

**UNIVERSIDAD COMPLUTENSE DE MADRID**

**FACULTAD DE CIENCIAS MATEMATICAS**  
**Departamento de Estadística e Investigación Operativa**



**SOBRE APROXIMACIONES ESTOCASTICAS  
APLICADAS A PROBLEMAS DE INFERENCIA  
ESTADISTICA CON INFORMACION PARCIAL**

**MEMORIA PARA OPTAR AL GRADO DE DOCTOR**

**PRESENTADA POR**

**Carlos Rivero Rodríguez**

Bajo la dirección del doctor:  
Teófilo Valdés Sánchez

**Madrid, 2005**

**ISBN: 84-669-1807-8**

T 24891

UNIVERSIDAD COMPLUTENSE DE MADRID  
FACULTAD DE CIENCIAS MATEMÁTICAS  
DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

Sobre Aproximaciones Estocásticas aplicadas a  
problemas de Inferencia Estadística con  
información parcial

CARLOS RIVERO RODRÍGUEZ



UNIVERSIDAD COMPLUTENSE



5314014976

Memoria para optar al grado de Doctor en  
Ciencias Matemáticas, realizada bajo la  
dirección del Dr. D. Teófilo Valdés Sánchez

Madrid, diciembre de 2000



BIBLIOTECA

618648356

125801107

**D. TEÓFILO VALDÉS SÁNCHEZ, PROFESOR DEL  
DEPARTAMENTO DE ESTADÍSTICA E  
INVESTIGACIÓN OPERATIVA DE LA  
UNIVERSIDAD COMPLUTENSE DE MADRID**

**CERTIFICA:**

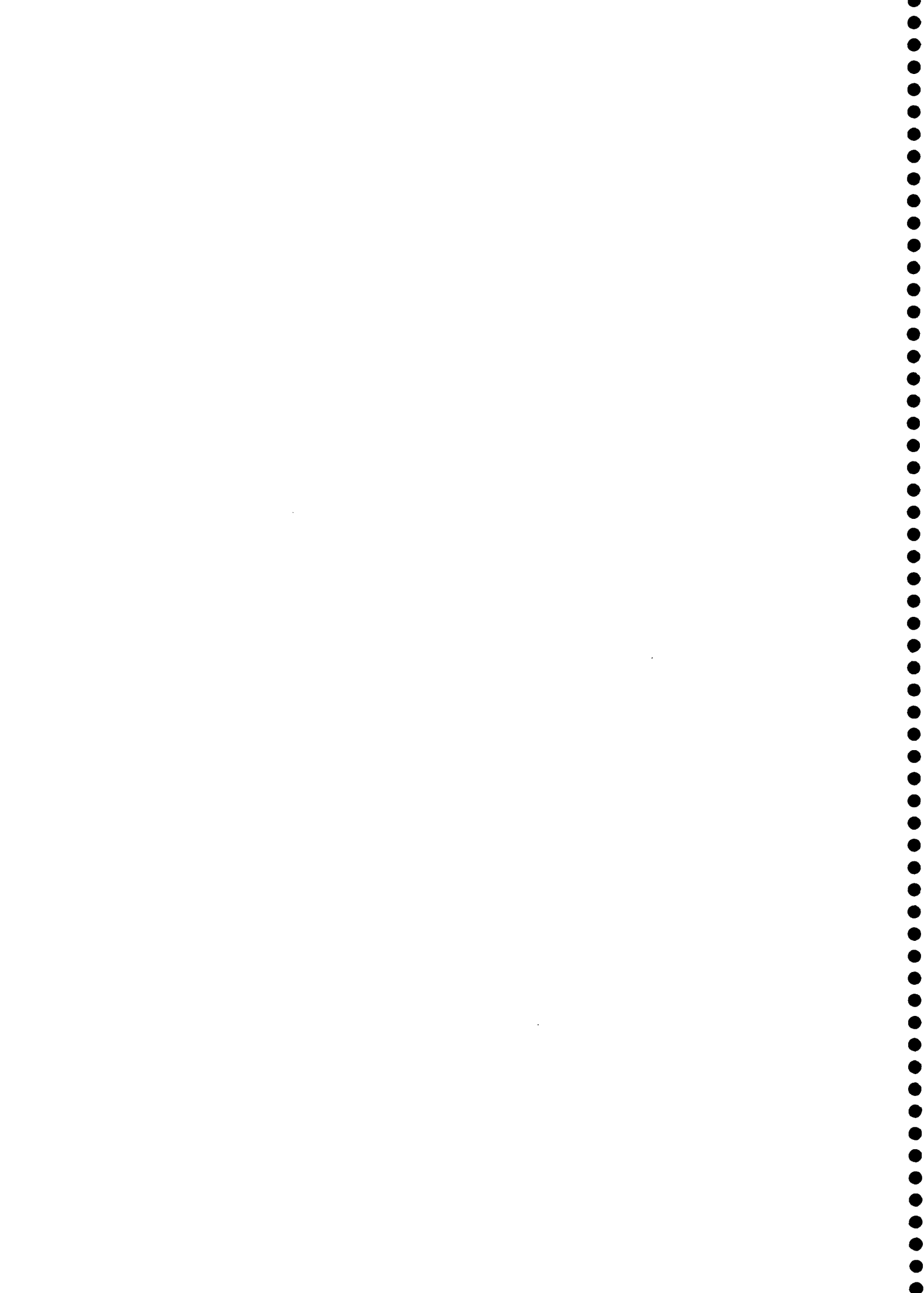
Que la presente memoria de título:

**Sobre Aproximaciones Estocásticas aplicadas a  
problemas de Inferencia Estadística con información  
parcial**

ha sido realizada bajo mi dirección por D. Carlos Rivero Rodríguez, Licenciado en Ciencias Matemáticas, y constituye su tesis para optar al título de Doctor en Ciencias Matemáticas.

Y para que así conste, en cumplimiento de la legislación vigente y a los efectos oportunos, firmo la presente en Madrid a 14 de diciembre de 2000.

*a mis abuelos, mis padres y mis hermanos*



No querría que se comenzase el desarrollo de la memoria sin reconocer que ésta no hubiera sido posible sin la atención de algunas personas a las cuales quiero agradecer la dedicación y el esfuerzo que han realizado.

Un agradecimiento especial al profesor Miguel Martín, porque si no le hubiese conocido durante el año que yo hacía el quinto curso de la licenciatura seguramente no estaría en este momento a punto de leer mi tesis y, muy probablemente, tampoco seguiría vinculado a la universidad. He de agradecerle que en aquellos días me animara a seguir este camino. Además, he de agradecerle todas sus explicaciones y las ideas que me ha aportado a lo largo de este tiempo.

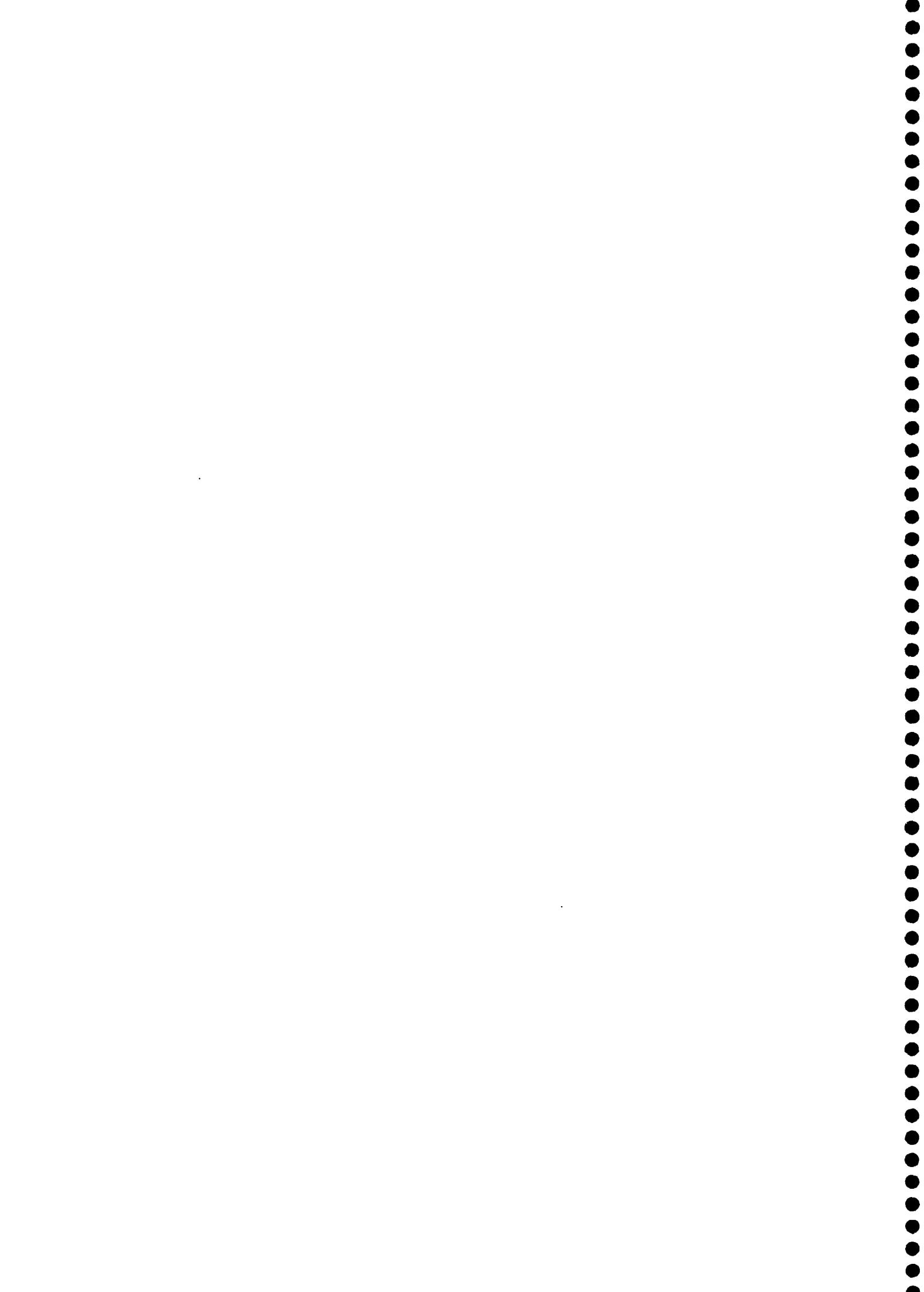
También un agradecimiento especial al profesor Teófilo Valdés por proponerme el tema de la censura, que a la postre constituye parte de la memoria, por sus ideas, por sus explicaciones y por haberme atendido durante los días finales de preparación de la tesis a pesar de su dolorosa situación familiar.

También he de agradecer de nuevo a Miguel Martín y Teófilo Valdés, así como a Pedro Zufiria, Ángela Castillo y Tomás Prieto, la ideas y las referencias que aportaron todos ellos en el seminario que formamos el año pasado. En especial a Pedro Zufiria por los múltiples temas y puntos de vista, novedosos para mí, que aportaba en esas reuniones y, sobre todo, por sacar a la luz el tema de las Aproximaciones Estocásticas.

Un agradecimiento a los compañeros del departamento, que han permitido que pudiera dedicar el tiempo necesario a la tesis.

Un agradecimiento especial a toda mi familia, por soportarme durante este tiempo.

Un agradecimiento también a todas las personas que de una manera u otra se sientan o sean partícipes de la realización de la memoria.





<b>1</b>	<b>Introducción</b>	<b>3</b>
	<b>Parte I</b>	<b>11</b>
<b>2</b>	<b>Algoritmo MD de estimación en modelos lineales con datos agrupados y correcciones en media</b>	<b>13</b>
2.1	Introducción	13
2.2	Modelo de regresión lineal, notación, algoritmo e hipótesis	14
2.3	Un primer resultado de convergencia	16
2.4	Resultados de convergencia estocástica	19
2.5	Estimación de la matriz de covarianzas asintótica	24
2.6	Simulaciones	26
<b>3</b>	<b>Algoritmo MdD de estimación en modelos lineales con datos agrupados y correcciones modales</b>	<b>35</b>
3.1	Introducción	35
3.2	Modelo de regresión lineal, notación, algoritmo e hipótesis	36
3.3	Un primer resultado de convergencia	38
3.4	Resultados de convergencia estocástica	40
3.5	Estimación de la matriz de covarianzas asintótica	46
	<b>Parte II</b>	<b>49</b>
<b>4</b>	<b>Algoritmos de una iteración basados en correcciones en media</b>	<b>51</b>
4.1	Introducción. Procesos de una iteración	51
4.2	Formalización del modelo	53
4.3	Un primer algoritmo de una iteración. Algoritmo MD de una iteración	56
4.4	Resultado de convergencia en distribución	64
4.5	Un segundo algoritmo de una iteración	68
4.6	Un tercer algoritmo de una iteración	73
	Observación sobre las hipótesis	81
4.7	Simulaciones	85
<b>5</b>	<b>Algoritmos de una iteración basados en correcciones en moda</b>	<b>105</b>
5.1	Introducción	105
5.2	Formalización del modelo	106
5.3	Un primer algoritmo de una iteración. Algoritmo MdD de una iteración	107
5.4	Resultado de convergencia en distribución	115
5.5	Un segundo algoritmo de una iteración	118

5.6	Un tercer algoritmo de una iteración	122
5.7	Sobre las hipótesis de no acotación	125
5.8	Aleatorización de los intervalos de clasificación	126
5.9	Otra forma de aleatorizar los intervalos de clasificación	130
5.10	Simulaciones	132
<b>Parte III</b>		<b>141</b>
<b>6</b>	<b>Corrección general</b>	<b>143</b>
6.1	Introducción	143
6.2	Formalización del modelo y diferentes tipos de correcciones	144
	Algunos ejemplos de correcciones	145
6.3	Método general de estimación en dos iteraciones	150
	Convergencia del proceso secundario	151
	Convergencias estocásticas del proceso iterativo primario	152
6.4	Métodos generales de estimación en una iteración	153
	Método completo	153
	Método basado en aproximaciones estocásticas	154
6.5	Observaciones y otros algoritmos	155
	Otro algoritmo alternativo al completo	155
	Sobre la corrección en media	156
	Correcciones aleatorias	156
6.6	Aleatorización de las correcciones en media	158
<b>7</b>	<b>Análisis de otras situaciones colaterales</b>	<b>163</b>
7.1	Introducción	163
7.2	Regresión no lineal. Caso unidimensional	164
	Un modelo de censura alternativo	173
7.3	Estimación paramétrica	176
	Modelo de censura alternativo	176
	Modelo de censura original	185
7.4	Regresión no lineal. Caso multidimensional	188
	Primeros resultados	189
	Un nuevo algoritmo	193
<b>8</b>	<b>Conclusiones finales</b>	<b>197</b>
	<b>Referencias</b>	<b>203</b>

# 1. Introducción

En los problemas de Inferencia Estadística es frecuente que por razones ajenas, o no, al experimento la información que el estadístico reciba sea parcial en algún sentido. A este respecto, es claro que si la información que se obtiene proviene de la medición de un fenómeno con un cierto aparato, pueda ocurrir que el valor real y preciso de dicha medición venga alterado por problemas de precisión, o bien que la medición exceda, superior o inferiormente, los límites de la escala de medibilidad del aparato. También es posible que, al recoger la información, ésta se encuentre agrupada, disponiéndose solamente de una tabla de frecuencias. En cualquiera de estos casos la información es incompleta.

En toda esta memoria se asume que la información se encuentra censurada en alguna medida. En sentido amplio, diremos que un dato es censurado si toda la información de que se dispone sobre el mismo es su pertenencia a un cierto subconjunto, pero no su valor exacto. Las dos situaciones mencionadas arriba constituyen dos ejemplos diferentes de censura. En primer lugar, el aparato de medida no podía medir por encima de una cota máxima  $c_{\max}$ , que determina su límite superior de detección.

En la segunda situación la censura podría aparecer, por ejemplo, cuando los datos provienen de diferentes fuentes. Algunas de estas fuentes pueden proporcionar los valores exactos de sus elementos muestrales. Por el contrario, otras fuentes quizás no proporcionen los valores exactos, sino solamente una agrupación ad hoc con diferentes intervalos de agrupación.

Básicamente serán estos dos tipos de censura los que se utilizarán a lo largo de la memoria. Estos dos tipos de censura son claramente diferentes, en el sentido que se explicará más adelante en esta introducción.

En cuanto a los problemas de inferencia a los que se refiere el título de la memoria, debe indicarse que se tratan de problemas de estimación paramétrica en modelos de regresión lineal y no lineal.

Ante una muestra proveniente de un modelo paramétrico, existen diferentes procedimientos muy conocidos de estimación puntual (e.g., máxima verosimilitud, momentos,...). Sin embargo, dichos métodos suelen estar concebidos para una muestra completa, no existiendo en muchos casos una manera clara de adaptarlos ante situaciones de datos incompletos o censurados. Despreciar la información censurada y quedarse sólo con la información no censurada, para después aplicar los métodos usuales a esta muestra reducida, es claramente inadecuado.

Por ejemplo, en el caso de los datos provenientes de distintas fuentes eventualmente agrupados en clases, si se desprecia la información agrupada se pueden producir sesgos y, además, se estará perdiendo una información útil. De igual forma, en el caso citado de aparatos sometidos a límites de detección  $c_{\max}$  no tiene sentido despreciar la información censurada. La utilización, pues, de esta última resulta imprescindible, no pudiendo ser arbitraria ni obviada.

Debe hacerse notar que, en todos los problemas que se tratarán en esta memoria, la información disponible constará de una parte con datos censurados y de otra con datos no censurados. La proporción de datos censurados queda libre pudiendo llegar a ser muy próxima a la unidad, en el peor de los casos.

Para introducir la técnica iterativa propuesta en esta memoria, se presentará un ejemplo simple. Supongamos que se quiere estimar la media de una normal  $N(\mu, 1)$  a partir de una muestra aleatoria, eventualmente censurada, a través del estimador habitual, es decir, la media muestral. Si la información, relativa a un

dato censurado es que su valor está en un determinado intervalo  $C$ , ¿cómo se puede utilizar dicha información? Inicialmente, se podría pensar en sustituir el valor faltante por un valor arbitrario dentro del intervalo  $C$  y calcular la media muestral de la muestra completa resultante de esa imputación. Claramente, esta sustitución por un valor arbitrario puede producir resultados nefastos. Así pues, se podría pensar, en segunda opción, sustituir el valor faltante por el valor medio de la variable truncada sobre el intervalo  $C$ . Esto, sin embargo, no es posible, puesto que, desconociéndose el valor de  $\mu$ , está indeterminada la distribución truncada.

Nuestra propuesta consiste en un proceso iterativo en el cual se parte de un valor  $\mu_0$  arbitrario para después sustituir todos los datos faltantes por la media de la variable en estudio supuesto que  $\mu_0$  sea el verdadero valor del parámetro. Una vez obtenida de esta forma la muestra completa, se estima el parámetro con ella mediante, por ejemplo, la media muestral, como se ha indicado. Así se obtiene una nueva estimación del parámetro  $\mu_1$  con el que se reitera el proceso.

Ésta es la idea básica que sustenta los algoritmos tratados en esta memoria y que puede tener su precedente en los algoritmos EM, los cuales se utilizan, como es sabido, para aproximar iterativamente los máximos de la función de verosimilitud de un modelo con datos incompletos (DEMPSTER, LAIRD Y RUBIN (1977); LITTLE Y RUBIN (1987); TANNER (1993); WU (1983); SCHMEE Y HAHN (1979)).

En el capítulo 2 se desarrolla el primer algoritmo utilizando el esquema anterior para la estimación del parámetro de un modelo lineal con datos censurados. Dicho algoritmo se identificará en lo sucesivo como algoritmo en dos iteraciones con correcciones en media. El término correcciones en media proviene de la utilización de la esperanza condicionada para sustituir los datos faltantes por un valor concreto. La razón del calificativo *dos iteraciones* se basa en la existencia de un proceso iterativo primario y un proceso iterativo secundario. El proceso

secundario consiste en la iteración que se debe llevar a cabo para obtener la estimación del parámetro cuando se fija un tamaño de muestra. Dicho proceso secundario se obtendrá a partir de un paso de proyección, propio de los modelos lineales, y de un paso de sustitución, utilizando las correcciones en media. La iteración primaria surge de la necesidad de obtener resultados asintóticos cuando el tamaño de la muestra crece. Obviamente, las propiedades de insesgadez o de mínima varianza para la estimación obtenida con un tamaño muestral fijo, no consideran el proceso recursivo primario. En esta memoria el interés estará centrado en la obtención de buenas propiedades asintóticas de los estimadores (ya sean de consistencia, convergencia en media, casi segura o en distribución a variables normales) sobre la citada iteración primaria.

A partir de este punto de arranque, la memoria presenta dos líneas fundamentales de desarrollo.

1. La primera línea consiste en simplificar los métodos de dos iteraciones (incluidos en los capítulos 2 y 3) y proponer métodos de una sola iteración (capítulos 4 y 5). El interés de esta simplificación radica en dos puntos.
  - (a) La mayor sencillez de los métodos de una iteración. De esta forma, computacionalmente al menos, se ahorrará tiempo.
  - (b) La imposibilidad de llevar, en muchas situaciones, el proceso secundario hasta el final e, incluso, de realizar un alto número de iteraciones (por razones de coste o tiempo) para obtener una buena aproximación del punto límite. Este último proporciona, en definitiva, la estimación del parámetro a tamaño de muestra fijo. Si la citada aproximación no es muy precisa es posible que las propiedades asintóticas del proceso primario demostradas en esta memoria, se alejen de aquellas que afectan realmente al valor obtenido. En todo caso, las iteraciones anidadas siempre son poco

atractivas por principio.

Los métodos de una iteración, básicamente, consisten en mantener el proceso primario y realizar una única iteración del proceso secundario (o un número reducido de ellas).

A raíz del primer método propuesto de una sola iteración surgirán dos más, igualmente basados en una iteración. Todos ellos siguen la línea de las aproximaciones estocásticas y serán muy útiles para probar las convergencias estocásticas que afectan a todas las generalizaciones ante problemas de regresión no lineal contenidas en el capítulo 7 de la memoria. Para tener una idea sobre las aproximaciones estocásticas se puede consultar KUSHNER Y YIN (1997); KUSHNER Y CLARK (1978).

2. La segunda línea de desarrollo de la memoria se centra en la sustitución de los métodos de corrección en media condicionada (capítulos 2, 4) por correcciones en moda condicionada (capítulos 3, 5), así como otros tipos de correcciones muy generales (capítulo 6). Así se intentan evitar los tediosos cálculos que pueden suponer las evaluaciones de las esperanzas condicionadas. En ocasiones, dichas esperanzas no admiten una expresión cerrada, no siendo calculables fácilmente. En resumen, es preciso utilizar métodos de cuadratura, lo cual puede suponer un aumento excesivo de tiempo de cálculo a la hora de implementar el algoritmo. Este cálculo de integrales será sustituido por la evaluación bien de las modas condicionadas citadas (que, en situaciones habituales, sólo dependen de la forma de las distribuciones involucradas) o bien de los otros tipos de correcciones más generales mencionados arriba.

Se verá que las propiedades asintóticas con este último tipo de correcciones son similares a las demostradas con correcciones en media, tanto para los algoritmos de una iteración como para los de dos iteraciones. Sin embargo, el tiempo de cálculo de los primeros será muy inferior al de los segundos. Esto se comprobará

mediante algunas simulaciones de situaciones concretas.

Estas simulaciones también servirán para comparar la eficacia de los métodos de dos iteraciones frente a los de una iteración. Adicionalmente, se usarán para mostrar numéricamente las propiedades de convergencia que teóricamente se han demostrado en esta memoria.

Como ya se ha indicado, en el capítulo 6 se defiende la posibilidad de utilizar tipos generales de corrección, e.g., corrección en mediana, sustitución por un valor intermedio del intervalo de censura, etc. Más aún, se verá que las sustituciones no tienen por qué ser necesariamente deterministas, pudiendo igualmente asignarse valores aleatorios. Con esto, se quiere mostrar la gran variedad de posibilidades que existe a la hora de imputar la información censurada. Con todas ellas se pueden conseguir buenas propiedades de convergencia estocástica y ahorro de tiempo de cálculo.

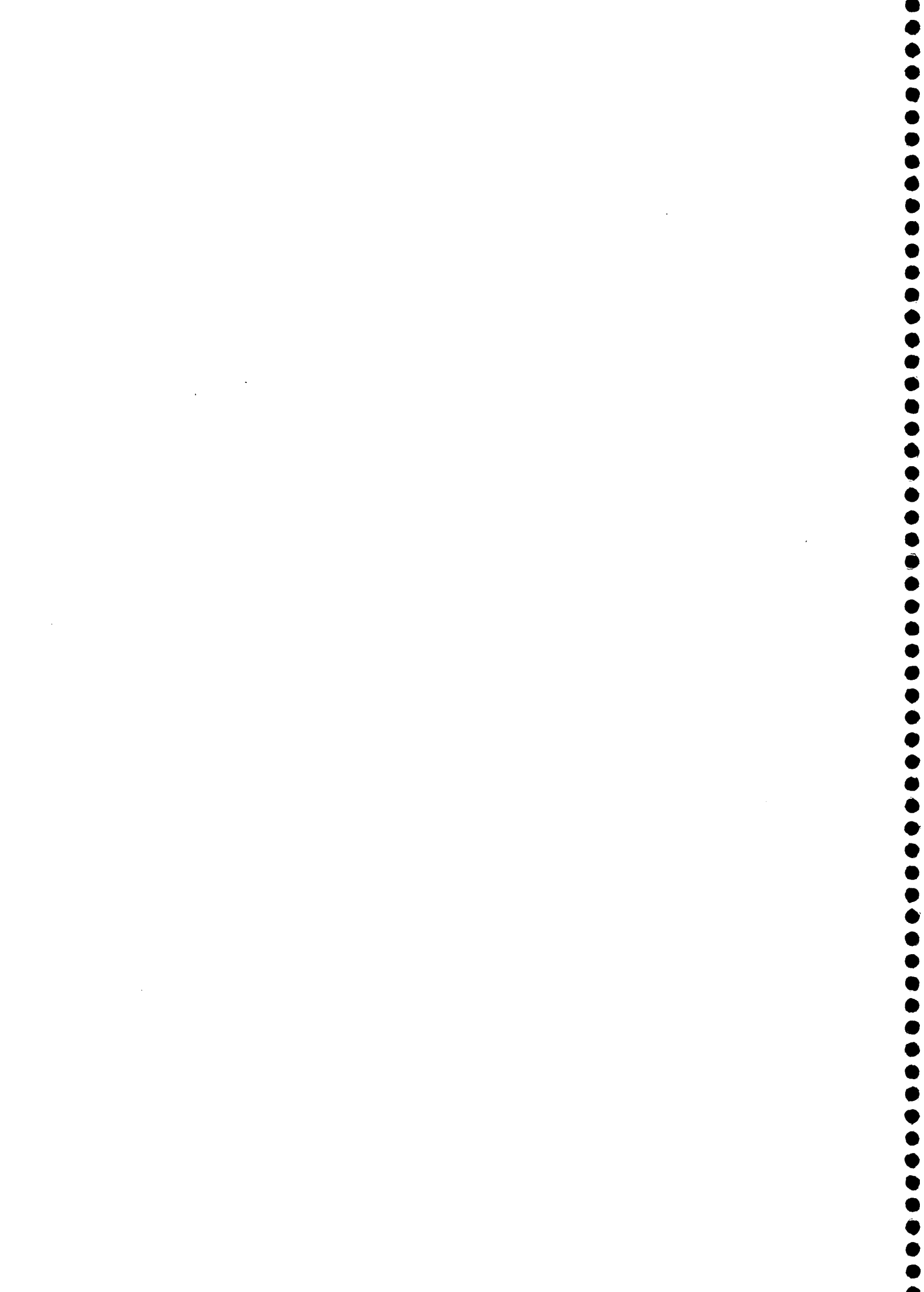
Por último, en el capítulo 7 se estudiará el problema de la estimación paramétrica en modelos no lineales. Dicha estimación se podrá llevar a cabo con cualquiera de los mencionados tipos de corrección de los datos censurados (en media, moda, mediana, etc.). Aquí lo fundamental es el tipo de técnicas de demostración empleadas. Todas ellas están basadas en aproximaciones estocásticas. La razón hay que buscarla en el hecho de que en los modelos no lineales no existe el paso claro de proyección, propio de los lineales. Las aproximaciones estocásticas no utilizan ningún tipo de proyección. Por el contrario, dependen de un tamaño de paso  $\alpha_n > 0$  decreciendo hacia cero, el cual deberá seleccionarse de forma adecuada para obtener convergencia.

En definitiva, para concluir se puede decir que los algoritmos más general entre los aquí estudiados son aquellos que se incluyen en el capítulo 7 que se está comentando. Por ser los más generales requerirán ciertas condiciones de regularidad que deberán imponerse sobre las funciones regresoras no lineales



## 1. Introducción

(además de la ya citada selección adecuada de los tamaños de paso que caracteriza las técnicas de aproximaciones estocásticas) para llegar a obtener convergencia global.



**Parte I**  
**Métodos de dos iteraciones**

## **2. Algoritmo MD de estimación en modelos lineales con datos agrupados y correcciones en media**

### **2.1 Introducción**

En este capítulo se introduce un algoritmo de estimación iterativa basado en imputaciones en media condicionada válido para ajustar modelos lineales, cuando la información es parcial. En particular, cuando algunos datos están agrupados en intervalos y asumiéndose que aquellos se extraen de diferentes fuentes. De esta forma, los intervalos de agrupación pueden variar para los diferentes elementos de la muestra. En resumen, se supondrá que los criterios de agrupación pueden ser diferentes para los distintos valores muestreados, aunque en todo caso en número finito.

El algoritmo propuesto se inspira en el EM. Éste, como es sabido, consta de un primer paso E que consiste en una evaluación en media (como el aquí propuesto), seguido de un paso M de maximización. Este último coincide con la proyección habitual por mínimos cuadrados cuando el error se distribuye normalmente. Sin embargo, el algoritmo de estimación aquí propuesto no exige la citada distribución normal para los errores. Las convergencias estocásticas demostradas en este capítulo son válidas ante una clase bastante general de distribuciones.

Se probará la convergencia del algoritmo iterativo secundario a un punto que no tiene necesariamente que coincidir con el estimador máximo verosímil del

modelo censurado, tal como ocurre con el EM. En todo caso, se demostrará que el estimador aquí propuesto tiene buenas propiedades estocásticas asintóticas, muy similares a las del propio estimador máximo verosímil. En particular, se probará que el estimador tiene distribución asintótica normal y se propondrá una estimación consistente de su matriz de covarianzas asintótica.

En resumen, la idea original del algoritmo aquí propuesto hay que buscarla en el propio EM. A este respecto, las referencias DEMPSTER, LAIRD Y RUBIN (1977), TANNER (1993) y MCLACHLAN Y KRISHNAN (1997) son obligadas como antecedentes inmediatos al trabajo aquí expuesto. Finalmente, el estudio de las convergencias estocásticas asintóticas (tanto del proceso iterativo primario como del secundario, los cuales en conjunto determinan, como se verá, el algoritmo) tiene, fundamentalmente, como antecedentes WU (1983) y MCLACHLAN Y KRISHNAN (1997).

## 2.2 Modelo de regresión lineal, notación, algoritmo e hipótesis

Considérese un modelo lineal usual, para una muestra de tamaño  $n$ , de la forma

$$z_i = a^t x_i + \nu_i, \quad i = 1, \dots, n \quad (1)$$

donde tanto el parámetro  $a$ , que debe ser estimado, como las variables independientes  $x_i$ 's son  $m$ -dimensionales, y los errores  $\nu_i$  son independientes e idénticamente distribuidos con función de densidad  $f(x) > 0, \forall x \in \mathcal{R}$ , admitiéndose que tiene varianza finita. Sin pérdida de generalidad, se supondrá que los errores tienen media cero, como es habitual, y varianza uno.

Se asumirá que la variable dependiente ha sido obtenida de diferentes fuentes, pudiendo ser bien agrupada (posiblemente con diferentes intervalos de clasificación dentro de cada fuente) o no agrupada. Así pues, el conjunto

2. Algoritmo MD de estimación en modelos lineales con datos agrupados y correcciones en media

$I = \{1, \dots, n\}$  puede partitionarse en  $I^g$  y en  $I^{ng}$ , conteniendo respectivamente aquellos  $i$ 's cuyos valores muestrales  $z_i$  han sido agrupados o no lo han sido. Más aún, como se indicó arriba, el criterio de clasificación puede variar de una fuente a otra en un número finito. En consecuencia, puede asumirse la partición  $I^g = I_1 \cup \dots \cup I_s$ , donde dentro de cada  $I_j$  los intervalos de agrupación están dados por los puntos extremos del  $j$ -ésimo criterio de agrupación o censura

$$-\infty = c_{j,0} < c_{j,1} < \dots < c_{j,r_j-1} < c_{j,r_j} = \infty,$$

con  $r_j \geq 1$ . Para cada  $i \in I^g$ , solamente se conoce el criterio de clasificación y el intervalo que contiene a  $z_i$ . Por el contrario si  $i \in I^{ng}$  se conoce el valor exacto  $z_i$ . Se asumirá, además, que valores relacionados con los conjuntos  $I^{ng}, I_1, \dots, I_s$  aparecen en la población completa con probabilidades positivas  $\pi_0, \pi_1, \dots, \pi_s$ .

Finalmente, de aquí en adelante se supondrá que  $X^t X$  es una matriz de rango completo, donde  $X^t = (x_1, \dots, x_n)$ . Se tratará de buscar una estimación del parámetro vectorial  $a$  en el modelo (1). Para ello, se sugiere un procedimiento recursivo basado en medias condicionadas, según el cual secuencialmente se imputan primero los valores agrupados (usando la información disponible hasta el momento) y después se mejora la actual estimación de  $a$  mediante mínimos cuadrados.

Fijada la muestra de tamaño  $n$ , con datos censurados y no censurados, el método iterativo propuesto es el siguiente:

INICIALIZACIÓN: Fíjese una estimación vectorial inicial arbitraria  $a_0$ .

ITERACIÓN: Asumiendo que la estimación actual conocida es  $a_p$ , la siguiente estimación viene definida por

$$a_{p+1} = (X^t X)^{-1} X^t y(a_p),$$

donde para cada  $i \in I$

$$y_i(a_p) = z_i, \quad \text{si } i \in I^{ng}$$

$$= a_p^t x_i + \gamma(-a_p^t x_i + c_{j,r}, -a_p^t x_i + c_{j,r+1}), \text{ si } i \in I_j, z_i \in (c_{j,r}, c_{j,r+1}].$$

En la última expresión, la función  $\gamma$  se define para parejas  $(t, w)$ , con  $-\infty \leq t < w \leq \infty$ , mediante

$$\gamma(t, w) = E(\nu | \nu \in (t, w)),$$

siendo  $\nu$  una variable aleatoria con función de densidad  $f$ . Se escribirán siempre intervalos de la forma  $(t, w]$  entendiéndose que si  $w = \infty$  el intervalo es en realidad de la forma  $(t, \infty)$ . En todo caso, como se trabaja con distribuciones continuas, podrían ponerse siempre intervalos abiertos. El anterior procedimiento recursivo se llamará en lo sucesivo algoritmo MD.

Denotando  $\delta(\beta) = \gamma(\beta + t, \beta + w)$ , la siguiente condición técnica se asumirá de aquí en adelante:

$$\delta(\beta) \text{ y } \beta - \delta(\beta) \text{ son no decrecientes} \quad (\text{CONDICION C1})$$

cualquiera que sea  $t < w$ . La necesidad de imponer esta condición, que afecta al comportamiento de las medias truncadas de la función de densidad  $f$ , hace que se restrinja el conjunto de densidades  $f$  sobre las cuales se pueden aplicar los resultados de convergencia demostrados en este capítulo.

OBSERVACION: Puede comprobarse fácilmente que una buena parte de las distribuciones usuales cumplen la CONDICION C1. Entre ellas, por ejemplo, la distribución normal, Laplace, logística, una distribución plana alrededor del cero y con colas normales o exponenciales, etc.

### 2.3 Un primer resultado de convergencia

Fijada una muestra de tamaño  $n$  proveniente del modelo (1), la sucesión determinista  $a_p$  generada por el algoritmo MD converge a un punto que es independiente de la estimación inicial que se proponga  $a_0$ .

**Teorema 2.1** Si  $(X^t X)^{ng} = \sum_{i \in I^{ng}} x_i x_i^t$  es una matriz definida positiva,

2. Algoritmo MD de estimación en modelos lineales con datos agrupados y correcciones en media

entonces, para cualquier vector inicial  $a_0$ , la sucesión  $a_p$  generada por el algoritmo MD converge a un vector  $a^*$  que satisface la ecuación implícita

$$a^* = (X^t X)^{-1} X^t y(a^*). \quad (2)$$

Adicionalmente, el punto  $a^*$  es la única solución de la anterior ecuación.

DEMOSTRACION: Puesto que la muestra está fijada, se puede suponer que cada  $i \in I^g$  tiene asignado un intervalo de clasificación fijo y conocido, que llamaremos  $(l_i, u_i]$ , recubriendo el dato no observado  $z_i$ . Se sigue de la primera parte de la CONDICION C1 que la función integral

$$\Gamma(\beta) = \int_0^\beta \delta(u) du$$

es convexa en  $\mathcal{R}$ . En consecuencia, la función auxiliar

$$H(a) = \frac{1}{2} \sum_{i \in I^{ng}} (z_i - a^t x_i)^2 + \sum_{i \in I^g} \Gamma(-a^t x_i)$$

es estrictamente convexa y cumple la condición de cono

$$H(a) \geq \eta \|a\| + \chi$$

para ciertos valores  $\eta > 0$  y  $\chi \in \mathcal{R}$ , puesto que  $H(a) \rightarrow \infty$  si  $\|a\| \rightarrow \infty$ . Puede asegurarse, pues, que  $H(a)$  tiene un único punto crítico  $a^*$  donde su gradiente se anula. La igualdad

$$\frac{\partial H}{\partial a}(a^*) = - \sum_{i \in I^{ng}} x_i (z_i - a^{*t} x_i) - \sum_{i \in I^g} x_i \delta(-a^{*t} x_i) = 0$$

es equivalente a (2).

Es claro que cualquier punto límite de cualquier sucesión  $\{a_p\}$  generada por el algoritmo tiene que verificar la mencionada expresión (2). Se probará ahora que siempre existe el límite de la sucesión generada por el método MD. Escribese

$$a_{p+1} - a_p = (X^t X)^{-1} X^t (y(a_p) - y(a_{p-1})).$$

Obsérvese entonces que

$$y(a) = (z^{ng}, X^g a)^t + (0, D(a) 1^g)^t,$$



donde  $X = (X^{ng}, X^g)^t$ ,  $z = (z^{ng}, z^g)^t$ ,  $D(a) = \text{diag}(d_i(a))$ , con  $d_i(a) = \gamma(-a^t x_i + l_i, -a^t x_i + u_i)$  para cada  $i \in I^g$ , y, finalmente, el vector  $1^g = (1, \dots, 1)^t$ , es de orden igual al cardinal de  $I^g$ . Así pues

$$y(a_p) - y(a_{p-1}) = (0, X^g(a_p - a_{p-1}))^t + (0, (D(a_p) - D(a_{p-1})) 1^g)^t.$$

Asumiendo  $i \in I^g$ , se sigue, a partir de la CONDICION C1, que

$$d_i(a_p) - d_i(a_{p-1}) = -m_i^{(p)}(a_p - a_{p-1})^t x_i,$$

para algún  $0 \leq m_i^{(p)} \leq 1$ . Por otro lado, hágase  $m_i^{(p)} = 1$  cuando  $i \in I^{ng}$ . Todo esto produce

$$a_{p+1} - a_p = (X^t X)^{-1} X^t (I - M^{(p)}) X (a_p - a_{p-1}),$$

donde  $M^{(p)} = \text{diag}(m_i^{(p)})$ .

Esta última igualdad permite escribir

$$\begin{aligned} & \left\| (X^t X)^{\frac{1}{2}} (a_{p+1} - a_p) \right\| \\ &= \left\| (X^t X)^{-\frac{1}{2}} (X^t X)^{-\frac{1}{2}} X^t (I - M^{(p)}) X (X^t X)^{-\frac{1}{2}} (X^t X)^{\frac{1}{2}} (a_p - a_{p-1}) \right\| \\ &\leq \tau_p \left\| (X^t X)^{\frac{1}{2}} (a_p - a_{p-1}) \right\|, \end{aligned}$$

donde  $(X^t X)^{\frac{1}{2}}$  representa la raíz cuadrada de  $X^t X$ , que es una matriz simétrica y definida positiva, y donde  $\tau_p$  es el radio espectral de la matriz simétrica  $(X^t X)^{-\frac{1}{2}} X^t (I - M^{(p)}) X (X^t X)^{-\frac{1}{2}}$  que coincide con

$$\tau_p = \max_{\|u\|=1} \left| \frac{u^t X^t (I - M^{(p)}) X u}{u^t X^t X u} \right| = \max_{\|u\|=1} \left| \frac{\sum_{i \in I^g} (1 - m_i^{(p)}) u^t x_i x_i^t u}{u^t X^t X u} \right|.$$

Recordando que  $0 \leq 1 - m_i^{(p)} \leq 1$ , la última expresión está uniformemente acotada en  $p$ , puesto que

$$\max_{\|u\|=1} \frac{\sum_{i \in I^g} u^t x_i x_i^t u}{u^t X^t X u} \leq \left( 1 + \frac{\min_{\|u\|=1} \sum_{i \in I^{ng}} u^t x_i x_i^t u}{\max_{\|u\|=1} \sum_{i \in I^g} u^t x_i x_i^t u} \right)^{-1} = \tau < 1.$$

Esto concluye la demostración y garantiza la convergencia de cualquier  $\{a_p\}$  a  $a^*$ , al menos, a una tasa de convergencia lineal.  $\square$

Este punto límite  $a^*$  de cualquier sucesión generada por el algoritmo MD definirá la estimación propuesta del parámetro de regresión. Aunque  $a^*$  no tiene que coincidir con el estimador máximo verosímil del modelo censurado, se puede visualizar como un estimador de Huber, al igual que aquél, al venir definido por una ecuación implícita.

## 2.4 Resultados de convergencia estocástica

Hasta este momento,  $n$  se ha supuesto fijo. Ahora se denotará  $a^* = a_n$ , para hacer explícita la dependencia del estimador de la muestra de tamaño  $n$ . Esta recursividad en  $n$  determina el proceso primario de iteración del algoritmo, el cual presenta en cada etapa la iteración anidada explicada en las dos secciones anteriores. Se investigará en esta sección la convergencia y las propiedades estocásticas de  $a_n$  cuando  $n \rightarrow \infty$ . De hecho, se comprobarán propiedades que asemejan  $a_n$  con los estimadores máximo verosímiles, los cuales pueden obtenerse, como es sabido, con los métodos EM.

Sea  $a$  el verdadero valor del parámetro de regresión en el modelo (1). Se asumirá en lo sucesivo una cuestión técnica que no supone en realidad pérdida en generalidad. Para cada  $j$  y  $r$ , se supondrá que  $\delta_{j,r}(\beta) = \gamma(\beta + c_{j,r}, \beta + c_{j,r+1})$  es continuamente diferenciable en un entorno de  $-a^t x_i$ , para cada  $i \in \mathcal{N}$ . Si la función de densidad  $f$  de los errores  $\nu$  es continua, esta condición, obviamente, se cumple. Por el contrario, si  $f$  tiene un conjunto numerable de saltos, la anterior hipótesis ocurre con probabilidad uno, suponiendo simplemente que la distribución subyacente de las variables independientes  $x_i$  es continua.

La primera propiedad del estimador  $a_n$  es la de ser consistente, es decir, converge en probabilidad al verdadero valor  $a$ . De hecho, la convergencia probada en el siguiente teorema es en media de orden 2.

**Teorema 2.2** *Si  $(X^t X)^{ng}$  es definida positiva y se cumplen las dos condiciones*

técnicas siguientes

$$\inf_n \lambda_n = \lambda > 0, \quad (3)$$

$$\max_{i \leq n} \|x_i\|^2 = O(1), \quad (4)$$

donde  $\lambda_n$  denota el mínimo autovalor de la matriz  $n^{-1}(X^t X)^{ng}$ , entonces  $a_n \xrightarrow{L_2} a$ , de donde,  $a_n$  es un estimador consistente de  $a$ . Además, la sucesión  $n^{\frac{1}{2}}(a_n - a)$  es acotada en  $L_2$ .

OBSERVACION: La hipótesis (4) equivale a que existe una cota superior  $K < \infty$ , para las normas de las variables independientes. Es decir, para cada  $i \in \mathcal{N}$ , se cumple

$$\|x_i\| \leq K.$$

DEMOSTRACION DEL TEOREMA: Se ha asumido que  $a$  es el verdadero valor del parámetro de regresión. Definamos las funciones indicadoras  $\delta_{i,0}$  ( $= 1$  si  $i \in I^{ng}$ ) y  $\eta_{i,j,h}$  ( $= 1$  si  $i \in I_j \subset I^g$  y  $z_i \in (c_{j,h-1}, c_{j,h}]$ ), para cada  $i \in \mathcal{N}$ ,  $j = 1, \dots, s$  y  $h = 1, \dots, r_j$ .

Obsérvese que  $y(a)$  puede escribirse en la forma

$$y(a) = Xa + \varepsilon,$$

donde las componentes de  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^t$  son independientes y definidas como

$$\varepsilon_i = \delta_{i,0}\nu_i + \sum_{j=1}^s \sum_{h=1}^{r_j} \eta_{i,j,h} \gamma(-a^t x_i + c_{j,h-1}, -a^t x_i + c_{j,h}).$$

Si  $F$  es la función de distribución de probabilidad de los errores  $\nu$ , está claro que  $\varepsilon_i = \nu_i$  si  $i \in I^{ng}$  y, por el contrario, si  $i \in I^g$

$$\varepsilon_i = E(\nu | \nu \in (-a^t x_i + c_{j,h-1}, -a^t x_i + c_{j,h})) \text{ con probabilidad } \pi_j q_{j,h},$$

donde  $q_{j,h} = F(-a^t x_i + c_{j,h}) - F(-a^t x_i + c_{j,h-1})$ . Nótese que las esperanzas de  $\varepsilon_i$  son nulas, debido a las correcciones empleadas en media condicionada:

$$E(\varepsilon_i) = \pi_0 E(\nu_i) + \sum_{j=1}^s \sum_{h=1}^{r_j} \pi_j q_{j,h} \gamma(-a^t x_i + c_{j,h-1}, -a^t x_i + c_{j,h}) = 0.$$

Igualmente, se puede afirmar que  $V(\varepsilon_i) \leq V(\nu) = 1$ .

2. Algoritmo MD de estimación en modelos lineales con datos agrupados y correcciones en media

Al igual que se hizo al probar el teorema 2.1, escribáse

$$y(a_n) - y(a) = (I - M^*) X(a_n - a),$$

donde la matriz aleatoria  $M^*$  está asociada a  $(a_n, a)$  de la misma manera que  $M^{(p)}$  lo está a la pareja  $(a_p, a_{p-1})$  en la demostración anterior. Todos los  $a_n$  cumplen (2), de donde

$$X^t \varepsilon = X^t y(a) - X^t X a - X^t y(a_n) + X^t X a_n,$$

es decir,

$$X^t \varepsilon = X^t M^* X(a_n - a). \quad (5)$$

A partir de (3), se obtiene que el mínimo autovalor de  $X^t M^* X$  no puede ser inferior a  $n\lambda_n \geq n\lambda > 0$ , de donde

$$a_n - a = (X^t M^* X)^{-1} X^t \varepsilon.$$

Así pues,

$$E \|a_n - a\|^2 \leq \lambda^{-2} n^{-2} m \sum_{i,k=1}^n \|x_i\| |E(\varepsilon_i \varepsilon_k)| \|x_k\|.$$

Teniendo en cuenta que, si  $i \neq j$  es  $|E(\varepsilon_i \varepsilon_k)| = |E(\varepsilon_i)E(\varepsilon_k)| = 0$ , que  $|E(\varepsilon_i^2)| = V(\varepsilon_i) \leq 1$  y que, además,  $\|x_i\| \leq K$ , se puede concluir directamente que

$$E \|a_n - a\|^2 \leq \lambda^{-2} n^{-2} m \sum_{i=1}^n \|x_i\|^2 \leq \lambda^{-2} n^{-1} m K^2 = O(n^{-1}),$$

probándose así la primera parte del teorema.

Finalmente, obsérvese que, repitiendo un argumento similar al de antes, se puede escribir

$$E \left\| n^{\frac{1}{2}} (a_n - a) \right\|^2 \leq n \lambda^{-2} m n^{-2} \sum_i^n \|x_i\|^2 \leq \lambda^{-2} m K^2 = O(1),$$

con lo que se finaliza la demostración.  $\square$

A continuación se formulará un resultado que refleja la ley de probabilidad límite del estimador propuesto  $a_n$ . Se verá que la convergencia es a una

distribución normal, cuya matriz de covarianzas desconocida, después, se tratará de estimar. El resultado es el siguiente:

**Teorema 2.3** *Bajo las mismas hipótesis del teorema 2.2, la sucesión  $n^{\frac{1}{2}}(a_n - a)$  converge en distribución a una normal, es decir,*

$$n^{\frac{1}{2}}(a_n - a) \xrightarrow[n \rightarrow \infty]{D} N(0, \Lambda)$$

para alguna matriz de covarianzas  $\Lambda$ . En consecuencia, supuesto  $n$  suficientemente grande, la distribución de  $a_n - a$  puede ser aproximada por

$$(a_n - a) \approx N(0, n^{-1}\Lambda).$$

DEMOSTRACION: Defínase la matriz  $M = \text{diag}(m_i)$ , donde  $m_i = \frac{d}{d\beta} \delta_{j,r}(\beta)|_{\beta=-a^t x_i}$ , si  $i \in I_j$  y  $z_i \in (c_{j,r}, c_{j,r+1}]$ , y  $m_i = 1$ , si  $i \in I^{ng}$ . Se probará que se producen las siguientes dos convergencias estocásticas

$$n^{-\frac{1}{2}} X^t (M - E(M)) X (a_n - a) \xrightarrow[n \rightarrow \infty]{L_1} 0 \quad (6)$$

y

$$n^{-\frac{1}{2}} X^t (M^* - M) X (a_n - a) \xrightarrow[n \rightarrow \infty]{P} 0. \quad (7)$$

Para ver la primera de ellas, denótese  $(p_{jk}) = n^{-1} X^t (M - E(M)) X$  y obsérvese que

$$\left\| n^{-\frac{1}{2}} X^t (M - E(M)) X (a_n - a) \right\| \leq \sum_{j,k=1}^m |p_{jk}| \left\| n^{\frac{1}{2}} (a_n - a) \right\|.$$

A partir del teorema 2.2, será suficiente con demostrar que  $p_{jk} \xrightarrow{L_2} 0$ , ya que por la desigualdad de Hölder

$$E \left| \sum_{j,k=1}^m |p_{jk}| \left\| n^{\frac{1}{2}} (a_n - a) \right\| \right| \leq \left( E \left( \sum_{j,k=1}^m |p_{jk}| \right)^2 E \left\| n^{\frac{1}{2}} (a_n - a) \right\|^2 \right)^{\frac{1}{2}}$$

de donde se concluirá que la parte derecha de la desigualdad converge a cero. Para todo  $j, k$ , veamos, pues, que  $p_{jk} \xrightarrow{L_2} 0$ . Se tiene que

$$p_{jk} = n^{-1} \sum_{i=1}^n x_{ij} x_{ik} (m_i - E(m_i)).$$

Recordando que  $0 \leq m_i \leq 1$  y también la independencia de las  $m_i$ , puede

concluirse a partir de (4) que

$$\begin{aligned} E(p_{jk}^2) &= n^{-2} \sum_{i=1}^n x_{ij}^2 x_{ik}^2 E|m_i - E(m_i)|^2 \leq n^{-2} \sum_{i=1}^n \|x_i\|^4 \\ &\leq n^{-1} K^4 = O(n^{-1}), \end{aligned}$$

lo cual implica la convergencia en cuestión.

Con el fin de probar el segundo límite en probabilidad (7), nótese que, a partir de las hipótesis de diferenciabilidad establecidas al principio de la sección junto con la propiedad ya probada  $a_n \xrightarrow{P} a$ , se cumple que  $m_i - m_i^* \xrightarrow{P} 0$ . Defínase  $(r_{jk}) = n^{-1} X^t (M^* - M) X$  y escribáse

$$\left\| n^{-\frac{1}{2}} X^t (M^* - M) X (a_n - a) \right\|^2 \leq \left( \sum_{j,k=1}^m r_{jk}^2 \right) \left\| n^{\frac{1}{2}} (a_n - a) \right\|^2,$$

donde

$$|r_{jk}| \leq n^{-1} \max_{i \leq n} \|x_i\|^2 \sum_{i=1}^n |m_i^* - m_i| \leq K^2 \frac{1}{n} \sum_{i=1}^n |m_i^* - m_i|.$$

Puesto que todas las  $m_i - m_i^*$  están acotadas, se sigue, por la ley débil de los grandes números, que  $r_{jk} \xrightarrow{P} 0$ , lo cual implica el límite en probabilidad (7).

