y Rubin (1992a) muestran la importancia que tienen las condiciones iniciales en la velocidad a la que converge el algoritmo en un problema en el que el espacio paramétrico tiene una dimensión alta. Matthews (1993) proporciona un ejemplo en el que el Gibbs Sampling aparentemente converge cuando realmente no se ha alcanzado la convergencia. Smith y Roberts (1993) y Mengersen y Robert (1995) puntualizan que cuando la distribución de los parámetros es bimodal, el Gibbs Sampling puede quedar atrapado en una de las modas a lo largo de muchas iteraciones, reduciéndose así la probabilidad de que se alcance la convergencia. Hills y Smith (1992) destacan que el número de iteraciones necesarias para alcanzar la convergencia es una función de los valores iniciales y de la estructura de correlación de los procesos estocásticos generados por el Gibbs Sampling; la conclusión a la que llegan es que la correlación es el problema más serio para alcanzar la convergencia. Polson (1995) analiza una cota sobre la tasa de convergencia que puede usarse para elegir el número de iteraciones que garantizan la precisión necesaria para realizar la estimación con la muestra obtenida mediante Gibbs Sampling; el número de iteraciones necesarias depende nuevamente de la correlación y la dimensión del espacio paramétrico. Además,

en los capítulos 2 y 4 se muestra como la covergencia es muy lenta en la aplicación del Gibbs Sampling en modelos de regresión y series temporales con datos heterogéneos. Por tanto, es muy importante contar con algún mecanismo que nos permita decidir el número de iteraciones necesario en cada aplicación del algoritmo. En general, este problema no ha sido resuelto de manera concluyente, aunque se han propuesto diversos métodos basados tanto en el seguimiento de una serie como en el seguimiento conjunto de varias de ellas.

El diagnóstico basado en una única secuencia juzga la convergencia del algoritmo usando técnicas propias del análisis de series temporales univariantes. En esta línea se encuentra el método propuesto por Geweke (1992) para evaluar la aproximación en el cálculo de momentos de la distribución a posteriori usando Gibbs Sampling. Esta propuesta está basada en técnicas habituales en el análisis espectral de series temporales. Hastings (1970), Raftery y Lewis (1992) y Robert (1994) proponen otros métodos.

Dentro de las propuestas multi-secuenciales, Fosdick (1959) sugiere simular varias secuencias y fijar el criterio de parada cuando la diferencia entre las medias de las series sea menor que una cierta cota prefijada. Ripley (1987) propone comparar como mínimo tres series combinando métodos gráficos con métodos propios de series temporales. El método propuesto por Gelfand *et al.* (1990) consiste en estimar las funciones de densidad de algunas marginales variando el número de iteraciones y estudiar gráficamente si las densidades son indistinguibles. El Gibbs-Stopper de Ritter y Tanner (1992) utiliza una distribución de *importancia* para construir una sucesión de pesos que deben converger a una variable aleatoria con distribución degenerada y cuyo seguimiento nos informa de la distribución de las variables generadas. Gelman y Rubin (1992b) también proponen un método en el que se comparan varias secuencias mediante estadísticos basados en medias y varianzas. La convergencia se juzga por el *potencial de reducción*

de variabilidad definido en Gelman y Rubin (1992b), que se estima con un estadístico que incorpora la variabilidad de las series y entre series y que converge a uno cuando el número de iteraciones tiende a infinito. Los valores iniciales deben generarse a partir de una distribución sobredispersa con relación a la verdadera distribución marginal. Las principales ventajas de este método son que es completamente cuantitativo y que también sirve para detectar problemas de multimodalidad, en la proximidad de una moda el potencial de reducción no continúa aproximándose a uno.

1.4.4 Comparación con otros mecanismos de simulación

Evidentemente, no todos los algoritmos y métodos de remuestreo que han aparecido en los últimos años son comparables con el Gibbs Sampling, y en algunos casos dicha comparación carece de interés debido a la poca relación existente entre los problemas que permiten resolver. En esta revisión se comenta brevemente la relación con los métodos que se emplean en el mismo contexto en el que se aplica el Gibbs Sampling.

Existe una estrecha relación entre el Gibbs Sampling y el algoritmo de Sustitución propuesto por Tanner y Wong (1987), y en el caso de dos variables ambos métodos coinciden. En general, se necesitan conocer menos distribuciones condicionadas para aplicar el Gibbs Sampling, aunque si se dispone de alguna distribución adicional es posible aplicar el algoritmo de Sustitución de manera que se acelera la convergencia, llegando a ser más eficiente que el Gibbs Sampling. Chib (1992) combina en un ejemplo estos dos métodos obteniendo resultados más satisfactorios que si emplea sólo uno de los algoritmos.

Las hipótesis de partida para poder aplicar el "muestreo de importancia" propuesto por Rubin (1987) varían sustancialmente de las requeridas para aplicar el Gibbs Sampling y se basan en el conocimiento de la distribución conjunta, así como de alguna

distribución de importancia. Las conclusiones derivadas de resultados empíricos (véase Gelfand y Smith, 1990) muestran que el Gibbs Sampling genera las variables aleatorias deseadas de forma más eficiente.

Por último, los métodos más habituales para el cálculo de las distribuciones y momentos de interés son el método de integración de Monte Carlo (Kloek y van Dijk, 1978) y el método de Laplace (Tierney y Kadane, 1986). Ambos requieren un estudio analítico previo y en muchos aspectos son complementarios. Son muchas las cuestiones que no han sido estudiadas acerca de la relación entre Gibbs Sampling, Integración de Monte Carlo y Expansiones de Laplace, no existiendo una respuesta precisa cuando se plantea la elección entre cualquiera de estos métodos para resolver un determinado problema. Una discusión más extensa sobre estos métodos puede encontrarse en Geweke (1992).

Capítulo 2

Gibbs Sampling en problemas de regresión con datos heterogéneos

2.1 Introducción

En este capítulo se extiende la aplicación del Gibbs Sampling propuesta por Verdinelli y Wasserman (1991), para la identificación de valores atípicos en una muestra, a un modelo de regresión lineal. En el contexto de regresión, si existen valores atípicos que se enmascaran o señalan como atípicas a otras observaciones que no lo son, los parámetros del modelo estarán altamente correlacionados y la convergencia no se alcanzará en un número razonable de iteraciones. El algoritmo puede proporcionar una idea falsa de las probabilidades a posteriori ya que los resultados que se obtienen aumentando el número de iteraciones permanecen estables en torno a valores distintos de los verdaderos.

Este capítulo se organiza del siguiente modo. En la sección 2.2 se presenta la aplicación del Gibbs Sampling a la identificación de valores atípicos en problemas de regresión lineal. Se supone un modelo normal de contaminación de escala y se examina la convergencia del algoritmo en algunos ejemplos con datos reales y simulados. En la

sección 2.3 se analizan los problemas de convergencia asociados a muestras con valores atípicos muy influyentes. En la sección 2.4 se presenta un modelo semiparamétrico bayesiano para la identificación de atípicos en el que se dan los mismos problemas de convergencia, poniendo de manifiesto que la velocidad de convergencia no depende del modelo seleccionado.

2.2 Gibbs Sampling para la identificación de datos atípicos

2.2.1 El modelo de contaminación de escala

La falta de homogeneidad en la muestra se asocia muy frecuentemente a la idea de que la distribución de los datos es una mezcla de distribuciones. Los datos proceden de una distribución central con alta probabilidad, y con una probabilidad pequeña de una distribución que contamina. En este capítulo se supone el modelo normal de contaminación de escala para identificar valores atípicos. Este modelo fue propuesto por Tukey (1960) y estudiado, entre otros, por Box y Tiao (1968), Freeman (1980), Pettit (1992) y Peña y Tiao (1992).

En el modelo normal de contaminación de escala las observaciones $\boldsymbol{y} = (y_1, \ldots, y_n)'$ son generadas por el modelo

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i \qquad i = 1, \ldots, n, \tag{2.1}$$

donde n es el tamaño muestral, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ son variables fijas, $\boldsymbol{\beta}$ es un vector $p \times 1$ de parámetros desconocidos, y u_i son variables aleatorias que se distribuyen según una mixtura de normales dada por

$$u_i \sim (1-\alpha)\, N(0, \sigma^2) + \alpha\, N(0, k^2\sigma^2) \qquad i = 1, \ldots, n. \tag{2.2}$$

Se supone que el parámetro de escala k es conocido y que $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ es una matriz de rango completo.

La mezcla de distribuciones de la ecuación (2.2) indica que existe una probabilidad α de que cada dato haya sido espúreamente generado a partir de la distribución contaminante. Aquellos datos generados por la distribución alternativa se consideran valores atípicos.

El modelo de contaminación de escala, a diferencia de los de contaminación en la media de Abraham y Box (1978) o Guttman, Dutter y Freeman (1978), permite detectar valores atípicos en ambas colas de la distribución y el número de parámetros es constante ya que no depende de la cantidad de atípicos que haya en la muestra. La ventaja de este modelo frente a los robustos que emplean distribuciones con colas pesadas (Box y Tiao, 1973 y West, 1984), o frente a los que no suponen modelo alternativo, es que no sólo facilita un procedimiento de identificación de atípicos, sino que además proporciona una estimación robusta y eficiente de los parámetros. Además, cuando k es suficientemente grande, Peña y Guttman (1993) demuestran que la identificación de valores atípicos es esencialmente la misma con este modelo y con los de contaminación en la media o los que no suponen modelo alternativo y emplean el método de la ordenada de la distribución predictiva condicionada (Geisser, 1980).

Box y Tiao (1968) obtienen las distribuciones a posteriori en el modelo (2.1) y (2.2) cuando se suponen distribuciones a priori no informativas e independientes para los parámetros de localización y escala, la distribución conjunta a priori de $\boldsymbol{\beta}$ y σ^2 es

$$P(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}. \tag{2.3}$$

Cuando el parámetro de contaminación α es conocido, la probabilidad a posteriori de que sólo los datos pertenecientes a un determinado conjunto de n_{I} observaciones,

indexadas por $I = \{i_1, \ldots, i_{n_I}\}$, sean atípicos viene dada por la expresión

$$p_I \propto \left(\frac{\alpha}{1-\alpha}\right)^{n_I} k^{-n_I} \left(\frac{\mid \boldsymbol{X}'\boldsymbol{X} \mid}{\mid \boldsymbol{X}'\boldsymbol{X} - \phi \boldsymbol{X}_I'\boldsymbol{X}_I \mid}\right)^{\frac{1}{2}} \left(\frac{s^2}{s^2_{(I)}}\right)^{\frac{n-p}{2}}, \tag{2.4}$$

donde $\phi = 1 - k^{-2}$, $\boldsymbol{X}_I$ es una matriz de dimensión $n_I \times p$ cuyas filas corresponden a las filas indexadas por I de la matriz $\boldsymbol{X}$, s^2 es el estimador insesgado de la varianza residual y $s^2_{(I)}$ es el estimador que se obtiene teniendo en cuenta los n_I datos considerados como atípicos.

Cuando k es suficientemente grande, la probabilidad (2.4) se puede aproximar por

$$p_I \propto \omega^{n_I} \mid \boldsymbol{I} - \boldsymbol{H}_I \mid^{-1/2} s^{-n+p}_{(I)}, \tag{2.5}$$

donde $\omega = \alpha/k(1-\alpha)$, $\boldsymbol{H}_I = \boldsymbol{X}_I(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_I'$, y ahora $s^2_{(I)}$ es el estimador insesgado de σ^2 que se obtiene eliminando de la muestra las observaciones $(\boldsymbol{y}_I, \boldsymbol{X}_I)$.

Las ecuaciones (2.4) y (2.5) pueden utilizarse fácilmente para comprobar si existe un único dato atípico en la muestra, o si existe un determinado grupo de observaciones que han sido espúreamente generadas. Sin embargo, cuando la cantidad y posición de los valores atípicos es desconocida, que es lo más habitual con datos reales, la identificación de estos datos requiere el cálculo de las 2^n probabilidades a posteriori correspondientes a todas las configuraciones posibles de datos buenos y atípicos. Por ejemplo, si tenemos una muestra de tamaño $n = 40$, tenemos que calcular 2^{40} probabilidades (aproximadamente 10^{12}).

Siguiendo a Verdinelli y Wasserman (1991) se puede extender la aplicación del Gibbs Sampling para la identificación de valores atípicos, a la estimación de las probabilidades a posteriori de que cada observación sea atípica en el modelo (2.1) y (2.2). La aplicación del Gibbs Sampling se lleva a cabo aumentando el vector de parámetros con un conjunto

de variables de clasificación $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)'$, que se definen como

$$\delta_i = \begin{cases} 1 & \text{si } y_i \sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, k^2\sigma^2) \\ 0 & \text{si } y_i \sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2). \end{cases}$$

Diremos que $(y_i, \boldsymbol{x}_i')$ es un dato atípico cuando la probabilidad marginal a posteriori p_i (la probabilidad a posteriori de que la variable de clasificación i-ésima sea igual a uno) sea mayor que 0.5. Así α es la probabilidad a priori que tiene cada observación de ser atípica.

2.2.2 Aplicación del Gibbs Sampling

Para poder aplicar el Gibbs Sampling en la resolución del modelo (2.1) y (2.2) es necesario generar muestras de las distribuciones condicionadas a posteriori de todos los parámetros. En este caso las distribuciones de los parámetros se pueden obtener fácilmente cuando se conoce qué observaciones contaminan la muestra, es decir, cuando se especifica el vector $\boldsymbol{\delta}$; del mismo modo, la expresión de cada una de las probabilidades de que un dato sea atípico se pueden calcular si los parámetros del modelo son conocidos. Con las distribuciones a priori para $\boldsymbol{\beta}$ y σ^2 dadas en la ecuación (2.3), y suponiendo para el parámetro de contaminación α una distribución a priori $Beta(\gamma_1, \gamma_2)$ con media $\alpha_0 = E(\alpha) = \gamma_1/(\gamma_1 + \gamma_2)$, las distribuciones condicionadas son:

1. La distribución condicionada de $\boldsymbol{\beta}$ es

$$\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{\delta}, \sigma^2 \sim N_p\left(\tilde{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\right),$$

donde

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$$

y $\boldsymbol{V}$ es una matriz diagonal con elementos en la diagonal principal $v_{ii} = k^2$ si $\delta_i = 1$ y $v_{ii} = 1$ en caso contrario.

2. Sean u_i^* los errores estándar definidos como $u_i^* = (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})/(1 + \delta_i(k-1))$ para $i = 1, \ldots, n$. La distribución de la precisión de los errores es

$$\sigma^{-2} \mid \boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{\beta} \sim \ Gamma \left(\frac{n}{2}, \frac{1}{2} \sum_{i=1}^{n} u_i^{*2} \right)$$

y, por tanto,

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} u_i^{*2} \sim \chi_n^2.$$

3. La distribución condicionada de δ_i es *Bernoulli* con probabilidad de éxito

$$P(\delta_i = 1 \mid \boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, \alpha) = \frac{\exp\left(-u_i^2/2k^2\sigma^2\right)\alpha}{\exp\left(-u_i^2/2k^2\sigma^2\right)\alpha + \exp\left(-u_i^2/2\sigma^2\right)(1-\alpha)k}. \quad (2.6)$$

4. La distribución condicionada de α depende únicamente del vector $\boldsymbol{\delta}$ y es

$$\alpha \mid \boldsymbol{\delta} \sim \ Beta\left(\gamma_1 + n\bar{\delta}, \gamma_2 + n(1-\bar{\delta})\right), \quad (2.7)$$

donde $n\bar{\delta} = \sum_{i=1}^{n} \delta_i$. La media de la distribución (2.7) es una combinación lineal de la media a priori y la media muestral

$$E(\alpha \mid \boldsymbol{\delta}) = \frac{\gamma_1 + \gamma_2}{\gamma_1 + \gamma_2 + n}\alpha_0 + \frac{n}{\gamma_1 + \gamma_2 + n}\bar{\delta}.$$

Estas distribuciones condicionadas pertenecen a familias de distribuciones paramétricas conocidas y para las que se han desarrollado diversos métodos de generación de muestras aleatorias (véase, por ejemplo, Devroye, 1986).

Las iteraciones del Gibbs Sampling comienzan con un vector arbitrario de valores iniciales $(\alpha^{(0)}, \sigma^{2(0)}, \boldsymbol{\delta}^{(0)}, \boldsymbol{\beta}^{(0)})$. Siguiendo un número S de iteraciones el esquema descrito en el capítulo de introducción , se obtiene la secuencia $(\alpha^{(0)}, \sigma^{2(1)}, \boldsymbol{\delta}^{(1)}, \boldsymbol{\beta}^{(1)}), \ldots,$ $(\alpha^{(s)}, \sigma^{2(s)}, \boldsymbol{\delta}^{(s)}, \boldsymbol{\beta}^{(s)})$, que converge en distribución a la conjunta de $(\alpha, \sigma^2, \boldsymbol{\delta}, \boldsymbol{\beta})$. Cuando se ejecuta el Gibbs Sampling R veces en paralelo se puede estimar la media, la varianza o cualquier otra característica de las distribuciones a posteriori a partir de las muestras

independientes e idénticamente distribuidas obtenidas en la última iteración

$$\alpha_1^{(S)}, \ldots, \alpha_R^{(S)}$$
$$\sigma_1^{2(S)}, \ldots, \sigma_R^{2(S)}$$
$$\boldsymbol{\delta}_1^{(S)}, \ldots, \boldsymbol{\delta}_R^{(S)}$$
$$\boldsymbol{\beta}_1^{(S)}, \ldots, \boldsymbol{\beta}_R^{(S)}.$$

En particular, los estimadores de la probabilidad marginal a posteriori de que cada observaciones sea atípica son

$$\hat{p}_{i_R}^{(S)} = \frac{1}{R} \sum_{r=1}^{R} \delta_{ir}^{(S)} \qquad i = 1, \ldots, n. \tag{2.8}$$

El estimador de la densidad de $\boldsymbol{\beta}$ es un estimador no paramétrico tipo núcleo. Alternativamente, es posible estimar p_i con las últimas R iteraciones de una única secuencia. Aunque ejecutar el algoritmo sólo una vez puede suponer un ahorro computacional, tiene el inconveniente de que las muestras así obtenidas son idénticamente distribuidas pero no independientes. Considerando que en los ejemplos que se muestran a continuación la dimensión del espacio paramétrico (el tamaño muestral más los parámetros del modelo) es moderada, siempre se ejecuta el algoritmo en paralelo y se emplea el estimador (2.8) para estimar p_i. Además, en la sección 2.3 se discute la sensibilidad del Gibbs Sampling a la especificación de las condiciones iniciales, ejecutando el algoritmo en paralelo se evita que las conclusiones dependan de la selección de un único vector de valores iniciales.

Los estimadores (2.8) de las probabilidades a posteriori para cada dato, calculados en cada iteración constituyen las series de probabilidades que se emplearán para controlar la convergencia.

2.2.3 Ejemplos

El comportamiento del Gibbs Sampling en el problema de identificación de datos atípicos en regresión se ilustra en cuatro ejemplos. En el primero se aplica el algoritmo a un conjunto de datos frecuentemente analizados con distintos procedimientos de identificación de grupos de atípicos. En este ejemplo la convergencia es muy rápida y los datos atípicos se identifican inmediatamente. Sin embargo, como se muestra en los siguientes ejemplos, el algoritmo puede no converger en un número razonable de iteraciones. Además, el Gibbs Sampling puede proporcionar una idea falsa de las probabilidades ya que las series de probabilidades presentan un comportamiento estable en torno a ciertos valores límite que no se corresponden con las verdaderas probabilidades a posteriori de que una observación sea atípica.

El algoritmo se ejecuta en todos los ejemplos 1000 veces (en paralelo) con distintos valores iniciales. Los valores de los parámetros que se obtienen en la última iteración de cada ejecución constituyen la muestra a partir de la cual se estiman las probabilidades a posteriori $\hat{p}_{i\mathrm{R}}^{(\mathrm{S})}$. Estas probabilidades se representan en los gráficos mediante una barra para cada dato muestral, y en el caso de las observaciones que son atípicas también con un punto.

En todos los ejemplos $k = 10$ y el criterio seguido para seleccionar los valores iniciales es: (1) $\alpha_0 = \gamma_1/(\gamma_1 + \gamma_2) = 0.2$ y $\gamma_1 + \gamma_2 = n$, lo que implica que en todas las iteraciones $E(\alpha \mid \boldsymbol{\delta}) = 1/2\alpha_0 + 1/2\bar{\delta}$; (2) seleccionar $\delta_i^{(0)} = 1$ con probabilidad α_0; y (3) $\boldsymbol{\beta}^{(0)}$ es el estimador de mínimos cuadrados generalizados, $\boldsymbol{\beta}^{(0)} = (\boldsymbol{X}'\boldsymbol{V}^{(0)^{-1}}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{(0)^{-1}}\boldsymbol{y}$, donde $\boldsymbol{V}^{(0)}$ es una matriz diagonal con elementos en la diagonal principal $v_{ii} = k^2$ si $\delta_i^{(0)} = 1$, y $v_{ii} = 1$ en caso contrario. No es necesario especificar valores iniciales ni para la varianza σ^2 porque es el primer parámetro que se calcula en cada iteración, ni para α porque depende únicamente de $\boldsymbol{\delta}$.

i	y_i	x_{1_i}	x_{2_i}	x_{3_i}	i	y_i	x_{1_i}	x_{2_i}	x_{3_i}	i	y_i	x_{1_i}	x_{2_i}	x_{3_i}
1	42	80	27	89	8	20	62	24	93	15	8	50	18	89
2	37	80	27	88	9	15	58	23	87	16	7	50	18	86
3	37	75	25	90	10	14	58	18	80	17	8	50	19	72
4	28	62	24	87	11	14	58	18	89	18	8	50	19	79
5	18	62	22	87	12	13	58	17	88	19	9	50	20	80
6	18	62	23	87	13	11	58	18	82	20	15	56	20	82
7	19	62	24	93	14	12	58	19	93	21	15	70	20	91

Tabla 2.1: Datos "Stack Loss".

Ejemplo 1: Los datos "Stack Loss"

El conjunto de datos conocido en inglés como *Stack Loss data* (Brownlee, 1965) se recoge en la tabla 2.1. Se trata de 21 observaciones diarias de un proceso que se realiza en una planta para la oxidación del amoníaco a ácido nítrico. Los datos corresponden a tres variables explicativas y una variable respuesta (además el modelo incluye un término constante). Estos datos han sido estudiados con distintos procedimientos para la identificación de observaciones atípicas y se han encontrado 4 valores atípicos en los datos 1, 3, 4 y 21 (véase, por ejemplo, Daniel y Wood, 1980 o Rousseeuw y van Zomeren, 1990). Algunos autores también incluyen la observación 2 en la lista de atípicos. Aunque no se consideran unos datos especialmente difíciles, los métodos tradicionales de diagnóstico únicamente permiten identificar las observaciones 4 y 21 como atípicas.

Después de realizar 500 iteraciones del algoritmo, las probabilidades a posteriori de que cada observación sea atípica se representan en la figura 2.1(a). Los resultados que se observan confirman que los datos 1, 3, 4 y 21 son atípicos, con probabilidades

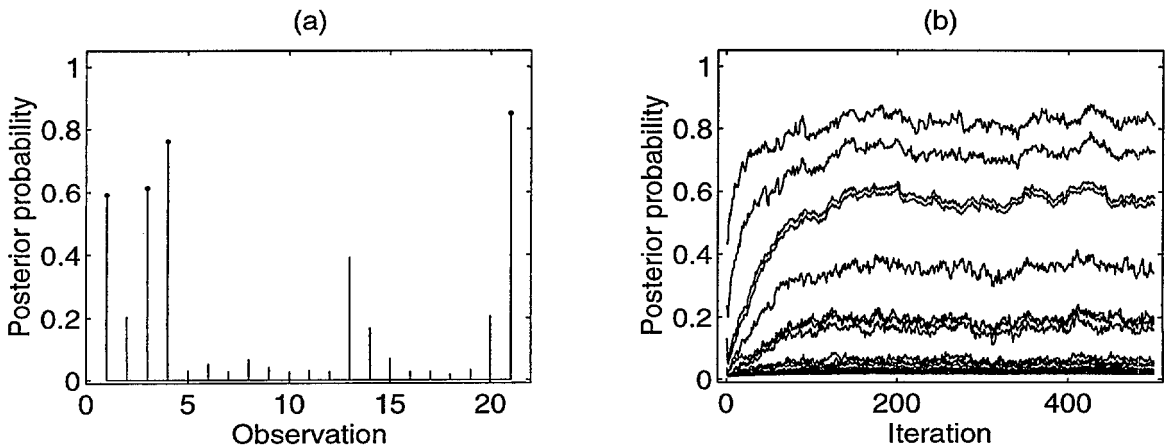

Figura 2.1: Resultados del Gibbs Sampling con los datos *Stack Loss*: (a) probabilidades a posteriori de que cada dato sea atípico con 500 iteraciones; (b) probabilidades a posteriori en función del número de iteración.

superiores a 0.5. En la figura 2.1(b) se representan las series de probabilidad a posteriori para cada dato en función del número de iteraciones. Se observa que la convergencia se alcanza en pocas iteraciones, menos de 200.

Ejemplo 2: Datos de Hawkins, Bradu y Kass

Los datos construidos por Hawkins, Bradu y Kass (1984) son un ejemplo muy típico de enmascaramiento. Se recogen en la tabla 2.2 y corresponden a 75 observaciones de cuatro variables en un modelo con término constante. En la figura 2.2 se muestran las proyecciones de la nube de puntos sobre todos los planos que se obtienen tomando como ejes dos variables distintas. Los datos que aparecen más alejados en todos los gráficos corresponden a las 14 primeras observaciones. Estos datos tienen un alto potencial, pero únicamente las observaciones de la 1 a la 10 son atípicas. En este ejemplo los valores atípicos no son fácilmente identificables con procedimientos de detección

i	y_i	x_{1_i}	x_{2_i}	x_{3_i}	i	y_i	x_{1_i}	x_{2_i}	x_{3_i}	i	y_i	x_{1_i}	x_{2_i}	x_{3_i}
1	9.7	10.1	19.6	28.3	26	-0.8	0.9	3.3	2.5	51	0.7	2.3	1.5	0.4
2	10.1	9.5	20.5	28.9	27	-0.7	3.3	2.5	2.9	52	-0.5	3.3	0.6	1.2
3	10.3	10.7	20.2	31.0	28	0.3	1.8	0.8	2.0	53	0.7	0.3	0.4	3.3
4	9.5	9.9	21.5	31.7	29	0.3	1.2	0.9	0.8	54	0.7	1.1	3.0	0.3
5	10.0	10.3	21.1	31.1	30	-0.3	1.2	0.7	3.4	55	0.0	0.5	2.4	0.9
6	10.0	10.8	20.4	29.2	31	0.0	3.1	1.4	1.0	56	0.1	1.8	3.2	0.9
7	10.8	10.5	20.9	29.1	32	-0.4	0.5	2.4	0.3	57	0.7	1.8	0.7	0.7
8	10.3	9.9	19.6	28.8	33	-0.6	1.5	3.1	1.5	58	-0.1	2.4	3.4	1.5
9	9.6	9.7	20.7	31.0	34	-0.7	0.4	0.0	0.7	59	-0.3	1.6	2.1	3.0
10	9.9	9.3	19.7	30.3	35	0.3	3.1	2.4	3.0	60	-0.9	0.3	1.5	3.3
11	-0.2	11.0	24.0	35.0	36	-1.0	1.1	2.2	2.7	61	-0.3	0.4	3.4	3.0
12	-0.4	12.0	23.0	37.0	37	-0.6	0.1	3.0	2.6	62	0.6	0.9	0.1	0.3
13	-0.7	12.0	26.0	34.0	38	0.9	1.5	1.2	0.2	63	-0.3	1.1	2.7	0.2
14	0.1	11.0	34.0	34.0	39	-0.7	2.1	0.0	1.2	64	-0.5	2.8	3.0	2.9
15	-0.4	3.4	2.90	2.10	40	-0.5	0.5	2.0	1.2	65	0.6	2.0	0.7	2.7
16	0.6	3.1	2.2	0.3	41	-0.1	3.4	1.6	2.9	66	-0.9	0.2	1.8	0.8
17	-0.2	0.0	1.6	0.2	42	-0.7	0.3	1.0	2.7	67	-0.7	1.6	2.0	1.2
18	0.0	2.3	1.6	2.0	43	0.6	0.1	3.3	0.9	68	0.6	0.1	0.0	1.1
19	0.1	0.8	2.9	1.6	44	-0.7	1.8	0.5	3.2	69	0.2	2.0	0.6	0.3
20	0.4	3.1	3.4	2.2	45	-0.5	1.9	0.1	0.6	70	0.7	1.0	2.2	2.9
21	0.9	2.6	2.2	1.9	46	-0.4	1.8	0.5	3.0	71	0.2	2.2	2.5	2.3
22	0.3	0.4	3.2	1.9	47	-0.9	3.0	0.1	0.8	72	-0.2	0.6	2.0	1.5
23	-0.8	2.0	2.3	0.8	48	0.1	3.1	1.6	3.0	73	0.4	0.3	1.7	2.2
24	0.7	1.3	2.3	0.5	49	0.9	3.1	2.5	1.9	74	-0.9	0.0	2.2	1.6
25	-0.3	1.0	0.0	0.4	50	-0.4	2.1	2.8	2.9	75	0.2	0.3	0.4	2.6

Tabla 2.2: Datos de Hawkins, Bradu y Kass.

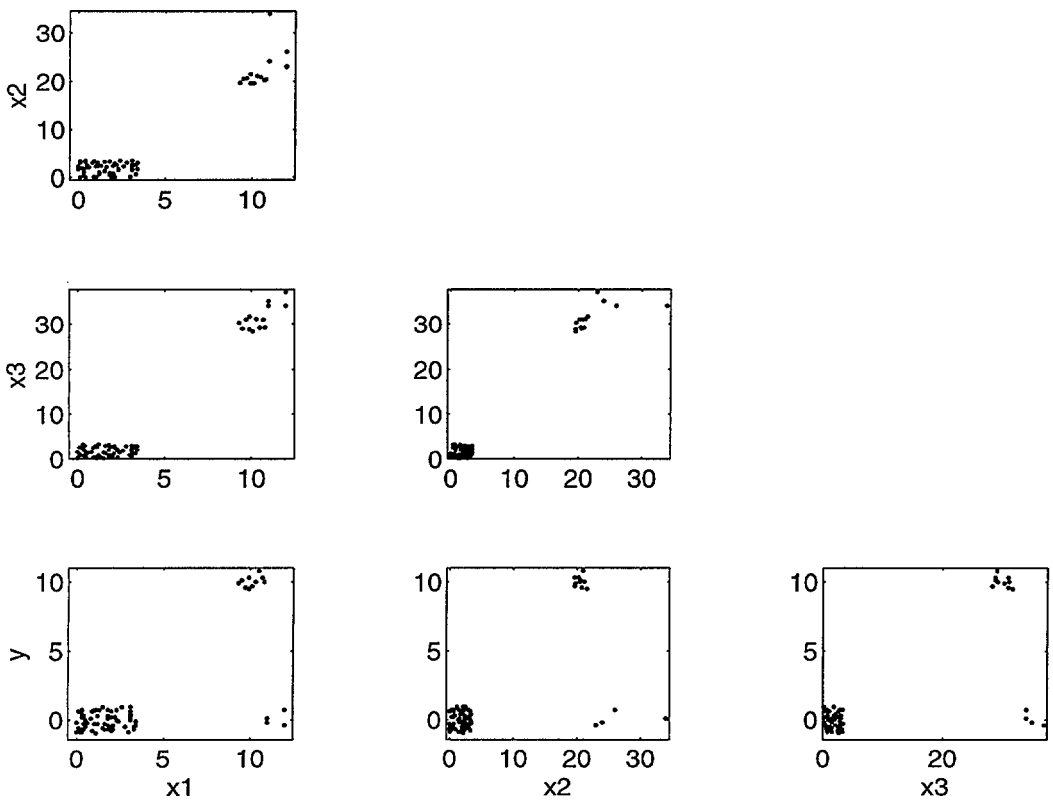

Figura 2.2: Proyecciones de la nube de puntos de los datos de Hawkins, Bradu y Kass.

individual debido al enmascaramiento: los diez datos atípicos se enmascaran y señalan como atípicas las observaciones 11 a 14 que son buenas.

