UNIVERSITAT DE VALÈNCIA

Aportaciones al análisis bayesiano semiparamétrico de datos de supervivencia

Eduardo Beamonte Córdoba

Memoria para optar al grado de doctor dirigida por:
José D. Bermúdez Edo.

Departamento de Estadística e Investigación Operativa. Junio de 1998.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

MATEMATICAS

Aportaciones al análisis bayesiano semiparamétrico de datos de supervivencia

Eduardo Beamonte Córdoba

Dirigida por José D. Bermúdez Edo

Junio de 1998

D. José Domingo Bermúdez Edo, profesor titular del Departamento de Estadística e Investigación Operativa de la Universitat de València,

CERTIFICA:

que la presente memoria,

"Aportaciones al análisis bayesiano semiparamétrico de datos de supervivencia",

ha sido realizada bajo su dirección por *Eduardo Beamonte Córdoba* para optar al grado de Doctor en Ciencias Matemáticas por la Universitat de València.

Y para que así conste, en cumplimiento de la legislación vigente, firma el presente certificado en Burjasot, a veintinueve de Junio de 1998.

Fdo. José D. Bermúdez Edo.

Agradecimientos

La presente memoria es el fruto del trabajo desarrollado durante los últimos seis años en el Departamento de Estadística e Investigación Operativa de la Universitat de València. Quiero agradecer a todos sus miembros su apoyo y la disposición que siempre han tenido a brindarme su ayuda.

También debo expresar mi gratitud a mis compañeros del Departamento de Economía Aplicada de la Universitat de València, especialmente a los de la Unidad Docente de Estadística. Siendo éste el centro donde desarrollo mi trabajo, ellos han sido los que no han dudado en ofrecerme su colaboración, especialmente en los últimos meses, en todo aquello que he precisado.

Desde estas líneas quiero dejar constancia de mi especial gratitud al principal artífice de que este proyecto de Tesis Doctoral saliera adelante. Si obligado agradecimiento es el del director en todas las tesis doctorales, en este caso no es un tópico y sí un sincero reconocimiento el que debo a José Bermúdez, que desde sus amplios conocimientos de la Estadística bayesiana ha sabido guiarme sabiamente en todo momento para poder culminar este proyecto.

Finalmente, quiero agradecer a mi familia su comprensión y dedicación, y especialmente a Ana, que ha tenido que soportarme y convivir conmigo pacientemente durante la gestación de esta memoria.

Eduardo Beamonte Córdoba
Valencia, Junio de 1998.

A mi madre, a Ana.

Índice General

1 Introducción y antecedentes **1**

1.1 Introducción 3

1.2 Soluciones no paramétricas 5

1.2.1 El modelo de Cox 5

1.2.2 Otras soluciones 7

1.3 Modelo Gamma para incorporar covariables 13

1.3.1 Modelo jerárquico Gamma 13

1.3.2 Función de azar Gamma 16

1.4 Métodos MCMC 21

1.4.1 Preliminares 21

1.4.2 Métodos de simulación dinámica 23

1.4.3 Problemas prácticos en la implementación de métodos MCMC 27

2 Modelo semiparamétrico aditivo 35

2.1 Introducción 37

2.2 Presentación del modelo 38

2.2.1 Modelo Gamma-poligonal aditivo 42

2.3 Análisis del modelo 47

2.3.1 Función de verosimilitud 47

2.3.2 Distribución inicial y distribución final 53

2.4 Estudio de la distribución final mediante simulación 56

3 Aplicación al análisis de datos de supervivencia 63

3.1 Introducción 65

3.2 Implementación del modelo Gamma-poligonal aditivo 66

3.3 Datos simulados 72

3.3.1 Banco de datos 1 72

3.3.2 Banco de datos 2 83

3.4 Datos reales 91

3.4.1 Banco de datos de Stanford 91

3.4.2 Banco de datos de una encuesta a licenciados en Matemáticas 97

4 Conclusiones y futuras líneas de investigación **107**

A Definiciones y propiedades en supervivencia **113**

B Procesos estocásticos. Cadenas de Markov **117**

C Bancos de datos de Stanford y de una encuesta a licenciados en Matemáticas **123**