Con la misma notación de arriba, escribáse ahora (después de sumar y restar  $(M + M^*)$ )

$$n^{-\frac{1}{2}} X^t E(M) X (a_n - a) = n^{-\frac{1}{2}} X^t M^* X (a_n - a) + \zeta_n,$$

con  $\zeta_n \rightarrow 0$  en probabilidad. Puesto que  $E(m_i) = 1$  para todo  $i \in I^{ng}$ , el mínimo autovalor de la matriz simétrica  $n^{-1} X^t E(M) X$  no puede ser inferior a  $\lambda$ . Esto implica la no singularidad de la matriz de arriba y también la de su límite cuando  $n \rightarrow \infty$  (si tal límite existe). Teniendo en cuenta (5), se puede escribir

$$n^{-1} X^t E(M) X n^{\frac{1}{2}} (a_n - a) = n^{-\frac{1}{2}} X^t \varepsilon + \zeta_n,$$

de donde, por la no singularidad citada,

$$n^{\frac{1}{2}} (a_n - a) = (X^t E(M) X)^{-1} n^{\frac{1}{2}} X^t \varepsilon + \zeta_n^*. \quad (8)$$

Obsérvese que  $\|\zeta_n^*\| \leq \lambda^{-1} \|\zeta_n\| \xrightarrow{P} 0$ . En consecuencia, las distribuciones asintóticas de  $n^{\frac{1}{2}}(a_n - a)$  y de  $n^{\frac{1}{2}}(X^t E(M)X)^{-1} X^t \varepsilon$  tienen que coincidir. Nótese que la  $j$ -ésima componente del último vector aleatorio es

$$\left[ n^{\frac{1}{2}} (X^t E(M)X)^{-1} X^t \varepsilon \right]_j = n^{-\frac{1}{2}} \sum_{i=1}^n f_j x_i \varepsilon_i,$$

con  $f_j$  denotando la  $j$ -ésima fila de  $(n^{-1} X^t E(M)X)^{-1}$ . Es suficiente con probar que todas estas componentes se distribuyen en el límite normalmente. Para este fin, obsérvese que la varianza de cada término de la suma está acotada por

$$n^{-1} (f_j x_i)^2 V(\varepsilon_i) \leq n^{-1} \left\| (n^{-1} X^t E(M)X)^{-1} x_i \right\|^2 \leq n^{-1} \lambda^{-2} \|x_i\|^2.$$

En conclusión, las dos siguientes relaciones se siguen directamente de la acotación de las variables independientes

$$\max_{i \leq n} V \left( n^{-\frac{1}{2}} f_j x_i \varepsilon_i \right) \leq \lambda^{-2} n^{-1} (\max_{i \leq n} \|x_i\|^2) \leq \lambda^{-2} n^{-1} K^2 = O(n^{-1})$$

y

$$\sum_{i=1}^n V \left( n^{-\frac{1}{2}} f_j x_i \varepsilon_i \right) \leq \lambda^{-2} n^{-1} \sum_{i=1}^n \|x_i\|^2 \leq \lambda^{-2} K^2 = O(1).$$

Ambas relaciones implican las condiciones del teorema central del límite de Lindeberg-Feller, completándose así la demostración (ver LAHA Y ROHATGI (1979)).□

## 2.5 Estimación de la matriz de covarianzas asintótica

En esta sección se tratará de obtener una estimación, a partir de la información de que se dispone en cada momento, de la matriz de covarianzas asintótica  $\Lambda$ , que aparece en el último resultado de la sección anterior. Las hipótesis de partida son las mismas allí citadas, en concreto las relativas a la diferenciabilidad de  $\delta_{jr}$ .

Considérese  $\varepsilon(\alpha) = (\varepsilon_1(\alpha), \dots, \varepsilon_n(\alpha))^t$ , donde  $\alpha \in \mathcal{R}^m$  y

$$\varepsilon_i(\alpha) = \delta_{i,0} \nu_i + \sum_{j=1}^s \sum_{h=1}^{r_j} \eta_{i,j,h} \gamma (-\alpha^t x_i + c_{j,h-1}, -\alpha^t x_i + c_{j,h}).$$

Nótese que el vector  $\varepsilon$  definido a lo largo de la demostración del teorema 2.2 corresponde a  $\varepsilon(a)$ . Sea  $\Sigma(\varepsilon(\alpha))$  la matriz diagonal de covarianzas de  $\varepsilon(\alpha)$ . Asíumase la existencia de los siguiente dos límites:

$$\Phi = \lim_n \Phi_n = \lim_n n^{-1} X^t \Sigma(\varepsilon) X \quad (9)$$

$$W = \lim_n W_n = \lim_n n^{-1} X^t E(M) X \quad (10)$$

Mirando la expresión (8), es claro que  $\Lambda = W^{-1} \Phi W^{-1}$ . Defínase la matriz, de tamaño  $n \times n$ , dada por  $M_n = \text{diag} \left( m_i^{(n)} \right)$  donde  $m_i^{(n)} = \frac{d}{d\beta} \delta_{j,r}(\beta) |_{\beta = -a_i^+, x_i}$ , si  $i \in I_j$  y  $z_i \in (c_{j,r}, c_{j,r+1}]$ , y  $m_i^{(n)} = 1$ , si  $i \in I^{ng}$ . Cuando la última derivada no existe, defínase  $m_i^{(n)} = \theta$ , como un valor arbitrario. Puesto que  $a_n \xrightarrow{P} a$ , también  $m_i^{(n)} - m_i \xrightarrow{P} 0$ , bajo las hipótesis asumidas de diferenciabilidad. Denótese  $\Theta_n = n^{-1} X^t M_n X$  y  $\Pi_n = n^{-1} X^t \Sigma(\varepsilon(a_n)) X$ . Se afirma que las matrices  $\Theta_n^{-1} \Pi_n \Theta_n^{-1}$  estiman consistentemente la matriz de covarianzas  $\Lambda$ , puesto que los siguiente límite en probabilidad ocurren, como se verá a continuación,

$$\|\Theta_n - W_n\| \xrightarrow{P} 0$$

y

$$\|\Pi_n - \Phi_n\| \xrightarrow{P} 0.$$

Recuérdese, de la demostración del teorema 2.3, que  $n^{-1} X^t (M - E(M)) X \xrightarrow{P} 0$ , de donde el primer límite se concluiría a partir de que  $n^{-1} X^t (M^{(n)} - M) X = (s_{jk}) \xrightarrow{P} 0$ . Finalmente, obsérvese que

$$|s_{jk}| \leq \max_{i \leq n} \|x_i\|^2 n^{-1} \sum_{i=1}^n \left| m_i^{(n)} - m_i \right| \leq K^2 \frac{1}{n} \sum_{i=1}^n \left| m_i^{(n)} - m_i \right|.$$

Puesto que  $m_i^{(n)} - m_i \xrightarrow{P} 0$ , siendo todas las  $m_i^{(n)} - m_i$  acotadas, se sigue que  $\frac{1}{n} \sum_{i=1}^n \left| m_i^{(n)} - m_i \right| \xrightarrow{P} 0$ , completándose la demostración del primer límite.

Para el segundo, obsérvese que  $\varepsilon_i(a_n) \xrightarrow{P} \varepsilon_i(a)$ . Puesto que  $\{\varepsilon_i(a_n)\}$  son uniformemente acotadas, puede escribirse  $\text{Var}(\varepsilon_i(a_n)) \rightarrow \text{Var}(\varepsilon_i(a))$ . Un



argumento similar al empleado arriba permite asegurar que

$$n^{-1}X^t(\Sigma(\varepsilon(a_n)) - \Sigma(\varepsilon))X \xrightarrow{P} 0.$$

En conclusión, la siguiente proposición ha sido demostrada:

**Teorema 2.4** *Bajo las hipótesis del teorema 2.3 de la sección anterior y asumiendo la existencia de los límites (9) y (10), las matrices  $\Theta_n^{-1}\Pi_n\Theta_n^{-1}$  estiman consistentemente la matriz de covarianzas asintótica  $\Lambda$ .*

En definitiva, supuesto  $n$  suficientemente grande, la distribución de  $a_n$  es aproximadamente  $N(a, n^{-1}\Theta_n^{-1}\Pi_n\Theta_n^{-1})$ .

## 2.6 Simulaciones

En esta sección se muestra mediante simulaciones el comportamiento del algoritmo MD expuesto en este capítulo, el cual utiliza, como se ha explicado, correcciones en media.

Se ha asumido que en el modelo lineal

$$z_i = a^t x_i + \nu_i, \quad i = 1, \dots, N,$$

los términos de error  $\nu_i$  son variables aleatorias independientes e idénticamente distribuidas como normales  $N(0, 1)$ . El parámetro a estimar se ha fijado como  $a = \begin{pmatrix} -4 \\ 10 \end{pmatrix}$ . Buscando ajustar una recta que no pase por el origen, las variables independientes acotadas  $x_i$  se toman de forma que  $x_{i1} = 1$  y  $x_{i2}$  sea extraído dentro del intervalo (1.2, 2.7). Algunos de los valores  $z_i$  no son observados y han sido censurados. En concreto, el 90% de los datos se censuran, lo cual indica que la información real es muy limitada. Cuando  $z_i$  es censurado se clasifica su valor en alguno de los siguientes intervalos:

$$(-\infty, 0], (0, 10], (10, 15], (15, 20], (20, 30], (30, \infty).$$

Fijado este modelo lineal con censura, se extrae una muestra concreta, con información parcial, de tamaño  $N = 1000$ . Las primeras 10 realizaciones de

## 2. Algoritmo MD de estimación en modelos lineales con datos agrupados y correcciones en media

esta muestra aparecen en la tabla 1 del final de la sección, mostrándose, además, los correspondientes intervalos de censura para los datos agrupados. En la tabla citada, la última columna refleja la variable  $p_i$ , indicatriz de la censura de  $z_i$ , de forma que  $p_i = 1$  cuando el dato  $z_i$  es censurado y  $p_i = 0$  en otro caso. Para cada dato censurado, la tabla 1 incluye su intervalo de censura  $(l_i, u_i]$ . Se asume que este intervalo es conocido; no así el verdadero valor de  $z_i$ , si bien en la tabla 1 aparece dicho valor, con el fin de ilustrar como se ejerce la censura.

La estimación  $a_N$  en la etapa  $N = 1000$  del proceso iterativo primario es  $a_N = (-4.423, 10.271)$ . La tabla 2 muestra el resultado de las iteraciones del proceso secundario tomando distintos puntos de arranque, en concreto  $(25, 50)$ ,  $(1000, -1000)$  y  $(0, 0)$ . En todos los casos los puntos de convergencia coinciden, tal como se demostró en el teorema 2.1. Puede observarse que, si se considera como criterio de parada del método iterativo secundario el que la distancia entre cada una de las componentes de dos valores sucesivos sea menor que  $10^{-3}$ , el número de iteraciones que se deben realizar es de aproximadamente 40, para todos los puntos de arranque antes citados.

A continuación, se ha simulado también el mismo modelo anterior cuando se censuran solamente un 50% de los datos. La tabla 3 muestra la evolución de las primeras 15 iteraciones del proceso secundario, tomando como puntos de arranque  $(50, -30)$ ,  $(0, 0)$  y  $(-50, 80)$ . La convergencia se produce al punto  $(-6.155, 11.130)$ . Con el criterio mencionado de parada, el número de iteraciones necesarias para obtener convergencia se reduce hasta aproximadamente 8 iteraciones.

En cualquier caso, uno de los objetivos de esta sección se centra en comprobar numéricamente que la convergencia en  $L_2$  demostrada en el teorema 2.2 es efectiva. Para comprobar que  $E \|a_n - a\|^2 \xrightarrow{n \rightarrow \infty} 0$  se realiza el proceso de estimación descrito para un total de  $M = 60$  muestras diferentes. Para cada

etapa  $n = 1, \dots, 50$ , del proceso iterativo primario, se estiman las esperanzas  $E \|a_n - a\|^2$ , empíricamente mediante

$$E \widehat{\|a_n - a\|^2} = \frac{1}{60} \sum_{m=1}^{60} \|a_n(m) - a\|^2,$$

donde  $a_n(m)$  denota la estimación en la etapa  $n$ , para la  $m$ -ésima muestra,  $m = 1, \dots, 60$ . La figura 1 muestra como las citadas esperanzas cuadráticas empíricas tienen una tendencia de decrecimiento hacia cero a medida que  $n$  aumenta.

Por último, se ha querido comprobar la normalidad asintótica que establece el teorema 2.3 y estimar, además, la matriz de covarianzas implicada en esa convergencia. Fijado un tamaño de muestra  $N = 200$ , el teorema 2.3 establece que

$$\sqrt{N}(a_N - a) \approx N(0, \Lambda), \quad (11)$$

siendo  $\Lambda$  una matriz de covarianzas asintótica de tamaño  $2 \times 2$ . Como antes se han tomado  $M = 60$  muestras diferentes del modelo lineal de censura y se han calculado para cada una de ellas las estimaciones correspondientes a la etapa primaria  $a_N$ , utilizando el algoritmo MD de dos iteraciones.

Los contrastes de Kolmogorov-Smirnov o de la  $\chi^2$  pueden servir para comprobar la hipótesis de normalidad. Con el primero de ellos, se han obtenido p-valores de 0.95 y 0.9346 para la normalidad de cada una de las componentes de  $a_N$ . Con los segundos, los p-valores se reducen ligeramente. Finalmente, considerando (11) se obtiene como estimación de  $\Lambda$  la matriz

$$\hat{\Lambda} = \begin{pmatrix} 185.18 & -87.16 \\ -87.16 & 41.29 \end{pmatrix}.$$

#### CONCLUSIONES

En síntesis:

1. Se observa que a medida que el porcentaje de datos censurados aumenta el número de iteraciones necesarias para obtener convergencia del proceso

## 2. Algoritmo MD de estimación en modelos lineales con datos agrupados y correcciones en media

secundario también aumenta. De hecho, con un 90% de datos censurados se precisan 40 iteraciones, las cuales se reducen a 8 con un 50% de censuras.

2. La figura 1 muestra que hay una clara tendencia de decrecimiento hacia cero del error cuadrático de estimación a medida que el número de iteraciones primarias crece. Sin embargo, la figura no es siempre decreciente, observándose en ocasiones ligeros aumentos de los errores cuadráticos medios estimados. Estos ligeros aumentos pueden deberse al hecho de que  $E \|a_n - a\|^2$  no sea estrictamente decreciente. En todo caso, los ligeros aumentos podrían, también, justificarse a través de la aleatorización latente en la aproximación, pudiendo evitarse simplemente aumentando el número  $M = 60$  de muestras. Por último, se quiere hacer notar que la realización de esta simulación exige un enorme tiempo de computación (alrededor de 110 minutos). Ello está motivado por el gran número de iteraciones que conlleva la obtención de las estimaciones y porque cada iteración supone un cálculo de esperanzas condicionadas que se han llevado a cabo con un método de Gauss-Kronrod de cuadratura (y en consecuencia, muchas nuevas iteraciones y cálculos). En los siguientes capítulos se verá como se reduce este tiempo de cálculo utilizando otro tipo de correcciones distintas de las basadas en esperanzas condicionadas (e.g., moda, mediana,...).
3. La normalidad asintótica está garantizada. En este sentido, simplemente remito a los p-valores que se han obtenido con los contrastes habituales de bondad de ajuste antes citados.

	$x_{i1}$	$x_{i2}$	$z_i$	$l_i$	$u_i$	$p_i$
1	1.0	1.5	12.209			0
2	1.0	1.7	11.837	10.0	15.0	1
3	1.0	1.9	14.346			0
4	1.0	2.2	19.657	15.0	20.0	1
5	1.0	1.5	13.092	10.0	15.0	1
6	1.0	1.7	11.752	10.0	15.0	1
7	1.0	1.9	15.306			0
8	1.0	2.2	18.275	15.0	20.0	1
9	1.0	1.5	12.273	10.0	15.0	1
10	1.0	1.7	13.521	10.0	15.0	1

Tabla 1: 10 valores muestrales del modelo lineal censurado de la sección 2.6

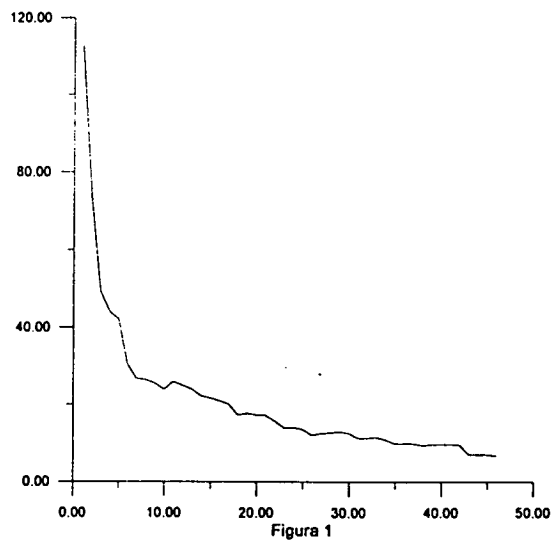


Figura 1: Error cuadrático de estimación del algoritmo MD

<i>p</i>	Pto. inicio (25,50)		Pto. inicio (1000,-1000)		Pto. inicio (0,0)	
	25	50	1000	-1000	0	0
1	-5.49862	11.51941	-0.35003	7.18298	-0.87555	7.46832
2	-7.07464	12.09423	0.445012	7.07964	0.16142	7.24334
3	-7.01913	11.93961	-0.13582	7.55728	-0.31961	7.66773
4	-6.59551	11.63954	-0.96244	8.10687	-1.09530	8.18832
5	-6.14536	11.34843	-1.72713	8.59325	-1.82826	8.65578
6	-5.80094	11.13016	-2.31872	8.96513	-2.39739	9.01395
7	-5.51621	10.95185	-2.79818	9.26451	-2.85993	9.30288
8	-5.28757	10.80928	-3.18041	9.50259	-3.22904	9.53282
9	-5.10592	10.69620	-3.48324	9.69103	-3.52159	9.71487
10	-4.96220	10.60678	-3.72260	9.83992	-3.75285	9.85873
11	-4.84866	10.53616	-3.92925	9.96495	-3.95312	9.97979
12	-4.75902	10.48041	-4.08132	10.05845	-4.05401	10.04214
13	-4.68826	10.43640	-4.15182	10.10264	-4.13239	10.09077
14	-4.63241	10.40167	-4.20856	10.13804	-4.19388	10.12897
33	-4.42564	10.27308	-4.42092	10.27015	-4.42076	10.27005
34	-4.42515	10.27278	-4.42142	10.27046	-4.42130	10.27038
35	-4.42476	10.27254	-4.42182	10.27071	-4.42172	10.27065
36	-4.42446	10.27235	-4.42213	10.27090	-4.42206	10.27086
37	-4.42422	10.27220	-4.42238	10.27106	-4.42232	10.27102
38	-4.42403	10.27208	-4.42258	10.27118	-4.42253	10.27115
39	-4.42388	10.27199	-4.42274	10.27128	-4.42270	10.27125
40	-4.42376	10.27192	-4.42286	10.27135	-4.42283	10.27133
41	-4.42367	10.27186	-4.42295	10.27141	-4.42293	10.27140
42	-4.42359	10.27181	-4.42303	10.27146	-4.42301	10.27145
43	-4.42353	10.27177	-4.42309	10.27150	-4.42307	10.27149
44	-4.42349	10.27175	-4.42314	10.27153	-4.42312	10.27152
45	-4.42345	10.27172	-4.42317	10.27155	-4.42317	10.27155
93	-4.42331	10.27164	-4.42331	10.27164	-4.42331	10.27164
94	-4.42331	10.27164	-4.42331	10.27164	-4.42331	10.27164
95	-4.42331	10.27164	-4.42331	10.27164	-4.42331	10.27164
96	-4.42331	10.27164	-4.42331	10.27164	-4.42331	10.27164
97	-4.42331	10.27164	-4.42331	10.27164	-4.42331	10.27164
98	-4.42331	10.27164	-4.42331	10.27164	-4.42331	10.27164
99	-4.42331	10.27164	-4.42331	10.27164	-4.42331	10.27164

Tabla 2: Iteraciones del proceso secundario con diferentes puntos de inicio.  
90% de censura

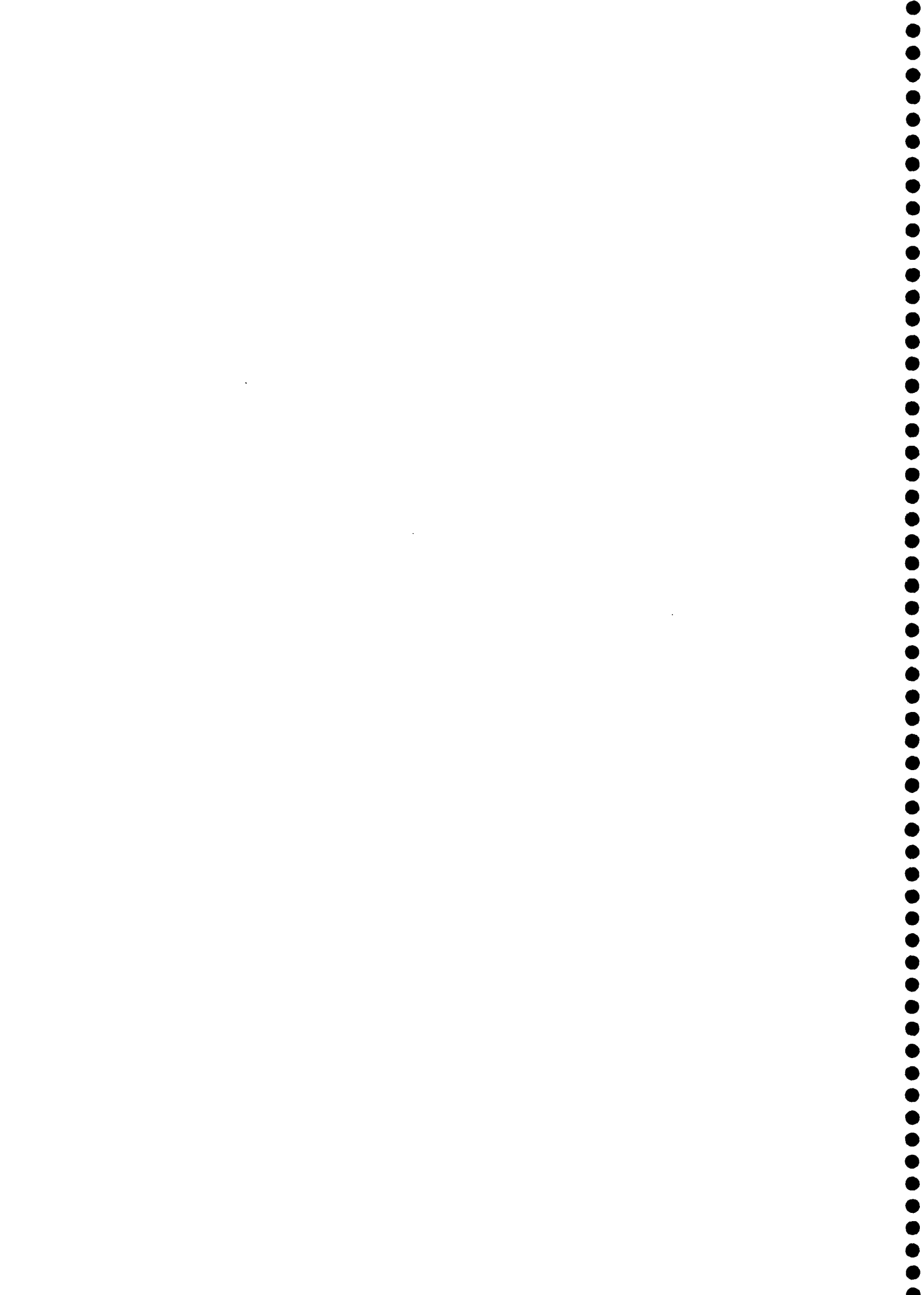
<i>n</i>	Punto inicio (50,-30)		Punto inicio (0,0)		Punto inicio (-50,80)	
	50	-30	0	0	-50	80
1	-5.44796	10.49727	-5.46073	10.50627	-6.65461	11.84927
2	-5.94308	10.97499	-5.94622	10.97703	-6.29683	11.28603
3	-6.11469	11.10063	-6.11521	11.10097	-6.19772	11.16654
4	-6.14829	11.12471	-6.14838	11.12477	-6.16544	11.13775
5	-6.15439	11.12911	-6.15441	11.12912	-6.15773	11.13160
6	-6.15550	11.12991	-6.15550	11.12991	-6.15613	11.13038
7	-6.15570	11.13005	-6.15570	11.13005	-6.15582	11.13014
8	-6.15573	11.13008	-6.15573	11.13008	-6.15576	11.13010
9	-6.15574	11.13009	-6.15574	11.13009	-6.15574	11.13009
10	-6.15574	11.13009	-6.15574	11.13009	-6.15574	11.13009
11	-6.15574	11.13009	-6.15574	11.13009	-6.15574	11.13009
12	-6.15574	11.13009	-6.15574	11.13009	-6.15574	11.13009

Tabla 3: Iteraciones del proceso secundario con diferentes puntos de inicio.  
50% de censura

### Notas y comentarios

- Aunque, como se ha dicho en la introducción, el algoritmo MD tiene relación con el EM, éste último presenta un antecedente histórico, el cual podría referenciarse, igualmente, como antecedente del algoritmo aquí propuesto. Se trata de ORCHARD Y WOODBURY (1972), donde por primera vez se presenta el así llamado principio de información incompleta.
- La CONDICION C1 que figura al comienzo del capítulo es obligada para poder garantizar todos los resultados contenidos en él. Dicha condición afecta al peso de las colas de la distribución de los términos de error. Cuanto menos pesadas sean éstas más fácilmente se podrá garantizar dicha condición. En el ejemplo iii) de la sección 6.2.1 se incluye algún resultado que garantiza la condición en cuestión.
- Como se ha indicado, el punto débil del algoritmo, cuando se consideran términos generales de error, hay que buscarlo en las cuadraturas que conlleva. A este respecto, las referencias TIERNEY Y KADANE (1986) y GEWEKE (1989) resultan obligadas, en mi opinión.





### **3. Algoritmo MdD de estimación en modelos lineales con datos agrupados y correcciones modales**

#### **3.1 Introducción**

En este capítulo se desarrolla un algoritmo iterativo de estimación basado en correcciones en moda, el cual continúa siendo válido para ajustar modelos lineales con datos agrupados en intervalos. Se seguirá asumiendo, como en el capítulo anterior, que los datos pueden ser extraídos de diferentes fuentes, de forma que los intervalos de agrupación, antes citados, pueden variar para los diferentes elementos de la muestra.

La idea de las imputaciones en moda surge como respuesta a los problemas computacionales que implican el cálculo de esperanzas condicionadas del algoritmo visto en el capítulo anterior. El cálculo de las integrales involucradas en el anterior procedimiento MD puede ser abordado por los distintos métodos de cuadratura existentes. Pese a ello, la utilización de estos últimos supone de facto añadir un nuevo proceso iterativo anidado a los dos que componen, como se sabe, el MD. Hoy día está aceptado con generalidad que las iteraciones anidadas tienden a degenerar los resultados, computacionalmente hablando, y deben ser evitadas en la medida de lo posible.

En la línea antes citada, se propone sustituir el paso relativo a las correcciones en media condicionada por un paso alternativo de corrección en modas

condicionadas. Este último resulta mucho más sencillo de calcular puesto que, como se verá más adelante, depende fundamentalmente de la forma de las distribuciones de los términos de error. Por lo demás, se mantiene el esquema operativo del algoritmo MD. En este sentido, el segundo paso del mismo continúa consistiendo en la proyección ya explicada en la sección 2.2.

Por lo tanto, encontramos dos ventajas fundamentales en esta propuesta: primera la reducción computacional de tiempo; y segunda la aplicabilidad a una familia de distribuciones más amplia que la utilizada con la corrección en media. De hecho será un método válido para cualquier distribución unimodal en cero. Todo esto se complementa con la permanencia de los resultados asintóticos, en la línea de los del capítulo anterior, en particular los de consistencia y convergencia en ley a la normal.

Es de resaltar que la utilización de correcciones en moda no tiene precedentes. El nuevo procedimiento iterativo de estimación se separa completamente del EM, el cual, como se ha indicado en la introducción del capítulo anterior, constituyó nuestro punto de partida inicial. Las referencias mencionadas DEMPSTER, LAIRD Y RUBIN (1977), TANNER (1993) y MCLACHLAN Y KRISHNAN (1997) pierden el carácter citado como antecedentes del trabajo que sigue. No así ORCHARD Y WOODBURY (1972), WU (1983) y LOUIS (1982) en lo relativo al principio subyacente a la información incompleta, la búsqueda de convergencia única (independiente del punto de arranque) para el proceso iterativo secundario y el intento de aceleración global.

### **3.2 Modelo de regresión lineal, notación, algoritmo e hipótesis**

El modelo inicial coincide, salvo en la forma de los errores, con el expuesto en la

sección 2.2, es decir, considérese un modelo lineal

$$z_i = a^t x_i + \nu_i, \quad i = 1, \dots, n \quad (12)$$

con las mismas condiciones que el modelo (1). Tan sólo se impondrá aquí una hipótesis distribucional que afecta a los errores  $\nu_i$ . Se asumirá que su densidad  $f > 0$  es simétrica, unimodal en cero y con varianza finita. En esta situación, la esperanza de los errores es cero y, sin pérdida de generalidad, puede aceptarse que tienen varianza uno.

Las variables dependientes se asumirá que pueden ser agrupadas o no agrupadas, existiendo diferentes particiones en intervalos de clasificación en el caso de ser agrupadas. Se asumirá, en este sentido, la misma notación usada en la sección 2.2, en lo que afecta tanto a los conjuntos de índices  $I, I^g, I^{ng}$  como a los intervalos de agrupación de los datos dentro de cada fuente:

$$-\infty = c_{j,0} < c_{j,1} < \dots < c_{j,r_j} = \infty.$$

Así mismo, se asumirá que  $X^t X$  es una matriz de rango completo, donde  $X^t = (x_1, \dots, x_n)$ . Para la estimación del vector  $a$  en el modelo (12), se sugiere emplear un procedimiento recursivo similar al MD del capítulo anterior. La única diferencia entre ambos afectará a las correcciones o imputaciones de los valores censurados. Las correcciones en media condicionada se sustituirán por correcciones en moda condicionada, mucho más simple que aquéllas ante las condiciones de forma impuestas sobre la densidad subyacente a los errores.

Formalmente, el proceso iterativo secundario, fijada la muestra de tamaño  $n$ , es el siguiente:

INICIALIZACIÓN: Fijese una estimación vectorial inicial arbitraria  $a_0$ .

ITERACIÓN: Asumiendo que la estimación actual conocida es  $a_p$ , la siguiente estimación esta definida por:

$$a_{p+1} = (X^t X)^{-1} X^t y(a_p)$$

donde para cada  $i \in I$

$$\begin{aligned} y_i(a_p) &= z_i, & \text{si } i \in I^{ng} \\ &= a_p^t x_i + \gamma(-a_p^t x_i + c_{j,r}, -a_p^t x_i + c_{j,r+1}), & \text{si } i \in I_j, z_i \in (c_{j,r}, c_{j,r+1}]. \end{aligned}$$

En la última expresión, la función  $\gamma$  se define para parejas  $(t, w)$ , siendo  $t < w$ , mediante

$$\gamma(t, w) = \text{Moda}(\nu | \nu \in (t, w]).$$

Como se indicó en el capítulo anterior si  $w = \infty$  se entenderá que el intervalo es de la forma  $(t, \infty)$ . En lo sucesivo llamaremos a este procedimiento algoritmo MdD.

Supuesto  $t, w$  fijos, denotando por  $\delta(\beta) = \gamma(\beta + t, \beta + w)$ , la expresión explícita de esta función  $\delta$  con el tipo de errores fijado es clara:

$$\delta(\beta) = \begin{cases} \beta + w & \text{si } \beta < -w \\ 0 & \text{si } -w < \beta < -t \\ \beta + t & \text{si } \beta > -t. \end{cases}$$

Se observa, pues, que  $\delta(\beta)$  es una función no decreciente, al igual que  $\beta - \delta(\beta)$ . Ambas funciones son, además, independientes de la distribución unimodal en cero asumida para el término de error  $\nu$ .

### 3.3 Un primer resultado de convergencia

Veamos, en primer lugar, que cualquier sucesión  $a_p$  generada por el proceso secundario del algoritmo MdD converge a un punto que es independiente de la estimación inicial  $a_0$  que se proponga.

**Teorema 3.1** *Asúmase que  $(X^t X)^{ng} = \sum_{i \in I^{ng}} x_i x_i^t$  es una matriz definida positiva. Para cualquier vector inicial  $a_0$ , la sucesión  $a_p$  generada por el algoritmo MdD converge a un vector  $a^*$  que satisface la ecuación implícita*

$$a^* = (X^t X)^{-1} X^t y(a^*). \quad (13)$$

*Adicionalmente, el punto  $a^*$  es la única solución de la anterior ecuación.*

DEMOSTRACION: Teniendo en cuenta que, una vez que se dispone de una

realización del experimento, cada dato tiene asignado un intervalo fijo de clasificación y siguiendo el argumento de demostración utilizado en el teorema 2.1, se puede afirmar que la ecuación

$$a = (X^t X)^{-1} X^t y(a), \quad a \in \mathcal{R}^m$$

tiene una única solución, en un cierto punto  $a^*$ .

El punto límite de cualquier sucesión determinista  $\{a_p\}$  generada por el algoritmo debe verificar la expresión (13). Sólo resta probar que siempre existe el límite de esa sucesión generada por el método MdD. Para ello, escríbase

$$a_{p+1} - a_p = (X^t X)^{-1} X^t (y(a_p) - y(a_{p-1})).$$

Utilizando la forma de la función  $\delta(\beta)$  asociada a las correcciones en moda empleadas, se puede escribir, equivalentemente, como

$$a_{p+1} - a_p = (X^t X)^{-1} X^t (I - M^{(p)}) X (a_p - a_{p-1}),$$

donde  $M^{(p)} = \text{diag}(m_i^{(p)})$  y  $0 \leq m_i^{(p)} \leq 1$ . De esta expresión, siguiendo nuevamente la demostración del teorema 2.1, se puede concluir que

$$\left\| (X^t X)^{\frac{1}{2}} (a_{p+1} - a_p) \right\| \leq \tau \left\| (X^t X)^{\frac{1}{2}} (a_p - a_{p-1}) \right\| \quad (14)$$

para un valor  $\tau \in (0, 1)$ . En consecuencia,

$$\left\| (X^t X)^{\frac{1}{2}} (a_{p+1} - a_p) \right\| \leq \tau^p \left\| (X^t X)^{\frac{1}{2}} (a_1 - a_0) \right\|.$$

Se sigue, para cada  $r \geq 1$ ,

$$\begin{aligned} \left\| (X^t X)^{\frac{1}{2}} (a_{p+r} - a_p) \right\| &\leq \left\| (X^t X)^{\frac{1}{2}} (a_1 - a_0) \right\| \sum_{i=1}^r \tau^{p+i} \\ &\leq \left\| (X^t X)^{\frac{1}{2}} (a_1 - a_0) \right\| \sum_{i=1}^{\infty} \tau^{p+i}, \end{aligned}$$

cuyo último término converge hacia cero cuando  $p \rightarrow \infty$ . Esto implica que la sucesión  $\{a_p\}$  es de Cauchy, convergiendo, pues, al vector  $a^*$ . La desigualdad (14) indica que la tasa de convergencia de cada sucesión secundaria  $a_p$  es, cuanto

menos, lineal.  $\square$

Se propone al punto  $a^*$ , límite único de cualquier sucesión generada por el proceso secundario del algoritmo MdD, como estimación del vector  $a$  del modelo (12) a tamaño muestral fijo. El vector  $a^*$  se puede ver como un estimador de Huber, igual que ocurría con el algoritmo MD, por ser solución de la ecuación implícita (13). Esto permitirá demostrar las propiedades de convergencia asintótica (en el proceso primario de iteración) que se probarán a continuación.

### 3.4 Resultados de convergencia estocástica

Como en el capítulo anterior, denotemos  $a^* = a_n$ , para hacer explícita la dependencia del estimador del tamaño muestral  $n$ . Se investigará en esta sección la convergencia y las propiedades estocásticas de  $a_n$  cuando  $n \rightarrow \infty$ . Se comprobará que, pese a la simplificación que supone emplear correcciones en moda,  $a_n$  continúa siendo un estimador consistente de  $a$  asintóticamente distribuido como una normal. Obsérvese, pues, que sus propiedades son equiparables con las de los estimadores de máxima verosimilitud del modelo censurado (obtenidos, por ejemplo, vía el algoritmo EM) e, incluso, con las de los estimadores derivados del algoritmo MD con correcciones en media.

Recordemos que  $a$  denota el verdadero valor del parámetro de regresión a estimar en el modelo (12). Obsérvese que para cada  $j = 1, \dots, s$  y  $r = 0, \dots, r_j - 1$ ,  $\delta_{j,r}(\beta) = \gamma(\beta + c_{j,r}, \beta + c_{j,r+1})$  es continuamente diferenciable, salvo en  $-c_{j,r}$  y en  $-c_{j,r+1}$ . Asumamos en lo sucesivo que  $c_{j,r} \neq a^t x_i$ , para cada  $j$  y  $r$ . Esta condición no supone en la práctica gran restricción. Por ejemplo, si se supone que la distribución subyacente de las  $x_i$ 's es continua, al ser el conjunto de puntos extremos  $c_{j,r}$  finito, la condición se cumplirá con probabilidad uno.

La primera propiedad que se probará para el estimador  $a_n$  es la de consistencia, es decir, la convergencia en probabilidad de  $a_n$  al verdadero valor  $a$ . De hecho,

como en el algoritmo MD, puede garantizarse, incluso, la convergencia más restrictiva en media cuadrática.

**Teorema 3.2** Si  $(X^t X)^{ng}$  es definida positiva y existe un valor real  $\rho > 1$  para el cual se cumplen las dos condiciones técnicas siguientes

$$\inf_n \lambda_n = \lambda > 0, \quad (15)$$

$$\max_{i \leq n} \|x_i\|^2 = O(n^{\rho-1}), \quad (16)$$

donde  $\lambda_n$  denota el mínimo autovalor de la matriz  $n^{-\rho}(X^t X)^{ng}$ , entonces  $a_n \xrightarrow{L_2} a$ , es decir,  $a_n$  es un estimador consistente de  $a$ . Además, la sucesión  $n^{\frac{\rho}{2}}(a_n - E(a_n))$  es acotada en  $L_2$ .

OBSERVACIONES: Las hipótesis (15) y (16) implican que

$$0 < \lambda \leq \lambda_n \leq n^{-\rho} \sum_{i=1}^n \|x_i\|^2 \leq n^{1-\rho} \max_{i \leq n} \|x_i\|^2$$

y

$$n^{-\rho} \sum_{i=1}^n \|x_i\| \leq n^{-\rho} \left[ \sum_{i=1}^n \|x_i\|^2 \right]^{\frac{1}{2}} \leq n^{-\frac{\rho}{2}} \left( n^{1-\rho} \max_{i \leq n} \|x_i\|^2 \right)^{\frac{1}{2}}.$$

En consecuencia, las siguientes dos condiciones también se cumplen

$$\sum_{i=1}^n \|x_i\|^2 = O(n^\rho), \quad (17)$$

$$n^{-\rho} \sum_{i=1}^n \|x_i\| = O\left(n^{-\frac{\rho}{2}}\right). \quad (18)$$

DEMOSTRACION DEL TEOREMA: Para cada  $i \in \mathcal{N}$ , definamos las funciones indicadoras  $\delta_{i,0}$  ( $= 1$  si  $i \in I^{ng}$ ) y  $\eta_{i,j,h}$  ( $= 1$  si  $i \in I_j$  y  $z_i \in (c_{j,h-1}, c_{j,h})$ ),  $j = 1, \dots, s$  y  $h = 1, \dots, r_j$ .

Obsérvese que  $y(a)$  puede escribirse en la forma

$$y(a) = Xa + \varepsilon,$$

donde las componentes de  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^t$  son independientes y definidas como

$$\varepsilon_i = \delta_{i,0}\nu_i + \sum_{j=1}^s \sum_{h=1}^{r_j} \eta_{i,j,h} \gamma(-a^t x_i + c_{j,h-1}, -a^t x_i + c_{j,h}).$$

Sea  $F$  la función de distribución de probabilidad de los errores  $\nu$ . Está claro



que  $\varepsilon_i = \nu_i$  si  $i \in I^{ng}$  y, por el contrario, si  $i \in I^g$

$$\varepsilon_i = \gamma(-a^t x_i + c_{j,h-1}, -a^t x_i + c_{j,h}) \text{ con probabilidad } \pi_j q_{j,h}$$

(para  $j = 1, \dots, s, h = 1, \dots, r_j$ ), donde

$$q_{j,h} = F(-a^t x_i + c_{j,h}) - F(-a^t x_i + c_{j,h-1}).$$

Nótese, en primer lugar, que las esperanzas

$$E(\varepsilon_i) = \sum_{j=1}^s \sum_{h=1}^{r_j} \pi_j q_{j,h} \gamma(-a^t x_i + c_{j,h-1}, -a^t x_i + c_{j,h})$$

están acotadas, es decir,  $|E(\varepsilon_i)| \leq H < \infty, \forall i \in \mathcal{N}$ . En efecto, esto ocurre puesto que la función

$$g(z) = \sum_{j=1}^s \sum_{h=1}^{r_j} \pi_j [F(z + c_{j,h}) - F(z + c_{j,h-1})] \gamma(z + c_{j,h-1}, z + c_{j,h})$$

es continua y  $g(z) \rightarrow 0$  cuando  $|z| \rightarrow \infty$ . Igualmente, está claro que también las varianzas están acotadas, es decir,  $V(\varepsilon_i) \leq \bar{H} < \infty$ , para cada  $i \in \mathcal{N}$ .

Al igual que se hizo al probar el teorema anterior, escribáse

$$y(a_n) - y(a) = (I - M^*) X(a_n - a),$$

donde la matriz aleatoria  $M^*$  está asociada a  $(a_n, a)$  de la misma manera que  $M^{(p)}$  lo está a la pareja  $(a_p, a_{p-1})$  en la demostración del teorema 3.1. Todos los  $a_n$  cumplen (13), de donde

$$X^t \varepsilon = X^t M^* X(a_n - a).$$

A partir de (16), se obtiene que el mínimo autovalor de  $X^t M^* X$  no puede ser inferior a  $n^\rho \lambda_n \geq n^\rho \lambda > 0$ , por lo cual puede escribirse

$$a_n - a = (X^t M^* X)^{-1} X^t \varepsilon.$$

Así pues,

$$E \|a_n - a\|^2 \leq \lambda^{-2} n^{-2\rho} m \sum_{i,k} \|x_i\| |E(\varepsilon_i \varepsilon_k)| \|x_k\|.$$

Combinando (18) junto con que  $|E(\varepsilon_i \varepsilon_k)| = |E(\varepsilon_i)| |E(\varepsilon_k)| \leq \max\{H^2, \bar{H}\}$ , se concluye directamente que

$$E \|a_n - a\|^2 \leq \lambda^{-2} m \max\{H^2, 1\} n^{-2\rho} \left[ \sum_{i=1}^n \|x_i\| \right]^2 = O(n^{-\rho}). \quad (19)$$

Finalmente, obsérvese que

$$a_n - E(a_n) = (X^t M^* X)^{-1} X^t (\varepsilon - E(\varepsilon)) \quad (20)$$

y repítase un argumento similar al de antes (teniendo en cuenta que  $|E(\varepsilon_i - E(\varepsilon_i))(\varepsilon_k - E(\varepsilon_k))| = |E(\varepsilon_i - E(\varepsilon_i))| |E(\varepsilon_k - E(\varepsilon_k))| = 0$  y utilizando la propiedad (17)) para poder escribir

$$\begin{aligned} E \left\| n^{\frac{\rho}{2}} (a_n - E(a_n)) \right\|^2 &\leq n^\rho \lambda^{-2} n^{-2\rho} m \sum_{i=1}^n \|x_i\|^2 V(\varepsilon_i) \\ &\leq \lambda^{-2} m \bar{H} n^{-\rho} \sum_{i=1}^n \|x_i\|^2 \leq T < \infty, \end{aligned}$$

concluyendo así la demostración.  $\square$

Observemos que las condiciones de este teorema difieren de las impuestas para el algoritmo MD fundamentalmente en la no acotación de las variables independientes a medida que el tamaño de la muestra aumenta. En todo caso, esta hipótesis es plausible en situaciones prácticas en las que, por ejemplo, interviene el tiempo como variable del modelo, o alguna variable acumulativa.

A continuación se formulará un teorema central que refleja la ley de probabilidad límite del estimador propuesto  $a_n$ . La convergencia a una distribución normal del proceso primario del algoritmo modal MdD continúa cumpliéndose, en semejanza con los resultados obtenidos para el algoritmo más complejo MD. El resultado es el siguiente:

**Teorema 3.3** *Bajo las mismas hipótesis del teorema anterior, la sucesión  $n^{\frac{\rho}{2}}(a_n - E(a_n))$  converge en distribución*

$$n^{\frac{\rho}{2}}(a_n - E(a_n)) \xrightarrow[n \rightarrow \infty]{D} N(0, \Lambda)$$

para alguna matriz de covarianzas  $\Lambda$ . Así pues, supuesto  $n$  suficientemente grande, la distribución de  $a_n - a$  puede ser aproximada mediante

$$(a_n - a) \approx N(0, n^{-\rho}\Lambda).$$

DEMOSTRACION: Defínase la matriz  $M = \text{diag}(m_i)$  donde  $m_i = \frac{d}{d\beta} \delta_{j,r}(\beta)|_{\beta=-a^t x_i}$ , si  $i \in I_j$  y  $z_i \in (c_{j,r}, c_{j,r+1}]$ , y  $m_i = 1$ , si  $i \in I^{ng}$ . Observemos que, al contrario de lo que ocurre en el teorema 2.3,  $\frac{d}{d\beta} \delta_{j,r}(\beta)|_{\beta=-a^t x_i}$  toma los valores 0 ó 1, y que, por hipótesis, existe para todo  $i \in I^g$ .

Aceptemos que se cumplen los posteriores lema 3.4 y lema 3.5, cuya demostración se pospone para no desviar la atención sobre cuestiones adyacentes.

Se puede escribir a partir de los lemas citados

$$n^{-\frac{\rho}{2}} X^t E(M) X (a_n - E(a_n)) = n^{-\frac{\rho}{2}} X^t M^* X (a_n - E(a_n)) + \zeta_n,$$

con  $\zeta_n \xrightarrow{P} 0$ . De la misma forma que se vio en el teorema 2.3, la matriz simétrica  $n^{-\rho} X^t E(M) X$ , y su límite, si existe, son no singulares. A partir de la expresión (20), se tiene que

$$n^{-\frac{\rho}{2}} X^t E(M) X (a_n - E(a_n)) = n^{-\frac{\rho}{2}} X^t (\varepsilon - E(\varepsilon)) + \zeta_n.$$

Se sigue que

$$n^{\frac{\rho}{2}} (a_n - E(a_n)) = n^{\frac{\rho}{2}} (X^t E(M) X)^{-1} X^t (\varepsilon - E(\varepsilon)) + \zeta_n^*,$$

siendo  $\|\zeta_n^*\| = \left\| n^{\rho} (X^t E(M) X)^{-1} \zeta_n \right\| \leq \lambda^{-1} \|\zeta_n\| \xrightarrow{P} 0$ . Esto implica que la distribución asintótica del vector  $n^{\frac{\rho}{2}} (a_n - E(a_n))$  y la de  $n^{\frac{\rho}{2}} (X^t E(M) X)^{-1} X^t (\varepsilon - E(\varepsilon))$  tienen que coincidir. La  $j$ -ésima componente del último vector aleatorio es

$$\left[ n^{\frac{\rho}{2}} (X^t E(M) X)^{-1} X^t (\varepsilon - E(\varepsilon)) \right]_j = \sum_{i=1}^n n^{-\frac{\rho}{2}} f_j x_i (\varepsilon_i - E(\varepsilon_i)),$$

con  $f_j$  denotando la  $j$ -ésima fila de  $(n^{-\rho} X^t E(M) X)^{-1}$ . Operando de forma análoga a como se hizo en la última parte del teorema 2.3 y utilizando (16) y (17),

se puede llegar a que

$$\max_{i \leq n} V \left( n^{-\frac{\rho}{2}} f_j x_i (\varepsilon_i - E(\varepsilon_i)) \right) \leq \lambda^{-2} n^{-\rho} (\max_{i \leq n} \|x_i\|^2) = O(n^{-1})$$

y

$$\sum_{i=1}^n V \left( n^{-\frac{\rho}{2}} f_j x_i (\varepsilon_i - E(\varepsilon_i)) \right) \leq \lambda^{-2} n^{-\rho} \sum_{i=1}^n \|x_i\|^2 = O(1).$$

Esto indica que las condiciones del teorema central del límite de Lindeberg-Feller se cumplen, completándose así la demostración.  $\square$

**Lema 3.4** *Se verifica la siguiente convergencia en media de orden uno*

$$n^{-\frac{\rho}{2}} X^t (M - E(M)) X (a_n - E(a_n)) \xrightarrow[n \rightarrow \infty]{L_1} 0,$$

siendo  $M$  la matriz cuya definición figura al principio de la demostración del teorema 3.3.

DEMOSTRACION: Denótese en corto la matriz  $(p_{jk}) = n^{-\rho} X^t (M - E(M)) X$  y obsérvese que

$$\left\| n^{-\frac{\rho}{2}} X^t (M - E(M)) X (a_n - E(a_n)) \right\| \leq \sum_{j,k=1}^m |p_{jk}| \left\| n^{\frac{\rho}{2}} (a_n - E(a_n)) \right\|.$$

Por el teorema 3.2, será suficiente con demostrar que  $p_{jk} \xrightarrow{L_2} 0$ , puesto que por la desigualdad de Hölder

$$E \left[ \sum_{j,k} |p_{jk}| \left\| n^{\frac{\rho}{2}} (a_n - E(a_n)) \right\| \right] \leq \left( E \left( \sum_{j,k} |p_{jk}| \right)^2 E \left\| n^{\frac{\rho}{2}} (a_n - E(a_n)) \right\|^2 \right)^{\frac{1}{2}}$$

de donde se concluiría, por el teorema 3.2, que la parte derecha de la desigualdad converge a cero. Para todo  $j, k \in \{1, \dots, m\}$

$$p_{jk} = n^{-\rho} \sum_{i=1}^n x_{ij} x_{ik} (m_i - E(m_i)).$$

Finalmente, recordando que  $0 \leq m_i \leq 1$  y también la independencia de las  $m_i$ , puede concluirse, a partir de (16) y (17), que

$$E(p_{jk}^2) = n^{-2\rho} \sum_{i=1}^n x_{ij}^2 x_{ik}^2 E|m_i - E(m_i)|^2 \leq n^{-2\rho} \sum_{i=1}^n \|x_i\|^4$$

$$\leq n^{-2\rho} \max_{i \leq n} \|x_i\|^2 \sum_{i=1}^n \|x_i\|^2 = O(n^{-1}),$$

lo cual implica que  $p_{jk} \xrightarrow{L_2} 0$ .  $\square$

**Lema 3.5** *Se produce la siguiente convergencia en probabilidad*

$$n^{-\frac{\rho}{2}} X^t (M^* - M) X (a_n - E(a_n)) \xrightarrow[n \rightarrow \infty]{P} 0,$$

siendo  $M$  y  $M^*$  las matrices que figuran respectivamente en las demostraciones de los teoremas 3.3 y 3.2.

DEMOSTRACION: Nótese que, a partir de las hipótesis de diferenciabilidad establecidas al principio de la sección y también a partir de que  $a_n \xrightarrow{P} a$ , sucede que  $m_i - m_i^* \xrightarrow{P} 0$ . Definase la matriz  $(r_{jk}) = n^{-\rho} X^t (M^* - M) X$  y escríbase

$$\left\| n^{-\frac{\rho}{2}} X^t (M^* - M) X (a_n - E(a_n)) \right\|^2 \leq \left( \sum_{j,k}^m r_{jk}^2 \right) \left\| n^{\frac{\rho}{2}} (a_n - E(a_n)) \right\|^2$$

donde

$$|r_{jk}| \leq n^{1-\rho} \max_{i \leq n} \|x_i\|^2 n^{-1} \sum_{i=1}^n |m_i^* - m_i|.$$

Puesto que todas las  $m_i - m_i^*$  están acotadas, se sigue que  $r_{jk} \xrightarrow{P} 0$ , lo cual implica, utilizando el teorema 3.2, el límite en probabilidad que formula el teorema.

### 3.5 Estimación de la matriz de covarianzas asintótica

En esta sección se dará una estimación, calculable a partir de la información de que se dispone en cada momento, de la matriz de covarianzas asintótica  $\Lambda$ , que figura en el teorema 3.3. Esta estimación será útil para confeccionar intervalos de confianza para el parámetro  $a$ , o para realizar contrastes de hipótesis sobre su valor.

Considérese el siguiente vector funcional de error  $\varepsilon(\alpha) = (\varepsilon_1(\alpha), \dots, \varepsilon_n(\alpha))^t$ ,

dependiente del valor del vector  $\alpha \in \mathcal{R}^m$ , donde

$$\varepsilon_i(\alpha) = \delta_{i,0}\nu_i + \sum_{j=1}^s \sum_{h=1}^{r_j} \eta_{i,j,h} \gamma(-\alpha^t x_i + c_{j,h-1}, -\alpha^t x_i + c_{j,h}),$$

siendo  $\gamma$  la función de corrección en moda condicionada utilizada a lo largo de todo este capítulo. El vector  $\varepsilon$  definido en el curso de la demostración del teorema 3.2 coincide con  $\varepsilon(a)$ . Sea  $\Sigma(\varepsilon)$  la matriz de covarianzas de  $\varepsilon$  y asúmase la existencia de los siguiente dos límites:

$$\Phi = \lim_n \Phi_n = \lim_n n^{-\rho} X^t \Sigma(\varepsilon) X \quad (21)$$

$$W = \lim_n W_n = \lim_n n^{-\rho} X^t E(M) X. \quad (22)$$

Es fácil darse cuenta de que  $\Lambda = W^{-1} \Phi W^{-1}$ . Defínase la matriz  $n \times n$  dada por  $M_n = \text{diag} \left( m_i^{(n)} \right)$  donde  $m_i^{(n)} = \frac{d}{d\beta} \delta_{j,r}(\beta) |_{\beta=-a_i^t x_i}$ , si  $i \in I_j$  y  $z_i \in (c_{j,r}, c_{j,r+1}]$ , y  $m_i^{(n)} = 1$ , si  $i \in I^{ng}$ . Cuando la última derivada no exista, considérese que  $m_i^{(n)}$  es un valor arbitrario. Puesto que  $a_n \xrightarrow{P} a$ , también  $m_i^{(n)} - m_i \xrightarrow{P} 0$ , bajo las hipótesis hechas al principio de la sección 3.4.

Para concluir, denótese  $\Theta_n = n^{-\rho} X^t M_n X$  y  $\Pi_n = n^{-\rho} X^t \Sigma(\varepsilon(a_n)) X$ . Con todos los anteriores elementos se llega a la siguiente estimación consistente de  $\Lambda$ .