La figura 2.3(a) muestra las probabilidades correspondientes a cada observación después de realizar 2000 iteraciones del Gibbs Sampling. Se observa claramente que los estimadores de las probabilidades están afectados por el efecto del enmascaramiento y que los valores atípicos no se identifican. También se observa que el señalamiento de datos buenos por parte de los atípicos provoca la detección errónea de las observaciones 11 a 14, con probabilidades casi iguales a uno. Estos resultados se alcanzan en muy pocas iteraciones como muestra la figura 2.3(b) y producen una falsa idea de convergencia al presentar las series un comportamiento muy estable en las primeras

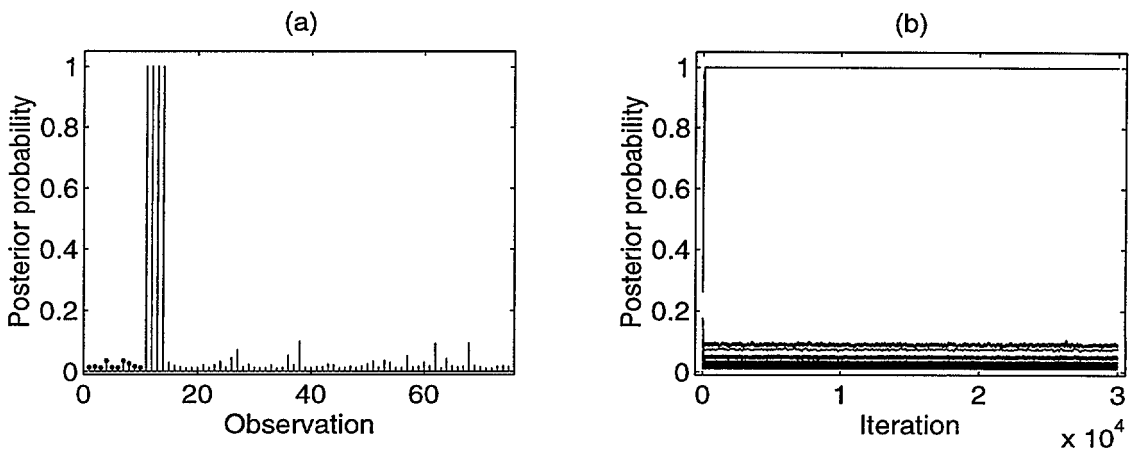

Figura 2.3: Resultados del Gibbs Sampling con los datos de Hawkins, Bradu y Kass: (a) probabilidades a posteriori de que cada dato sea atípico con 2000 iteraciones; (b) probabilidades a posteriori en función del número de iteración.

30000 iteraciones.

Ejemplo 3: Datos de Rousseeuw

El tercer ejemplo se construye siguiendo el esquema de simulación propuesto por Rousseeuw (1984). Se trata de 50 observaciones generadas en dos grupos (véase la figura 2.4) y cuyos valores numéricos se recogen en la tabla 2.3. La muestra contiene 20 datos atípicos (en la parte derecha del gráfico de la figura 2.4) que se generan a partir de una normal bivariante con vector de medias $\mu = (7, 2)'$, desviaciones típicas 0.5 y correlación 0. Los 30 datos buenos se generan a partir del modelo lineal dado por la ecuación $y_i = 2 + x_i + u_i$, donde las x_i siguen una distribución uniforme en el intervalo $(1, 4)$ y los errores u_i se distribuyen normales con media cero y desviación típica 0.2. La contaminación representa un 40 por ciento de los datos. La recta de regresión que se estima por mínimos cuadrados es la línea que aparece en la figura 2.4 y los residuos

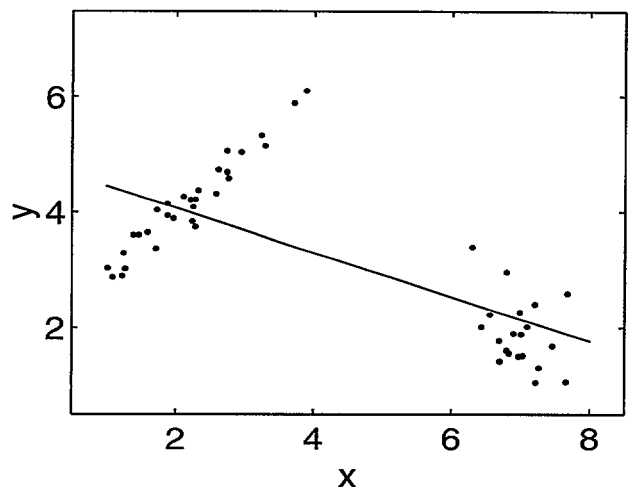

Figura 2.4: Datos simulados de Rousseeuw y estimador de mínimos cuadrados de la regresión

estudentizados construidos con esta estimación no permiten identificar valores atípicos entre el grupo de la derecha. Únicamente las observaciones 32 y 33 (parte superior del grupo de datos buenos) tienen un residuo estudentizado más grande al ser señaladas por el grupo de atípicos ($t_{32} = 2.77$ y $t_{33} = 3.16$).

Las probabilidades de que cada observación sea atípica después de 30000 iteraciones y las series correspondientes se muestran en la figura 2.5(a) y la figura 2.5(b), respectivamente. Se observa que las primeras 20 observaciones —las atípicas— no son identificadas cuando las series se estabilizan, reproduciéndose de nuevo los mismos problemas que aparecen en la identificación con los métodos para atípicos individuales.

Ejemplo 4: Diagrama de Hertzsprung-Russell

El diagrama de Hertzsprung-Russell de las estrellas del grupo CYG OB1 que se muestra en la figura 2.6 es un ejemplo con datos reales. Se observan dos variables de 47 estrellas situadas en la dirección de Cygnus. La variable independiente es el logaritmo

i	y_i	x_i	i	y_i	x_i	i	y_i	x_i	i	y_i	x_i	i	y_i	x_i
1	1.67	7.45	11	2.23	6.55	21	5.04	2.73	31	4.32	2.58	41	3.30	1.24
2	1.89	6.89	12	1.04	7.22	22	3.84	2.23	32	5.87	3.71	42	3.38	1.71
3	2.27	6.99	13	1.43	6.69	23	4.72	2.61	33	6.09	3.88	43	5.02	2.94
4	2.96	6.79	14	2.59	7.67	24	4.04	1.72	34	3.89	1.96	44	2.87	1.09
5	1.88	7.00	15	1.60	6.79	25	2.89	1.22	35	3.03	1.01	45	5.13	3.28
6	1.52	7.03	16	3.41	6.30	26	4.09	2.25	36	4.58	2.76	46	4.21	2.21
7	2.01	7.09	17	2.01	6.43	27	3.61	1.46	37	4.26	2.10	47	4.38	2.32
8	1.51	6.97	18	1.76	6.69	28	3.94	1.87	38	3.65	1.59	48	3.02	1.27
9	1.31	7.26	19	1.05	7.65	29	4.68	2.73	39	5.32	3.23	49	4.14	1.87
10	1.55	6.83	20	2.41	7.20	30	3.75	2.28	40	3.61	1.38	50	4.22	2.28

Tabla 2.3: Datos simulados de Rousseeuw.

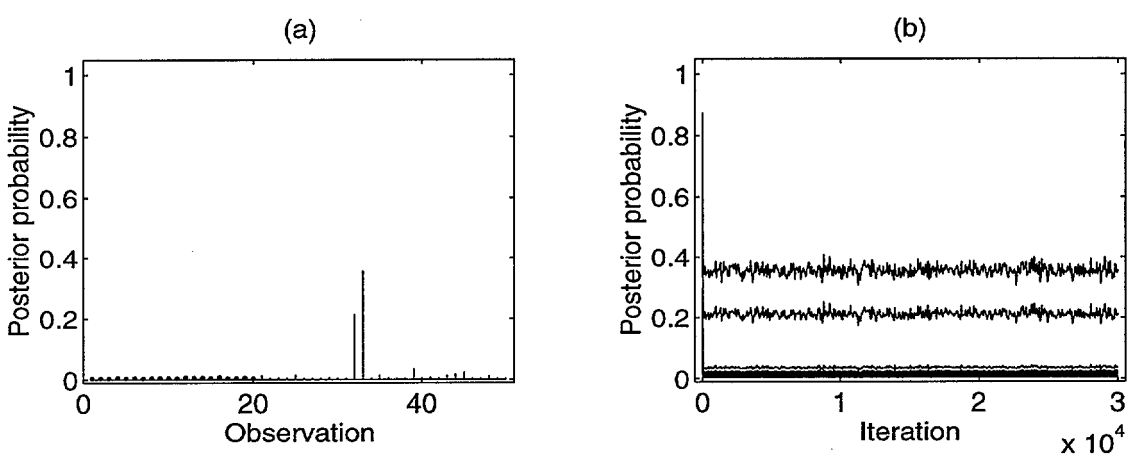

Figura 2.5: Resultados del Gibbs Sampling con los datos simulados de Rousseeuw: (a) probabilidades a posteriori de que cada dato sea atípico con 30000 iteraciones; (b) probabilidades a posteriori en función del número de iteración.

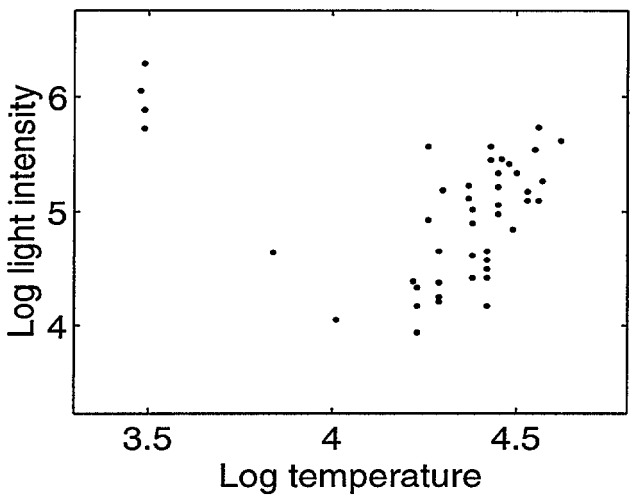

Figura 2.6: Diagrama de Hertzsprung-Russell de las estrellas del grupo CYG OB1.

de la temperatura efectiva en la superficie de la estrella y la variable dependiente es el logaritmo de la intensidad de la luz (el modelo incluye constante). Los datos los proporcionan Rousseeuw y Leroy (1987) y se recogen en la tabla 2.4. Observando la nube de puntos de la figura 2.6 se identifican cuatro valores atípicos en las posiciones 11, 20, 30 y 34, que corresponden a estrellas gigantes.

Este ejemplo sirve para ilustrar como los problemas de convergencia del Gibbs Sampling detectados en los ejemplos anteriores con datos artificiales, también se producen con datos reales. Se observa en la figura 2.7(a) y en la figura 2.7(b) que después de 10,000 iteraciones las estrellas gigantes no son identificadas como atípicas y, sin embargo, las series parece que convergen.

2.3 Análisis de la convergencia del Gibbs Sampling

Los ejemplos analizados en la sección anterior demuestran que la aplicación estándar del Gibbs Sampling puede ser un mal procedimiento para la identificación de valores

i	y_i	x_i	i	y_i	x_i	i	y_i	x_i	i	y_i	x_i	i	y_i	x_i
1	5.23	4.37	11	5.73	3.49	21	4.38	4.29	31	4.42	4.38	41	4.62	4.38
2	5.74	4.56	12	5.45	4.43	22	4.22	4.29	32	5.10	4.56	42	5.06	4.45
3	4.93	4.26	13	5.42	4.48	23	4.42	4.42	33	5.22	4.45	43	5.34	4.50
4	5.74	4.56	14	4.05	4.01	24	4.85	4.49	34	6.29	3.49	44	5.34	4.45
5	5.19	4.30	15	4.26	4.29	25	5.02	4.38	35	4.34	4.23	45	5.54	4.55
6	5.46	4.46	16	4.58	4.42	26	4.66	4.42	36	5.62	4.62	46	4.98	4.45
7	4.65	3.84	17	3.94	4.23	27	4.66	4.29	37	5.10	4.53	47	4.50	4.42
8	5.27	4.57	18	4.18	4.42	28	4.90	4.38	38	5.22	4.45			
9	5.57	4.26	19	4.18	4.23	29	4.39	4.22	39	5.18	4.53			
10	5.12	4.37	20	5.89	3.49	30	6.05	3.48	40	5.57	4.43			

Tabla 2.4: Datos del diagrama de Hertzsprung-Russell.

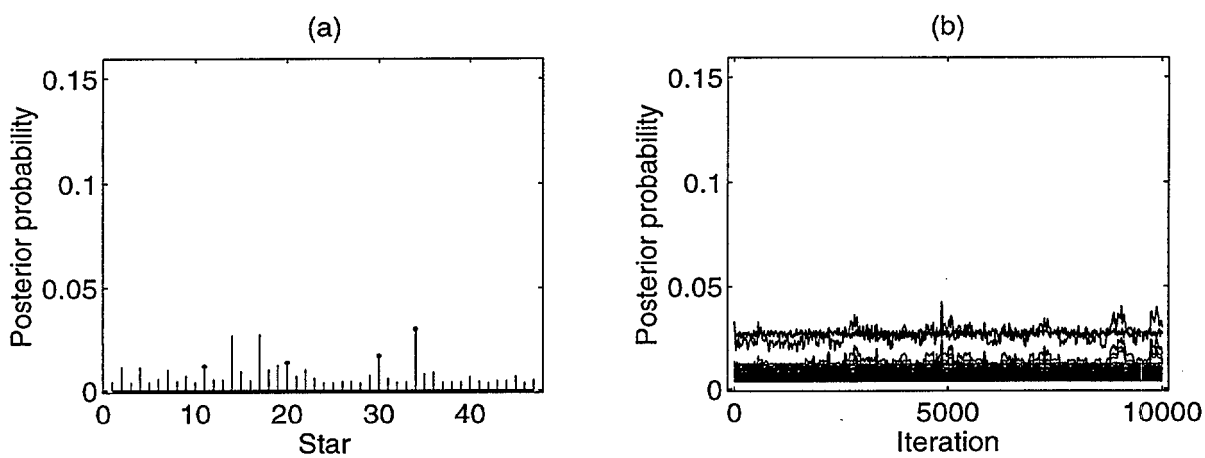

Figura 2.7: Resultados del Gibbs Sampling con los datos del diagrama de Hertzsprung-Russell: (a) probabilidades a posteriori de que cada dato sea atípico con 10000 iteraciones; (b) probabilidades a posteriori en función del número de iteración.

atípicos en ciertos conjuntos de datos: la estabilidad de las series de probabilidades de que cada observación sea atípica indican una aparente convergencia a distribuciones de los parámetros distintas de las verdaderas. En esta sección se demuestra que el comportamiento engañoso del Gibbs Sampling es debido al problema del enmascaramiento causado por la presencia de grupos de observaciones atípicas muy influyentes, y se estudia el efecto en la estimación de las probabilidades y de los parámetros del modelo.

Smith y Roberts (1993) sugieren que una dimensión elevada del espacio paramétrico y una correlación alta pueden provocar que la convergencia del Gibbs Sampling sea lenta. Estos dos problemas se presentan cuando existe enmascaramiento ya que las variables de clasificación de los atípicos que se enmascaran estarán muy correlacionadas y, en general, la dimensión del espacio paramétrico crece con el tamaño muestral; sin embargo, el problema es más grave que el mencionado por Smith y Roberts (1993). Por ejemplo, los datos de la figura 2.8 son una muestra de una mezcla de dos normales con un 30 por ciento de contaminación en la que estos dos problemas se presentan. Las probabilidades que se observan en la figura 2.9(a) y las series de probabilidades de la figura 2.9(b) muestran que la convergencia es lenta, como se esperaba, pero finalmente el algoritmo converge. Este no es el caso de los ejemplos con datos de regresión analizados en la sección anterior. La principal diferencia entre estas dos situaciones es el papel que juega el potencial en los modelos de regresión. Si datos atípicos con alto potencial (atípicos influyentes) y que provocan enmascaramiento se consideran inicialmente como buenos, la estimación de los coeficientes de la regresión estará sesgada, los residuos correspondientes a estos datos serán pequeños y la probabilidad de clasificarlos como atípicos en la siguiente iteración será muy baja.

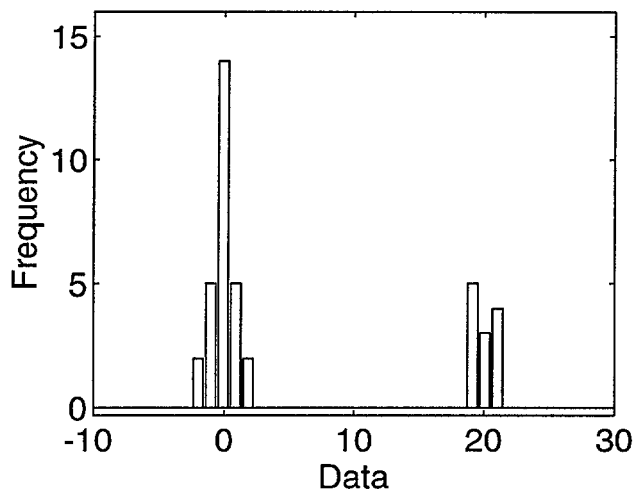

Figura 2.8: Histograma de una muestra de tamaño $n = 40$ de una mezcla de normales.

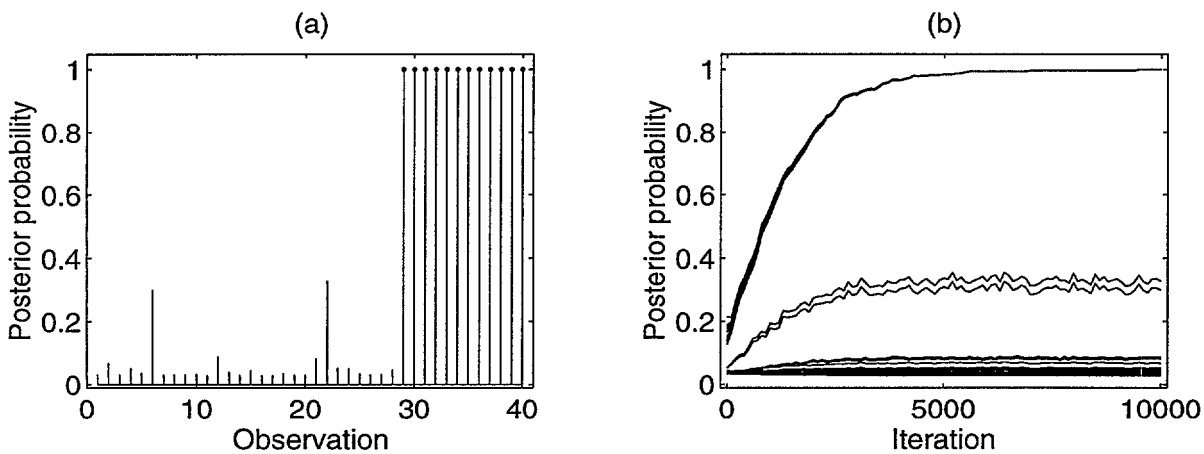

Figura 2.9: Resultados del Gibbs Sampling con los datos generados con una mezcla de normales: (a) probabilidades a posteriori de que cada dato sea atípico con 10000 iteraciones; (b) probabilidades a posteriori en función del número de iteración.

2.3.1 Estimación de las probabilidades

Sea $\boldsymbol{\delta}^{(0)}$ la configuración de atípicos con la que se inicia la ejecución del Gibbs Sampling, y sea $\boldsymbol{\beta}^{(0)}$ el estimador de mínimos cuadrados generalizados para la configuración $\boldsymbol{\delta}^{(0)}$. La probabilidad de que $\delta_i^{(1)} = 1$ en la primera iteración viene dada por (2.6), sustituyendo $\boldsymbol{\beta}$ por $\boldsymbol{\beta}^{(0)} = (\boldsymbol{X}'\boldsymbol{V}^{(0)^{-1}}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{V}^{(0)^{-1}}\boldsymbol{y}$, σ^2 por la varianza que se genera en la primera iteración y α por la contaminación generada también en la primera iteración. La probabilidad definida por (2.6) se puede expresar en la primera iteración como

$$P(\delta_i^{(1)} = 1 \mid \boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, \alpha) = \left(1 + \left(\frac{1-\alpha^{(1)}}{\alpha^{(1)}}\right) F_{10}^{(1)}(i)\right)^{-1}, \tag{2.9}$$

donde $F_{10}^{(1)}$ es el factor de Bayes, que para la observación i-ésima viene dado por

$$F_{10}(i) = k \cdot \exp\left(-\frac{1}{2\phi^{-1}\sigma^{2(1)}} u_i^{(0)2}\right),$$

donde $u_i^{(0)} = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}^{(0)}$ es el residuo inicial, y $\phi = 1 - k^{-2}$. La probabilidad (2.9) depende para cada observación del residuo $u_i^{(0)}$ ($\sigma^{(1)}$ y $\alpha^{(1)}$ son los mismos para todos los datos) y, para k suficientemente grande, será próxima a uno cuando $u_i^{(0)}$ sea grande y próxima a cero cuando $u_i^{(0)}$ tienda a cero.

Sea $\boldsymbol{S}_0 = (\boldsymbol{X}_0, \boldsymbol{y}_0)$ el conjunto de observaciones que se clasifican como buenas en las condiciones iniciales. Para k suficientemente grande, $\boldsymbol{V}^{(0)}$ es aproximadamente la matriz identidad y, por tanto, $u_i^{(0)}$ será el estimador de mínimos cuadrados de la submuestra $(\boldsymbol{X}_0, \boldsymbol{y}_0)$. Si $\boldsymbol{S}_0$ contiene observaciones atípicas con alto potencial, el vector inicial de coeficientes $\boldsymbol{\beta}^{(0)}$ estará sesgado y los residuos mínimo cuadráticos correspondientes a estos datos serán pequeños. Por tanto, la probabilidad de identificarlos como atípicos en la siguiente iteración será muy pequeña. Por el contrario, si estos atípicos señalan observaciones buenas como atípicas, estas tendrán un residuo grande y con probabilidad próxima a uno se identificaran como atípicas en la siguiente iteración. Este

hecho se puede ver fácilmente en el caso extremo en el que la muestra contiene un grupo I de n_I observaciones atípicas idénticas. Supongamos que el conjunto $\boldsymbol{S}_0 = (\boldsymbol{y}_0, \boldsymbol{X}_0)$ incluye todo el grupo de atípicos y que $h = \boldsymbol{x}_i'(\boldsymbol{X}_{0(I)}'\boldsymbol{X}_{0(I)})^{-1}\boldsymbol{x}_i$ es el *potencial fuera de la muestra* para $i \in I$, donde el subíndice (I) en $\boldsymbol{X}_{0(I)}$ significa que las filas indexadas por I han sido eliminadas de la matriz $\boldsymbol{X}_0$; del mismo modo en $\boldsymbol{y}_{0(I)}$ se eliminan las observaciones indexadas por I. Peña y Yohai (1995) prueban que $u_i^{(0)}$ se puede expresar como

$$u_i^{(0)} = \frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_{(I)}^{(0)}}{1 + n_I h} \qquad \text{para } i \in I, \tag{2.10}$$

donde, para valores suficientemente grandes de k, $\boldsymbol{\beta}_{(I)}^{(0)}$ se puede aproximar por el estimador de mínimos cuadrados cuando se eliminan de $\boldsymbol{S}_0$ las observaciones indexadas por I, $\hat{\boldsymbol{\beta}}_{0(I)} \simeq (\boldsymbol{X}_{0(I)}'\boldsymbol{X}_{0(I)})^{-1}\boldsymbol{X}_{0(I)}'\boldsymbol{y}_{0(I)}$. A partir de la ecuación (2.10) es inmediato ver que el residuo $u_i^{(0)}$ será pequeño si el potencial h es elevado (nótese que h no está acotado) y su efecto crece cuando aumenta el número de atípicos n_I. Por tanto, para los valores atípicos con alto potencial los residuos $u_i^{(0)}$ serán muy pequeños y la probabilidad (2.9) muy próxima a cero. Por el contrario, si el conjunto $\boldsymbol{S}_0$ no contiene valores atípicos, los residuos $u_i^{(0)} = \boldsymbol{y}_i - \boldsymbol{x}_i'\boldsymbol{\beta}^{(0)}$ para $i \in I$, que son residuos fuera de la muestra, tenderán a ser grandes y la probabilidad (2.9) próxima a uno. La única posibilidad de detectar los atípicos en la siguiente iteración es cuando todos ellos son clasificados como atípicos al generar los valores de la distribución condicionada del vector $\boldsymbol{\delta}$. Por ejemplo, si tenemos una muestra con 10 datos atípicos y la probabilidad (2.9) es 0.01, entonces la probabilidad de identificar todos los atípicos es 10^{-20}.

Este error se reproduce en las siguientes iteraciones, produciéndose una aparente convergencia a las probabilidades que se obtendrían si la verdadera recta de regresión es la que incluye los datos atípicos. La solución a este problema se discute en el siguiente capítulo y depende de una asignación correcta en las condiciones iniciales de las variables de clasificación del grupo de atípicos enmascarados. Los gráficos de la

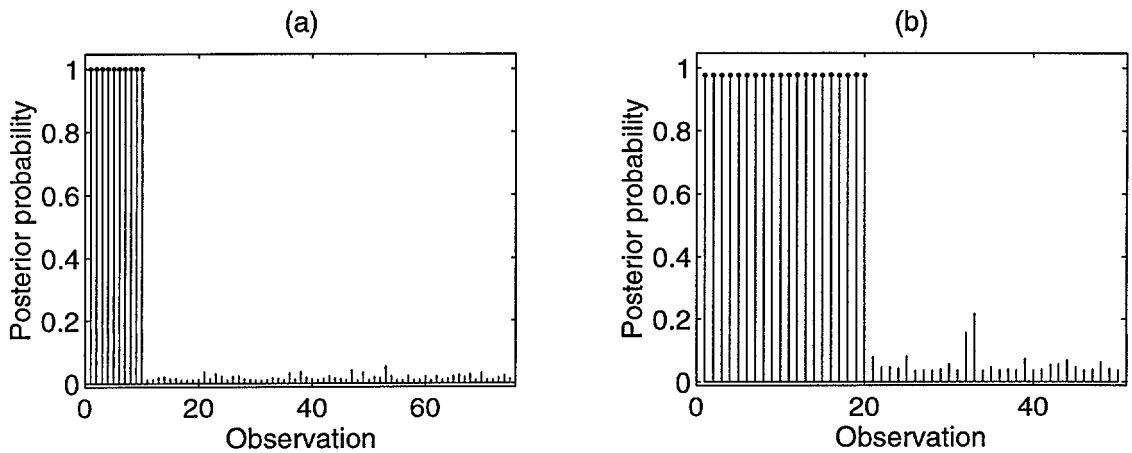

Figura 2.10: Probabilidades a posteriori después de 200 iteraciones cuando los atípicos se asignan inicialmente a la distribución contaminante: (a) datos de Hawkins, Bradu y Kass; (b) datos simulados de Rousseeuw.

figura 2.10 muestran las probabilidades de que cada dato sea atípico en los ejemplos 2 y 3 cuando se considera inicialmente que al menos los atípicos han sido generados por la distribución contaminante. El número de iteraciones necesarias para alcanzar la convergencia es pequeño y las probabilidades se han estimado con la muestra obtenida después de 200 iteraciones.

2.3.2 Estimación de los parámetros

Como consecuencia de la sección anterior es razonable asumir la siguiente *propiedad de dependencia inicial:*

i) Si el conjunto $\boldsymbol{S}_0$ no contiene datos atípicos, los atípicos que existan en la muestra siempre se identifican y los datos buenos nunca son señalados como atípicos;

ii) Si el conjunto $\boldsymbol{S}_0$ contiene datos atípicos influyentes, la probabilidad de identificar todos los atípicos de la muestra es pequeña y será muy próxima a cero si el número

de valores atípicos en $\boldsymbol{S}_0$ es elevado.

Cuando las series de probabilidades se estabilizan después de pocas iteraciones, la propiedad de dependencia inicial implica que si $\boldsymbol{S}_0$ contiene uno o más valores atípicos con alto potencial, estos atípicos no serán identificados. Supongamos que las series se estabilizan a partir de la iteración S-ésima y que q es la probabilidad de que no haya datos atípicos en $\boldsymbol{S}_0$. Si consideramos el caso particular de una muestra que contiene sólo un grupo de n_{I} datos atípicos (no necesariamente idénticos) que se enmascaran, entonces la densidad marginal a posteriori de $\boldsymbol{\beta}^{(\mathrm{S})}$ se puede escribir como

$$P(\boldsymbol{\beta}^{(\mathrm{S})} \mid \boldsymbol{y}) = w_{\mathrm{I}} \cdot P(\boldsymbol{\beta}^{(\mathrm{S})} \mid \boldsymbol{y}, \boldsymbol{\delta}_{\mathrm{I}}) + \sum_{\mathrm{J} \neq \mathrm{I}} w_{\mathrm{J}} \cdot P(\boldsymbol{\beta}^{(\mathrm{S})} \mid \boldsymbol{y}, \boldsymbol{\delta}_{\mathrm{J}}), \tag{2.11}$$

donde $\boldsymbol{\delta}_{\mathrm{J}}$ es un vector con componentes iguales a uno para los datos del conjunto J y cero en los restantes. Los pesos w_{J} son iguales a la probabilidad marginal a posteriori $P(\boldsymbol{\delta}^{(\mathrm{S})} = \boldsymbol{\delta}_{\mathrm{J}} \mid \boldsymbol{y})$. Esta distribución es una media ponderada de 2^n distribuciones t multivariantes con $\nu = n - p$ grados de libertad y vectores de medias $\hat{\boldsymbol{\beta}}^{(\mathrm{S})}_{(\mathrm{J})}$, dados por

$$\hat{\boldsymbol{\beta}}^{(\mathrm{S})}_{(\mathrm{J})} = \hat{\boldsymbol{\beta}} - \phi(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'_{\mathrm{J}}(\boldsymbol{I} - \phi\boldsymbol{X}_{\mathrm{J}}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'_{\mathrm{J}})^{-1}(\boldsymbol{y}_{\mathrm{J}} - \boldsymbol{X}_{\mathrm{J}}\hat{\boldsymbol{\beta}}), \tag{2.12}$$

donde $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ es el estimador de mínimos cuadrados y $\phi = 1 - k^2$. Las matrices de dispersión son $\hat{\boldsymbol{\Omega}}^{(\mathrm{S})}_{(\mathrm{J})} = s^2_{(\mathrm{J})}(\boldsymbol{X}'\boldsymbol{X} - \phi\boldsymbol{X}'_{\mathrm{J}}\boldsymbol{X}_{\mathrm{J}})^{-1}$, donde $s^2_{(\mathrm{J})}$ se puede expresar como

$$\nu s^2_{(\mathrm{J})} = \nu s^2 - \phi(\boldsymbol{y}_{\mathrm{J}} - \boldsymbol{X}_{\mathrm{J}}\hat{\boldsymbol{\beta}})'(\boldsymbol{I} - \phi\boldsymbol{X}_{\mathrm{J}}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'_{\mathrm{J}})^{-1}(\boldsymbol{y}_{\mathrm{J}} - \boldsymbol{X}_{\mathrm{J}}\hat{\boldsymbol{\beta}})$$

y s^2 es la varianza muestral.

Con las condiciones asumidas en la propiedad de dependencia inicial se deducen las conclusiones siguientes:

(1) El peso $w_{\mathrm{I}} = P(\boldsymbol{\delta}^{(\mathrm{S})} = \boldsymbol{\delta}_{\mathrm{I}} \mid \boldsymbol{y})$ es igual a q, la probabilidad de que $\boldsymbol{S}_0$ no contenga datos atípicos.

(2) Los pesos $w_J = P(\boldsymbol{\delta}^{(S)} = \boldsymbol{\delta}_J \mid \boldsymbol{y})$ son iguales a cero para todos los conjuntos J tales que $I \subset J$, esto se debe a que los datos buenos no son erróneamente señalados.

Por tanto, la densidad a posteriori $\boldsymbol{\beta}^{(S)}$ dada por la ecuación (2.11) se reduce a la expresión

$$P(\boldsymbol{\beta}^{(S)} \mid \boldsymbol{y}) = w_I \cdot P(\boldsymbol{\beta}^{(S)} \mid \boldsymbol{y}, \boldsymbol{\delta}_I) + \sum_{J \in \Omega_I} w_J \cdot P(\boldsymbol{\beta}^{(S)} \mid \boldsymbol{y}, \boldsymbol{\delta}_J),$$

donde $\Omega_I = \{J \mid J \cap I \neq I\}$.