Índice de figuras **133**

Índice de tablas **137**

Bibliografía **139**

Capítulo 1

Introducción y antecedentes

1.1 Introducción

El análisis estadístico de *datos de vida* ha experimentado un gran desarrollo en los últimos años, constituyendo materia de estudio de multitud de investigadores de diversas áreas de conocimiento relacionadas con: las ciencias de la salud (Gilks et al., 1993; Gómez y Lagakos, 1994; Tsai et al., 1994), las ciencias biológicas (Van der Laan et al., 1997; Pradel et al., 1997; Yue y Chan, 1997), la ingeniería (Meeker y LuValle, 1995; Lindsey, 1997; Hu et al., 1998) y las ciencias económicas y sociales, (Follmann et al., 1990; Jaggia y Thosar, 1995; Beenstock, 1996), entre otras.

Las peculiares características de este tipo de datos hacen que su tratamiento estadístico sea también particular. Es habitual en el análisis de datos de vida encontrar que algunos de ellos corresponden a *tiempos de supervivencia censurados* (ver apéndice A), esto es, para estos datos tan sólo se conoce que el verdadero tiempo de supervivencia es mayor (o menor, según el tipo de censura) que el observado. Ello complica notablemente el estudio independientemente del tipo de análisis que se lleve a cabo. Tal complicación se acentúa cuando se quiere realizar un análisis bayesiano bajo un modelo paramétrico o semiparamétrico.

En este trabajo se propone un modelo semiparamétrico para el análisis bayesiano de datos de vida con covariables. La idea principal es modelizar los mismos mediante la *función de azar* (ver apéndice A), considerando ésta como la suma de una parte no paramétrica y otra paramétrica. De este modo se obtiene una considerable riqueza en la forma funcional del azar que permite contemplar un amplio registro de situaciones de supervivencia. Además, con el modelo propuesto cabe la incorporación de covariables dentro de la parte paramétrica, lo cual suele ser fundamental en el análisis de supervivencia. La metodología que aquí se desarrolla para el tratamien-

to bayesiano de este tipo de datos está basada en las llamadas *técnicas de Monte Carlo en cadenas de Markov (MCMC)*. De profusa utilización en los últimos tiempos, estas técnicas han devenido en imprescindibles para el análisis de modelos realmente complejos.

La presente memoria está estructurada en cuatro capítulos y tres apéndices. En un primer capítulo introductorio se presentan las soluciones no paramétricas al problema de estimación en supervivencia más frecuentes en la literatura. Se presenta un modelo Gamma para el análisis bayesiano de datos de vida con covariables y se realiza una breve revisión de los métodos MCMC utilizados para llevar a cabo los estudios. En el segundo capítulo se propone un modelo semiparamétrico para la función de azar, en el que ésta es la suma de una componente paramétrica, relativa a una distribución conocida, y otra componente no paramétrica o libre de distribución. Se demuestran algunas de sus propiedades teóricas más interesantes y se obtienen las distribuciones condicionales completas necesarias para el estudio, mediante técnicas MCMC, de la distribución final y de las distribuciones predictivas. En el tercer capítulo se muestra cómo resolver los problemas técnicos que pueden aparecer en la implementación de los resultados teóricos del capítulo anterior y, utilizando la herramienta informática así desarrollada, se analizan varios bancos de datos simulados y reales. En el cuarto y último se reflexiona sobre las conclusiones obtenidas y se comentan las futuras líneas de investigación a desarrollar como continuación de este trabajo. Por último se recogen tres apéndices: en el apéndice A se dan las definiciones más importantes relativas a los datos de vida, así como algunas propiedades. En el apéndice B se proporcionan algunas definiciones y resultados relativos a procesos estocásticos y cadenas de Markov que se utilizan a lo largo del trabajo. En el apéndice C se detallan los bancos de datos reales utilizados en la presente memoria.