**Teorema 3.6** *Bajo las hipótesis del teorema 3.3 y asumiendo las condiciones límite (21) y (22), las matrices  $\Theta_n^{-1} \Pi_n \Theta_n^{-1}$  estiman consistentemente la matriz de covarianzas asintótica  $\Lambda$ .*

DEMOSTRACION: Es suficiente probar los siguientes límites en probabilidad

$$\|\Theta_n - W_n\| \xrightarrow{P} 0$$

y

$$\|\Pi_n - \Phi_n\| \xrightarrow{P} 0.$$


Ambas convergencias se prueban mediante un razonamiento similar al utilizado en la demostración del teorema 2.4.  $\square$

En definitiva, se puede decir que, supuesto que  $n$  sea suficientemente grande, la distribución de  $a_n$  es aproximadamente  $N(a, n^{-\rho} \Theta_n^{-1} \Pi_n \Theta_n^{-1})$ , distribución

útil, según se indicó, para hacer inferencias sobre el parámetro.

### Notas y comentarios

- Se han realizado distintas simulaciones con correcciones en moda, si bien su presentación será pospuesta hasta el final del capítulo 5, con el fin de mostrar, además, distintas comparaciones con otros procedimientos de corrección alternativos.
- En relación con las condiciones de no acotación para las variables independientes del teorema 3.1, es de resaltar que se podrían imponer condiciones de aleatorización sobre los intervalos de agrupación para relajar la no acotación citada. Véase al respecto las secciones 5.8 y 5.9. La misma condición de aleatorización podría conducir, adicionalmente, a que se mantuviera la igualdad  $E(\varepsilon_i) = 0$ , la cual, como se ha visto, se pierde a consecuencia de las correcciones modales aquí empleadas.



**Parte II**  
**Métodos de una iteración**



## **4. Algoritmos de una iteración basados en correcciones en media**

### **4.1 Introducción. Procesos de una iteración**

Los métodos iterativos MD y MdD (de dos iteraciones) presentados en la Parte I de esta memoria están concebidos en la misma línea que los estimadores usuales. Su novedad estriba en la concepción del proceso secundario, si bien las propiedades asintóticas de los estimadores resultantes se establecen siguiendo un esquema ortodoxo. Por ejemplo, los estimadores (usuales) máximo verosímiles pueden identificarse como algoritmos de iteración doble, en la mayoría de las situaciones reales donde se aplican. La iteración secundaria reflejaría el proceso secuencial por el que se ha optado para obtener el óptimo de la verosimilitud a tamaño de muestra dado, cuando dicho óptimo no admita expresión cerrada. Desde luego, esta última situación es habitual en la mayoría de los casos prácticos, cuando existen datos incompletos o censurados como los tratados en esta memoria. Por su parte, el proceso primario consistiría en hacer crecer el tamaño muestral  $n$ , buscando algún tipo de convergencia estocástica (por general, en ley), cuando  $n \rightarrow \infty$ .

En el mejor de los casos, los estimadores usuales presentan propiedades similares a las que han sido demostradas para los algoritmos MD y MdD antes citados. Por ejemplo, en el caso de los estimadores máximo verosímiles, la propiedad de trayectorias únicas (independientes de los puntos de arranque

elegidos del proceso secundario en cada etapa primaria) sólo está garantizada si la sucesión funcional de verosimilitudes cumple, por ejemplo, condiciones de concavidad y óptimo único, término a término. De no ser así, los procesos iterativos secundarios (obtenidos mediante algoritmos usuales, e. g., Newton-Raphson, EM, etc.) sólo garantizan convergencia a puntos críticos de las verosimilitudes, dependiendo del punto de arranque que se haya elegido. Los citados algoritmos secundarios pueden, incluso, ciclar, algo que nunca ocurre con los métodos MD y MdD presentados en los capítulos precedentes.

Todos los procesos iterativos con anidamiento secundario presentan desde un punto de vista práctico varios inconvenientes. Entre ellos:

1. Los procesos secundarios nunca se pueden prolongar infinitamente. Ello obliga a imponer una condición de parada que habitualmente se establece mediante la existencia de un umbral  $\varepsilon > 0$  (por ejemplo,  $\varepsilon = 10^{-6}$  ó  $\varepsilon = 10^{-3}$ ) de modo que se termina el proceso si la distancia entre dos términos sucesivos de la sucesión generada, digamos  $d(a_{p-1}, a_p)$ , no sobrepasa el citado umbral. Es evidente que este criterio de parada no garantiza un nivel de aproximación al punto buscado, digamos  $a$ , por debajo del umbral antes citado, puesto que ello implicaría que  $d(a_p, a) < \varepsilon$ . En resumen, la condición de parada sólo se justifica en términos pragmáticos (operativamente hablando) pudiendo ser arbitraria en ciertos casos. Ahora bien, puesto que los resultados asintóticos del proceso primario son válidos exclusivamente para la sucesión de puntos límite exactos del proceso secundario, se sigue que pueden existir discrepancias claras por esta vía a nivel práctico.
2. En sentido estricto, aun obviando la limitación anterior, el proceso de iteración primario se debería ejecutar en la misma línea. Así, cuando se utilizan estimadores máximo varosímiles (al igual que ocurre si se usaran los métodos MD y/o MdD), uno debería tener elementos muestrales de holgura para

contrastar (¡cuanto menos!) una cierta regla de parada. ¿Por qué no se hace así? Por un lado, porque habría que establecer dicha regla frente a convergencias estocásticas (¿cúal?). En segundo lugar, porque el proceso secundario no tiene memoria al iterar en el primario. Obsérvese: cuando se añade un nuevo elemento muestral, el nuevo óptimo de la verosimilitud ampliada no es calculable a partir del anterior. En consecuencia, los cálculos deben repetirse desde el principio, con el consiguiente deterioro en tiempo de cálculo que esto supone. El resultado final es que, por lo general, no se itera en el proceso primario, asumiéndose que, para un tamaño de muestra dado, uno se encuentra próximo al infinito que exige el resultado asintótico. ¿Por qué?

Una posibilidad para atajar este problema descansa en sustituir la doble iteración por una sola, de la misma forma que, existiendo un límite doble  $a_{m,n} \rightarrow a$ , uno intenta obtenerlo a través de un camino, digamos  $m = n$  o, en general,  $m = m(n)$ . Ésta es la idea matriz que subyace en los planteamientos llevados a cabo por la vía de las redes neuronales (véase WHITE (1992) y HAYKIN (1994), por ejemplo) y por las aproximaciones estocásticas tipo Kushner (véase KUSHNER Y YIN (1997), WASAN (1969) y BENVENISTE, METIVIER Y PRIOURET (1990), entre otros). En el contexto de estimación recursiva con datos incompletos o censurados, como se trata en esta memoria, no existen antecedentes. Los capítulos 4 y 5 que constituyen la Parte II de este trabajo, exploran distintas variantes de esta vía inédita de iteración única. En este sentido, el actual capítulo 4 prolonga el 2, considerando exclusivamente correcciones en media. Por su parte, el capítulo 5 constituye una similar extensión del 3.

## 4.2 Formalización del modelo

Como se señaló en la introducción, en este capítulo se presentan y analizan distintas variantes de una iteración para el proceso iterativo de estimación MD

introducido en el capítulo 2. Como viene siendo habitual, se centrará la atención sobre el modelo lineal

$$z_i = a^t x_i + \nu_i, \quad i = 1, \dots, n \quad (23)$$

donde tanto el parámetro  $a$ , que tiene que ser estimado, como las variables independientes  $x_i$  son  $m$ -dimensionales, y los errores  $\nu_i$  son independientes, con media cero y varianza finita. Supondremos que  $\nu_i$  son variables aleatorias con recorrido toda la recta real, o al menos un conjunto denso de ella. Sin pérdida de generalidad se supondrá que los errores tienen varianza uno. De nuevo, se asumirá que la información que se conoce de la muestra obtenida es parcial, pudiendo algunos datos  $z_i$  ser agrupados con criterios dispares. Se mantendrá la notación empleada en la sección 2.2. El conjunto de observaciones hasta la etapa  $n$ ,  $I_n = \{1, \dots, n\}$ , se dividirá en  $I_n^g$  y en  $I_n^{ng}$ , conteniendo respectivamente aquellos  $i$ 's cuyos valores muestrales  $z_i$  han sido agrupados y no lo han sido. Por otra parte, denotaremos por  $I_n^g = I_1^g \cup \dots \cup I_n^g$  los diferentes criterios de agrupación. Para cada  $i \in I_n^g$ , se asumirá que solamente se conoce el criterio de clasificación y el intervalo que contiene a  $z_i$ . Por el contrario si,  $i \in I_n^{ng}$ , se observa el valor exacto  $z_i$ . Obsérvese que  $I_j^n \subset I_j^{n+1}$  y  $I_n^{ng} \subset I_{n+1}^{ng}$ , y llamemos  $I_j = \lim_n I_j^n$ ,  $I^g = \lim_n I_n^g$  y  $I^{ng} = \lim_n I_n^{ng}$ .

Finalmente, de aquí en adelante se asumirá, siempre que sea necesario, que  $X^t X$  es una matriz de rango completo, donde  $X^t = (x_1, \dots, x_n)$ . Esta última condición no será necesaria, por ejemplo, para el algoritmo que se propone en la sección 4.6.

A lo largo del capítulo se sugiere buscar una estimación de  $a$ , en el modelo (23), mediante una serie de procedimientos recursivos de una iteración basados en correcciones en esperanzas condicionadas. De esta forma, se intenta atajar la problemática innata de los métodos en dos iteraciones citada en la introducción. La doble iteración propia del MD se evitará sustituyendo las infinitas iteraciones del

proceso secundario por una sola. Se probará que este método sigue manteniendo buenas propiedades asintóticas de convergencia. De hecho se probará la convergencia en media de orden uno, así como un teorema central similar al que se obtuvo para el procedimiento MD puro.

La simplificación puede llevarse aun más lejos. Obsérvese que, en el algoritmo propuesto en la sección 2.2, el paso principal de actualización del proceso secundario corresponde a

$$a_{p+1} = (X^t X)^{-1} X^t y(a_p).$$

Es decir, en cada iteración del proceso primario es necesario calcular la inversa de la matriz  $X^t X$ , y esto puede resultar costoso, sobre todo si  $m$  es grande. En la sección 4.5 se tratará de evitar ese cálculo, sustituyendo simplemente la matriz por una tamaño de paso real y positivo, que se elegirá adecuadamente. Esta idea es la que subyace en las técnicas de aproximaciones estocásticas del tipo de KUSHNER Y YIN (1997).

Por último, en la sección 4.6 se desarrolla un tercer algoritmo que simplifica aún más el anterior. Obsérvese que el vector  $y(a)$  es de tamaño  $n$ , involucrando todos los valores muestrales hasta esa etapa. En esta sección se trata de utilizar exclusivamente la información que aporta el último valor muestral obtenido, de nuevo en la línea de las técnicas de aproximaciones estocásticas. Son claras las ventajas computacionales de este método, puesto que en cada iteración sólo se necesita computar un producto con un total de  $m$  factores.

Pese a las enormes simplificaciones citadas, tanto en la sección 4.5 como en la 4.6 se continúa garantizando consistencia de las estimaciones, habiéndose probado, además, la convergencia en media de orden uno.

### 4.3 Un primer algoritmo de una iteración. Algoritmo

## MD de una iteración

Como se ha indicado, este primer método constituye una variante de iteración única del método MD desarrollado en el capítulo 2. Del proceso secundario se ejecuta una sola iteración, utilizándose su salida como punto inicial en la etapa primaria siguiente a la actual.

Formalmente, el nuevo método iterativo es el siguiente:

INICIALIZACIÓN: Tómese un vector inicial  $a_1$  arbitrario.

ITERACIÓN: Si  $a_n$  es la estimación en la etapa primaria  $n$  del verdadero vector  $a$ , la nueva estimación para la etapa  $n + 1$  será

$$a_{n+1} = (X^t X)^{-1} X^t y(a_n),$$

donde  $X^t = (x_1, \dots, x_n)$  es una matriz de tamaño  $m \times n$ , y el vector  $y(a_n) = (y_1(a_n), \dots, y_n(a_n))^t$  tiene por componentes

$$\begin{aligned} y_i(a_n) &= z_i, & \text{si } i \in I^{ng} \\ &= a_n^t x_i + \gamma(-a_n^t x_i + c_{j,r}, -a_n^t x_i + c_{j,r+1}), & \text{si } i \in I_j, z_i \in (c_{j,r}, c_{j,r+1}]. \end{aligned}$$

Obsérvese que en sentido estricto se debería escribir  $X = X_n$  para hacer explícita la dependencia de  $X$  del tamaño muestral  $n$ , si bien suprimiremos el subíndice de aquí en adelante. Para cada  $j = 1, \dots, s$  y  $r = 0, 1, \dots, r_j - 1$ , la función  $\gamma(-a_n^t x_i + c_{j,r}, -a_n^t x_i + c_{j,r+1})$  se corresponde con una corrección en esperanza condicionada, es decir, coincide con el valor esperado del error  $\nu$ , condicionado a que el parámetro coincide con su actual estimación  $a_n$  y a que  $\nu_i$  pertenece al intervalo correspondiente, cuyos extremos figuran como argumentos de la función  $\gamma$ . En concreto, si llamamos  $\delta_{jr}(\beta) = \gamma(\beta + c_{j,r}, \beta + c_{j,r+1})$  se tendrá que

$$\delta_{jr}(\beta) = E(\nu | \nu \in (\beta + c_{j,r}, \beta + c_{j,r+1})).$$

El algoritmo propuesto se denotará de aquí en adelante, pese a la inconsistencia del término, algoritmo MD de una iteración. Obsérvese al respecto que la D se

introdujo para referirse a la doble iteración que está implícita en el MD puro. El nombre MU sería más correcto para nombrar esta variante simplificadora (de iteración Única) contemplada en este capítulo. Pese a ello, se ha preferido hacer explícito con el nombre elegido que procede del MD.

Asúmase en lo sucesivo la condición técnica

$$\delta_{jr}(\beta) \text{ y } \beta - \delta_{jr}(\beta) \text{ son no decrecientes} \quad (\text{CONDICION C1})$$

cualquiera que sean  $j = 1, \dots, s$  y  $r = 0, 1, \dots, r_j - 1$ . Esta condición restringe en la práctica la clase de distribuciones de error  $\nu$  a las que se puede aplicar el método propuesto de estimación, en el sentido de poder asegurar la convergencia. En cualquier caso, una buena parte de las distribuciones usuales cumplen esta condición según se indicó en la sección 2.2.

El principal resultado de convergencia del método MD de una iteración se enuncia a continuación. Se asumen unas ciertas condiciones que afectan solamente a los valores  $x_i$  de las variables independientes, llegándose a asegurar la convergencia en media de orden uno y, en consecuencia, la convergencia en probabilidad. Recordemos que con el método MD puro de dos iteraciones se obtiene, al menos, convergencia en media de orden dos. En cualquier caso, el coste de una iteración con este nuevo método es mínimo, manteniéndose igualmente la consistencia.

**Teorema 4.1** *Aceptemos que  $(X^t X)^{ng} = \sum_{i \in I_n^{ng}} x_i x_i^t y (X^t X)^g = \sum_{i \in I_n^g} x_i x_i^t$  son matrices definidas positivas para cada  $n \in \mathcal{N}$ , y que, además, se cumplen las siguientes condiciones*

$$\inf_n \lambda_n = \lambda > 0,$$

$$\|x_i\| \leq K < +\infty, \quad \text{para cada } i \in \mathcal{N},$$

donde  $\lambda_n$  denota el mínimo autovalor de la matriz  $n^{-1}(X^t X)^{ng}$ . En estas condiciones,  $a_n \xrightarrow{L_1} a$  y, en consecuencia,  $a_n$  es un estimador consistente de  $a$ .

OBSERVACION: La matriz  $n^{-1}(X^t X)^g$  es cuadrada de orden  $m$ , simétrica y definida positiva. De hecho, su máximo autovalor coincide con su traza. Si

ponemos, por otra parte,

$$n^{-1}(X^t X)^g = n^{-1} \sum_{i \in I_n^g} x_i x_i^t,$$

se obtiene claramente que la traza de esa matriz coincide con

$$n^{-1} \sum_{i \in I_n^g} \|x_i\|^2.$$

Denotando por  $\delta_n$  el máximo autovalor de la matriz  $n^{-1}(X^t X)^g$ , se tendrá, a partir de la segunda hipótesis del teorema, que

$$\delta_n = n^{-1} \sum_{i \in I_n^g} \|x_i\|^2 \leq K^2 < \infty.$$

En definitiva, para la demostración posterior, escribábase

$$\sup_n \delta_n = M < \infty.$$

DEMOSTRACION DEL TEOREMA: Defínase las siguientes funciones indicatrices:  $\delta_{i,0} = 1$  si  $i \in I^{ng}$ ;  $\eta_{i,j,r} = 1$  si  $i \in I_j \subset I^g$  y  $z_i \in (c_{j,r}, c_{j,r+1}]$ , donde  $i \in \mathcal{N}$ ,  $j = 1, \dots, s$  y  $r = 0, 1, \dots, r_j - 1$ . Obsérvese que  $y(a)$  puede escribirse como

$$y(a) = Xa + \varepsilon(a),$$

donde las componentes de  $\varepsilon(a) = (\varepsilon_1(a), \dots, \varepsilon_n(a))^t$  son independientes y definidas como

$$\varepsilon_i(a) = \delta_{i,0} \nu_i + \sum_{j=1}^s \sum_{r=0}^{r_j-1} \eta_{i,j,r} \delta_{jr} (-a^t x_i).$$

Por analogía, se puede definir

$$\varepsilon(a_n) = y(a_n) - Xa_n.$$

Según la definición de la sucesión  $a_n$ , se puede escribir

$$a_{n+1} = (X^t X)^{-1} X^t y(a_n) = a_n + (X^t X)^{-1} X^t \varepsilon(a_n). \quad (24)$$

Teniendo en cuenta el lema 4.3 posterior, si  $i \in I_j \subset I^g$  y  $z_i \in (c_{j,r}, c_{j,r+1}]$

$$y_i(a_n) - y_i(a) = a_n^t x_i - a^t x_i + \delta_{jr} (-a_n^t x_i) - \delta_{jr} (-a^t x_i)$$



4. Algoritmos de una iteración basados en correcciones en media

$$\begin{aligned} &= a_n^t x_i - a^t x_i + m_i^* (-a_n^t x_i + a^t x_i) \\ &= (1 - m_i^*) (a_n^t x_i - a^t x_i). \end{aligned}$$

Si, por el contrario,  $i \in I^{ng}$

$$y_i(a_n) - y_i(a) = 0.$$

En definitiva, se puede concluir que

$$y(a_n) - y(a) = (I - M^*) X(a_n - a),$$

donde  $M^* = \text{diag}(m_i^*)$ , cumpliéndose que  $m_i^* = 1$ , si  $i \in I^{ng}$ , y  $0 \leq m_i^* \leq 1$ , si  $i \in I^g$ . La diferencia entre los errores en  $a_n$  y  $a$  quedaría

$$\begin{aligned} \varepsilon(a_n) - \varepsilon(a) &= -Xa_n + Xa + (I - M^*) X(a_n - a) \\ &= -M^* X(a_n - a). \end{aligned}$$

Por consiguiente,

$$a_{n+1} = a_n - (X^t X)^{-1} X^t M^* X(a_n - a) + (X^t X)^{-1} X^t \varepsilon(a),$$

o bien

$$a_{n+1} - a = \left[ I - (X^t X)^{-1} X^t M^* X \right] (a_n - a) + (X^t X)^{-1} X^t \varepsilon(a).$$

Esta última expresión es mejor escribirla, equivalentemente, en los siguientes términos

$$a_{n+1} - a = (X^t X)^{-1} X^t (I - M^*) X(a_n - a) + (X^t X)^{-1} X^t \varepsilon(a). \quad (25)$$

Obsérvese que  $E(\varepsilon_i(a)) = 0$ . Para verlo recuérdese que

$$E(\nu | \nu \in A) = \frac{1}{\text{Pr}(A)} \int_A x dF(x),$$

donde  $F$  es la función de distribución asociada al error  $\nu$ . En consecuencia, se puede poner

$$E(\varepsilon_i(a)) = \pi_0 E(\nu_i)$$

$$\begin{aligned}
 & + \sum_{j=1}^s \pi_j \sum_{r=0}^{r_j-1} \Pr(\nu_i \in (-a^t x_i + c_{j,r}, -a^t x_i + c_{j,r+1}]) \delta_{j,r}(-a^t x_i) \\
 = & \pi_0 E(\nu_i) + \sum_{j=1}^s \pi_j \sum_{r=0}^{r_j-1} \int_{(-a^t x_i + c_{j,r}, -a^t x_i + c_{j,r+1}]} x dF_{\nu_i}(x) \\
 = & \pi_0 E(\nu_i) + \sum_{j=1}^s \pi_j \int_R x dF_{\nu_i}(x) \\
 = & \pi_0 E(\nu_i) + \sum_{j=1}^s \pi_j E(\nu_i) \\
 = & E(\nu_i) = 0,
 \end{aligned}$$

sin más que tener en cuenta que los errores  $\nu_i$  son todos de media cero. Como se puede ver, utilizar las correcciones en media condicionada produce errores muestrales que tienen esperanza nula.

Calculando la norma euclídea de la expresión (25) y aplicando la desigualdad triangular, resulta que

$$\|a_{n+1} - a\| \leq \mu_n \|a_n - a\| + \left\| (X^t X)^{-1} X^t \varepsilon(a) \right\|,$$

siendo  $\mu_n$  el mayor autovalor de la matriz definida positiva

$$(X^t X)^{-1} X^t (I - M^*) X.$$

Los autovalores de esta matriz coinciden con los de

$$\begin{aligned}
 & (X^t X)^{\frac{1}{2}} (X^t X)^{-1} X^t (I - M^*) X (X^t X)^{-\frac{1}{2}} \\
 = & (X^t X)^{-\frac{1}{2}} X^t (I - M^*) X (X^t X)^{-\frac{1}{2}}.
 \end{aligned}$$

Resulta, pues, que se puede poner  $\mu_n$  de la siguiente forma

$$\mu_n = \max_{\|u\|=1} \left| \frac{u^t X^t (I - M^*) X u}{u^t X^t X u} \right| = \max_{\|u\|=1} \left| \frac{\sum_{i \in I_n^g} (1 - m_i^*) u^t x_i x_i^t u}{u^t X^t X u} \right|.$$

Recordando que para cada  $i \in I_n^g$  es  $0 \leq 1 - m_i^* \leq 1$ , y que tanto  $\sum_{i \in I_n^g} x_i x_i^t$  como  $X^t X$  son matrices definidas positivas, resulta que la última expresión está

acotada superiormente por

$$\begin{aligned} \max_{\|u\|=1} \frac{\sum_{i \in I_n^g} u^t x_i x_i^t u}{u^t X^t X u} &= \max_{\|u\|=1} \left( \frac{u^t X^t X u}{\sum_{i \in I_n^g} u^t x_i x_i^t u} \right)^{-1} \\ &= \max_{\|u\|=1} \left( 1 + \frac{\sum_{i \in I_n^g} u^t x_i x_i^t u}{\sum_{i \in I_n^g} u^t x_i x_i^t u} \right)^{-1} \\ &\leq \left( 1 + \frac{\min_{\|u\|=1} \sum_{i \in I_n^g} u^t x_i x_i^t u}{\max_{\|u\|=1} \sum_{i \in I_n^g} u^t x_i x_i^t u} \right)^{-1}. \end{aligned}$$

Puesto que

$$\min_{\|u\|=1} \sum_{i \in I_n^g} u^t x_i x_i^t u = \min_{\|u\|=1} u^t (X^t X)^{ng} u = n\lambda_n > n\lambda > 0$$

y

$$\max_{\|u\|=1} \sum_{i \in I_n^g} u^t x_i x_i^t u = \max_{\|u\|=1} u^t (X^t X)^g u = n\delta_n < nM < \infty,$$

basta tomar

$$\tau = \left( 1 + \frac{\lambda}{M} \right)^{-1}$$

para poder afirmar que

$$\mu_n \leq \tau < 1.$$

Así pues, se concluye que

$$\|a_{n+1} - a\| \leq \tau \|a_n - a\| + \left\| (X^t X)^{-1} X^t \varepsilon(a) \right\|.$$

El mínimo autovalor de la matriz definida positiva  $X^t X$  es no inferior al mínimo autovalor de  $(X^t X)^{ng}$ , el cual coincide con  $n\lambda_n$ . Observemos, por tanto, que si  $c_n = \left\| (X^t X)^{-1} X^t \varepsilon(a) \right\|$ , se tiene que

$$\begin{aligned} E(c_n^2) &\leq \lambda^{-2} n^{-2} E \left\| X^t \varepsilon(a) \right\|^2 \\ &= \lambda^{-2} n^{-2} E [\varepsilon(a)^t X X^t \varepsilon(a)] \\ &= \lambda^{-2} n^{-2} E \left[ \sum_{i,j=1}^n \varepsilon_i(a) \varepsilon_j(a) x_i^t x_j \right]. \end{aligned}$$

Finalmente, teniendo en cuenta primero que, por la independencia de las observaciones y por la propiedad de la esperanza nula de los errores de

observación, se verifica

$$E(\varepsilon_i(a)\varepsilon_j(a)) = E(\varepsilon_i(a))E(\varepsilon_j(a)) = 0,$$

y, utilizando después la hipótesis de acotación de las variables independientes  $x_i$ , resulta que

$$\begin{aligned} E(c_n^2) &\leq \lambda^{-2}n^{-2} \sum_{i=1}^n E(\varepsilon_i(a)^2) \|x_i\|^2 \\ &\leq \lambda^{-2}n^{-2}K^2 \sum_{i=1}^n E(\varepsilon_i(a)^2). \end{aligned}$$

Se puede también calcular la varianza de los errores de observación. Ésta coincide con

$$\begin{aligned} E(\varepsilon_i(a)^2) &= \pi_0 E(\nu_i^2) \\ &\quad + \sum_{j=1}^s \pi_j \sum_{r=0}^{r_j-1} \Pr(\nu_i \in -a^t x_i + (c_{j,r}, c_{j,r+1}]) (\delta_{j,h}(-a^t x_i))^2 \\ &\leq \pi_0 E(\nu_i^2) + \sum_{j=1}^s \pi_j \sum_{r=0}^{r_j-1} \int_{(-a^t x_i + c_{j,r}, -a^t x_i + c_{j,r+1}]} x^2 dF_{\nu_i}(x) \\ &= \pi_0 E(\nu_i^2) + \sum_{j=1}^s \pi_j \int_R x^2 dF_{\nu_i}(x) \\ &= \pi_0 E(\nu_i^2) + \sum_{j=1}^s \pi_j E(\nu_i^2) \\ &= E(\nu_i^2) = \text{Var}(\nu_i) = 1. \end{aligned}$$

La desigualdad que aparece arriba simplemente se deriva de la conocida propiedad de las esperanzas condicionadas, válida para cualquier suceso  $A$

$$E(\nu|A)^2 \leq E(\nu^2|A).$$

Teniendo en cuenta la anterior acotación de la varianza de los errores, se cumple que

$$E(c_n^2) \leq \lambda^{-2}n^{-1}K^2 \xrightarrow{n \rightarrow \infty} 0.$$

En definitiva, se tiene que

$$E \|a_{n+1} - a\| \leq \tau E \|a_n - a\| + E(c_n),$$

donde  $0 < \tau < 1$  y donde

$$d_n = E(c_n) \leq (E(c_n^2))^{\frac{1}{2}} \xrightarrow{n \rightarrow \infty} 0.$$

Iterando la desigualdad anterior  $n$  veces, resulta

$$E \|a_{n+1} - a\| \leq \tau^n E \|a_1 - a\| + \sum_{i=1}^n \tau^{n-i} d_i.$$

Finalmente, aplicando el lema 4.2 posterior, se puede concluir que

$$E \|a_n - a\| \xrightarrow{n \rightarrow \infty} 0,$$

con lo cual, también  $a_n \xrightarrow{P} a$  y se termina la demostración.  $\square$

**Lema 4.2** Si  $0 < \tau < 1$  y  $\{d_i\}$  es una sucesión de valores positivos, tal que  $d_i \xrightarrow{i \rightarrow \infty} 0$ , entonces

$$\sum_{i=1}^n \tau^{n-i} d_i \xrightarrow{n \rightarrow \infty} 0.$$

DEMOSTRACION: Se sabe que  $\sum_{i=1}^{\infty} \tau^i = \frac{1}{1-\tau} = M < \infty$ . Fijado  $\varepsilon > 0$ , se puede encontrar  $n_0$  tal que  $0 < d_n < \frac{\varepsilon}{2M}$ , para cada  $n \geq n_0$ . Tómesese después  $n_1 \geq n_0$  verificando para cada  $n \geq n_1$

$$d_1 \tau^{n-1} + d_2 \tau^{n-2} + \dots + d_{n_0-1} \tau^{n-n_0+1} < \frac{\varepsilon}{2}.$$

Por consiguiente, para cada  $n \geq n_1$

$$\begin{aligned} \sum_{i=1}^n \tau^{n-i} d_i &= \sum_{i=1}^{n_0-1} \tau^{n-i} d_i + \sum_{i=n_0}^n \tau^{n-i} d_i \\ &\leq \sum_{i=1}^{n_0-1} \tau^{n-i} d_i + \frac{\varepsilon}{2M} \sum_{i=n_0}^n \tau^{n-i} \\ &\leq \sum_{i=1}^{n_0-1} \tau^{n-i} d_i + \frac{\varepsilon}{2} < \varepsilon. \square \end{aligned}$$

**Lema 4.3** La CONDICION C1 asegura que cualquiera que sean  $j, r \in \mathcal{N}$  y  $a, b \in R$ , existe  $m^* \in [0, 1]$  tal que

$$\delta_{jr}(b) - \delta_{jr}(a) = m^*(b - a).$$

DEMOSTRACION: Si  $a > b$

$$a - \delta_{jr}(a) \geq b - \delta_{jr}(b)$$

y así

$$0 \leq \delta_{jr}(a) - \delta_{jr}(b) \leq (a - b),$$

con lo cual existirá un valor  $m^* \in [0, 1]$ , dependiente de  $a$  y  $b$ , tal que

$$\delta_{jr}(a) - \delta_{jr}(b) = m^*(a - b).$$

Idénticamente, se llega a la misma conclusión si  $a < b$ .  $\square$

#### 4.4 Resultado de convergencia en distribución

Se enunciará a continuación un teorema que demuestra que la estimación propuesta por el procedimiento iterativo de estimación MD en una sola etapa tiene distribución normal en el límite.

**Teorema 4.4** Bajo las condiciones del teorema 4.1 existe una matriz de covarianzas  $\Sigma$  tal que

$$\sqrt{n}(a_n - a) \xrightarrow[n \rightarrow \infty]{D} N(0, \Sigma).$$

DEMOSTRACION: Llamemos  $X_n^t = (x_1, \dots, x_n)$  para hacer explícita la dependencia de  $n$ . Definamos  $A_n^* = (X_n^t X_n)^{-1} X_n^t (I - M^*) X_n$  y  $A_n = E(A_n^*) = (X_n^t X_n)^{-1} X_n^t (I - E(M^*)) X_n$ . Teniendo en cuenta la expresión (25) de la demostración del teorema anterior y llamando  $\varepsilon^n(a) = (\varepsilon_1(a), \dots, \varepsilon_n(a))$ , de nuevo para notar su dependencia de  $n$ , se puede poner

$$\begin{aligned} a_{n+1} - a &= A_n^*(a_n - a) + (X_n^t X_n)^{-1} X_n^t \varepsilon^n(a) \\ &= A_n(a_n - a) + (X_n^t X_n)^{-1} X_n^t \varepsilon^n(a) \end{aligned}$$

4. Algoritmos de una iteración basados en correcciones en media

$$+ (A_n^* - A_n) (a_n - a).$$

Probaremos que  $n^{\frac{1}{2}} (A_n^* - A_n) (a_n - a) \xrightarrow{P} 0$ . En efecto, para ello, en primer lugar, denotemos por  $p_{jk} = n^{-1} \sum_{i=1}^n x_{ij} x_{ik} (m_i^* - E(m_i^*))$  a la componente  $j, k = 1, \dots, m$  de la matriz cuadrada  $n^{-1} X_n (M^* - E(M^*)) X_n^t$ . Como  $a_n \xrightarrow{P} a$ , resulta que

$$|p_{jk}| \leq K^2 n^{-1} \sum_{i=1}^n |m_i^* - E(m_i^*)| \xrightarrow{P} 0.$$

En segundo lugar, probaremos que  $\left\| n^{\frac{1}{2}} (a_n - a) \right\|$  esta acotado en  $L_1$ . Para ello recuérdese que se puede escribir

$$E \left\| n^{\frac{1}{2}} (a_{n+1} - a) \right\| \leq n^{\frac{1}{2}} \tau^n E \|a_1 - a\| + \sum_{i=1}^n \tau^{n-i} n^{\frac{1}{2}} d_i,$$

donde  $d_n \leq \lambda^{-1} K n^{-\frac{1}{2}} \xrightarrow{n \rightarrow \infty} 0$ . En la expresión anterior, el primer término converge a cero,  $n^{\frac{1}{2}} \tau^n E \|a_1 - a\| \xrightarrow{n \rightarrow \infty} 0$ , y el segundo término está acotado,  $\sum_{i=1}^n \tau^{n-i} n^{\frac{1}{2}} d_i \leq \lambda^{-1} K \sum_{i=1}^n \tau^{n-i} \leq \lambda^{-1} K (1 - \tau)^{-1}$ . En consecuencia, queda demostrada la acotación en  $L_1$  citada.

Por último, observemos que

$$\left\| n^{\frac{1}{2}} (A_n^* - A_n) (a_n - a) \right\| \leq \lambda^{-1} \left\| n^{\frac{1}{2}} (a_{n+1} - a) \right\| \sum_{j,k=1}^m |p_{jk}|$$

lo cual permite afirmar la convergencia a cero en probabilidad del primer término, como se pretendía demostrar.

Queda probado, pues, que la distribución asintótica de  $n^{\frac{1}{2}} (a_n - a)$  coincide con la de  $A_n (a_n - a) + (X_n^t X_n)^{-1} X_n^t \varepsilon^n(a)$ . Iterando  $n$  veces se puede ver que, además, coincide con la distribución asintótica de

$$\sqrt{n} \prod_{i=1}^n A_i (a_1 - a) + \sqrt{n} \sum_{i=1}^n \left[ \prod_{j=i}^n A_j \right] (X_i^t X_i)^{-1} X_i^t \varepsilon^i(a).$$

Obsérvese ahora que cualquiera que sea el vector  $b \in \mathcal{R}^m$  y  $n \in \mathcal{N}$  resulta que

$$\|A_n b\| \leq \tau \|b\|,$$

siendo  $\tau = \left(1 - \frac{\lambda}{M}\right)^{-1} < 1$  el valor definido a lo largo de la demostración del teorema 4.1. Entonces

$$\sqrt{n} \left\| \prod_{i=1}^n A_i b \right\| \leq \sqrt{n} \tau^n \|b\| \xrightarrow{n \rightarrow \infty} 0.$$

Directamente de esta propiedad se tiene que  $\sqrt{n} \prod_{i=1}^n A_i (a_1 - a) \xrightarrow{P} 0$ , resultando que la distribución límite de  $\sqrt{n} (a_{n+1} - a)$  coincide con la de  $\sqrt{n} \sum_{i=1}^n \left[ \prod_{j=i}^n A_j \right] (X_i^t X_i)^{-1} X_i^t \varepsilon^i(a)$ , que será la que se estudiará a continuación. Este último término se puede escribir de la siguiente manera:

$$\begin{aligned} & \sqrt{n} \sum_{i=1}^n \left[ \prod_{j=i}^n A_j \right] (X_i^t X_i)^{-1} X_i^t \varepsilon^i(a) & (27) \\ &= \sqrt{n} \sum_{i=1}^n \left[ \prod_{j=i}^n A_j \right] (X_i^t X_i)^{-1} \sum_{k=1}^n x_k \varepsilon_k(a) \\ &= \sqrt{n} \sum_{k=1}^n \sum_{i=k}^n \left[ \prod_{j=i}^n A_j \right] (X_i^t X_i)^{-1} x_k \varepsilon_k(a). \end{aligned}$$

Llamemos  $\sum_{i=k}^n \left[ \prod_{j=i}^n A_j \right] (X_i^t X_i)^{-1} = C_n^k$ . Se puede escribir la siguiente cadena de desigualdades

$$\begin{aligned} \|C_n^k x_k\| &\leq \sum_{i=k}^n \left\| \left[ \prod_{j=i}^n A_j \right] (X_i^t X_i)^{-1} x_k \right\| \\ &\leq \sum_{i=k}^n \tau^{n-i+1} \left\| (X_i^t X_i)^{-1} x_k \right\| \\ &\leq \sum_{i=k}^n \tau^{n-i+1} \lambda^{-1} i^{-1} \|x_k\| \\ &\leq \lambda^{-1} K \sum_{i=k}^n \tau^{n-i+1} i^{-1}. \end{aligned}$$

La tercera desigualdad es consecuencia de que el mínimo autovalor de  $X_i^t X_i$  no puede ser inferior al mínimo autovalor de  $(X_i^t X_i)^{ng}$ , el cual por hipótesis esta acotado inferiormente por  $\lambda i$ . A partir del lema 4.5 se puede concluir que



4. Algoritmos de una iteración basados en correcciones en media

$$\|C_n^k x_k\|^2 = O(n^{-2}).$$

Si  $f_{jn}^k$  denota la fila  $j$ -ésima de la matriz  $C_n^k$ ,  $j = 1, \dots, m$ , la componente  $j$ -ésima de (27) será

$$\sqrt{n} \sum_{k=1}^n f_{jn}^k x_k \varepsilon_k(a),$$

que tiene media cero y verifica, además, que

$$\begin{aligned} \text{Var} \left( \sqrt{n} f_{jn}^k x_k \varepsilon_k(a) \right) &= n \left( f_{jn}^k x_k \right)^2 E \left( \varepsilon_k(a)^2 \right) \\ &\leq n \|C_n^k x_k\|^2 E \left( \varepsilon_k(a)^2 \right). \end{aligned}$$

Teniendo en cuenta que la varianza de los errores de observación se acota por uno,

$\text{Var} \left( \varepsilon_k(a) \right) = E \left( \varepsilon_k(a)^2 \right) \leq 1$ , se puede escribir

$$\max_{k \leq n} \text{Var} \left( \sqrt{n} f_{jn}^k x_k \varepsilon_k(a) \right) \leq \lambda^{-2} K^2 n \left( \sum_{i=k}^n \tau^{n-i+1} i^{-1} \right)^2 = O(n^{-1})$$

y

$$\sum_{k=1}^n \text{Var} \left( \sqrt{n} f_{jn}^k x_k \varepsilon_k(a) \right) \leq \lambda^{-2} K^2 n^2 \left( \sum_{i=k}^n \tau^{n-i+1} i^{-1} \right)^2 = O(1).$$

Ambas expresiones implican la convergencia buscada de cada componente  $\sqrt{n} \sum_{k=1}^n f_{jn}^k x_k \varepsilon_k(a)$  a una distribución normal, y en consecuencia la convergencia de la variable  $m$ -dimensional  $\sqrt{n} (a_n - a)$  a una normal de media cero.  $\square$

**Lema 4.5** Si  $0 < \tau < 1$  y  $k \geq 1$ , entonces existe un valor  $C < \infty$  tal que para todo  $n \in \mathcal{N}$

$$\tau < n \sum_{i=k}^n \tau^{n-i+1} i^{-1} < C.$$

DEMOSTRACION: Tómesese  $\varepsilon > 0$ . Se cumple  $1 \leq \frac{n}{i} < 1 + \varepsilon(1 - \tau)$  si, y sólo si,  $i > \frac{n}{1 + \varepsilon(1 - \tau)}$ . Por otra parte se puede tomar  $n_0$  tal que para cada  $n \geq n_0$  se verifica

$$n^2 \tau^{\frac{\varepsilon(1-\tau)}{1+\varepsilon(1-\tau)} n} \leq \frac{(1 + \varepsilon(1 - \tau)) \varepsilon}{\tau}.$$

En esta situación se puede escribir

$$\begin{aligned}
 \sum_{i=k}^n \binom{n}{i} \tau^{n-i+1} &\leq \sum_{\substack{i=k \\ i > \frac{n}{1+\varepsilon(1-\tau)}}}^n (1 + \varepsilon(1 - \tau)) \tau^{n-i+1} + \sum_{\substack{i=k \\ i \leq \frac{n}{1+\varepsilon(1-\tau)}}}^n \binom{n}{i} \tau^{n-i+1} \\
 &\leq (1 + \varepsilon(1 - \tau)) \sum_{i=1}^{\infty} \tau^i + \sum_{\substack{i=k \\ i \leq \frac{n}{1+\varepsilon(1-\tau)}}}^n n \tau^{n - \frac{n}{1+\varepsilon(1-\tau)} + 1} \\
 &\leq \frac{1}{1 - \tau} + \varepsilon + \frac{n^2}{1 + \varepsilon(1 - \tau)} \tau^{n - \frac{n}{1+\varepsilon(1-\tau)} + 1} \\
 &\leq \frac{1}{1 - \tau} + 2\varepsilon \leq C,
 \end{aligned}$$

concluyéndose la demostración.  $\square$

#### 4.5 Un segundo algoritmo de una iteración

La expresión (24) permite contemplar el método iterativo MD de una iteración como un tipo de aproximación estocástica, en la cual la expresión de la estimación  $a_{n+1}$  en la etapa  $n + 1$  corresponde a la estimación en la etapa  $n$  sumado un factor aleatorio de búsqueda que depende de  $\varepsilon(a_n)$ . Este factor  $\varepsilon(a_n)$  depende, a su vez, de todas las observaciones hechas hasta la etapa  $n$ . Las aproximaciones estocásticas corrigen la última estimación mediante una variable cuya esperanza depende del valor a estimar multiplicada adecuadamente por un tamaño de paso. De esta forma, las aproximaciones estocásticas adoptan la forma de ecuaciones en diferencias estocásticas. En nuestro caso, el problema estriba en que la estimación es multidimensional, con lo cual los tamaños de paso necesario podrían variar en cada componente. De todas formas, el problema podrá reducirse a uno de tipo unidimensional al tomar normas. En todo caso, las técnicas de demostración clásicas de WASAN (1969) y las más novedosas de KUSHNER Y YIN (1997), podrían ser de utilidad en nuestro caso.

En concreto, las aproximaciones estocásticas permiten concebir una variante

del algoritmo antes visto, en la cual se sustituya la inversión de la matriz  $(X^t X)$ , distinta para cada paso  $n$ , por unos valores positivos  $\alpha_n$ , que llamaremos tamaño de paso. El método iterativo propuesto se puede escribir formalmente como sigue:

**INICIALIZACIÓN:** Tómese un vector inicial  $a_1$  arbitrario (el cual podría comenzar con  $k$  datos en lugar de uno solo).

**ITERACIÓN:** Si  $a_n$  es la estimación en la etapa  $n$  del verdadero vector  $a$ , la estimación para la etapa  $n + 1$  se actualiza mediante la expresión

$$a_{n+1} = a_n + \alpha_n X^t \varepsilon(a_n),$$

donde  $\varepsilon(a_n) = (\varepsilon_1(a_n), \dots, \varepsilon_n(a_n))^t$  y cada componente coincide con  $\varepsilon_i(a_n) = y_i(a_n) - a_n^t x_i$ . El valor de  $y_i(a_n)$  ya fue establecido en la sección 4.3, y se basa en las esperanzas de los errores condicionadas tanto a los intervalos de agrupación como a la asunción de que el verdadero valor del parámetro es  $a_n$ .

Para poder asegurar la convergencia de las estimaciones  $a_n$  al verdadero valor  $a$  del modelo lineal se deberán elegir adecuadamente los tamaños de paso  $\alpha_n$ . En todo caso, si la elección de  $\alpha_n$  es sencilla el ahorro de tiempo al utilizar este método frente al método que utiliza la inversión de  $X^t X$  es sustancial (ver al respecto las simulaciones presentadas al final de este capítulo). Obsérvese que este algoritmo de estimación (aun más simplificado que el anterior) continúa siendo de una sola iteración, con todas las ventajas ya expuestas al principio del capítulo. El siguiente lema hace referencia a la selección de los tamaños de paso.

**Lema 4.6** *Tómese una sucesión de pasos positivos  $\{\alpha_n\}$ , de tal forma que  $\alpha_n n \rightarrow 0$ ,  $\sum \alpha_n n = \infty$ ,  $\sum \alpha_n n^{\frac{1}{2}} < \infty$  y  $\alpha_n \max v_n < 1$ , siendo  $v_n$  el máximo autovalor de la matriz  $X^t X$ . Supóngase además que  $\inf \lambda_n = \lambda > 0$ , siendo  $\lambda_n$  el mínimo autovalor de la matriz  $n^{-1} (X^t X)^{ng}$ . Sea la matriz de tamaño  $n$ ,  $M^* = \text{diag}(m_i^*)$  definida de forma que  $m_i^* = 1$  si  $i \in I_n^{ng}$ , y  $0 \leq m_i^* \leq 1$  si  $i \in I_n^g$ . Con todo lo anterior, los autovalores de la matriz  $I - \alpha_n X^t M^* X$  son positivos y menores que  $1 - \alpha_n n \lambda < 1$ , a partir de un  $n$  en adelante.*

**DEMOSTRACION:** En efecto, el mínimo autovalor de la matriz simétrica y

definida positiva del enunciado se corresponde con

$$\begin{aligned}
 \min_{\|u\|=1} u^t (I - \alpha_n X^t M^* X) u &= 1 - \alpha_n \max_{\|u\|=1} u^t X^t M^* X u \\
 &= 1 - \alpha_n \max_{\|u\|=1} \sum_{i=1}^n m_i^* u^t x_i x_i^t u \\
 &\geq 1 - \alpha_n \max_{\|u\|=1} \sum_{i=1}^n u^t x_i x_i^t u \\
 &= 1 - \alpha_n v_n \geq 1 - \alpha_n \max v_n > 0.
 \end{aligned}$$

Por otra parte, el máximo de los autovalores coincide con

$$\begin{aligned}
 \max_{\|u\|=1} u^t (I - \alpha_n X^t M^* X) u &= 1 - \alpha_n \min_{\|u\|=1} u^t X^t M^* X u \\
 &= 1 - \alpha_n \min_{\|u\|=1} \sum_{i=1}^n m_i^* u^t x_i x_i^t u \\
 &\leq 1 - \alpha_n \min_{\|u\|=1} \sum_{i \in I^{ng}} u^t x_i x_i^t u \\
 &= 1 - \alpha_n n \lambda_n \leq 1 - \alpha_n n \lambda.
 \end{aligned}$$

Además,  $1 - \alpha_n n \lambda < 1$ , si  $n$  es suficientemente grande, lo cual concluye la demostración.  $\square$

OBSERVACION: Teniendo en cuenta que  $\alpha_n n \rightarrow 0$ , una condición suficiente para que se cumpla la hipótesis  $\alpha_n \max v_n < 1$  es que  $\sup \delta_n = M < \infty$ , siendo  $\delta_n$  el máximo autovalor de la matriz  $n^{-1} X^t X$ .

Si suponemos que  $\|x_i\| \leq K < +\infty$  para cada  $i \in \mathcal{N}$ , entonces  $\delta_n = n^{-1} \sum_{i=1}^n \|x_i\|^2 \leq K^2$  para cada  $n \in \mathcal{N}$ , con lo cual se cumple la susodicha hipótesis del lema 4.6.

Se enunciará ahora un teorema en el que se demuestra la convergencia del nuevo método propuesto de una iteración al verdadero valor buscado del modelo lineal. La convergencia probada es en  $L_1$ , tal como ocurría en el método MD de una iteración contemplado en la sección 4.3 anterior.

**Teorema 4.7** *Asumamos que  $(X^t X)^{ng}$  es una matriz definida positiva para cada*

4. Algoritmos de una iteración basados en correcciones en media

$n \in \mathcal{N}$ . Supongamos que las siguientes condiciones se cumplen

$$\inf_n \lambda_n = \lambda > 0,$$

$$\|x_i\| \leq K < +\infty, \quad \text{para cada } i \in \mathcal{N},$$

donde  $\lambda_n$  denota el mínimo autovalor de la matriz  $n^{-1}(X^t X)^{ng}$ . Si se selecciona

en cada etapa un tamaño de paso  $\alpha_n$  de forma que  $\alpha_n n \rightarrow 0$ ,  $\sum \alpha_n n = \infty$  y  $\sum \alpha_n n^{\frac{1}{2}} < \infty$ , entonces el método de una iteración cumple que  $a_n \xrightarrow{L_1} a$  y, en consecuencia,  $a_n$  es un estimador consistente de  $a$ .

DEMOSTRACION: Recordemos que los errores  $\nu$  son tales que para cada  $j, r \in \mathcal{N}$  se cumple la CONDICION C1. En consecuencia, de la misma forma que en el resultado principal de la sección anterior, se puede escribir

$$a_{n+1} = a_n - \alpha_n X^t M^* X (a_n - a) + \alpha_n X^t \varepsilon(a),$$

donde  $M^* = \text{diag}(m_i^*)$  verifica que  $m_i^* = 1$  si  $i \in I_n^{ng}$ , y  $0 \leq m_i^* \leq 1$  si  $i \in I_n^g$ .

Esta última expresión es más conveniente ponerla de la siguiente forma

$$a_{n+1} - a = [I - \alpha_n X^t M^* X] (a_n - a) + \alpha_n X^t \varepsilon(a).$$

Tomando normas euclideas se tiene

$$\|a_{n+1} - a\| \leq \mu_n \|a_n - a\| + \alpha_n \|X^t \varepsilon(a)\|,$$

siendo  $\mu_n$  el radio espectral de la matriz  $I - \alpha_n X^t M^* X$ . Según se probó en el lema 4.6, esta matriz es definida positiva y además  $\mu_n \leq 1 - \alpha_n n \lambda < 1$ , para  $n$  suficientemente grande.

Llamemos  $c_n = \alpha_n \|X^t \varepsilon(a)\|$ . La independencia, la media nula y la varianza acotada por uno de los errores de observación  $\varepsilon_i(a)$ , probada en la demostración del teorema 4.1 y válida en este contexto, permite afirmar que

$$\begin{aligned} E(c_n^2) &= \alpha_n^2 E(\|X^t \varepsilon(a)\|^2) = \alpha_n^2 E\left(\sum_{i,j=1}^n \varepsilon_i(a) \varepsilon_j(a) x_i^t x_j\right) \\ &= \alpha_n^2 \sum_{i=1}^n \|x_i\|^2 E(\varepsilon_i(a)^2) \end{aligned}$$

$$\leq \alpha_n^2 n K^2 \xrightarrow{n \rightarrow \infty} 0,$$

puesto que  $\alpha_n n \rightarrow 0$ .

En definitiva

$$E \|a_{n+1} - a\| \leq \mu_n E \|a_n - a\| + E(c_n).$$

Puesto que se supone que  $\sum_n \alpha_n n = \infty$ , se tiene que  $\prod_n (1 - \alpha_n n \lambda) = 0$ , con lo cual también  $\prod_n \mu_n = 0$ . Por consiguiente,

$$E \|a_{n+1} - a\| \leq \prod_{i=1}^n \mu_i E \|a_1 - a\| + \sum_{i=1}^n \tau_n^i E(c_i),$$

siendo  $\tau_n^i = \prod_{j=i+1}^n \mu_j$ . Así pues, teniendo en cuenta que i)  $\tau_n^i \xrightarrow{n \rightarrow \infty} 0$ , para cada  $i$  fijo, ii)  $\tau_n^i \leq 1$  y iii)

$$\sum_{n=1}^{\infty} E(c_n) \leq \sum_{n=1}^{\infty} E(c_n^2)^{\frac{1}{2}} \leq K \sum_{n=1}^{\infty} \alpha_n n^{\frac{1}{2}} < \infty,$$

se puede concluir, a partir del lema 4.8 posterior, que

$$E \|a_n - a\| \xrightarrow{n \rightarrow \infty} 0.$$

Queda así probada la convergencia en media de orden 1 y, en consecuencia, la consistencia del estimador propuesto  $a_n$ .  $\square$

**Lema 4.8** Sean  $d_i$  y  $\tau_n^i$  valores positivos tales que  $\sum_{i=1}^{\infty} d_i < \infty$ ,  $\tau_n^i \xrightarrow{n \rightarrow \infty} 0$ , para cada  $i \in \mathcal{N}$  y  $\tau_n^i \leq D < \infty$ , para cada  $i, n \in \mathcal{N}$ . Se verifica que

$$\sum_{i=1}^n \tau_n^i d_i \xrightarrow{n \rightarrow \infty} 0.$$

DEMOSTRACION: Supongamos que  $\sum_{i=1}^{\infty} d_i = M < \infty$ . Fijado  $\varepsilon > 0$  se puede encontrar  $n_0$  tal que para cada  $n \geq n_0$  se cumple  $\sum_{i=n_0}^n d_i < \varepsilon/2D$ . Para  $n \geq n_0$  será

$$\begin{aligned} \sum_{i=1}^n \tau_n^i d_i &= \sum_{i=1}^{n_0-1} \tau_n^i d_i + \sum_{i=n_0}^n \tau_n^i d_i \\ &\leq \sum_{i=1}^{n_0-1} \tau_n^i d_i + D \sum_{i=n_0}^n d_i \leq \sum_{i=1}^{n_0-1} \tau_n^i d_i + \frac{\varepsilon}{2}. \end{aligned}$$

Por otra parte, para cada  $i = 1, \dots, n_0 - 1$ , existe  $n_i$  tal que para  $n \geq n_i$  es  $\tau_n^i \leq \frac{\varepsilon}{2M}$ . Se sigue que

$$\sum_{i=1}^{n_0-1} \tau_n^i d_i \leq \frac{\varepsilon}{2M} \sum_{i=1}^{n_0-1} d_i \leq \frac{\varepsilon}{2}.$$

En conclusión,  $\sum_{i=1}^n \tau_n^i d_i \leq \varepsilon$ , para cada  $n \geq \max \{n_0, n_1, \dots, n_{n_0-1}\}$ .  $\square$

## 4.6 Un tercer algoritmo de una iteración

El método propuesto de la sección anterior se inspira en las técnicas de aproximaciones estocásticas tipo KUSHNER Y YIN (1997), existiendo como en éstos un tamaño de paso  $\alpha_n$ , que debe ser fijado adecuadamente. Pese a ello, no coincide exactamente con una aproximación estocástica del tipo citado. La razón está en el factor de búsqueda, que corresponde al vector  $X^t \varepsilon(a_n)$ . Obsérvese que  $\varepsilon(a_n)$  es un vector de tamaño  $n$ , que incorpora información de todos los datos observados hasta la etapa  $n$ . Es decir, que la información se va acumulando y utilizando repetidas veces en cada iteración. La idea de las aproximaciones estocásticas no es la de acumular la información, sino solamente utilizar la última estimación obtenida, en la misma línea que actúan las redes neuronales. Además, para calcular ese factor de búsqueda deben hacerse en cada etapa un total de  $n \times m$  productos y  $n \times m$  sumas, correspondientes a  $m$  productos escalares de vectores de tamaño  $n$  involucrados. Todo ello puede suponer un enorme tiempo de cálculo a medida que  $n$  crece.