Para k suficientemente grande y asumiendo que $\phi \approx 1$, se deduce de la ecuación (2.12) que el vector de medias $\hat{\boldsymbol{\beta}}^{(S)}_{(I)}$ es el estimador de mínimos cuadrados que se obtiene cuando se elimina de la muestra el grupo de datos atípicos I (véase Cook y Weisberg, 1982); $\hat{\boldsymbol{\beta}}^{(S)}_{(I)}$ es un estimador insesgado de $\boldsymbol{\beta}$. Para $J \neq I$, el vector de medias $\hat{\boldsymbol{\beta}}^{(S)}_{(J)}$ es un estimador sesgado de $\boldsymbol{\beta}$ debido a que J contiene datos atípicos, y la distribución $P(\boldsymbol{\beta}^{(S)} \mid \boldsymbol{y}, \boldsymbol{\delta}_J)$ se puede colapsar a una distribución t multivariante con media el estimador de mínimos cuadrados para la muestra completa (incluyendo los atípicos). La distribución a posteriori de $\boldsymbol{\beta}^{(S)}$ se puede considerar como una mixtura de dos distribuciones t multivariantes con vectores de medias $\hat{\boldsymbol{\beta}}_{(I)}$ y $\hat{\boldsymbol{\beta}}$, y con pesos q y $1-q$, respectivamente. En particular, cuando sólo existe una observación atípica $(y_i, \boldsymbol{x}_i')$ en la muestra, la probabilidad q puede ser muy próxima a uno y la distribución de $\boldsymbol{\beta}^{(S)}$ es prácticamente unimodal, una t multivariante con media $\hat{\boldsymbol{\beta}}_{(i)}$. En general, es posible que exista más de un grupo de observaciones atípicas y la distribución de $\boldsymbol{\beta}^{(S)}$ dado $\boldsymbol{y}$ podría ser multimodal.

La bimodalidad que se obtiene en la estimación de la distribución de $\boldsymbol{\beta}$ cuando el Gibbs Sampling no converge se puede ver claramente en el ejemplo de los datos artificiales de Hawkins, Bradu y Kass (1984). En la figura 2.11 se representa con una curva de trazo discontinuo el estimador tipo núcleo de la densidad del coeficiente β_1 de la regresión que se obtiene cuando se inicia el Gibbs Sampling con un conjunto $\boldsymbol{S}_0$ de

cuatro observaciones elegidas al azar y consideradas como buenas. La probabilidad de que entre las cuatro observaciones ninguna sea atípica, o lo que es lo mismo, ninguna este mal clasificada es

$$q = \binom{10}{0}\binom{65}{4} \Big/ \binom{75}{4} = 0.557. \tag{2.13}$$

La distribución que se obtiene es bimodal; con una moda en $\beta_1 = -0.9297$ aproximadamente, que es el estimador de mínimos cuadrados cuando se eliminan los cuatro datos que son señalados como atípicos; y la otra moda en $\beta_1 = -0.1452$ aproximadamente, que es el estimador de mínimos cuadrados cuando se eliminan los datos atípicos. Nótese que, como cabría esperar, los pesos de ambas distribuciones están próximos a 0.5 y la distribución que se obtiene cuando los datos atípicos son eliminados tiene menos dispersión. La curva con trazo continuo representa la densidad a posteriori cuando se ejecuta el Gibbs Sampling con las condiciones iniciales ideales para evitar el enmascaramiento, es decir, cuando todos las observaciones atípicas son consideradas como tales en las condiciones iniciales.

2.4 Modelo bayesiano semiparamétrico

En esta sección se discute si los fallos en la convergencia del Gibbs Sampling detectados en los ejemplos pueden atribuirse al modelo particular que se propone para la identificación de los atípicos. Como los métodos de detección de observaciones atípicas con el modelo de contaminación de escala, el modelo de contaminación en la media y el modelo que no supone distribución alternativa para los atípicos son equivalentes (véase Peña y Guttman, 1993), se considera un modelo más general que contempla contaminación tanto en la media como en la varianza.

Supondremos que las observaciones son generadas por el modelo (2.1) y la dis-

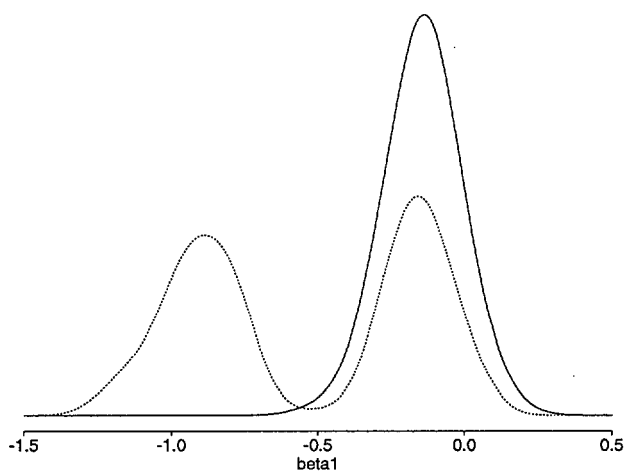

Figura 2.11: La curva con trazo continuo es la densidad a posteriori del parámetro β_1 del modelo de regresión para los datos de Hawkins, Bradu y Kass; la curva con trazo discontinuo es el estimador tipo núcleo de la densidad de β_1 construido con la muestra que se obtiene después de 500 iteraciones del Gibbs Sampling.

tribución de los errores es

$$u_i \sim (1-\alpha)\, N(0,\sigma^2) + \alpha\, N(h_i, \sigma^2\tau_i^2) \qquad i = 1,\ldots,n. \tag{2.14}$$

Como se dispone de una única observación para estimar la media y la varianza de la distribución contaminante asociada a cada dato, este modelo sólo será identificable cuando algunos parámetros sean iguales para distintas observaciones. Esto implica que la distribución de los pares $\theta_i = (h_i, \tau_i^2)$ debe ser discreta. Por tanto, para completar la estructura a priori se considera únicamente esta restricción y se supone que $\theta_1,\ldots,\theta_n$ son independientes y su distribución bivariante es desconocida.

Las distribuciones a priori que se describen a continuación junto con (2.1) y (2.14) definen un modelo bayesiano jerárquico que es semiparamétrico. Por un lado, la mayoría de las distribuciones a priori pertenecen a familias paramétricas conocidas y, por otro lado, la distribución de θ_i es desconocida por lo que el espacio paramétrico incluye una colección $\mathcal{F}$ de distribuciones de probabilidad. Las distribuciones necesarias para completar la estructura a priori son:

$$\begin{aligned}
\theta_i &\sim G \\
G &\sim DP\,(\mu G_0) \\
G_0 &\sim N(m,b) \times Gamma - Invertida\,(u/2, v/2) \\
\mu &\sim Gamma\,(a_0, b_0),
\end{aligned}$$

donde G es una distribución bivariante desconocida y DP es un proceso de Dirichlet con parámetro μG_0. La notación G se usa tanto para indicar una medida de probabilidad como una función de distribución.

La idea de suponer que a priori la distribución G es una realización aleatoria de un proceso de Dirichlet fue propuesta por Ferguson (1973), que define formalmente los procesos de Dirichlet como:

DEFINICIÓN 1 *Sea $\mathcal{A}$ una σ-álgebra de subconjuntos de Θ. Sea γ una medida finita, positiva y finitamente aditiva, en $(\Theta, \mathcal{A})$. Se dice que una medida de probabilidad aleatoria G en $(\Theta, \mathcal{A})$ es un proceso de Dirichlet con parámetro γ si para cualquier $k = 1, 2, \ldots$ y para cualquier partición medible $A_1, \ldots, A_k$ de Θ, la distribución conjunta del vector de probabilidades $(G(A_1), \ldots, G(A_k))$ es la distribución de Dirichlet con parámetros $(\gamma(A_1), \ldots, \gamma(A_k))$.*

El parámetro γ del proceso no es necesariamente una medida de probabilidad. Sin embargo, en el contexto en que estamos trabajando, el espacio paramétrico $\Theta = \mathbb{R} \times \mathbb{R}^+$ con la σ-álgebra de Borel, es frecuente encontrar en la literatura la descomposición $\gamma = \mu G_0$, donde μ es la medida del espacio total, $\mu = \gamma(\Theta)$, y la distribución $G_0(\cdot) = \gamma(\cdot)/\gamma(\Theta)$ representa la forma de la medida.

La importancia de los procesos de Dirichlet radica en que es una familia de distribuciones a priori conjugada cuando el parámetro desconocido es una medida de probabilidad. Si G se distribuye a priori como un proceso de Dirichlet con parámetro μG_0 y $\theta_1, \ldots, \theta_n$ es una muestra de G, entonces la distribución de $G \mid \theta_1, \ldots, \theta_n$ es también un proceso de Dirichlet con parámetro $\mu G_0 + \sum_{j=1}^{n} I_{\theta_j}$. En particular, suponer que la distribución a priori de G en el modelo (2.1) y (2.14) es un proceso de Dirichlet tiene dos ventajas importantes:

1. Las trayectorias del proceso son medidas de probabilidad discretas con probabilidad uno (véase Ferguson, 1973 y Blackwell, 1973).
2. Dada una muestra $y_1, \ldots, y_n$ de variables aleatorias i.i.d. con distribución F_θ, donde se supone que la distribución a priori G del parámetro θ es una trayectoria de un proceso de Dirichlet, entonces Antoniak (1974) prueba que la distribución a posteriori $G \mid y_1, \ldots, y_n$ es una mezcla de procesos de Dirichlet (MDP).

2.4.1 Distribuciones a posteriori

La distribución conjunta de los parámetros del modelo (2.1) y (2.14) es

$$p(\theta_1,\ldots,\theta_n \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \theta_1,\ldots,\theta_n)\, p(\theta_1,\ldots,\theta_n) = \prod_{i=1}^{n} p(y_i \mid \theta_i)\, p(\theta_i \mid \theta_1,\ldots,\theta_{i-1}), \tag{2.15}$$

y se obtiene aplicando una propiedad de "memoria" que verifican los procesos de Dirichlet: dados $\theta_1,\ldots,\theta_i$, la distribución de un nuevo θ_{i+1} es G_0 con probabilidad $q_{i+1} = \mu/(\mu+i)$ y es degenerada en alguno de los $\theta_1,\ldots,\theta_i$ con probabilidad $1-q_{i+1}$. Este resultado se demuestra en la siguiente proposición.

PROPOSICIÓN 1 *Sea $\theta \in \Theta$ una variable aleatoria con distribución G, donde G se distribuye según un proceso de Dirichlet con parámetro $\gamma = \mu G_0$. Sean $\theta_1,\ldots,\theta_n$ variables aleatorias i.i.d. con distribución G, entonces*

$$\theta \mid \theta_1,\ldots,\theta_n \sim \frac{\mu}{\mu+n} G_0 + \frac{1}{\mu+n}\sum_{j=1}^{n} I_{\theta_j}, \tag{2.16}$$

donde I_{θ_j} es la medida de probabilidad que otorga probabilidad 1 a θ_j.

Demostración. Para probar este resultado se utiliza el teorema 1 (pag. 217) y las proposiciones 1 (pag. 214) y 4 (pag. 216) de Ferguson (1973). Por el teorema 1, la distribución condicionada $G \mid \theta_1,\ldots,\theta_n$ sigue también un proceso de Dirichlet con parámetro $\mu G_0 + \sum_{j=1}^{n} I_{\theta_j}$. Entre las propiedades de los procesos de Dirichlet enunciadas en la proposición 1 se tiene que si F_n es la función de distribución empírica de $\theta_1,\ldots,\theta_n$, entonces

$$E(G(\theta) \mid \theta_1,\ldots,\theta_n) = \frac{\mu G_0(\theta) + nF_n(\theta)}{lim_{\theta\to+\infty}(\mu G_0(\theta) + nF_n(\theta))}$$

y, por la proposición 4, la distribución a priori condicionada es

$$\theta \mid \theta_1,\ldots,\theta_n \sim E(G \mid \theta_1,\ldots,\theta_n) = \frac{\mu}{\mu+n} G_0 + \frac{1}{\mu+n}\sum_{j=1}^{n} I_{\theta_j}. \qquad \square$$

Debido a que G es una distribución discreta, existe una probabilidad distinta de cero de obtener valores idénticos de los parámetros para distintas observaciones. El número de parámetros distintos se reduce a un cierto $k \leq n$ que se denota por $\theta^* = (\theta_1^*, \ldots, \theta_k^*)$, siendo n_j el número de parámetros iguales a θ_j^*. En consecuencia, la ecuación (2.16) se puede expresar como

$$\theta \mid \theta_1, \ldots, \theta_n \sim \frac{\mu}{\mu + n} G_0 + \frac{1}{\mu + n} \sum_{j=1}^{k} n_j I_{\theta_j^*}.$$

Por la proposición 1 y sustituyendo en la ecuación (2.15), se tiene que la distribución a posteriori de los parámetros es

$$p(\theta_1, \ldots, \theta_n \mid \boldsymbol{y}) \propto \prod_{i=1}^{n} f(y_i \mid \theta_i) \, \frac{\mu g_0(\theta_i) + \sum_{j<i} I_{\theta_j}^*(\theta_i)}{\mu + i - 1}, \tag{2.17}$$

donde g_0 es la función de masa o la densidad de la distribución G_0. La obtención de la distribución (2.17) es complicada si se pretende hacer directamente.

2.4.2 Distribuciones a posteriori con Gibbs Sampling

Escobar (1994) propone utilizar el Gibbs Sampling para obtener una muestra de la distribución a posteriori a partir de las condicionadas en problemas con distribuciones a priori que son procesos de Dirichlet, y prueba que

$$\theta_i \mid \boldsymbol{y}, \theta_{(i)} \sim \pi_{n+1} G_i + \sum_{j \neq i} \pi_j \, I_{(\theta_i = \theta_j)}, \tag{2.18}$$

donde $\theta_{(i)} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$ y $\pi_{n+1} + \sum_{j \neq i} \pi_j = 1$. La ecuación (2.18) significa que en cada iteración del Gibbs Sampling el parámetro θ_i es igual a uno de los valores de $\theta_{(i)}$ con probabilidad π_j, mientras que con probabilidad π_{n+1} es un nuevo valor generado a partir de la distribución a posteriori de θ_i dado el dato y_i y la distribución

a priori G_0, $G_i(\theta_i) \propto f(y_i \mid \theta_i)g_0(\theta_i)$. Los pesos π_i son proporcionales a

$$\begin{aligned} \pi_j &\propto f(y_i \mid \theta_j) \qquad j = 1, \ldots, n \\ \pi_{n+1} &\propto \int f(y_i \mid \theta)\, dG_0(\theta) \end{aligned}$$

Esta idea es utilizada por Escobar y West (1995) para la estimación no-paramétrica de densidades cuando la distribución de G_0 es la conjugada normal/gamma-invertida, en este caso integrar y simular de G_i no es muy complicado. El problema surge cuando las distribuciones $f(y \mid \theta)$ y G_0 no son conjugadas como sucede en el modelo (2.1) y (2.14), donde son conjugadas condicionalmente. Para facilitar la aplicación del Gibbs Sampling Müller, Erkanli y West (1992) proponen utilizar la modificación introducida por MacEachern (1994) y MacEachern y Müller (1994), que además se comporta mejor en términos de velocidad de convergencia que el esquema habitual. El vector de parámetros se aumenta con n indicadores de grupo $\boldsymbol{s} = (s_1, \ldots, s_n)'$, que toman el valor $s_i = s_{i'} = j$ si y sólo si $\theta_i = \theta_{i'} = \theta_j^\star$.

Las distribuciones condicionadas de los parámetros del modelo (2.1) y (2.14) se detallan a continuación. Se han eliminado en cada caso aquellos parámetros que no afectan a la determinación de la distribución.

1. La distribución condicionada de $\boldsymbol{\beta}$ es

$$\boldsymbol{\beta} \mid \boldsymbol{y}, \sigma^2, \boldsymbol{\delta}, \boldsymbol{s}, \boldsymbol{\theta}^\star \sim N_p(\hat{\boldsymbol{\beta}}_s, \sigma^2(\boldsymbol{X}'\boldsymbol{V}_s{}^{-1}\boldsymbol{X})^{-1})),$$

donde $\hat{\boldsymbol{\beta}}_s = (\boldsymbol{X}'\boldsymbol{V}_s{}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}_s{}^{-1}(\boldsymbol{y} - \boldsymbol{H}_s)$, $\boldsymbol{H}_s = (\delta_1 h^\star_{s_1}, \ldots, \delta_n h^\star_{s_n})'$ y $\boldsymbol{V}_s$ es una matriz diagonal con elementos en la diagonal principal $v_{ii} = 1 + \delta_i(\tau^{2\star}_{s_i} - 1)$.

2. La distribución condicionada de σ^2 es

$$\sigma^2 \mid \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{s}, \boldsymbol{\theta}^\star \sim Gamma - Invertida\ (n/2, \sigma_s^2/2),$$

donde $\sigma_s^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{H}_s)'\boldsymbol{V}_s{}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{H}_s)$.

3. La distribución condicionada de δ_i es una *Bernoulli* con probabilidad de éxito

$$P(\delta_i = 1 \mid \boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{s}, \theta^\star) = \frac{\alpha f_{\mathrm{N}}((u_i - h^\star_{s_i})/\sigma\tau^\star_{s_i})}{\alpha f_{\mathrm{N}}((u_i - h^\star_{s_i})/\sigma\tau^\star_{s_i}) + (1-\alpha)\tau^\star_{s_i} f_{\mathrm{N}}(u_i/\sigma)}.$$

4. La distribución condicionada de α es

$$\alpha \mid \boldsymbol{\delta} \sim \mathit{Beta}\,(\gamma_1 + n\bar{\delta}, \gamma_2 + n(1-\bar{\delta})),$$

donde $n\bar{\delta} = \sum_{i=1}^{n} \delta_i$.

5. Sea $\boldsymbol{s}_{(i)}$ el vector $\boldsymbol{s}$ cuando se elimina s_i, y sea n_{ij} el número de indicadores de grupo en $\boldsymbol{s}_{(i)}$ iguales a j. Entonces el número de indicadores diferentes es

$$k_{(i)} = \begin{cases} k-1 & \text{si } s_i \neq s_j \quad \text{y } j \neq i \\ k & \text{en caso contrario.} \end{cases}$$

Para calcular $\pi_{i,j} = P(s_i = j \mid \boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}, \boldsymbol{s}_{(i)}, \boldsymbol{\theta}^\star, \mu)$ se consideran dos casos:

(i) Si $\delta_i = 1$, la probabilidad $\pi_{i,j}$ es

$$\pi_{i,j} = \begin{cases} C\, n_{ij}\, \tau^\star_j\, f_{\mathrm{N}}((u_i - h^\star_j)/\sigma\tau^\star_j) & \text{para } j = 1, \ldots, k_{(i)} \\ C\, \mu\, \tau^\star_{s_i}\, f_{\mathrm{N}}((u_i - h^\star_{s_i})/\sigma\tau^\star_{s_i}) & \text{para } j = k_{(i)} + 1, \end{cases}$$

donde $C = (\pi_{i,k_{(i)}+1} + \sum_{j \neq i} \pi_{i,j})^{-1}$. Nótese que $\pi_{i,k_{(i)}+1}$ es proporcional a $\int f(y_i \mid \theta) dG_0(\theta)$. Esta integral puede aproximarse por Monte Carlo o para simplificar los cálculos se ha aproximado por la densidad de una $N(\boldsymbol{x}_i'\boldsymbol{\beta} - h^\star_{s_i}, \sigma^2\tau^{2\star}_{s_i})$ (véase Müller, Erkanli y West, 1992).

(ii) Si $\delta_i = 0$, la probabilidad $\pi_{i,j}$ es

$$\pi_{i,j} = \begin{cases} n_{ij}/(\mu + n - 1) & \text{para } j = 1, \ldots, k_{(i)} \\ \mu/(\mu + n - 1) & \text{para } j = k_{(i)} + 1. \end{cases}$$

6. Para $j = 1, \ldots, k$, se definen los conjuntos $I_j^\star = \{i \mid \delta_i = 1 \text{ y } s_i = j\}$ y se denota por $n_j^\star$ el tamaño de $I_j^\star$. Las distribuciones condicionadas de $h_j^\star$ and $\tau_j^{2\star}$ son:

$$\begin{aligned} h_j^\star \mid \boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}, \boldsymbol{s}, \tau_j^{2\star} &\sim N(m_j, b_j) \\ \tau_j^{2\star} \mid \boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}, \boldsymbol{s}, h_j^\star &\sim Gamma - Invertida\left(\frac{n_j^\star + u}{2}, \frac{v + v_j}{2}\right), \end{aligned}$$

donde $b_j = (b^{-2} + \tau_j^{-2\star}\sigma^{-2}n_j^\star)^{-1}$, $m_j = b_j\left(b^{-2}m + \tau_j^{-2\star}\sigma^{-2}\sum_{i\in I_j^\star} u_i\right)$ y $v_j = \sigma^{-2}\sum_{i\in I_j^\star}(u_i - h_j^\star)^2$.

7. La distribución condicionada de μ se obtiene aumentando el vector de parámetros con una nueva variable artificial η (véase Escobar y West, 1995). Las distribuciones condicionadas son

$$\eta \mid \boldsymbol{y}, \mu \sim Beta(\mu + 1, n)$$

$$\mu \mid \boldsymbol{y}, \boldsymbol{s}, \eta \sim \pi\, Gamma(a_1, b_1) + (1 - \pi)\, Gamma(a_1 - 1, b_1),$$

donde $\pi = (a_1 - 1)/(a_1 - 1 + nb_1)$, $a_1 = a_0 + k$ y $b_1 = b_0 - log(\eta)$.

Al pertenecer todas las distribuciones a posteriori condicionadas a familias paramétricas conocidas, y para las que existen diversos métodos de simulación, la aplicación del Gibbs Sampling es inmediata. El esquema de funcionamiento del Gibbs Sampling es igual al que siguen Müller, Erkanli y West (1992) en la estimación de la función de regresión, con las correspondientes modificaciones que se derivan de las diferencias entre su modelo y el modelo (2.1) y (2.14). Los pasos que se siguen son:

1. Se seleccionan valores iniciales de los parámetros: $\alpha^{(0)} = 0.2$; para cada i la variable de clasificación $\delta_i^{(0)} = 1$ con probabilidad $\alpha^{(0)}$; $h_i^{(0)} \sim N(m, b)$ y $\tau_i^{2(0)} \sim Gamma - Invertida(u/2, v/2)$; $\boldsymbol{\beta}^{(0)} = \hat{\boldsymbol{\beta}}_s^{(0)}$; $\sigma^{2(0)} = \sigma_s^{2(0)}$ y $\eta^{(0)} \sim Unif(0, 1)$. Así, $k^{(0)} = n$, y para $i = 1, \ldots, n$, $s_i^{(0)} = i$, $h_i^{\star(0)} = h_i^{(0)}$ y $\tau_i^{2\star(0)} = \tau_i^{2(0)}$.
2. Para $i = 1, \ldots, n$ se calcula la probabilidad de que $s_i = j$. Si algún $s_i = k_{(i)} + 1$ se genera un nuevo $\theta_{k_{(i)}+1}$. Implícitamente, los nuevos indicadores determinan el número k de componentes en cada iteración.

3. Se obtiene un nuevo vector $\boldsymbol{\theta}^\star = (\theta_1^\star, \ldots, \theta_k^\star)$ simulando los correspondientes $h_j^\star$ y $\tau_j^{2\star}$, para $j = 1, \ldots, k$.
4. Se clasifican las observaciones según el componente de la mixtura al que pertenecen con mayor probabilidad y se obtiene un nuevo $\boldsymbol{\delta}$.
5. Se obtienen nuevos valores del parámetro de contaminación, α, de los parámetros de la regresión, $\boldsymbol{\beta}$ y σ^2, y del parámetro del proceso de Dirichlet, μ.
6. Se repiten los pasos 2–5 hasta que el algoritmo converge.

El resultado que se obtiene al aplicar el Gibbs Sampling para la resolución del modelo semiparamétrico con los datos de los ejemplos analizados en este capítulo son los mismos que se obtienen con el modelo de contaminación de escala de Box y Tiao (1968). La presencia de observaciones atípicas que son muy influyentes provoca nuevamente que el algoritmo no converja en un número razonable de iteraciones. Podemos concluir que el fallo en la convergencia del Gibbs Sampling es debido al problema del potencial y no al modelo particular de generación de atípicos que se suponga.

Capítulo 3

Algoritmo adaptativo para identificar datos atípicos en regresión

3.1 Introducción

Como se vio en el capítulo 2, la aplicación del Gibbs Sampling a la detección de observaciones atípicas produce buenos resultados cuando se trata de identificar datos atípicos aislados, sin embargo, puede conducir a conclusiones erróneas cuando alguno de los datos atípicos está enmascarado. Si la muestra contiene un conjunto de observaciones atípicas que producen enmascaramiento, y el conjunto inicial $\boldsymbol{S}_0$ incluye alguno de estos datos, el algoritmo puede fallar. Por tanto, un objetivo claro es tratar de iniciar el algoritmo con un conjunto $\boldsymbol{S}_0$ que no contenga observaciones atípicas. Esta idea es similar a la que se emplea en los procedimientos de estimación robusta basados en remuestreo (Rousseeuw, 1984, y Hawkins, Bradu y Kass, 1984).

La única información inicial disponible para construir el conjunto $\boldsymbol{S}_0$ es que, por definición, las observaciones atípicas constituyen una fracción pequeña de los datos. Cuando se ejecuta el algoritmo y las series de probabilidades se estabilizan se obtiene

cierta información sobre la dependencia entre las variables de clasificación. Basándonos en este hecho proponemos un algoritmo adaptativo en el que las condiciones iniciales del Gibbs Sampling se van adaptando en un procedimiento que se realiza en dos etapas. En la primera se inicia el Gibbs Sampling con un conjunto pequeño de observaciones consideradas inicialmente como buenas que se depuran mediante un test de valores atípicos para eliminar algún posible dato atípico aislado. A continuación se ejecuta el algoritmo hasta que las series de probabilidades se estabilizan. El análisis de la estructura de dependencia entre las variables de clasificación, que se estima con la muestra obtenida en la última iteración, permite dividir el conjunto de datos en dos grupos que se utilizan para fijar los valores iniciales del Gibbs Sampling en la segunda etapa. El algoritmo que resulta de aplicar este proceso de aprendizaje y adaptación del Gibbs Sampling converge en pocas iteraciones a la distribución a posteriori de los parámetros.

Este capítulo se organiza del siguiente modo. En la sección 3.2 se presenta el nuevo algoritmo adaptativo de Gibbs Sampling para la identificación de atípicos en modelos de regresión. En la sección 3.3 se discute el comportamiento del algoritmo en conjuntos de datos reales y simulados que están contaminados por observaciones atípicas enmascaradas. También se compara el procedimiento con los métodos de detección de atípicos de Hadi y Simonoff (1993) y Peña y Yohai (1995).

3.2 Procedimiento para evitar el enmascaramiento

3.2.1 Valores iniciales en la primera etapa

El algoritmo se inicia en la primera etapa otorgando valor cero a las variables de clasificación de un conjunto inicial $\boldsymbol{S}_0 = (\boldsymbol{y}_0, \boldsymbol{X}_0)$, y valor uno en caso contrario. El conjunto

$\boldsymbol{S}_0$ debe ser una submuestra de tamaño n_0 tal que la probabilidad de que contenga más de una observación atípica sea muy baja. En este caso se garantiza que: (1) si $\boldsymbol{S}_0$ no contiene datos atípicos se obtienen estimadores insesgados de los parámetros de la regresión que permiten la identificación de las observaciones anómalas en la siguiente iteración; (2) si en $\boldsymbol{S}_0$ se encuentra un único valor atípico, este puede provocar estimaciones sesgadas de los parámetros pero, evidente, no puede producir enmascaramiento. En este caso, el dato atípico puede ser fácilmente identificado y eliminado del conjunto inicial mediante los métodos estándar de diagnóstico para observaciones atípicas aisladas. Por ejemplo, el factor de Bayes permite comparar un modelo que incluye una determinada observación atípica frente a un modelo en el cual todos los datos proceden de la distribución central, la comparación se realiza mediante el cociente entre las probabilidades de ambos modelos. El factor de Bayes es inversamente proporcional a la densidad predictiva $p(y_j \mid \boldsymbol{y}_{0(j)})$, que se puede expresar como

$$p(y_j \mid \boldsymbol{y}_{0(j)}) \propto s_{0(j)}^{-1}(1-h_{0_j})^{1/2}\left(1+\frac{t_j^2}{n_0-p-1}\right)^{-\frac{(n_0-p)}{2}} \qquad j \in \boldsymbol{S}_0 \qquad (3.1)$$

y t_j es el residuo estudentizado dado por

$$t_j = \frac{y_j - \boldsymbol{x}_j'\hat{\boldsymbol{\beta}}_0}{s_{0(j)}(1-h_{0_j})^{1/2}} \qquad j \in \boldsymbol{S}_0, \qquad (3.2)$$

donde $\hat{\boldsymbol{\beta}}_0 = (\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0'\boldsymbol{y}_0$ es el estimador de mínimos cuadrados de la muestra $\boldsymbol{S}_0$, $h_{0j} = \boldsymbol{x}_j'(\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}\boldsymbol{x}_j$ es el potencial de la observación j-ésima, y $s_{0(j)}$ es la desviación típica muestral cuando el dato j-ésimo es eliminado y la muestra es $\boldsymbol{S}_0$. La decisión de rechazar la hipótesis de que la observación j-ésima es atípica puede realizarse mediante la *escala de evidencia de Jeffreys* (Jeffreys, 1961, apéndice B) o, como la distribución predictiva (3.1) es una función monótona de los residuos estudentizados, se puede contrastar la hipótesis mediante el test clásico de valores atípicos basado en eliminar la observación j-ésima. Los residuos estudentizados tienen una distribución *t de Student*

n_0	2	3	4	5	6	7
P_{n_0}	0.990	0.971	0.946	0.915	0.879	0.840

Tabla 3.1: Probabilidad de que a lo sumo haya un valor atípico en cualquier submuestra de tamaño n_0 con los datos de Hawkins, Bradu y Kass y $\alpha_0 = 0.1$.

con $n_0 - p - 1$ grados de libertad. El nivel de significación del contraste puede elegirse mediante la desigualdad de Bonferroni, que conduce a un nivel de significación global $\alpha_T = \alpha_I / n_0$.

Sea P_{n_0} la probabilidad de que en una submuestra $\boldsymbol{S}_0$ de tamaño n_0 no exista más de una observación atípica. Como α es la probabilidad a priori de que una observación sea atípica y α_0 es la esperanza de la distribución a priori de este parámetro, entonces el número esperado de observaciones en la muestra que han sido generadas por el modelo central es $n(1-\alpha_0)$ y el número de observaciones generadas por el modelo alternativo es $n\alpha_0$. La probabilidad P_{n_0} se calcula mediante la expresión

$$P_{n_0} = \binom{\bar{n}_\alpha}{n_0}\binom{n}{n_0}^{-1} + \binom{\bar{n}_\alpha}{n_0 - 1}\binom{n_\alpha}{1}\binom{n}{n_0}^{-1}, \tag{3.3}$$

donde n_α es el entero más próximo a $n\alpha_0$ (en caso de empate, se toma el mayor) y $\bar{n}_\alpha = n - n_\alpha$. Por ejemplo, en la tabla 3.1 se presentan las probabilidades (3.3) para los datos de Hawkins, Bradu y Kass presentados en el capítulo anterior, con $\alpha_0 = 0.1$. Esta tabla indica que si se seleccionan aleatoriamente tres observaciones y se consideran buenas en las condiciones iniciales, esta submuestra no contendrá observaciones atípicas en 971 casos entre 1000.

La decisión final sobre el tamaño del conjunto inicial $\boldsymbol{S}_0$ dependerá de dos factores: la sensibilidad, que requiere la selección inicial de pocos datos buenos; y la potencia,

que requiere tener suficientes datos para estimar los parámetros. En cualquier caso, el mínimo tamaño que se considera es el de los *conjuntos elementales* (Hawkins *et al.*, 1984), que son subconjuntos de tamaño p.