1.2 Soluciones no paramétricas

Una de las primeras soluciones al problema de estimación no paramétrica de la *función de supervivencia* (ver apéndice A) es el llamado estimador *límite-producto* o de *Kaplan-Meier* (Kaplan y Meier, 1958), debido a los autores que primero comentaron sus propiedades.

Basado en la función de supervivencia empírica, tiene una sencilla e intuitiva interpretación, si bien las necesidades existentes de incorporar covariables o realizar el análisis desde la perspectiva bayesiana, hacen que sus limitaciones, en ese sentido, sean grandes.

Incluido habitualmente dentro del capítulo de soluciones no paramétricas o semiparamétricas, se encuentra el llamado *modelo de azares proporcionales de Cox*, que es el más utilizado para modelizar datos de vida con covariables.

1.2.1 El modelo de Cox

Cox (1972) propuso un modelo de regresión de azares proporcionales para el tratamiento de tiempos de supervivencia con covariables, en el que la función de azar de cada individuo es de la forma:

$$h(t) = h_0(t)\, exp(\boldsymbol{\beta}\mathbf{x}), \tag{1.1}$$

donde $\mathbf{x}$ es el vector de variables explicativas del individuo, $\boldsymbol{\beta}$ es el vector paramétrico y $h_0(t)$ es una función arbitraria llamada *función de azar base*, función de azar para un individuo con $\mathbf{x} = \mathbf{0}$.

En un posterior trabajo, Cox (1975) proporciona estimadores de los parámetros $\boldsymbol{\beta}$ basándose en una *función de verosimilitud parcial*, como al-

ternativa a los estimadores máximo verosímiles obtenidos con anterioridad para la verosimilitud marginal.

Este modelo de azares proporcionales de Cox ha sido comúnmente utilizado para el análisis de datos de vida con variables explicativas, debido en parte a su sencillez y a que resulta apropiado para muchos estudios de supervivencia. Ligeras variaciones del mismo (como utilizar una función positiva distinta a la exponencial para ligar covariables y parámetros en (1.1)) han sido desarrolladas por diferentes autores para potenciar la aplicabilidad del modelo. Sin embargo, las soluciones propuestas por Cox (1972, 1975) al problema de estimación de los parámetros β son también las utilizadas por la mayoría de los investigadores (Whitehead, 1980; Quantin et al., 1996; Wang et al., 1997).

Es calificado en la literatura como *modelo semiparamétrico* pues si bien no asume niguna distribución para el azar base, sí que modeliza la relación entre las covariables y la función de azar vía la función exponencial, asumiendo proporcionalidad entre las funciones de azar de dos individuos cualesquiera. Es precisamente esta circunstancia la que restringe, en cierto modo, su utilización.

Existen en la literatura gran cantidad de referencias al modelo de azares proporcionales de Cox, sin embargo, tan sólo una pequeña parte de ellas afronta el problema de estimación desde una perspectiva bayesiana. Dejando de lado otras razones, resulta evidente que la complejidad inherente al tratamiento bayesiano, sobre todo en las evaluaciones de complicadas integrales, motiva a realizar un análisis clásico del problema. En cualquier caso, los grandes avances tecnológicos en materia de computación y el gran desarrollo alcanzado por los métodos MCMC permite, cuando menos, abordar y en muchas ocasiones resolver este tipo de problema.

El tratamiento bayesiano del modelo de Cox pasa por considerar a los parámetros β, e incluso al azar base $h_0(t)$, como cantidades aleatorias. En una ligera modificación del modelo, algunos autores consideran también aleatoria la función que relaciona las covariables con la función de azar. Trabajos como los de Hjort (1990), Dellaportas y Smith (1993), Mallick y Gelfand (1994) y Gelfand y Mallick (1995) son excelentes referencias de la investigación en este sentido.

1.2.2 Otras soluciones

Numerosas han sido las aportaciones de diversos autores al problema de estimación no paramétrica de la función de supervivencia. Sin pretender realizar en este apartado una completa revisión de las mismas, sí que se comentan las más citadas en la literatura y se revisan más detalladamente las soluciones bayesianas semiparamétricas para el análisis de datos de vida con covariables.