En el método que se propone a continuación se utilizará en cada iteración únicamente la información que proporciona el último valor muestral, para actualizar la estimación actual del parámetro. De este modo, se intenta conseguir estrictamente una aproximación estocástica en el sentido de KUSHNER Y YIN (1997). Por otra parte, este nuevo procedimiento no necesita realizar ningún cálculo para obtener la dirección de búsqueda en la etapa  $n$ , puesto que ésta vendrá

dada por el propio vector de variables independientes de la observación  $n$ -ésima. Es decir, cualquiera que sea  $n$ , el número de operaciones a realizar en dicha etapa es el mismo, al contrario de lo que ocurría con el método anterior. La disminución que este hecho supone en complejidad computacional resulta evidente.

Estrictamente, se propone el siguiente tipo de estimación recursiva:

INICIALIZACIÓN: Tómese un vector inicial  $a_1$  arbitrario.

ITERACIÓN: Si  $a_n$  es la estimación en la etapa  $n$  del verdadero vector  $a$ , la estimación en la etapa  $n + 1$  se llevará a cabo mediante la ecuación recursiva

$$a_{n+1} = a_n + \alpha_n x_n \varepsilon_n(a_n),$$

donde  $\varepsilon_n(a_n) = y_n(a_n) - a_n^t x_n$ ,  $\alpha_n > 0$  y  $a_n, x_n \in R^m$ .

Obsérvese que se puede escribir  $X^t \varepsilon(a_n) = \sum_{i=1}^n x_i \varepsilon_i(a_n)$ , con lo cual este nuevo método propuesto simplifica el de la sección 4.5 al quedarnos exclusivamente con el último sumando. El método de aquella sección puede en ocasiones ser más eficiente que éste, porque el factor de búsqueda, al multiplicarlo por el tamaño del paso, se convierte en una ponderación de los errores utilizados en el nuevo algoritmo aquí propuesto.

En definitiva, si comparamos esta propuesta con los métodos vistos hasta el momento, encontramos que ésta supone una indudable reducción del número de operaciones por iteración. Es de esperar que la convergencia de este nuevo método sea más lenta, necesitándose más iteraciones para obtener una buena aproximación del límite buscado. En todo caso, como cada iteración conlleva menor número de cálculos, es posible que, en tiempo, sea este último un método más eficiente que los anteriores.

El interés se centrará en buscar una sucesión de tamaños de paso  $\{\alpha_n\} \subset R^+$  adecuada para poder garantizar la convergencia  $a_n \xrightarrow{P} a$ . Adicionalmente, manteniendo la notación de la sección 4.3, se puede cambiar algo la CONDICION C1, para ampliar el conjunto de distribuciones  $F_\nu$  sobre las cuales será aplicable



el resultado posterior de convergencia del algoritmo. En concreto, supongamos que se verifica la CONDICION C2, que viene dada por las dos hipótesis siguientes

$$\delta_{jr}(\beta) \text{ es no decreciente} \quad (28)$$

y que existe una constante  $M < \infty$  tal que

$$M\beta - \delta_{jr}(\beta) \text{ es no decreciente.} \quad (29)$$

**Lema 4.9** *Las condiciones (28) y (29) aseguran que, cualquiera que sean  $j = 1, \dots, s$ ,  $r = 0, 1, \dots, r_j - 1$  y  $a, b \in R$ , existe  $m^* \in [0, M]$  tal que  $\delta_{jr}(b) - \delta_{jr}(a) = m^*(b - a)$ .*

DEMOSTRACION: Si  $a > b$ ,

$$Ma - \delta_{jr}(a) \geq Mb - \delta_{jr}(b).$$

Así,

$$0 \leq \delta_{jr}(a) - \delta_{jr}(b) \leq M(a - b),$$

por lo cual existirá un valor  $m^* \in [0, M]$ , dependiente de  $a$  y  $b$ , tal que

$$\delta_{jr}(a) - \delta_{jr}(b) = m^*(a - b).$$

A la misma conclusión se llega si  $a < b$ .  $\square$

Puesto que las correcciones se siguen haciendo en esperanza condicionada, de la misma forma que en desarrollos anteriores se continúa verificando que  $E(\varepsilon_i^2(a)) \leq 1$ , para todo  $x_i \in R^m$ . No será preciso utilizar en este desarrollo el hecho de que las variables  $\varepsilon_i(a)$  son independientes y tienen media cero.

OBSERVACIONES: Se harán una serie de observaciones previas antes de imponer las condiciones precisas sobre los pasos  $\alpha_n$ .

i) La matriz  $x_n x_n^t$  es de tamaño  $m \times m$  y rango 1, cualquiera que sea el vector  $x_n \neq 0$ . Así pues, si  $m > 1$  el mínimo autovalor de  $x_n x_n^t$  es cero, y su máximo autovalor es  $\|x_n\|^2$ . Obsérvese que, en efecto,  $x_n x_n^t$  es semidefinida positiva y

$$\max_{\|u\|=1} u^t x_n x_n^t u = \max_{\|u\|=1} \|u^t x_n\|^2 = \max_{\|u\|=1} \|u\|^2 \|x_n\|^2 |\cos(\widehat{u, x_n})|^2 = \|x_n\|^2.$$

ii) Los autovalores de  $I - A$  son  $\{1 - \lambda_i\}$ , siendo  $\{\lambda_i\}$  los de  $A$ , puesto que para cada  $\lambda$

$$\det(I - A - \lambda I) = (-1)^m \det(A - (1 - \lambda)I).$$

Así pues, el máximo autovalor de  $I - \alpha x_n x_n^t$ , con  $\alpha > 0$ , será 1 y el mínimo será  $1 - \alpha \|x_n\|^2$ . Por tanto, se puede siempre encontrar  $\alpha > 0$  cumpliendo que  $\alpha \|x_n\|^2 < 1$ . Así,  $1 - \alpha \|x_n\|^2 > 0$ , con lo cual  $I - \alpha x_n x_n^t$  sería definida positiva y su máximo autovalor 1.

Procedamos como en los casos anteriores del siguiente modo. Se sabe que

$$\varepsilon_n(a_n) = y_n(a_n) - x_n^t a_n,$$

o bien,

$$\varepsilon_n(a_n) = -m_n^* x_n^t (a_n - a) + \varepsilon_n(a),$$

donde ahora  $m_n^* = 1$  cuando el dato  $n$ -ésimo es no agrupado, y  $0 \leq m_n^* \leq M$  cuando  $n \in I^g$ . Recuérdese que  $a \in \mathcal{R}^m$  es el verdadero valor del parámetro desconocido del modelo lineal. Se tiene, pues, que

$$x_n \varepsilon_n(a_n) = -m_n^* x_n x_n^t (a_n - a) + x_n \varepsilon_n(a),$$

de donde

$$a_{n+1} = a_n - \alpha_n m_n^* x_n x_n^t (a_n - a) + \alpha_n x_n \varepsilon_n(a).$$

Después de sumar  $a$  en cada uno de los dos miembros, se obtiene

$$a_{n+1} - a = [I - \alpha_n m_n^* x_n x_n^t] (a_n - a) + \alpha_n x_n \varepsilon_n(a). \quad (30)$$

Esta última es la expresión fundamental sobre la que se pivotará en nuestro desarrollo. Como  $x_n$  son valores conocidos, en todas las etapas se puede tomar  $\alpha_n$  positivo tal que  $\alpha_n \|x_n\|^2 M < 1$ , con lo cual la matriz  $I - \alpha_n m_n^* x_n x_n^t$  es definida positiva. Con esta condición sobre la elección de los pasos  $\alpha_n$ , se puede continuar poniendo, como en casos anteriores,

$$\|a_{n+1} - a\| \leq \mu_n \|a_n - a\| + \alpha_n \|x_n \varepsilon_n(a)\|,$$

donde  $\mu_n$  es el máximo autovalor  $I - \alpha_n m_n^* x_n x_n^t$ . Como sabemos es  $\mu_n = 1$ , llegándose, pues, a

$$\|a_{n+1} - a\| \leq \|a_n - a\| + \alpha_n \|x_n \varepsilon_n(a)\|,$$

no pudiéndose alcanzar por este camino la convergencia de  $\|a_n - a\|$  a cero.

La forma de proceder debe ser diferente a la que se acaba de explicar, la cual coincide con la utilizada en la sección 4.5. Se enunciará a continuación el resultado general que, como en los casos precedentes, asegura convergencia en media de orden uno. Las condiciones que se imponen hacen referencia a la adecuada selección de los tamaños de paso  $\alpha_n$ , establecidos de acuerdo a los valores de  $x_n$ . Sobre estos últimos, en principio, no se supone ninguna condición adicional de acotación. Más adelante se verán condiciones que aseguran las asumidas en el teorema que sigue. Por otra parte, en el capítulo 7 se analizará este mismo algoritmo desde otra perspectiva, asegurándose la convergencia en  $L_2$  bajo otras condiciones distintas a las actuales.

**Teorema 4.10** *Tómense los tamaños de paso  $\alpha_n > 0$  tales que, para cada  $n \in \mathcal{N}$ ,  $\alpha_n \|x_n\|^2 M < 1$ , y cumpliéndose, además,*

$$\sum_{n=1}^{\infty} \alpha_n \|x_n\| < \infty.$$

*Llamemos  $\eta_n^i$  al máximo autovalor de la matriz definida positiva*

$$\prod_{j=i}^n [I - \alpha_j m_j^* x_j x_j^t].$$

*Si para cada  $i \in \mathcal{N}$  existe una sucesión determinista  $\{\tau_n^i\}$ , convergente a cero cuando  $n \rightarrow \infty$ , tal que  $\eta_n^i \leq \tau_n^i$  c.s., entonces se cumple que  $E \|a_n - a\| \rightarrow 0$ . Es decir,  $a_n \xrightarrow{L_1} a$  y, en consecuencia,  $a_n$  es un estimador consistente de  $a$ .*

DEMOSTRACION: La existencia de las sucesiones  $\{\tau_n^i\}$  está ligada directamente a la elección de la sucesión de tamaños de paso  $\{\alpha_n\}$ , por lo que esta condición debe ser coherente con las anteriormente impuestas sobre los tamaños de paso.

Obsérvese que  $\eta_n^i \leq 1$ , con lo cual, si existe una sucesión  $\tau_n^i$  en las condiciones del enunciado, se puede suponer, sin pérdida de generalidad, que  $\tau_n^i \leq 1$ . En otro caso basta tomar  $\bar{\tau}_n^i = \max \{1, \tau_n^i\}$ , que verifica las condiciones del teorema y se encuentra uniformemente acotada por uno.

La demostración del resultado se basa en la línea argumental previa, y en concreto, en la iteración  $n$  veces de la ecuación (30), que produce

$$a_{n+1} - a = \prod_{i=1}^n [I - \alpha_i m_i^* x_i x_i^t] (a_1 - a) + \sum_{i=1}^n \prod_{j=i}^n [I - \alpha_j m_j^* x_j x_j^t] \alpha_i x_i \varepsilon_i(a).$$

Así, tomando normas euclideas en este momento resulta

$$\|a_{n+1} - a\| \leq \eta_n^1 \|a_1 - a\| + \sum_{i=1}^n \eta_n^i \alpha_i \|x_i\| |\varepsilon_i(a)|,$$

donde  $\eta_n^i$  es el máximo autovalor de la matriz definida positiva

$$\prod_{j=i}^n [I - \alpha_j m_j^* x_j x_j^t].$$

Por las hipótesis del enunciado, se puede asegurar que

$$\|a_{n+1} - a\| \leq \tau_n^1 \|a_1 - a\| + \sum_{i=1}^n \tau_n^i \alpha_i \|x_i\| |\varepsilon_i(a)| \quad c.s.$$

Llamando  $c_i = \alpha_i \|x_i\| |\varepsilon_i(a)|$  y tomando esperanzas en la última expresión se llega a

$$E \|a_{n+1} - a\| \leq \tau_n^1 E \|a_1 - a\| + \sum_{i=1}^n \tau_n^i E(c_i).$$

Como  $E(\varepsilon_i^2(a)) \leq 1$ , también es  $E(|\varepsilon_i(a)|) \leq 1$ , para todo  $x_i \in R^m$ . Puesto que ahora es  $E(c_i) \leq \alpha_i \|x_i\|$  resulta, por las hipótesis del teorema, que es  $\sum_{i=1}^{\infty} E(c_i) < \infty$ . Sólo queda aplicar el lema 4.8 de la sección anterior, teniendo en cuenta que  $\tau_n^i \leq 1$  y  $\tau_n^i \rightarrow 0$ , para llegar a que  $\sum_{i=1}^n \tau_n^i E(c_i) \xrightarrow{n \rightarrow \infty} 0$ . Esto asegura que  $a_n \xrightarrow{L_1} a$ , con lo cual se concluye la demostración.  $\square$

OBSERVACIONES: Denotemos por  $\max(A)$  y  $\min(A)$  al máximo y mínimo autovalor de la matriz cuadrada  $A$ , respectivamente. Se verifican los siguientes lemas.

**Lema 4.11** Si  $A$  es una matriz  $m \times m$  definida positiva, siempre es posible encontrar un vector  $y \in \mathcal{R}^m$  con  $\|y\| = 1$ , de tal forma que

$$\|Ay\| = \max(A).$$

De igual forma, existe  $x \in \mathcal{R}^m$ ,  $\|x\| = 1$ , tal que

$$\|Ax\| = \min(A).$$

DEMOSTRACION: Es evidente, sin más que tomar sendos autovalores asociados, respectivamente, a  $\max(A)$  y  $\min(A)$  y reescalarlos, después, a norma uno.  $\square$

**Lema 4.12** Si  $A$  y  $B$  son matrices definidas positivas resulta que

$$\max(AB) \leq \max(A) \max(B)$$

y

$$\min(AB) \geq \min(A) \min(B).$$

DEMOSTRACION: Tómese el vector  $y$  que proporciona el lema anterior para la matriz  $AB$ , con lo cual

$$\max(AB) = \|AB y\| \leq \max(A) \|B y\| \leq \max(A) \max(B).$$

Un razonamiento similar conduce a la segunda desigualdad del lema.  $\square$

**Lema 4.13** Si la matriz  $A$  es invertible, los autovalores de  $AB$  coinciden con los de  $BA$ .

DEMOSTRACION: Para cada valor complejo  $\lambda$ , directamente, se verifica que

$$\det(AB - \lambda I) = \det(A^{-1}(AB - \lambda I)A) = \det(BA - \lambda I). \square$$

Llamemos  $A_j = [I - \alpha_j m_j^* x_j x_j^t]$ . Para cada  $n > i > 1$ , si la matriz  $\prod_{j=1}^{i-1} A_j$  es definida positiva, se puede escribir

$$\prod_{j=i}^n A_j = \left[ \prod_{j=1}^{i-1} A_j \right]^{-1} \prod_{j=1}^n A_j.$$

Aplicando el lema 4.12, cualquiera que sea el vector  $y \in \mathcal{R}^m$  será

$$\left\| \prod_{j=i}^n A_j y \right\| \leq \eta_n^i \|y\| \leq \zeta_{i-1}^1 \eta_n^1 \|y\|,$$

siendo  $\zeta_{i-1}^1 > 0$  el máximo autovalor de la matriz  $\left[\prod_{j=1}^{i-1} A_j\right]^{-1}$ , que coincide con la inversa del mínimo autovalor de  $\prod_{j=1}^{i-1} A_j$ . Bajo las hipótesis del teorema 4.10, se tiene que  $\zeta_{i-1}^1 \eta_n^1 \leq \zeta_{i-1}^1 \tau_n^1$  c.s., donde el último término converge casi seguro a cero cuando  $n \rightarrow \infty$ . En todo caso, el desarrollo anterior no permite acotar por un valor determinista a  $\zeta_{i-1}^1 \eta_n^1$ , puesto que  $\zeta_{i-1}^1$  sigue siendo aleatorio. Por esta razón el siguiente teorema tiene interés.

**Teorema 4.14** *Tómense  $\alpha_n > 0$  tales que  $\alpha_n \|x_n\|^2 M < 1$ , para cada  $n \in \mathcal{N}$ , cumpliéndose, además, que*

$$\sum_{n=1}^{\infty} \alpha_n \|x_n\| < \infty.$$

Llamemos  $\eta_n^1$  al máximo autovalor de la matriz definida positiva

$$\prod_{j=1}^n [I - \alpha_j m_j^* x_j x_j^t].$$

Si existe una sucesión de números reales  $\{\tau_n^1\}$ , convergente a cero cuando  $n \rightarrow \infty$ , tal que  $\eta_n^1 \leq \tau_n^1$  c.s., entonces se cumple que  $a_n \xrightarrow{L_1} a$ .

DEMOSTRACION: Obsérvese que el resultado es cierto si existen números reales  $\{C_i\}$  tal que  $\zeta_{i-1}^1 \leq C_{i-1} < \infty$  c.s., puesto que en esa situación  $\eta_n^1 \leq \zeta_{i-1}^1 \eta_n^1 \leq C_{i-1} \tau_n^1 \xrightarrow{n \rightarrow \infty} 0$ .

El mínimo autovalor de  $A_j$  coincide con  $1 - \alpha_j \|x_j\| m_j^*$ , que está acotado inferiormente por  $1 - \alpha_j \|x_j\| M$ . En cualquier situación, se puede tomar el valor positivo  $C_{i-1} = \left[\prod_{j=1}^{i-1} (1 - \alpha_j \|x_j\| M)\right]^{-1} < \infty$ , como cota del valor  $\zeta_{i-1}^1$  descrito.  $\square$

Los lemas que se enunciaron anteriormente permiten aún formular otra versión alternativa del teorema 4.10.

**Teorema 4.15** *Elíjanse tamaños de paso  $\alpha_n > 0$  tales que  $\alpha_n \|x_n\|^2 M < 1$ , para cada  $n \in \mathcal{N}$ , y de forma que se cumpla, además, la condición*

$$\sum_{n=1}^{\infty} \alpha_n \|x_n\| < \infty.$$

Llamemos  $\rho_n^1$  al máximo autovalor de la matriz definida positiva

$$\prod_{\substack{j=1 \\ j \in I^{n_g}}}^n [I - \alpha_j x_j x_j^t].$$

Si existe una sucesión de números reales  $\{\tau_n^1\}$ , convergente a cero cuando  $n \rightarrow \infty$ , tal que  $\rho_n^1 \leq \tau_n^1$  c.s., entonces se cumple que  $\alpha_n \xrightarrow{L_1} a$ .

DEMOSTRACION: Este resultado permite trasladar las condiciones impuestas sobre las matrices aleatorias de la forma (31), en donde intervienen los valores  $m_i^*$  desconocidos, a las matrices del tipo (32), que sólo son aleatorias en el conjunto de índices cuyos datos son no agrupados. En todo caso, las condiciones siguen haciendo referencia a la elección del tamaño de paso  $\alpha_n$  de la aproximación estocástica.

Para la demostración, basta observar que el máximo autovalor de la matriz definida positiva

$$\prod_{j=1}^n [I - \alpha_j m_j^* x_j x_j^t] \quad (31)$$

es menor o igual que el máximo autovalor de la matriz definida positiva

$$\prod_{\substack{j=1 \\ j \in I^{n_g}}}^n [I - \alpha_j m_j^* x_j x_j^t] = \prod_{\substack{j=1 \\ j \in I^{n_g}}}^n [I - \alpha_j x_j x_j^t]. \quad \square \quad (32)$$

#### 4.6.1 Observación sobre las hipótesis

En esta subsección se comprobará en algún caso práctico las condiciones que han sido impuestas sobre la distribución de los errores, las cuales permiten aplicar los resultados de convergencia expuestos. Por poner un ejemplo, se verá que la distribución normal verifica la CONDICION C1 pero no las verifica la distribución t-Student.

En concreto, con la notación utilizada a lo largo del capítulo, se han fijado los

valores  $c_{j,r} = 0$  y  $c_{j,r+1} = 4$ . La función que se debe estudiar corresponde a

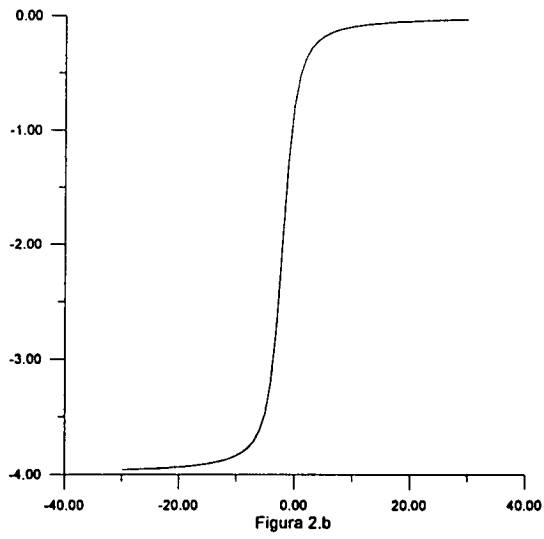
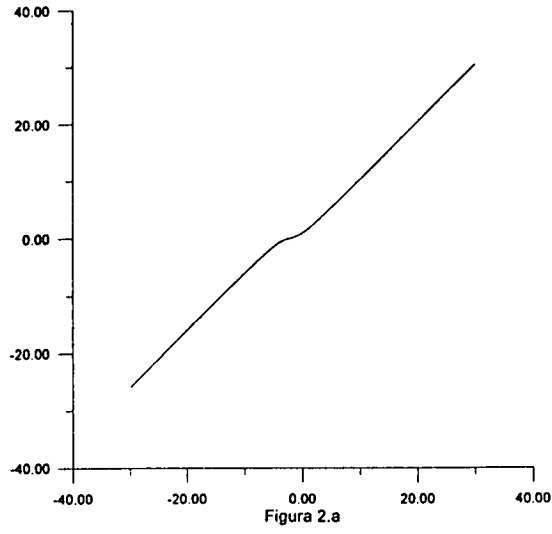
$$\delta_{jr}(\beta) = E(\nu | \nu \in (\beta, \beta + 4]).$$

Tómese inicialmente  $\nu$  como una variable aleatoria distribuida según una normal  $N(0, 1)$ . La figura 2.a, al final de la subsección, muestra el comportamiento de  $\delta_{jr}(\beta)$ , cuando  $-30 \leq \beta \leq 30$ . Se observa claramente que es una función creciente. De la misma forma la figura 2.b muestra el comportamiento de la función  $\beta - \delta_{jr}(\beta)$  cuando  $-30 \leq \beta \leq 30$ . También se observa claramente que es una función creciente. Así se muestra que la distribución normal verifica la CONDICION C1.

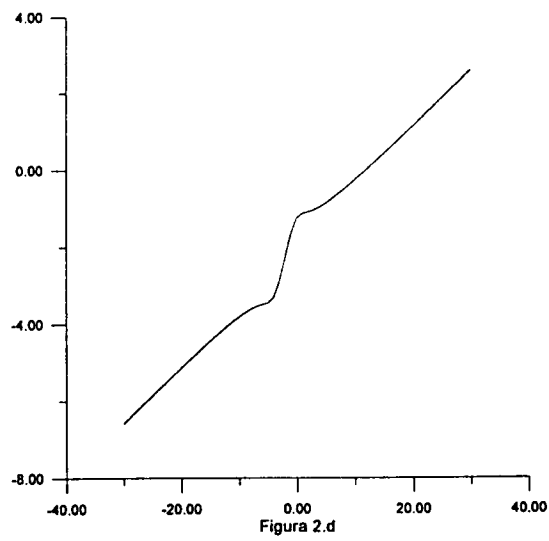
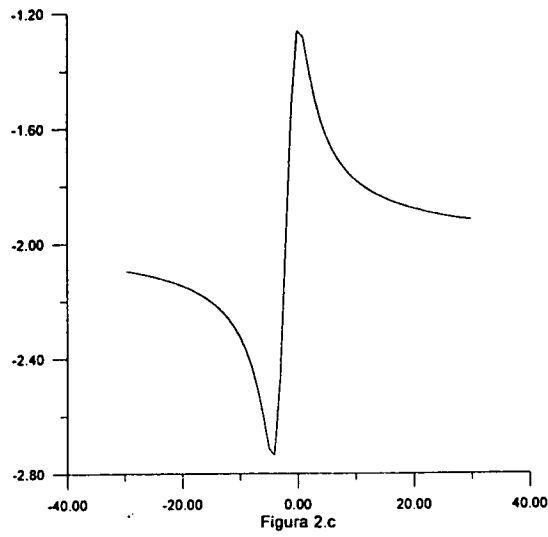
Sin embargo, si se toma  $\nu$  distribuida según una t-Student  $t(2)$ , resulta que, si bien la función  $\delta_{jr}(\beta)$  tiene un comportamiento similar al de la normal (no se ha dibujado), no así la función  $\beta - \delta_{jr}(\beta)$ , cuya gráfica se muestra en la figura 2.c, para  $-30 \leq \beta \leq 30$ . Se ve que no es una función creciente y por lo tanto la distribución t-Student no verifica la CONDICION C1 y no se puede asegurar la convergencia de los algoritmos que dependen de esta condición (aquellos incluidos en la sección 4.3 y 4.5).

Por otra parte, en la última sección se ha propuesto la CONDICION C2, dada por (28) y (29), como alternativa a la CONDICION C1. De hecho, C2 es menos restrictivas que C1. Para comprobarlo, tómese  $M = 1.1$ . Se verifica que la función  $M\beta - \delta_{jr}(\beta)$  es creciente para la distribución t-Student, según se muestra en la figura 2.d. Se desprende de ello que el algoritmo de esta sección 4.6, basado en las aproximaciones estocásticas, se puede utilizar cuando la distribución subyacente a los errores del modelo es una t-Student. Evidentemente, este algoritmo también es válido para la normal tomando  $M = 1$ .





Figuras 2.a, b: Funciones  $\delta(\beta)$  y  $\beta\delta(\beta)$  para la distribución normal



Figuras 2.c, d: Funciones  $\delta(\beta)$  y  $1.1\beta - \delta(\beta)$  para la distribución *t-Student*

## 4.7 Simulaciones

En esta sección se expondrán los resultados que se han obtenido por simulación tras haber aplicado los algoritmos de una iteración con correcciones en media tratados en este capítulo. Dichas simulaciones servirán, además, para comprobar los resultados teóricos vistos. Adicionalmente, se incluirán en esta sección distintas comparaciones entre los algoritmos de este capítulo con el embrión original de todos ellos: el MD puro de dos iteraciones, presentado en el capítulo 2 de esta memoria.

### PRIMERA SIMULACIÓN

Los datos y el modelo de censura de los que se parten son los mismos que en la sección 2.6. Debe hacerse notar que, en el método MD de dos iteraciones del capítulo 2, se puede calcular la estimación en la etapa  $N$  sin necesidad de calcular la estimación en las etapas anteriores. Esta es una de las causas, como ya fue indicado, por las que los métodos de dos iteraciones demandan una gran cantidad de tiempo, puesto que en cierta medida los cálculos se rehacen el completo en cada etapa primaria. Por el contrario, los métodos de una iteración no permiten obtener la estimación en la etapa  $N$  sin calcular previamente las de las etapas anteriores. Por esta razón, se han querido comparar ambos métodos, habiéndose calculado sus respectivas estimaciones en las etapas  $n = 1, \dots, N$ , con  $N = 1000$ . Los cálculos con el método de dos iteraciones se han extendido sobre un número  $P = 100$  de iteraciones del proceso secundario para cada etapa primaria. Las citadas iteraciones son suficientes para obtener en cada etapa una buena aproximación del punto límite buscado.

Las secuencia de estimaciones obtenidas haciendo uso del método de dos iteraciones y el de una iteración de la sección 4.3, ambos con punto de partida  $(25, 50)$ , aparece en la tabla 4. Las figuras 3.a, b, c, d muestran gráficamente las

trayectorias de las estimaciones para tamaños de muestra comprendidos entre 1 y 100. En concreto, las figuras 3.a y 3.b se corresponden con las trayectorias del proceso primario para la primera y segunda componente del vector  $a = (-4, 10)^t$  con el método MD puro de dos iteraciones. Por su parte, las figuras 3.c y 3.d muestran las trayectorias similares a las anteriores obtenidas con el algoritmo de una iteración presentado en la sección 4.3.

Adicionalmente, se ha simulado el comportamiento del método de una iteración presentado en la sección 4.5. Éste ha sido aplicado a la misma muestra de datos, con idéntico punto de arranque  $(25, 50)$ . Como se sabe, las iteraciones de este proceso de estimación vienen dadas por la ecuación

$$a_{n+1} = a_n + \alpha_n X^t \varepsilon(a_n).$$

Los tamaños de paso que se han elegido han sido

$$\alpha_n = \frac{1}{\|X^t \varepsilon(a_n)\|},$$

para que en cada iteración el salto sea de norma unidad. Las figuras 4.a y 4.c grafican las primeras 1000 iteraciones del proceso, respectivamente, para la primera y la segunda componente del parámetro  $a = (-4, 10)^t$ . Las figuras 4.b y 4.d muestran, por su parte, una ampliación de las anteriores figuras 4.a y 4.c, para las etapas finales posteriores a la 900. Tras las 1000 iteraciones anteriores, se han cambiado los tamaños de paso a  $\alpha_n = n^{-1.7}$  y se ha proseguido el proceso. Las figuras 4.e y 4.f muestran las trayectorias obtenidas para otras 1000 iteraciones con los nuevos pasos, para la primera y segunda componente, respectivamente.

## CONCLUSIONES

1. Con  $n$  pequeño las estimaciones con el método MD de dos iteraciones y el de una iteración de la sección 4.3 (figuras 3) se alejan considerablemente del punto de convergencia para después volver rápidamente a sus proximidades.

Aunque este comportamiento es similar en ambos métodos, la estimación de una iteración parece algo más regular, lo cual es alentador por la simplificación que éste supone frente al MD puro. El comportamiento citado no se produce, sin embargo, en la simulación del método presentado en la sección 4.5 (figuras 4). Éste, aparentemente, refleja una cierta regularidad que no muestran los otros dos métodos.

2. Si bien se tiene asegurada convergencia en  $L_1$  de las estimaciones, cosa que no se puede apreciar observando una única trayectoria, al menos, las figuras 3 y 4 muestran que las trayectorias tienden a estabilizarse, si bien no lo hacen exactamente en el punto  $(-4, 10)^t$ , que se esperaría con la convergencia casi segura.
3. A partir de esta simulación se puede apreciar un comportamiento que no se había visto teóricamente. En las figuras 5.a, b, c y d el trazo grueso corresponde a la estimación con el método de una iteración de la sección 4.3 y el trazo fino al de dos iteraciones. En concreto, las figuras 5.a y 5.b presentan las trayectorias superpuestas de ambos métodos para la primera componente del vector  $a$ , entre rangos de 0 a 100 iteraciones y de 500 a 1000, respectivamente. Las figuras 5.c y 5.d son similares para la componente segunda. Según se visualiza en dichas figuras, cuando  $n$  es grande ( $n > 50$ ) las estimaciones con ambos métodos de dos y de una iteración tienden a ser la misma. Además, las estimaciones del método de dos iteraciones son algo más irregulares que en el caso de una iteración. Esto permitiría conjeturar que la eficacia de ambos métodos sea asintóticamente la misma, pese a la simplificación que supone el método de una iteración. Este tema se tratará más adelante.
4. Si en el algoritmo de la sección 4.5 se pretenden tomar tamaños de paso de la forma

$$\alpha_n = n^p,$$

las condiciones del teorema 4.7 exigen que se cumpla

$$n^{\rho+1} \rightarrow 0, \quad \sum n^{\rho+1} = \infty \text{ y } \sum n^{\rho+\frac{1}{2}} < \infty.$$

Esto implica que se deba seleccionar  $\rho \in [-2, -\frac{3}{2})$ , para que el algoritmo converja. Sin embargo, en este caso, la distancia entre dos saltos consecutivos decrece rápidamente, razón por la cual el número de iteraciones necesarias para obtener una buena estimación del punto  $a = (-4, 10)^t$  puede hacerse muy grande. Para evitar esta situación es posible hacer una selección dinámica de  $\alpha_n$ . El objetivo es seleccionar para las primeras iteraciones tamaños de paso grande, para llegar en pocas iteraciones a las proximidades del punto  $(-4, 10)^t$ . Después, se decrece sucesivamente el tamaño de paso  $\alpha_n$ , para garantizar las propiedades asintóticas que son exigidas a los tamaños de paso. Así, se intentaría aumentar la eficacia del método.

5. En relación con el método de una iteración de la sección 4.5, se observa que, cuando se toma un tamaño de paso grande, las trayectorias oscilan, pero siempre en las proximidades del punto  $(-4, 10)^t$  (véase figuras 4.b y 4.d). Esto mismo ocurre cuando el tamaño de paso disminuye (véase figuras 4.e y 4.f), si bien la elongación de las oscilaciones tiende a decrecer a medida que el número de iteraciones aumenta. En todo caso, lo anterior es consistente con la convergencia en  $L_1$  al punto  $(-4, 10)^t$ , demostrada en el teorema 4.7.

## SEGUNDA SIMULACIÓN

En el método de una iteración se garantiza convergencia estocástica en probabilidad (vía  $L_1$  o  $L_2$ ) y en ley. Desde un punto de vista inferencial, dichas convergencias son suficientes, puesto que la última, más débil, es la utilizada habitualmente con el fin citado. Sin embargo, no está garantizada la independencia de la trayectoria con respecto al punto de arranque a diferencia de lo que ocurría con el MD puro de dos iteraciones. Esta segunda simulación ha pretendido

contrastar la dependencia o independencia de los algoritmos de una iteración con respecto al punto de arranque. Las figuras 6 muestran las trayectorias obtenidas con el algoritmo de una iteración presentado en la sección 4.3 cuando los puntos de arranque son  $(50, -30)^t$ ,  $(-50, 80)^t$  y  $(0, 0)^t$ , todos ellos muy alejados entre sí. Las figuras 6.a y 6.b muestran las trayectorias a corto y largo plazo para la primera componente del parámetro. Análogamente, las figuras 6.c y 6.d lo hacen para la segunda. En ambos casos, puede comprobarse que todas las trayectorias citadas coinciden a partir de la iteración 11, a menos de un error de  $10^{-6}$ .

#### CONCLUSION

Aparentemente, se produce una independencia de las trayectorias desde un  $n$  ( $n = 11$ , en nuestro caso) en adelante, tal como ocurría con el MD puro de dos iteraciones. De nuevo, este resultado refleja una conducta muy alentadora a favor de los métodos de una iteración. ¡Pese a la simplificación que ellos suponen, aparentemente, no se pierden las buenas propiedades de los métodos basados en dos iteraciones!

#### TERCERA SIMULACIÓN

Con los mismos datos de las dos simulaciones anteriores, en esta tercera simulación se pretende mostrar la evolución de los errores en  $L_1$  de las secuencias generadas por los algoritmos de una iteración. Para todos ellos la convergencia en  $L_1$  ha sido demostrada teóricamente. En consecuencia,  $E \|a_n - a\| \rightarrow 0$ , cuando  $n \rightarrow \infty$ . La figura 7 muestra este hecho para los algoritmos MD puro (de dos iteraciones) y MD de una iteración presentado en la sección 4.3. Para cada etapa (primaria en el MD puro) se ha evaluado  $E \|a_n - a\|$  empíricamente a partir de  $M = 60$  muestras diferentes, mediante la expresión habitual

$$E \|\widehat{a_n} - a\| = \frac{1}{60} \sum_{i=1}^{60} \|a_n(m) - a\|,$$

siendo  $a_n(m)$  la estimación para la  $m$ -ésima muestra. El trazo grueso de la citada figura se corresponde con el algoritmo MD de dos iteraciones. Éste sigue, como se sabe, trayectorias independientes del punto de arranque. No así el de una iteración (trazo fino de la figura 7) para el cual se ha graficado la trayectoria con punto de arranque  $a_1 = (34, 32)^t$ .

### CONCLUSIONES

1. La figura 7 confirma la convergencia  $E \|a_n - a\| \rightarrow 0$ , en todos los casos. Pese a ello, la rapidez de la convergencia parece ser ligeramente mayor para el MD puro, si bien para un tamaño  $n > 50$  prácticamente coinciden. Incluso para  $n > 30$  se producen cruces, lo cual, aparentemente, indica que las tasas de convergencia a cero de los errores en  $L_1$  son similares para ambos algoritmos (de nuevo, ¡pese a la simplificación que suponen los algoritmos de una iteración!).
2. Pese a que en los métodos de una iteración las trayectorias pueden depender del punto de arranque, las figuras 6 indican que esta dependencia es mínima. En consecuencia, haber tomado como punto de inicio  $(34, 32)^t$  es prácticamente irrelevante frente a las conclusiones citadas.

### CUARTA SIMULACIÓN

En esta última simulación se comprueba la normalidad asintótica de las estimaciones, demostrada en el teorema 4.4, estimándose, también, las matrices de covarianzas asociadas. Para ello, se ha calculado  $a_N$ , para  $N = 200$ , a partir de  $M = 60$  muestras diferentes.

Los contrastes habituales de bondad de ajuste (Kolmogorov-Smirnov y de la  $\chi^2$ ) pueden emplearse con los 60 valores de  $a_N$  obtenidos. Los p-valores resultantes usando el contraste de la  $\chi^2$  han sido 0.9453 y 0.9290 para cada una de las coordenadas del parámetro  $a$ . Con el contraste de Kolmogorov-Smirnov los



#### 4. Algoritmos de una iteración basados en correcciones en media

p-valores son del mismo orden.

Por último, la matriz de varianzas-covarianzas empírica para la citada muestra de tamaño 60 de  $a_N$  es

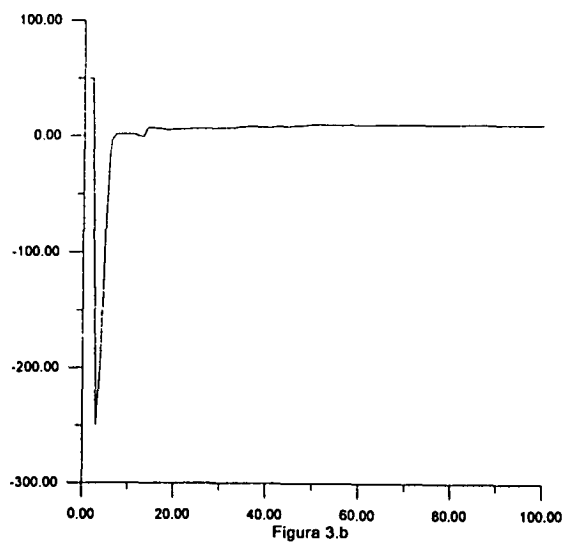
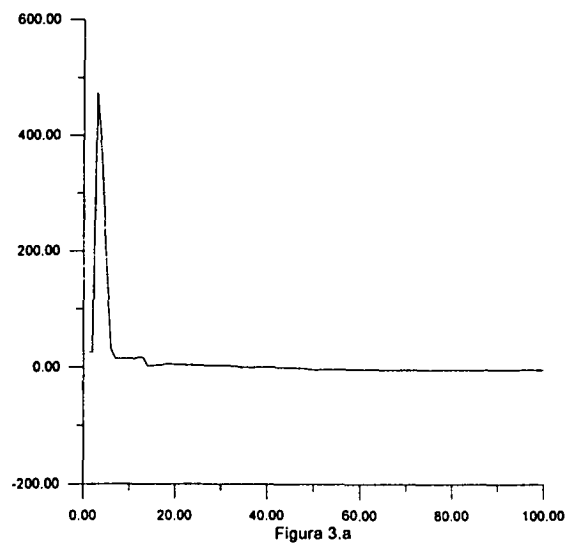
$$\hat{\Sigma} = \begin{pmatrix} 185.2 & -87.2 \\ -87.2 & 41.3 \end{pmatrix}.$$

#### CONCLUSION

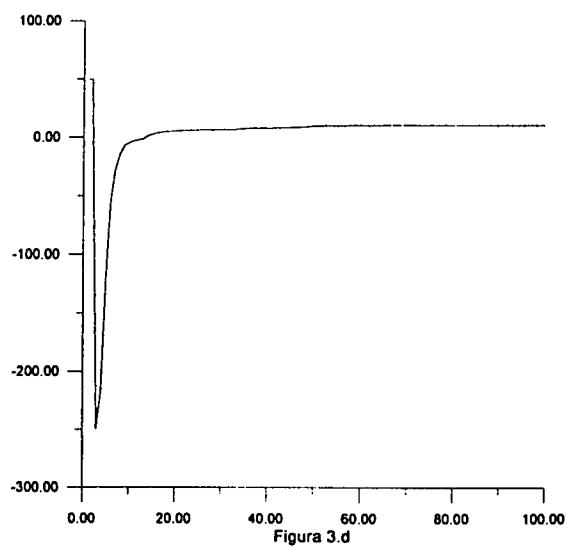
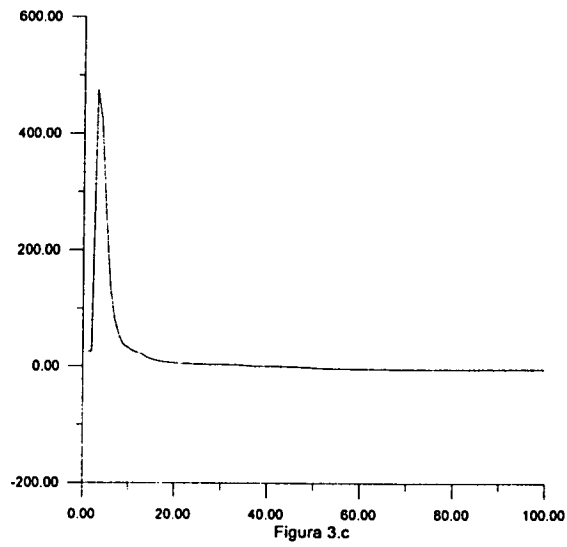
Puede comprobarse que la matriz de covarianzas anterior es similar a la obtenida para el proceso de dos iteraciones en la sección 2.6. De nuevo, aparentemente, lo anterior indica que pese a la simplificación que supone el método de una iteración frente el MD puro, las variabilidades asintóticas de ambos son del mismo rango.

<i>n</i>	Proceso de dos iteraciones		Proceso de una iteración	
			25	50
1	474.99542	-249.99725	474.99542	-249.99725
2	369.41260	-187.49756	424.07901	-217.49739
3	174.12247	-72.54400	259.86374	-118.33829
4	30.622290	-4.53251	137.36356	-54.64499
5	14.09649	1.10129	80.61238	-28.02524
6	14.09649	1.10129	51.49039	-14.24667
7	14.09649	1.10129	36.23812	-7.19063
8	14.09649	1.10129	31.09683	-4.99985
9	14.09649	1.10129	26.56226	-3.23885
10	16.35790	-0.70784	23.59492	-2.46498
11	16.35791	-0.70784	20.99416	-1.58322
12	1.69823	7.00738	15.35925	0.88987
13	1.69400	7.00999	11.970343	2.40073
14	2.94461	6.17625	9.46344	3.42061
44	-2.04358	9.18527	-1.15210	8.66886
45	-2.04358	9.18527	-1.39474	8.79960
46	-2.55304	9.52491	-1.58843	8.92453
47	-3.60512	10.17074	-1.937692	9.14084
48	-4.68194	10.73749	-2.68979	9.54773
49	-4.68194	10.73749	-3.15857	9.81258
50	-4.18377	10.40538	-3.44887	9.96965
51	-4.18377	10.40538	-3.62370	10.06764
52	-4.21197	10.42022	-3.76190	10.14616
53	-4.21197	10.42022	-3.85678	10.20194
54	-3.80090	10.14618	-3.89749	10.21527
55	-3.80090	10.14618	-3.90337	10.21366
56	-4.51353	10.51124	-4.19780	10.35530
995	-4.35349	10.22796	-4.34117	10.22028
996	-4.36305	10.23433	-4.34440	10.22260
997	-4.36305	10.23433	-4.34791	10.22487
998	-4.38348	10.24509	-4.36147	10.23198
999	-4.383489	10.24509	-4.36796	10.23561
1000	-4.42331	10.27164	-4.37459	10.24085

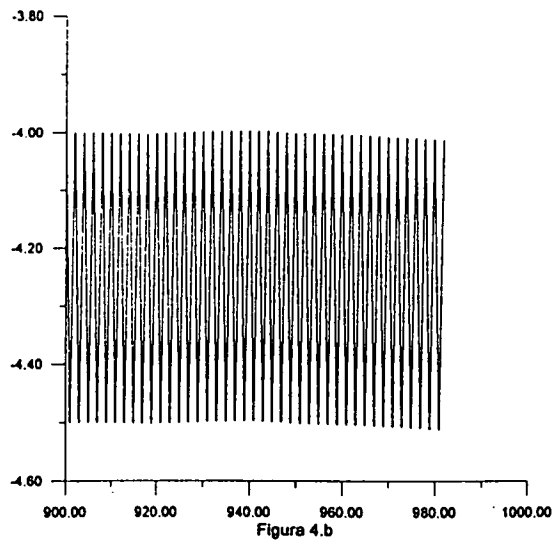
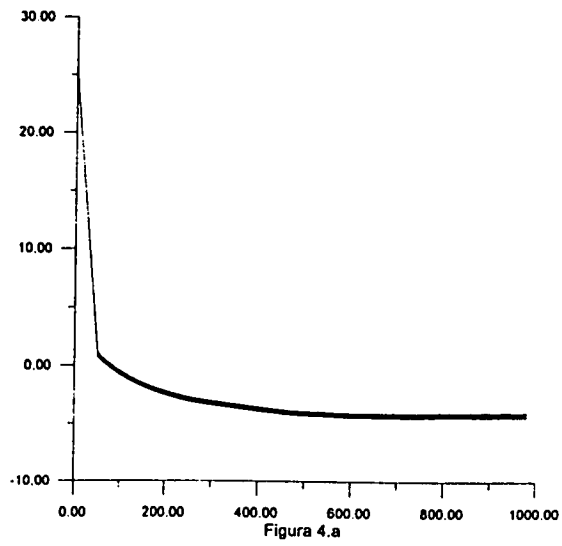
Tabla 4: *Estimaciones del proceso MD de dos y una iteración*



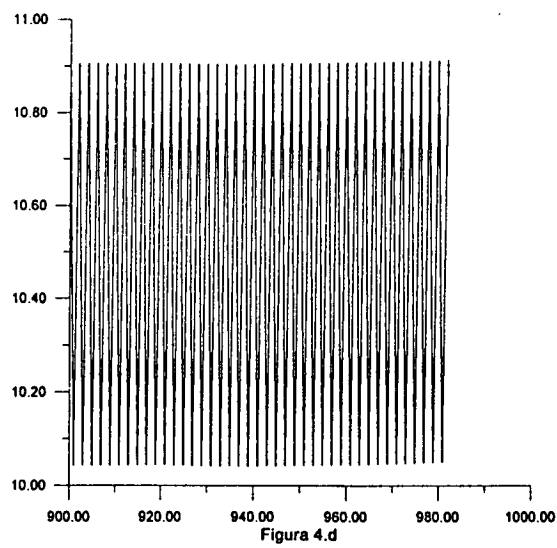
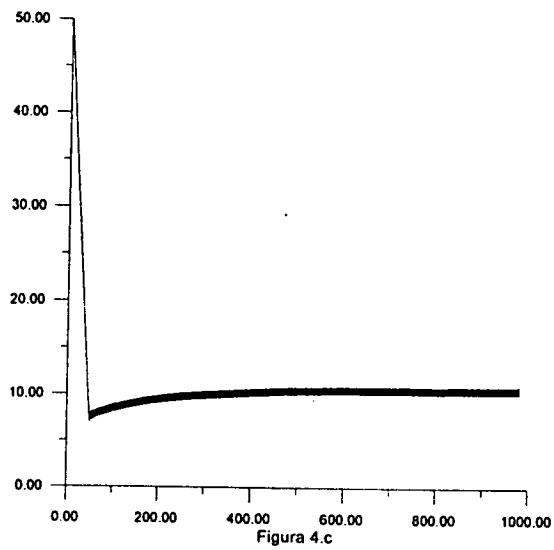
Figuras 3.a, b: *Estimación del vector  $a=(-4,10)$  mediante el algoritmo MD de dos iteraciones*



Figuras 3.c, d: Estimación del vector  $a=(-4,10)$  mediante el algoritmo MD de una iteraciones de la sección 4.3



Figuras 4: Estimaciones del vector  $a=(-4,10)$  mediante el algoritmo de una iteraciones de la sección 4.5



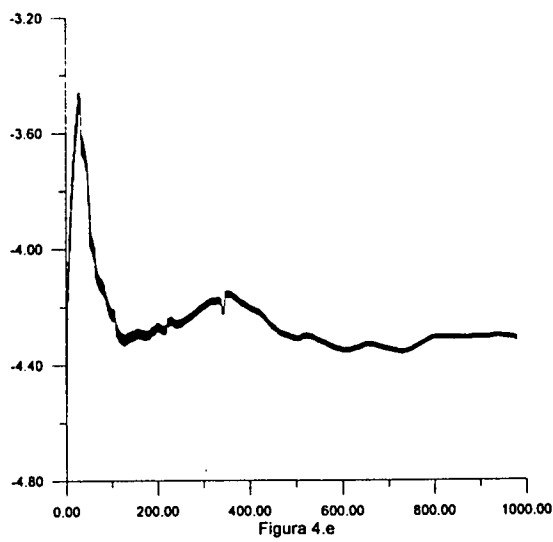


Figura 4.e

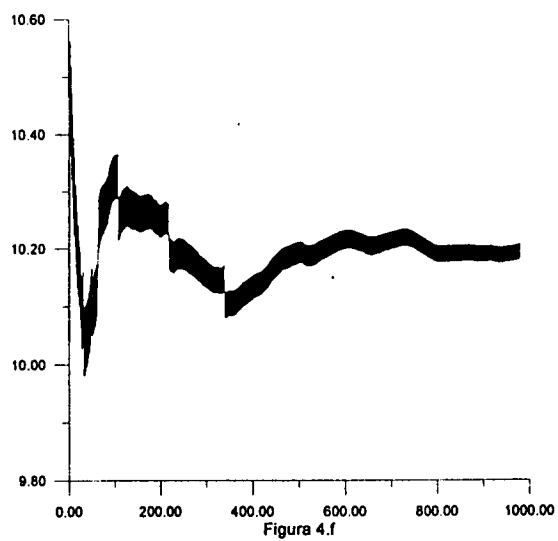
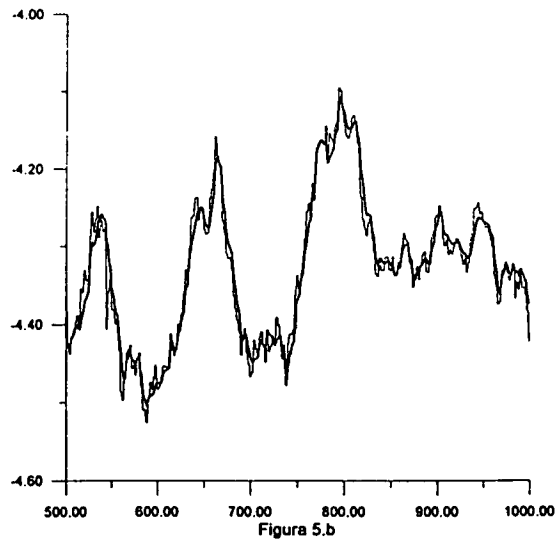
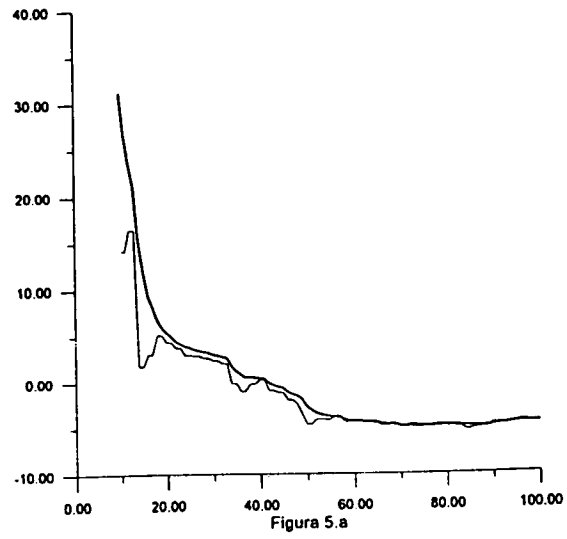
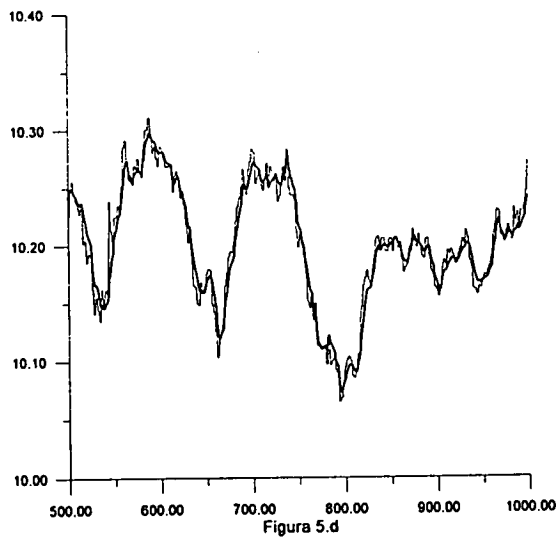
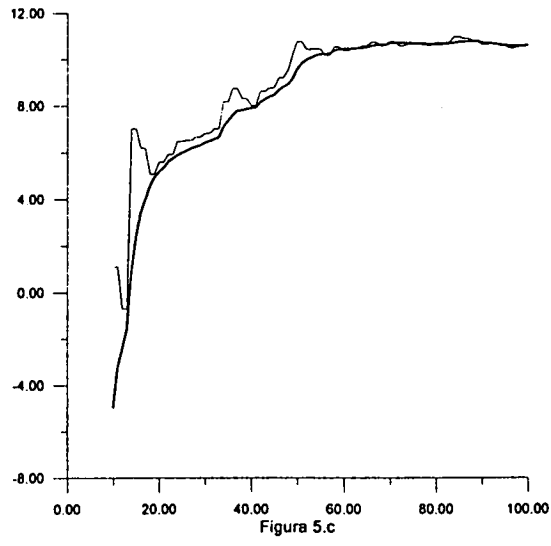


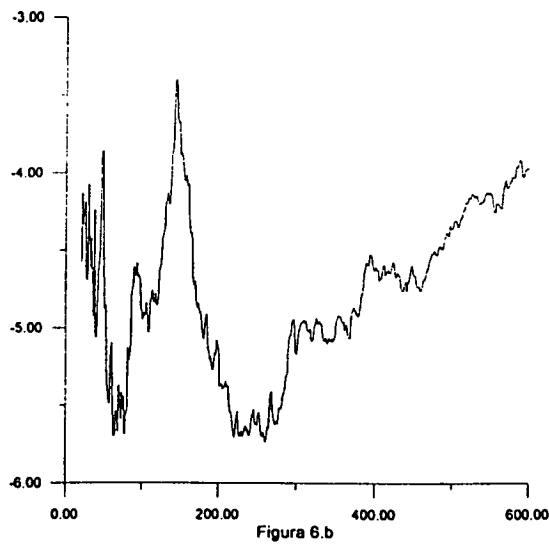
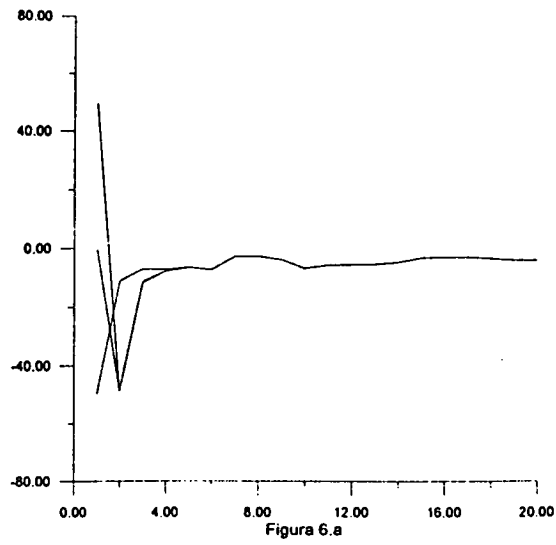
Figura 4.f



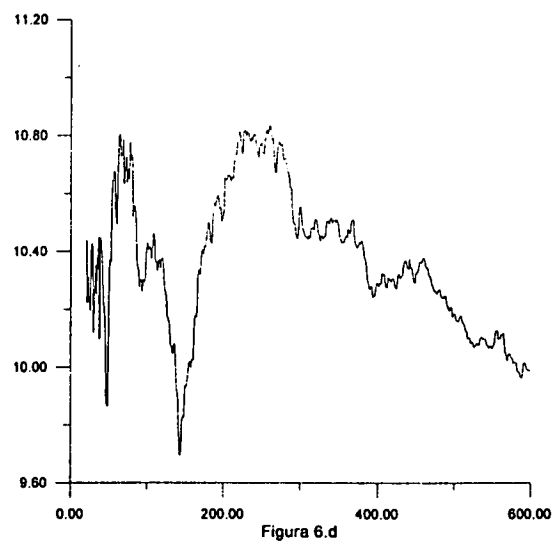
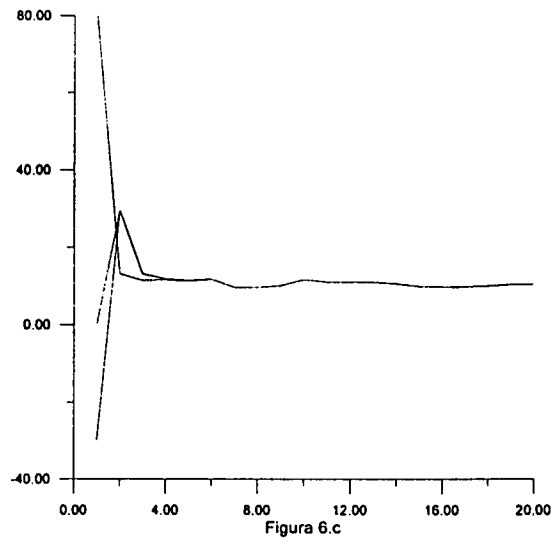
Figuras 5: Comparación de las estimaciones entre el algoritmo MD de una iteración de la sección 4.3 (trazo grueso) y el algoritmo MD de dos iteraciones (trazo fino)







Figuras 6: Estimaciones del punto  $a=(-4,10)$  con el algoritmo MD de una iteración de la sección 4.3 para diferentes puntos de arranque.