3.2.2 La matriz de covarianzas

El procedimiento propuesto para seleccionar las condiciones iniciales en la primera etapa no garantiza que $\boldsymbol{S}_0$ no contenga observaciones atípicas ya que el método es probabilístico y además no se conoce la proporción exacta de contaminación en la muestra. Si el conjunto inicial $\boldsymbol{S}_0$ contiene datos atípicos con alto potencial (atípicos influyentes), la probabilidad de que en las siguientes iteraciones se identifiquen estos datos y los que enmascaran será muy baja (véase la sección 2.3 del capítulo 2). La probabilidad de identificar todos los datos atípicos de la muestra cuando las series se estabilizan (supongamos que sucede a partir de la iteración S-ésima) será q, la probabilidad de que no haya atípicos en $\boldsymbol{S}_0$, que no se conoce. Lo que si se sabe es que las variables de clasificación correspondientes a grupos de datos atípicos enmascarados, o de datos buenos señalados, deben tener generalmente los mismos valores cuando las series se estabilizan. Por tanto, la matriz de covarianzas del vector $\boldsymbol{\delta}^{(S)}$ contiene información acerca de la dependencia entre las variables de clasificación que puede emplearse para identificar grupos de observaciones con efectos similares. Se espera que pares de observaciones que pertenecen a los grupos de datos enmascarados o señalados tengan covarianzas altas en valor absoluto, mientras que la covarianza entre un dato atípico y un dato bueno, y la covarianza entre dos datos buenos será muy próxima a cero. Este hecho sugiere estimar la matriz de covarianzas a posteriori de $\boldsymbol{\delta}^{(S)}$ y buscar conjuntos de datos con covarianzas altas en valor absoluto que deben corresponder a datos enmascarados o señalados.

Sea C la matriz de covarianzas de las variables binarias $\boldsymbol{\delta}^{(S)}$. Su elemento (i,j) es

$$c_{ij} = P(\delta_i^{(S)} = 1, \delta_j^{(S)} = 1 \mid \boldsymbol{y}) - P(\delta_i^{(S)} = 1 \mid \boldsymbol{y}) \cdot P(\delta_j^{(S)} = 1 \mid \boldsymbol{y}),$$

que puede estimarse calculando las probabilidades después de S iteraciones de R ejecuciones en paralelo del Gibbs Sampling. Los estimadores que se obtienen son

$$\hat{c}_{ij} = \hat{p}_{ij_R}^{(S)} - \hat{p}_{i_R}^{(S)} \cdot \hat{p}_{j_R}^{(S)},$$

donde $\hat{p}_{i_R}^{(S)}$ es el estimador de $P(\delta_i^{(S)} = 1 \mid \boldsymbol{y})$ dado en la ecuación (2.8), y $\hat{p}_{ij_R}^{(S)}$ es el estimador de $P(\delta_i^{(S)} = 1, \delta_j^{(S)} = 1 \mid \boldsymbol{y})$ dado por

$$\hat{p}_{ij_R}^{(S)} = \frac{1}{R} \sum_{r=1}^{R} \delta_{ir}^{(S)} \, \delta_{jr}^{(S)}.$$

Esta matriz de covarianzas está relacionada con la matriz de interacciones que proponen Peña y Tiao (1992) para calcular las curvas BROC y SEBROC. La matriz de interacciones no se construye con las probabilidades marginales como la matriz de covarianzas, sino que sus elementos son la diferencia entre la probabilidades conjuntas de que: (1) una observación sea atípica y las restantes buenas; y (2) dos observaciones sean atípicas y el resto buenas.

Para localizar conjuntos de observaciones con estructura similar de dependencia, es natural tratar de identificarlos mediante el estudio de los valores y vectores propios de la matriz $\hat{C}$. Esta misma idea es utilizada por Peña y Yohai (1995) para la identificación de datos atípicos mediante la matriz de influencia. Para estudiar los autovalores de la matriz $\hat{C}$, se denomina $\boldsymbol{D}$ a la matriz de datos de las variables de clasificación que se obtiene después de S iteraciones. Esta matriz es

$$\boldsymbol{D} = (\boldsymbol{\delta}_1^{(S)}, \ldots, \boldsymbol{\delta}_R^{(S)})', \tag{3.4}$$

donde las columnas son muestras aleatorias de las variables de clasificación δ_i en la iteración S. Entonces, la matriz $\hat{\boldsymbol{C}}$ se puede escribir como

$$\hat{\boldsymbol{C}} = \frac{1}{R}\boldsymbol{D}'\boldsymbol{D} - \frac{1}{R^2}\boldsymbol{D}'\mathbf{1}_{\mathrm{R}}\mathbf{1}_{\mathrm{R}}'\boldsymbol{D},$$

y los vectores propios asociados a los valores propios no nulos de $\hat{\boldsymbol{C}}$ son los coeficientes de las componentes principales de $\boldsymbol{D}$.

Consideremos de nuevo el caso en el que existe sólo un grupo de valores atípicos que se enmascaran y analicemos la estructura de valores y vectores propios de la matriz de covarianzas $\hat{\boldsymbol{C}}$ en esta situación. Llamemos $\boldsymbol{d}_i$ a la columna i-ésima de $\boldsymbol{D}$ y, sin pérdida de generalidad, se supone que el grupo de atípicos corresponde a las últimas columnas de la matriz $\boldsymbol{D}$. Además, se supone que los atípicos señalan a un conjunto H de observaciones buenas. Finalmente, se denomina G al conjunto de datos buenos no señalados. Supongamos que los tamaños de estos conjuntos son n_{I}, n_{H} y n_{G} ($n = n_{\mathrm{G}} + n_{\mathrm{H}} + n_{\mathrm{I}}$), y que los datos señalados corresponden a las columnas anteriores a las de los atípicos.

Cuando las series de probabilidades son estables se denominan $J_1^{(\mathrm{S})}, \ldots, J_{\mathrm{R}}^{(\mathrm{S})}$ a los conjuntos que el Gibbs Sampling identifica como atípicos al final de cada secuencia. Por la condición de dependencia inicial, $J_r^{(\mathrm{S})}$ es igual a I cuando el conjunto inicial $\boldsymbol{S}_0$ no contiene atípicos, lo que se espera que suceda en $Q = qR$ de estos conjuntos, mientras que será distinto de I en $\bar{Q} = R - Q$. Con el objeto de poder describir los elementos de la matriz $\boldsymbol{D}$ en el caso particular que estamos considerando, supongamos que, sin pérdida de generalidad, las Q primeras ejecuciones en paralelo del Gibbs Sampling corresponden a las que se inician con un conjunto $\boldsymbol{S}_0$ que no contiene datos atípicos. Bajo estas condiciones se caracterizan los siguientes tipos de columnas en la matriz $\boldsymbol{D}$:

(a) Las columnas correspondientes a datos buenos que no son señalados como atípicos

son de la forma

$$\boldsymbol{d}_j = \begin{pmatrix} \mathbf{0}_{\mathrm{Q}} \\ \boldsymbol{g}_j \end{pmatrix} \qquad j = 1, \ldots, n_{\mathrm{G}},$$

donde, por la propiedad de dependencia inicial, $\mathbf{0}_{\mathrm{Q}}$ es un vector no nulo de dimensión $Q \times 1$. El vector $\boldsymbol{g}_j = (g_{1j}, \ldots, g_{\bar{\mathrm{Q}}j})'$ puede tener unos pocos elementos distintos de cero debido a que la probabilidad de identificar como atípicos a datos buenos es muy pequeña, pero no nula. Podemos suponer que no existen diferencias importantes en la proporción de unos (datos buenos mal clasificados) entre estas columnas, y que están acotadas por un valor muy pequeño π, tal que

$$\frac{1}{R} \sum_{i=1}^{\bar{\mathrm{Q}}} g_{ij} \leq \pi \quad \text{para todo } j = 1, \ldots, n_{\mathrm{G}}. \tag{3.5}$$

(b) Las columnas correspondientes a datos buenos señalados como atípicos son de la forma

$$\boldsymbol{d}_j = \begin{pmatrix} \mathbf{0}_{\mathrm{Q}} \\ \mathbf{1}_{\bar{\mathrm{Q}}} \end{pmatrix} \qquad j = n_{\mathrm{G}} + 1, \ldots, n_{\mathrm{G}} + n_{\mathrm{H}},$$

donde $\mathbf{1}_{\bar{\mathrm{Q}}}$ es un vector unitario de dimensión $\bar{Q} \times 1$.

(c) Las columnas correspondientes a datos atípicos son de la forma

$$\boldsymbol{d}_j = \begin{pmatrix} \mathbf{1}_{\mathrm{Q}} \\ \boldsymbol{g}_j \end{pmatrix} \qquad j = n - n_{\mathrm{I}}, \ldots, n, \tag{3.6}$$

donde

$$0 \leq \sum_{j=n-n_{\mathrm{I}}}^{n} g_{ij} < n_{\mathrm{I}} \qquad i = 1, \ldots, \bar{Q}.$$

El número de unos en $\boldsymbol{g}_j$ depende del grado de enmascaramiento. Los dos casos extremos son: (1) los datos atípicos del conjunto I son atípicos aislados, lo que implica que $\boldsymbol{g}_j = \mathbf{1}_{\bar{\mathrm{Q}}}$; y (2) los datos atípicos del conjunto I son idénticos y con un potencial elevado, lo que implica que $\boldsymbol{g}_j = \mathbf{0}_{\bar{\mathrm{Q}}}$. Consideremos el segundo caso

en el cual el Gibbs Sampling falla completamente como se vio en el capítulo 2. Para esta estructura de atípicos los vectores columna (3.6) son

$$\boldsymbol{d}_j = \begin{pmatrix} \mathbf{1}_{\mathrm{Q}} \\ \mathbf{0}_{\bar{\mathrm{Q}}} \end{pmatrix} \qquad j = n - n_{\mathrm{I}}, \ldots, n.$$

Teniendo en cuenta (a)-(c), la matriz $\boldsymbol{D}$ se puede expresar como la matriz por bloques

$$\boldsymbol{D} = \begin{pmatrix} \mathbf{0} & \vdots & \mathbf{0} & \vdots & \mathbf{1}_{\mathrm{Q}}\mathbf{1}'_{n_{\mathrm{I}}} \\ \cdots & & \cdots & & \cdots \\ \boldsymbol{G} & \vdots & \mathbf{1}_{\bar{\mathrm{Q}}}\mathbf{1}'_{n_{\mathrm{H}}} & \vdots & \mathbf{0} \end{pmatrix},$$

donde $\boldsymbol{G} = (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_{n_{\mathrm{G}}})$ es una matriz de dimensión $\bar{Q} \times n_{\mathrm{G}}$. La matriz de covarianzas $\hat{\boldsymbol{C}}$ se puede escribir como

$$\hat{\boldsymbol{C}} = \begin{pmatrix} \frac{1}{R}\boldsymbol{G}'\boldsymbol{G} - \frac{1}{R^2}\boldsymbol{G}'\mathbf{1}_{\bar{\mathrm{Q}}}\mathbf{1}'_{\bar{\mathrm{Q}}}\boldsymbol{G} & \vdots & \frac{Q}{R^2}\boldsymbol{G}'\mathbf{1}_{\bar{\mathrm{Q}}}\mathbf{1}'_{n_{\mathrm{H}}} & -\frac{Q}{R^2}\boldsymbol{G}'\mathbf{1}_{\bar{\mathrm{Q}}}\mathbf{1}'_{n_{\mathrm{I}}} \\ \cdots & & \cdots & \cdots \\ \frac{Q}{R^2}\mathbf{1}_{n_{\mathrm{H}}}\mathbf{1}'_{\bar{\mathrm{Q}}}\boldsymbol{G} & \vdots & \frac{Q\bar{Q}}{R^2}\mathbf{1}_{n_{\mathrm{H}}}\mathbf{1}'_{n_{\mathrm{H}}} & -\frac{Q\bar{Q}}{R^2}\mathbf{1}_{n_{\mathrm{H}}}\mathbf{1}'_{n_{\mathrm{I}}} \\ -\frac{Q}{R^2}\mathbf{1}_{n_{\mathrm{I}}}\mathbf{1}'_{\bar{\mathrm{Q}}}\boldsymbol{G} & \vdots & -\frac{Q\bar{Q}}{R^2}\mathbf{1}_{n_{\mathrm{I}}}\mathbf{1}'_{n_{\mathrm{H}}} & \frac{Q\bar{Q}}{R^2}\mathbf{1}_{n_{\mathrm{I}}}\mathbf{1}'_{n_{\mathrm{I}}} \end{pmatrix}.$$

Asumiendo que π es una cantidad muy próxima a cero, esta matriz de covarianzas se puede aproximar por

$$\hat{\boldsymbol{C}} = \begin{pmatrix} \frac{1}{R}\boldsymbol{G}'\boldsymbol{G} & \vdots & \mathbf{0} \\ \cdots & & \cdots \\ \mathbf{0} & \vdots & \hat{\boldsymbol{C}}_{22} \end{pmatrix},$$

donde $\hat{\boldsymbol{C}}_{22}$ es una matriz de dimensión $(n_{\mathrm{H}} + n_{\mathrm{I}}) \times (n_{\mathrm{H}} + n_{\mathrm{I}})$, dada por

$$\hat{\boldsymbol{C}}_{22} = \frac{Q\bar{Q}}{R^2}\begin{pmatrix} \mathbf{1}_{n_{\mathrm{H}}}\mathbf{1}'_{n_{\mathrm{H}}} & \vdots & -\mathbf{1}_{n_{\mathrm{H}}}\mathbf{1}'_{n_{\mathrm{I}}} \\ \cdots & & \cdots \\ -\mathbf{1}_{n_{\mathrm{I}}}\mathbf{1}'_{n_{\mathrm{H}}} & \vdots & \mathbf{1}_{n_{\mathrm{I}}}\mathbf{1}'_{n_{\mathrm{I}}} \end{pmatrix}.$$

Es inmediato comprobar que los valores propios de $\hat{\boldsymbol{C}}$ coinciden con los valores propios de las matrices $\frac{1}{R}\boldsymbol{G}'\boldsymbol{G}$ y $\hat{\boldsymbol{C}}_{22}$. Por la ecuación (3.5) los valores propios de $\frac{1}{R}\boldsymbol{G}'\boldsymbol{G}$ cumplen la relación

$$\sum_{i=j}^{n_{\mathrm{G}}} \lambda_j = \mathrm{traza}\left(\frac{1}{R}\boldsymbol{G}'\boldsymbol{G}\right) = \frac{1}{R}\sum_{j=1}^{n_{\mathrm{G}}}\sum_{i=1}^{\bar{Q}} g_{ij}^2 \leq \pi\, n_{\mathrm{G}}.$$

La matriz $\hat{C}_{22}$ tiene un único valor propio distinto de cero, que es

$$\lambda_{\mathrm{I}} = q(1-q)(n_{\mathrm{H}} + n_{\mathrm{I}}). \tag{3.7}$$

Entonces, la matriz $\hat{C}$ tiene un valor propio igual a λ_{I} y n_{G} valores propios adicionales cuya suma es menor o igual que πn_{G}, donde π es una probabilidad muy próxima a cero. Además, para cualquier valor no nulo de a, $\boldsymbol{v}_a = (\mathbf{0}'_{n_{\mathrm{G}}}, a\mathbf{1}'_{n_{\mathrm{H}}}, -a\mathbf{1}'_{n_{\mathrm{I}}})'$ es un vector propio de la matriz $\hat{C}$ asociado al valor propio λ_{I}.

El valor propio λ_{I} será próximo a cero (no es posible identificar los atípicos) cuando la probabilidad q sea muy pequeña o casi uno. Como q es la probabilidad de que no haya atípicos en el conjunto $\boldsymbol{S}_0$, se obtendrá un valor de q cercano al cero cuando existe un porcentaje alto de contaminación en la muestra y el tamaño de $\boldsymbol{S}_0$ no es suficientemente pequeño. Este problema se evita con el procedimiento propuesto para seleccionar los valores iniciales en la primera etapa, que tiene por objeto conseguir que con alta probabilidad el conjunto $\boldsymbol{S}_0$ no contenga observaciones atípicas. Por otro lado, un valor de q cercano a uno corresponde al caso en el que no existen valores atípicos en la muestra, o existe sólo un porcentaje muy bajo y el tamaño de $\boldsymbol{S}_0$ es pequeño. En este caso los atípicos se identifican directamente por las estimaciones de sus probabilidades que proporciona el Gibbs Sampling. La situación más interesante es cuando $0 < q < 1$ y n_{I} (y muy posiblemente n_{H}) es relativamente elevado, que corresponde al caso más difícil en el que los valores atípicos no se identifican al final de muchas secuencias y además suelen provocar que aparezcan como atípicas observaciones que no lo son. Entonces el valor propio λ_{I} será relativamente grande y el vector propio asociado indicará correctamente que datos están enmascarados o señalados: las observaciones con coordenadas relativamente altas (en valor absoluto) en el vector propio $\boldsymbol{v}_a$ son atípicas o señaladas como atípicas. Este análisis permite dividir la muestra en dos grupos: (1) el conjunto que contiene las observaciones con coordenadas distintas

de cero en el vector propio o con alta probabilidad individual $\hat{p}_i^{(s)}$; y (2) el resto de las observaciones. Llamaremos al primer conjunto *conjunto de posibles atípicos*, que abreviaremos con las siglas PO (del inglés, *potential outlier*).

Por ejemplo, en la tabla 3.2 se presenta la matriz de covarianzas para los datos de Hawkins, Bradu y Kass, y en la tabla 3.3 sus autovalores más grandes. Se observa en esta tabla que existe un autovalor claramente mayor que los restantes que son muy próximos a cero. En las coordenadas del vector propio asociado a este autovalor (figura 3.1) se aprecia que sólo las coordenadas de las primeras 14 observaciones son claramente distintas de cero. Además, la tabla 3.2 muestra que en la matriz estimada las covarianzas entre estos datos son grandes (en valor absoluto). En consecuencia, se deben incluir en el grupo PO de posibles atípicos las observaciones 1 a 14. La matriz $\hat{C}$ se ha construido con las probabilidades estimadas después de 500 iteraciones. Nótese que para este conjunto de datos $q = 0.557$ se calculó en (2.13), $n_{\text{I}} = 10$, $n_{\text{H}} = 4$ y, de acuerdo con la ecuación (3.7) el valor esperado del mayor autovalor es igual a 3.45, que es muy similar al que se obtiene y se muestra en la tabla 3.3.

Cuando la muestra contiene varios grupos de atípicos, estos sólo pueden producir p efectos diferentes de enmascaramiento en $\mathbf{R}^p$. Por tanto, el número máximo de valores propios que hay que examinar es p. Una generalización inmediata del análisis realizado previamente indica que los efectos independientes aparecerán en distintos vectores propios de la matriz de covarianzas estimada $\hat{C}$. Este resultado es la base del procedimiento que se presenta a continuación.

3.2.3 Algoritmo adaptativo de Gibbs Sampling I

El método propuesto para seleccionar las condiciones iniciales en la primera etapa, junto con la información que proporciona la matriz de covarianzas de las variables de

2	.22									
3	.23	.22								
4	.22	.22	.22							
5	.22	.22	.23	.22						
6	.23	.22	.23	.22	.23					
7	.22	.22	.22	.21	.22	.22				
8	.22	.22	.22	.22	.22	.22	.21			
9	.22	.22	.22	.22	.22	.22	.22	.22		
10	.23	.22	.23	.22	.23	.23	.22	.22	.22	
11	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22
12	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22
13	-.22	-.22	-.22	-.22	-.22	-.22	-.21	-.22	-.22	-.22
14	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22
15	.	.	.	.	.	.	.	.	.	.
16	.	.	.	.	.	.	.	.	.	.
17	.	.	.	.	.	.	.	.	.	.
18	.	.	.	.	.	.	.	.	.	.

11	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22
12	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22
13	-.22	-.22	-.22	-.22	-.22	-.22	-.21	-.22	-.22	-.22
14	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22
15	.	.	.	.	.	.	.	.	.	.
16	.	.	.	.	.	.	.	.	.	.
17	.	.	.	.	.	.	.	.	.	.
18	.	.	.	.	.	.	.	.	.	.

11	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22
12	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22
13	-.22	-.22	-.22	-.22	-.22	-.22	-.21	-.22	-.22	-.22
14	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22
15	.	.	.	.	.	.	.	.	.	.
16	.	.	.	.	.	.	.	.	.	.
17	.	.	.	.	.	.	.	.	.	.
18	.	.	.	.	.	.	.	.	.	.

Tabla 3.2: Matriz de covarianzas con los datos de Hawkins, Bradu y Kass. Sólo se representan valores superiores a 0.01 en valor absoluto.

Componente	Autovalor	Variabilidad explicada	Porcentaje acumulado variabilidad explicada
1	3.4297	0.7849	78.49
2	0.0391	0.0089	79.38
3	0.0338	0.0077	80.15

Tabla 3.3: Autovalores de la matriz de covarianzas con los datos de Hawkins, Bradu y Kass.

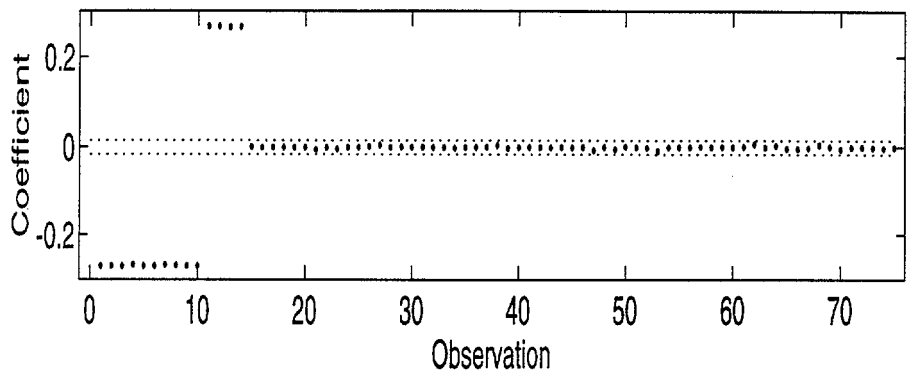

Figura 3.1: Coordenadas del vector propio asociado al mayor valor propio λ_1 de la matriz de covarianzas con los datos de Hawkins, Bradu y Kass.

clasificación, permite dividir el conjunto de datos en dos grupos: PO y $\overline{\text{PO}}$. Si se inicia el algoritmo de Gibbs Sampling dando valor 1 a las variables de clasificación correspondientes a los datos del grupo PO, entonces los valores que se obtienen cuando se ejecuta el algoritmo con pocas iteraciones son una muestra de la distribución a posteriori del vector $\boldsymbol{\delta}$. La estimación de las probabilidades $p_i = P(\delta_i = 1 \mid \boldsymbol{y})$ para $i = 1, \ldots, n$, permite identificar las observaciones atípicas, siendo estas las que tienen una probabilidad superior a 0.5. Con estas indicaciones el algoritmo adaptativo de Gibbs Sampling que proponemos se lleva a cabo en las dos etapas siguientes:

Etapa 1: Se ejecuta el algoritmo de Gibbs Sampling en paralelo hasta que las series de probabilidades de que cada dato sea atípico permanezcan estables. Para cada secuencia se seleccionan las condiciones iniciales siguiendo los pasos:

i. Sea n_0 el máximo entero tal que la probabilidad (3.3) de no encontrar más de un dato atípico en cualquier submuestra de tamaño n_0 es mayor o igual que c_1. Se seleccionan $m = \max\{n_0, p\}$ números aleatorios $i_1, \ldots, i_m$ entre $1, \ldots, n$.

ii. Se construye el conjunto inicial $\boldsymbol{S}_0 = \{(y_{i_1}, \boldsymbol{x}_{i_1}), \ldots, (y_{i_m}, \boldsymbol{x}_{i_m})\}$. Si $m > p$, se calculan los residuos estudentizados $t_{i_1}, \ldots, t_{i_m}$ dados por la ecuación (3.2).

iii. Cuando $m = p$, las variables de clasificación iniciales son:

$$\delta_j^{(0)} = \begin{cases} 0 & \text{si } j = i_1, \ldots, i_m \\ 1 & \text{en caso contrario} \end{cases}$$

Cuando $m > p$, las variables de clasificación iniciales son:

$$\delta_j^{(0)} = 1 \qquad \text{para } j \neq i_1, \ldots, i_m$$

$$\delta_j^{(0)} = \begin{cases} 0 & \text{si } t_j < t_{m-p-1,\alpha_1/n_0} \quad (t \textit{ de Student}) \\ 1 & \text{en caso contrario} \end{cases} \qquad \text{para } j = i_1, \ldots, i_m$$

iv. Si $n - \sum_{j=1}^{n} \delta_j < p$, o la submatriz de $\boldsymbol{X}$ formada por las filas correspondientes a las observaciones para las que $\delta_j^{(0)} = 0$ no es definida positiva, entonces repetir los pasos *(i)-(iii)*. En caso contrario, $\beta^{(0)} = (\boldsymbol{X}'\boldsymbol{V}^{(0)^{-1}}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{V}^{(0)^{-1}}\boldsymbol{y}$, donde $\boldsymbol{V}^{(0)}$ es una matriz diagonal con elementos en la diagonal principal $v_{jj}^{(0)} = 1 + \delta_j^{(0)}(k-1)$.

Con los valores de las variables de clasificación que se obtienen en la última iteración de cada secuencia se estima la matriz de covarianzas $\hat{\boldsymbol{C}}$ y los vectores propios asociados a los c_2 autovalores mayores $(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots)$. Se divide la muestra en los grupos PO y $\overline{\text{PO}}$ siguiendo los pasos:

a. Si $\hat{p}_i^{(s)} > 0.5$, entonces $(y_j, \boldsymbol{x}_j') \in$ PO.

b. Para $i = 1, \ldots, c_2$ y $j = 1, \ldots, n$, se calcula $m_i =$ mediana $\mid v_{ij} \mid /0.6475$. Si $\mid v_{ij} \mid > c_3\, m_i$, entonces $(y_j, \boldsymbol{x}_j') \in$ PO.

c. Si el par $(y_j, \boldsymbol{x}_j')$ no está en PO, entonces esta en $\overline{\text{PO}}$.

Etapa 2. Se reinicia el algoritmo de Gibbs Sampling hasta que las series de probabilidades de que cada dato sea atípico permanezcan estables. Para cada secuencia se seleccionan las condiciones iniciales siguientes:

1. $\delta_j^{(0)} = 1$ si $(y_j, \boldsymbol{x}_j') \in$ PO, y $\delta_j^{(0)} = 0$ en caso contrario.
2. $\boldsymbol{\beta}^{(0)} = (\boldsymbol{X}'\boldsymbol{V}^{(0)^{-1}}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{(0)^{-1}}\boldsymbol{y}$, donde $\boldsymbol{V}^{(0)}$ es una matriz diagonal con elementos $v_{jj}^{(0)} = 1 + \delta_j^{(0)}(k-1)$.

Antes de ejecutar las dos etapas del algoritmo adaptativo se tienen que determinar los valores de las constantes c_1, c_2 y c_3. El criterio con el que debe seleccionarse un valor adecuado de c_1 ya fue discutido en la sección 3.2.1. Teniendo en cuenta el doble objetivo de obtener sensibilidad y potencia, valores de c_1 superiores a 0.9 e inferiores a 0.975 parecen los más adecuados. Para decidir el número de vectores propios que se examinan, el entero c_2 tiene que ser igual al número de valores propios distintos de cero. Se sugiere elegir el mínimo valor entre (p, c_2^*), donde c_2^* es el número de autovalores que superan cinco veces una medida robusta de dispersión entre los valores propios λ_i de $\hat{C}$ y que puede ser $mediana(\lambda_i)/0.6475$. La constante c_3 se utiliza para decidir qué coordenadas de los vectores propios son significativamente distintas de cero y, por tanto, qué observaciones pueden ser atípicas. También se ha optado por construir una medida robusta para medir la dispersión de las observaciones alrededor del cero, que es el valor de las coordenadas que se espera para los datos buenos. Se recomienda no elegir c_3 muy pequeño porque la medida de dispersión puede ser muy pequeña ya que generalmente las observaciones atípicas y señaladas no superan el 50 por ciento de la muestra. Las otras cantidades que se deben determinar previamente son el número de secuencias y el número de iteraciones; el número de secuencias en paralelo depende de las propiedades asintóticas de los estimadores que se proponen; y el número de iteraciones en ambas etapas puede decidirse por cualquiera de los métodos multisecuenciales para controlar la convergencia del Gibbs Sampling que se discuten en el capítulo de introducción. En particular, el procedimiento que se propone en este trabajo es más simple y proporciona un buen criterio de parada en los ejemplos en

los que se ha aplicado. Se ejecuta el Gibbs Sampling hasta la iteración S-ésima tal que, para un cierto $\epsilon > 0$, se verifica que $\mid \hat{p}_{i_R}^{(S+1)} - \hat{p}_{i_R}^{(S)} \mid < \epsilon$ para todo $i = 1, \ldots, n$. Finalmente, para reducir el esfuerzo computacional se puede ejecutar en la segunda etapa una única secuencia, ya que las condiciones iniciales son siempre las mismas, y elegir alguno de los criterios unisecuenciales.

Como resumen del funcionamiento del algoritmo se tiene que los pasos que se siguen en la primera etapa pretenden que el conjunto inicial $\boldsymbol{S}_0$ no contenga datos atípicos y que la información que aporta la matriz de covarianzas permita aislar los datos que pueden ser atípicos. Los datos con coordenadas altas en los vectores propios se identifican con una medida robusta de desviaciones al cero que es el valor de las coordenadas que se espera para los datos buenos. Finalmente, en la segunda etapa se ejecuta el algoritmo con unas condiciones iniciales fijas. El procedimiento concluye cuando las series de probabilidades presentan un comportamiento estable.

3.3 Comportamiento del algoritmo adaptativo

Se aplica el algoritmo adaptativo de Gibbs Sampling a algunos de los conjuntos de datos discutidos en los ejemplos del capítulo 2. Nos centraremos en aquellos casos en los que no se identifican las observaciones atípicas con el Gibbs Sampling estándar. Antes de ejecutar el algoritmo se fijan los valores de las constantes $c_1 = 0.95$ y $c_3 = 5$, y del nivel de significación individual $\alpha_I = 0.05$. El número de vectores propios c_2 que se examinan varía en cada ejemplo según el criterio establecido en la sección anterior. En las dos etapas se ejecutan 300 secuencias del Gibbs Sampling en paralelo y el número de iteraciones se decide con $\epsilon = 0.002$. En todos los ejemplos $k = 10$, $\alpha_0 = 0.2$ y $\gamma_1 + \gamma_2 = n$, de manera que como $\alpha_0 = \gamma_1/(\gamma_1 + \gamma_2)$ se tiene que en cada iteración $E(\alpha \mid \boldsymbol{\delta}) = 1/2E(\alpha) + 1/2\bar{\delta}$. En las dos etapas se estiman las probabilidades de que

cada dato sea atípico con las muestras que proporcionan las últimas iteraciones de cada secuencia.

En todos los ejemplos se comparan los resultados que proporciona el algoritmo adaptativo con los resultados de aplicar a las dos variantes del procedimiento para identificar grupos de atípicos de Hadi y Simonoff (1993) y el procedimiento de Peña y Yohai (1995), brevemente descritos en la introducción. En los dos procedimientos las observaciones atípicas son las que tienen un residuo estudentizado grande. En algunos casos los residuos se calculan fuera de la muestra, pero en las tablas donde se presentaran los resultados no se diferenciarán.

3.3.1 Ejemplo 1: Diagrama de Hertzsprung-Russell

En el primer ejemplo se aplica el algoritmo adaptativo de Gibbs Sampling a los datos del diagrama de Hertzsprung-Russell que se muestran en la tabla 2.4 y en la figura 2.6. Los datos atípicos se encuentran en las posiciones 11, 20, 30 y 34; las otras observaciones más alejadas de la nube de puntos son la 7, 9 y 18. Los residuos estudentizados que se obtienen con los procedimientos de Hadi y Simonoff (1993) y Peña y Yohai (1995) se muestran en las tres primeras columnas de la tabla 3.5. Es inmediato comprobar que los cuatro datos atípicos se identifican con cualquiera de los tres métodos.