En los estudios de supervivencia es habitual presentar los modelos mediante la función de azar. La intuitiva y sencilla interpretación de dicha función es la causa principal de tal proceder (Cox y Oakes, 1984). Casi inmediatamente después de la aparición del modelo de Cox, diversos autores desarrollaron métodos alternativos de regresión para la incorporación de covariables al análisis de supervivencia. Muchas de estas técnicas están basadas en el modelo lineal y sus desarrollos, desde un punto de vista no bayesiano, pueden consultarse, por ejemplo, en Miller (1976), Buckley y James (1979) y Koul et al. (1981). Miller y Halpern (1982) proporcionan una buena revisión de estos métodos y efectúan una comparación entre los mismos.

Debido a su importancia y profusa utilización en diferentes áreas de conocimiento, cabe mencionar expresamente el *modelo de tiempos de fallo acelerados*, un caso particular de regresión lineal que modeliza el logaritmo de los tiempos de vida, de modo que las covariables tienen un efecto multiplicativo en dichos tiempos. Modelización muy utilizada en ingeniería, recibe este nombre pues en este área es habitual acelerar el tiempo de fallo en ciertos estudios para obtener porcentajes de datos censurados razonables. El análisis de supervivencia en este área de conocimiento recibe el nombre de *fiabilidad* y una excelente referencia como manual de análisis de fiabilidad es Barlow y Proschan (1996). Martz y Waller (1982) fueron los primeros autores en tratar este tipo de datos bajo un contexto bayesiano. Su libro permanece como un clásico en la materia.

En los últimos tiempos han merecido especial interés por parte de muchos investigadores los modelos semiparamétricos para el análisis bayesiano de datos de vida con covariables. En ellos, suele modelizarse la función de azar, o la *función de azar acumulado* (ver apéndice A), según una parte no paramétrica y otra paramétrica. Bajo la perspectiva bayesiana, se asume que la parte no paramétrica es la realización de un proceso estocástico y se considera una distribución inicial, con posibles hiperparámetros desconocidos, para la parte paramétrica.

El principal atractivo de estos modelos es su notable capacidad de adecuación a complejas situaciones de supervivencia. A continuación, se efectúa una breve revisión de los modelos semiparamétricos desarrollados para el tratamiento de diversos tipos de datos, desde un punto de vista clásico y después se comenta la metodología bayesiana semiparamétrica con mayor detalle.

Los ya comentados modelos de regresión de Cox y de regresión lineal

son los modelos semiparamétricos comúnmente utilizados para la inclusión de covariables en el *análisis de supervivencia univariante.* Ver, por ejemplo, Cox y Oakes (1984) para una consulta más detallada de los métodos no bayesianos más utilizados en el análisis de este tipo de datos. En el análisis de *datos de vida multivariantes*, el modelo mayoritariamente utilizado es el llamado *modelo de fragilidad* (Vaupel et al., 1979). En Andersen et al. (1993) y en Oakes (1989, 1994) se realizan excelentes revisiones de los métodos frecuentistas empleados en el estudio de estos datos, utilizando el modelo de fragilidad. Como alternativa, Wei et al. (1989) modelizan el azar marginal de cada tiempo de supervivencia mediante el propio modelo de Cox. Cuando el segundo suceso que define el tiempo de supervivencia puede experimentarse más de una vez, hablamos de *datos de vida de suceso múltiple.* Oakes (1992) proporciona una completa revisión del análisis no bayesiano de estos tiempos utilizando modelos de fragilidad. Los *datos de supervivencia doblemente censurados en un intervalo* han adquirido un especial protagonismo en los últimos tiempos, paralelo a la evolución de los estudios del síndrome de inmunodeficiencia adquirida. Es habitual en datos provenientes de esta enfermedad que no se observe el primer suceso de definición del tiempo de supervivencia (seropositividad, por ejemplo) y sólo se tenga certeza de que ha tenido lugar dentro de un cierto intervalo temporal y que ocurra lo mismo para el segundo de los sucesos (desarrollo del virus, por ejemplo). Pueden consultarse en De Gruttola y Lagakos (1989) y Gómez y Lagakos (1994) algunos modelos no paramétricos para el análisis de este tipo de datos. Para su estudio incorporando covariables ver, por ejemplo, Kim et al. (1993).