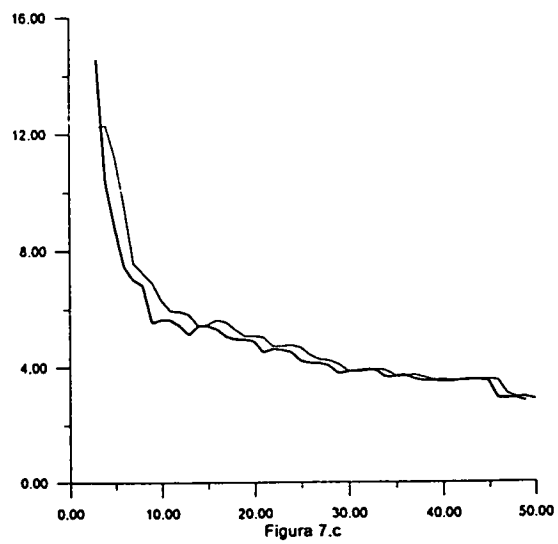


Figura 7: Comparación de los errores absolutos de estimación entre el algoritmo de una iteración de la sección 4.3 (trazo fino) y el MD de dos iteraciones (trazo grueso)

### Notas y comentarios

- Los algoritmos presentados en este capítulo constituyen versiones particulares de una iteración del algoritmo MD de dos iteraciones. Este último es específico dentro del contexto de estimación de modelos lineales con datos censurados. Sin embargo, los algoritmos de dos iteraciones aparecen de facto en otros muchos contextos de estimación (e. g., máxima verosimilitud, estimación bayesiana,...). Así pues, se podría intentar adaptar métodos de una iteración en todos ellos.
- Los algoritmos de una iteración presentados en las secciones 4.3, 4.5 y 4.6, por este orden, representan niveles cada vez más cercanos a los esquemas propios de las redes neuronales y las aproximaciones estocásticas tipo Kushner. El último de ellos coincide exactamente con los esquemas citados. Véase al respecto WHITE (1992) y KUSHNER Y YIN (1997).
- La elección de los tamaños de paso variables (teoremas 4.7 y 4.10) puede entrañar alguna dificultad. En este sentido, las elecciones adaptativas podrían aumentar la velocidad de las convergencias estocásticas de los algoritmos. En la primera simulación se incluye un ejemplo de ello. Al respecto, la utilización de los habituales tamaños de paso constantes podría ser conveniente en alguna fase de la ejecución del algoritmo. Véase, en este sentido, KUSHNER Y YIN (1997) y KUSHNER Y CLARK (1978).

## **5. Algoritmos de una iteración basados en correcciones en moda**

### **5.1 Introducción**

Como se ha indicado en la introducción del capítulo anterior, se centrará ahora la atención en simplificar los algoritmos modales de dos iteraciones, buscando, de nuevo, una iteración simple. La razón que justifica esta actuación ya ha sido mencionada en la citada introducción, a la cual remitimos en este momento. Adicionalmente, los comentarios vertidos en el capítulo 3, sobre la conveniencia de utilizar correcciones en moda en lugar de correcciones en media, continúan aquí siendo válidos. En primer lugar, pese a la simplificación que suponen los algoritmos de una iteración frente a los de iteración doble, los algoritmos presentados en el capítulo anterior no evitan los cálculos de cuadratura asociados a las correcciones en media, cuando las distribuciones de los errores son generales. En el capítulo 3 se observó que la sustitución de correcciones en media por correcciones modales relajaba los citados cálculos de cuadratura bajo las condiciones habituales de simetría y moda única de las densidades de los errores. En estos casos, las correcciones en moda son prácticamente libres de distribución, simplificándose los cálculos enormemente. Al respecto, se pueden consultar los resultados comparativos presentados en la sección 5.10 de este mismo capítulo. Pese a la simplificación citada, los resultados del capítulo 3 mostraban que las propiedades de convergencia de los algoritmos modales eran similares a las de los

algoritmos MD, con correcciones en media. El único coste afectaba a la rapidez de convergencia, deteriorándose ésta ligeramente. En todo caso, los resultados presentados al final de este capítulo indican que la pérdida en velocidad de convergencia afecta, tan sólo, al número de iteraciones necesarias para obtener una precisión dada, el cual aumenta ligeramente. Pese a ello, en términos de tiempo la relación se invierte, puesto que el mayor número de iteraciones que exigen los algoritmos modales se compensa con creces por la citada menor complejidad en los cálculos de las correcciones que deben llevarse a cabo en cada iteración.

En resumen, en este capítulo, como en el anterior, el esquema de actuación busca acercarse a planteamientos próximos a las redes neuronales y las aproximaciones estocásticas tipo Kushner. Puesto que el tipo de algoritmos propuestos son inéditos en el contexto de censura tratado en esta memoria, las referencias que enmarcan el desarrollo que figura a continuación coinciden con las del capítulo 4 anterior. En particular, WHITE (1992), HAYKIN (1994), KUSHNER Y YIN (1997), WASAN (1969) y BENVENISTE, METIVIER Y PRIOURET (1990).

## 5.2 Formalización del modelo

De nuevo, se tratará un modelo de regresión lineal del tipo

$$z_i = a^t x_i + \nu_i, \quad i = 1, \dots, n. \quad (33)$$

El parámetro  $a$  debe ser estimado y la variable independiente  $x_i$ , al igual que  $a$ , es  $m$ -dimensional. Los errores  $\nu_i$  son independientes e idénticamente distribuidos. Se asumirá que la variable dependiente ha sido obtenida de diferentes fuentes, pudiendo ser bien agrupada (posiblemente con diferentes intervalos de clasificación dentro de cada fuente) o no agrupada. El conjunto de observaciones  $I = \{1, 2, 3, \dots\}$  puede partitionarse en la forma  $I = I^g \cup I^{ng} = I_1 \cup \dots \cup I_s \cup I^{ng}$ ,

con la notación habitualmente usada hasta ahora. Dentro de cada  $I_j$ , las clases de agrupación están dadas por los puntos extremos

$$-\infty = c_{j,0} < c_{j,1} < \dots < c_{j,r_j-1} < c_{j,r_j} = \infty$$

$j = 1, \dots, s$ ,  $s < \infty$ . Para cada  $i \in I^g$ , solamente se conoce el criterio de clasificación y el intervalo que contiene a  $z_i$ . Por el contrario, si  $i \in I^{ng}$ , se conoce y se observa el valor exacto  $z_i$ . Asumamos, además, que los valores muestrales relacionados con los conjuntos  $I^{ng}$ ,  $I_1, \dots, I_s$  aparecen en la población con probabilidades positivas  $\pi_0, \pi_1, \dots, \pi_s$ .

Adicionalmente, de aquí en adelante se asumirá que  $X^t X$  es una matriz de rango completo, donde  $X^t = (x_1, \dots, x_n)$ . Por último, para simplificar los cálculos, en lo sucesivo se supondrá que los errores  $\nu_i$  son variables aleatorias continuas con función de densidad simétrica, unimodal en cero y con colas no crecientes (podría asumirse, pese a ello, que los errores tienen otra forma diferente, obteniéndose también convergencia, según se explicará más adelante).

### 5.3 Un primer algoritmo de una iteración. Algoritmo MmD de una iteración

Por similitud con el método propuesto en la sección 4.3, se propone a continuación un primer algoritmo de una única iteración que frente a aquél simplemente cambia las correcciones en media por correcciones en moda. Formalmente, es el siguiente.

INICIALIZACIÓN: Tómese un vector inicial  $a_1$  arbitrario.

ITERACIÓN: Si  $a_n$  es la estimación en la etapa  $n$  del verdadero vector  $a$ , la estimación para la etapa  $n + 1$  sigue la ecuación en diferencias

$$a_{n+1} = (X^t X)^{-1} X^t y(a_n),$$

donde  $X^t = (x_1, \dots, x_n)$  es una matriz de tamaño  $m \times n$  y el vector  $y(a_n) =$



$(y_1(a_n), \dots, y_n(a_n))^t$  tiene por componentes

$$y_i(a_n) = z_i, \quad \text{si } i \in I^{ng}$$

$$= a_n^t x_i + \gamma(-a_n^t x_i + c_{j,r}, -a_n^t x_i + c_{j,r+1}), \text{ si } i \in I_j, z_i \in (c_{j,r}, c_{j,r+1}].$$

Para cada  $j = 1, \dots, s$  y  $r = 0, 1, \dots, r_j - 1$ , la función  $\gamma(-a_n^t x_i + c_{j,r}, -a_n^t x_i + c_{j,r+1})$  iguala el valor modal del error  $\nu$ , condicionado a que i) el verdadero parámetro es el valor actual de la estimación  $a_n$  y ii)  $\nu$  está en el intervalo correspondiente asociado. En concreto, manteniendo la notación del capítulo anterior,  $\gamma(\beta + c_{j,r}, \beta + c_{j,r+1}) = \delta_{jr}(\beta)$  con

$$\delta_{jr}(\beta) = \text{Moda}(\nu | \nu \in (\beta + c_{j,r}, \beta + c_{j,r+1})).$$

El algoritmo propuesto se llamará, de aquí en adelante, algoritmo MdD de una iteración (pese a la transgresión semántica ya apuntada al presentar el anterior algoritmo MD de una iteración).

Las funciones  $\delta_{jr}$  ahora son calculables explícitamente. Teniendo en cuenta la forma de los errores antes citados, es claro que para cada  $j = 1, \dots, s$  y  $r = 0, 1, \dots, r_j - 1$ ,

$$\delta_{jr}(\beta) = \begin{cases} \beta + c_{j,r+1}, & \text{si } \beta < -c_{j,r+1} \\ 0, & \text{si } -c_{j,r+1} < \beta < -c_{j,r} \\ \beta + c_{j,r}, & \text{si } \beta > -c_{j,r} \end{cases},$$

donde pudiera ocurrir que  $c_{j,r+1} = \infty$  y/o  $c_{j,r} = -\infty$ . Esta función  $\delta_{jr}$  es continua, lineal a trozos y con pendiente uno o cero (por consiguiente, no decreciente). Adicionalmente,

$$\beta - \delta_{jr}(\beta) = \begin{cases} -c_{j,r+1}, & \text{si } \beta < -c_{j,r+1} \\ \beta, & \text{si } -c_{j,r+1} < \beta < -c_{j,r} \\ -c_{j,r}, & \text{si } \beta > -c_{j,r} \end{cases},$$

es, de nuevo, una función continua, lineal a trozos y no decreciente con pendiente uno o cero. En definitiva, por construcción se cumple que

$$\delta_{jr}(\beta) \text{ y } \beta - \delta_{jr}(\beta) \text{ son no decrecientes,} \quad (34)$$

cualquiera que sean  $j = 1, \dots, s$  y  $r = 0, 1, \dots, r_j - 1$ . Es posible que esta condición

se cumpla aun cuando los errores no tengan la forma particular asumida. De hecho, podrían no ser unimodales y seguirse cumpliendo la condición (34). En todo caso, por simplicidad seguiremos aceptando que los errores tienen la forma propuesta, ya que no supone una gran limitación en la práctica, al coincidir con las hipótesis generalmente aceptadas.

El principal resultado de convergencia del método MdD de una iteración se enuncia a continuación. Se asumirán determinadas condiciones técnicas relativas a los valores  $x_i$  de las variables independientes, para asegurar la convergencia en media de orden uno y, en consecuencia, la convergencia en probabilidad. La principal diferencia con el resultado del capítulo precedente afecta a la no acotación de los valores  $x_i$ , aunque más adelante se tratará de resolver ese problema.

**Teorema 5.1** Sean  $(X^t X)^{ng}$  y  $(X^t X)^g$  matrices definidas positivas para cada  $n \in \mathcal{N}$ . Supongamos que existe  $\rho > 1$  para el cual las siguientes condiciones técnicas se cumplen

$$\inf_n \lambda_n = \lambda > 0, \\ \max_{i \leq n} \|x_i\|^2 = O(n^{\rho-1}),$$

donde  $\lambda_n$  denota el mínimo autovalor de la matriz  $n^{-\rho}(X^t X)^{ng}$ . En estas condiciones,  $a_n \xrightarrow{L_1} a$  y, en consecuencia,  $a_n$  es un estimador consistente de  $a$ .

OBSERVACIONES: La segunda propiedad indica que existe  $K < \infty$  tal que

$$n^{1-\rho} \max_{i \leq n} \|x_i\|^2 \leq K, \text{ para cada } n \in \mathcal{N}.$$

El máximo autovalor de la matriz  $n^{-\rho}(X^t X)^g = n^{-\rho} \sum_{i \in I_n^g} x_i x_i^t$  coincide con su traza que es igual a

$$n^{-\rho} \sum_{i \in I_n^g} \|x_i\|^2.$$

Denotando por  $\delta_n$  el máximo autovalor de la matriz  $n^{-\rho}(X^t X)^g$ , se tendrá, a partir

de la segunda hipótesis del teorema, que

$$\delta_n = n^{-\rho} \sum_{i \in I_n^g} \|x_i\|^2 \leq n^{1-\rho} \max_{i \leq n} \|x_i\|^2 \leq K < \infty.$$

En definitiva, podemos escribir

$$\sup_n \delta_n = M < \infty.$$

Por otra parte, las hipótesis del teorema implican que

$$0 < \lambda \leq \lambda_n \leq n^{-\rho} \sum_{i=1}^n \|x_i\|^2 \leq n^{1-\rho} \max_{i \leq n} \|x_i\|^2$$

y

$$n^{-\rho} \sum_{i=1}^n \|x_i\| \leq n^{-\rho} \left[ \sum_{i=1}^n \|x_i\|^2 \right]^{\frac{1}{2}} \leq n^{-\frac{\rho}{2}} \left( n^{1-\rho} \max_{i \leq n} \|x_i\|^2 \right)^{\frac{1}{2}}.$$

Adicionalmente, la siguiente condición también se cumple

$$n^{-\rho} \sum_{i=1}^n \|x_i\| = O\left(n^{-\frac{\rho}{2}}\right). \quad (35)$$

DEMOSTRACION DEL TEOREMA: Defínense las siguientes variables binarias con valores unitarios (siendo nulas en caso contrario):  $\delta_{i,0} = 1$  si  $i \in I^{ng}$ ;  $\eta_{i,j,h} = 1$  si  $i \in I_j \subset I^g$  y  $z_i \in (c_{j,h-1}, c_{j,h}]$ , donde  $i \in \mathcal{N}$ ,  $j = 1, \dots, s$  y  $h = 1, \dots, r_j$ . Obsérvese que  $y(a)$  puede escribirse como

$$y(a) = Xa + \varepsilon(a),$$

donde las componentes de  $\varepsilon(a) = (\varepsilon_1(a), \dots, \varepsilon_n(a))^t$  son independientes y definidas como

$$\varepsilon_i(a) = \delta_{i,0} \nu_i + \sum_{j=1}^s \sum_{h=1}^{r_j} \eta_{i,j,h} \gamma(-a^t x_i + c_{j,h-1}, -a^t x_i + c_{j,h}).$$

Por otra parte, definiendo

$$\varepsilon(a_n) = y(a_n) - Xa_n,$$

se puede escribir

$$a_{n+1} = (X^t X)^{-1} X^t y(a_n) = a_n + (X^t X)^{-1} X^t \varepsilon(a_n). \quad (36)$$

Teniendo en cuenta la propiedad (34), razonando como en el capítulo anterior, se puede escribir

$$y(a_n) - y(a) = (I - M^*) X(a_n - a),$$

donde  $M^* = \text{diag}(m_i^*)$ , siendo  $m_i^* = 1$  si  $i \in I_n^{mg}$ , y  $0 \leq m_i^* \leq 1$  si  $i \in I_n^g$ . Se puede concluir, pues,

$$\begin{aligned} \varepsilon(a_n) - \varepsilon(a) &= -Xa_n + Xa + (I - M^*) X(a_n - a) \\ &= -M^* X(a_n - a). \end{aligned}$$

Sumando a cada término el vector  $a$ , se llega a

$$a_{n+1} - a = \left[ I - (X^t X)^{-1} X^t M^* X \right] (a_n - a) + (X^t X)^{-1} X^t \varepsilon(a),$$

o equivalentemente

$$a_{n+1} - a = (X^t X)^{-1} X^t (I - M^*) X(a_n - a) + (X^t X)^{-1} X^t \varepsilon(a). \quad (37)$$

Es fácil observar que la condición  $E(\varepsilon_i(a)) = 0$ , que se verificaba en el teorema 4.1, ahora no tiene por qué cumplirse. Sin embargo, se probará que  $|E(\varepsilon_i(a))| \leq H$ , para cada  $x_i \in R^m$ . En efecto, denotemos por  $F_\nu$  a la función de distribución asociada al error  $\nu$ . Es claro que si  $E(\nu^r) < \infty$ ,  $r \geq 1$ , entonces  $\int_z^\infty y^r dF_\nu(y) \xrightarrow{z \rightarrow \infty} 0$ . Por tanto, si  $-\infty < a < b$  y  $z \geq -a$  se cumple que

$$\begin{aligned} 0 &\leq (z+a)^r (F_\nu(z+b) - F_\nu(z+a)) \leq (z+a)^r (1 - F_\nu(z+a)) \\ &= (z+a)^r \int_{z+a}^\infty dF_\nu(y) \leq \int_{z+a}^\infty y^r dF_\nu(y) \xrightarrow{z \rightarrow \infty} 0. \end{aligned}$$

Tomando  $z$  suficientemente grande, lo anterior conduce a que, cuando  $-\infty < c_{j,h-1} < c_{j,h} \leq \infty$ ,

$$\begin{aligned} g_{jh}^r(z) &= \text{Pr}(\nu_i \in (z + c_{j,h-1}, z + c_{j,h}]) (\gamma(z + c_{j,h-1}, z + c_{j,h}))^r \\ &= \text{Pr}(\nu_i \in (z + c_{j,h-1}, z + c_{j,h}]) (z + c_{j,h-1})^r \xrightarrow{z \rightarrow \infty} 0. \end{aligned}$$

Si por el contrario  $-\infty = c_{j,h-1} < c_{j,h}$ , de nuevo para  $z$  suficientemente grande

es  $\gamma(z + c_{j,h-1}, z + c_{j,h}) = 0$ , con lo cual también

$$g_{jh}^r(z) = \Pr(\nu_i \in (z + c_{j,h-1}, z + c_{j,h}]) (\gamma(z + c_{j,h-1}, z + c_{j,h}))^r \xrightarrow{z \rightarrow \infty} 0.$$

En definitiva, se ha demostrado que

$$g^r(z) = \sum_{j=1}^s \pi_j \sum_{h=1}^{r_j} g_{jh}^r(z)$$

es una función continua que converge a cero cuando  $z \rightarrow \infty$ . Un razonamiento idéntico al realizado permite ver que también converge a cero cuando  $z \rightarrow -\infty$ . En conclusión, las funciones  $g^r(z)$  están todas acotadas. La esperanza en cuestión coincide con

$$\begin{aligned} E(\varepsilon_i(a)) &= \pi_0 E(\nu_i) \\ &+ \sum_{j=1}^s \pi_j \sum_{h=1}^{r_j} \Pr(\nu_i \in -a^t x_i + (c_{j,h-1}, c_{j,h}]) \delta_{jh-1}(-a^t x_i) \\ &= \sum_{j=1}^s \pi_j \sum_{h=1}^{r_j} g_{jh}^1(-a^t x_i) = g^1(-a^t x_i), \end{aligned}$$

de donde se desprende que debe existir  $H < \infty$  (cota de  $g^1$ ) tal que  $|E(\varepsilon_i(a))| \leq H$ , cualquiera que sea el vector independiente  $x_i \in \mathcal{R}^m$  de la observación  $i$ -ésima.

Queda claro que utilizar las correcciones en modas condicionadas no produce errores muestrales con esperanza nula, a diferencia de lo que ocurría en caso de utilizar correcciones en medias condicionadas.

Observemos también que

$$E(\varepsilon_i(a)^2) = g^2(-a^t x_i),$$

con lo cual, de nuevo, debe existir  $H^* < \infty$  cumpliendo que  $0 \leq E(\varepsilon_i(a)^2) \leq H^*$  cualquiera que sea el vector  $x_i \in \mathcal{R}^m$ .

Calculando la norma euclídea, y aplicando la desigualdad triangular, en la expresión (37), resulta que

$$\|a_{n+1} - a\| \leq \mu_n \|a_n - a\| + \left\| (X^t X)^{-1} X^t \varepsilon(a) \right\|,$$

donde  $\mu_n$  es un valor real que denota el mayor autovalor de la matriz definida positiva  $(X^t X)^{-1} X^t (I - M^*) X$ . Los autovalores de la matriz anterior coinciden, por semejanza, con los de

$$(X^t X)^{-\frac{1}{2}} X^t (I - M^*) X (X^t X)^{-\frac{1}{2}},$$

con lo cual  $\mu_n$  es

$$\mu_n = \max_{\|u\|=1} \left| \frac{u^t X^t (I - M^*) X u}{u^t X^t X u} \right| = \max_{\|u\|=1} \left| \frac{\sum_{i \in I_n^g} (1 - m_i^*) u^t x_i x_i^t u}{u^t X^t X u} \right|.$$

Recordemos que, para cada  $i \in I_n^g$ , es  $0 \leq 1 - m_i^* \leq 1$ , y que tanto  $\sum_{i \in I_n^g} x_i x_i^t$  como  $X^t X$  son matrices definidas positivas. Resulta, pues, que la última expresión está acotada superiormente por

$$\begin{aligned} \max_{\|u\|=1} \frac{\sum_{i \in I_n^g} u^t x_i x_i^t u}{u^t X^t X u} &= \max_{\|u\|=1} \left( \frac{u^t X^t X u}{\sum_{i \in I_n^g} u^t x_i x_i^t u} \right)^{-1} \\ &= \max_{\|u\|=1} \left( 1 + \frac{\sum_{i \in I_n^g} u^t x_i x_i^t u}{\sum_{i \in I_n^g} u^t x_i x_i^t u} \right)^{-1} \\ &\leq \left( 1 + \frac{\min_{\|u\|=1} \sum_{i \in I_n^g} u^t x_i x_i^t u}{\max_{\|u\|=1} \sum_{i \in I_n^g} u^t x_i x_i^t u} \right)^{-1}. \end{aligned}$$

Puesto que

$$\min_{\|u\|=1} \sum_{i \in I_n^g} u^t x_i x_i^t u = \min_{\|u\|=1} u^t (X^t X)^{ng} u = \lambda_n n^\rho > \lambda n^\rho > 0$$

y

$$\max_{\|u\|=1} \sum_{i \in I_n^g} u^t x_i x_i^t u = \max_{\|u\|=1} u^t (X^t X)^g u = \delta_n n^\rho < M n^\rho < \infty,$$

basta tomar

$$\tau = \left( 1 + \frac{\lambda}{M} \right)^{-1}$$

para poder afirmar que

$$\mu_n \leq \tau < 1.$$

Así pues, se puede escribir

$$\|a_{n+1} - a\| \leq \tau \|a_n - a\| + \left\| (X^t X)^{-1} X^t \varepsilon(a) \right\|.$$

El mínimo autovalor de la matriz definida positiva  $X^t X$  es no inferior al mínimo

autovalor de  $(X^t X)^{ng}$ , que coincide con  $n^\rho \lambda_n$ . Por consiguiente, llamando  $c_n = \left\| (X^t X)^{-1} X^t \varepsilon(a) \right\|$  se llega a

$$\begin{aligned} E(c_n^2) &\leq \lambda_n^{-2} n^{-2\rho} E \|X^t \varepsilon(a)\|^2 \\ &= \lambda^{-2} n^{-2\rho} E [\varepsilon(a)^t X X^t \varepsilon(a)] \\ &= \lambda^{-2} n^{-2\rho} E \left[ \sum_{i,j=1}^n \varepsilon_i(a) \varepsilon_j(a) x_i^t x_j \right]. \end{aligned}$$

Finalmente, por la independencia de las observaciones y por la acotación de las esperanzas de los errores de observación, es claro que

$$|E(\varepsilon_i(a) \varepsilon_j(a))| = |E(\varepsilon_i(a))| |E(\varepsilon_j(a))| \leq H^2,$$

resultando

$$\begin{aligned} E(c_n^2) &\leq \lambda^{-2} n^{-2\rho} \sum_{i,j=1}^n |E(\varepsilon_i(a) \varepsilon_j(a))| \|x_i\| \|x_j\| \\ &\leq \lambda^{-2} n^{-2\rho} \max\{H^2, H^*\} \sum_{i,j=1}^n \|x_i\| \|x_j\| \\ &= \lambda^{-2} n^{-2\rho} \max\{H^2, H^*\} \left[ \sum_{i=1}^n \|x_i\| \right]^2. \end{aligned}$$

La condición (35), asegura

$$E(c_n^2) \leq O(n^{-\rho}) \xrightarrow{n \rightarrow \infty} 0.$$

En definitiva, se tiene que

$$E \|a_{n+1} - a\| \leq \tau E \|a_n - a\| + E(c_n),$$

donde  $0 < \tau < 1$  y

$$d_n = E(c_n) \leq (E(c_n^2))^{\frac{1}{2}} \leq O(n^{-\frac{\rho}{2}}) \xrightarrow{n \rightarrow \infty} 0.$$

Por tanto,

$$E \|a_{n+1} - a\| \leq \tau^n E \|a_1 - a\| + \sum_{i=1}^n \tau^{n-i} d_i.$$

Aplicando el lema 4.2 del capítulo precedente, se concluye que

$$E \|a_n - a\| \xrightarrow{n \rightarrow \infty} 0,$$

de donde, también  $a_n \xrightarrow{P} a$ , terminándose así la demostración.  $\square$

## 5.4 Resultado de convergencia en distribución

Se enunciará en esta sección un teorema de convergencia en distribución a la normal del estimador propuesto  $a_n$ . Una posible estimación de la matriz de covarianzas asintótica proporcionaría un medio para construir intervalos de confianza y realizar contrastes de hipótesis sobre el valor del parámetro  $a$ .

**Teorema 5.2** *Bajo las condiciones del teorema 5.1 existe una matriz de covarianzas  $\Sigma$  tal que*

$$n^{\frac{\rho}{2}} (a_n - E(a_n)) \xrightarrow[n \rightarrow \infty]{D} N(0, \Sigma).$$

DEMOSTRACION: Denótese  $X_n^t = (x_1, \dots, x_n)$  para notar la dependencia de  $n$ . Definamos  $A_n^* = (X_n^t X_n)^{-1} X_n^t (I - M^*) X_n$  y  $A_n = E(A_n^*) = (X_n^t X_n)^{-1} X_n^t (I - E(M^*)) X_n$ . Teniendo en cuenta la expresión (37) y llamando  $\varepsilon^n(a) = (\varepsilon_1(a), \dots, \varepsilon_n(a))$ , de nuevo para hacer explícita su dependencia de  $n$ , se puede poner

$$a_{n+1} - a = A_n(a_n - a) + (X_n^t X_n)^{-1} X_n^t \varepsilon^n(a) + (A_n^* - A_n)(a_n - a).$$

Probaremos que  $n^{\frac{\rho}{2}} (A_n^* - A_n)(a_n - a) \xrightarrow{P} 0$ . En efecto, para ello, en primer lugar, denotemos por  $p_{jk} = n^{-\rho} \sum_{i=1}^n x_{ij} x_{ik} (m_i^* - E(m_i^*))$  la componente  $j, k$  ( $= 1, \dots, m$ ) de la matriz cuadrada  $n^{-\rho} X_n (M^* - E(M^*)) X_n^t$ . Como  $a_n \xrightarrow{P} a$ , teniendo en cuenta que  $n^{1-\rho} \max_{i \leq k} \|x_i\|^2 = O(n^{\rho-1})$ , resulta que

$$|p_{jk}| \leq n^{1-\rho} \max_{i \leq k} \|x_i\|^2 n^{-1} \sum_{i=1}^n |m_i^* - E(m_i^*)| \xrightarrow{P} 0.$$

En segundo lugar, se puede probar que  $\|n^{\frac{\rho}{2}} (a_n - a)\|$  esta acotado en  $L_1$ . Para



ello, recordemos que se puede escribir

$$E \left\| n^{\frac{\epsilon}{2}} (a_{n+1} - a) \right\| \leq n^{\frac{\epsilon}{2}} \tau^n E \|a_1 - a\| + \sum_{i=1}^n \tau^{n-i} n^{\frac{\epsilon}{2}} d_i,$$

donde  $d_n \leq O(n^{-\frac{\epsilon}{2}}) \xrightarrow{n \rightarrow \infty} 0$ . Como  $n^{\frac{\epsilon}{2}} \tau^n E \|a_1 - a\| \xrightarrow{n \rightarrow \infty} 0$  y, también,  $\sum_{i=1}^n \tau^{n-i} n^{\frac{\epsilon}{2}} d_i = O(1)$ , queda demostrada la acotación deseada. Por último, observemos que

$$\left\| n^{\frac{\epsilon}{2}} (A_n^* - A_n) (a_n - a) \right\| \leq \lambda^{-1} \left\| n^{\frac{\epsilon}{2}} (a_{n+1} - a) \right\| \sum_{j,k=1}^m |p_{jk}|,$$

lo cual permite afirmar la convergencia  $n^{\frac{\epsilon}{2}} (A_n^* - A_n) (a_n - a) \xrightarrow{P} 0$ .

Si la distribución asintótica de  $n^{\frac{\epsilon}{2}} (a_n - a)$  coincide con la de  $A_n (a_n - a) + (X_n^t X_n)^{-1} X_n^t \epsilon^n(a)$ , iterando  $n$  veces se puede ver que, además, coincide con la distribución asintótica de

$$n^{\frac{\epsilon}{2}} \prod_{i=1}^n A_i (a_1 - a) + n^{\frac{\epsilon}{2}} \sum_{i=1}^n \left[ \prod_{j=i}^n A_j \right] (X_i^t X_i)^{-1} X_i^t \epsilon^i(a).$$

Observemos ahora que

$$n^{\frac{\epsilon}{2}} \left\| \prod_{i=1}^n A_i (a_1 - a) \right\| \leq n^{\frac{\epsilon}{2}} \tau^n \|a_1 - a\| \xrightarrow{n \rightarrow \infty} 0,$$

siendo  $\tau = (1 - \frac{\lambda}{M})^{-1} < 1$  el valor definido a lo largo de la demostración del teorema anterior. Se sigue que la distribución límite de  $n^{\frac{\epsilon}{2}} (a_{n+1} - a)$  coincide con la de  $n^{\frac{\epsilon}{2}} \sum_{i=1}^n \left[ \prod_{j=i}^n A_j \right] (X_i^t X_i)^{-1} X_i^t \epsilon^i(a)$ , de igual forma que la de  $n^{\frac{\epsilon}{2}} (a_{n+1} - E(a_{n+1}))$  coincide con la distribución de  $n^{\frac{\epsilon}{2}} \sum_{i=1}^n \left[ \prod_{j=i}^n A_j \right] (X_i^t X_i)^{-1} X_i^t (\epsilon^i(a) - E(\epsilon^i(a)))$ , que estudiaremos a continuación. Este último término se puede poner de la siguiente manera:

$$\begin{aligned} & n^{\frac{\epsilon}{2}} \sum_{i=1}^n \left[ \prod_{j=i}^n A_j \right] (X_i^t X_i)^{-1} X_i^t (\epsilon^i(a) - E(\epsilon^i(a))) \quad (38) \\ &= n^{\frac{\epsilon}{2}} \sum_{i=1}^n \left[ \prod_{j=i}^n A_j \right] (X_i^t X_i)^{-1} \sum_{k=1}^n x_k (\epsilon^i(a) - E(\epsilon^i(a))) \end{aligned}$$

5. Algoritmos de una iteración basados en correcciones en moda

$$= n^{\frac{p}{2}} \sum_{k=1}^n \sum_{i=k}^n \left[ \prod_{j=i}^n A_j \right] (X_i^t X_i)^{-1} x_k (\varepsilon^i(a) - E(\varepsilon^i(a))).$$

Llamemos  $\sum_{i=k}^n \left[ \prod_{j=i}^n A_j \right] (X_i^t X_i)^{-1} = C_n^k$  y observemos que se puede escribir la siguiente cadena de desigualdades

$$\begin{aligned} \|C_n^k x_k\| &\leq \sum_{i=k}^n \left\| \left[ \prod_{j=i}^n A_j \right] (X_i^t X_i)^{-1} x_k \right\| \\ &\leq \sum_{i=k}^n \tau^{n-i+1} \left\| (X_i^t X_i)^{-1} x_k \right\| \\ &\leq \sum_{i=k}^n \tau^{n-i+1} \lambda^{-1} i^{-\rho} \|x_k\| \\ &= \lambda^{-1} \|x_k\| \sum_{i=k}^n \tau^{n-i+1} i^{-\rho}. \end{aligned}$$

La tercera desigualdad es consecuencia de que el mínimo autovalor de  $X_i^t X_i$  no puede ser inferior al mínimo autovalor de  $(X_i^t X_i)^{ng}$  que, por hipótesis, esta acotado inferiormente por  $\lambda i^\rho$ .

Si  $f_{jn}^k$  denota la fila  $j$ -ésima de la matriz  $C_n^k$ ,  $j = 1, \dots, m$ , la componente  $j$ -ésima de (38) sería

$$n^{\frac{p}{2}} \sum_{k=1}^n f_{jn}^k x_k (\varepsilon_k(a) - E(\varepsilon_k(a)))$$

que tiene media cero y verifica lo siguiente:

$$\begin{aligned} \text{Var} \left( f_{jn}^k x_k \varepsilon_k(a) \right) &= \left( f_{jn}^k x_k \right)^2 E \left( (\varepsilon_k(a) - E(\varepsilon_k(a)))^2 \right) \\ &\leq \left\| C_n^k x_k \right\|^2 E \left( (\varepsilon_k(a) - E(\varepsilon_k(a)))^2 \right) \\ &\leq (1 - \tau)^{-2} \lambda^{-2} n^{-2\rho} \|x_k\|^2 \text{Var}(\varepsilon_k(a)). \end{aligned}$$

La última desigualdad es consecuencia del lema 5.3 posterior. Teniendo en cuenta que la varianza de los errores de observación está acotada,  $\text{Var}(\varepsilon_k(a)) \leq H^*$ , se

puede escribir

$$\begin{aligned} \max_{k \leq n} \text{Var} \left( n^{\frac{\rho}{2}} f_{jn}^k x_k (\varepsilon_k(a) - E(\varepsilon_k(a))) \right) &\leq \left( \frac{\lambda}{1-\tau} \right)^{-2} H^* n^{-\rho} \max_{k \leq n} \|x_k\|^2 \\ &= O(n^{-1}) \end{aligned}$$

y

$$\sum_{k=1}^n \text{Var} \left( n^{\frac{\rho}{2}} f_{jn}^k x_k (\varepsilon_k(a) - E(\varepsilon_k(a))) \right) \leq \left( \frac{\lambda}{1-\tau} \right)^{-2} n^{-\rho} \sum_{k=1}^n \|x_k\|^2 = O(1),$$

que, en conjunto, implican la convergencia en ley de la sucesión de variables reales  $n^{\frac{\rho}{2}} \sum_{k=1}^n f_{jn}^k x_k (\varepsilon_k(a) - E(\varepsilon_k(a)))$  a una distribución normal, y en consecuencia la convergencia de la variable  $m$ -dimensional  $n^{\frac{\rho}{2}} (a_n - E(a_n))$  a una normal de media cero.  $\square$

**Lema 5.3** Si  $0 < \tau < 1$  y  $\rho > 1$ , entonces

$$\sum_{i=k}^n \tau^{n-i+1} i^{-\rho} = O(n^{-\rho}).$$

DEMOSTRACION: Similar a la demostración del lema 4.5.  $\square$

## 5.5 Un segundo algoritmo de una iteración

Basándonos en la expresión (36) y con la misma idea original de la sección 4.5, se puede proponer un algoritmo donde las correcciones se hagan en moda (evitándose así las posibles aproximaciones de cuadratura propias de las correcciones en media) y se eviten, además, los cálculos que implican las inversiones de las matrices  $X^t X$ . Basta sustituir dicha matriz inversa por un tamaño de paso  $\alpha_n > 0$ , el cual deberá elegirse adecuadamente para continuar manteniendo las buenas propiedades de convergencia estocástica obtenidas en la sección anterior.

La idea de las aproximaciones estocásticas permite proponer el siguiente algoritmo iterativo:

INICIALIZACIÓN: Tómese un vector inicial  $a_1$  arbitrario.

ITERACIÓN: Siendo  $a_n$  la estimación en la etapa  $n$  del verdadero vector  $a$ , su actualización en la etapa  $n + 1$  tomará la forma

$$a_{n+1} = a_n + \alpha_n X^t \varepsilon(a_n),$$

donde el vector  $\varepsilon(a_n) = (\varepsilon_1(a_n), \dots, \varepsilon_n(a_n))^t$  tiene por componentes  $\varepsilon_i(a_n) = y_i(a_n) - a_n^t x_i$ , siendo  $y_i(a_n)$  el valor imputado para el dato censurado  $i$ -ésimo de acuerdo con el criterio modal y considerando que  $a_n$  es el verdadero valor del parámetro. Es decir,

$$\begin{aligned} y_i(a_n) &= z_i, & \text{si } i \in I^{ng} \\ &= a_n^t x_i + \gamma(-a_n^t x_i + c_{j,r}, -a_n^t x_i + c_{j,r+1}), & \text{si } i \in I_j, z_i \in (c_{j,r}, c_{j,r+1}], \end{aligned}$$

siendo

$$\gamma(-a_n^t x_i + c_{j,r}, -a_n^t x_i + c_{j,r+1}) = \text{Moda}(\nu | \nu \in (-a_n^t x_i + c_{j,r}, -a_n^t x_i + c_{j,r+1})).$$

Las condiciones que se impondrán sobre los tamaños de paso tienen cierta relación con las vistas en su momento en el caso de la corrección en media. Propongamos un lema cuya demostración está en línea con la del lema 4.6 de la sección 4.5.

**Lema 5.4** *Sea un valor  $\rho > 1$ . Tómesese una sucesión  $\{\alpha_n\}_{n \in \mathcal{N}}$  de valores positivos, de tal forma que  $\alpha_n n^\rho \rightarrow 0$ ,  $\sum \alpha_n n^\rho = \infty$ ,  $\sum \alpha_n n^{\frac{\rho}{2}} < \infty$  y  $\alpha_n n^\rho \sup \delta_n < 1$ , siendo  $\delta_n$  el máximo autovalor de la matriz  $n^{-\rho} X^t X$ . Supóngase además que  $\inf \lambda_n = \lambda > 0$ , siendo  $\lambda_n$  el mínimo autovalor de la matriz  $n^{-\rho} (X^t X)^{ng}$ . Sea la matriz de tamaño  $n$ ,  $M^* = \text{diag}(m_i^*)$  definida de forma que  $m_i^* = 1$  si  $i \in I_n^{ng}$ , y  $0 \leq m_i^* \leq 1$  si  $i \in I_n^g$ . En estas condiciones, se verifica que los autovalores de la matriz  $I - \alpha_n X^t M^* X$  son positivos y menores que  $1 - \alpha_n n^\rho \lambda < 1$ , a partir de un  $n$  en adelante.*

DEMOSTRACION: Similar a la vista en el lema 4.6 con  $\rho = 1$ .  $\square$

OBSERVACION: Teniendo en cuenta que  $\alpha_n n^\rho \rightarrow 0$ , una condición suficiente para que se cumpla  $\alpha_n n^\rho \sup \delta_n < 1$  es que  $\sup \delta_n = M < \infty$ . Si suponemos que existe  $K < +\infty$  tal que  $\max_{i \leq n} \|x_i\|^2 \leq K n^{\rho-1}$  para cada  $n \in \mathcal{N}$ , entonces

$\delta_n = n^{-\rho} \sum_{i=1}^n \|x_i\|^2 \leq n^{-\rho} n \max_{i \leq n} \|x_i\|^2 \leq K$  para cada  $n \in \mathcal{N}$ , con lo cual se cumple la condición del lema.

Se enunciará a continuación un teorema en el que se demuestra la convergencia del nuevo método de una iteración con corrección en moda, al verdadero valor  $a$  desconocido del modelo lineal (33). Las condiciones sobre las variables independientes son similares a las del teorema de convergencia del algoritmo completo y la selección de los tamaños de paso debe estar acorde con las condiciones del lema 5.4.

**Teorema 5.5** Sean  $(X^t X)^{ng}$  y  $(X^t X)^g$  matrices definidas positivas para cada  $n \in \mathcal{N}$ . Supongamos que existe  $\rho > 1$  para el cual las siguientes condiciones técnicas se cumplen

$$\inf_n \lambda_n = \lambda > 0, \\ \max_{i \leq n} \|x_i\|^2 = O(n^{\rho-1}),$$

donde  $\lambda_n$  denota el mínimo autovalor de la matriz  $n^{-\rho}(X^t X)^{ng}$ . Selecciónese en cada etapa un tamaño de paso  $\alpha_n$  de forma que  $\alpha_n n^\rho \rightarrow 0$ ,  $\sum \alpha_n n^\rho = \infty$  y  $\sum \alpha_n n^{\frac{\rho}{2}} < \infty$ . En estas condiciones, el método de una iteración con correcciones en moda converge en  $L_1$ , es decir,  $a_n \xrightarrow{L_1} a$  y, en consecuencia,  $a_n$  es un estimador consistente de  $a$ .

DEMOSTRACION: Recordemos que las correcciones en moda implicaban, para cada  $j, r \in \mathcal{N}$ , las condiciones (34), cualquiera que fueran los errores  $\nu$  unimodales en cero y simétricos. En consecuencia, operando como en resultados anteriores, se puede escribir lo siguiente

$$a_{n+1} = a_n - \alpha_n X^t M^* X (a_n - a) + \alpha_n X^t \varepsilon(a),$$

donde  $M^* = \text{diag}(m_i^*)$  verifica que  $m_i^* = 1$  si  $i \in I_n^{ng}$ , y  $0 \leq m_i^* \leq 1$  si  $i \in I_n^g$ .

Esta última expresión conviene escribirla en la siguiente forma

$$a_{n+1} - a = [I - \alpha_n X^t M^* X] (a_n - a) + \alpha_n X^t \varepsilon(a).$$

Tomando normas euclideas se tiene

$$\|a_{n+1} - a\| \leq \mu_n \|a_n - a\| + \alpha_n \|X^t \varepsilon(a)\|,$$

siendo  $\mu_n$  el máximo autovalor de la matriz  $I - \alpha_n X^t M^* X$ . Según se probó en el lema 5.4, esta matriz es definida positiva y, además,  $\mu_n \leq 1 - \alpha_n n^\rho \lambda < 1$ , para  $n$  grande. Se supone, sin pérdida de generalidad, que esa acotación por 1 se cumple para cada  $n \geq 1$ .

La acotación, ya vista en la sección 5.3, de los dos primeros momentos de los errores de observación  $\varepsilon_i(a)$  se sigue cumpliendo aquí. Basta razonar como allí se hizo. Es decir, existen constantes  $H, H^* < \infty$ , tales que  $|E(\varepsilon_i(a))| \leq H$  y  $E(\varepsilon_i(a)^2) \leq H^*$ , cualquiera que sean los valores  $x_i \in \mathcal{R}^m$ . Llamando  $c_n = \alpha_n \|X^t \varepsilon(a)\|$ , a partir de la independencia y las acotaciones citadas de los momentos de los errores  $\varepsilon_i(a)$ , se llega a

$$\begin{aligned} E(c_n^2) &= \alpha_n^2 E(\|X^t \varepsilon(a)\|^2) = \alpha_n^2 E\left(\sum_{i,j=1}^n \varepsilon_i(a) \varepsilon_j(a) x_i^t x_j\right) \\ &\leq \alpha_n^2 \max\{H^2, H^*\} \sum_{i,j=1}^n \|x_i\| \|x_j\| \\ &= \alpha_n^2 \max\{H^2, H^*\} \left[\sum_{i=1}^n \|x_i\|\right]^2. \end{aligned}$$

Teniendo en cuenta que la condición (35) sigue siendo cierta como consecuencia directa de las hipótesis establecidas sobre las  $x_i$ , resulta que existe  $K < \infty$  tal que

$$E(c_n^2) \leq \alpha_n^2 \max\{H, H^*\} K n^\rho \xrightarrow{n \rightarrow \infty} 0,$$

puesto que  $\alpha_n n^\rho \rightarrow 0$  por hipótesis.

En definitiva,

$$E\|a_{n+1} - a\| \leq \mu_n E\|a_n - a\| + E(c_n).$$

Puesto que  $\sum_n \alpha_n n^\rho = \infty$  se tiene que  $\prod_n (1 - \alpha_n n^\rho \lambda) \mu_n = 0$ , y así también es  $\prod_n \mu_n = 0$ . Así pues,

$$E\|a_{n+1} - a\| \leq \prod_{i=1}^n \mu_i E\|a_1 - a\| + \sum_{i=1}^n \tau_n^i E(c_i), \quad (39)$$

siendo  $\tau_n^i = \prod_{j=i+1}^n \mu_j$ . Finalmente, teniendo en cuenta que  $1 \geq \tau_n^i \xrightarrow{n \rightarrow \infty} 0$ , para cada  $i$  fijo, y considerando que

$$\sum_{n=1}^{\infty} E(c_n) \leq \sum_{n=1}^{\infty} E(c_n^2)^{\frac{1}{2}} \leq K \sum_{n=1}^{\infty} \alpha_n n^{\frac{p}{2}} < \infty,$$

se puede concluir, a través del lema 4.8 del capítulo precedente, que

$$E \|a_n - a\| \xrightarrow{n \rightarrow \infty} 0,$$

puesto que convergen a cero los dos sumandos a la derecha de la expresión (39), concluyéndose la demostración.  $\square$

## 5.6 Un tercer algoritmo de una iteración

En esta sección desarrollaremos el último algoritmo inspirado en el desarrollo de la sección 4.6, aunque usando ahora correcciones en moda. Las razones que inspiran el establecimiento de este nuevo algoritmo son similares a las expresadas en aquel momento y se omiten aquí. El algoritmo estrictamente es el siguiente:

**INICIALIZACIÓN:** Tómese un vector inicial  $a_1$  arbitrario.

**ITERACIÓN:** Sea  $a_n$  la estimación en la etapa  $n$  del verdadero vector  $a$ . La nueva estimación para la etapa  $n + 1$  será entonces

$$a_{n+1} = a_n + \alpha_n x_n \varepsilon_n(a_n),$$

donde  $\varepsilon_n(a_n) = y_n(a_n) - a_n^t x_n$ ,  $\alpha_n > 0$ ,  $a_n \in R^m$  y  $x_n \in R^m$ , con

$$\begin{aligned} y_n(a_n) &= z_n, & \text{si } n \in I^{ng} \\ &= a_n^t x_n + \gamma(-a_n^t x_n + c_{j,r}, -a_n^t x_n + c_{j,r+1}), & \text{si } n \in I_j, z_n \in (c_{j,r}, c_{j,r+1}]. \end{aligned}$$

Como es natural  $\gamma$  es aquí la función de corrección en moda

$$\gamma(-a_n^t x_n + c_{j,r}, -a_n^t x_n + c_{j,r+1}) = \text{Moda}(\nu | \nu \in (-a_n^t x_n + c_{j,r}, -a_n^t x_n + c_{j,r+1}]).$$

Obsérvense las implicaciones simplificadoras de este nuevo algoritmo. Cada dato se utiliza una única vez en el proceso. Es decir, cuando el dato es censurado

solamente se imputa su valor faltante una vez, a diferencia de cómo ocurría en los algoritmos anteriores. En estos últimos, de las secciones 5.3 y 5.5, según se actualizaba la estimación del parámetro en cada etapa, era preciso imputar repetidamente todos los datos censurados. De esta forma, cada iteración obligaba al cálculo del vector completo de datos censurados que a su vez se utilizaba para la actualización del parámetro. En resumen, la eventual (?) menor eficacia de este método puede verse compensada con el menor esfuerzo de cálculo, al no requerir el almacenamiento de toda la información pasada, sino solamente el almacenamiento de la última estimación, que es en definitiva la idea de las aproximaciones estocásticas. Ésta consiste, como se sabe, en mejorar la estimación actual mediante la información que proporciona el último valor muestral.

Nuestro objetivo será buscar una sucesión de pasos  $\{\alpha_n\} \subset R^+$  adecuada para garantizar convergencia del tipo  $a_n \xrightarrow{L_1} a$ .

Puesto que las correcciones se hacen en moda, recordemos que se cumple que  $\delta_{jr}(\beta)$  y  $\beta - \delta_{jr}(\beta)$  son no decrecientes, con lo cual si el dato  $n$ -ésimo es censurado,  $z_n \in (c_{j,r}, c_{j,r+1}]$  y la estimación actual es  $a_n$ , se puede encontrar un valor  $m_n^* \in [0, 1]$ , dependiente de  $a_n$  y  $a$ , tal que

$$\begin{aligned} & \gamma(-a_n^t x_n + c_{j,r}, -a_n^t x_n + c_{j,r+1}) - \gamma(-a^t x_n + c_{j,r}, -a^t x_n + c_{j,r+1}) \\ &= \delta_{jr}(-a_n^t x_n) - \delta_{jr}(-a^t x_n) = -m_n^* (a_n^t x_n - a^t x_n) = -m_n^* x_n^t (a_n - a). \end{aligned} \quad (40)$$

El teorema que asegura la convergencia de  $a_n$  es el siguiente.