En la figura 3.3(a) se presentan las probabilidades estimadas en la etapa 1 de que cada observación sea atípica. Ya en esta primera etapa se identifica correctamente que observaciones son atípicas, con probabilidades superiores a 0.5. El Gibbs Sampling se inicia en la primera etapa con tres observaciones consideradas como buenas y el número de vectores propios que se examinan es $p = 2$ (véase la tabla 3.4 para los valores numéricos de los autovalores). Los dos vectores propios asociados a los mayores autovalores se muestran en la figura 3.2. Los puntos representan las coordenadas

Componente	Autovalor	Variabilidad explicada	Porcentaje acumulado variabilidad explicada
1	0.9717	0.3612	36.12
2	0.2131	0.0792	44.04
3	0.1241	0.0461	48.65
4	0.1090	0.0405	52.70

Tabla 3.4: Autovalores de la matriz de covarianzas con los datos del diagrama de Hertzsprung-Russell.

de cada dato en el vector propio y la líneas discontinuas representan las bandas de confianza para el cero que se construyen tal y como se indica en el paso (b) de la etapa 1. Se puede ver que claramente las coordenadas correspondientes a los datos atípicos están fuera de las bandas de confianza en los dos vectores propios. Existen otros tres puntos fuera de las bandas del segundo vector propio, con signos opuestos a los puntos del grupo de atípicos y que corresponden a los datos 7, 9 y 18. Además, las coordenadas de las observaciones 5, 14 y 40 están en el límite de las bandas y también se incluyen en el grupo de posibles atípicos PO. Con esta información se fijan las condiciones iniciales con las que se inicia el Gibbs Sampling en la etapa 2 y cuyos resultados pueden verse en la figura 3.3(b). Se confirma que las probabilidades a posteriori de que sean atípicas cada una de las cuatro estrellas gigantes son superiores a 0.5, aunque levemente más bajas que las que se obtienen al final de la primera etapa.

3.3.2 Ejemplo 2: Datos de Hawkins, Bradu y Kass

Los datos de Hawkins, Bradu y Kass se muestran en la tabla 2.2 y en la figura 2.2. Recordemos que las primeras 10 observaciones son atípicas y las 4 siguientes son

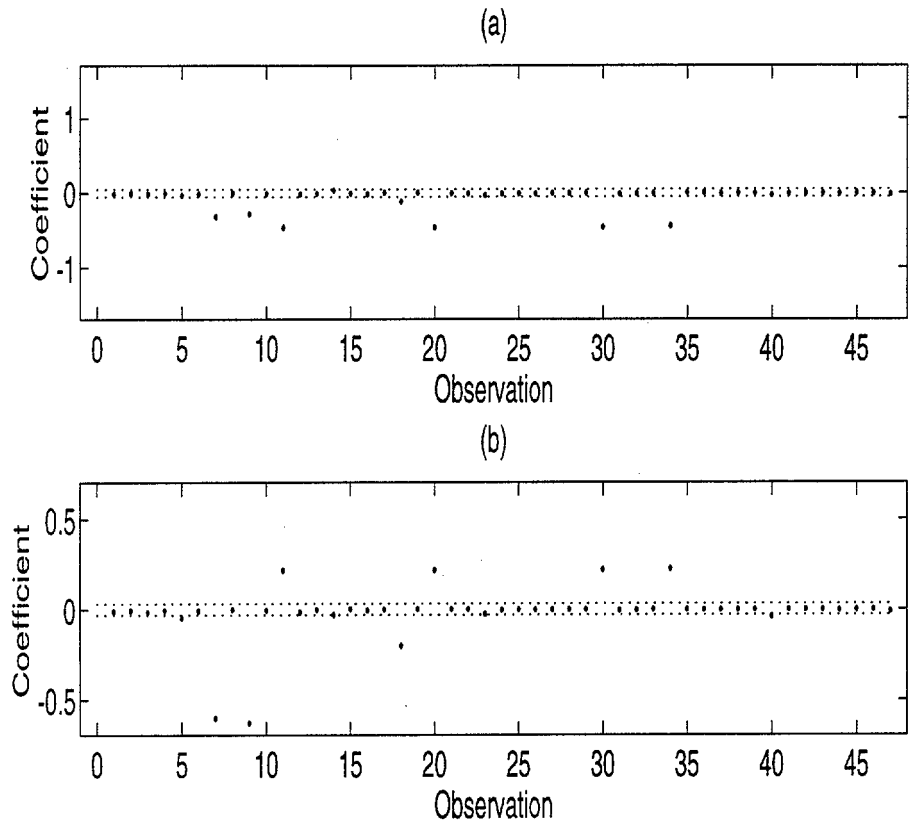

Figura 3.2: Coordenadas del vector propio asociado a los valores propios λ_1 (en (a)) y λ_2 (en (b)) de la matriz de covarianzas con los datos del diagrama de Hertzsprung-Russell.

	Res. estudentizado			Prob.		Res. estudentizado			Prob.
Dato	HS-I	HS-II	PY	AAGS	Dato	HS-I	HS-II	PY	AAGS
1	0.264	-1.582	0.861	0.028	26	-2.686	0.805	-0.829	0.025
2	1.670	0.025	1.195	0.042	27	-1.930	-1.268	-0.159	0.019
3	-0.578	-2.421	0.677	0.027	28	-1.403	-0.500	-0.019	0.017
4	1.670	0.025	1.195	0.042	29	-2.730	-1.653	-0.483	0.024
5	0.480	-2.497	1.133	0.045	**30**	5.607	-16.421	5.328	0.627
6	0.862	-0.797	0.982	0.030	31	-3.576	0.846	-1.238	0.039
7	0.868	-8.112	2.726	0.337	32	-1.485	1.803	-0.447	0.021
8	-0.707	1.497	-0.067	0.019	33	-0.246	-0.278	0.429	0.019
9	2.386	-4.064	2.437	0.296	**34**	6.310	-16.420	5.750	0.643
10	-0.272	-1.271	0.582	0.022	35	-3.018	-1.355	-0.662	0.026
11	4.673	-15.943	4.745	0.620	36	0.740	1.346	0.570	0.026
12	0.986	-1.252	1.113	0.036	37	-1.302	1.327	-0.289	0.019
13	0.553	-0.361	0.775	0.025	38	-0.246	-0.278	0.422	0.019
14	-2.845	-3.741	-0.273	0.051	39	-0.910	1.106	-0.087	0.018
15	-3.765	-0.155	-1.180	0.039	40	1.570	-1.603	1.430	0.056
16	-3.059	1.031	-1.035	0.031	41	-2.644	0.279	-0.722	0.023
17	-4.821	-0.229	-1.727	0.092	42	-1.024	0.166	0.022	0.017
18	-4.927	2.012	-2.126	0.162	43	0.048	0.184	0.469	0.020
19	-3.739	-0.913	-1.077	0.039	44	0.336	-0.616	0.726	0.024
20	5.141	-16.131	5.032	0.623	45	0.741	0.430	0.723	0.025
21	-3.214	-0.493	-0.868	0.028	46	-1.414	0.388	-0.177	0.018
22	-3.948	-0.041	-1.286	0.044	47	-3.433	1.260	-1.244	0.040
23	-3.807	1.492	-1.456	0.055					
24	-2.182	1.385	-0.713	0.023					
25	-0.818	-0.833	0.279	0.018					

Tabla 3.5: Resultados de los procedimientos de Hadi y Simonoff (HS-I y HS-II), Peña y Yohai (PY) y el algoritmo adaptativo de Gibbs Sampling (AAGS) con los datos del diagrama de Hertzsprung-Russell.

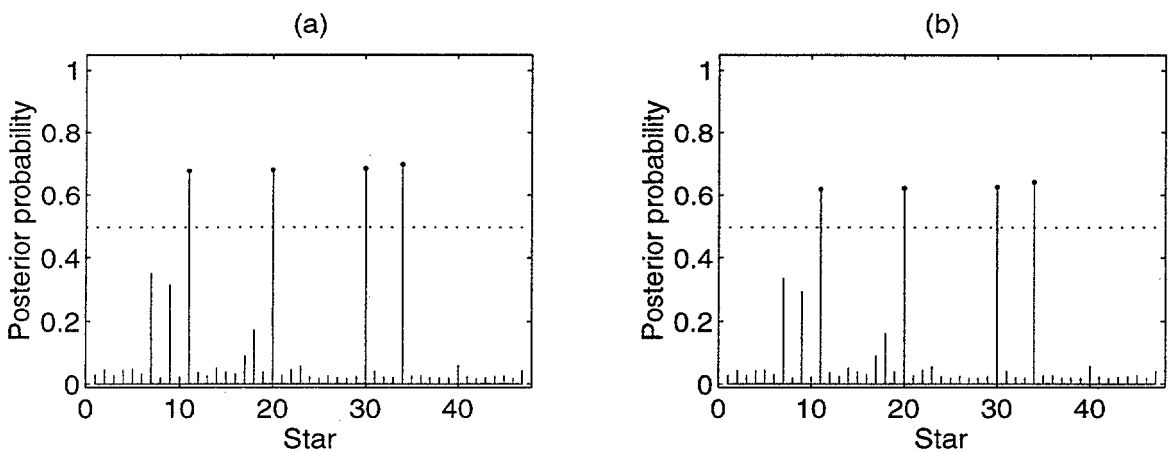

Figura 3.3: Resultados del Gibbs Sampling con los datos del diagrama de Hertzsprung-Russell: (a) probabilidades a posteriori de que cada dato sea atípico en la primera etapa; (b) probabilidades a posteriori en la segunda etapa.

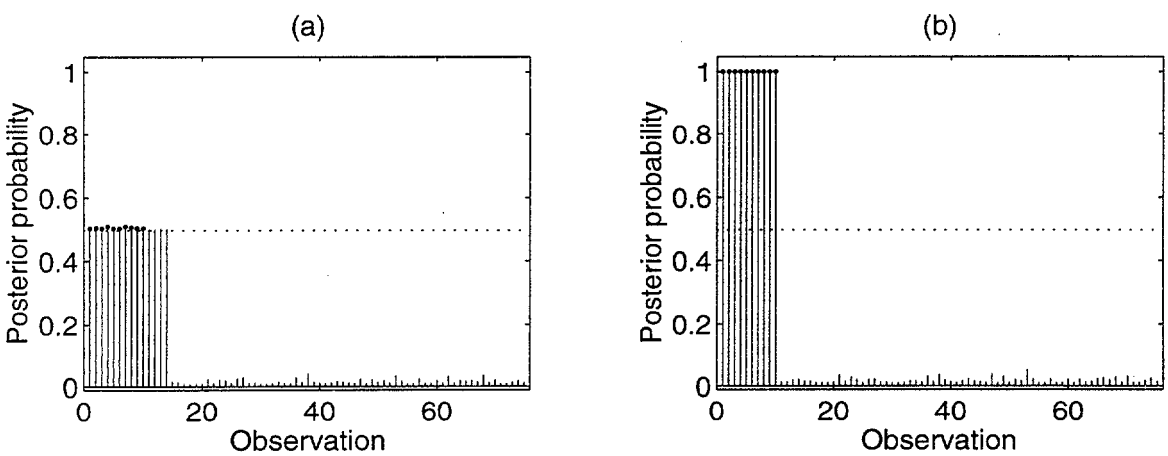

Figura 3.4: Resultados del Gibbs Sampling con los datos de Hawkins, Bradu y Kass: (a) probabilidades a posteriori de que cada dato sea atípico en la primera etapa; (b) probabilidades a posteriori en la segunda etapa.

observaciones buenas señaladas como atípicas y con potencial alto. En este conjunto de datos los procedimientos de Hadi y Simonoff (1993) fallan debido a que los datos atípicos son también muy influyentes, mientras que el procedimiento de Peña y Yohai (1995) identifica adecuadamente los datos atípicos. En la tabla 3.6 se observa que los residuos más grandes que proporcionan los métodos de Hadi y Simonoff (1993) corresponden a las observaciones señaladas, quedando los atípicos enmascarados.

Cuando se aplica el algoritmo adaptativo a estos datos el conjunto de observaciones buenas con el que se inicia el Gibbs Sampling en la etapa 1 incluye cuatro observaciones, que es el tamaño de los conjuntos elementales. Existe un único valor propio de la matriz de covarianzas significativamente distinto de cero (véase tabla 3.3) y las coordenadas del vector propio asociado se muestran en la figura 3.1. Las conclusiones que se obtienen al final de la primera etapa son las mismas tanto si se consideran las probabilidades individuales estimadas en la última iteración y que se muestran en la figura 3.4(a), como si se considera la estructura de vectores propios de la matriz de covarianzas que fue discutida en la sección 2.3; el grupo de posibles atípicos PO está formado por las observaciones 1 a 14, que son los datos atípicos enmascarados y los datos buenos señalados. En la etapa 2 estos datos son los únicos que se consideran como atípicos en las condiciones iniciales. Cuando las series se estabilizan las probabilidades a posteriori de que cada dato sea atípico se muestran en la figura 3.4(b), se observa que los 10 primeros datos se identifican correctamente como atípicos con probabilidades iguales a uno. Además, las probabilidades de que sean atípicos los cuatro datos que anteriormente eran señalados como atípicos son muy pequeñas (véase la cuarta columna de la tabla 3.6).

	Res. estudentizado			Prob.
Dato	HS-I	HS-II	PY	AAGS
1	1.0762	-0.5244	5.3525	1.0000
2	2.2188	-0.7264	5.4420	1.0000
3	0.1100	0.8893	5.3188	1.0000
4	-1.5237	0.9149	4.8893	1.0000
5	-0.1409	0.3787	5.1448	1.0000
6	0.7106	-0.9386	5.3135	1.0000
7	2.9565	-1.6929	5.6465	1.0000
8	2.2196	-0.1829	5.5893	1.0000
9	-0.6850	1.1203	5.0402	1.0000
10	0.8538	1.5551	5.3079	1.0000
11	-26.6269	3.1950	0.9464	0.0117
12	-28.7513	4.9007	0.9020	0.0117
13	-25.1989	1.7252	0.6873	0.0185
14	-11.8374	-3.7948	0.8719	0.0194

Tabla 3.6: Resultados de los procedimientos de Hadi y Simonoff (HS-I y HS-II), Peña y Yohai (PY) y el algoritmo adaptativo de Gibbs Sampling (AAGS) con los datos de Hawkins, Bradu y Kass.

3.3.3 Ejemplo 3: Datos de Rousseeuw

Los datos simulados de Rousseeuw que se encuentran en la tabla 2.3 y en la figura 2.4 son los más interesantes porque muestran el alto punto de ruptura que exhibe el procedimiento basado en el algoritmo adaptativo de Gibbs Sampling. La contaminación en la muestra es del 40 por ciento y los procedimientos de Hadi y Simonoff (1993) y de Peña y Yohai (1995) no son capaces de identificar el grupo de atípicos como puede verse en la tabla 3.8.

Los resultados que proporciona el método que proponemos son muy buenos. Iniciando la etapa 1 del algoritmo con un conjunto de 4 observaciones consideradas como buenas, las probabilidades que se obtienen son superiores a 0.5 sólo para las observaciones 32 y 33. A pesar de que con el Gibbs Sampling no sólo no se identifican los atípicos, sino que además dos observaciones que son buenas son señaladas como atípicas, la información que proporciona la matriz de covarianzas permite dividir los datos en dos grupos, incluyéndose todos los atípicos en PO. La matriz de covarianzas tiene dos valores propios no nulos que se encuentran en la tabla 3.7. Las coordenadas de los vectores propios asociados se muestran en la figura 3.5. Si nos fijamos en el primer vector propio, los resultados son los que se esperan: (1) aparecen con valores distintos de cero las coordenadas correspondientes a los datos enmascarados y señalados; y (2) los signos de las coordenadas de los datos enmascarados y los señalados son opuestos. De esta forma se inicia el Gibbs Sampling en la etapa 2 considerando como atípicos las primeras 20 observaciones, más la 32 y la 33. Las probabilidades a posteriori que se muestran en la figura 3.6(b) indican claramente que existe un grupo de 20 observaciones espúreamente generadas con probabilidad uno.

Componente	Autovalor	Variabilidad explicada	Porcentaje acumulado variabilidad explicada
1	0.5264	0.3184	31.84
2	0.3113	0.1883	50.67
3	0.1576	0.0953	60.20
4	0.0671	0.0406	64.26

Tabla 3.7: Autovalores de la matriz de covarianzas con los datos simulados de Rousseeuw.

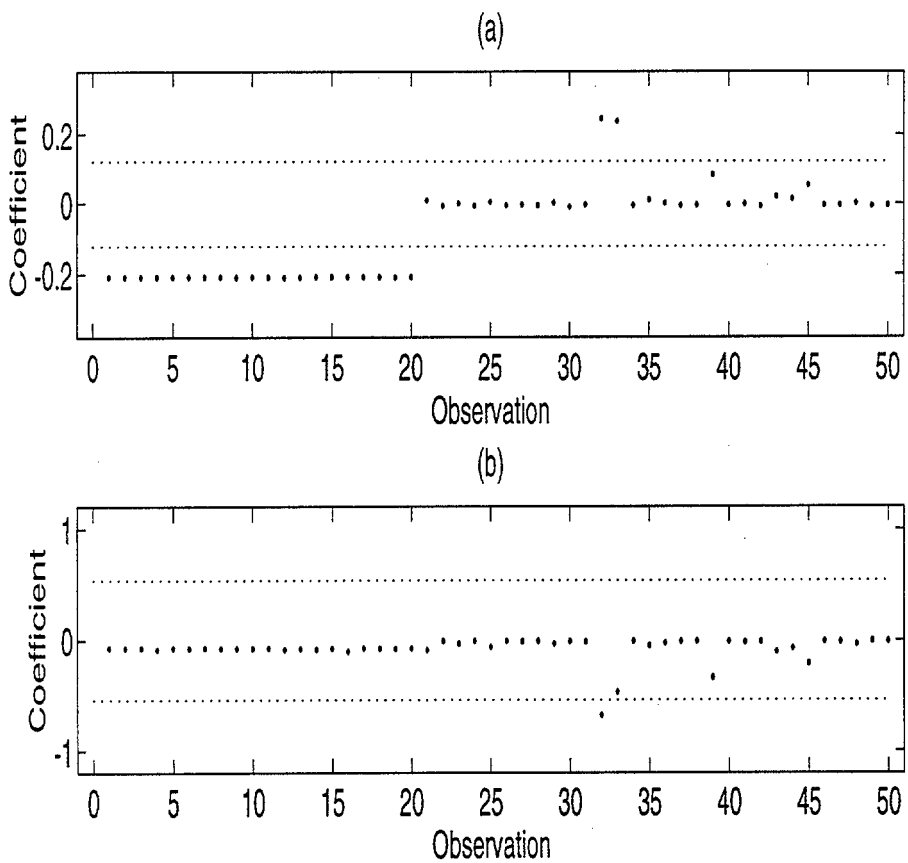

Figura 3.5: Coordenadas del vector propio asociado a los valores propios λ_1 (en (a)) y λ_2 (en (b)) de la matriz de covarianzas con los datos simulados de Rousseeuw.

	Res. estudentizado			Prob.		Res. estudentizado			Prob.
Dato	HS-I	HS-II	PY	AAGS	Dato	HS-I	HS-II	PY	AAGS
1	0.963	1.083	-0.124	1.000	26	0.233	0.222	-0.653	0.025
2	0.714	0.824	-0.232	1.000	27	-2.916	-2.862	-1.825	0.025
3	2.211	2.283	0.311	1.000	28	-1.092	-1.078	-1.118	0.024
4	4.407	4.419	1.074	1.000	29	3.275	3.197	0.440	0.024
5	0.895	1.004	-0.168	1.000	30	-1.049	-1.024	-1.077	0.066
6	-0.478	-0.328	-0.614	1.000	31	1.778	1.736	-0.127	0.027
7	1.576	1.667	0.056	1.000	32	9.540	9.323	2.868	0.059
8	-0.648	-0.494	-0.673	1.000	33	10.934	10.450	3.420	0.081
9	-0.845	-0.677	-0.723	1.000	34	-1.131	-1.113	-1.126	0.024
10	-0.742	-0.590	-0.712	1.000	35	-5.750	-5.633	-2.911	0.026
11	1.384	1.465	-0.031	1.000	36	2.952	2.883	0.329	0.024
12	-1.843	-1.794	-1.110	1.000	37	0.656	0.630	-0.523	0.028
13	-1.499	-1.329	-0.971	1.000	38	-2.536	-2.489	-1.675	0.024
14	4.602	4.636	1.233	1.000	39	6.590	6.437	1.648	0.035
15	-0.614	-0.467	-0.672	1.000	40	-3.067	-3.012	-1.888	0.027
16	5.165	5.143	1.305	1.000	41	-4.400	-4.313	-2.386	0.024
17	0.259	0.369	-0.403	1.000	42	-3.314	-3.244	-1.941	0.041
18	-0.190	-0.590	-0.538	1.000	43	4.924	4.807	1.028	0.030
19	-1.105	-0.916	-0.788	1.000	44	-6.201	-6.070	-3.064	0.039
20	3.099	3.155	0.640	1.000	45	5.993	5.858	1.426	0.025
21	4.616	4.502	0.912	0.052	46	0.645	0.622	-0.520	0.024
22	-0.788	-0.771	-0.992	0.041	47	1.532	1.489	-0.224	0.025
23	3.208	3.128	0.412	0.029	48	-5.357	-5.243	-2.729	0.039
24	-1.023	-1.015	-1.106	0.035	49	-0.283	-0.291	-0.848	0.033
25	-5.911	-5.784	-2.937	0.054	50	0.811	0.786	-0.461	0.024

Tabla 3.8: Resultados de los procedimientos de Hadi y Simonoff (HS-I y HS-II), Peña y Yohai (PY) y el algoritmo adaptativo de Gibbs Sampling (AAGS) con los datos simulados de Rousseeuw.

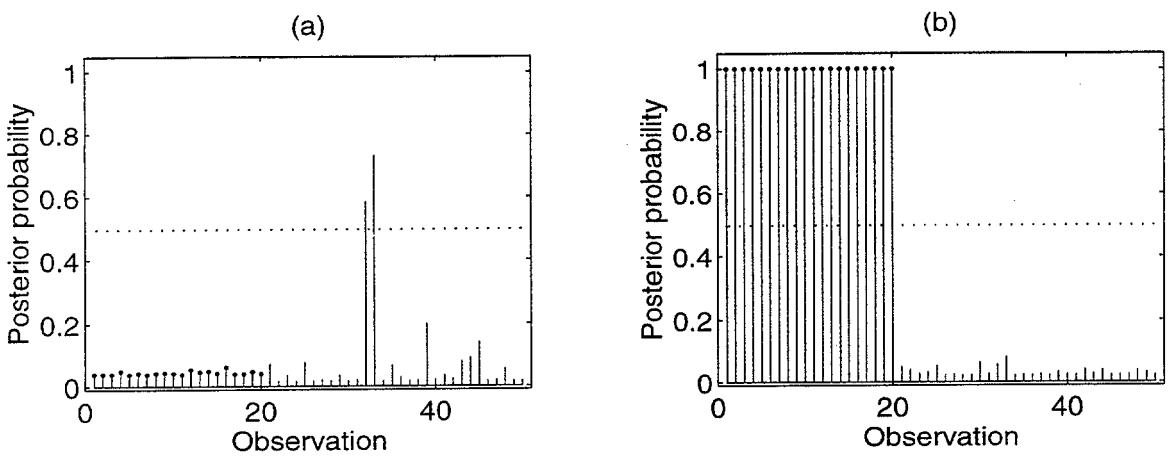

Figura 3.6: Resultados del Gibbs Sampling con los datos simulados de Rousseeuw: (a) probabilidades a posteriori de que cada dato sea atípico en la primera etapa; (b) probabilidades a posteriori en la segunda etapa.

Capítulo 4

Detección de datos atípicos en series temporales

4.1 Introducción

El análisis de series temporales está orientado a identificar la estructura de dependencia temporal que rige el fenómeno objeto del estudio. Muy frecuentemente los datos están contaminados por observaciones atípicas o sujetos a cambios estructurales. La presencia de estas alteraciones en la estructura de dependencia puede afectar a la identificación del modelo y provocar sesgos importantes en la estimación de los parámetros. Cuando la posición y el tipo de atípico o cambio estructural son conocidos, el análisis de intervención de Box y Tiao (1975) permite reducir el sesgo en la estimación. Sin embargo, es difícil que se disponga de esta información. Para resolver el problema Chang y Tiao (1983), Chang, Tiao y Chen (1988) y Tsay (1988) proponen un procedimiento iterativo de búsqueda de atípicos, Denby y Martin (1979), Martin *et al.* (1983) y Bustos y Yohai (1986) emplean métodos robustos para la estimación, Peña (1987b, 1990, 1991) propone estadísticos diagnósticos para medir la influencia de las observaciones, y McCulloch y Tsay (1994a) identifican los atípicos y estiman las distribuciones a posteriori de los parámetros del modelo mediante Gibbs Sampling.

Entre los datos atípicos se distinguen dos tipos, el innovativo y el aditivo (véase Fox, 1972). El atípico innovativo se produce cuando existe contaminación en un factor externo, y el aditivo se debe a una contaminación en la serie observada. La existencia de datos atípicos aditivos puede producir sesgos importantes en la estimación de los parámetros del modelo, mientras que los atípicos innovativos tienen un efecto mucho menor en general. En este capítulo se trata el problema de la identificación de observaciones atípicas de tipo aditivo en procesos autorregresivos.

Cuando los atípicos no se presentan en la serie aisladamente sino que se produce una racha de observaciones atípicas, los procedimientos mencionados no siempre son capaces de identificarlos debido a que se produce un efecto de enmascaramiento, al que se añade un problema de datos buenos que son señalados como atípicos. Chen y Liu (1993) proponen un nuevo procedimiento iterativo para evitar el enmascaramiento con el que se estiman conjuntamente los parámetros del modelo y los tamaños de los valores atípicos. El problema que presenta este procedimiento es que se inicia con la estimación de los parámetros del modelo suponiendo que no existen observaciones atípicas. Al igual que sucede con datos independientes las medidas de influencia basadas en eliminar un dato (y tratarlo como faltante en series temporales) no sirven para identificar los atípicos cuando se produce enmascaramiento. Bruce y Martin (1989) proponen medir la influencia eliminando bloques de observaciones consecutivas. Aunque en series temporales la ordenación temporal reduce el número de bloques posibles, aun sigue siendo necesario un esfuerzo computacional demasiado grande. Bruce y Martin (1988) sugieren un mecanismo para reducir estos cálculos.

Aplicando el algoritmo de Gibbs Sampling estándar para resolver modelos de series temporales con datos atípicos se obtienen muy buenos resultados en la identificación de atípicos aislados. Sin embargo, en este capítulo se muestra como la presencia de rachas de atípicos aditivos en procesos autorregresivos conduce al algoritmo hacia una solución

errónea a lo largo de muchas iteraciones. Los mismos fenómenos de enmascaramiento y señalamiento se observan cuando el Gibbs Sampling presenta un comportamiento estable. Se propone un nuevo algoritmo adaptativo de Gibbs Sampling que proporciona: (1) un método para identificar la posición y magnitud de los datos atípicos; y (2) una muestra de la distribución a posteriori de los parámetros del proceso autorregresivo y la varianza de las perturbaciones. La decisión de si una observación es atípica se basa en las distribuciones a posteriori de las variables de clasificación que se estiman mediante Gibbs Sampling.

Este capítulo se organiza del siguiente modo. En la sección 4.2 se discute la aplicación del Gibbs Sampling estándar en la detección de datos atípicos en procesos autorregresivos, tanto para atípicos aislados, como para rachas de atípicos. En la sección 4.3 se propone un nuevo algoritmo adaptativo de Gibbs Sampling que permite identificar rachas de atípicos. La ejecución del algoritmo requiere la obtención de las distribuciones condicionadas para bloques de observaciones que se resuelve en esta misma sección. El procedimiento se ilustra con un ejemplo.

4.2 Detección de datos atípicos en procesos autorregresivos

4.2.1 Modelo autorregresivo con datos atípicos

Sea $\{x_t\}$ un proceso autorregresivo de orden p y supongamos que se observan $y_1, \ldots, y_n$ tales que

$$y_t = \delta_t \beta_t + x_t \qquad t = 1, \ldots, n$$

siendo $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)'$ un vector de variables binarias; $\delta_t = 1$ si la observación t es un dato atípico y $\delta_t = 0$ si la observación es buena, entendiéndose por buena que ha sido

generada por el proceso autorregresivo $\{x_t\}$. Si el dato en el instante t es atípico, β_t es su tamaño o magnitud. A priori parece lógico suponer que todas las observaciones tienen la misma probabilidad de ser atípicas, así

$$P(\delta_t = 1) = \alpha \qquad t = 1, \ldots, n.$$

Suponiendo que $x_1, \ldots, x_p$ son fijas y que entre las primeras p observaciones no hay datos atípicos, de manera que $x_t = y_t$ para $t = 1, \ldots, p$, el proceso autorregresivo $\{x_t\}$ se puede expresar como

$$x_t = \phi_0 + \sum_{i=1}^{p} \phi_i x_{t-i} + a_t \qquad t = p+1, \ldots, n,$$

donde las perturbaciones $\{a_t\}$ son un proceso de ruido blanco y la distribución de a_t es $N(0, \sigma_a^2)$. Si se define el vector $\boldsymbol{X}_t = (1, x_{t-1}, \ldots, x_{t-p})'$, la serie observada se puede expresar como un modelo de regresión múltiple de la forma

$$y_t = \delta_t \beta_t + \boldsymbol{\phi}' \boldsymbol{X}_t + a_t \qquad t = p+1, \ldots, n. \tag{4.1}$$

Abraham y Box (1979) obtienen las distribuciones a posteriori de los parámetros del modelo (4.1) cuando se suponen distribuciones a priori para el vector $\boldsymbol{\phi}$ y la varianza σ_a^2 independientes y no informativas

$$P(\boldsymbol{\phi}, \sigma_a^2) \propto \sigma_a^{-2}.$$

La distribución a posteriori de $\boldsymbol{\phi}$ es una mezcla de distribuciones t multivariantes, dada por

$$P(\boldsymbol{\phi} \mid \boldsymbol{y}) = \sum w_r P(\boldsymbol{\phi} \mid \boldsymbol{y}, \boldsymbol{\delta}_r), \tag{4.2}$$

donde $w_r = P(\boldsymbol{\delta}_r \mid \boldsymbol{y})$, y la suma es en las 2^n combinaciones posibles del vector $\boldsymbol{\delta}$; el vector $\boldsymbol{\delta}_r$ representa cualquier combinación de $\boldsymbol{\delta}$ que tenga exactamente r atípicos. Si

existe un atípico en el instante t la distribución a posteriori del tamaño es una mezcla de distribuciones t de Student, dada por

$$P(\boldsymbol{\beta}_t \mid \boldsymbol{y}) = \sum w_r P(\boldsymbol{\beta}_t \mid \boldsymbol{y}, \boldsymbol{\delta}_r), \tag{4.3}$$

donde la suma es también en las 2^n combinaciones del vector $\boldsymbol{\delta}$. Las distribuciones marginales a posteriori del vector $\boldsymbol{\delta}$ permiten identificar la posición de los atípicos y se obtienen a partir de

$$p_t = P(\delta_t = 1 \mid \boldsymbol{y}) = \sum w_{r_t} \qquad t = 1, \ldots, n, \tag{4.4}$$

donde la suma es en las 2^{n-1} combinaciones de $\boldsymbol{\delta}$ para las que $\delta_t = 1$.

Calcular las distribuciones a posteriori (4.2), (4.3) y (4.4) es muy costoso computacionalmente para tamaños muestrales simplemente moderados ya que son mezclas de 2^n o 2^{n-1} distribuciones.

4.2.2 Gibbs sampling para la identificación de atípicos aislados

McCulloch y Tsay (1994a) proponen calcular las distribuciones a posteriori (4.2), (4.3) y (4.4) mediante Gibbs sampling. Para su aplicación es necesario conocer todas las distribuciones condicionadas de los parámetros del modelo (4.1), condicionadas a la muestra y a los restantes parámetros. Estas distribuciones se presentan en la proposición 2 cuando se supone para el parámetro de contaminación α una distribución a priori $Beta(\gamma_1, \gamma_2)$ con media $\alpha_0 = E(\alpha) = \gamma_1/(\gamma_1 + \gamma_2)$. Los tamaños de los valores atípicos β_t se suponen a priori independientes y con distribución $N(0, \tau^2)$. La prueba de la proposición 2 se desarrolla en el apéndice final de este capítulo.

PROPOSICIÓN 2 *Sea* $\boldsymbol{y} = (y_1, \ldots, y_n)'$ *un vector de observaciones que siguen el modelo de la ecuación (4.1) con distribuciones a priori*

$$\delta_t \sim Bernoulli(\alpha) \qquad t = 1, \ldots, n$$

y

$$P(\boldsymbol{\phi}, \sigma_a^2, \boldsymbol{\beta}, \alpha) \propto \frac{\alpha^{\gamma_1 - 1}(1-\alpha)^{\gamma_2 - 1}}{\sigma_a^2} \exp\left(-\frac{1}{2\tau^2}\sum_{t=1}^{n} \beta_t^2\right),$$

donde los hiperparámetros γ_1, γ_2 *y* τ^2 *son conocidos. Las distribuciones de cada parámetro, condicionadas al resto de los parámetros y a la muestra* $\boldsymbol{y}$, *son:*

1. *La distribución del vector autorregresivo* $\boldsymbol{\phi}$ *es*

$$\boldsymbol{\phi} \mid \boldsymbol{y}, \sigma_a^2, \boldsymbol{\delta}, \boldsymbol{\beta} \sim N\left(\boldsymbol{\phi}^*, \sigma_a^2\, \boldsymbol{\Omega}_\phi\right),$$

donde la matriz $\boldsymbol{\Omega}_\phi$ *es*

$$\boldsymbol{\Omega}_\phi = \left(\sum_{t=p+1}^{n} \boldsymbol{X}_t \boldsymbol{X}_t'\right)^{-1} \tag{4.5}$$

y el vector de medias es el estimador de mínimos cuadrados de los parámetros de un proceso autorregresivo de orden p

$$\boldsymbol{\phi}^* = \boldsymbol{\Omega}_\phi \sum_{t=p+1}^{n} \boldsymbol{X}_t x_t, \tag{4.6}$$

siendo $x_t = y_t - \delta_t \beta_t$.

2. *La distribución de la precisión de las perturbaciones es*

$$\sigma_a^{-2} \mid \boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\delta}, \boldsymbol{\beta} \sim Gamma\left(\frac{n-p}{2}, \frac{1}{2}\sum_{t=p+1}^{n} a_t^2\right),$$

y, por tanto, para la varianza de las perturbaciones se tiene que

$$\frac{1}{\sigma_a^2}\sum_{t=p+1}^{n} a_t^2 \sim \chi^2_{n-p}.$$

3. *La distribución de* δ_j *es una Bernoulli con parámetro*

$$P(\delta_j = 1 \mid \boldsymbol{y}, \boldsymbol{\phi}, \sigma_a^2, \boldsymbol{\delta}_{(j)}, \boldsymbol{\beta}, \alpha) = \frac{\exp\left(-\frac{1}{2\sigma_a^2}\sum_{t=j}^{T_j}(e_t^* + \pi_{t-j}\beta_j)^2\right)}{\exp\left(-\frac{1}{2\sigma_a^2}\sum_{t=j}^{T_j}(e_t^* + \pi_{t-j}\beta_j)^2\right) + \frac{1-\alpha}{\alpha}\exp\left(-\frac{1}{2\sigma_a^2}\sum_{t=j}^{T_j} e_t^{*2}\right)}, \tag{4.7}$$

donde $e_t^\star$ *es el residuo en el instante* t *para la serie corregida por los valores atípicos ya identificados*

$$e_t^\star = x_t - \left(\phi_0 + \sum_{i=1}^{t-j-1} \phi_i x_{t-i} + \phi_{t-j} y_j + \sum_{i=t-j+1}^{p} \phi_i x_{t-i}\right) \quad si\ t \neq j$$

y

$$e_j^\star = y_j - \left(\phi_0 + \sum_{i=1}^{p} \phi_i x_{j-i}\right),$$

siendo $T_j = \min(n, j+p)$, $\pi_0 = -1$ *y* $\pi_j = \phi_j$ *para* $j = 1, \ldots, p$.

La distribución del tamaño del valor atípico es

$$\beta_j \mid \boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\delta}, \sigma_a^2 \sim N(\delta_j \beta_j^\star, \sigma_j^2), \tag{4.8}$$

con parámetros,

$$\sigma_j^2 = \frac{\tau^2 \sigma_a^2}{\tau^2 \nu_{T_j-j}^2 \delta_j + \sigma_a^2}, \tag{4.9}$$

siendo $\nu_{T_j-j}^2 = (1 + \phi_1^2 + \cdots, +\phi_{T_j-j}^2)$, *y*

$$\beta_j^\star = \frac{\sigma_j^2}{\sigma_a^2}\left(e_j^\star - \phi_1 e_{j+1}^\star - \ldots \phi_{T_j-j} e_{T_j}^\star\right). \tag{4.10}$$

4. La distribución de α *depende únicamente del vector* $\boldsymbol{\delta}$ *y es*

$$\alpha \mid \boldsymbol{\delta} \sim Beta\left(\gamma_1 + (n-p)\bar{\delta}\ ,\ \gamma_2 + (n-p)(1-\bar{\delta})\right),$$

donde $(n-p)\bar{\delta} = \sum_{t=p+1}^{n} \delta_t$ *y la media condicionada a posteriori de* α *puede escribirse como una combinación lineal de la media a priori y la media de los datos,*

$$E(\alpha \mid \boldsymbol{\delta}) = \omega E(\alpha) + (1-\omega)\bar{\delta},$$

donde $\omega = (\gamma_1 + \gamma_2)/(\gamma_1 + \gamma_2 + n - p)$.

La ecuación (4.7) tiene una interpretación muy simple, la hipótesis de que $\delta_j = 1$ (y_j es un dato atípico) dados los parámetros afecta únicamente a los residuos $e_j, \ldots, e_{T_j}$.

Suponiendo que los parámetros son conocidos el problema es: (1) calcular los residuos cuando se advierte que $\delta_j = 1$, estos son $e_t^* + \pi_{t-j}\beta_j$; (2) calcular los residuos cuando $\delta_j = 0$, estos son e_t^*; y (3) compararlos. Esto es lo que hace (4.7) en la forma habitual, comparando las verosimilitudes de ambas hipótesis. La probabilidad (4.7) se puede escribir como

$$P(\delta_j = 1 \mid \boldsymbol{y}, \boldsymbol{\phi}, \sigma_a^2, \boldsymbol{\delta}_{(j)}, \boldsymbol{\beta}, \alpha) = \left(1 + \frac{(1-\alpha)}{\alpha} F_{10}(j)\right)^{-1}, \tag{4.11}$$

donde F_{10} es el factor de Bayes, que para la observación j-ésima es

$$F_{10}(j) = \frac{f(\boldsymbol{y} \mid \theta_{\delta_j}; \delta_j = 0)}{f(\boldsymbol{y} \mid \theta_{\delta_j}; \delta_j = 1)}. \tag{4.12}$$

Tomando logaritmos y teniendo en cuenta (4.23) y (4.24) se tiene que

$$\log F_{10}^2(j) = \frac{1}{\sigma_a^2}\left(\sum_{t=j}^{T_j}(e_t^* + \pi_{t-j}\beta_j)^2 - \sum_{t=j}^{T_j} e_t^{*2}\right). \tag{4.13}$$

Como los residuos son el error de predicción un periodo adelante, se puede decir que (4.13) equivale a prever la serie con $\delta_j = 1$, prever la serie con $\delta_j = 0$, y comparar las sumas de los errores de predicción en los periodos $j, j+1, \ldots, T_j$. Esto es equivalente al test clásico de Chow (Chow, 1960) para contrastar cambio estructural cuando la varianza es conocida. Si la observación y_j no es atípica, la suma de los errores de predicción suponiendo que y_j es un dato atípico es mayor que la suma de los errores suponiendo que no lo es y, por tanto, la probabilidad (4.11) tenderá a cero. Lo contrario ocurrirá cuando exista un valor atípico en y_j, en este caso (4.11) tenderá a uno. Cuando no existe información a priori sobre la magnitud de los valores atípicos ($\tau^2 \to \infty$) y se identifica la observación y_j como atípica, la media de la distribución a posteriori de β_j tiende a $\hat{\beta}_j$,

$$\beta_j^* \to \hat{\beta}_j = \nu_{T_j-j}^{-2}\left(e_j^* - \phi_1 e_{j+1}^* - \ldots \phi_{T_j-j} e_{T_j}^*\right),$$

que es la estimación clásica del efecto del atípico cuando los parámetros son conocidos (véase Chang, Tiao y Chen, 1988). Además, la varianza de la distribución condicionada dada por la ecuación (4.9) coincide con la varianza del estimador $\hat{\beta}_j$.

Otra forma de escribir (4.10) consiste en expresar la media condicionada a posteriori del tamaño del valor atípico como una combinación lineal de la media a priori y la estimada a partir de los datos. Esta estimación puede demostrarse (véase Peña, 1990) que es la diferencia entre el dato observado y_j y el predictor lineal $\hat{y}_j$ que minimiza el error cuadrático medio. La ecuación (4.10) se puede expresar como

$$\beta_j^* = \frac{\tau^2 \nu_{T_j-j}^2}{\tau^2 \nu_{T_j-j}^2 + \sigma_a^2}(y_j - \hat{x}_{j|n}) + \frac{\sigma_a^2}{\tau^2 \nu_{T_j-j}^2 + \sigma_a^2}\,\beta_0, \tag{4.14}$$

donde β_0 es la media a priori, en este caso $\beta_0 = 0$, y $\hat{x}_{j|n}$ es la esperanza condicionada de y_j dadas las restantes observaciones. El predictor lineal óptimo $\hat{y}_j = \hat{x}_{j|n}$ se obtiene como combinación de las p observaciones anteriores y posteriores a y_j, y viene dado por

$$\hat{x}_{j|n} = \phi_0 \nu_{T_j-j}^{-2} \tilde{\pi}_{T_j-j} - \nu_{T_j-j}^{-2} \left(\sum_{i=1}^{p} \sum_{t=0}^{T_j-j-i} \pi_t \pi_{t+i} x_{j-i} + \sum_{i=1}^{T_j-j} \sum_{t=0}^{T_j-j-i} \pi_t \pi_{t+i} x_{j+i} \right), \tag{4.15}$$

donde $\tilde{\pi}_t = 1 - \phi_1 - \ldots - \phi_t$. Cuando las observaciones están lejos del final de la serie $(j + p < n)$ el filtro (4.15) se puede escribir como

$$\hat{x}_{j|n} = \phi_0 \nu_p^{-2} \tilde{\pi}_p - (1 - \rho^{\mathrm{D}}(B)) x_j,$$

donde $\rho^{\mathrm{D}}(B) = \nu_p^{-2} \pi(B)\pi(B^{-1})$ es la función de autocorrelación del proceso dual

$$x_t^{\mathrm{D}} = \phi_0 \tilde{\pi}_p + a_t - \phi_1 a_{t-1} - \ldots \phi_p a_{t-p} \tag{4.16}$$

introducida por Cleveland (1972). En este caso, ν_p^2 es la varianza del proceso dual. En general, a partir del polinomio autorregresivo truncado $\pi_{T_j-j}(B) = (1 - \pi_1 B - \ldots \pi_{T_j-j} B^{T_j-j})$ y la varianza truncada del proceso dual $\nu_{T_j-j}^2 = (1 + \pi_1^2 + \ldots + \pi_{T_j-j}^2)$,

el estimador (4.15) se puede expresar en función de $\rho^{\mathrm{D}}_{T_j-j}(B) = \nu^{-2}_{T_j-j}\pi_p(B)\pi_{T_j-j}(B^{-1})$, que es la función de autocorrelación "truncada" del proceso dual. Por tanto, para cualquier valor de j el predictor lineal óptimo de y_j es

$$\hat{x}_{j|n} = \phi_0 \nu^{-2}_{T_j-j}\tilde{\pi}_{T_j-j} - (1 - \rho^{\mathrm{D}}_{T_j-j}(B))x_j.$$

Sea $\boldsymbol{\theta}$ el vector de todos los parámetros desconocidos del modelo

$$\boldsymbol{\theta} = (\boldsymbol{\phi}, \sigma_a^2, \delta_{p+1}, \ldots, \delta_n, \beta_{p+1}, \ldots, \beta_n, \alpha)'.$$

Ejecutando el algoritmo de Gibbs sampling con un vector de parámetros iniciales $\boldsymbol{\theta}^{(0)}$ elegido arbitrariamente, se estiman las distribuciones a posteriori de todos los parámetros generando iterativamente valores de las distribuciones condicionadas de cada uno de ellos a la muestra y a los restantes parámetros. Por los resultados probados en la proposición 2 las distribuciones a posteriori se obtienen simulando de distribuciones conocidas como son la Normal Multivariante, la Gamma Invertida o la Beta.

Cuando las series contienen únicamente valores atípicos aislados, con pocas iteraciones del Gibbs Sampling estándar se obtienen las distribuciones a posteriori de los parámetros y se identifican los valores atípicos. El problema surge cuando existe una racha de valores atípicos. En este caso se puede producir enmascaramiento de los valores centrales de la racha, e identificarse como atípicos los datos anteriores y posteriores a la racha (este fenómeno también se conoce en inglés como *smearing*). El ejemplo siguiente muestra estas dos situaciones.

Ejemplo: Serie artificial

La serie que se muestra en la figura 4.1(a) corresponde a los datos de la tabla 4.1. La serie es un AR(3) con $n = 50$ datos generados con parámetros $\boldsymbol{\phi} = (0, 2.1, -1.46, 0.336)'$

y $\sigma_a^2 = 1$ (las raíces del polinomio autorregresivo son 0.6, 0.7 y 0.8). Se incluye un dato atípico aditivo en el instante $t = 27$ de tamaño $\beta_{27} = 7$. Ejecutando el Gibbs Sampling una única secuencia de $S = 3000$ iteraciones, se estiman las probabilidades a posteriori de que cada dato sea atípico y de su tamaño con la muestra que proporcionan las $R = 1000$ últimas iteraciones. Los dos parámetros se estiman con la media muestral

$$\hat{p}_t^{(S)} = \frac{1}{R} \sum_{r=S-R+1}^{S} \delta_t^{(r)}$$

y

$$\hat{\beta}_t^{(S)} = \frac{1}{R} \sum_{r=S-R+1}^{S} \beta_t^{(r)}.$$

Los valores de los hiperparámetros que se seleccionan son $\gamma_1 = 5$, $\gamma_2 = 95$ y $\tau^2 = 9$. De esta forma la media de la distribución a priori del parámetro de contaminación es $\alpha_0 = 0.05$ y la desviación típica de la distribución a priori de los β_t es 3 veces la residual. Los resultados se muestran en la figura 4.2. El estimador de la probabilidad de que sea atípico el dato 27 es $\hat{p}_{27}^{(S)} = 1$ y el de su tamaño es $\hat{\beta}_{27}^{(S)} = 7.09$. Los estimadores de las medias a posteriori del resto de parámetros en el modelo (medias muestrales) son $\hat{\alpha}^{(S)} = 0.04$, $\hat{\boldsymbol{\phi}}^{(S)} = (-0.15, 2.05, -1.49, 0.38)'$ y $\hat{\sigma}_a^{2(S)} = 1.41$.

En la figura 4.1(b) se representa la misma serie a la que se ha añadido una racha de 4 valores atípicos a partir del instante $t = 38$. Los tamaños de los atípicos son $\beta_{38} = 20$, $\beta_{39} = 20$, $\beta_{40} = 17$ y $\beta_{41} = 15$. En la figura 4.3 se muestran los resultados de ejecutar el Gibbs Sampling durante $S = 50000$ iteraciones. Como puede observarse en la figura 4.3(a) las observaciones que están en el centro de la racha no se identifican como atípicas y si se identifican las observaciones 37 y 42 que son buenas, con probabilidades casi iguales a uno. Los estimadores de las probabilidades para los datos de los extremos de la racha son $\hat{p}_{38}^{(S)} = 1$ y $\hat{p}_{41}^{(S)} = 0.975$, pero se observa un sesgo importante en la estimación de sus tamaños $\hat{\beta}_{38}^{(S)} = 5.11$ y $\hat{\beta}_{41}^{(S)} = 5.61$ (véase la figura 4.3(b)). Además, los tamaños correspondientes a las observaciones fuera de la racha teóricamente son

t	y_t	t	y_t	t	y_t	t	y_t	t	y_t
1	-0.1274	11	-14.0664	21	4.3355	31	1.4819	41	-4.9735
2	0.5541	12	-14.4163	22	5.0138	32	-3.4934	42	-7.3584
3	-1.0973	13	-12.9740	23	5.2705	33	-8.6292	43	-9.7666
4	-3.8683	14	-9.8910	24	5.6541	34	-10.6260	44	-9.4669
5	-4.9750	15	-6.6001	25	5.5110	35	-10.1068	45	-7.8670
6	-5.8816	16	-3.3033	26	5.4403	36	-8.0770	46	-5.2135
7	-6.2186	17	-1.6699	27	7.8908	37	-5.4360	47	-2.8650
8	-7.8022	18	-1.3128	28	10.0692	38	-5.9219	48	-0.0087
9	-9.7464	19	-0.1391	29	9.9644	39	-5.5209	49	0.9087
10	-12.0397	20	1.9578	30	6.3423	40	-4.8099	50	1.9833

Tabla 4.1: Serie de datos simulados.

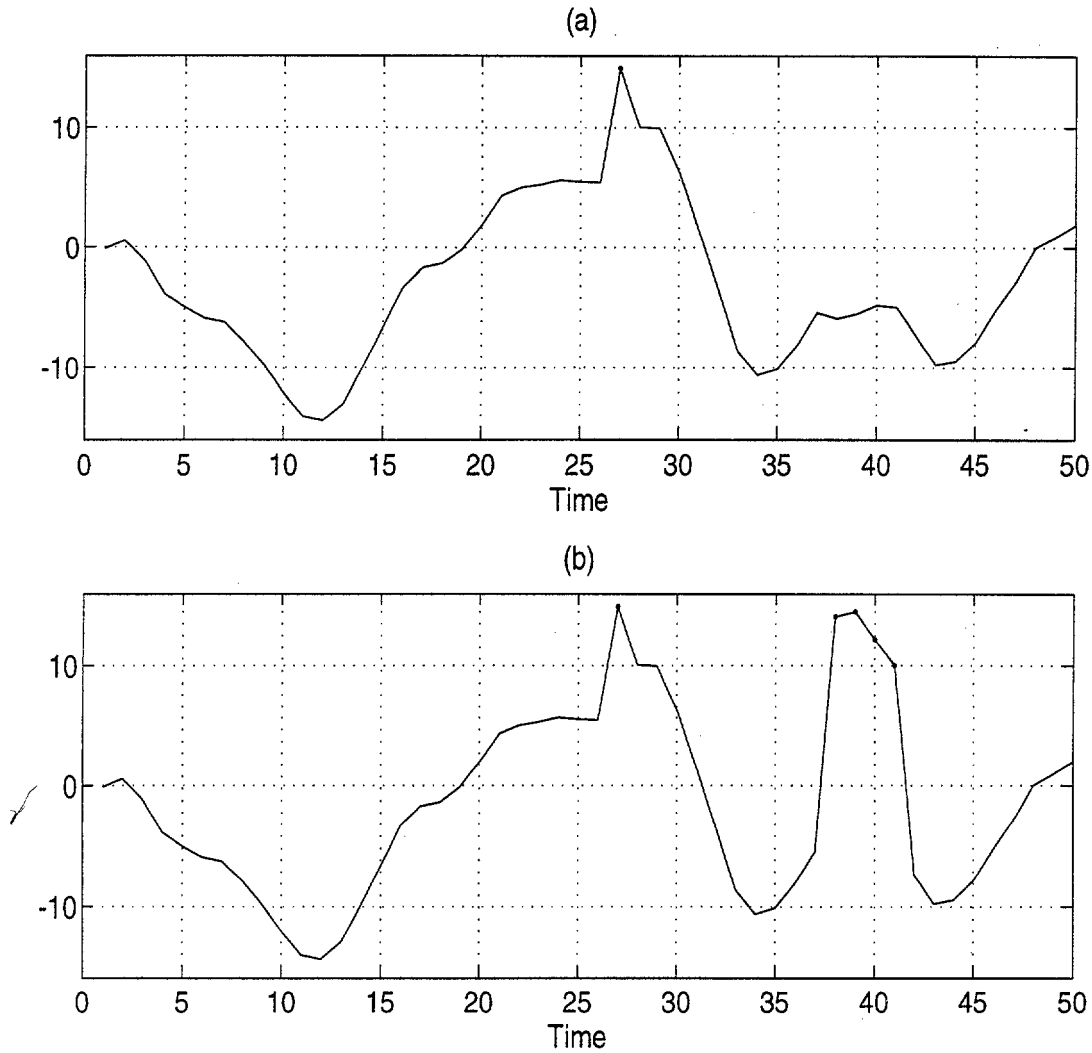

Figura 4.1: AR(3) con parámetros $\phi = (0, 2.1, -1.46, 0.336)'$ y $\sigma_a^2 = 1$: (a) con un dato atípico aditivo en $t = 27$ de tamaño $\beta_{27} = 7$; (b) con 5 datos atípicos en t =27, 38, 39, 40 y 41 de tamaños 7, 20, 20, 17 y 15 respectivamente

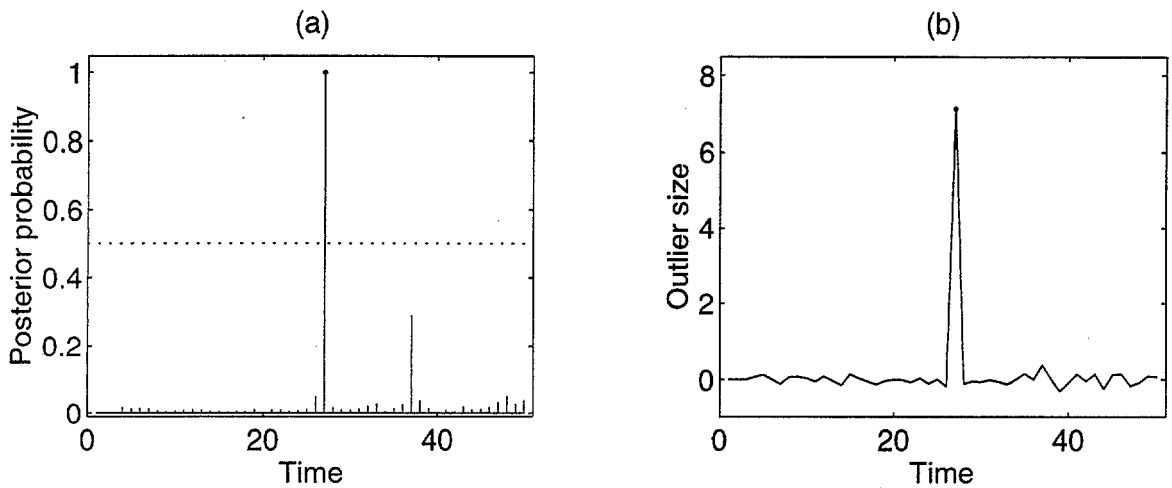

Figura 4.2: Resultados del Gibbs Sampling con la serie AR(3) con un dato atípico aditivo en $t = 27$ de tamaño $\beta_{27} = 7$: (a) probabilidades a posteriori de que cada dato sea atípico con 3000 iteraciones; (b) estimación de los tamaños de los atípicos para cada dato.

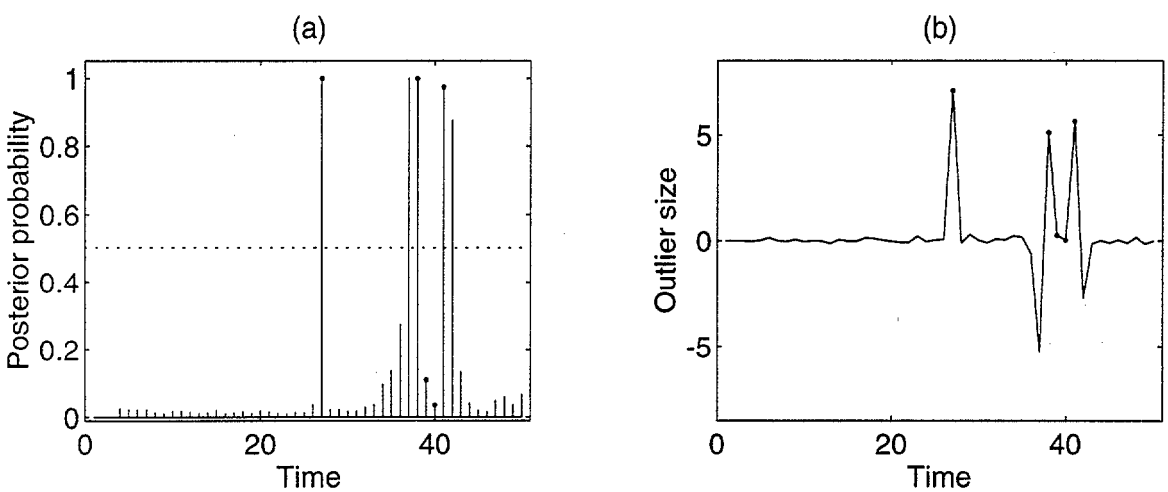

Figura 4.3: Resultados del Gibbs Sampling con la serie AR(3) con 5 datos atípicos en t =27, 38, 39, 40 y 41 de tamaños 7, 20, 20, 17 y 15 respectivamente: (a) probabilidades a posteriori de que cada dato sea atípico con 50000 iteraciones; (b) estimación de los tamaños de los atípicos para cada dato.

cero excepto para el atípico aislado y, sin embargo, los estimadores que se obtienen son $\hat{\beta}_{37}^{(s)} = -5.24$ y $\hat{\beta}_{42}^{(s)} = -2.71$. Es claro que en esta serie los datos atípicos 39 y 41 están enmascarados y los datos buenos 37 y 42 están señalados como atípicos. Los estimadores de la media a posteriori para el resto de los parámetros son $\hat{\alpha}^{(s)} = 0.07$, $\hat{\boldsymbol{\phi}}^{(s)} = (-0.09, 2.16, -1.82, 0.58)'$ y $\hat{\sigma}_a^{2(s)} = 2.15$.

4.3 Detección de rachas de atípicos

Para evitar el efecto del enmascaramiento en series temporales con rachas de atípicos Chen y Liu (1993) demuestran que es mejor estimar los tamaños de los atípicos conjuntamente. El problema es que inicialmente no se conoce la posición exacta de la racha. La solución que se propone en esta sección consiste en ejecutar el algoritmo de Gibbs Sampling estándar para determinar bloques de observaciones que pueden estar contaminadas y, posteriormente, ejecutarlo de nuevo agrupando los parámetros de clasificación y tamaño de los atípicos, de manera que en cada iteración se generan muestras de vectores de la dimensión de los bloques. La información que aporta la primera ejecución del algoritmo permite adaptar las condiciones iniciales en la segunda ejecución y modificar las distribuciones a priori del modelo. Esta modificación es necesaria para reducir el sesgo que se produce en cada iteración en la media condicionada de los tamaños de los atípicos. En efecto, cuando se aplica el Gibbs Sampling se supone que todos los atípicos tienen tamaños medios a priori iguales a cero y, por la ecuación (4.14), se tiene que la media condicionada $\boldsymbol{\beta}_j^{\star}$ es una combinación lineal de la media a priori y el estimador del tamaño. Por tanto, la especificación de la media a priori como $\beta_0 = 0$ introduce un sesgo en la media condicionada.

El nuevo algoritmo adaptativo de Gibbs Sampling consta de tres etapas:

i. En la *Etapa 1* se identifican todos los posibles datos atípicos que contiene la

muestra.

ii. En la *Etapa 2* se estiman conjuntamente los efectos de los atípicos consecutivos que se han identificado.

iii. En la *Etapa 3* se identifican las posiciones definitivas de los atípicos y se estiman conjuntamente las distribuciones a posteriori de los parámetros del proceso AR(p), la contaminación y las magnitudes de los atípicos.

La ventaja de considerar conjuntamente los parámetros de observaciones que forman parte de una racha es que cuando están muy correlacionados el Gibbs Sampling se mueve hacia la distribución a posteriori más directamente. Si se genera de la conjunta los movimientos en cada iteración son en la dirección de los ejes principales en lugar de paralelos a los ejes de coordenadas.

4.3.1 Localización de rachas de atípicos

El enmascaramiento de observaciones atípicas depende del sesgo en la media de la distribución condicionada de $\boldsymbol{\beta}$ que introducen dos factores: (1) simular los parámetros uno a uno; y (2) el sesgo que produce la media a priori.

El problema de simular las probabilidades y los tamaños uno a uno, como se hace en cada iteración del Gibbs Sampling estándar, es que las distribuciones condicionadas de las variables de clasificación y de los tamaños de los atípicos dependen de los p valores anteriores y los p posteriores a cada dato. La presencia de un atípico en la muestra afecta a las distribuciones condicionadas de $2p+1$ datos como se deduce inmediatamente de las ecuaciones (4.7) y (4.10). Chen y Liu (1993) demuestran que los estimadores de los tamaños que se obtienen separadamente pueden ser muy distintos de los que se obtienen conjuntamente, dependiendo de la estructura del proceso.

En una serie que tenga un bloque de k observaciones atípicas consecutivas, existe

posibilidad de confusión en $2p + k$ datos. El objetivo es ejecutar el Gibbs Sampling e identificar estas rachas. Una vez que se localiza una observación con una probabilidad superior a 0.5 de ser atípica, se lleva a cabo una búsqueda de observaciones con probabilidades de ser atípicas superiores a 0.5 entre los $2p$ datos anteriores y posteriores. Si en la búsqueda se identifica algún punto, estos pueden ser los extremos de una racha y enmascarar a las observaciones intermedias. Debido a los problemas de enmascaramiento y señalamiento que se pueden producir con determinadas estructuras de dependencia temporal, es conveniente rebajar este requerimiento y buscar observaciones con probabilidad superior a una determinada cota que puede ser 0.2 ó 0.3.

Cuando los parámetros del modelo y la posición de los atípicos son conocidos se puede calcular la media de la distribución conjunta a posteriori, que es el estimador de máxima verosimilitud cuando no existe información a priori sobre la magnitud de los atípicos ($\tau^2 \to \infty$). Se denota por $\tilde{\boldsymbol{\beta}}^{(S)}$ al estimador de máxima verosimilitud del vector $\boldsymbol{\beta}$ para las rachas identificadas con la información que aporta el Gibbs Sampling en S iteraciones ($\boldsymbol{\delta}^{(S)}$ y $\boldsymbol{\phi}^{(S)}$). Su expresión viene dada por la ecuación (4.22) que se deriva del teorema 2 (que se presenta en la sección siguiente). El estimador conjunto proporciona información más precisa de los tamaños de los atípicos y se utiliza para determinar las condiciones iniciales en la nueva ejecución del Gibbs Sampling.

La segunda razón por la que no funciona el Gibbs Sampling estándar en la detección de atípicos consecutivos es porque la especificación a priori de los tamaños no es apropiada para los datos atípicos. Para resolver este problema se propone modificar el modelo antes de ejecutar el Gibbs Sampling de nuevo, incorporando la información de los tamaños estimados $\tilde{\boldsymbol{\beta}}^{(S)}$ para las posibles rachas de atípicos. Esta alteración del modelo se refleja en una nueva distribución a priori que permite reducir el sesgo. Las distribuciones a priori que se proponen para los β_t del modelo (4.1) son independientes con distribuciones: (a) para observaciones que pertenecen a una racha la distribución

es $N(\tilde{\boldsymbol{\beta}}_j^{(S)}, \tau^2)$, donde $\tilde{\boldsymbol{\beta}}_j^{(S)}$ es la coordenada correspondiente a la observación j-ésima en el estimador conjunto de todos los tamaños de la racha; (b) para las observaciones atípicas aisladas la media de la distribución a priori se puede sustituir por el estimador que proporciona la primera ejecución del Gibbs Sampling $\hat{\beta}_j^{(S)}$; y (c) para las observaciones buenas no se modifica la distribución a priori.