**Teorema 5.6** *Tómense  $\alpha_n > 0$  tales que  $\alpha_n \|x_n\|^2 < 1$ , para cada  $n \in N$ , cumpliéndose, además, que*

$$\sum_{n=1}^{\infty} \alpha_n \|x_n\| < \infty.$$

*Llamemos  $\eta_n^i$  al máximo autovalor de la matriz definida positiva*

$$\prod_{j=i}^n [I - \alpha_j m_j^* x_j x_j^t],$$



donde  $m_j^*$  se ha definido en (40). Si para cada  $i \in \mathcal{N}$  existe una sucesión determinista  $\{\tau_n^i\}$ , convergente a cero cuando  $n \rightarrow \infty$ , tal que  $\eta_n^i \leq \tau_n^i$  c.s., entonces se cumple que  $E \|a_n - a\| \rightarrow 0$ , (es decir,  $a_n$  converge en  $L_1$ , y en consecuencia converge en probabilidad).

DEMOSTRACION: La existencia de las sucesiones  $\{\tau_n^i\}$  está ligada directamente a la elección de la sucesión de tamaños de paso  $\{\alpha_n\}$ , por lo que esta condición debe ser coherente con las anteriormente impuestas sobre los tamaños de paso. Obsérvese que  $\eta_n^i \leq 1$ . Por tanto, si existe una sucesión  $\tau_n^i$  en las condiciones del enunciado, puede suponerse uniformemente acotada por uno.

La demostración del resultado parte de escribir, tras considerar (40),

$$\varepsilon_n(a_n) = -m_n^* x_n^t (a_n - a) + \varepsilon_n(a),$$

donde  $m_n^* = 1$  cuando el dato  $n$ -ésimo es no agrupado, y  $0 \leq m_n^* \leq 1$  cuando el dato  $n$ -ésimo es agrupado. Recuérdese que en el último caso  $m_n^*$  depende de  $a_n$  y del verdadero valor del parámetro desconocido  $a$ .

En consecuencia,

$$a_{n+1} = a_n - \alpha_n m_n^* x_n x_n^t (a_n - a) + \alpha_n x_n \varepsilon_n(a),$$

resultando, después de sumar  $a$  en cada uno de los dos miembros, que

$$a_{n+1} - a = [I - \alpha_n m_n^* x_n x_n^t] (a_n - a) + \alpha_n x_n \varepsilon_n(a). \quad (41)$$

El valor seleccionado de  $\alpha_n > 0$  hace que la matriz  $I - \alpha_n m_n^* x_n x_n^t$  sea definida positiva, puesto que su mínimo autovalor corresponde a  $1 - \alpha_n \|x_n\|^2 > 0$ .

Iterando  $n$  veces la expresión (41) resulta,

$$a_{n+1} - a = \prod_{i=1}^n [I - \alpha_i m_i^* x_i x_i^t] (a_1 - a) + \sum_{i=1}^n \prod_{j=i}^n [I - \alpha_j m_j^* x_j x_j^t] \alpha_i x_i \varepsilon_i(a).$$

Así pues, tomando normas euclídeas en este momento se llega a

$$\|a_{n+1} - a\| \leq \eta_n^1 \|a_1 - a\| + \sum_{i=1}^n \eta_n^i \alpha_i \|x_i\| |\varepsilon_i(a)|,$$

donde  $\eta_n^i$  se define en el enunciado como el máximo autovalor de la matriz definida

positiva

$$\prod_{j=i}^n [I - \alpha_j m_j^* x_j x_j^t].$$

Por las hipótesis de acotación sobre  $\eta_n^i$  se puede poner

$$\|a_{n+1} - a\| \leq \tau_n^1 \|a_1 - a\| + \sum_{i=1}^n \tau_n^i \alpha_i \|x_i\| |\varepsilon_i(a)| \quad c.s.$$

Llamando  $c_i = \alpha_i \|x_i\| |\varepsilon_i(a)|$  y tomando esperanza en la última expresión se llega a

$$E \|a_{n+1} - a\| \leq \tau_n^1 E \|a_1 - a\| + \sum_{i=1}^n \tau_n^i E(c_i).$$

Como  $E(\varepsilon_i^2(a)) \leq H^*$ , también  $E(|\varepsilon_i(a)|) \leq \sqrt{H^*}$ , para todo  $x_i \in R^m$ . Dado que  $E(c_i) \leq \sqrt{H^*} \alpha_i \|x_i\|$ , una de las hipótesis del teorema conduce a  $\sum_{i=1}^{\infty} E(c_i) < \infty$ . Sólo queda aplicar el lema 4.8 del capítulo anterior, teniendo en cuenta que  $\tau_n^i \leq 1$  y  $\tau_n^i \xrightarrow{n \rightarrow \infty} 0$ , para llegar a la conclusión de que  $\sum_{i=1}^n \tau_n^i E(c_i) \xrightarrow{n \rightarrow \infty} 0$ . Se sigue, como se deseaba probar, que  $a_n \xrightarrow{L_1} a$ , siendo, en consecuencia,  $a_n$  es un estimador consistente de  $a$ .  $\square$

Las observaciones hechas en la sección 4.6 en este punto con respecto a la posibilidad de mejorar las hipótesis de teorema son válidas también ahora, y no se repiten.

## 5.7 Sobre las hipótesis de no acotación

Las hipótesis del teorema principal de la sección 5.3, relativas al algoritmo MdD de una iteración, implican la no acotación de las variables independientes, ya que al ser  $\rho > 1$  resulta  $\max_{i \leq n} \|x_i\|^2 = O(n^{\rho-1}) \rightarrow \infty$ . Por el contrario, las hipótesis del capítulo 4 hacían referencia a variables  $x_i$  acotadas, que suele ser una condición más común en la práctica. Sin embargo, un problema con variables acotadas se puede transformar fácilmente en uno que no lo sea, pudiéndose aplicar

los resultados con correcciones en moda al nuevo problema. En efecto, sea

$$z_i = a^t x_i + \nu_i,$$

con  $\nu_i$  variables aleatorias independientes e idénticamente distribuidas y  $x_i \in \mathcal{R}^m$ , tal que  $\|x_i\| \leq K$  para cada  $i \in \mathcal{N}$ . Es suficiente poner

$$z_i + c_i = a^t x_i + c_i + \nu_i$$

donde  $\{c_i\} \subset \mathcal{R}$  es una sucesión no acotada, y escribir el nuevo modelo como

$$z_i^* = b^t x_i^* + \nu_i$$

donde  $z_i^* = z_i + c_i$ ,  $x_i^* = (x_i, c_i)^t \in \mathcal{R}^{m+1}$  y el parámetro a estimar es  $b = (a, 1)^t \in \mathcal{R}^{m+1}$ . En esta situación, hay que dejar que el algoritmo estime  $b$  sin considerar la condición de contorno  $b_{n+1} = 1$ . En todo caso, la elección de  $c_i$  puede necesitar algún esfuerzo en vistas a que se cumplan de forma adecuada las dos condiciones del enunciado del teorema 5.1.

## 5.8 Aleatorización de los intervalos de clasificación

En esta sección se estudia otra manera de salvar el problema de la no acotación de las variables independientes, que consiste en replantear el modelo inicial de forma que se aleatoricen las clases de agrupación de los datos censurados. Esto permitirá modificar el teorema 5.1 e imponer condiciones de acotación manteniendo la misma conclusión. La técnica será extensible al teorema 5.5.

Como ya se comentó en la sección anterior, los resultados con correcciones en moda no son aplicables cuando las variables  $x_i$  son acotadas. La razón fundamental subyacente es que los errores muestrales  $\varepsilon_i(a)$ , para los índices censurados, no es seguro que tengan media cero, excepto cuando la corrección es en media. Ahora bien, incluso no siendo cero esta media, aún es posible aleatorizar sobre los intervalos de clasificación de los datos censurados de forma que los valores positivos y negativos de las citadas medias (con intervalos de clasificación

fijos) se compensen. Esta es la idea que se propone a lo largo de esta sección. Las hipótesis de aleatorización de los intervalos dependiendo del problema serán comprobables con mayor o menor facilidad.

De nuevo, considérese el modelo lineal

$$z_i = a^t x_i + \nu_i, \quad i = 1, \dots, n$$

donde  $a, x_i \in \mathcal{R}^m$  y  $\nu_i$  son variables aleatorias independientes e idénticamente distribuidas, simétricas, unimodales en cero y de varianza uno. Supongamos que el dato  $i$ -ésimo  $z_i$  es no censurado con probabilidad  $\pi_0$  y es censurado con probabilidad  $1 - \pi_0$ . De esta forma, el conjunto de índices se puede dividir en  $I^{ng}$  e  $I^g$  conteniendo los índices no censurados y censurados, respectivamente. Cuando  $i \in I^g$ , no se conoce, como en anteriores ocasiones, el valor exacto de  $z_i$ , sino un intervalo de agrupación  $(l_i, u_i)$ , donde  $-\infty \leq l_i < u_i \leq \infty$ . Se supondrá que los límites de censura  $(l_i, u_i)$  no son fijos, como ha sido usual hasta ahora, sino aleatorios. El algoritmo MdD de una iteración con correcciones en moda de la sección 5.3 se puede escribir de la siguiente forma:

INICIALIZACIÓN: Tómesese un vector inicial  $a_1$  arbitrario.

ITERACIÓN: Si  $a_n$  es la estimación en la etapa  $n$  del verdadero vector  $a$ , su actualización en la etapa  $n + 1$  se lleva a cabo mediante

$$a_{n+1} = (X^t X)^{-1} X^t y(a_n).$$

El vector  $y(a_n) = (y_1(a_n), \dots, y_n(a_n))^t$  tiene por componentes

$$\begin{aligned} y_i(a_n) &= z_i, & \text{si } i \in I^{ng} \\ &= a_n^t x_i + \gamma(-a_n^t x_i + l_i, -a_n^t x_i + u_i), & \text{si } i \in I^g, z_i \in (l_i, u_i), \end{aligned}$$

donde

$$\gamma(-a_n^t x_i + l_i, -a_n^t x_i + u_i) = \text{Moda}(\nu | \nu \in (-a_n^t x_i + l_i, -a_n^t x_i + u_i)).$$

A continuación se enunciará un teorema de convergencia para el anterior

algoritmo. Las condiciones de acotación sobre las variables independientes  $x_i$  coinciden con las asumidas en los teoremas vistos en el caso de correcciones en media. El aspecto esencial en este caso es la aleatorización de las clases de censura, la cual permite, bajo ciertas condiciones de simetría, obtener errores de observación de esperanza cero, con las consiguientes ventajas técnicas ya apuntadas con antelación.

Sean  $L_i, U_i$  las variables aleatorias representando los extremos inferior y superior del intervalo de clasificación de dato  $i$ -ésimo. Se asumirá, como es lógico, que  $L_i < U_i$  c.s., para cada dato censurado.

**Teorema 5.7** Sean  $(X^t X)^{ng}$  y  $(X^t X)^g$  matrices definidas positivas para cada  $n \in \mathcal{N}$ . Supongamos que las siguientes condiciones se cumplen

$$\inf_n \lambda_n = \lambda > 0,$$

$$\|x_i\| \leq K < +\infty, \quad \text{para cada } i \in \mathcal{N},$$

donde  $\lambda_n$  denota el mínimo autovalor de la matriz  $n^{-1}(X^t X)^{ng}$ . Supongamos

que, para cada  $i \in \mathcal{N}$ , se cumple que

$$(L_i, U_i) - a^t x_i \stackrel{d}{=} a^t x_i - (U_i, L_i). \quad (42)$$

y

$$E(L_i^2) \leq C < \infty. \quad (43)$$

En estas condiciones,  $a_n \xrightarrow{L_1} a$  (en consecuencia,  $a_n$  es un estimador consistente de  $a$ ).

DEMOSTRACION: Fijados  $l_i < u_i$  se sabe que la función en  $\beta$

$$\delta_{l_i, u_i}(\beta) = \gamma(\beta + l_i, \beta + u_i)$$

verifica que es no decreciente, al igual que  $\beta - \delta_{l_i, u_i}(\beta)$ . La expresión (37) continúa cumpliéndose, de donde

$$a_{n+1} - a = (X^t X)^{-1} X^t (I - M^*) X (a_n - a) + (X^t X)^{-1} X^t \varepsilon(a),$$

siendo  $M^* = \text{diag}(m_i^*)$ , con  $m_i^* = 1$  si  $i \in I^{ng}$  y  $0 \leq m_i^* \leq 1$  si  $i \in I^g$ .

Se probará ahora que  $E(\varepsilon_i(a)) = 0$ , al contrario de lo que ocurría en el caso

de correcciones modales con intervalos de censura fijos. En efecto,

$$\varepsilon_i(a) = \delta_{l_i, u_i}(-a^t x_i), \quad \text{si } i \in I^g \text{ y } z_i \in (l_i, u_i).$$

En consecuencia,

$$\begin{aligned} E(\varepsilon_i(a) | i \in I^g) &= E(\delta_{L_i, U_i}(-a^t x_i)) \\ &= \int_{\{l_i > a^t x_i\}} (-a^t x_i + l_i) dF_{L_i}(l_i) \\ &\quad + \int_{\{u_i < a^t x_i\}} (-a^t x_i + u_i) dF_{U_i}(u_i). \end{aligned}$$

Teniendo en cuenta la condición de simetría (42), resulta que

$$\begin{aligned} \int_{\{x > a^t x_i\}} (-a^t x_i + x) dF_{L_i}(x) &= \int_{\{x > a^t x_i\}} (-a^t x_i + x) dF_{2a^t x_i - U_i}(x) \\ &= \int_{\{2a^t x_i - u > a^t x_i\}} (a^t x_i - u) dF_{U_i}(u) \\ &= \int_{\{u < a^t x_i\}} (a^t x_i - u) dF_{U_i}(u). \end{aligned}$$

Se sigue que  $E(\varepsilon_i(a) | i \in I^g) = 0$ , concluyéndose que

$$\begin{aligned} E(\varepsilon_i(a)) &= \pi_0 E(\varepsilon_i(a) | i \in I^{ng}) + (1 - \pi_0) E(\varepsilon_i(a) | i \in I^g) \\ &= \pi_0 E(\nu_i) = 0, \end{aligned}$$

puesto que  $E(\nu_i) = 0$  por la simetría de los errores.

También se puede probar que existe  $H^* < \infty$ , tal que  $E(\varepsilon_i(a)^2) \leq H^*$ , para cada  $x_i$ . En efecto, obsérvese que, por la acotación de las  $x_i$  y de  $E(L_i^2)$ , se verifica, para algún valor  $H$  real, que

$$E\left((L_i - a^t x_i)^2\right) \leq H < \infty.$$

Si el dato  $i$ -ésimo es censurado, resulta

$$\begin{aligned} E(\varepsilon_i(a)^2 | i \in I^g) &= E(\delta_{L_i, U_i}^2(-a^t x_i)) \\ &= \int_{\{l_i > a^t x_i\}} (-a^t x_i + l_i)^2 dF_{L_i}(l_i) \end{aligned}$$

$$\begin{aligned}
 & + \int_{\{u_i < a^t x_i\}} (-a^t x_i + u_i)^2 dF_{U_i}(u_i) \\
 = & 2 \int_{\{l_i > a^t x_i\}} (-a^t x_i + l_i)^2 dF_{L_i U_i}(l_i, u_i) \\
 = & 2 \int_{\{z_i < 0\}} z_i^2 dF_{L_i - a^t x_i}(z_i) \\
 \leq & 2 \int_{\mathcal{R}} z_i^2 dF_{L_i - a^t x_i}(z_i) \leq 2H.
 \end{aligned}$$

En consecuencia, el segundo momento de los errores  $\varepsilon_i(a)$  se puede escribir como

$$\begin{aligned}
 E(\varepsilon_i(a)^2) & = \pi_0 E(\varepsilon_i(a)^2 | i \in I^{ng}) + (1 - \pi_0) E(\varepsilon_i(a)^2 | i \in I^g) \\
 & \leq \pi_0 E(\nu_i^2) + (1 - \pi_0) 2H = \pi_0 + (1 - \pi_0) 2H,
 \end{aligned}$$

habiéndose probado su acotación uniforme por  $H^* = \pi_0 + (1 - \pi_0) 2H$ , cualquiera que sea la variable independiente  $x_i$ .

La existencia de la matriz  $M^*$ , la esperanza cero y la varianza acotada de los errores de observación permite concluir el resultado, siguiendo los mismos pasos empleados en la demostración del teorema 4.1 del capítulo 4.  $\square$

OBSERVACION: La expresión (42) es una condición de simetría en la aleatorización de los intervalos de clasificación. En concreto, afirma que si el dato  $i$ -ésimo es censurado, la distribución de los intervalos de censura es simétrica respecto al valor medio del dato  $a^t x_i$ .

## 5.9 Otra forma de aleatorizar los intervalos de clasificación

Lo que se aleatoriza en este caso es el conjunto de posibles particiones de agrupación de los datos censurados. Sea  $\mathcal{P}$  el conjunto de posibles particiones de  $\mathcal{R}$  en un conjunto finito de intervalos. Se puede representar  $\mathcal{P} = \{(c_1, \dots, c_r) | r \in \mathcal{N}, -\infty < c_1 < \dots < c_r < \infty\}$ , donde  $(c_1, \dots, c_r)$  representa a la

partición

$$(-\infty, c_1], (c_1, c_2], \dots, (c_{r-1}, c_r], (c_r, \infty).$$

Sea  $\mu$  una medida de probabilidad sobre  $(\mathcal{P}_{\mathcal{N}}, \sigma(\mathcal{P}_{\mathcal{N}}))$ , siendo  $\sigma(\mathcal{P}_{\mathcal{N}})$  la  $\sigma$ -álgebra de Borel asociada a  $\mathcal{P}_{\mathcal{N}}$ . La manera de obtener un intervalo de clasificación, si el dato  $i$ -ésimo es censurado, es seleccionando una partición de  $\mathcal{P}_{\mathcal{N}}$  mediante la medida de probabilidad  $\mu$ , observando después en qué intervalo de esa partición cae el valor  $z_i$ . De esta forma, fijado el elemento  $(c_1, \dots, c_r) \in \mathcal{P}_{\mathcal{N}}$ , resulta que para  $j = 0, \dots, r - 1$

$$\Pr(z_i \in (c_j, c_{j+1})) = \Pr(\nu_i \in (-a^t x_i + c_j, -a^t x_i + c_{j+1})).$$

Obsérvese que el modelo original de censura desarrollado a lo largo de la memoria coincide con un caso particular de éste. Para verlo, sea  $1 - \pi_0$  la probabilidad de que el dato  $i$ -ésimo sea censurado. Sean  $P_1, \dots, P_s$  un conjunto finito de  $s$  particiones de la recta real. Basta tomar  $\mu$  una medida de probabilidad discreta sobre  $P_1, \dots, P_s$  de forma que

$$\mu(P_i) = \frac{\pi_i}{1 - \pi_0}.$$

Tanto el algoritmo MD con correcciones en media como el Mdd con correcciones en moda, de una iteración, se pueden particularizar a este tipo de aleatorizaciones de la siguiente forma:

INICIALIZACIÓN: Tómesese un vector inicial  $a_1$  arbitrario.

ITERACIÓN: Si  $a_n$  es la estimación en la etapa  $n$  del verdadero vector  $a$ , la nueva estimación para la etapa  $n + 1$  será

$$a_{n+1} = (X^t X)^{-1} X^t y(a_n),$$

donde el vector  $y(a_n) = (y_1(a_n), \dots, y_n(a_n))^t$  tiene por componentes

$$\begin{aligned} y_i(a_n) &= z_i, & \text{si } i \in I^{ng} \\ &= a_n^t x_i + \gamma(-a_n^t x_i + c_{ij}, -a_n^t x_i + c_{ij+1}), & \text{si } i \in I^g, z_i \in (c_{ij}, c_{ij+1}). \end{aligned}$$



La función  $\gamma$  puede establecerse de la forma habitual, tanto en términos de la media como de la moda, habiéndose seleccionado  $(c_{i1}, \dots, c_{ir})$  de acuerdo a la medida de probabilidad  $\mu$ .

Los resultados de convergencia vistos en las secciones 4.3 y 5.3 siguen siendo válidos aquí.

**Teorema 5.8** *Cualquiera que sea la medida de probabilidad  $\mu$ , los teoremas 4.1 y 5.1 se verifican para el algoritmo anterior.*

DEMOSTRACION: Sólo es necesario tener en cuenta el comportamiento de los errores de observación en lo relativo a su media y su varianza. En efecto,

$$\begin{aligned} & E(\varepsilon_i(a) | (c_{i1}, \dots, c_{ir}), i \in I^g) \\ &= \sum_{h=1}^r \Pr(\nu_i \in -a^t x_i + (c_{i,h-1}, c_{i,h}]) \gamma(-a^t x_i + c_{i,h-1}, -a^t x_i + c_{i,h}). \end{aligned}$$

En el caso de la media  $E(\varepsilon_i(a) | (c_{i1}, \dots, c_{ir}), i \in I^g) = 0$  y en el caso de la moda  $E(\varepsilon_i(a) | (c_{i1}, \dots, c_{ir}), i \in I^g) \leq H$ , según se probó en los respectivos teoremas.

Así pues,

$$\begin{aligned} E(\varepsilon_i(a)) &= \pi_0 E(\nu_i) + (1 - \pi_0) \int_{\mathcal{P}_N} E(\varepsilon_i(a) | (c_{i1}, \dots, c_{ir}), i \in I^g) d\mu \\ &= (1 - \pi_0) \int_{\mathcal{P}_N} E(\varepsilon_i(a) | (c_{i1}, \dots, c_{ir}), i \in I^g) d\mu. \end{aligned}$$

Se sigue que, en el caso de la media,  $E(\varepsilon_i(a)) = 0$ , mientras que, ante correcciones en moda,  $E(\varepsilon_i(a)) \leq (1 - \pi_0) H$ .

De la misma forma se demuestra que, en ambos casos,  $E((\varepsilon_i(a))^2) \leq H^*$ .

Estas dos acotaciones permiten seguir las demostraciones respectivas del teorema 4.1 y 5.1 para concluir el resultado.  $\square$

## 5.10 Simulaciones

En esta sección se muestran los resultados obtenidos por simulación tras aplicar alguno de los algoritmos de una iteración con correcciones en moda, expuestos en

este capítulo. Como en anteriores ocasiones, se ha partido del modelo lineal

$$z_i = a^t x_i + \nu_i,$$

con  $a = \begin{pmatrix} -4 \\ 10 \end{pmatrix}$ . Los errores  $\nu_i$  se han simulado asumiéndose que son variables aleatorias independientes idénticamente distribuidas según normales estándar  $N(0, 1)$ . Los valores de la variable independiente  $x_i$  se han elegido no acotados, cumpliéndose que

$$\max_{i \leq n} \|x_i\|^2 = O(\sqrt{n}). \quad (44)$$

Finalmente, el 80% de los datos se han censurado sobre los intervalos de censura

$$(-\infty, 0], (0, 20], (20, 50], (50, 100], (100, 200], (200, \infty).$$

Con las características citadas, se ha obtenido un número  $N = 1000$  de datos, siguiendo el modelo de censura anterior. Algunos de los valores obtenidos aparecen de forma ilustrativa en la tabla 5. La última columna refleja el indicador de censura. Si  $p_i = 0$  el dato  $i$ -ésimo es observado, mientras que si  $p_i = 1$  el dato es censurado, indicando las columnas  $l_i, u_i$  su correspondiente intervalo de censura. Puede observarse cómo cada componente de  $x_i$  crece de acuerdo con la expresión en notación Landau (44).

Las figuras 8 muestran las trayectorias del algoritmo modal de una iteración, presentado en la sección 5.3, para los puntos de arranque  $(100, 100)$ ,  $(-100, -100)$  y  $(0, 0)$ . La figura 8.a refleja la secuencia obtenida para la primera componente de  $a_n$ , mientras que la figura 8.b hace lo propio para la segunda.

En todos los algoritmos MdD, tanto de una como de dos iteraciones, se ha demostrado, cuanto menos, la convergencia en  $L_1$  al verdadero valor del parámetro de la sucesión generada por el algoritmo. Exactamente lo mismo sucede con los algoritmos MD, de dos o una iteración, en donde se emplean correcciones en media. Las figuras 9 muestran, con fines comparativos entre las correcciones en media y en moda, las trayectorias (empíricas) de la sucesión  $E \|a_n - a\| \rightarrow 0$ , para

las estimaciones proporcionadas por los algoritmos MD y MdD de una iteración. La evaluación empírica de las anteriores diferencias esperadas en norma en cada etapa se han obtenido a partir de  $M = 5000$  reiteraciones de la secuencia  $\{a_n\}$ , mediante

$$E \|\widehat{a_n} - a\| = \frac{1}{5000} \sum_{m=1}^{5000} \|a_n(m) - a\|,$$

donde  $a_n(m)$  denota el valor de  $a_n$  en la reiteración  $m$ -ésima. La figura 9.a superpone las sucesiones  $E \|a_n - a\|$  para los dos algoritmos citados hasta la iteración 50, tomando como punto de arranque el (20, 20) en el algoritmo de una iteración. La figura 9.b, por su parte, es similar dentro del rango de iteraciones de 1 a 500 (de facto, ambas coinciden debido a la reducción de escala).

El algoritmo de una iteración presentado en la sección 5.3 converge en ley a la normal (teorema 5.2). Ésta se puede contrastar mediante los test usuales de bondad de ajuste ( $\chi^2$ , Kolmogorov-Smirnov,...). Con este fin, se han obtenido para la etapa  $N = 500$ , un número  $M = 600$  de valores reiterados de  $a_{500}$ . Con esta muestra el contraste de la  $\chi^2$  obtiene como p-valores 0.9505 y 0.9316. El teorema 5.2, antes citado, establece que

$$500^{\frac{1}{2}} (a_{500} - a) \approx N(0, \Sigma).$$

Los datos  $x_i$ , mostrados parcialmente en la tabla 5, se han obtenido, según (44), para  $\rho = \frac{3}{2}$ . Con este valor, la matriz de varianzas-covarianzas empírica  $\widehat{\Sigma}$  obtenida con la muestra citada de 600 valores de  $a_{500}$  coincidió con

$$\widehat{\Sigma} = \begin{pmatrix} 509 & -241 \\ -241 & 110 \end{pmatrix}.$$

#### CONCLUSIONES

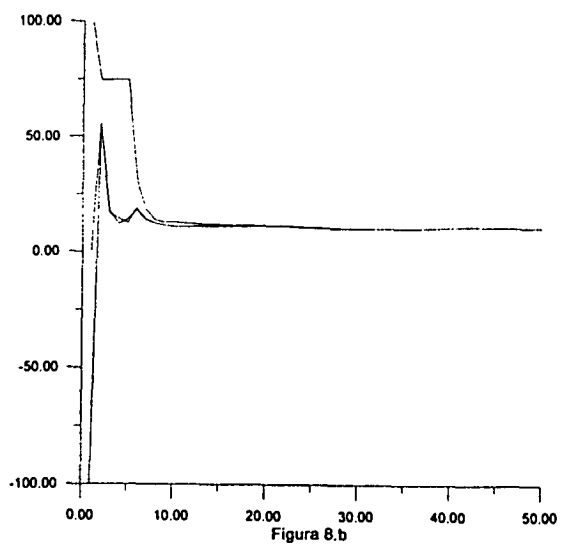
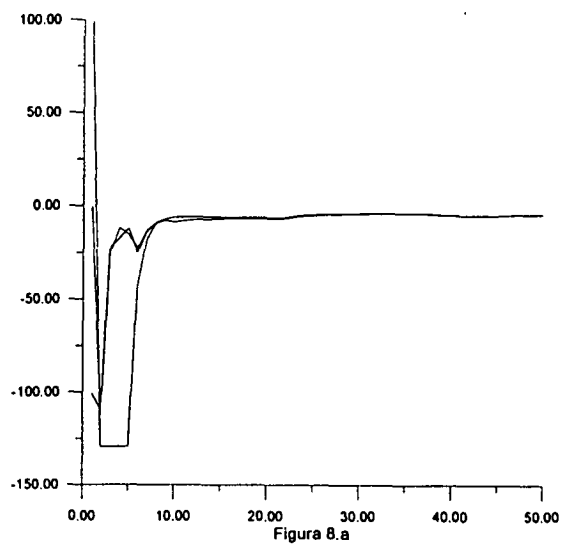
1. Como ocurría con las correcciones en media, los algoritmos modales de una iteración no garantizan la unicidad de las trayectorias con respecto al punto de arranque (a diferencia de los de iteración doble). Las figuras 8 indican que, pese

a lo anterior, las trayectorias correspondientes a puntos de arranque distintos tienden a confundirse a medida que  $n$  crece. De hecho, en las simulaciones realizadas todas las trayectorias a partir de  $n = 46$  difieren en menos de  $10^{-3}$ . En este sentido, la simplificación que suponen los algoritmos de una iteración apenas limita de facto la propiedad de trayectorias únicas.

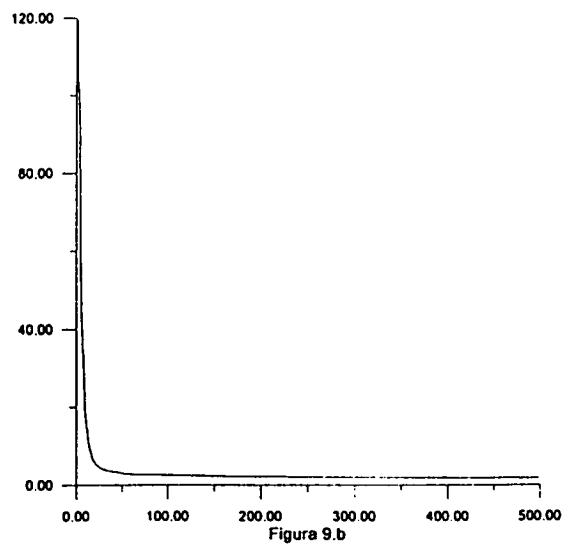
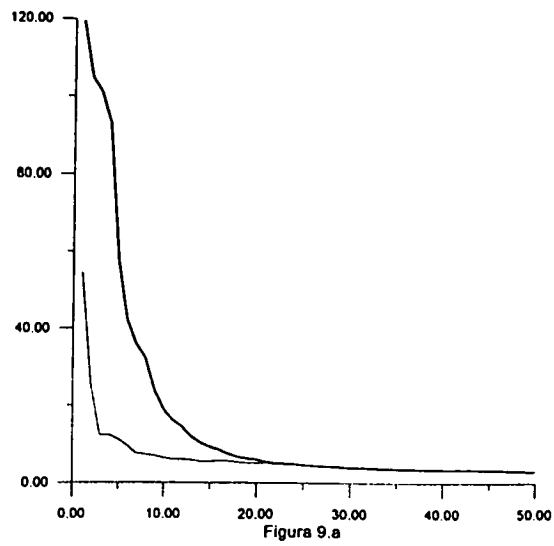
2. Los algoritmos en moda suponen, como ya fue indicado en el capítulo 3, una gran reducción de tiempo de cálculo frente a los de correcciones en media, puesto que los primeros evitan cálculos de cuadratura propios de los últimos. En este sentido, los tiempos empleados para llevar a cabo las simulaciones precisadas para obtener la figura 9 han puesto de manifiesto una reducción de tiempos 2200 veces a favor de los algoritmos modales. Pese a ello, puede observarse que las tasas de convergencia en  $L_1$  para ambos son similares. Obsérvese que, prácticamente, la figura 9.b identifica las trayectorias obtenidas para los algoritmos en media y moda. La figura 9.a, por su parte, indica que la pérdida en precisión derivada de la simplificación citada sólo es visible en las, aproximadamente, 20 primeras iteraciones.
3. En resumen, al igual que en el capítulo 3, todo apunta a que los algoritmos modales suponen una enorme ventaja frente a los que utilizan correcciones en media, cuando los errores son unimodales y simétricos, al menos.

	$x_{i1}$	$x_{i2}$	$z_i$	$l_i$	$u_i$	$p_i$
1	1.0	2.00	15.22	0	20	1
2	1.2	2.74	23.18	20	50	1
3	1.7	3.29	26.18	20	50	1
4	1.8	3.82	32.04	20	50	1
5	1.5	2.99	23.26	20	50	1
6	1.6	3.60	29.99	20	50	1
7	2.1	4.07	31.88			0
8	2.2	4.54	37.76	20	50	1
9	1.7	3.46	26.82			0
10	1.8	4.09	32.99	20	50	1
11	2.4	4.55	34.08	20	50	1
12	2.4	5.03	41.85	20	50	1
13	1.9	3.80	31.89	20	50	1
14	1.9	4.45	34.29	20	50	1
15	2.6	4.92	38.45	20	50	1
16	2.6	5.40	44.66	20	50	1
17	2.0	4.06	33.20	20	50	1
18	2.1	4.74	39.18	20	50	1
19	2.7	5.22	44.10	20	50	1
20	2.7	5.71	46.94			0

Tabla 5: 20 valores de una muestra del modelo de la sección 5.10



Figuras 8: Estimaciones del punto  $a=(-4,10)$  con el algoritmo MdD de una iteración de la sección 5.3 para diferentes puntos de arranque.



Figuras 9: Comparación de los errores absolutos de estimación entre los algoritmos con correcciones en media (trazo fino) y en moda (trazo grueso)

### Notas y comentarios

- La aleatorización sobre los intervalos de censura (véanse secciones 5.8 y 5.9) constituyen una vía para abordar el problema técnico derivado del hecho de que los errores  $\varepsilon(a)$  no tengan por qué estar centrados en medias, cuando las correcciones se ejercen en moda. Aunque en el capítulo 4 no se hizo mención alguna sobre la citada aleatorización, con correcciones en media la condición  $E(\varepsilon(a)) = 0$  está asegurada. En consecuencia, esta última igualdad continúa siendo cierta cualquiera que fuera el tipo de aleatorización que se eligiera. Se sigue que todos los resultados del capítulo 4 continúan siendo válidos en este caso sin necesidad de imponer ninguna condición de simetría.





**Parte III**  
**Generalizaciones**

## 6. Corrección general

### 6.1 Introducción

En los capítulos 2 a 5 se han estudiado distintos algoritmos de estimación con dos o una iteración, cuando la información es parcial por haberse sometido a censura. La idea central de todos los algoritmos citados radica en la existencia de un proceso de corrección. En éste se estima, a partir de la información actual, el valor de los datos censurados, con el objetivo de obtener una muestra completa particular. Con ésta se opera, después, como si el problema fuera con información completa. Hasta ahora, sólo se han considerado dos tipos de correcciones: en media y en moda.

Con ambas correcciones se obtienen resultados de convergencia estocástica similares, pese a las diferencias significativas existentes entre ellos (por ejemplo, la modal es libre de distribución y la otra no). Este hecho induce a pensar si con otros tipos de correcciones, dependientes o no de distribución, se podrían seguir garantizando convergencias estocásticas similares a las establecidas en los teoremas de los capítulos 2 a 5. Si así fuera, las posibilidades de corrección se ampliarían y, en cada caso concreto, se podría optar por una u otra, atendiendo a criterios de facilidad de cálculo, rapidez de convergencia, precisión, etc. Por ejemplo, cuando un dato estuviese censurado en el intervalo  $(a, b)$ , se podría imputar su valor desconocido (en lugar de por la media o la moda condicionada a dicho intervalo y a la estimación actual del parámetro) por la mediana o,

simplemente, por un punto intermedio entre  $a$  y  $b$ , etc. En esta sección se analizarán las distintas funciones de corrección posibles, así como las condiciones que será necesario imponer sobre ellas para que el estimador resultante sigan manteniendo buenas propiedades asintóticas, e. g., consistencia, normalidad asintótica, etc.

En este sentido, se verá a lo largo del capítulo que las correcciones en mediana se comportan de una manera muy similar a las correcciones en moda, bajo determinadas condiciones (lema 6.2). Cuando éstas se cumplen y, además,  $F^{-1}$  admite forma explícita (siendo  $F$  la distribución de los términos de error), es fácil comprobar que las correcciones en mediana son muy sencillas de aplicar. Así se evitan las aproximaciones por cuadratura involucradas en las correcciones en media que justificaron, en su momento, su sustitución por correcciones modales. Lo mismo ocurre cuando las correcciones dentro de un intervalo se efectúan a través de puntos intermedios al mismo. Puesto que estas dos nuevas correcciones se equiparan a la moda, los comentarios vertidos en la introducción del capítulo 5 podrían ser repetidos ahora.

Finalmente, en la última parte de este capítulo (sección 6.6) se verá que, incluso, las correcciones aleatorias podrían tener un comportamiento similar a las correcciones fijas, pudiendo aquéllas estar justificadas en ciertos casos.

## 6.2 Formalización del modelo y diferentes tipos de correcciones

Como en los capítulos precedentes, volvamos a considerar el modelo de regresión lineal

$$z_i = a^t x_i + \nu_i, \quad i = 1, \dots, n, \quad (45)$$

donde  $a$  y  $x_i$  son  $m$ -dimensionales, y los errores  $\nu_i$  son independientes e idénticamente distribuidos con media cero y varianza uno. La forma de agrupación

o censura de datos coincide con la descrita en secciones anteriores, existiendo  $s$  criterios diferentes de agrupación, que aparecen en la población con probabilidades  $\pi_1, \dots, \pi_s$ . El criterio  $j$ -ésimo se determina por los puntos extremos de sus intervalos de clasificación

$$-\infty = c_{j,0} < c_{j,1} < \dots < c_{j,r_j-1} < c_{j,r_j} = \infty.$$

Se propone llevar a cabo la estimación del parámetro  $a$  mediante un procedimiento recursivo, según el cual se imputan los datos censurados usando la información disponible hasta el momento, a través de una función de corrección seleccionada dentro de una familia, más o menos amplia, de posibilidades. La citada familia de funciones reales con valores en  $\mathcal{R}$  será denotada por

$$\Delta = \{ \delta_{t,w} : t = c_{j,h-1}, w = c_{j,h} \text{ para algún } j = 1, \dots, s \text{ y } h = 1, \dots, r_j \}.$$

Se supondrá que todos sus miembros satisfacen las siguientes tres condiciones, las cuales serán mantenidas como hipótesis a lo largo de todo el capítulo:

CONDICION 1:  $\beta + t \leq \delta_{t,w}(\beta) \leq \beta + w$

CONDICION 2:  $\delta_{t,w}(\beta)$  y  $\beta - \delta_{t,w}(\beta)$  son no decrecientes en todo  $\beta \in \mathcal{R}$ .

CONDICION 3:  $\delta_{t,w}(\beta)$  es diferenciable con continuidad localmente en cada punto  $-a^t x_i$ , cualquiera que sea  $i \in I^g$ , excepto para un número finito de dichos índices.

A continuación se indican distintas familias  $\Delta$  que satisfacen las CONDICIONES 1-3.

### 6.2.1 Algunos ejemplos de correcciones

i) IMPUTACIONES DE DISTRIBUCION-LIBRE: Para cada  $t, w$  finitos, tómesese algún valor arbitrario  $0 \leq \alpha_{t,w} \leq 1$ , y definamos  $\delta_{t,w}(\beta)$  como el valor intermedio entre  $(\beta + t)$  y  $(\beta + w)$  dado por

$$\delta_{t,w}(\beta) = (\beta + t)\alpha_{t,w} + (\beta + w)(1 - \alpha_{t,w}).$$

Adicionalmente, definase  $\delta_{t,\infty}(\beta) = \beta + t + a_1$  y  $\delta_{-\infty,w}(\beta) = \beta + w - a_2$ , donde  $a_1 > 0$  y  $a_2 > 0$  son dos valores previamente fijados. Es claro que todas las funciones definidas de este modo son lineales y con pendiente uno, por lo que claramente satisfacen las CONDICIONES 1-3.

ii) IMPUTACION EN MODA: Tómese

$$\delta_{t,w}(\beta) = \text{Moda}(\nu | \nu \in (\beta + t, \beta + w)).$$

Cuando los errores tienen función de densidad unimodal en cero, estas funciones toman la forma explícita

$$\delta_{t,w}(\beta) = (w + \beta)I_{(-\infty, -w]}(\beta) + (t + \beta)I_{[-t, \infty)}(\beta)$$

que es lineal a trozos y claramente cumplen las CONDICIONES 1-2. La tercera condición no es muy restrictiva y, por ejemplo, se cumple con probabilidad uno, si la distribución subyacente de las  $x_i$ 's es continua. Este caso se corresponde exactamente con el tratado en los capítulos 3 y 5.

iii) IMPUTACION EN MEDIA: En este caso, tómese

$$\delta_{t,w}(\beta) = E(\nu | \nu \in (\beta + t, \beta + w)) \quad (46)$$

con lo cual  $\Delta$  obviamente satisface la CONDICION 1. Si la función de densidad de los errores  $\nu$  es continua y positiva en todo  $\mathcal{R}$ , la CONDICION 3 también se cumple y, además,  $\delta_{t,w}(\beta)$  crece estrictamente. Así pues, la única hipótesis que debe ser comprobada es la de no decrecimiento de  $\beta - \delta_{t,w}(\beta)$ . El siguiente lema establece una condición suficiente para que esto ocurra. Llamemos  $f(x)$  a la función de densidad del error  $\nu$  y supongamos que  $f(x)$  es una función absolutamente continua, es decir, que (casi seguro Lebesgue) existe  $f'(x)$  tal que, para cada  $t < w$ ,

$$\int_t^w f(x)dx = f(w) - f(t).$$

Defínase en esta situación la función  $\varphi(x) = -f(x)/f'(x)$ ,  $x \in \mathcal{R}$ .

**Lema 6.1** Si  $f$  es una función de densidad absolutamente continua y, para todo  $w > 0$  y  $\beta > 0$ , las tres siguientes covarianzas condicionadas

$$\begin{aligned} &Cov((\nu, \varphi(\nu)) | \beta - w < \nu < \beta + w), \\ &Cov((\nu, \varphi(\nu)) | -\infty < \nu < \beta), \\ &Cov((\nu, \varphi(\nu)) | \beta < \nu < \infty) \end{aligned}$$

son no negativas, entonces se cumplen las CONDICIONES 1-3, y reciprocamente.

DEMOSTRACION: Puesto que se ha asumido que  $f$  es continua, solamente se precisa demostrar que  $\delta_{t,w}(\beta) - \beta$  es una función no creciente en  $\beta$  para todo  $t < w$ , donde al menos uno de estos dos valores es finito. Para valores dados  $t < w$  donde ambos son finitos,  $\delta_{t+d,w+d}(\beta) = \delta_{t,w}(\beta + d)$ . Se sigue que si el lema es cierto para los extremos del intervalo  $[t, w]$ , entonces también es cierto para los extremos de cualquier otro intervalo  $[t^0, w^0]$  de la misma longitud. En consecuencia, sin pérdida de generalidad, se puede asumir que  $t = -w$ . Omitiendo el subíndice ( $\delta_{-w,w} = \delta$ ) para mayor simplicidad, es fácil comprobar que  $\delta$  es una función par y continua. Así pues, basta demostrar que  $\delta(\beta) - \beta$  no crece estrictamente en  $\beta > 0$ . A continuación, escríbase

$$\delta(\beta) = \frac{\int_{-w}^w (x + \beta) f(x + \beta) dx}{\int_{-w}^w f(x + \beta) dx}$$

y diferénciese dentro del signo integral para concluir que

$$\frac{d\delta(\beta)}{d\beta} = 1 - cov((\nu, \varphi(\nu)) | \beta - w < \nu < \beta + w).$$

Esto significa que si la covarianza de la parte derecha es positiva, entonces  $\beta - \delta(\beta)$  es creciente. Sólo resta probar que  $\delta_{-\infty,w}(\beta) - \beta$  y  $\delta_{t,\infty}(\beta) - \beta$  no crece, cualesquiera que sean los valores finitos  $t$  y  $w$ . Como antes, obsérvese que  $\delta_{-\infty,w+d}(\beta) = \delta_{-\infty,w}(\beta + d)$ , (existe una expresión similar para  $\delta_{t,\infty}(\beta)$ ) de donde la condición deseada se cumple para todo  $t, w$  si, y sólo si, se verifica para cualquier valor particular, por ejemplo para  $t = w = 0$ . Finalmente, obsérvese que  $\delta_{-\infty,0}(\beta) - \beta = -\delta_{0,\infty}(\beta) - \beta$ . Así pues, sólo una de las dos condiciones aún no probadas tiene que ser demostrada, pero sólo para  $\beta > 0$ . Un

razonamiento similar al utilizado más arriba al tratar el caso de  $\delta_{-w,w}$ , conduce a afirmar que  $\delta_{-\infty,0}(\beta) - \beta$  y  $\delta_{0,\infty}(\beta) - \beta$  son ambas funciones no decrecientes en  $\beta > 0$  si las covarianzas condicionadas  $cov((\nu, \varphi(\nu) | -\infty < \nu < \beta))$  y  $cov((\nu, \varphi(\nu) | \beta < \nu < \infty))$  son no negativas, lo cual completa la demostración.  $\square$

Puede comprobarse fácilmente que las siguientes densidades, entre otras, cumplen las condiciones del lema 6.1: normal  $N(0, \sigma_0)$ ; Laplace;  $f(x) = 2^{-1} (1 + |x|)^{-1}$ ,  $x \in \mathcal{R}$ .

iv) IMPUTACIONES EN MEDIANA: Defínase ahora

$$\delta_{t,w}(\beta) = \text{Mediana}(\nu | \nu \in (\beta + t, \beta + w)). \quad (47)$$

La CONDICION 1 se verifica obviamente. Adicionalmente, como en el ejemplo anterior, si la función de densidad de los errores  $\nu$  es continua, la CONDICION 3, así como la primera parte de la CONDICION 2, también se cumplen. El siguiente lema proporciona una condición suficiente para que  $\beta - \delta_{t,w}(\beta)$  sea no decreciente (y, por consiguiente, las tres condiciones antes citadas se verifiquen), cuando las densidades de los errores son simétricas y unimodales en cero.

**Lema 6.2** *Si  $f(x) > 0$  es una densidad simétrica y unimodal en cero y, además, para cada  $a > 0$ ,  $\varphi_a(x) = f(x)/f(x+a)$  es una función no decreciente en cualquier  $x > 0$ , entonces se cumplen las CONDICIONES 1-3.*

DEMOSTRACION: Como se ha dicho, solamente es necesario comprobar que  $\beta - \delta_{t,w}(\beta)$  es no decreciente. En efecto, la hipótesis de enunciado garantiza que  $f(z)$  alcanza su máximo en  $z = 0$ , y decrece cuando  $|z| \rightarrow \infty$ . Supóngase que  $t$  y  $w$  son finitos. Puesto que  $\delta_{t+d,w+d}(\beta) = \delta_{t,w}(\beta + d)$ , es obvio comprobar que si el lema es cierto para los extremos del intervalo  $[t, w]$ , también lo es para los extremos de cualquier otro intervalo  $[t^0, w^0]$  de la misma longitud. Así pues, sin pérdida de generalidad, se puede asumir que  $t = -w$  y se omitirán los subíndices ( $\delta_{-w,w} = \delta$ ) puesto que no se hará uso de ellos en lo sucesivo de la demostración. Adicionalmente, puesto que  $\delta(-\beta) = -\delta(\beta)$  y  $\delta(0) = 0$ , es suficiente con probar

el lema para  $\beta > 0$ . Si  $0 < \beta \leq t$ , la propiedad es claramente cierta. Para  $\beta > t$ , se debe probar que para todo  $d > 0$

$$\delta(\beta + d) - \delta(\beta) \leq d.$$

Para verlo, nótese que la mediana de  $\nu_i | (\beta - t \leq \nu_i \leq \beta + t)$  no puede ser menor que la mediana de  $\nu_i - d | (\beta - t + d \leq \nu_i \leq \beta + t + d)$ , puesto que sus respectivas densidades son no nulas en el intervalo  $[\beta - t, \beta + t]$ , ambas son no crecientes y  $f(z)/f(z+a)$  no decrece. Esto significa que la gráfica de la segunda densidad comienza por encima de la de la primera, y se cruzan solamente una vez, lo cual implica la mencionada propiedad de las medianas. Esto concluye la demostración puesto que para  $t = -\infty$  or  $w = \infty$ , es posible apelar a un argumento similar.  $\square$

Los siguientes son algunos ejemplos de funciones de densidad para las cuales el lema 6.2 es útil, garantizándose así que las CONDICIONES 1-3 se cumplen: normal  $N(0, \sigma_0)$ ; distribución de Laplace; distribución logística;  $f$  plana alrededor del cero y teniendo colas normales o exponenciales.

La condición impuesta sobre  $f(z)/f(z+a)$  significa que la tasa de crecimiento relativo de la función de densidad del error sobre cualquier intervalo positivo de longitud  $a$  no decrece cuando el intervalo de desplaza hacia la derecha.

Existen condiciones suficientes adicionales para verificar que  $\beta - \delta_{t,w}(\beta)$  es no decreciente. Una, que podría ser útil cuando  $F^{-1}$  admite una expresión explícita, simplemente obliga a  $fF^{-1}$  a ser una función cóncava en  $(0, 1)$ . Para demostrar esto último basta simplemente diferenciar.

v) MIXTURAS: En los ejemplos previos, todas las funciones de la familia  $\Delta$  corresponden al mismo tipo de imputaciones, pero no existe ningún problema en que algunos de los elementos de  $\Delta$  sean, por ejemplo, imputaciones en moda y otros lo sean en mediana. O bien, en otro sentido, que un elemento  $\delta_{t,w}$  sea una combinación lineal convexa de las correcciones, por ejemplo, en moda, mediana



y del punto medio. De esta forma, se pone de manifiesto que las posibilidades de elección de la familia  $\Delta$  son muy amplias.

Propondremos a continuación un método de dos iteraciones utilizando una familia arbitraria de correcciones  $\Delta$ , y después se extenderán los resultados al caso de algoritmos de una iteración. Así, se seguirá el mismo orden de exposición de los capítulos 2 a 5.

### 6.3 Método general de estimación en dos iteraciones

Recuérdese que en los métodos de dos iteraciones, la obtención de la estimación del parámetro desconocido  $a$  en la etapa primaria  $n$ , suponía la iteración sucesiva de un proceso secundario, cuyo límite proporcionaba dicha estimación. El anteriormente mencionado método de estimación en dos iteraciones consiste estrictamente en los siguientes pasos:

**ITERACIÓN PRIMARIA:** La etapa  $n$ -ésima de este proceso iterativo corresponde a tomar  $n$  valores muestrales, que pueden ser censurados o no, de acuerdo con las probabilidades de censura establecidas en la sección 6.2.

**ITERACIÓN SECUNDARIA:**

**INICIALIZACION:** Fíjese un valor arbitrario de la estimación inicial  $a_0$ , del vector  $a$ .

**ITERACION:** Asumiendo que la estimación en la etapa  $p$  corresponde a  $a_p$ , calcúlese un vector de datos, imputando cada uno de los valores faltantes mediante un método de corrección extraído de la familia  $\Delta$ , es decir, para cada  $i \in I$  tómese

$$\begin{aligned} y_i(a_p) &= z_i, & \text{si } i \in I^{ng} \\ &= a_p^t x_i + \delta_{c_{j,r}, c_{j,r+1}} (-a_p^t x_i), & \text{si } i \in I_j, z_i \in (c_{j,r}, c_{j,r+1}]. \end{aligned}$$

Después, actualícese la estimación en la etapa  $p$  mediante el proceso de proyección

habitual, utilizando el vector de datos completo obtenido, es decir,

$$a_{p+1} = (X^t X)^{-1} X^t y(a_p).$$

Este método recursivo se referirá en lo sucesivo como el  $\Delta$ -algoritmo de dos iteraciones, para hacer ver la dependencia de la familia de correcciones  $\Delta$  utilizada. Nótese que a lo largo del proceso secundario la matriz de datos  $X^t = (x_1 \dots x_n)$  permanece invariable, cambiando sólo en la iteración primaria.

Puesto que la clase  $\Delta$  está abierta a muchas posibilidades, se puede disponer de una gran variedad de correcciones para imputar los datos censurados. Usualmente, pero no siempre (tal como se vio en los ejemplos), se definirá cada elemento de  $\Delta$  como una función dependiente de los errores  $\nu$  que aparecen en el modelo lineal (45). En cualquiera de los casos, dependan o no de  $\nu$ , se demostrará que el proceso iterativo secundario converge siempre, cualquiera que sea el punto inicial, a un único punto fijo, que no depende de este vector de arranque. Adicionalmente, se verán diferentes convergencias estocásticas del proceso iterativo primario, según el tamaño muestral aumenta.

### 6.3.1 Convergencia del proceso secundario

El siguiente teorema refleja la convergencia del proceso secundario del  $\Delta$ -algoritmo, cualquiera que sea el vector inicial.

**Teorema 6.3** *Asúmase que  $(X^t X)^{ng} = \sum_{i \in I^{ng}} x_i x_i^t$  es definida positiva. Para cualquier vector inicial  $a_0$ , la sucesión  $a_p$  generada por el  $\Delta$ -algoritmo converge a un vector  $a^*$ , que satisface la ecuación implícita*

$$a^* = (X^t X)^{-1} X^t y(a^*). \quad (48)$$

*Además, el punto  $a^*$  es la única solución de la anterior ecuación.*

DEMOSTRACION: Similar a la del teorema 2.1.  $\square$

La existencia de este único punto límite  $a^*$  de cualquier sucesión generada por el  $\Delta$ -algoritmo permite definir este valor como la estimación del parámetro de

regresión en la etapa  $n$ , el cual se corresponde con un M-estimador de Huber, al estar determinado mediante la ecuación implícita (48). Obsérvese que esta estimación es independiente de la estimación inicial, al contrario de lo que sucede en los algoritmos de una iteración.

### 6.3.2 Convergencias estocásticas del proceso iterativo primario

Ahora se harán iteraciones en el índice  $n$  del proceso primario. Denotemos  $a^* = a_n$  (para notar la dependencia de la estimación del tamaño muestral  $n$ ). En los siguientes teoremas se investigará la convergencia de  $a_n$  cuando  $n \rightarrow \infty$ , asumiendo que  $a$  es el verdadero valor del parámetro de regresión. En todos ellos, se omitirán las demostraciones por ser similares a las desarrolladas en teoremas previos.