4.3.2 Distribuciones condicionadas de bloques de observaciones

Cuando se han identificado las posibles rachas de atípicos que contiene la muestra, la forma más eficiente de ejecutar el Gibbs Sampling es generando muestras de las distribuciones conjuntas por bloques de observaciones. Supongamos que se identifica una racha de k datos atípicos que se inicia en el instante j-ésimo. Sean $\boldsymbol{\delta}_{j,k} = (\delta_j, \ldots, \delta_{j+k-1})'$ y $\boldsymbol{\beta}_{j,k} = (\beta_j, \ldots, \beta_{j+k-1})'$ los vectores que contienen las variables de clasificación y tamaño del bloque completo. En este caso, el vector de parámetros del modelo es

$$\boldsymbol{\theta}_B = (\boldsymbol{\phi}, \sigma_a^2, \delta_{p+1}, \ldots, \boldsymbol{\delta}_{j,k}, \delta_{j+k}, \ldots, \delta_n, \beta_{p+1}, \ldots, \boldsymbol{\beta}_{j,k}, \beta_{j+k}, \ldots, \beta_n, \alpha)'.$$

Para poder aplicar el Gibbs Sampling estándar es necesario conocer las distribuciones condicionadas a posteriori de todos los parámetros de $\boldsymbol{\theta}_B$. Junto con las distribuciones de la proposición 2 las distribuciones de $\boldsymbol{\delta}_{j,k}$ y $\boldsymbol{\beta}_{j,k}$ son las que se emplean en la segunda ejecución del Gibbs Sampling. Las distribuciones condicionadas de los vectores k dimensionales $\boldsymbol{\delta}_{j,k}$ y $\boldsymbol{\beta}_{j,k}$ se obtienen en los teoremas 1 y 2 que se enuncian a continuación. Las demostraciones de los teoremas se presentan en el apéndice final de este capítulo.

TEOREMA 1 *Si $\boldsymbol{y} = (y_1, \ldots, y_n)'$ es un vector de observaciones que siguen el modelo de la ecuación (4.1), y suponiendo distribuciones a priori independientes Bernoulli(α)*

para $\delta_j, \ldots, \delta_{j+k-1}$, *la distribución de* $\boldsymbol{\delta}_{j,k}$ *condicionada a la muestra y al resto de los parámetros* $\boldsymbol{\theta}_{\delta_{j,k}} = (\boldsymbol{\phi}, \boldsymbol{\delta}_{1,j-1}, \boldsymbol{\delta}_{j+k,n}, \boldsymbol{\beta}, \sigma_a^2, \alpha)'$ *es*

$$P(\boldsymbol{\delta}_{j,k} \mid \boldsymbol{y}, \boldsymbol{\theta}_{\delta_{j,k}}) = C\ \alpha^{\boldsymbol{s}_{j,k}}(1-\alpha)^{k-\boldsymbol{s}_{j,k}} \times$$

$$\times \exp\left(-\frac{1}{2\sigma_a^2}\left(\sum_{t=j}^{j+k-1}(e_t^* + \sum_{i=0}^{t-j}\pi_i\delta_{t-i}\beta_{t-i})^2 + \sum_{t=j+k}^{T_{j,k}}(e_t^* + \sum_{i=t-j-k+1}^{t-j}\pi_i\delta_{t-i}\beta_{t-i})^2\right)\right) \quad (4.17)$$

donde $\boldsymbol{s}_{j,k} = \sum_{t=j}^{j+k-1}\delta_t$, $T_{j,k} = \min\{n, j+k+p-1\}$, $\pi_0 = -1$, $\pi_i = \phi_i$ *para* $i = 1, \ldots, p$ *y* $\pi_i = 0$ *para* $i < 0$ *e* $i > p$, C *es la constante de normalización para que la suma de las probabilidades de las* 2^k *combinaciones de* $\boldsymbol{\delta}_{j,k}$ *sea 1, y* e_t^* *es el residuo en el periodo* t *para la serie corregida por los valores atípicos identificados en periodos fuera del intervalo* $[j, j+k-1]$

$$e_t^* = \begin{cases} y_t - \phi_0 - \sum_{i=1}^{t-j}\pi_i y_{t-i} - \sum_{i=t-j+1}^{p}\pi_i x_{t-i} & si\ \ t = j, \ldots, j+k-1 \\ x_t - \phi_0 - \sum_{i=1}^{t-j-k}\pi_i x_{t-i} - \sum_{i=t-j-k+1}^{t-j}\pi_i y_{t-i} - \sum_{i=t-j+1}^{p}\pi_i x_{t-i} & si\ \ t > j+k-1. \end{cases} \quad (4.18)$$

TEOREMA 2 *Si* $\boldsymbol{y} = (y_1, \ldots, y_n)'$ *es un vector de observaciones que siguen el modelo de las ecuaciones (4.1), y suponiendo distribuciones a priori independientes* $N(\beta_0, \tau^2)$ *para el vector* $\boldsymbol{\beta}_{j,k}$, *la distribución de* $\boldsymbol{\beta}_{j,k}$ *condicionada a la muestra y al resto de los parámetros es*

$$\boldsymbol{\beta}_{j,k} \mid \boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\delta}, \sigma_a^2, \alpha \sim N_k\left(\boldsymbol{\beta}_{j,k}^{\star}, \boldsymbol{\Omega}_{j,k}\right), \quad (4.19)$$

con parámetros

$$\boldsymbol{\Omega}_{j,k} = \left(\boldsymbol{D}_{j,k}\left(\frac{1}{\sigma_a^2}\sum_{t=j}^{T_{j,k}}\boldsymbol{\Pi}_{t-j}\boldsymbol{\Pi}_{t-j}'\right)\boldsymbol{D}_{j,k} + \frac{1}{\tau^2}\boldsymbol{I}\right)^{-1} \quad (4.20)$$

y

$$\boldsymbol{\beta}^{*}_{j,k} = \boldsymbol{\Omega}_{j,k} \cdot \left(-\frac{1}{\sigma_a^2} \sum_{t=j}^{T_{j,k}} e_t^{*} \boldsymbol{D}_{j,k} \boldsymbol{\Pi}_{t-j} + \frac{1}{\tau^2} \boldsymbol{\beta}_0 \right), \tag{4.21}$$

donde $T_{j,k} = \min\{n, j+k+p-1\}$, $\pi_0 = -1$, $\pi_i = \phi_i$ *para* $i = 1, \ldots, p$ *y* $\pi_i = 0$ *para* $i < 0$ *e* $i > p$. $\boldsymbol{D}_{j,k}$ *es una matriz diagonal* $k \times k$ *con elementos* $\delta_j, \ldots, \delta_{j+k-1}$ *y* $\boldsymbol{\Pi}_t$ *es un vector* $k \times 1$ *que se define como* $\boldsymbol{\Pi}_t = (\pi_t, \pi_{t-1}, \ldots, \pi_{t-k+1})'$. *Los residuos* e_t^{*} *vienen dados por la ecuación (4.18).*

Si definimos las matrices $\boldsymbol{W}_1 = \sigma_a^{-2} \cdot \boldsymbol{\Omega}_{j,k} \cdot \left(\boldsymbol{D}_{j,k} \sum_{t=j}^{T_{j,k}} \boldsymbol{\Pi}_{t-j} \boldsymbol{\Pi}'_{t-j} \boldsymbol{D}_{j,k} \right)$ y $\boldsymbol{W}_2 = \tau^{-2} \cdot \boldsymbol{\Omega}_{j,k}$ se tiene que

$$\boldsymbol{\beta}^{*}_{j,k} = \boldsymbol{W}_1 \cdot \tilde{\boldsymbol{\beta}}_{j,k} + \boldsymbol{W}_2 \cdot \boldsymbol{\beta}_0,$$

donde $\boldsymbol{W}_1 + \boldsymbol{W}_2 = \boldsymbol{I}$. Esto quiere decir que la media a posteriori de la distribución condicionada de $\boldsymbol{\beta}_{j,k}$ es una combinación lineal entre el vector de medias a priori $\boldsymbol{\beta}_0$ y el estimador clásico de máxima verosimilitud o de mínimos cuadrados generalizados del vector de tamaños de los atípicos que es

$$\tilde{\boldsymbol{\beta}}_{j,k} = \left(\boldsymbol{D}_{j,k} \sum_{t=j}^{T_{j,k}} \boldsymbol{\Pi}_{t-j} \boldsymbol{\Pi}'_{t-j} \boldsymbol{D}_{j,k} \right)^{-1} \left(-\sum_{t=j}^{T_{j,k}} e_t^{*} \boldsymbol{D}_{j,k} \boldsymbol{\Pi}_{j,k} \right). \tag{4.22}$$

Maravall y Peña (1995) demuestran que el estimador (4.22) es equivalente al vector de diferencias entre las observaciones y_t $(t = j, \ldots, j+k-1)$ y las predicciones $\hat{y}_t = E(y_t \mid y_1, \ldots, y_{j-1}, y_{j+k}, \ldots, y_n)$ cuando $\delta_t = 1$, y es cero en caso contrario. La matriz $\boldsymbol{\Pi} = \sum_{t=j}^{T_{j,k}} \boldsymbol{\Pi}_{t-j} \boldsymbol{\Pi}'_{t-j}$ es la submatriz $k \times k$ de la matriz de autocovarianzas "truncadas" del proceso dual (4.16),

$$\mathbf{\Pi} = \begin{pmatrix} \nu^2_{T_{j,k}-j} & \gamma^{\mathrm{D}}_{1,T_{j,k}-j-1} & \cdots & \gamma^{\mathrm{D}}_{k-1,T_{j,k}-j-k+1} \\ \gamma^{\mathrm{D}}_{-1,T_{j,k}-j} & \nu^2_{T_{j,k}-j-1} & \cdots & \gamma^{\mathrm{D}}_{k-2,T_{j,k}-j-k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma^{\mathrm{D}}_{-k+1,T_{j,k}-j} & \gamma^{\mathrm{D}}_{-k+2,T_{j,k}-1} & \cdots & \nu^2_{T_{j,k}-j-k+1} \end{pmatrix},$$

donde $\gamma^{\mathrm{D}}_{i,j} = \nu^2_j \rho^{\mathrm{D}}_{i,j}$, ν^2_j es la varianza "truncada" del proceso dual y $\rho^{\mathrm{D}}_{i,j}$ es el coeficiente de B^i en la función de autocorrelación "truncada" del proceso dual, $\rho^{\mathrm{D}}_j(B) = \nu_j^{-2}\pi_p(B)\pi_j(B^{-1})$.

Para simplificar la programación de las probabilidades (4.17) y las distribuciones (4.19), otra forma de escribir los residuos (4.18) es definir una nueva variable $x^\star_t$, tal que

$$x^\star_t = \begin{cases} y_t & \text{si } t = j, \ldots, j+k-1 \\ x_t & \text{si } t \neq j, \ldots, j+k-1. \end{cases}$$

Entonces, los residuos (4.17) se pueden escribir como

$$e^\star_t = x^\star_t - \phi_0 - \sum_{i=1}^{p} \phi_i x^\star_{t-i} \qquad \text{para } t > j.$$

4.3.3 Algoritmo adaptativo de Gibbs Sampling II

El algoritmo adaptativo de Gibbs Sampling para series temporales se ejecuta en tres etapas, cuya descripción se detalla a continuación:

Etapa 1: Se ejecuta el algoritmo de Gibbs Sampling con una única secuencia de S iteraciones, hasta que el comportamiento de las series resultantes sea estable. Con la muestra que proporcionan las R últimas iteraciones se estiman los parámetros $\hat{\delta}^{(\mathrm{S})}_t$ y $\hat{\beta}^{(\mathrm{S})}_t$ para $t = p+1, \ldots, n$, y el vector $\hat{\boldsymbol{\phi}}^{(\mathrm{S})}$.

Etapa 2: Se identifican las rachas de observaciones atípicas mediante los siguientes pasos:

1. Identificar las observaciones correspondientes a los instantes $t_1, \ldots, t_m$, tales que $\hat{p}_{t_i}^{(s)} > 0.5$.
2. Si existe alguna observación a una distancia k_i de t_i tal que $-2p \leq k_i \leq 2p$ y $\hat{p}_{t_i+k_i}^{(s)} > 0.3$, se forma un bloque entre y_{t_i} e $y_{t_i+k_i}$.

Se calcula el estimador de máxima verosimilitud $\tilde{\boldsymbol{\beta}}_{t_i,k_i+1}^{(s)}$ dado por la ecuación (4.22), para los instantes t_i tales que $k_i > 0$.

Se modifican las distribuciones a priori de los tamaños β_t del siguiente modo:

a. Si una observación pertenece a un bloque identificado en 2, entonces la distribución a priori es $N(\tilde{\beta}_j^{(s)}, \tau^2)$, siendo $\tilde{\boldsymbol{\beta}}_j^{(s)}$ la coordenada correspondiente a la observación j-ésima de $\tilde{\boldsymbol{\beta}}_{t_i,k_i+1}^{(s)}$.
b. Si una observaciones es atípica y está aislada ($k_i = 0$), la distribución a priori es $N(\hat{\beta}_j^{(s)}, \tau^2)$, donde $\hat{\beta}_j^{(s)}$ es el estimador de β_j que se obtiene en la primera etapa.
c. Si una observación no se identifica como atípica ni pertenece a alguna racha, la distribución a priori es $N(0, \tau^2)$.

Etapa 3: Se ejecuta el algoritmo de Gibbs Sampling con una única secuencia de S iteraciones, hasta que el comportamiento de las series resultantes sea estable. El algoritmo se ejecuta para el vector de parámetros $\boldsymbol{\theta}_B$, con tantos bloques como hayan sido identificados en la etapa 2. Las condiciones iniciales son:

1. $\delta_t^{(0)} = 1$ si $\hat{p}_t^{(s)} > 0.5$ (estimado en la etapa 1), ó si y_t pertenece a algún bloque de los identificados en la etapa 2; $\delta_t^{(0)} = 0$ en caso contrario.
2. $\beta_t^{(0)} = \tilde{\beta}_t^{(s)}$ si y_t está en algún bloque; $\beta_t^{(0)} = \hat{\beta}_t^{(s)}$ si y_t es una observación atípica aislada; y $\beta_t^{(0)} = 0$ si no ocurre ninguna de estas dos opciones.

Una diferencia importante entre este algoritmo y el que se propuso en el capítulo 3 para regresión es que no existen indicaciones para seleccionar las condiciones iniciales en la primera etapa, éstas se pueden elegir libremente. En la aplicación del Gibbs Sampling en series temporales para la resolución del modelo (4.1) no tiene sentido suponer en las condiciones iniciales que la mayoría de las observaciones son atípicas. La razón es que para cada dato que se supone atípico inicialmente hay que proponer un tamaño adecuado sobre el que no se tiene ninguna información antes de ejecutar el algoritmo. Por este mismo motivo no supone una ventaja adicional ejecutar el Gibbs Sampling en paralelo y se ha optado por ejecutar una única secuencia y utilizar las últimas $R = 1000$ iteraciones para estimar las distribuciones a posteriori de los parámetros del modelo.

El número S de iteraciones que se deben ejecutar en cada etapa se puede determinar por cualquiera de los métodos unisecuenciales para controlar la convergencia. El procedimiento que se ha seguido en este trabajo se adapta fácilmente al programa principal y se basa en el control de los estimadores de las probabilidades de que cada observación sea atípica. En cada iteración $s > 1000$ se calcula $\hat{p}_{p+1}^{(s)}, \ldots, \hat{p}_n^{(s)}$ con la muestra que proporcionan las $s - 1000$ últimas iteraciones. El criterio de parada consiste en buscar el mínimo número de iteraciones $S > 2000$ tal que, para un cierto $\epsilon > 0$, se verifica que $\mid \hat{p}_t^{(S+1)} - \hat{p}_t^{(S)} \mid < \epsilon$ para todo $t = p+1, \ldots, n$.

4.3.4 Ejemplo: Serie con una racha de atípicos

Para ilustrar el comportamiento del algoritmo adaptativo de Gibbs Sampling II se considera nuevamente la serie de la figura 4.1(b). La observación atípica individual que se encuentra en el instante $t = 27$ se detecta fácilmente, mientras que la racha que comienza en la observación $t = 38$ produce enmascaramiento y no se detecta con el

Gibbs Sampling estándar. El número de iteraciones del Gibbs Sampling que se ejecutan en las etapas 1 y 3 dependen de $\epsilon = 10^{-5}$.

Los resultados de la primera etapa del algoritmo adaptativo de Gibbs Sampling se mostraron en la figura 4.3. A partir de las probabilidades estimadas para cada dato se identifica una racha de posibles atípicos entre las observaciones 37 y 42. En la figura 4.4(a) se representan los tamaños de los atípicos en la racha $\tilde{\boldsymbol{\beta}}_{37,6}^{(s)}$, estimados conjuntamente por máxima verosimilitud. Los estimadores se obtienen eliminando los datos de la racha y considerándolos como faltantes; se calculan como la diferencia entre los valores observados y los interpolados. En la figura 4.4(b) se muestran la serie observada con trazo continuo, y la serie interpolada con trazo discontinuo.

Adaptando las condiciones iniciales se ejecuta de nuevo el Gibbs Sampling un número $S = 10700$ de iteraciones. Los estimadores de las probabilidades a posteriori de que cada dato sea atípico, $\hat{p}_t^{(s)}$, se muestran en la figura 4.5(a). El algoritmo identifica sin ninguna duda las observaciones que son atípicas y ninguna más. Los tamaños de las observaciones atípicas que se estiman se representan en la figura 4.5(a). Los valores que se obtienen son $\hat{\beta}_{27}^{(s)} = 7.28$ y $\hat{\boldsymbol{\beta}}_{37,6}^{(s)} = (-0.09, 17.35, 16.78, 15.01, 14.73, 0.02)'$. Cuando el algoritmo identifica correctamente los atípicos la varianza residual estimada disminuye, siendo al final de las iteraciones $\hat{\sigma}_a^{2(s)} = 1.16$. Los estimadores de los restantes parámetros son $\hat{\alpha}^{(s)} = 0.07$ y $\hat{\boldsymbol{\phi}}^{(s)} = (-0.12, 2.13, -1.65, 0.45)'$. En esta etapa todos los estimadores que se calculan son medias muestrales.

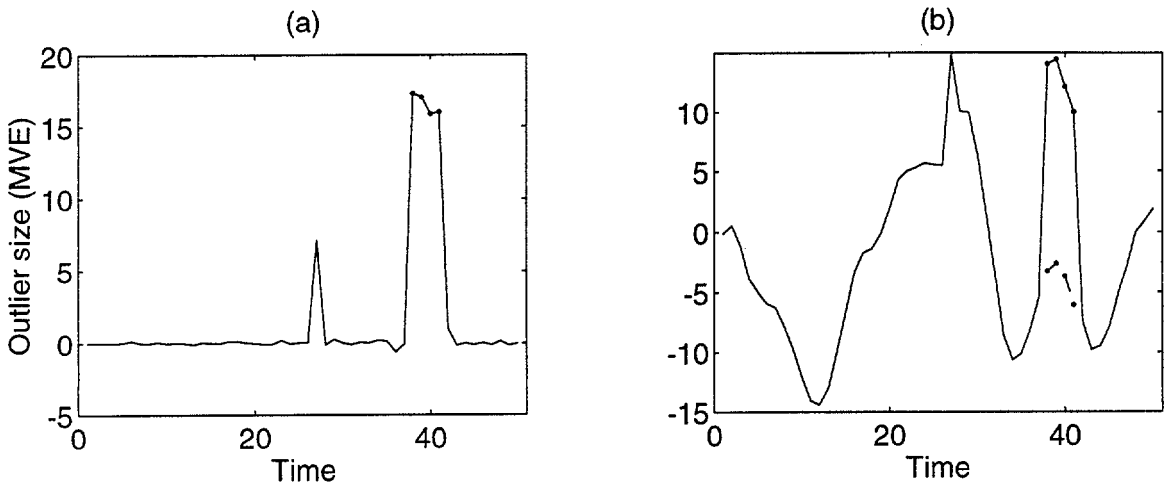

Figura 4.4: Resultados de la segunda etapa del algoritmo adaptativo de Gibbs Sampling con la serie AR(3) con 5 datos atípicos: (a) estimación máximo verosímil conjunta de los tamaños de los atípicos entre los datos 37 y 42 (el resto son los estimadores del Gibbs Sampling); (b) serie observada (trazo continuo) y serie interpolada cuando se eliminan los datos 37 a 42 (trazo discontinuo).

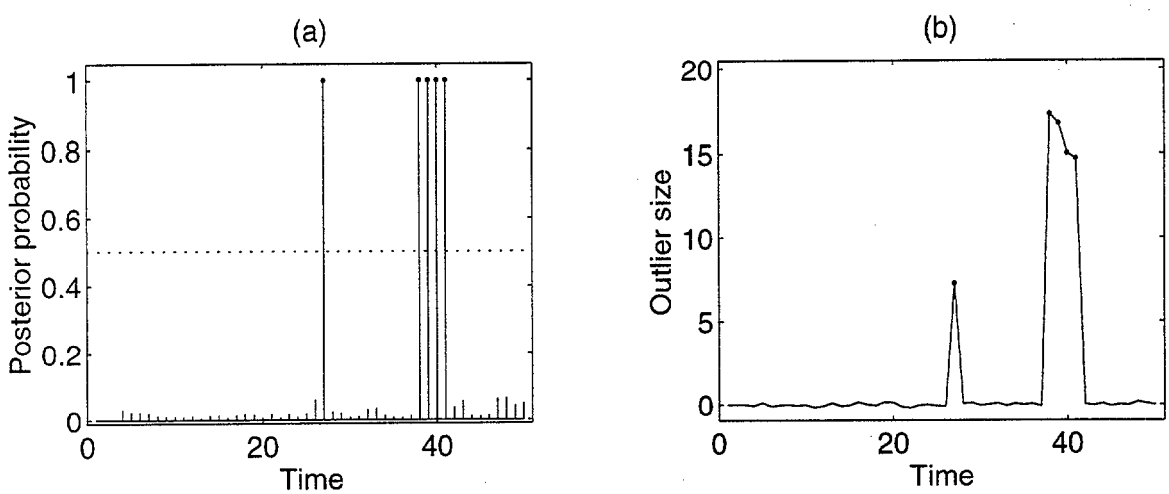

Figura 4.5: Resultados del Gibbs Sampling con la serie AR(3) con 5 datos atípicos: (a) probabilidades a posteriori de que cada dato sea atípico con 10700 iteraciones; (b) estimación de los tamaños de los atípicos para cada dato.

Apéndice

Demostración de la proposición 2. Sea $\boldsymbol{\theta}_\phi = (\boldsymbol{\delta}, \boldsymbol{\beta}, \sigma_a^2, \alpha)'$, como $P(\boldsymbol{\phi} \mid \boldsymbol{\theta}_\phi)$ es localmente uniforme se tiene que

$$P(\boldsymbol{\phi} \mid \boldsymbol{y}, \boldsymbol{\theta}_\phi) \propto f(y_{p+1} \mid \boldsymbol{\theta}_\phi; \boldsymbol{\phi}) \cdots f(y_n \mid y_{p+1}, \ldots, y_{n-1}, \boldsymbol{\theta}_\phi; \boldsymbol{\phi}) \cdot P(\boldsymbol{\phi} \mid \boldsymbol{\theta}_\phi)$$

$$\begin{aligned} &\propto \exp\left(-\frac{(y_{p+1} - \delta_{p+1}\beta_{p+1} - \boldsymbol{\phi}'\boldsymbol{X}_{p+1})^2}{2\sigma_a^2}\right) \cdots \exp\left(-\frac{(y_n - \delta_n\beta_n - \boldsymbol{\phi}'\boldsymbol{X}_n)^2}{2\sigma_a^2}\right) \\ &\propto \exp\left(-\frac{1}{2\sigma_a^2}\left(\boldsymbol{\phi}'\left(\sum_{t=p+1}^{n} \boldsymbol{X}_t\boldsymbol{X}_t'\right)\boldsymbol{\phi} - 2\boldsymbol{\phi}'\left(\sum_{t=p+1}^{n} \boldsymbol{X}_t x_t\right)\right)\right). \end{aligned}$$

Por tanto,

$$\boldsymbol{\phi} \mid \boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{\beta}, \sigma_a^2 \sim N\left(\boldsymbol{\phi}^\star, \sigma_a^2 \cdot \boldsymbol{\Omega}_\phi\right),$$

donde Ω_ϕ y $\boldsymbol{\phi}^\star$ coinciden con (4.5) y (4.6).

Sea $\boldsymbol{\theta}_{\sigma_a} = (\phi, \boldsymbol{\delta}, \boldsymbol{\beta}, \alpha)'$, entonces

$$\begin{aligned} P(\sigma_a^{-2} \mid \boldsymbol{y}, \boldsymbol{\theta}_{\sigma_a}) &\propto f(y_{p+1} \mid \boldsymbol{\theta}_{\sigma_a}; \sigma_a^{-2}) \cdots f(y_n \mid y_{p+1}, \ldots, y_{n-1}, \boldsymbol{\theta}_{\sigma_a}; \sigma_a^{-2}) \cdot P(\sigma_a^{-2} \mid \boldsymbol{\theta}_{\sigma_a}) \\ &\propto \left(\sigma_a^{-2}\right)^{\frac{n-p}{2}} \exp\left(-\frac{1}{2}\sigma_a^{-2} \sum_{t=p+1}^{n} (y_t - \delta_t\beta_t - \boldsymbol{\phi}'\boldsymbol{X}_t)^2\right)\left(\sigma_a^{-2}\right)^{-1} \end{aligned}$$

y la distribución de σ_a^{-2} es

$$\sigma_a^{-2} \mid \boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\delta}\boldsymbol{\beta} \sim Gamma\left(\frac{n-p}{2}, \frac{1}{2}\sum_{t=p+1}^{n} a_t^2\right).$$

Teniendo en cuenta que, si Z es una variable aleatoria con distribución $Gamma(n/2, m/2)$, entonces mZ tiene una distribución χ_n^2 (Johnson y Kotz, 1970, pag. 167), es inmediato comprobar que

$$\frac{1}{\sigma_a^2}\sum_{t=p+1}^{n} a_t^2 \sim \chi_{n-p}^2.$$

Sea $\boldsymbol{\theta}_{\delta_j} = (\boldsymbol{\phi}, \boldsymbol{\delta}_{(-j)}, \boldsymbol{\beta}, \sigma_a^2, \alpha)'$. Si definimos el vector $\boldsymbol{y}_j^k$ como la parte de la muestra comprendida entre los instantes j y k, $\boldsymbol{y}_j^k = (y_j, \ldots, y_k)'$, la función de verosimilitud puede factorizarse del siguiente modo:

$$f(\boldsymbol{y} \mid \theta_{\delta_j}; \delta_j) = f(\boldsymbol{y}_{p+1}^{j-1} \mid \theta_{\delta_j}) \cdot f(\boldsymbol{y}_j^{T_j} \mid \boldsymbol{y}_{p+1}^{j-1}, \theta_{\delta_j}; \delta_j) \cdot f(\boldsymbol{y}_{T_j+1}^{n} \mid \boldsymbol{y}_{p+1}^{T_j}, \theta_{\delta_j}),$$

siendo $T_j = \min(n, j+p)$. El único término que depende de δ_j es $f(\boldsymbol{y}_j^{T_j} \mid \boldsymbol{y}_{p+1}^{j-1}, \theta_{\delta_j}; \delta_j)$ que es el producto de las densidades condicionadas

$$\begin{aligned}
f(y_j \mid \boldsymbol{y}_{p+1}^{j-1}, \boldsymbol{\theta}_{\delta_j}; \delta_j) &\propto \exp\left(-\frac{1}{2\sigma_a^2}(e_j^* - \delta_j\beta_j)^2\right) \\
f(y_{j+1} \mid \boldsymbol{y}_{p+2}^{j}, \theta_{\delta_j}; \delta_j) &\propto \exp\left(-\frac{1}{2\sigma_a^2}(e_{j+1}^* + \phi_1\delta_j\beta_j)^2\right) \\
&\vdots \\
f(y_{T_j} \mid \boldsymbol{y}_{T_j-p}^{T_j-1}, \boldsymbol{\theta}_{\delta_j}; \delta_j) &\propto \exp\left(-\frac{1}{2\sigma_a^2}(e_{T_j}^* + \phi_{T_j-j}\delta_j\beta_j)^2\right).
\end{aligned}$$

Por tanto,

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\delta_j}; \delta_j = 0) \propto \exp\left(-\frac{1}{2\sigma_a^2}\sum_{t=j}^{T_j} e_t^{*2}\right) \tag{4.23}$$

y

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\delta_j}; \delta_j = 1) \propto \exp\left(-\frac{1}{2\sigma_a^2}\sum_{t=j}^{T_j}(e_t^* + \pi_{t-j}\beta_j)^2\right). \tag{4.24}$$

Aplicando el teorema de Bayes se obtiene la probabilidad (4.7).

Sea $\boldsymbol{\theta}_{\beta_j} = (\boldsymbol{\phi}, \boldsymbol{\delta}, \boldsymbol{\beta}_{(-j)}, \sigma_a^2, \alpha)'$, entonces

$$P(\beta_j \mid \boldsymbol{y}, \boldsymbol{\theta}_{\beta_j}) \propto f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\beta_j}; \beta_j) \cdot P(\beta_j \mid \boldsymbol{\theta}_{\beta_j}).$$

Si $\delta_j = 0$, la verosimilitud (4.23) no depende de β_j y, por tanto,

$$P(\beta_j \mid \boldsymbol{y}, \boldsymbol{\theta}_{\beta_j}) = P(\beta_j) = f_{\mathrm{N}}(\beta_j \mid 0, \tau^2),$$

donde $f_{\mathrm{N}}(\cdot \mid \mu, \sigma^2)$ es la función de densidad de una normal con media μ y varianza σ^2.

Si $\delta_j = 1$, la verosimilitud es (4.24) y

$$\begin{aligned} P(\beta_j \mid \boldsymbol{y}, \boldsymbol{\theta}_{\beta_j}) &\propto \exp\left(-\frac{1}{2\sigma_a^2}\sum_{t=j}^{T_j}(e_t^\star + \pi_{t-j}\beta_j)^2 - \frac{1}{2\tau^2}\beta_j^2\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\left(\frac{1}{\sigma_a^2}\sum_{t=j}^{T_j}\pi_{t-j}^2 + \frac{1}{\tau^2}\right)\beta_j^2 + \frac{2}{\sigma_a^2}\sum_{t=j}^{T_j}\pi_{t-j}\beta_j e_t^\star\right)\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{\tau^2\nu_{T_j-j}^2 + \sigma_a^2}{\sigma_a^2\tau^2}\right)\left(\beta_j + \left(\frac{\tau^2}{\tau^2\nu_{T_j-j}^2 + \sigma_a^2}\right)\sum_{t=j}^{T_j}\pi_{t-j}e_t^\star\right)^2\right). \end{aligned}$$

Por tanto, la distribución condicionada de la magnitud del dato atípico dados los restantes parámetros es

$$\beta_j \mid \boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\delta}, \sigma_a^2 \sim N(\delta_j\beta_j^\star, \sigma_j^2),$$

donde la media y la varianza coinciden con las expresiones (4.9) y (4.10).