**Teorema 6.4** *Si  $(X^t X)^{ng}$  es definida positiva y existe  $\rho > 1$  para el cual se verifican las siguientes propiedades*

$$\inf_n \lambda_n = \lambda > 0, \quad (49)$$

$$\max_{i \leq n} \|x_i\|^2 = O(n^{\rho-1}), \quad (50)$$

donde  $\lambda_n$  denota el mínimo autovalor de la matriz  $n^{-\rho}(X^t X)^{ng}$ , entonces  $a_n \xrightarrow{L_2} a$ , y, así,  $a_n$  es un estimador consistente de  $a$ . Además, la sucesión  $n^{\frac{\rho}{2}}(a_n - E(a_n))$  está acotada en  $L_2$ .

DEMOSTRACION: Similar a la del teorema 3.2.  $\square$

Llegados a este punto, es interesante formular un teorema central del límite para  $a_n$ , con el fin de poder obtener, aparte de la estimación puntual de  $a$ , intervalos de confianza y realizar contrastes de hipótesis.

**Teorema 6.5** *Bajo las hipótesis del teorema anterior, la sucesión  $n^{\frac{\rho}{2}}(a_n - E(a_n))$  converge en distribución a una normal, es decir,*

$$n^{\frac{\rho}{2}}(a_n - E(a_n)) \xrightarrow[n \rightarrow \infty]{D} N(0, \Lambda)$$

para alguna matriz de covarianzas  $\Lambda$ . Esto quiere decir que, supuesto  $n$  suficientemente grande, la distribución de  $a_n - a$  puede ser aproximada por

$$(a_n - a) \approx N(0, n^{-\rho}\Lambda).$$

DEMOSTRACION: Similar a la del teorema 3.3.  $\square$

A continuación se pasará a exponer algunos resultados para los algoritmos de una iteración.

## 6.4 Métodos generales de estimación en una iteración

### 6.4.1 Método completo

En semejanza con las secciones 4.3 y 5.3, se propondrá aquí un algoritmo de una iteración con correcciones generales. Las demostraciones que afectan a las convergencias estocásticas se omitirán, referenciando sólo sus precedentes argumentales, tal como se ha hecho en la sección anterior.

El proceso iterativo de estimación será el siguiente:

INICIALIZACIÓN: Fijese un valor arbitrario para la estimación inicial  $a_1$ , del vector  $a$ .

ITERACIÓN: Asumiendo que la estimación en la etapa  $n$  es  $a_n$ , calcúlese un vector de datos, imputando cada uno de los valores faltantes mediante un método de corrección extraído de la familia  $\Delta$ , esto es, para  $i \in I = \{1, \dots, n\}$

$$\begin{aligned} y_i(a_n) &= z_i, & \text{si } i \in I^{ng} \\ &= a_n^t x_i + \delta_{c_{j,r}, c_{j,r+1}} (-a_n^t x_i), & \text{si } i \in I_j, z_i \in (c_{j,r}, c_{j,r+1}]. \end{aligned}$$

Después, actualícese la estimación en la etapa  $n$  mediante la ecuación

$$a_{n+1} = (X^t X)^{-1} X^t y(a_n).$$

Este método recursivo mantiene unas propiedades de convergencia análogas a las del método de dos iteraciones, las cuales se establecen en el siguiente teorema.

**Teorema 6.6** Sean  $(X^t X)^{ng}$  y  $(X^t X)^g$  matrices definidas positivas para cada  $n \in \mathcal{N}$  y supóngase que existe un número real  $\rho > 1$  para el cual las siguientes condiciones se cumplen

$$\inf_n \lambda_n = \lambda > 0, \\ \max_{i \leq n} \|x_i\|^2 = O(n^{\rho-1}),$$

donde  $\lambda_n$  denota el mínimo autovalor de la matriz  $n^{-\rho}(X^t X)^{ng}$ . En estas

condiciones,  $a_n \xrightarrow{L_1} a$  y, además, existe una matriz de covarianzas  $\Sigma$  tal que  $a_n$  converge en distribución a la normal

$$n^{\frac{\rho}{2}} (a_n - E(a_n)) \xrightarrow[n \rightarrow \infty]{D} N(0, \Sigma).$$

DEMOSTRACION: Basta argumentar como en los teoremas 5.1 y 5.2.  $\square$

### 6.4.2 Método basado en aproximaciones estocásticas

En este caso, se debe disponer de una sucesión de valores reales, o tamaños de paso,  $\alpha_n$  de forma que, elegidos de forma adecuada, se consiga que el algoritmo propuesto converja al verdadero valor del parámetro  $a$ .

INICIALIZACIÓN: Tómese un vector inicial  $a_1$  arbitrario.

ITERACIÓN: Partiendo de la estimación  $a_n$  en la etapa  $n$ , su actualización en la etapa  $n + 1$  tomará la forma

$$a_{n+1} = a_n + \alpha_n X^t \varepsilon(a_n),$$

donde el vector  $\varepsilon(a_n) = (\varepsilon_1(a_n), \dots, \varepsilon_n(a_n))$  tiene por componentes  $\varepsilon_i(a_n) = y_i(a_n) - a_n^t x_i$ , siendo  $y_i(a_n)$  el valor imputado para el dato censurado  $i$ -ésimo de acuerdo con la función de la familia  $\Delta$  correspondiente.

El siguiente teorema propone algunas condiciones suficientes, aunque no necesarias, para que se produzca la convergencia citada.

**Teorema 6.7** Sean  $(X^t X)^{ng}$  y  $(X^t X)^g$  matrices definidas positivas para cada  $n \in \mathcal{N}$ . Supongamos que existe  $\rho > 1$  para el cual las siguientes condiciones técnicas se cumplen

$$\inf_n \lambda_n = \lambda > 0,$$

$$\max_{i \leq n} \|x_i\|^2 = O(n^{\rho-1}),$$

donde  $\lambda_n$  denota el mínimo autovalor de la matriz  $n^{-\rho}(X^t X)^{ng}$ . Selecciónese en cada etapa un tamaño de paso  $\alpha_n$  de forma que  $\alpha_n n^\rho \rightarrow 0$ ,  $\sum \alpha_n n^\rho = \infty$  y  $\sum \alpha_n n^{\frac{\rho}{2}} < \infty$ . En estas condiciones, el  $\Delta$ -algoritmo de una iteración enunciado inmediatamente converge en  $L_1$ , es decir,  $a_n \xrightarrow{L_1} a$ .

DEMOSTRACION: Sigue la misma línea argumental del teorema 5.5.  $\square$

## 6.5 Observaciones y otros algoritmos

### 6.5.1 Otro algoritmo alternativo al completo

En el método en dos iteraciones, el llamado proceso secundario era un proceso iterativo que convergía a un punto que se consideraba el valor de la estimación en la etapa primaria actual. En el método completo anterior conveníamos en hacer solamente una iteración de ese proceso secundario. Pese a la simplificación, se obtenía de hecho convergencia en los mismos términos. Lo que se propone ahora es un método intermedio entre estos dos, de forma que en el proceso secundario hagamos un número  $k > 1$  de iteraciones, en lugar de infinitas. Este número de iteraciones  $k$  podría, incluso, variar dependiendo de la etapa primaria  $n$ . En todo caso, la razón de su aparición viene marcada por el hecho de que, a nivel práctico, en el proceso secundario nunca se podrán ejecutar las infinitas iteraciones de que consta, pudiendo, además, ser muy costoso realizar aquéllas necesarias para llegar a una buena aproximación del límite.

El algoritmo puede escribirse de la siguiente forma:

INICIALIZACIÓN: Fíjese un valor inicial arbitrario como estimación de  $a$ , digamos  $a_1$ .

ITERACIÓN: Si la estimación en la etapa primaria  $n$  es  $a_n$ , realícense  $k_n < \infty$  iteraciones secundarias como sigue:

INICIALIZACION: Tómesese  $a_n^1 = a_n$ .

ITERACION: Partiendo de  $a_n^p$ , calcúlese la siguiente iteración en  $p$  mediante

$$a_n^{p+1} = (X^t X)^{-1} X^t y(a_n^p), \quad p = 1, \dots, k_n,$$

donde  $X^t = (x_1 \dots x_n)$  y el vector  $y(a_n^p)$  se define como en la sección anterior, utilizando la familia de correcciones  $\Delta$ .

La actualización de la estimación en la etapa  $n + 1$  iguala al último vector de la iteración anidada, es decir,

$$a_{n+1} = a_n^{k_n+1}.$$

Para este proceso cabe esperar que se sigan verificando resultados de convergencia en línea con los de la sección 6.4. En concreto,  $a_{n+1} = a_n^{k_n+1} \xrightarrow{L_2} a$  y un resultado de convergencia en distribución similar al que aparece en el teorema 6.6.

## 6.5.2 Sobre la corrección en media

En el ejemplo iii) de la sección 6.2.1, al igual que en los capítulos 2 y 4 se propone utilizar correcciones en media. En ambos casos, se obtienen convergencias estocásticas similares bajo ciertas condiciones. Las impuestas en este capítulo no coinciden con las que se incluyen en los dos restantes capítulos citados. La razón es que en esta parte de la memoria se hace un planteamiento general para la clase amplia de correcciones  $\Delta$ , el cual no explota la especificidad de la corrección particular en media. En todo caso, los resultados de este capítulo son compatibles con todos los anteriores, si bien, repito, no coinciden.

## 6.5.3 Correcciones aleatorias

Según se adelantó en la introducción, las correcciones no tienen necesariamente que ser funciones deterministas, pudiendo ser, también, aleatorias. Considérese la

familia de variables aleatorias

$$\Delta^* = \{ \delta_{t,w}^* : t = c_{j,h-1}, w = c_{j,h}, \text{ para algún } j = 1, \dots, s \text{ y } h = 1, \dots, r_j \},$$

donde  $\delta_{t,w}^*(\beta)$  es una variable aleatoria con valores en el intervalo  $[\beta + t, \beta + w]$  de media  $\delta_{t,w}(\beta)$ , para cada  $\beta \in \mathcal{R}$ , es decir,

$$E(\delta_{t,w}^*(\beta)) = \delta_{t,w}(\beta).$$

Obsérvese que  $\delta_{t,w}(\beta)$  debe verificar las CONDICIONES 1-3. Adicionalmente, supongase que esas variables aleatorias tienen varianzas acotadas

$$Var(\delta_{t,w}^*(\beta)) \leq \sigma^2,$$

para cada  $t, w$ .

Con estas condiciones se puede plantear un  $\Delta^*$ -algoritmo en una iteración de la siguiente forma:

INICIALIZACIÓN: Fijese el valor arbitrario de la estimación inicial  $a_1$ .

ITERACIÓN: Asumiendo que la estimación en la etapa  $n$  es  $a_n$ , calcúlese el vector de datos completo, es decir, para  $i \in I = \{1, \dots, n\}$  calcúlese

$$\begin{aligned} y_i(a_n) &= z_i, & \text{si } i \in I^{ng} \\ &= a_n^t x_i + \delta_{c_{j,r}, c_{j,r+1}}^* (-a_n^t x_i), & \text{si } i \in I_j, z_i \in (c_{j,r}, c_{j,r+1}], \end{aligned}$$

donde aquí  $\delta_{c_{j,r}, c_{j,r+1}}^* (-a_n^t x_i)$  representa esa propia variable aleatoria realizada.

Después, actualícese la estimación mediante la ecuación

$$a_{n+1} = (X^t X)^{-1} X^t y(a_n).$$

Con esta modificación se sigue manteniendo el resultado de convergencia visto para el algoritmo no aleatorio equivalente.

**Teorema 6.8** *El teorema 6.6 continúa siendo válido cuando la sucesión  $a_n$  proviene del  $\Delta^*$ -algoritmo de una iteración.*

DEMOSTRACION: Sígase la del teorema 6.6 y las subsiguientes observaciones de la demostración del teorema 6.9.  $\square$



Resultados análogos se podrían enunciar para los algoritmos de dos iteraciones y para los algoritmos de una iteración basados en aproximaciones estocásticas (utilizando distintos tamaños de paso  $\alpha_n$ ), ambos con correcciones aleatorias.

## 6.6 Aleatorización de las correcciones en media

Cuando las correcciones son en media, los datos censurados se han venido sustituyendo hasta ahora por su media condicionada al intervalo de censura y a la actual estimación del parámetro. En esta sección, en la misma línea que en la anterior 6.5.3, se propone sustituir el dato censurado no por su media sino por una variable aleatoria cuya esperanza coincida con ella.

En concreto, para cada  $t < w$  fijos, considérese una variable aleatoria tomando valores en el intervalo  $[t, w]$ , digamos  $\gamma^*(t, w)$ , cumpliéndose

$$E(\gamma^*(t, w)) = \gamma(t, w) = E(\nu | \nu \in [t, w])$$

y

$$Var(\gamma^*(t, w)) \leq \sigma^2 < \infty.$$

Se puede plantear un algoritmo MD de una iteración con correcciones en media aleatorias de la siguiente forma:

INICIALIZACIÓN: Fijese el valor arbitrario de la estimación inicial  $a_1$ .

ITERACIÓN: Asumiendo que la estimación en la etapa  $n$  es  $a_n$ , calcúlese el vector de datos completo

$$\begin{aligned} y_i(a_n) &= z_i, & \text{si } i \in I^{ng} \\ &= a_n^t x_i + \gamma^*(-a_n^t x_i + c_{j,r}, -a_n^t x_i + c_{j,r+1}), & \text{si } i \in I_j, z_i \in (c_{j,r}, c_{j,r+1}] \end{aligned}$$

donde ahora  $\gamma^*(-a_n^t x_i + c_{j,r}, -a_n^t x_i + c_{j,r+1})$  denotan realizaciones bajo independencia de las variables aleatorias análogas. Después, actualícese la estimación en la etapa  $n$  mediante la ecuación

$$a_{n+1} = (X^t X)^{-1} X^t y(a_n).$$

Con esta modificación, se sigue manteniendo un resultado de convergencia similar al de la versión equivalente no aleatorizada del algoritmo.

**Teorema 6.9** *El teorema 4.1 continúa siendo válido cuando la sucesión  $a_n$  proviene del anterior algoritmo MD de una iteración con correcciones aleatorias.*

DEMOSTRACION: Siguiendo la línea de la demostración del teorema 4.1, se puede escribir

$$y(a) = Xa + \bar{\varepsilon}(a),$$

donde

$$\bar{\varepsilon}(a) = \varepsilon(a) + \sum_{j=1}^s \sum_{r=0}^{r_j-1} \eta_{i,j,r} R_{jr}(-a^t x_i).$$

El vector  $\varepsilon(a)$  está definido en el citado teorema 4.1 y

$$R_{jr}(-a^t x_i) = \gamma^*(-a_n^t x_i + c_{j,r}, -a_n^t x_i + c_{j,r+1}) - \gamma(-a_n^t x_i + c_{j,r}, -a_n^t x_i + c_{j,r+1}).$$

Llamemos  $R_i(a) = \sum_{j=1}^s \sum_{r=0}^{r_j-1} \eta_{i,j,r} R_{jr}(-a^t x_i)$ , siendo una variable aleatoria nula si  $i \in I^{n_g}$ . Definiendo

$$\bar{\varepsilon}(a_n) = y(a_n) - Xa_n,$$

se puede concluir, a partir de las propiedades de  $\gamma$ , que

$$y(a_n) - y(a) = (I - M^*)X(a_n - a) + R(a_n) - R(a),$$

donde  $R = (R_1, \dots, R_n)$ . En definitiva,

$$\bar{\varepsilon}(a_n) - \bar{\varepsilon}(a) = -M^*X(a_n - a) + R(a_n) - R(a),$$

donde  $M^*$  se define aquí como en el teorema 4.1. Por último, escribamos

$$a_{n+1} - a = \left[ I - (X^t X)^{-1} X^t M^* X \right] (a_n - a) + (X^t X)^{-1} X^t \zeta(a_n, a),$$

siendo  $\zeta(a_n, a) = \bar{\varepsilon}(a) + R(a_n) - R(a) = \varepsilon(a) + R(a_n)$ . Obsérvese que

$$E(R_i(a_n) | a_n) = 0,$$

con lo cual, al ser  $E(\varepsilon(a)) = 0$ , resulta que

$$E(\zeta_i(a_n, a)) = E(E(\zeta_i(a_n, a) | a_n)) = E(E(R_i(a_n) | a_n)) + E(\varepsilon_i(a)) = 0.$$

Por otra parte,

$$E (R_i^2 (a_n) | a_n) \leq \sigma^2,$$

de donde, por la desigualdad de Minkowski, resulta

$$E (\zeta_i^2 (a_n, a) | a_n) \leq \left[ E (R_i^2 (a_n) | a_n)^{\frac{1}{2}} + E (\varepsilon_i(a))^{\frac{1}{2}} \right]^2 \leq (\sigma + 1)^2.$$

Por último, las componentes de  $\varepsilon(a)$  son independientes, así como también las de  $R(a_n)$  condicionadas por  $a_n$ .

Llamando  $c_n = \left\| (X^t X)^{-1} X^t \zeta(a_n, a) \right\|$ , todo lo anterior permite escribir que

$$E (c_n^2) = E (E (c_n^2) | a_n) \leq \lambda^{-2} n^{-1} K^2 (1 + \sigma)^2.$$

Se concluye la demostración razonando como en el teorema 4.1, teniendo en cuenta que las esperanzas se deben tomar primero condicionadas por  $a_n$ .  $\square$

Se puede concluir, a la vista de esta demostración, que los algoritmos que utilizan correcciones aleatorias se comportan idénticamente a los que utilizan correcciones deterministas. Por ello, se puede afirmar que, utilizando dichas correcciones aleatorias, se siguen verificando todos los teoremas de los capítulos 2 a 5, referidos a algoritmos con dos y una iteración. En todo caso, no se enunciarán explícitamente.

## Notas y comentarios

- Los resultados teóricos de este capítulo parecen indicar que se pueden emplear distintos criterios de corrección (siempre que éstos se establezcan con cierta lógica), garantizándose con todos ellos que los estimadores finales disfrutan de buenas propiedades asintóticas en términos de convergencia estocástica.

## 7. Análisis de otras situaciones colaterales

- En el algoritmo alternativo presentado en la sección 6.5.1, se sugiere sustituir las infinitas iteraciones secundarias en cada etapa primaria por un número finito de ellas. Esta misma idea fue sugerida por LANGE (1995) en contextos del EM. En este último trabajo, la recomendación final consistió en realizar una sola, convirtiéndose así de facto en un algoritmo de iteración única. Pese a ello, se comprueba, en el trabajo citado, que la eficiencia asintótica de ambos algoritmos resulta ser *localmente equivalente*.

## **7. Análisis de otras situaciones colaterales**

### **7.1 . Introducción**

En este capítulo se intenta mostrar cómo se podrían extender las aproximaciones estocásticas sobre modelos lineales con datos censurados, vistas hasta ahora, al caso de modelos no lineales. Ante las dificultades que se presentan en la regresión no lineal con parámetros multidimensionales, se desarrollará inicialmente en detalle el caso unidimensional, exponiéndose después sus posibilidades de ampliación al caso multivariante. Los algoritmos que se proponen se basan en los que han sido desarrollados en los capítulos precedentes.

Cuando las funciones de regresión son generales no se dispone de un paso claro de proyección, similar al lineal, para actualizar la estimación presente. Por esta causa, los algoritmos de aproximaciones estocásticas, similares a los desarrollados en las secciones 4.5, 4.6, 5.5 y 5.6, serán los únicos viables ahora. Es de resaltar que los resultados incluidos en las citadas secciones, al igual que los formulados en este capítulo, establecen convergencias en media cuadrática o en media de orden uno, asegurándose de esta forma la consistencia o convergencia en probabilidad. En KUSHNER Y YIN (1997) muchos resultados se establecen en términos de convergencia casi seguro, si bien muy habitualmente para subsucesiones de la secuencia generada por los algoritmos. Es conocido que, bajo condiciones de acotación, por ejemplo, la convergencia casi seguro implica la convergencia en media y que la convergencia en media sólo implica la convergencia casi seguro

de una subsucesión, pero no de toda la sucesión al completo. Por estas razones es previsible que los resultados aquí establecidos en contextos de censura sean similares a los de KUSHNER Y YIN (1997) y KUSHNER Y CLARK (1978).

En este capítulo, además, se considerarán distintos tipos de censura, mostrándose con algún ejemplo su factibilidad ante situaciones reales. Así se hará presente que las técnicas desarrolladas son aplicables a contextos muy dispares.

## 7.2 Regresión no lineal. Caso unidimensional

Se comenzará considerando un modelo no lineal del tipo

$$z_i = g_i(a) + \nu_i, \quad i = 1, 2, \dots, \quad (51)$$

donde  $z_i \in \mathcal{R}$ ,  $a \in \mathcal{R}$ ,  $g_i : \mathcal{R} \rightarrow \mathcal{R}$  y  $\nu_i$  es un error aleatorio real. Como siempre, se supone que algunos datos son no agrupados, si  $i \in I^{ng}$ , en cuyo caso se observa el valor  $z_i$ , mientras que otros valores  $z_i$  se encuentran agrupados, cuando  $i \in I^g$ . Por simplicidad, aunque sin pérdida de generalidad, se supondrá que existe un único criterio de clasificación determinado por  $r$  puntos extremos

$$-\infty = c_0 < c_1 < c_2 < \dots < c_r = \infty,$$

de forma que, si  $i$  es agrupado, tan sólo se sabe que  $z_i \in (c_j, c_{j+1}]$ , para algún  $j \in \{0, 1, \dots, r-1\}$ . Así, se supone, por último, que la probabilidad de  $I^{ng}$  sobre el conjunto poblacional de índices es  $\pi_0$ . Se trata de estimar el parámetro  $a \in \mathcal{R}$  del modelo (51) a partir de esa muestra sometida al proceso de censura citado mediante un procedimiento inspirado en las aproximaciones estocásticas, en la línea de las secciones 4.6 y 5.6 anteriores. Se supondrá, como es habitual, que los errores  $\nu_i$  tienen media cero y varianza finita.

Se propone el siguiente algoritmo de aproximación estocástica con correcciones aleatorias.

INICIALIZACIÓN: Tómese un valor inicial arbitrario  $a_1 \in \mathcal{R}$ .

ITERACIÓN: Supuesta conocida la estimación  $a_n$  de la etapa  $n$ , su actualización  $n + 1$  se calcula mediante la expresión

$$a_{n+1} = a_n + \beta_n (y_n(a_n) - g_n(a_n)), \quad (52)$$

donde

$$\begin{aligned} y_n(a_n) &= z_n, & \text{si } n \in I^{ng} \\ &= g_n(a_n) - \delta_j^*(-g_n(a_n)), & \text{si } n \in I^g, z_n \in (c_j, c_{j+1}]. \end{aligned}$$

En la última igualdad  $\delta_j^*(-g_n(a_n))$  denota una realización de una variable aleatoria definida sobre el intervalo  $(-g_n(a_n) + c_j, -g_n(a_n) + c_{j+1}]$ . Se supondrá que dicha variable aleatoria cumple que

$$\begin{aligned} E(\delta_j^*(-g_n(a_n)) | a_n) &= \delta_j(-g_n(a_n)) \\ &= E(\nu | \nu \in (-g_n(a_n) + c_j, -g_n(a_n) + c_{j+1}]) \end{aligned}$$

y tienen varianzas uniformemente acotadas

$$\text{Var}(\delta_j^*(-g_n(a_n)) | a_n) \leq \sigma^2,$$

para todo  $a_n \in \mathcal{R}$ . Obsérvese que en el planteamiento anterior, se podría suponer que  $\delta_j^*$  es una variable degenerada en  $\delta_j$ . En este caso, el algoritmo anterior llevaría a cabo correcciones no aleatorias.

Se verá a lo largo del capítulo que, bajo ciertas condiciones, se puede conseguir que la secuencia  $a_n$  generada por el algoritmo converja, en algún sentido estocástico, al verdadero valor  $a$ . Las hipótesis que se impondrán sobre los tamaños de paso son las usuales (véase KUSHNER Y YIN (1997)). Adicionalmente, se impondrán ciertas condiciones de acotación que surgen como consecuencia de la no linealidad del modelo y de la propia censura. Se verá que dichas condiciones no son muy restrictivas, pudiéndose, incluso, relajar para obtener mayor generalidad.

Por último, las funciones  $g_n$  podrían depender de una variable observable, de forma que  $g_n(a) = g(a, x_n)$ , con  $x_n \in \mathcal{R}^m$ , como es habitual. En lo sucesivo, cuando sea factible, se utilizará la notación Leibniz  $\dot{g}_n(\alpha) = \frac{dg_n}{d\alpha}(\alpha)$  y  $\dot{\delta}_j(\alpha) = \frac{d\delta_j}{d\alpha}(\alpha)$ .

El siguiente teorema establece un primer resultado de convergencia en media cuadrática del algoritmo descrito arriba.

**Teorema 7.1** *Supongamos que  $g_n(\cdot)$  y  $\delta_j(\cdot)$  son derivables y admiten derivadas uniformemente acotadas superior e inferiormente, es decir, existen  $\varepsilon > 0$  y  $M < \infty$  tales que*

$$\varepsilon \leq \dot{g}_n(\cdot) \leq M$$

y

$$\varepsilon \leq \dot{\delta}_j(\cdot) \leq M.$$

Si  $\sum \beta_n = \infty$  y  $\sum \beta_n^2 < \infty$ , entonces  $a_n \xrightarrow{L_2} a$  (concluyéndose que  $a_n$  es un estimador consistente de  $a$ ).

DEMOSTRACION: El algoritmo aquí propuesto generaliza los de las secciones 4.6 y 5.6. Se puede escribir en los mismos términos que en las secciones citadas como

$$a_{n+1} = a_n + \beta_n \varepsilon_n(a_n),$$

siendo  $\varepsilon_n(a_n) = y_n(a_n) - g_n(a_n)$ . Es fácil ver que si el modelo es lineal, y se asumen correcciones no aleatorias con tamaño de paso  $\beta_n = \alpha_n x_n$ , el algoritmo propuesto coincide exactamente con el de la sección 4.6.

Como consecuencia directa de las hipótesis del enunciado, resulta que

$$g_n(a_n) - g_n(a) = \dot{g}_n(\xi_n)(a_n - a), \quad (53)$$

siendo  $\xi_n$  un valor intermedio del segmento de extremos  $a_n$  y  $a$ . Por la misma razón, también es cierto que

$$\delta_j(-g_n(a_n)) - \delta_j(-g_n(a)) = -\dot{\delta}_j(\zeta_n)(g_n(a_n) - g_n(a)), \quad (54)$$

donde  $\zeta_n$  pertenece al segmento de extremos  $-g_n(a_n)$  y  $-g_n(a)$ . En definitiva,



si  $n \in I^{ng}$ ,

$$y_n(a_n) - y_n(a) = 0,$$

mientras que, si  $n \in I^g$  y  $z_n \in (c_j, c_{j+1}]$ ,

$$y_n(a_n) - y_n(a) = g_n(a_n) - g_n(a) - \delta_j(\zeta_n) \dot{g}_n(\xi_n)(a_n - a).$$

En consecuencia, la diferencia de los errores muestrales se puede escribir como

$$\begin{aligned} \varepsilon_n(a_n) - \varepsilon_n(a) &= y_n(a_n) - g_n(a_n) - y_n(a) + g_n(a) \\ &= -m_n^*(a_n - a), \end{aligned}$$

donde

$$\begin{aligned} m_n^* &= \dot{g}_n(\xi_n), & \text{si } n \in I^{ng} \\ &= \delta_j(\zeta_n) \dot{g}_n(\xi_n), & \text{si } n \in I^g \text{ y } z_n \in (c_j, c_{j+1}]. \end{aligned}$$

La acotación de las derivadas de  $g_n$  y  $\delta_j$  garantiza la acotación de la variable  $m_n^*$ .

Por tanto,

$$0 < \bar{\varepsilon} \leq m_n^* \leq H < \infty,$$

siendo  $H = \max\{M, M^2\}$  y  $\bar{\varepsilon} = \min\{\varepsilon, \varepsilon^2\}$ . El algoritmo propuesto se puede escribir, pues, en la forma

$$a_{n+1} = a_n - \beta_n m_n^*(a_n - a) + \beta_n \varepsilon_n(a).$$

Se sigue que

$$\begin{aligned} (a_{n+1} - a)^2 &= (a_n - a)^2 + \beta_n^2 m_n^{*2} (a_n - a)^2 + \beta_n^2 \varepsilon_n^2(a) \\ &\quad - 2\beta_n m_n^* (a_n - a)^2 + \beta_n \varepsilon_n(a) (a_n - a) \\ &\quad - \beta_n^2 m_n^* \varepsilon_n(a) (a_n - a). \end{aligned}$$

Tomando esperanzas resulta

$$\begin{aligned} E((a_{n+1} - a)^2) &= E((a_n - a)^2) - E(\beta_n m_n^* (2 - \beta_n m_n^*) (a_n - a)^2) \\ &\quad + E(\beta_n^2 \varepsilon_n^2(a)), \end{aligned}$$

puesto que  $E(\varepsilon_n(a)) = 0$  y  $\varepsilon_n(a)$  es independiente de  $a_n$ . Como  $m_n^* \leq H$ , se

puede tomar  $\beta_n \leq \frac{2}{H}$ , para todo  $n \in \mathcal{N}$ , con lo cual  $2 - m_n^* \beta_n > 0$ . Como  $m_n^* \beta_n \leq H \beta_n \rightarrow 0$ , se puede considerar que  $2 - \beta_n m_n^* \geq 1$ , a partir de un  $n_0$  en adelante. Además, se sabe que  $m_n^* \geq \bar{\epsilon}$ , con lo cual se concluye que

$$\beta_n m_n^* (2 - \beta_n m_n^*) (a_n - a)^2 \geq \beta_n m_n^* (a_n - a)^2 \geq \bar{\epsilon} \beta_n (a_n - a)^2,$$

para todo  $n$  desde un  $n_0$  en adelante. Si llamamos  $b_n = E \left( (a_n - a)^2 \right)$ , después de iterar  $n$  veces la expresión anterior se obtiene

$$\begin{aligned} b_{n+1} &= b_{n_0} - \sum_{i=n_0}^n E \left( \beta_i m_i^* (2 - \beta_i m_i^*) (a_i - a)^2 \right) \\ &\quad + \sum_{i=n_0}^n E \left( \beta_i^2 \epsilon_i^2 (a) \right) \\ &\leq b_{n_0} - \sum_{i=n_0}^n \bar{\epsilon} \beta_i E \left( (a_i - a)^2 \right) + \sum_{i=n_0}^n \beta_i^2 E \left( \epsilon_i^2 (a) \right). \end{aligned}$$

Se afirma que  $E \left( \epsilon_i^2 (a) \right) \leq 1 + \sigma^2$ . En efecto, por una parte

$$E \left( \epsilon_i^2 (a) \mid i \in I^{ng} \right) = Var \left( \nu \right),$$

y, por otra,

$$\begin{aligned} E \left( \epsilon_i^2 (a) \mid i \in I^g \right) &= \sum_{j=1}^r \Pr \left( z_i \in (c_j, c_{j+1}] \right) E \left( \delta_j^* (a)^2 \mid i \in I^g \right) \\ &= \sum_{j=1}^r \Pr \left( z_i \in (c_j, c_{j+1}] \right) \left( Var \left( \delta_j^* (a) \right) + \delta_j (a)^2 \right) \\ &\leq \sum_{j=1}^r \Pr \left( z_i \in (c_j, c_{j+1}] \right) \left( \sigma^2 + \delta_j (a)^2 \right) \\ &\leq \sigma^2 + 1. \end{aligned}$$

Teniendo en cuenta que  $b_{n+1} \geq 0$ , se puede deducir que

$$\begin{aligned} \sum_{i=n_0}^n \bar{\epsilon} \beta_i b_i &\leq b_{n_0} + \sum_{i=n_0}^n \beta_i^2 E \left( \epsilon_i^2 (a) \right) \\ &\leq b_{n_0} + (1 + \sigma^2) \sum_{i=n_0}^n \beta_i^2 \end{aligned}$$

$$< b_{n_0} + (1 + \sigma^2) \sum_{i=n_0}^{\infty} \beta_i^2 < \infty.$$

Esto significa, en conclusión, que

$$\sum_{i=n_0}^{\infty} \beta_i b_i < \infty$$

lo cual, junto a que  $\sum_{i=n_0}^{\infty} \beta_i = \infty$ , implica que

$$b_n = E \left( (a_n - a)^2 \right) \xrightarrow{n \rightarrow \infty} 0,$$

es decir,  $a_n \xrightarrow{L_2} a$ .  $\square$

A continuación, se verá que las condiciones de derivabilidad no son obligadas.

Lo esencial es que  $g_n$  se encuentre encerrada en un cono, al igual que las  $\delta_j$ .

**Corolario 7.2** *Supongamos que existen  $M < \infty$  y  $\varepsilon > 0$  tales que para cada  $a_n \neq a$  es*

$$\varepsilon \leq \frac{g_n(a_n) - g_n(a)}{a_n - a} \leq M$$

y, para cada  $j = 0, \dots, r - 1$  y  $\beta_1 \neq \beta_2$  es

$$\varepsilon \leq \frac{\delta_j(\beta_1) - \delta_j(\beta_2)}{\beta_1 - \beta_2} \leq M.$$

En estas condiciones, si  $\sum \beta_n = \infty$  y  $\sum \beta_n^2 < \infty$  se cumple que  $a_n \xrightarrow{L_2} a$ .

DEMOSTRACION: Basta observar que se siguen verificando ecuaciones similares a (53) y (54), incluso aunque  $g_n$  y  $\delta_j$  no sean derivables. En este caso, los anteriores valores  $\dot{g}_n(\xi_n)$  y  $\dot{\delta}_j(\zeta_n)$ , representan los cocientes de incrementos del enunciado en lugar de derivadas.  $\square$

OBSERVACION: La condición impuesta en este corolario sobre la función  $g_n$  significa que existe un cono con vértice en el punto  $(a, g_n(a))$ , conteniendo al grafo de la función  $g_n$ . Las condiciones  $\varepsilon > 0$  y  $M < \infty$  implican que dicho cono está generado por dos vectores de coordenadas estrictamente positivas. Un comentario semejante afecta a las funciones  $\delta_j$ .

Finalmente, en el siguiente corolario se relajan las condiciones impuestas de

acotación uniforme.

**Corolario 7.3** *Supongamos que existen dos sucesiones  $\{M_n\}$  y  $\{\varepsilon_n\}$  tales que para cada  $a_n \neq a$  se cumple*

$$0 < \varepsilon_n \leq \frac{g_n(a_n) - g_n(a)}{a_n - a} \leq M_n < \infty$$

y, para cada  $j = 0, \dots, r - 1$  y  $\beta_1 \neq \beta_2$  es

$$0 < \varepsilon_n \leq \frac{\delta_j(\beta_1) - \delta_j(\beta_2)}{\beta_1 - \beta_2} \leq M_n < \infty.$$

*Supongamos, además, que para cada  $n \in \mathcal{N}$ ,  $M_n \beta_n < \varepsilon < 2$  y  $\sum \varepsilon_n \beta_n = \infty$ . En estas condiciones, si  $\sum \beta_n = \infty$  y  $\sum \beta_n^2 < \infty$  se cumple que  $a_n \xrightarrow{L_2} a$ .*

DEMOSTRACION: Similar a la del corolario 7.2.  $\square$

Es de resaltar que, en el algoritmo (52), la dirección a lo largo de la cual se mueve la actualización  $a_{n+1}$  a partir de  $a_n$ , depende solamente del signo de  $y_n(a_n) - g_n(a_n) = \varepsilon_n(a_n)$ . Después, en el teorema 7.1, se vio que el algoritmo converge si  $g_n$  es creciente. Si  $g_n$  fuera decreciente se podría cambiar  $\varepsilon_n(a_n)$  por  $-\varepsilon_n(a_n)$ , para llegar a obtener convergencia de igual forma. En resumen, si  $g_n$  fueran monótonas, bastaría multiplicar el error  $\varepsilon_n(a_n)$  por el signo de la derivada de  $g_n$  (e, incluso, por la misma derivada si ésta es acotada) para poder concluir que el algoritmo converge. Esta idea será fundamental en el caso multidimensional, ya que en este caso se debe buscar, no sólo un signo, sino una dirección de cambio adecuada.

A continuación, se verá un resultado que será útil más adelante, a la hora de hablar del caso multidimensional. Lo importante aquí es que, aunque las hipótesis son menos restrictivas que las del teorema 7.1, se obtiene una conclusión, no idéntica, pero en muchos casos suficiente, similar a las que se conseguirán en el caso multidimensional.

**Teorema 7.4** *Supongamos que  $g_n(\cdot)$  y  $\delta_j(\cdot)$  son derivables y existen  $\varepsilon > 0$  y  $M < \infty$  tales que*

$$0 \leq \dot{g}_n(\cdot) \leq M$$

7. Análisis de otras situaciones colaterales

y

En estas condiciones, si  $\varepsilon \leq \delta_j(\cdot) \leq M$ ,  
 si  $\sum \beta_n = \infty$  y  $\sum \beta_n^2 < \infty$  se cumple que  $g_n(a_n) - g_n(a) \xrightarrow{L_2} 0$ .

DEMOSTRACION: De la misma forma que en el teorema 7.1 se puede escribir

$$a_{n+1} = a_n - \beta_n m_n^* (g_n(a_n) - g_n(a)) + \beta_n \varepsilon_n(a),$$

donde

$$\begin{aligned} m_n^* &= 1, & \text{si } n \in I^{ng} \\ &= \delta_j(\zeta_n), & \text{si } n \in I^g \text{ y } z_n \in (c_j, c_{j+1}]. \end{aligned}$$

Por la acotación de la derivada de  $\delta_j$ , se tiene una acotación equivalente para  $m_n^*$  que, en concreto, se pondrá como

$$0 < \bar{\varepsilon} \leq m_n^* \leq H < \infty.$$

Restando  $a$ , elevando al cuadrado y tomando esperanzas resulta que

$$\begin{aligned} E\left((a_{n+1} - a)^2\right) &= E\left((a_n - a)^2\right) + E\left(\beta_n^2 \varepsilon_n^2(a)\right) \\ &\quad + E\left(\beta_n^2 m_n^{*2} (g_n(a_n) - g_n(a))^2\right) \\ &\quad - E\left(2\beta_n m_n^* (g_n(a_n) - g_n(a))(a_n - a)\right). \end{aligned}$$

Aplicando el teorema del valor medio, existe un valor  $\xi_n \in [a_n, a]$  tal que

$$g_n(a_n) - g_n(a) = g_n'(\xi_n)(a_n - a),$$

de donde, al ser la derivada de  $g_n$  positiva,

$$(g_n(a_n) - g_n(a))(a_n - a) \geq \frac{1}{M} (g_n(a_n) - g_n(a))^2.$$

Se sigue de esta desigualdad que

$$\begin{aligned} E\left((a_{n+1} - a)^2\right) &\leq E\left((a_n - a)^2\right) + E\left(\beta_n^2 \varepsilon_n^2(a)\right) \\ &\quad - E\left(\beta_n m_n^* \left(\frac{2}{M} - \beta_n m_n^*\right) (g_n(a_n) - g_n(a))^2\right). \end{aligned}$$

Tomando  $n \geq n_0$  se puede conseguir que

$$\beta_n m_n^* \left( \frac{2}{M} - \beta_n m_n^* \right) \geq \frac{\bar{\varepsilon}}{M} \beta_n,$$

con lo cual, iterando  $n - n_0$  veces el proceso anterior, se llega a

$$\begin{aligned} E \left( (a_{n+1} - a)^2 \right) &\leq E \left( (a_{n_0} - a)^2 \right) - \sum_{i=n_0}^n \frac{\bar{\varepsilon}}{M} \beta_i E \left( (g_i(a_i) - g_i(a))^2 \right) \\ &\quad + \sum_{i=n_0}^n E \left( \beta_i^2 \varepsilon_i^2(a) \right) \\ &\leq E \left( (a_{n_0} - a)^2 \right) - \sum_{i=n_0}^n \frac{\bar{\varepsilon}}{M} \beta_i E \left( (g_i(a_i) - g_i(a))^2 \right) \\ &\quad + (1 + \sigma^2) \sum_{i=n_0}^n \beta_i^2. \end{aligned}$$

Teniendo en cuenta que  $E \left( (a_{n+1} - a)^2 \right) \geq 0$ , se puede deducir que

$$\frac{\bar{\varepsilon}}{M} \sum_{i=n_0}^n \beta_i E \left( (g_i(a_i) - g_i(a))^2 \right) \leq E \left( (a_{n_0} - a)^2 \right) + (1 + \sigma^2) \sum_{i=n_0}^n \beta_i^2 < \infty,$$

lo cual implica que

$$E \left( (g_n(a_n) - g_n(a))^2 \right) \xrightarrow{i \rightarrow \infty} 0,$$

como se quería demostrar.  $\square$

El próximo resultado también es una consecuencia del teorema 7.1 anterior. Las condiciones de monotonía impuestas sobre las  $g_n$  se pueden modificar, exigiendo sólo monotonía lateral a la izquierda y a la derecha de  $a$ , aunque no necesariamente del mismo tipo (en dichos lados y en las sucesivas  $g_n$ ). Puesto que, como se ha indicado, el signo de los errores es determinante, se propondrá el siguiente algoritmo de estimación bajo derivabilidad de las  $g_n$  (ya comentado tras la demostración del corolario 7.3).

INICIALIZACIÓN: Tómese un valor real arbitrario  $a_1$ .

ITERACIÓN: Actualícese la estimación  $a_n$  mediante la ecuación en diferencias

$$a_{n+1} = a_n + \beta_n \dot{g}_n(a_n) (a_n - a),$$

siendo  $\dot{g}_n(a_n)$  la derivada de  $g_n$  evaluada en el punto  $a_n$  (positiva o negativa, dependiendo del tipo de monotonía).

**Teorema 7.5** *Supongamos que las  $\delta_j(\cdot)$  son derivables y que existen  $\varepsilon > 0$  y  $M < \infty$  tales que*

$$\varepsilon \leq \dot{\delta}_j(\cdot) \leq M.$$

*Adicionalmente, supongamos que las  $g_n(\cdot)$  son monótonas y derivables, verificando*

$$\varepsilon \leq \left| \dot{g}_n(\cdot) \right| \leq M.$$

*En estas condiciones, si  $\sum \beta_n = \infty$  y  $\sum \beta_n^2 < \infty$  se cumple que  $a_n \xrightarrow{L_2} a$ .*

DEMOSTRACION: Sigue la misma línea argumental de la demostración de los últimos resultados.  $\square$

Las condiciones de acotación de las derivadas impuestas implican que  $g_n$  es monótona decreciente o creciente en todo  $\mathcal{R}$ . Una monotonía lateral distinta a la derecha e izquierda de  $a$  podría ser posible sustituyendo las condiciones de derivabilidad por condiciones como, en la línea de los corolarios 7.2 y 7.3. En este caso, la derivada  $\dot{g}_n(a_n)$  en el algoritmo debería sustituirse simplemente por un signo, positivo o negativo dependiendo de si  $g_n(a_n)$  está por encima o por debajo de  $g_n(a)$ . De esta forma, la convergencia establecida en el teorema 7.5 se mantendría. En todo caso, la aplicabilidad de esta última formulación exigiría conocer  $g_n(a)$ , siendo  $a$  desconocido. Aunque se podría disponer de esta información en algún caso, se ha omitido su formulación explícita por considerar que tiene un interés más bien referencial.

### 7.2.1 Un modelo de censura alternativo

Los resultados obtenidos se mantienen bajo otras muchas circunstancias. A título de ejemplo, se planteará a continuación una situación de censura distinta a la considerada hasta ahora. Esta nueva censura se referenciará en lo sucesivo como *modelo alternativo*, frente al *modelo original* que se ha venido utilizando hasta

ahora en esta memoria.

El modelo alternativo de censura citado, asumirá que los datos no serán observados siempre que caigan dentro de un cierto intervalo  $C$  fijo. Esto implica que la muestra observada está truncada en el conjunto complementario de  $C$ . En este caso, el mecanismo de pérdida no es ignorable, no pudiéndose utilizar solamente los datos no censurados. Este nuevo tipo de censura difiere del original, puesto que en este último la muestra no agrupada no era, en absoluto, truncada.

En resumen, si  $C = [a, b]$  es un intervalo fijado, se considera el modelo no lineal

$$z_n = g_n(a) + \nu_n$$

con las mismas características (excepto el tipo de censura) expuestas al comienzo de la sección 7.2. Como se ha indicado, si  $z_n \notin C$  se observa su valor exacto, a diferencia de cuando  $z_n \in C$ . De nuevo, los índices naturales se particionan en  $I^g$  e  $I^{ng}$ , conteniendo los índices censurados y no censurados, respectivamente.

El método iterativo que se propone es el siguiente:

INICIALIZACIÓN: Tómese un  $a_1$  inicial.

ITERACIÓN: Actualícese la estimación  $a_n$  de la etapa  $n$  mediante

$$a_{n+1} = a_n + \beta_n (y_n(a_n) - g_n(a_n)),$$

siendo

$$\begin{aligned} y_n(a_n) &= z_n, && \text{si } z_n \notin C \ (n \in I^{ng}) \\ &= g_n(a_n) + \delta(-g_n(a_n)), && \text{si } z_n \in C \ (n \in I^g) \end{aligned}$$

y

$$\delta(\beta) = E(\nu | \nu \in [\beta + a, \beta + b]).$$

Como antes, también ahora podría sustituirse  $\delta$  por una corrección aleatoria  $\delta^*$ , con media  $\delta$  y varianza acotada. Por simplicidad, se utilizará solamente  $\delta$  determinista. Obsérvese que la forma de evaluar las iteraciones de este algoritmo se inspira en



(52), si bien la situación aquí tratada es formalmente diferente.

**Teorema 7.6** *Asúmase que las funciones reales  $g_n(\cdot)$  y  $\delta(\cdot)$  admiten derivadas y éstas son uniformemente acotadas en una banda, es decir, existen  $\varepsilon > 0$  y  $M < \infty$  tales que*

$$\varepsilon \leq \dot{g}_n(\cdot) \leq M$$

y

$$\varepsilon \leq \dot{\delta}(\cdot) \leq M.$$

*Si, además, se toman tamaños de paso de forma que  $\sum \beta_n = \infty$  y  $\sum \beta_n^2 < \infty$ , entonces se verifica que  $a_n \xrightarrow{L_2} a$ .*

DEMOSTRACION: Puede emplearse un razonamiento similar al del teorema 7.1. Para ello, obsérvese que se cumplen dos propiedades análogas a las vistas en el citado teorema, aunque sus respectivas demostraciones difieren. En primer lugar, si  $\varepsilon_n(a) = y_n(a) - g_n(a)$ , resulta que

$$\begin{aligned} E(\varepsilon_n(a)) &= E(z_n - g_n(a) | z_n \notin C) \Pr(z_n \notin C) + \delta(-g_n(a)) \Pr(z_n \in C) \\ &= E(\nu_n | \nu_n \notin -g_n(a) + C) \Pr(\nu_n \notin -g_n(a) + C) \\ &\quad + E(\nu | \nu \in -g_n(a) + C) \Pr(\nu_n \in -g_n(a) + C) \\ &= 0. \end{aligned}$$

En segundo lugar, también se cumple que

$$\begin{aligned} E(\varepsilon_n^2(a)) &= E(\nu_n^2 | \nu_n \notin -g_n(a) + C) \Pr(\nu_n \notin -g_n(a) + C) \\ &\quad + (E(\nu | \nu \in -g_n(a) + C))^2 \Pr(\nu_n \in -g_n(a) + C) \\ &\leq E(\nu_n^2 | \nu_n \notin -g_n(a) + C) \Pr(\nu_n \notin -g_n(a) + C) \\ &\quad + E(\nu^2 | \nu \in -g_n(a) + C) \Pr(\nu_n \in -g_n(a) + C) \\ &= \text{Var}(\nu). \end{aligned}$$

Estas condiciones son suficientes para concluir que

$$E(a_n - a)^2 \xrightarrow{n \rightarrow \infty} 0. \square$$

**Corolario 7.7** *(Relajación de la diferenciabilidad) Supongamos que existen dos*

valores  $M < \infty$  y  $\varepsilon > 0$  cumpliéndose que, para cada  $a_n \neq a$ ,

$$\bar{\varepsilon} \leq \frac{g_n(a_n) - g_n(a)}{a_n - a} \leq M$$

y, para cada  $\beta_1 \neq \beta_2$ ,

$$\bar{\varepsilon} \leq \frac{\delta(\beta_1) - \delta(\beta_2)}{\beta_1 - \beta_2} \leq M.$$

En estas condiciones, si  $\sum \beta_n = \infty$  y  $\sum \beta_n^2 < \infty$ , se sigue verificando la conclusión del teorema anterior, es decir,  $a_n \xrightarrow{L_2} a$ .

DEMOSTRACION: Procédase como en el corolario 7.2.  $\square$

Podría incluso enunciarse un corolario que sustituyera los valores fijos  $M$  y  $\varepsilon$  por otras variables en cada iteración  $M_n$  y  $\varepsilon_n$ , en términos similares a los del corolario 7.3, si bien no será enunciado.

### 7.3 Estimación paramétrica

A continuación se verá cómo las aproximaciones estocásticas vistas en la sección anterior pueden aplicarse a problemas de estimación paramétrica. Si las funciones de regresión  $g_n$  son constantes (es decir,  $g_n = g$ , para cada  $n \in \mathcal{N}$ ), el modelo no lineal visto se convierte en

$$z_n = g(\theta) + \nu_n, \quad (55)$$

propio de un problema de estimación paramétrica (se ha cambiado el parámetro  $a$  por  $\theta$ ). Estos problemas de estimación en presencia de censura pueden, por tanto, ser abordados mediante las técnicas de aproximaciones estocásticas. Por simplicidad, se considerará un modelo paramétrico unidimensional, si bien no necesariamente  $g(\theta)$  tiene que ser una reparametrización de posición, como aparece en (55).

#### 7.3.1 Modelo de censura alternativo

Sea  $\mathcal{F} = \{F_\theta : \theta \in \Theta \subset \mathcal{R}\}$  una familia paramétrica de funciones de distribución.

Se comenzará asumiendo un modelo de censura similar al de la sección 7.2.1, si bien, después, se considerará el tipo de censura original. Así pues, existe un conjunto fijo de censura  $C$  (no necesariamente un intervalo) perteneciente a la  $\sigma$ -álgebra de Borel en  $\mathcal{R}$ . Las variables de observación  $\{Z_n\}$  son independientes e idénticamente distribuidas con función de distribución  $F_\theta \in \mathcal{F}$ , para un valor  $\theta \in \Theta$  desconocido. Se asumirá que esta muestra es solamente observable fuera de  $C$ . (es decir, si  $Z_n \notin C$ ), no siéndolo si  $Z_n \in C$ . Además, se supondrá que conocemos la función  $g : \mathcal{R} \rightarrow \mathcal{R}$ , siendo

$$g(\theta) = E(Z_n).$$

El algoritmo que se propone (en la línea de las aproximaciones estocásticas) para estimar  $\theta$  es el siguiente.

INICIALIZACIÓN: Tómese un valor inicial  $\theta_1 \in \Theta$ .

ITERACIÓN: La estimación en la etapa  $n + 1$  se calculará a partir de la estimación de la etapa  $n$  y de la información que proporciona únicamente el valor muestral  $Z_n$ , mediante la ecuación recurrente

$$\theta_{n+1} = \theta_n + \varepsilon_n (Z_n - g(\theta_n)) I(Z_n \notin C) + \varepsilon_n (Z_n(\theta_n) - g(\theta_n)) I(Z_n \in C),$$

siendo  $Z_n(\theta_n)$  es una variable aleatoria con distribución  $F_{\theta_n}$  truncada en  $C$ . El proceso iterativo se puede escribir como

$$\theta_{n+1} = \theta_n + \varepsilon_n (y_n(\theta_n) - g(\theta_n)), \quad (56)$$

donde

$$\begin{aligned} y_n(\theta_n) &= Z_n, & \text{si } n \in I^{ng} \\ &= Z_n(\theta_n), & \text{si } n \in I^g, \end{aligned}$$

teniendo  $I^g$  e  $I^{ng}$  el mismo significado asumido en todas las secciones anteriores.

Si  $\Theta \neq \mathcal{R}$ , es posible que la aplicación de (56) no garantice que  $\theta_{n+1} \in \Theta$ , cuando  $\theta_n \in \Theta$ . En este caso, sería necesario llevar a cabo un proceso  $\Pi_\Theta$  de

proyección sobre  $\Theta$ , convirtiéndose (56) en

$$\theta_{n+1} = \Pi_{\Theta} (\theta_n + \varepsilon_n (y_n (\theta_n) - g (\theta_n))).$$

Sin embargo, no se estudiará esta situación y se supondrá que el algoritmo está bien definido, siendo (56) la ecuación que determina su iteración.

La citada aproximación estocástica permite garantizar la convergencia casi seguro, en lugar de la convergencia en  $L_2$ , como hasta ahora. Para ver que  $\theta_n \xrightarrow{c.s.} \theta$  interesa calcular la esperanza de la variable aleatoria  $y_n (\theta_n)$ , condicionada por  $\theta_n$  y por todas las realizaciones anteriores, para analizar su comportamiento como función de  $\theta_n$ . Esta esperanza es:

$$E (y_n (\theta_n) | \theta_n) = \int_{C^c} x dF_{\theta}(x) + \Pr (Z_n \in C) E (Z_n (\theta_n) | \theta_n).$$

Definanse los valores  $A = \int_{C^c} x dF_{\theta}(x)$  y  $B = \Pr (Z_1 \in C)$ , ambos constantes, y  $h(\theta_n) = E (Z_n (\theta_n) | \theta_n)$  como función en  $\theta_n$ . Si  $\theta_n = \theta$ , es claro que  $A + Bh(\theta) = g(\theta)$ , pues

$$\begin{aligned} A + Bg(\theta) &= \int_{C^c} x dF_{\theta}(x) + \Pr (Z_1 \in C) E (Z_n (\theta) | \theta) \\ &= \int_{C^c} x dF_{\theta}(x) + \int_C x dF_{\theta}(x) = g(\theta). \end{aligned}$$

Supóngase, por último, que  $A + Bh(\cdot) - g(\cdot)$  es continua y que tiene un único cero en  $\theta$  (verdadero valor del parámetro). Posteriormente se verán distintas situaciones donde estas condiciones son factibles. Se tiene entonces que

$$\underset{\theta_n < \theta}{\text{signo}} (A + Bh(\theta_n) - g(\theta_n)) = - \underset{\theta_n > \theta}{\text{signo}} (A + Bh(\theta_n) - g(\theta_n)).$$

En consecuencia, la ecuación diferencial

$$\frac{d}{dt} \theta(t) = A + Bh(\theta(t)) - g(\theta(t))$$

tiene, en general, una única solución constante que corresponde a  $\theta(t) = \theta$ . Si este punto  $\theta$  es una solución asintóticamente estable para la ecuación diferencial anterior y se seleccionan los tamaños de paso adecuadamente, se puede conseguir que  $\theta_n \xrightarrow{c.s.} \theta$  (véase KUSHNER Y YIN (1997), capítulo 5). Si  $\theta$  es una solución

asintóticamente inestable para la ecuación diferencial es suficiente con cambiar  $\varepsilon_n$  por  $-\varepsilon_n$  para que  $\theta_n \xrightarrow{c.s.} \theta$ .