Sea $\boldsymbol{\theta}_\alpha = (\boldsymbol{\phi}, \boldsymbol{\delta}, \boldsymbol{\beta}, \sigma_a^2)'$, como la verosimilitud no depende de α,

$$P(\alpha \mid \boldsymbol{\theta}_\alpha) \propto P(\boldsymbol{\theta}_\alpha \mid \alpha) \cdot P(\alpha) = P(\boldsymbol{\delta} \mid \alpha) \cdot P(\alpha),$$

siendo

$$P(\delta_t \mid \alpha) = \alpha^{\delta_t}(1-\alpha)^{1-\delta_t} \qquad t = 1, \ldots, n.$$

La distribución condicionada de α es

$$\begin{aligned} P(\alpha \mid \boldsymbol{\delta}) &\propto \alpha^{\gamma_1 - 1}(1-\alpha)^{\gamma_2 - 1}\prod_{t=p+1}^{n}\alpha^{\delta_t}(1-\alpha)^{1-\delta_t} \\ &\propto \alpha^{\gamma_1 + (n-p)\bar{\delta} - 1}(1-\alpha)^{\gamma_2 + (n-p)(1-\bar{\delta}) - 1}. \end{aligned}$$

y, por tanto,

$$\alpha \mid \boldsymbol{\delta} \sim Beta\left(\gamma_1 + (n-p)\bar{\delta}, \gamma_2 + (n-p)(1-\bar{\delta})\right).$$

□

Demostración del teorema 1. La distribución del vector $\boldsymbol{\delta}_{j,k}$ condicionada a la muestra y al resto de los parámetros es

$$P(\boldsymbol{\delta}_{j,k} \mid \boldsymbol{y}, \boldsymbol{\theta}_{\delta_{j,k}}) \propto f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) \cdot \alpha^{\boldsymbol{s}_{j,k}}(1-\alpha)^{k-\boldsymbol{s}_{j,k}}, \tag{4.25}$$

donde la función de verosimilitud se puede factorizar como

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) = f(\boldsymbol{y}_{p+1}^{j-1} \mid \boldsymbol{\theta}_{\delta_{j,k}}) \cdot f(\boldsymbol{y}_{j}^{T_{j,k}} \mid \boldsymbol{y}_{p+1}^{j-1}, \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) \cdot f(\boldsymbol{y}_{T_{j,k}+1}^{n} \mid \boldsymbol{y}_{p+1}^{T_{j,k}}, \boldsymbol{\theta}_{\delta_{j,k}}),$$

siendo el vector $\boldsymbol{y}_j^k = (y_j, \ldots, y_k)'$. El único término que depende de $\boldsymbol{\delta}_{j,k}$ es $f(\boldsymbol{y}_j^{T_{j,k}} \mid \boldsymbol{y}_{p+1}^{j-1}, \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k})$ que es el producto de las densidades condicionadas

$$\begin{aligned}
f(y_j \mid \boldsymbol{y}_{p+1}^{j-1}, \boldsymbol{\theta}_{\delta_{j,k}}; \delta_j) &\propto \exp\left(-\frac{1}{2\sigma_a^2}(e_j^* - \delta_j\beta_j)^2\right) \\
&\vdots \\
f(y_{j+k-1} \mid \boldsymbol{y}_{p+1}^{j+k-2}, \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) &\propto \exp\left(-\frac{1}{2\sigma_a^2}(e_{j+k-1}^* - \delta_{j+k-1}\beta_{j+k-1} + \ldots + \pi_{k-1}\delta_j\beta_j)^2\right) \\
f(y_{j+k} \mid \boldsymbol{y}_{p+1}^{j+k-1}, \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) &\propto \exp\left(-\frac{1}{2\sigma_a^2}(e_{j+k}^* + \pi_1\delta_{j+k-1}\beta_{j+k-1} + \ldots + \pi_k\delta_j\beta_j)^2\right) \\
&\vdots \\
f(y_{T_{j,k}} \mid \boldsymbol{y}_{p+1}^{T_{j,k}-1}, \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) &\propto \exp\left(-\frac{1}{2\sigma_a^2}(e_{T_{j,k}}^* + \pi_{T_{j,k}-j-k+1}\delta_{j+k-1}\beta_{j+k-1} + \ldots + \right. \\
&\quad \left. + \pi_{T_{j,k}-j}\delta_j\beta_j)^2\right).
\end{aligned}$$

Por tanto, la función de verosimilitud se puede expresar como

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) \propto \exp\left(-\frac{1}{2\sigma_a^2}\left(\sum_{t=j}^{j+k-1}(e_t^* + \sum_{i=0}^{t-j}\pi_i\delta_{t-i}\beta_{t-i})^2 + \sum_{t=j+k}^{T_{j,k}}(e_t^* + \sum_{i=t-j-k+1}^{t-j}\pi_i\delta_{t-i}\beta_{t-i})^2\right)\right) \tag{4.26}$$

y sustituyendo en (4.25) se obtiene la probabilidad (4.17) para cualquier configuración del vector $\boldsymbol{\delta}_{j,k}$. □

Demostración del teorema 2. Sea $\boldsymbol{\theta}_{\beta_{j,k}} = (\boldsymbol{\phi}, \boldsymbol{\delta}, \sigma_a^2, \alpha)'$. La distribución condicionada de $\boldsymbol{\beta}_{j,k}$ es

$$P(\boldsymbol{\beta}_{j,k} \mid \boldsymbol{y}, \boldsymbol{\theta}_{\beta_{j,k}}) \propto f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\beta_{j,k}}; \boldsymbol{\beta}_{j,k}) \cdot P(\boldsymbol{\beta}_{j,k}).$$

La función de verosimilitud $f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\beta_{j,k}}; \boldsymbol{\beta}_{j,k})$ se calcula en la demostración del teorema 1 y la ecuación (4.26) se puede escribir como

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\beta_{j,k}}; \boldsymbol{\beta}_{j,k}) \propto \exp\left(-\frac{1}{2\sigma_a^2}\sum_{t=j}^{T_{j,k}}(e_t^\star + \boldsymbol{\Pi}'_{t-j}\boldsymbol{D}_{j,k}\boldsymbol{\beta}_{j,k})'(e_t^\star + \boldsymbol{\Pi}'_{t-j}\boldsymbol{D}_{j,k}\boldsymbol{\beta}_{j,k})\right).$$

Por tanto,

$$\begin{aligned}
P(\boldsymbol{\beta}_{j,k} \mid \boldsymbol{y}, \boldsymbol{\theta}_{\beta_{j,k}}) \;\propto\; & \exp\left(-\frac{1}{2\sigma_a^2}\sum_{t=j}^{T_{j,k}}(e_t^\star + \boldsymbol{\Pi}'_{t-j}\boldsymbol{D}_{j,k}\boldsymbol{\beta}_{j,k})'(e_t^\star + \boldsymbol{\Pi}'_{t-j}\boldsymbol{D}_{j,k}\boldsymbol{\beta}_{j,k})\right) \cdot \\
& \cdot \exp\left(-\frac{1}{2\tau^2}(\boldsymbol{\beta}_{j,k} - \boldsymbol{\beta}_0)'(\boldsymbol{\beta}_{j,k} - \boldsymbol{\beta}_0)\right) \\
\propto\; & \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}'_{j,k}\left(\frac{1}{\sigma_a^2}\sum_{t=j}^{T_{j,k}}\boldsymbol{D}_{j,k}\boldsymbol{\Pi}_{t-j}\boldsymbol{\Pi}'_{t-j}\boldsymbol{D}_{j,k} + \frac{1}{\tau^2}\boldsymbol{I}\right)\boldsymbol{\beta}_{j,k} - \right.\right. \\
& \left.\left. -2\left(-\frac{1}{\sigma_a^2}\sum_{t=j}^{T_{j,k}} e_t^\star\boldsymbol{\Pi}'_{t-j}\boldsymbol{D}_{j,k} + \frac{1}{\tau^2}\boldsymbol{\beta}'_0\right)\boldsymbol{\beta}_{j,k}\right)\right) \\
\propto\; & \exp\left(-\frac{1}{2}(\boldsymbol{\beta}_{j,k} - \boldsymbol{\beta}^\star_{j,k})'\boldsymbol{\Omega}^{-1}_{j,k}(\boldsymbol{\beta}_{j,k} - \boldsymbol{\beta}^\star_{j,k})\right),
\end{aligned}$$

donde $\boldsymbol{\Omega}_{j,k}$ y $\boldsymbol{\beta}^\star_{j,k}$ se definen en (4.20) y (4.21) respectivamente. □

Capítulo 5

Conclusiones

La investigación desarrollada en esta tesis doctoral se centra en la aplicación y mejora del algoritmo de Gibbs Sampling para la identificación de observaciones atípicas en regresión y series temporales.

La aplicación del Gibbs Sampling que proponen Verdinelli y Wasserman (1991) para el problema de identificación de valores atípicos en una muestra se puede extender a la identificación de atípicos en modelos de regresión. Cuando los atípicos están aislados, el Gibbs Sampling evita que se tengan que realizar los 2^n cálculos necesarios para obtener las distribuciones a posteriori de los parámetros en el modelo de contaminación de escala. Sin embargo, cuando los datos contienen observaciones atípicas que enmascaran atípicos o señalan como atípicas observaciones que lo son, los estimadores que se obtienen con la muestra que proporciona el Gibbs Sampling pueden estar muy sesgados. El análisis de otros modelos más generales que el de contaminación de escala indica que la convergencia tampoco mejora cambiando el modelo. Los ejemplos muestran que en regresión el principal problema para que el Gibbs Sampling no converja es el de las observaciones atípicas que son influyentes porque tienen un potencial alto.

El procedimiento basado en el nuevo algoritmo adaptativo de Gibbs Sampling com-

bina, en un proceso de aprendizaje que se realiza en dos etapas, el Gibbs Sampling estándar con el análisis de la matriz de covarianzas de las variables de clasificación. Los vectores propios asociados a los autovalores no nulos de la matriz estimada con Gibbs Sampling proporcionan información sobre qué datos pueden ser atípicos. El Gibbs Sampling se ejecuta en las dos etapas y las condiciones iniciales se van adaptando con la información que se incorpora en cada una de ellas. El procedimiento se puede usar de manera automática e incluye: (1) un criterio para seleccionar las condiciones iniciales cuando no se tiene información sobre la contaminación de la muestra; y (2) un método basado en la matriz de covarianzas para dividir la muestra en dos grupos, uno de ellos que contenga al menos todos los datos atípicos. La aplicación a algunos de los ejemplos más frecuentemente analizados, y en los que fallan tanto el Gibbs Sampling estándar como otros procedimientos de identificación de atípicos propuestos recientemente en la literatura, muestra la gran potencia del procedimiento para identificar observaciones atípicas enmascaradas. La aplicación del Gibbs Sampling con las condiciones iniciales que se determinan en la primera etapa del algoritmo adaptativo es suficiente en determinados conjuntos de datos para evitar el enmascaramiento. Sin embargo, cuando las observaciones atípicas son muy influyentes, o el porcentaje de contaminación es muy alto, es necesario completar las dos etapas para identificar los verdaderos datos atípicos que contiene la muestra.

En series temporales aparecen los problemas de enmascaramiento cuando se presentan varios atípicos consecutivos Es frecuente entonces que aparezcan como atípicas únicamente las observaciones iniciales y finales de la racha. El método que se propone se inscribe en el grupo de los que de forma unificada detectan atípicos, corrigen sus efectos y estiman los parámetros del modelo. El algoritmo adaptativo propuesto consta de tres etapas en las que: (1) se ejecuta el Gibbs Sampling para facilitar la aparición de las observaciones aisladas y los extremos de las rachas; (2) para cada dato tal que la

probabilidad estimada de ser atípico sea superior a 0.5 se exploran los p datos anteriores y posteriores; y (3) se ejecuta el Gibbs Sampling muestreando en cada iteración de las distribuciones multivariantes correspondientes a las variables de clasificación y los tamaños de los atípicos en la racha. La diferencia entre este algoritmo y el propuesto para regresión es que en la segunda ejecución del Gibbs Sampling se modifican además de las condiciones iniciales las distribuciones a priori de los tamaños de los atípicos, para reducir el sesgo en la media condicionada de este parámetro.

El algoritmo adaptativo para regresión se basa en una forma concreta de extraer la información de la matriz de covarianzas a través de los componentes principales. Una extensión inmediata del método es la exploración de otras formas de identificar los conjuntos de atípicos, como por ejemplo aplicar técnicas de análisis de conglomerados sobre esta matriz. Cuanto mayor sea la capacidad de discriminación del procedimiento mayor será la fiabilidad de los resultados que aporta el algoritmo adaptativo.

El modelo de series temporales considerado en esta tesis se limita a un proceso autorregresivo con valores atípicos aditivos. Se proponen las siguientes posibles líneas de investigación futuras:

1. Desarrollo de nuevos algoritmos adaptativos de Gibbs sampling para procesos ARMA(p,q). La obtención de las distribuciones condicionadas en procesos MA(q) es más complicada y se puede resolver aumentando el vector de parámetros con las q perturbaciones anteriores al primer periodo que se observa.
2. Considerar otros tipos de valores atípicos y cambios estructurales.
3. Eliminar la restricción de que no existan atípicos en las primeras p observaciones y que estas variables sean fijas. Un procedimiento consiste en aumentar el vector de parámetros con las variables no observadas del proceso p periodos antes de tomar la muestra.
4. Generalizar los métodos expuestos en esta tesis para otros modelos de series

temporales y, en particular, para modelos lineales dinámicos en el espacio de los estados.

Finalmente, es importante señalar que los algoritmos adaptativos de Gibbs Sampling desarrollados en esta tesis pueden adaptarse a la resolución de otros problemas en los que la convergencia del Gibbs Sampling estándar sea lenta. En particular, las ideas desarrolladas para adaptar las condiciones iniciales en el caso de regresión pueden ser útiles en análisis de conglomerados o en la identificación de atípicos en problemas como la regresión logística, modelos econométricos dinámicos o modelos de ecuaciones simultáneas.

Referencias

Abraham, B. y Box, G.E.P. (1978). "Linear models and spurious observations". *Applied Statistics,* 27, 131–138.

Abraham, B. y Box, G.E.P. (1979). "Bayesian analysis of some outlier problems in time series". *Biometrika,* 66, 229–236.

Albert, J.H. y Chib, S. (1993). "Bayes inference via Gibbs Sampling of autoregressive time series subject to Markov mean and variance shifts". *Journal of Business and Economic Statistics,* 11, 1–15.

Antoniak, C.E. (1974). "Mixtures of Dirichlet process with applications to nonparametric problems". *Annals of Statistics,* 2, 1152–1174.

Barnett, V. y Lewis, T. (1984). *Outliers in Statistical Data* (segunda edición). John Wiley.

Beckman, R.J. y Cook, R.D. (1983). "Outlier...s". *Technometrics,* 25, 119–163.

Belsley, D.A., Kuh, E. y Welsch, R.E. (1980). *Regression Diagnostics.* John Wiley.

Besag, J. (1974). "Spatial interaction and the statistical analysis of lattice systems" (con discusión). *Journal of the Royal Statistical Society, Serie B,* 36, 192–326.

Blackwell, D. (1973). "Discreteness of Ferguson selections". *Annals of Statistics,* 1, 356–358.

Box, G.E.P. (1980). "Sampling and Bayesian inference in scientific modelling and

robustness" (con discusión). *Journal of the Royal Statistical Society, A,* 143, 383–430.

Box, G.E.P. y Tiao, C.G. (1968). "A Bayesian approach to some outlier problems". *Biometrika,* 55, 119–129.

Box, G.E.P. y Tiao, C.G. (1973). *Bayesian Inference in Statistical Analysis.* Addison-Wesley.

Box, G.E.P. y Tiao, C.G. (1975). "Intervention analysis with applications to economic and environmental problems". *Journal of the American Statistical Association,* 70, 70–79.

Brownlee, K.A. (1965). *Statistical Theory and Methodology in Science and Engineering.* Segunda edición. New York: John Wiley.

Bruce, A.G. y Martin, D. (1989). "Leave-k-out diagnostics for time series" (con discusión). *Journal of the Royal Statistical Society, B,* 51, 363–424.

Bustos, O.H. y Yohai, V.J. (1986). "Robust estimates for ARMA models". *Journal of the American Statistical Association,* 81, 155–168.

Carlin, B.P. y Polson, N.G. (1991a). "An expected utility approach to influence diagnostics". *Journal of the American Statistical Association,* 86, 1013–1021.

Carlin, B.P. y Polson, N.G. (1991b). "Inference for nonconjugate bayesian models using the Gibbs sampler". *The Canadian Journal of Statistics,* 19, 399–405.

Carlin, B.P., Polson, N.G. y Stoffer, D.S. (1992). "A Monte Carlo approach to non-normal and nonlinear state-space modeling". *Journal of the American Statistical Association,* 87, 493–500.

Carter, C.K. y Kohn, R. (1994). "On Gibbs Sampling for state space models". *Biometrika,* 81, 541–553.

Chang, I. y Tiao, G.C. (1983). "Estimation of time series parameters in the presence of outliers". Technical Report 8, Statistics Research Center, University of Chicago.

Chang, I., Tiao, G.C. y Chen, C. (1988). "Estimation of time series parameters in the presence of outliers". *Technometrics,* 3, 193–204.

Chatterjee, S. y Hadi A.S. (1986). "Influential observations, high leverage points, and outliers in linear regression". *Statistical Science,* 1, 379–416.

Chatterjee, S. y Hadi A.S. (1988). *Sensitivity Analysis in Linear Regression.* John Wiley.

Chen, C. y Liu, L. (1993). "Joint estimation of model parameters and outlier effects in time series". *Journal of the American Statistical Association,* 88, 284–297.

Chib, S. (1992). "Bayes inference in the Tobit censored regression model". *Journal of Econometrics,* 51, 79–99.

Chow, G.C. (1960). "A test for equality between sets of observations in two linear regressions". *Econometrica,* 28, 591–605.

Cleveland, W.P. (1972). "The inverse autocorrelations of a time series and their applications". *Technometrics,* 14, 277–298.

Cook, R.D. (1977). "Detection of influential observations in linear regression". *Technometrics,* 19, 15–18.

Cook, R.D. y Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman and Hall.

Daniel, C. y Wood, F.S. (1980). *Fitting Equations to Data.* New York: John Wiley.

Denby, L. y Martin, D. (1979). "Robust estimation of the first-order autoregressive parameter". *Journal of the American Statistical Association,* 74, 140–146.

Devroye, L. (1986). *Non-uniform Random Variate Generation.* New York: Springer-Verlag.

Diebolt, J. y Robert, C.P. (1994). "Estimation of finite mixture distributions by Bayesian sampling". *Journal of the Royal Statistical Society, B,* 56, 363–375.

Eddy, W.F. (1980). Discusión del artículo de P.R. Freeman. *Bayesian Statistics 1,*

370–373. Ed. J.M. Bernardo *et al.*, Oxford University Press.

Escobar, M.D. (1994). "Estimating normal means with a Dirichlet process prior". *Journal of the American Statistical Association,* 89, 268–277.

Escobar, M.D. y West, M. (1995). "Bayesian density estimation and inference using mixtures". *Journal of the American Statistical Association,* 90, 577–588.

Ferguson, T.S. (1973). "A Bayesian analysis of some nonparametric problems". *Annals of Statistics,* 1, 209–230.

Fosdick, L.D. (1959). "Calculation of order parameters in a binary alloy by the Monte Carlo method". *Physical Review,* 116, 565–573.

Fox, A.J. (1972). "Outliers in time series". *Journal of the Royal Statistical Society, B,* 34, 350–363.

Freeman, P.R. (1980). "On the number of outliers in data from a linear model". *Bayesian Statistics 1,* 349–365. Ed. J.M. Bernardo *et al.*, Oxford University Press.

Frühwirth-Schnatter, S. (1994). "Data augmentation and dynamic linear models". *Journal of Time Series Analysis,* 15, 183–202.

Geisser, S. (1980). Discusión del artículo de G.E.P. Box. *Journal of the Royal Statistical Society, A,* 143, 416-417.

Geisser, S. (1985). "On the predicting of observables: a selective update". *Bayesian Statistics 2,* 203–230. Ed. J.M. Bernardo *et al.*, North Holland.

Gelfand, A.E., Hills, S.E., Racine-Poon, A., y Smith, A.F.M. (1990). "Illustration of bayesian inference in normal data models using Gibbs Sampling". *Journal of the American Statistical Association,* 85, 972–985.

Gelfand, A.E. y Smith, A.F.M. (1990). "Sampling-based approaches to calculating marginal densities". *Journal of the American Statistical Association,* 85, 398–409.

Gelman, A. y Rubin, D.B. (1992a). "A single series from the Gibbs sampler provides

a false sense of security". *Bayesian Statistics 4,* 625–631. Ed. J. Bernardo *et al.*, Oxford University Press.

Gelman, A. y Rubin, D.B. (1992b). "Inference from iterative simulation using multiple sequences" (con discusión). *Statistical Science,* 7, 457–511.

Geman, S. y Geman, D. (1984). "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images". *IEEE Transaction on Pattern Analysis and Machine Intelligence,* 6, 721–741.

Geweke, J. (1992). "Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments". *Bayesian Statistics 4,* 169–193. Ed. J.M. Bernardo *et al.*, Oxford University Press.

Gilks, W.R. y Wild, P. (1992). "Adaptive rejection sampling for Gibbs Sampling". *Applied Statistics,* 41, 337–348.

Guttman, I., Dutter, R. y Freeman, P.R. (1978). "Care and handing of univariate outliers in the general linear model to detect spurosity — A Bayesian approach". *Technometrics,* 20, 187–193.

Guttman, I. y Peña, D. (1985). "On robust filtering". *Journal of the American Statistical Association,* 80, 91–92.

Guttman, I. y Peña, D. (1988a). "Outliers and influence: evaluation by posteriors of parameters in the linear model". *Bayesian Statistics 3,* 631–640. Ed. J.M. Bernardo *et al.*, Oxford University Press.

Guttman, I y Peña, D. (1988b). "Bayesian approach to robustifying the Kalman filter". *Bayesian Analysis of Time Series and Dynamic Models,* 227–255. Ed. J. Spall, Marcel Dekker.

Guttman, I. y Peña, D. (1993). "A bayesian look at diagnostic in the univariate linear model". *Statistica Sinica,* 3, 367–390.

Hadi, A.S. y Simonoff, J.S. (1993). "Procedures for the identification of multiple

outliers in linear models". *Journal of the American Statistical Association,* 88, 1264–1272.

Harrison, P.J. y Stevens, C.F. (1976). "Bayesian forecasting" (con discusión). *Journal of the Royal Statistical Society, B,* 38, 205–247.

Harrison, P.J. y West, M. (1991). "Dynamic linear model diagnostics". *Biometrika,* 78, 797–808.

Hastings, W.K. (1970). "Monte-Carlo sampling methods using Markov chains and their applications". *Biometrika,* 57, 97–109.

Hawkins, D.M., Bradu, D. y Kass, G.V. (1984). "Location of several outliers in multiple regression data using elemental sets". *Technometrics,* 26, 321–323.

Hills, S.E. y Smith, A.F.M. (1991). "Parametrization issues in Bayesian inference". *Bayesian Statistics 4,* 227–246. Ed. J.M. Bernardo *et al.*, Oxford University Press.

Hotta, L.K. y Neves, M.M.C. (1992). "A brief review on tests for detection of time series outliers". *Estadística,* 44, 142–143.

Jeffreys, H. (1961). *Theory of Probability.* Tercera edición. Oxford University Press.

Johnson, W. y Geisser, S. (1983). "A predictive view of the detection and characterization of influential observations in regression analysis". *Journal of the American Statistical Association,* 78, 137–144.

Johnson, W. y Geisser, S. (1985). "Estimative influence measures for the multivariate general model". *Journal of Statistical Planning and Inference,* 11, 33–56.

Johnson, N.C. y Kotz, S. (1970). *Distributions in Statistics: Continuous univariate distributions I.* New York: John Wiley.

Justel, A. y Peña, D. (1995a). "Gibbs Sampling will fail in outlier problems with strong masking". Working Paper 95–21(5). Departamento de Estadística y Econometría, Universidad Carlos III de Madrid.

Justel, A. y Peña, D. (1995b). "Improving Gibbs Sampling for multiple outlier detection in linear models". Working Paper. Departamento de Estadística y Econometría, Universidad Carlos III de Madrid.

Justel, A., Peña, D. y Sánchez, M.J. (1993). "Grupos de atípicos en modelos econométricos". *Cuadernos Económicos de I.C.E.*, 55, 285–325.

Kalman, R.E. (1960). "A new approach to linear filtering and prediction problems". *Trans. ASME,* 82D, 35–45.

Kass, R.F., Tierney, L. y Kadane, J.B. (1989). "Approximate methods for assesing influence and sensitivity in Bayesian analysis". *Biometrika,* 76, 663–674.

Kempthorne, P.J. (1986). "Decision-theoretic measures of influence in regression". *Journal of the Royal Statistical Society, B,* 48, 370–378.

Kloek, T. y van Dijk, H.K. (1978). "Bayesian estimates of equation system parameters: an application of integration by Monte Carlo". *Econometrica,* 46, 1–20.

Kullback, S. y Leibler, R.A. (1951). "On information and sufficiency". *Annals of Mathematical Statistics,* 22, 79–86.

Liu, J., Wong, W.H. y Kong, A. (1992). "Correlation structure and convergence rate of the Gibbs sampler with various scans". Technical Report 304, Department of Statistics, University of Chicago.

Liu, J., Wong, W.H. y Kong, A. (1994). "Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes". *Biometrika,* 81, 27–40.

MacEachern, S.N. (1994). "Estimating normal means with a conjugate style Dirichlet process prior". *Communications in Statistics, Simulation and Computing,* 23, 727–741.

MacEachern, S.N. y Müller, P. (1994). "Estimating mixture of Dirichlet process models". ISDS Discussion Paper 94–11, Duke University.

Maravall, A. y Peña, D. (1995). "Missing observations and additive outliers in time series models". *Advances in Statistical Analysis and Statistical Computing.* JAI Press (en prensa).

Marriott, J.M., Ravishanker, N. y Gelfand, A.E. (1992). "Bayesian inference in stationary autoregressive models using Gibbs Sampling". Manuscrito.

Marriott, J.M., Ravishanker, N., Gelfand, A.E. y Pai, J.S. (1994). "Bayesian analysis for ARMA processes: Cmplete sampling based inference under exact likelihoods". *Bayesian Statistics and Econometrics: Essays in honour of Arnold Zellner.* Ed. D. Barry *et al.*

Martin, R.D., Samarov, A. y Vandaele, W. (1983). "Robust methods for ARIMA models". *Applied Time Series Analysis of Economic Data,* 153–169. Ed. A. Zellner, Bureau of the Census.

Matthews, P. (1993). "A slowly mixing Markov chain with implications for Gibbs sampling". *Statistics and Probability Letters,* 17, 231–236.

McCulloch, R.E. y Tsay R.S. (1993). "Bayesian inference and prediction for mean and variance shifts in autoregressive time series". *Journal of the American Statistical Association,* 88, 968–978.

McCulloch, R.E. y Tsay R.S. (1994a). "Bayesian analysis of autoregressive time series via the Gibbs sampler". *Journal of Time Series Analysis,* 15, 235–250.

McCulloch, R.E. y Tsay R.S. (1994b). "Statistical analysis of economic time series via Markov switching models". *Journal of Time Series Analysis,* 15, 523–539.

Meinhold, R.J. y Singpurwalla, N.D. (1989). "Robustification of Kalman filter models". *Journal of the American Statistical Association,* 84, 479–486.

Mengersen, K.L. y Robert, C.P. (1995) "Testing for mixtures: a Bayesian entropic approach". *Bayesian Statistics 5.* Ed. J.M. Bernardo *et al.* (en prensa).

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. y Teller, E. (1953).

"Equations of state calculations by fast computing machines". *Journal of Chemical Physics,* 21, 1087–1091.

Müeller, P. (1991). "A generic approach to posterior integration and Gibbs Sampling". Technical Report 91–09, Department of Statistics, Purdue University.

Müller, P. Erkanli, A. y West, M. (1992). "Bayesian curve fitting using multivariate normal mixtures". ISDS Discussion Paper 92–A09, Duke University.

Peña, D. (1987a). "Observaciones influyentes en modelos econométricos". *Investigaciones Económicas,* 1, 3–24.

Peña, D. (1987b). "Measuring the importance of outliers in ARIMA models". *New Perspectives in Theoretical and Applied Statistics,* 109–118. Ed. Puri *et al.*, John Wiley.

Peña, D. (1990). "Influential observations in time series". *Journal of Business & Economic Statistics,* 8, 235–241.

Peña, D. (1991). "Measuring influence in dynamic regression models". *Technometrics,* 33, 93–101.

Peña, D. (1993). *Estadística. Modelos y Métodos. Tomo II: Modelos Lineales* (segunda edición). Alianza Editorial.

Peña, D. y Guttman, I. (1993). "Comparing probabilistic methods for outlier detection in linear models". *Biometrika,* 80, 603–610.

Peña, D. y Ruiz Castillo, J. (1984). "Robust methods for building regression models". *Journal of Business and Economic Statistics,* 2, 10–20.

Peña, D. y Sánchez Albornoz, N. (1984). "Wheat prices in Spain 1857–1890: an application of the Box-Jenkins methodology". *Journal of European Economic History,* 13, 353–373.

Peña, D. y Tiao, G.C. (1992). "Bayesian robustness functions for linear models". *Bayesian Statistics 4,* 365–388. Ed. J.M. Bernardo *et al.*, Oxford University

Press.

Peña, D. y Yohai, V.J. (1995). "The detection of influential subsets in linear regression using an influence matrix". *Journal of the Royal Statistical Society, B*, 57, 145–156.

Pettit, L.I. (1990). "The conditional predictive ordinate for the normal distribution". *Journal of the Royal Statistical Society, B,* 52, 175–184.

Pettit, L.I. (1992). "Bayes factor for outliers models using the device of imaginary observations". *Journal of the American Statistical Association,* 87, 541–545.

Pettit, L.I. y Smith, A.F.M. (1985). "Outliers and influential observations in linear models". *Bayesian Statistics 2,* 473–494. Ed. J.M. Bernardo *et al.*, Oxford University Press.

Polson, N.G. (1995). "Convergence of Markov Chain Monte Carlo algorithms". *Bayesian Statistics 5.* Ed. J.M. Bernardo *et al.* (en prensa).

Raftery, A.E. y Lewis, S. (1992). "How many iterations in the Gibbs sampler?". *Bayesian Statistics 4,* 763–773. Ed. J.M. Bernardo em et al., Oxford University Press.

Ripley, B.D. (1987). *Stochastic Simulation.* John Wiley.

Ritter, C. y Tanner, M.A. (1992). "Faciliting the Gibbs sampler: the Gibbs stopper and the Griddy-Gibbs sampler". *Journal of the American Statistical Association,* 87, 861–868.

Robert, C.P. (1994). "Convergence assessments for Markov Chain Monte-Carlo methods". Working Paper. Crest, Laboratoire de Statistique, Insee, Paris.

Rousseeuw, P.J. (1984). "Least median of squares regression". *Journal of the American Statistical Association,* 79, 871–880.

Rousseeuw, P.J. y Leroy, A.M. (1987). *Robust Regression and Outlier detection.* New York: John Wiley.

Rousseeuw, P.J. y van Zomeren, B.C. (1990) "Unmaskig multivariate outliers and

leverage points" (con discusión). *Journal of the American Statistical Association,* 85, 633–651.

Rubin, D.B. (1987). Comentarios a "The calculation of posterior distributions by data argumentation" de M. Tanner y W. Wong. *Journal of the American Statistical Association,* 82, 528–550.

Rubin, D.B. (1988). "Using de SIR algorithm to simulate posterior distributions". *Bayesian Statistics 3.* Ed. J.M. Bernardo *et al.*, Oxford University Press.

Schervish, M.J. y Carlin, B.P. (1992). "On the convergence of successive substitution sampling". *Journal of Computational and Graphical Statistics,* 1, 111–127.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Londres: Chapman and Hall.

Smith, A.F.M. y Gelfand, A.E. (1992). "Bayesian statistics without tears: a sampling-resampling perspective". *The American Statistician,* 46, 84–88.

Smith, A.F.M. y Roberts, G.O. (1993). "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods" (con discusión). *Journal of the Royal Statistical Society, B,* 55, 3–24.

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

Spiegelhalter, D.J. y Smith, A.F.M. (1982). "Bayes factors for linear and log-linear models with vague prior information". *Journal of the Royal Statistical Society, B,* 44, 377–387.

Steel, M.F.J. (1991). "A Bayesian analysis of simultaneous equation models by combining recursive analytical and numerical approaches". *Journal of Econometrics,* 48, 83–117.

Tanner, M.A. y Wong, W.H. (1987). "The calculation of posterior distributions by data augmentation" (con discusión). *Journal of the American Statistical Association,*

82, 528–550.

Tierney, L. (1994). "Markov chains for exploring posterior distributions" (con discusión). *Annals of Statistics,* 22, 1701–1762.

Tierney, L. y Kadane, J.B. (1986). "Accurate aproximations for posterior moments and marginal densities". *Journal of the American Statistical Association,* 81, 82–86.

Tsay, R.S. (1988). "Outliers, level shifts, and variance change in time series". *Journal of Forecasting,* 7, 1–20.

Tukey, J.W. (1960). "A survey of sampling from contaminated distributions". *Contributions to Probability and Statistics: Volume Dedicated to Harold Hotelling,* Stanford: University Press.

Verdinelli, I. y Wasserman, L. (1991). "Bayesian analysis of outlier problems using the Gibbs sampler". *Statistics and Computing,* 1, 105–117.

West, M. (1981). "Robust sequential approximate Bayesian estimation". *Journal of the Royal Statistical Society, B,* 43, 157–166.

West, M. (1984). "Outlier models and prior distribution in Bayesian linear models". *Journal of the Royal Statistical Society, B,* 46, 431–439.

West, M. y Harrison, P.J. (1986). "Monitoring and adaptation in Bayesian forecasting models". *Journal of the American Statistical Association,* 81, 741–750.

Zellner, A. (1976). "Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms". *Journal of the American Statistical Association,* 71, 400–405.