Se verá, finalmente, que las condiciones señaladas sobre continuidad y cero único en  $\theta$  de la función  $A + Bh(\cdot) - g(\cdot)$  son casi coincidentes, en muchos casos, con las hipótesis impuestas para garantizar los resultados de convergencia de la sección anterior. Con este objetivo, se comenzará analizando la función en  $\theta_n$ .

$$h(\theta_n) = E(Z_n(\theta_n) | \theta_n) = \frac{\int_C x dF_{\theta_n}(x)}{\int_C dF_{\theta_n}(x)}$$

en ciertos ejemplos y casos particulares.

**Ejemplos** i) Sea  $C = (c, \infty)$ , para un  $c > 0$ , y asúmase que las  $F_\theta$  son distribuciones exponenciales de parámetro  $\theta \in \Theta = (0, \infty)$ . Es claro que  $g(\theta) = E(Z_\theta) = \theta$ , donde  $Z_\theta$  es una variable aleatoria distribuida como  $F_\theta$ . Se puede calcular fácilmente que  $h(\theta) = E(Z_\theta | Z_\theta > c) = \theta + c$ , con lo cual  $A + Bh(\theta_n) - \theta_n$ , como función  $\theta_n$ , es continua y tiene un único cero si, y sólo si,  $B = \Pr(Z_\theta > c) \neq 1$ , lo cual siempre ocurre por ser  $c > 0$ . El hecho de que  $B < 1$  significa que no todos los datos son censurados. Además, la ecuación diferencial citada arriba adopta la forma

$$\frac{d}{dt}\theta(t) = (A + Bc) + (B - 1)\theta(t).$$

Dicha ecuación es asintóticamente estable por ser  $B - 1 < 0$ , concluyéndose que  $\theta_n \xrightarrow{c.s.} \theta$ .

ii) Sea  $Z_\alpha$  una variable distribuida como una Beta( $\alpha, 1$ ) y  $C = (c, 1)$ , con  $\alpha > 0$  y  $0 < c < 1$ . Reparametricemos la distribución beta mediante  $\theta = \alpha(1 - \alpha)^{-1}$ ,  $0 < \theta < 1$ . La densidad de  $Z_\theta$  puede escribirse como

$$f(x, \theta) = \frac{\theta}{1 - \theta} x^{\frac{2\theta - 1}{1 - \theta}}, \quad 0 < x < 1.$$

Así pues,  $g(\theta) = E(Z_\theta) = \theta$ . Es fácil comprobar que la función

$$h(\theta) = E(Z_\theta | c < Z_\theta < 1) = \theta \frac{1 - c^{\frac{1}{1-\theta}}}{1 - c^{\frac{\theta}{1-\theta}}}, \quad 0 < \theta < 1$$

es derivable con continuidad, cumpliéndose  $0 < \frac{d}{d\theta}h(\theta) < 1$ , para todo  $0 < \theta < 1$ . En consecuencia,  $A + Bh(\theta_n) - \theta_n$  tiene derivada negativa, admitiendo, por ello, una raíz única en  $(0, 1)$ , que coincide con  $\theta_n = \theta$ . Así pues, de nuevo, se verifica que  $\theta_n \xrightarrow{c.s.} \theta$ .

**Caso de parámetro de localización** Obsérvese que en los dos ejemplos anteriores  $g(\theta)$  no era un parámetro de localización del modelo. Supóngase ahora que sí lo fuera, es decir:  $z_n = g(\theta) + \nu_n$ , donde las  $\nu_n$  son variables aleatorias independientes e idénticamente distribuidas. Supongamos, por analogía con el teorema 7.6, que  $g(\theta_n)$  y  $h(\theta_n)$  son derivables, como funciones en  $\theta_n$ . Se cumple, así, que  $h(\theta_n) = g(\theta_n) + \delta(-g(\theta_n))$ , donde  $\delta(-g(\theta_n))$  es la función

$$\delta(-g(\theta_n)) = E(\nu | \nu \in -g(\theta_n) + C),$$

siendo  $-g(\theta_n) + C$  una traslación del conjunto de censura  $C \subset \mathcal{R}$ . Interesa que  $A + (B - 1)g(\theta_n) + B\delta(-g(\theta_n))$  sea una función con una única raíz en  $\theta_n = \theta$ . En particular, esto se cumple si es monótona en  $\theta_n$ .

Si  $\frac{d}{d\theta_n}g(\theta_n) > 0$  (condición que se impone en el teorema 7.6), una condición suficiente para que se cumpla la monotonía citada es que  $\frac{d}{d\theta_n}\delta(\theta_n) > 1 - B^{-1}$ , en cuyo caso la solución es asintóticamente estable y se verifica que  $\theta_n \xrightarrow{c.s.} \theta$ . En la situación degenerada  $B = 0$  la condición de monotonía resulta inmediata, si bien aquélla indica que la probabilidad de censura es cero, siendo la muestra completamente observada con probabilidad uno.

Si  $\frac{d}{d\theta_n}g(\theta_n) < 0$ , la condición  $\frac{d}{d\theta_n}\delta(\theta_n) > 1 - B^{-1}$  implica que la solución sea asintóticamente inestable, por lo cual se debe cambiar el signo del tamaño del paso para obtener convergencia.

En cualquiera de los dos casos, si  $0 < B < 1$ , resulta que una condición

suficiente para la convergencia casi seguro citada es que  $\frac{d}{d\theta_n} \delta(\theta_n) > 0 \geq 1 - B^{-1}$ , en concordancia con el tipo de hipótesis que fueron impuestas en las secciones precedentes. Hay que hacer notar que si  $B = 1$ , la probabilidad de censura es uno, de donde  $A = 0$  y  $\delta(\theta_n) \equiv 0$ . La ecuación  $A + (B - 1)g(\theta_n) + B\delta(-g(\theta_n)) = 0$  se verifica, pues, para cualquier  $\theta_n$ , no teniendo solución única. En conclusión, no se puede abordar con este método la estimación de  $\theta$  si todos los datos son censurados (con probabilidad uno).

**Situación práctica** A continuación, se expondrá una situación práctica en donde el tipo de censuras aquí tratado es factible.

Supóngase que la duración de una cierta pieza sigue una distribución exponencial de densidad  $f(t) = \lambda e^{-\lambda t}$ , con  $t, \lambda > 0$ . Si cuando una pieza se rompe es sustituida inmediatamente por otra igual, el número  $N_t$  de piezas rotas en el intervalo de tiempo  $[0, t]$  es un proceso de Poisson de parámetro  $\lambda$ . Es conocido que la distribución conjunta de  $N_t$  y de los instantes de ruptura  $T_1, T_2, \dots, T_{N_t}$  hasta el instante  $t$ , sólo depende de  $t$ . En concreto,

$$\begin{aligned} f_{(N_t, T_1, \dots, T_{N_t})}(n, t_1, \dots, t_n) &= \lambda^n e^{-\lambda t} \text{ si } 0 < t_1 < \dots < t_n < t \\ &= 0 \quad \text{en otro caso.} \end{aligned}$$

De lo anterior se deduce que el estimador máximo verosímil de  $\lambda$  es

$$\hat{\lambda} = \frac{N_t}{t},$$

el cual es suficiente si  $t$  es fijo. Además, es fácil ver que  $\frac{N_t}{t} \rightarrow \lambda$  c.s., cuando  $t \rightarrow \infty$ .

Si la sustitución de la pieza rota por una nueva requiere la intervención de un operario, lo natural es que éste no pueda estar todo el tiempo pendiente de la máquina (posiblemente porque tenga otras máquinas que revisar). En consecuencia, una política de mantenimiento plausible podría ser la siguiente. Cuando el operario coloca una pieza nueva se ausenta (tras anotar la hora de

reparación) y regresa al cabo de un tiempo  $a$ , vigilando el funcionamiento de la pieza durante un período de tiempo  $b$ , si aquella no se estropea antes. Por tanto, cuando una pieza deja de funcionar o bien es cambiada instantáneamente (si en el instante de ruptura el operario está presente) o bien existe un tiempo de inoperatividad hasta que el operario regresa. De esta forma éste sólo sabe el tiempo exacto de duración de una pieza si se rompe en su presencia. En otro caso, la citada duración será censurada, por razones obvias.

Ante la política de mantenimiento inmediatamente explicada, denótese por  $R_i$  el instante de ruptura de la  $i$ -ésima pieza y por  $S_i$  el tiempo que transcurre entre  $R_i$  y el instante de reparación. Adicionalmente, el proceso puntual indicando el número de piezas rotas hasta  $t$ , será denotado por  $X_t$ . Es claro que existe una relación entre los anteriores procesos y el de Poisson puro  $N_t$  y sus tiempos  $T_1, T_2, \dots$  asociados, antes citados. De hecho, puede razonarse fácilmente que (con  $S_0 = 0$ )

$$T_{i+1} = R_{i+1} - S_1 - \dots - S_i, \quad i = 0, 1, 2, \dots$$

Después de la ruptura de la pieza  $i + 1$  habrá un tiempo positivo de inoperatividad de la máquina si, y sólo si,

$$T_{i+1} - T_i \in \bigcup_{k=0}^{\infty} (2k(a+b), 2k(a+b) + a) = \bigcup_{k=0}^{\infty} C_k = C.$$

En otro caso, la pieza es sustituida inmediatamente y la máquina no dejará de funcionar en ningún momento. En concreto, en el primero de los casos, si

$$T_{i+1} - T_i \in (2k_{i+1}(a+b), 2k_{i+1}(a+b) + a),$$

entonces

$$S_i = 2k_{i+1}(a+b) + a - (T_{i+1} - T_i) > 0.$$

Finalmente, obsérvese que para  $i = 0, 1, 2, \dots$

$$\begin{aligned} X_t &= N_{t-S_1-\dots-S_i} = i, \text{ si } t \in [T_i + S_1 + \dots + S_i, T_{i+1} + S_1 + \dots + S_i) \\ &= N_{T_{i+1}} = i + 1, \text{ si } t \in [T_{i+1} + S_1 + \dots + S_i, T_{i+1} + S_1 + \dots + S_{i+1}). \end{aligned}$$



7. Análisis de otras situaciones colaterales

Se comprobará que la densidad conjunta de  $X_t$  y de los instantes de ruptura es de la forma siguiente:

$$\begin{aligned} f_{(X_t, R_1, \dots, R_{N_t})}(n, r_1, \dots, r_n) &= \lambda^n e^{-\lambda(t-s)}, \text{ si } 0 < r_1 < \dots < r_n < t \\ &= 0, \quad \text{en otro caso,} \end{aligned}$$

donde  $s$  es el tiempo que la máquina está inoperativa en el intervalo  $[0, t]$ , es decir,

$$\begin{aligned} s &= s_1 + \dots + s_n, & \text{si } t \geq r_n + s_n \\ &= s_1 + \dots + s_{n-1} + t - r_n, & \text{si } t < r_n + s_n. \end{aligned}$$

Para verlo, basta observar que  $S_i$  depende solamente de  $R_1, \dots, R_i$ . Esto significa que  $S$  depende de  $R_1, \dots, R_n$ . En forma expandida se puede escribir, pues,

$$\begin{aligned} T_1 &= R_1 \\ T_2 &= R_2 - S_1(R_1) \\ &\dots \\ T_n &= R_n - S_1(R_1) - \dots - S_{n-1}(R_1, \dots, R_{n-1}). \end{aligned}$$

Se sigue que

$$\begin{aligned} &f_{(X_t, R_1, \dots, R_{N_t})}(n, r_1, \dots, r_n) \\ &= f_{(N_{t-s}, T_1, \dots, T_{N_{t-s}})}(n, r_1, r_2 - s_1, \dots, r_n - s_1 - \dots - s_{n-1}) \\ &= \lambda^n e^{-\lambda(t-s)}, \end{aligned}$$

siempre que  $0 < r_1 < \dots < r_n < t$ .

En conclusión, la distribución conjunta de  $X_t, R_1, \dots, R_{N_t}$  sólo depende de  $t$  y de  $S$ , por lo que un estimador bidimensional suficiente, si se dispusiera de la información completa, sería  $(N_t, S) = (N_t, S(R_1, \dots, R_n))$ . En la citada situación, el estimador de máxima verosimilitud del parámetro  $\lambda$  es

$$\hat{\lambda} = \frac{N_t}{t - S}.$$

Como se ha indicado el valor  $S$  no es observable, porque las piezas se pueden

romper cuando el operario no esté presente. Este hecho invalida la utilización del anterior estimador  $\hat{\lambda}$ . Así pues, el método clásico no sirve para estimar  $\lambda$ .

La estimación se puede abordar, sin embargo, mediante las aproximaciones estocásticas expuestas, considerando las duraciones de las sucesivas piezas como una muestra de una variable aleatoria exponencial eventualmente sometida a censura. Reparametrícese mediante  $\theta = \frac{1}{\lambda}$ , para que  $\theta$  coincida con la esperanza de la exponencial asociada. Fijado  $t$ , se propone la siguiente aproximación estocástica que consta de  $N_t$  pasos, con la cual se proporcionará la estimación de  $\theta$ . Fijado un  $\theta_1$  inicial, en cada uno de los pasos  $n$  se actualizará el valor de  $\theta_n$  mediante

$$\theta_{n+1} = \theta_n + \varepsilon_n (Y_n - \theta_n) I(Y_n \notin C) + \sum_{k=0}^{\infty} \varepsilon_n (Y_{\theta_n}^k - \theta_n) I(Y_n \in C_k),$$

donde  $Y_n = T_{n+1} - T_n$  e  $Y_{\theta_n}^k$  es una realización de una variable aleatoria exponencial de parámetro  $\theta_n$  truncada en  $C_k$ . Es inmediato observar que si  $Y$  es una variable exponencial con densidad  $f(t) = \frac{1}{\theta} e^{-\frac{1}{\theta}t}$ , entonces

$$E(Y|Y \in [a, b]) = \theta + b - \frac{b-a}{1 - e^{-\frac{b-a}{\theta}}}.$$

Esto sugiere proponer una versión del algoritmo anterior con correcciones no aleatorizadas, en la cual el valor aleatorio  $Y_{\theta_n}^k$  se sustituirá por el fijo

$$Y_{\theta_n}^k = \theta_n + (2k(a+b) + a) - \frac{a}{1 - e^{-\frac{a}{\theta_n}}} = h_k(\theta_n).$$

Para analizar la convergencia de la citada versión, obsérvese que  $N_t \rightarrow \infty$  c.s., si  $t \rightarrow \infty$ , por lo que se analizará la convergencia de  $\theta_n$  cuando  $n \rightarrow \infty$ . Para ello, si  $F_\theta$  es la función de distribución de una variable aleatoria exponencial de parámetro  $\lambda = \frac{1}{\theta}$ , resulta que

$$h(\theta_n) = \int_{C^c} x dF_\theta(x) + \sum_{k=0}^{\infty} \int_{C_k} dF_\theta(x) g_k(\theta_n) - \theta_n$$

tiene una única solución en  $\theta_n = \theta$ . Para ello, es suficiente comprobar que  $0 < \frac{d}{d\theta_n} h_k(\theta_n) < 1$ , con lo cual  $\frac{d}{d\theta_n} h(\theta_n) < 0$ . Así pues,  $\theta$  es la única solución

constante de la ecuación diferencial

$$\frac{d}{dt}\theta(t) = h(\theta(t))$$

la cual es asintóticamente estable, por lo que se concluye que  $\theta_n \xrightarrow{c.s.} \theta$ .

### 7.3.2 Modelo de censura original

Por último, se expondrán brevemente las aproximaciones estocásticas para abordar un problema de estimación paramétrica, similar al anterior, cuando el modelo de censura es el original, utilizado en todos los capítulos anteriores. Sea una familia paramétrica de funciones de distribución en  $\mathcal{R}$ ,  $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ , siendo  $\Theta \subseteq \mathcal{R}$ . La experimentación  $\{Z_n\}_{n=1}^\infty$  se identificará con una sucesión de variables aleatorias con función de distribución  $F_\theta \in \mathcal{F}$ , para un valor  $\theta \in \Theta$  desconocido. Se supondrá que la sucesión  $\{Z_n\}$  no es completamente observada, pudiendo algunos valores estar agrupados dentro de intervalos con extremos

$$-\infty = c_0 < c_1 < \dots < c_{r-1} < c_r = \infty.$$

Particionemos el conjunto de índices en  $I^g$  e  $I^{ng}$ , de forma que si  $n \in I^g$  solamente se conoce el intervalo de clasificación que contiene a  $Z_n$ , mientras que si  $n \in I^{ng}$  se observa el valor exacto de  $Z_n$ . Asíumase, por último, que las probabilidades de  $I^{ng}$  e  $I^g$  sobre la población de índices son  $\pi_0$  y  $1 - \pi_0$ . Como en la subsección anterior, se trata de estimar el valor  $\theta \in \Theta$  a partir de la citada experimentación censurada.

Se asumirá que cada distribución  $F_\theta$  tiene esperanza conocida,

$$g(\theta) = \int_{\mathcal{R}} x dF_\theta(x), \text{ para cada } \theta \in \Theta.$$

La aproximación estocástica que se propone para abordar el problema de estimación citado es la siguiente.

INICIALIZACIÓN: Tómesese  $\theta_1 \in \Theta$  como valor inicial arbitrario.

ITERACIÓN: Partiendo de la estimación  $\theta_n$ , actualícese ésta mediante

$$\theta_{n+1} = \theta_n + \varepsilon_n (y_n(\theta_n) - g(\theta_n)).$$

En esta ecuación recursiva

$$\begin{aligned} y_n(\theta_n) &= Z_n, & \text{si } n \in I^{ng} \\ &= Z_n^j(\theta_n), & \text{si } n \in I^g, Z_n \in (c_j, c_{j+1}] \end{aligned}$$

y  $Z_n^j(\theta_n)$  es cualquier realización de una variable aleatoria distribuida como  $F_{\theta_n}$  truncada sobre el intervalo  $(c_j, c_{j+1}]$ . Como en la sección anterior, se supondrá que el algoritmo está bien definido, obteniéndose en cada iteración  $n$  un valor  $\theta_n \in \Theta$ .

La convergencia a  $\theta$  del algoritmo planteado puede analizarse, como antes, a través del comportamiento de la esperanza de  $y_n(\theta_n)$ , condicionada por todas las realizaciones anteriores. Teniendo en cuenta que el valor en la etapa  $n$  sólo depende del valor en la etapa anterior y de la última realización, dicha esperanza es:

$$G(\theta_n) = E(y_n(\theta_n) | \theta_n) = g(\theta) \pi_0 + (1 - \pi_0) \sum_{j=0}^{r-1} \Pr(Z_n \in (c_j, c_{j+1}]) h_j(\theta_n),$$

siendo

$$h_j(\theta_n) = E(Z_n^j(\theta_n) | \theta_n) = \frac{\int_{c_j}^{c_{j+1}} x dF_{\theta_n}(x)}{\int_{c_j}^{c_{j+1}} dF_{\theta_n}(x)}.$$

Es claro que, si  $\theta_n = \theta$  resulta

$$\begin{aligned} G(\theta) &= g(\theta) \pi_0 + (1 - \pi_0) \sum_{j=0}^{r-1} \Pr(Z_1 \in (c_j, c_{j+1}]) h_j(\theta) \\ &= g(\theta) \pi_0 + (1 - \pi_0) \sum_{j=0}^{r-1} \int_{c_j}^{c_{j+1}} x dF_{\theta}(x) \\ &= g(\theta) \pi_0 + (1 - \pi_0) g(\theta) = g(\theta). \end{aligned}$$

Supóngase que la ecuación

$$G(\theta_n) - g(\theta_n) = g(\theta) \pi_0 + (1 - \pi_0) \sum_{j=0}^{r-1} \Pr(Z_1 \in (c_j, c_{j+1}]) h_j(\theta_n) - g(\theta_n) = 0$$

tiene una única solución en  $\theta_n = \theta$ . Bajo esta hipótesis, la ecuación diferencial

$$\frac{\partial}{\partial t} \theta(t) = G(\theta(t)) - g(\theta(t))$$

tiene una única solución constante, la cual es  $\theta(t) = \theta$ . Si esta solución es asintóticamente estable entonces  $\theta_n \xrightarrow{c.s.} \theta$  (véase de nuevo KUSHNER Y YIN (1997), capítulo 5). Si es asintóticamente inestable es suficiente con tomar tamaños de pasos negativos, es decir, cambiar  $\varepsilon_n$  por  $-\varepsilon_n$ , para conseguir que  $\theta_n \xrightarrow{c.s.} \theta$ . El siguiente ejemplo puede ser clarificador.

**Caso de parámetro de localización** Supóngase que  $g(\theta)$  es un parámetro de posición, cumpliéndose que

$$Z_n = g(\theta) + \nu_n,$$

donde  $\nu_n$  son variables aleatorias independientes e idénticamente distribuidas, con media cero y varianza finita. Las funciones definidas previamente vienen dadas por

$$h_j(\theta_n) = g(\theta_n) + \delta_j(-g(\theta_n)),$$

donde, para cada valor  $\beta \in \mathcal{R}$ , se define

$$\delta_j(\beta) = E(\nu | \nu \in (\beta + c_j, \beta + c_{j+1}]).$$

Denótese, además,  $B_j = \Pr(Z_1 \in (c_j, c_{j+1}])$ . En esta situación, se tiene que

$$\begin{aligned} G(\theta_n) - g(\theta_n) &= \pi_0 g(\theta) + (1 - \pi_0) \sum_{j=0}^{r-1} B_j (g(\theta_n) + \delta_j(-g(\theta_n))) - g(\theta_n) \\ &= \pi_0 g(\theta) - \pi_0 g(\theta_n) + \sum_{j=0}^{r-1} B_j \delta_j(-g(\theta_n)). \end{aligned}$$

Si  $\delta_j(\cdot)$  y  $g(\cdot)$  son funciones derivables

$$\frac{\partial}{\partial \theta_n} (G(\theta_n) - g(\theta_n)) = -\pi_0 \frac{\partial g(\theta_n)}{\partial \theta_n} - \sum_{j=0}^{r-1} B_j \frac{\partial g(\theta_n)}{\partial \theta_n} \frac{\partial \delta_j(\beta)}{\partial \beta} \Big|_{\beta=-g(\theta_n)}.$$

Suponiendo, por ejemplo, que  $\frac{\partial g(\beta)}{\partial \beta} > 0$  y  $\frac{\partial \delta_j(\beta)}{\partial \beta} > 0$ , para cada  $\beta \in \mathcal{R}$ , se

verifica que  $G(\theta_n) - g(\theta_n)$  es estrictamente monótona decreciente, con lo que la solución  $\theta_n = \theta$  es asintóticamente estable de la ecuación diferencial

$$\frac{d}{dt}\theta(t) = G(\theta(t)) - g(\theta(t)). \quad (57)$$

Se sigue, por tanto, que  $\theta_n \xrightarrow{c.s.} \theta$ .

Si, por el contrario, se supone que  $\frac{\partial g(\beta)}{\partial \beta} < 0$  y  $\frac{\partial \delta_i(\beta)}{\partial \beta} > 0$ , para cada  $\beta \in \mathcal{R}$ , se verifica que  $G(\theta_n) - g(\theta_n)$  es estrictamente creciente, por lo que  $\theta_n = \theta$  es una solución de la ecuación diferencial (57) asintóticamente inestable. En consecuencia, deberían tomarse tamaños de paso negativos para obtener la convergencia.

## 7.4 Regresión no lineal. Caso multidimensional

Considérese el modelo no lineal dado por

$$z_i = g_i(a) + \nu_i, \quad i = 1, 2, \dots,$$

donde  $z_i \in \mathcal{R}$ ,  $a \in \mathcal{R}^m$ ,  $g_i : \mathcal{R}^m \rightarrow \mathcal{R}$  y  $\nu_i$  son variables aleatorias reales, independientes e idénticamente distribuidas. Se supone que los criterios de agrupación de los datos son los mismos que se propusieron en la sección 7.2.

Los primeros métodos de aproximaciones estocásticas (véase ROBBINS Y MONRO (1951) y KIEFER Y WOLFOWITZ (1952)) tratan la estimación de parámetros unidimensionales, tal como se ha hecho hasta este momento en el capítulo. Para actualizar la estimación actual  $a_n$  se debe corregir este valor real a partir del resultado de una variable aleatoria real  $Y_n$ , de forma que

$$a_{n+1} = a_n + \alpha_n Y_n,$$

donde  $\alpha_n$  es un valor real positivo. Dependiendo de si  $Y_n > 0$  ó  $Y_n < 0$  la dirección de cambio a partir de  $a_n$  es una u otra. En este caso lo único que hace  $\alpha_n > 0$  es controlar la magnitud del cambio, si bien no la dirección.

En nuestro problema de regresión no lineal,  $Y_n$  es un error de observación, es

decir, la diferencia entre el valor observado de una variable aleatoria  $Z_n$  y su valor estimado a partir del valor actual del parámetro  $g_n(a_n)$  (recordar la notación de la sección 7.2). Por tanto,  $Y_n = Z_n - g_n(a_n)$  es un valor real. Si se supone que  $a_n$  es  $m$ -dimensional, con  $m > 1$ , y la aproximación estocástica es del tipo

$$a_{n+1} = a_n + \alpha_n Y_n,$$

se debe tomar  $\alpha_n$  también  $m$ -dimensional. En esta situación,  $\alpha_n$  determina claramente la dirección de cambio en la actualización y no sólo su amplitud. Esta es la razón por la que el tratamiento del caso multidimensional es diferente al del caso unidimensional.

En las referencias de KUSHNER se plantean distintas situaciones multidimensionales desde una perspectiva distinta a la anterior. Allí  $Y_n$  es una variable aleatoria  $m$ -dimensional, con lo cual se puede tomar  $\alpha_n$  real, no debiéndonos preocupar de la elección de su dirección, sino sólo de su magnitud. En este capítulo, por el contrario, se mantendrá  $Y_n$  como el error de observación real y se buscará en cada momento una dirección de cambio adecuada. En el caso lineal, dicha dirección de cambio se iguala al vector  $x_n$  de variables independientes del modelo de regresión, si bien en el caso no lineal esta dirección no es tan fácil de asignar. Se verá, sin embargo, que la elección del gradiente de la función de regresión en el punto de estimación actual como dirección de cambio puede ser conveniente.

#### 7.4.1 Primeros resultados

Se comenzará viendo que siempre existe un algoritmo, del tipo de las aproximaciones estocásticas usadas en las anteriores secciones, que converge al verdadero valor del parámetro multidimensional  $a \in \mathcal{R}^m$ . El algoritmo general es el siguiente:

INICIALIZACIÓN: Tómesese un vector inicial  $a_1 \in \mathcal{R}^m$  arbitrario.

ITERACIÓN: Actualícese  $a_n$  mediante

$$a_{n+1} = a_n + \bar{\beta}_n (y_n(a_n) - g_n(a_n)),$$

donde  $y_n(a_n)$  es la corrección en media habitual (aleatorizada o no aleatorizada) y  $\bar{\beta}_n \in \mathcal{R}^m$  es un vector que hay que elegir adecuadamente. En principio, se supondrá que dicho vector es aleatorio, dependiente del valor de  $a_n$ .

Supongamos que  $g_n : \mathcal{R}^m \rightarrow \mathcal{R}$  es diferenciable en todo su dominio y denotemos su vector gradiente en el punto  $\alpha \in \mathcal{R}^m$  por  $\dot{g}_n(\alpha)$ . Del teorema del valor medio se desprende que existe un vector  $\xi_n = \xi_n(a_n, a)$  que está en el segmento cerrado de extremos  $a_n$  y  $a$  cumpliendo

$$g_n(a_n) - g_n(a) = \dot{g}_n(\xi_n)^t (a_n - a).$$

En el algoritmo descrito, la elección de los pasos vectoriales se tomará

$$\bar{\beta}_n = \beta_n \dot{g}_n(\xi_n),$$

siendo  $\beta_n \in \mathcal{R}$  un tamaño de paso real, que verificará unas ciertas condiciones similares a las impuestas en el caso unidimensional. La igualdad anterior significa que la dirección de actualización del algoritmo cuando se está en  $a_n$  viene dada por el vector gradiente de  $g_n$  en el punto  $\xi_n$ . Como es evidente, el punto  $\xi_n$  no es calculable, puesto que se desconoce el valor de  $a$ . En todo caso, se verá que el algoritmo converge en  $L_2$  y después se tratará de estimar el vector  $\dot{g}_n(\xi_n)$  en cada iteración.

**Teorema 7.8** *Supongamos que las funciones reales con valores reales  $\delta_j(\cdot)$  son derivables para cada  $j = 0, 1, \dots, r - 1$ , y que existen  $\varepsilon > 0$  y  $M < \infty$  cumpliéndose que*

$$\|\dot{g}_n(\cdot)\| \leq M$$

y

$$\varepsilon \leq \delta_j(\cdot) \leq M.$$

*Si se toman tamaños de paso positivos tales que  $\sum \beta_n = \infty$  y  $\sum \beta_n^2 < \infty$ , entonces,  $g_n(a_n) - g_n(a) \xrightarrow{L_2} 0$ .*



DEMOSTRACION: Siguiendo el argumento del caso unidimensional, se llega fácilmente a

$$a_{n+1} = a_n - \beta_n m_n^* (g_n(a_n) - g_n(a)) \dot{g}_n(\xi_n) + \beta_n \varepsilon_n(a) \dot{g}_n(\xi_n),$$

donde

$$\begin{aligned} m_n^* &= 1, & \text{si } n \in I^{ng} \\ &= \delta_j(\zeta_n), & \text{si } n \in I^g \text{ y } z_n \in (c_j, c_{j+1}], \end{aligned}$$

para algún valor  $\zeta_n$  del segmento de extremos  $-g_n(a_n)$  y  $-g_n(a)$ . En definitiva, teniendo en cuenta la esperanza nula de los errores  $\varepsilon_n(a)$ , resulta

$$\begin{aligned} E \|a_{n+1} - a\|^2 &= E \|a_n - a\|^2 + \beta_n^2 E \left\| \varepsilon_n(a_n) \dot{g}_n(\xi_n) \right\|^2 \\ &\quad + E \left( \beta_n^2 m_n^{*2} (g_n(a_n) - g_n(a))^2 \left\| \dot{g}_n(\xi_n) \right\|^2 \right) \\ &\quad - 2E \left( \beta_n m_n^* (g_n(a_n) - g_n(a)) \dot{g}_n(\xi_n)^t (a_n - a) \right) \\ &= E \|a_n - a\|^2 + \beta_n^2 E \left\| \varepsilon_n(a_n) \dot{g}_n(\xi_n) \right\|^2 \\ &\quad - E \left( \beta_n m_n^* \left( 2 - \beta_n m_n^* \left\| \dot{g}_n(\xi_n) \right\|^2 \right) (g_n(a_n) - g_n(a))^2 \right). \end{aligned}$$

Para valores de  $n$  suficientemente grande ( $n \geq n_0$ ), teniendo en cuenta que

$$\min \{1, \varepsilon\} \leq m_n^* \leq \max \{1, M\},$$

se puede conseguir que

$$\beta_n m_n^* (2 - \beta_n m_n^* M^2) \geq \bar{\varepsilon} \beta_n,$$

siendo  $\bar{\varepsilon} = \min \{1, \varepsilon\}$ . Por otra parte, también se puede afirmar que  $E |\varepsilon_n(a_n)|^2 \leq 1 + \sigma^2$ , donde  $\sigma^2 = 0$  si las correcciones son no aleatorizadas, y  $\sigma^2 > 0$  si, por el contrario, las correcciones son aleatorizadas. A partir de la acotación de las normas del gradiente de  $g_n$ , se pueden igualmente acotar las esperanzas mediante

$$E \|a_{n+1} - a\|^2 \leq E \|a_n - a\|^2 + \beta_n^2 M^2 E |\varepsilon_n(a_n)|$$

$$\begin{aligned} & -\bar{\varepsilon}\beta_n E \left( (g_n(a_n) - g_n(a))^2 \right) \\ \leq & E \|a_n - a\|^2 + \beta_n^2 M^2 (1 + \sigma^2) \\ & -\bar{\varepsilon}\beta_n E \left( (g_n(a_n) - g_n(a))^2 \right). \end{aligned}$$

Iterando en esta desigualdad  $n - n_0$  veces resulta

$$\begin{aligned} 0 \leq E \|a_{n+1} - a\|^2 & \leq E \|a_{n_0} - a\|^2 + M^2 (1 + \sigma^2) \sum_{i=n_0}^n \beta_i^2 \\ & - \sum_{i=n_0}^n \bar{\varepsilon}\beta_i E \left( (g_i(a_i) - g_i(a))^2 \right). \end{aligned}$$

De aquí se deriva que la última serie es convergente y, por el mismo argumento que en el caso unidimensional (ver teorema 7.1), se concluye que

$$E \left( (g_i(a_i) - g_i(a))^2 \right) \xrightarrow{i \rightarrow \infty} 0. \square$$

OBSERVACIONES: i) La condición  $g_n(a_n) - g_n(a) \xrightarrow{L_2} 0$  no implica que  $a_n \xrightarrow{L_2} a$ . Por ejemplo, supóngase que se trata de estimar un parámetro bidimensional  $(\theta_1, \theta_2)$  a partir de unas variables de esperanza desconocida  $g(\theta_1, \theta_2) = \theta_1 + \theta_2 = c$ , que pueden ser censuradas o no serlo. Puesto que el algoritmo sólo depende del resultado de las variables y de  $g(\theta_1 + \theta_2) = c$ , parece intuitivo pensar que lo más que se puede estimar es la suma  $\theta_1 + \theta_2 = c$ , no siendo identificable cada uno de los parámetros.

ii) Por otro lado, obsérvese que, si bien en general  $\dot{g}_n(\xi_n)$  no es calculable, podrá serlo en algún caso particular. Así sucede, por ejemplo, cuando las funciones regresoras  $g_n$  son lineales. En concreto, supongamos que  $g_n(a) = x_n^t a$ , para cada  $a \in \mathcal{R}^m$ . El gradiente es un vector constante  $\dot{g}_n \equiv x_n$ , con lo cual el método es aplicable y se obtiene convergencia en los términos descritos. El paso de ITERACIÓN del algoritmo adopta la forma

$$a_{n+1} = a_n + \beta_n (y_n(a_n) - x_n^t a_n) x_n,$$

donde  $\beta_n$  es un tamaño de paso real y positivo. Este algoritmo coincide con uno de

los vistos en el capítulo 4, en concreto en la sección 4.6, si bien allí las condiciones de convergencia impuestas diferían de las vistas aquí. También diferían las técnicas de demostración allí empleadas y, de hecho, la conclusión  $x_n^t(a_n - a) \xrightarrow{L^2} 0$ , que se obtiene aquí, tampoco coincide con el tipo de convergencia obtenida en aquel lugar.

### 7.4.2 Un nuevo algoritmo

Para algunos tipos concretos de funciones  $g_n$  se puede proponer un algoritmo sin indeterminación alguna (es decir, calculable siempre) que consigue un tipo de convergencia similar a la establecida en el teorema 7.8. Bastaría tomar  $a_n$  como estimación de  $\xi_n$  en el algoritmo anterior, o bien en el algoritmo general tomar el paso vectorial

$$\bar{\beta}_n = \beta_n \dot{g}_n(a_n),$$

con  $\beta_n > 0$ . Según esto, el algoritmo progresa en la dirección del gradiente de  $g_n$  en el punto  $a_n$ . Para probar la convergencia serán necesarias condiciones de monotonía similares a las vistas en el caso unidimensional. Éstas, informalmente, asegurarán que la dirección del gradiente en  $a_n$  es la misma que en  $\xi_n$ , si bien ambos gradientes no coincidirán en norma. El siguiente teorema establece un resultado de utilidad, para el cual se sigue asumiendo la diferenciabilidad de las funciones  $g_n$ .

**Teorema 7.9** *Supongamos que existen  $\varepsilon > 0$  y  $M < \infty$  cumpliéndose que*

$$0 < \varepsilon \leq \frac{\dot{g}_n(a_n)^t(a_n - a)}{g_n(a_n) - g_n(a)},$$

para cada  $g_n(a_n) \neq g_n(a)$ , y

$$\|\dot{g}_n(\cdot)\| \leq M.$$

Además, supongamos que las  $\delta_j$  son derivables, verificando

$$\varepsilon \leq \dot{\delta}_j(\cdot) \leq M.$$

Si se toman tamaños de paso reales y positivos tales que  $\sum \beta_n = \infty$  y  $\sum \beta_n^2 < \infty$ , se verifica que  $g_n(a_n) - g_n(a) \xrightarrow{L_2} 0$ .

DEMOSTRACION: Como antes (véase el teorema 7.4), se puede llegar a que

$$\begin{aligned} E \|a_{n+1} - a\|^2 &= E \|a_n - a\|^2 + \beta_n^2 E \left\| \varepsilon_n(a_n) \dot{g}_n(a_n) \right\|^2 \\ &\quad + E \left( \beta_n^2 m_n^{*2} (g_n(a_n) - g_n(a))^2 \left\| \dot{g}_n(a_n) \right\|^2 \right) \\ &\quad - 2E \left( \beta_n m_n^* (g_n(a_n) - g_n(a)) \dot{g}_n(a_n)^t (a_n - a) \right) \\ &\leq E \|a_n - a\|^2 + \beta_n^2 M^2 (1 + \sigma^2) \\ &\quad - E \left( \beta_n m_n^* (2\varepsilon - \beta_n m_n^* M^2) (g_n(a_n) - g_n(a))^2 \right). \end{aligned}$$

Además, existe un  $n_0 \in \mathcal{N}$  tal que para cada  $n \geq n_0$  se cumple que

$$\beta_n m_n^* (2\varepsilon - \beta_n m_n^* M^2) \geq \bar{\varepsilon} \beta_n.$$

Iterando, lo anterior conduce a

$$\begin{aligned} 0 \leq E \|a_{n+1} - a\|^2 &\leq E \|a_{n_0} - a\|^2 + M^2 (1 + \sigma^2) \sum_{i=n_0}^n \beta_i^2 \\ &\quad - \sum_{i=n_0}^n \bar{\varepsilon} \beta_i E \left( (g_i(a_i) - g_i(a))^2 \right), \end{aligned}$$

lo cual implica la convergencia

$$E \left( (g_i(a_i) - g_i(a))^2 \right) \xrightarrow{i \rightarrow \infty} 0. \square$$

**Ejemplos** Veamos dos casos particulares donde se cumplen las hipótesis del teorema anterior.

i) Si las funciones  $g_n$  son lineales,  $g_n(a) = x_n^t a$ , basta tomar  $\varepsilon = 1$  para que

$$\frac{\dot{g}_n(a_n)^t (a_n - a)}{g_n(a_n) - g_n(a)} = \frac{x_n^t (a_n - a)}{x_n^t a_n - x_n^t a} = 1.$$

ii) Considérense las funciones  $h_n : \mathcal{R} \rightarrow \mathcal{R}$ , verificando las condiciones del teorema 7.1. Defínase después

$$g_n(a) = h_n(x_n^t a),$$

de donde

$$\dot{g}_n(a) = \dot{h}_n(x_n^t a) x_n.$$

Aplicando el teorema del valor medio, se asegura que

$$g_n(a_n) - g_n(a) = \dot{g}_n(\xi_n)(a_n - a) = \dot{h}_n(x_n^t \xi_n) x_n^t (a_n - a),$$

donde  $\xi_n$  está en el segmento definido por los extremos  $a_n$  y  $a$ . Se sigue que

$$\frac{\dot{g}_n(a_n)^t (a_n - a)}{g_n(a_n) - g_n(a)} = \frac{\dot{h}_n(x_n^t a_n)}{\dot{h}_n(x_n^t \xi_n)} \geq \frac{\varepsilon}{M} > 0,$$

sin más que recordar las condiciones de monotonía de  $h_n$  en el teorema 7.1 ( $h_n$  corresponde a la  $g_n$  de dicho teorema). Si, además, se impone que  $\|x_n\| \leq M$ , se cumplirá que

$$\left\| \dot{g}_n(\alpha) \right\| = \left| \dot{h}_n(x_n^t \alpha) \right| \|x_n\| \leq M^2.$$

Todas las condiciones impuestas sobre la función  $g_n$  en el teorema 7.9 han sido comprobadas, pues, en este caso.

### Notas y comentarios

- Las convergencias casi seguro establecidas en la sección 7.3 son las únicas que presentan tal carácter a lo largo de toda esta memoria. Las técnicas de demostración empleadas están basadas en el método ODE (por ordinary differential equations) que, en esencia, permite estudiar las convergencias de ecuaciones en diferencias estocásticas mediante la estabilidad asintótica de ciertas ecuaciones diferenciales deterministas. La referencia KUSHNER Y YIN (1997) es, en este sentido, obligada.

- Una buena parte de las convergencias en  $L_p$  contenidas en este capítulo (por ejemplo, teoremas 7.1 y 7.9) precisan de ciertas condiciones de monotonía sobre las funciones regresoras. Las convergencias globales obtenidas se convertirían en locales si las monotonías antes citadas tuvieran este mismo carácter.

## 8. Conclusiones finales

En este apartado se apuntan escuetamente los resultados más relevantes obtenidos durante el desarrollo de esta memoria. Se comentarán capítulo a capítulo, excluyendo el primero que, como se sabe, es introductorio.

En el capítulo 2:

- Se propone y analiza el algoritmo de dos iteraciones MD (M por corrección en Media y D por Dos iteraciones). Éste es válido para modelos lineales con información censurada y errores cualesquiera. Sigue la línea del EM, si bien sólo coincide con él bajo normalidad.
- En el teorema 2.1 se demuestra que el proceso de iteración secundario converge a un único punto, independientemente del punto de arranque elegido.
- El teorema 2.3 demuestra un Teorema Central del Límite que afecta al proceso primario de iteración. Dicho teorema puede servir para llevar a cabo inferencias a partir de la distribución asintótica estimada (véase teorema 2.4).
- El punto débil de los algoritmos propuestos radica en los métodos de cuadratura que se deben utilizar para la evaluación de las esperanzas condicionadas de corrección, ante distribuciones generales de los términos de error.

En el capítulo 3:

- Se proponen los algoritmos MdD (Md por corrección en Moda y D por dos iteraciones). Con ellos se intentan solucionar los problemas de cuadratura antes citados, separándonos totalmente del EM. Pese a la simplificación que

suponen los algoritmos MdD los resultados asintóticos que se han demostrado son similares a los obtenidos en el capítulo 2. El teorema 3.1 afecta a la unicidad del punto límite del proceso secundario. Por su parte, los teoremas 3.2 y 3.3 se refieren a las convergencias estocásticas relativas al proceso primario de iteración.

- Las correcciones en moda suponen una enorme facilidad de cálculo ante errores generales, asumiéndose simplemente condiciones de forma sobre sus densidades.
- Como subproducto de las demostraciones de los teoremas contenidos en los capítulos 2 y 3, tanto los algoritmos MD como MdD mantienen sus propiedades de convergencia estocástica ante situaciones de falta de homocedasticidad.

En el capítulo 4:

- Con los algoritmos de una iteración aquí propuestos se tratan de evitar las iteraciones anidadas propias de los de dos iteraciones. La versión más simple de los algoritmos aquí tratados está contenida en la sección 4.6 y apunta en su forma hacia las redes neuronales.
- La simplificación que suponen es enorme. Pese a ello, las propiedades asintóticas se mantienen. Los teoremas 4.1, 4.4, 4.7 y 4.10 constituyen las aportaciones propias más relevantes que afectan a la convergencia estocástica de los algoritmos propuestos.
- Los resultados computacionales muestran que estos algoritmos suponen, frente a los de dos iteraciones, enormes ventajas de cálculo, sin apenas pérdida de eficiencia (véase la figura 7).
- En las simulaciones que se han realizado (véanse las figuras 6) se aprecia que las trayectorias particulares del algoritmo aparentemente son independientes del punto de arranque (a partir de un  $n$  en adelante). Este resultado es



sorprendente y, sin duda, alentador.

En el capítulo 5:

- Como en el capítulo 3, aquí los algoritmos de estimación de una iteración considerados sustituyen las correcciones en esperanza condicionada por correcciones en moda. La simplificación que esto supone ya ha sido explicada y se omite.
- Las propiedades asintóticas se mantienen. Los resultados propios más relevantes sobre convergencias estocásticas están contenidos en los teoremas 5.1, 5.2, 5.5 y 5.6.
- Las simulaciones llevadas a cabo en la sección 5.10 de esta memoria muestran que los tiempos de cálculo se dividen por aproximadamente 2200 sin pérdida significativa de eficiencia.

En el capítulo 6:

- La escasa dependencia distribucional de las correcciones en moda induce un tipo de correcciones generales. Se permiten, incluso, correcciones de tipo aleatorio.
- En conclusión, todo parece indicar que cualquier tipo de corrección es bueno si se hace con cierta lógica. Los resultados propios más significativos están contenidos en los teoremas 6.8 y 6.9 (para correcciones aleatorias) y en los teoremas 6.3, 6.4, 6.5, 6.6 y 6.7 (para las correcciones deterministas generales).

En el capítulo 7:

- Se utilizan las aproximaciones estocásticas para el tratamiento de modelos no lineales.
- Los teoremas 7.8 y 7.9 constituyen las aportaciones propias que considero más relevante del capítulo. Las técnicas de demostración de las convergencias se

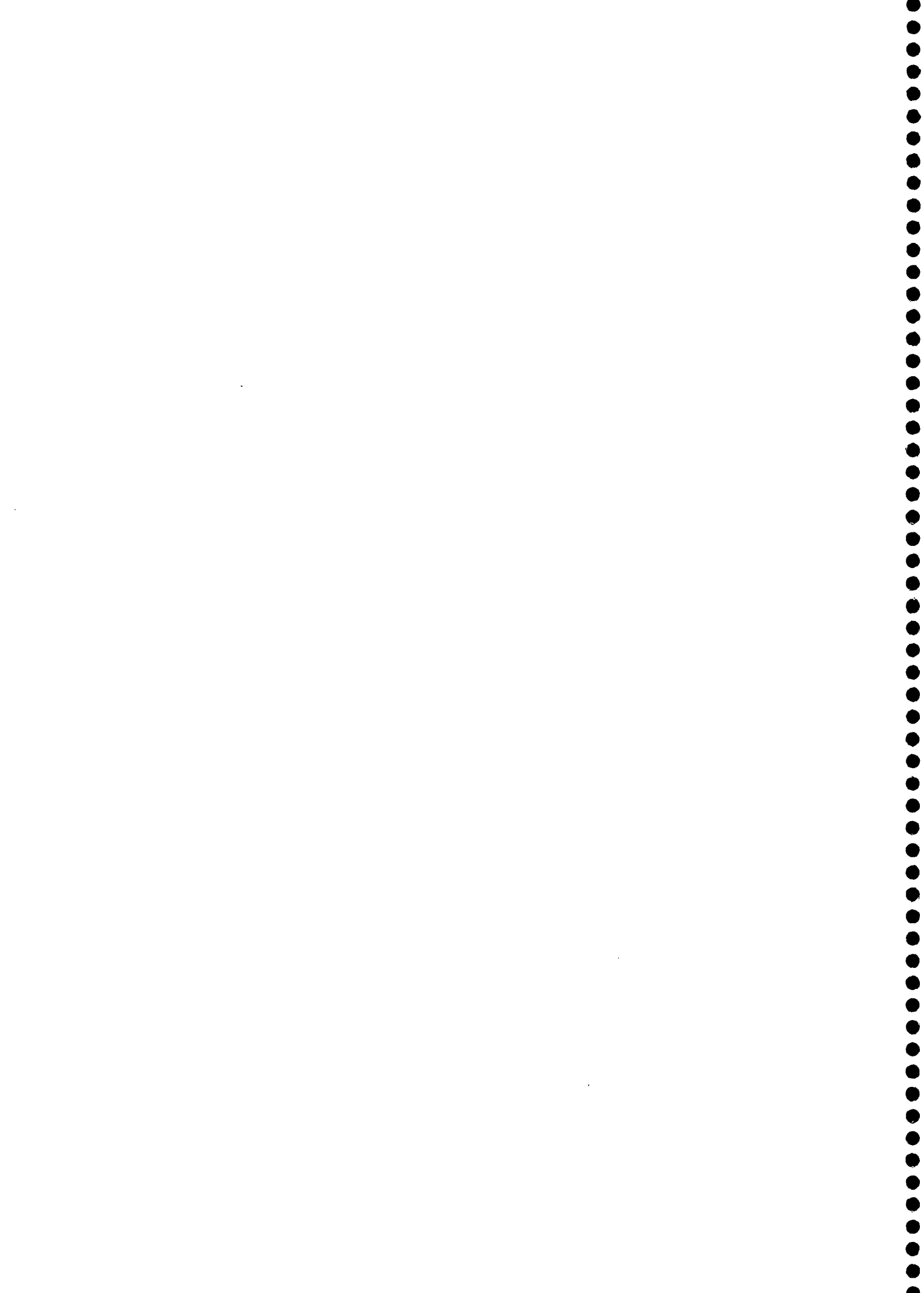
basan en aproximaciones estocásticas y los algoritmos propuestos son los más generales entre los contenidos en esta memoria. Por este motivo, se exigen ciertas condiciones de regularidad (para las funciones regresoras no lineales y para los tamaños de paso) con el fin de poder garantizar las convergencias estocásticas globales que constituyen la tesis de los teoremas citados.

**DESARROLLOS FUTUROS:** Los algoritmos de dos iteraciones son habituales en distintos contextos de estimación, con o sin información parcial (e.g., máxima verosimilitud, métodos bayesianos, etc.). Por ejemplo, es típico que a la hora de calcular un estimador máximo verosímil, con tamaño de muestra dado, se empleen métodos iterativos Newton-Raphson (iteración secundaria), para luego utilizar la distribución asintótica del estimador de máxima verosimilitud (iteración primaria). Ésta última constituye de facto el límite estocástico de la iteración primaria.

El precedente estudiado en esta memoria, en el sentido de sustituir dos iteraciones por una única, podría ser extendido a todos los contextos de estimación en dos iteraciones antes citados. La posibilidad de ampliación de esta memoria a otras situaciones usuales es prácticamente ilimitada y supondrá sin duda una línea futura de trabajo.



## Referencias



- [1] A. BENVENISTE, M. METIVIER Y P. PRIOURET, *Adaptive Algorithms and Stochastic Approximation*. Springer-Verlag, Berlin and New York, 1990.
- [2] A. P. DEMPSTER, N. M. LAIRD Y D. B. RUBIN, Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society*, Vol. **B 39**, 1977, pp. 1-38.
- [3] J. GEWEKE, Bayesian Inference in econometric models using Monte Carlo integrations. *Econometric*, Vol. **24**, 1989, pp. 1317-1339.
- [4] P. HALL Y C. C. HEYDE, *Martingale Limit Theory and Its Application*, Academic Press, 1980.
- [5] S. HAYKIN, *Neural Networks: A comprehensive Foundation*. Macmillan, New York, 1994.
- [6] J. KIEFER Y J. WOLFOWITZ, Stochastic estimation of the maximum of a regression function. *Ann. Math. Stat.*, Vol. **23**, 1952, pp. 462-466.
- [7] H. J. KUSHNER Y D. S. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, 1978.
- [8] H. J. KUSHNER Y G. G. YIN, *Stochastic Approximation Algorithms and Applications*, Springer, 1997.
- [9] R. G. LAHA Y V. K. ROHATGI, *Probability Theory*, Wiley, 1979.
- [10] K. LANGE, A gradient algorithm locally equivalent to de EM algorithm. *Journal of the Royal Statistical Society*, Vol. **B 57**, 1995, pp. 425-437.
- [11] R. J. A. LITTLE Y D. B. RUBIN, *Statistical analysis with missing data*, Wiley, 1987.
- [12] T. A. LOUIS, Finding observed information using the EM algorithm. *Journal of the Royal Statistical Society*, Vol. **B 44**, 1982, pp. 98-130.
- [13] G. J. MCLACHLAN Y T. KRISHNAN, *The EM Algorithm and Extensions*, Wiley, 1997.

- [14] T. ORCHARD Y M. A. WOODBURY, A missing information principle: Theory and applications. *Proceeding of the Sixth Berkeley Symposium on Mathematical Statistics*, Vol. 1, 1972, pp. 697-715.
- [15] M. M. RAO, *Linear Statistical Inference and Its Applications*, Wiley, 1973.
- [16] H. ROBBINS Y S. MONRO, A stochastics approximation methods, *Ann. Math. Stat.*, Vol. 22, 1951, pp. 400-407.
- [17] J. SCHMEE Y G. J. HAHN, A simple method for regression analysis with censored data. *Technometrics*, Vol. 21, 1979, pp. 417-432.
- [18] M. A. TANNER, *Tools for Statistical Inference. Observed Data and Data augmentation Methods*. Springer, 1993.
- [19] L. TIERNEY Y J.B. KADANE, Acurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*. Vol. 81, 1986, pp. 82-86.
- [20] M. T. WASAN, *Stochastic Approximation*, Cambridge at the University Press, 1969.
- [21] H. WHITE, *Artificial Neural Networks*. Blackwell, Oxford, UK, 1992.
- [22] C. F. J. WU, On the convergence of the EM algorithm. *Ann. Statist.*, Vol. 11, 1983, pp. 95-103.