Tal y como se comenta con anterioridad, bajo la perspectiva bayesiana y dentro de los modelos semiparamétricos, se ha de especificar un proceso inicial para la parte no paramétrica. Si el tratamiento del problema se realiza con covariables, éstas pueden incorporarse dentro de la parte paramétrica.

A continuación se detallan las modelizaciones bayesianas no paramétricas más discutidas en la literatura.

El análisis de supervivencia desde una perspectiva bayesiana también data de la década de los setenta. Ferguson (1973) fue pionero en proponer soluciones bayesianas al problema de estimación no paramétrica introduciendo *procesos de Dirichlet.* Verdaderamente, si la modelización se realiza vía la función de azar este tipo de procesos pierden bastante interpretabilidad. Los trabajos posteriores de Susarla y Van Ryzin (1976) y Ferguson y Phadia (1979) son obligadas referencias y suponen notables avances de la investigación en este sentido. Asímismo, Kalbfleisch (1978) y Burridge (1981) utilizan *procesos iniciales Gamma* para modelizar la función de azar acumulado, mientras que Hjort (1990) utiliza *procesos Beta.* Berliner y Hill (1988) propusieron una distribución predictiva no paramétrica como alternativa al estimador límite-producto de Kaplan y Meier. Chen et al. (1985), en uno los primeros trabajos en el cálculo de distribuciones predictivas utilizando covariables, realizaron un análisis bayesiano aplicado al estudio de la supervivencia en el cáncer de mama, utlizando una generalización del modelo de Berkson y Gage (1952). En el trabajo de Christensen y Johnson (1988) se obtienen predictivas para futuros individuos utilizando un proceso de Dirichlet inicial en el modelo de tiempos de fallo acelerados. Morales et al. (1991) también utilizan un proceso inicial Dirichlet en un modelo de azares proporcionales para la estimación no paramétrica de la función de supervivencia con *datos censurados por la izquierda y por la derecha.* Otra clase de procesos ampliamente utilizados en la modelización bayesiana no paramétrica son los llamados *procesos correlados.* Introducidos por Leonard (1978), trabajos posteriores como los de Gamerman (1991) y Fahrmeir (1994) han demostrado su potencial aplicación, sobre todo, en el análisis de supervivencia univariante con *covariables dependientes del tiempo.* Debido a su importancia en el desarrollo del presente trabajo, comentamos con

mayor detalle alguna versión de este modelo.

Consideramos una partición del eje temporal en intervalos $I_j = (a_{j-1}, a_j]$, $j = 1, \ldots, g$, de modo que $0 = a_0 < a_1 < a_2 < \cdots < a_{g-1} < a_g$. Dada la función de azar escalonada $h(t) = \lambda_j$, para $t \in I_j$, el *proceso autocorrelado de primer orden* para $\alpha_j = \log \lambda_j$ es:

$$\alpha_{j+1} = \alpha_j + e_{j+1}, \quad j = 1, \ldots, g-1,$$

donde $e_j \sim N(e_j \,|\, 0, \sigma^2)$, $j = 2, \ldots, g$, independientes e independientes de los α_j's previos. Una interesante variación de este modelo consiste en considerar:

$$\begin{aligned} \alpha_{j+1} &= \alpha_j + \delta_j + e_j \\ \delta_{j+1} &= \delta_j + e'_j\,, \qquad j = 1, \ldots, g-1, \end{aligned}$$

con $e_j \sim N(e_j \,|\, 0, \sigma_1^2)$ y $e'_j \sim N(e'_j \,|\, 0, \sigma_2^2)$ mutuamente independientes. Algunos autores han propuesto distribuciones paramétricas distintas de la log-Normal para λ_j (Arjas y Gasbarra, 1996, utilizan una distribución Gamma para λ_j dados $\lambda_1, \ldots, \lambda_{j-1}$, de modo que $E(\lambda_j \,|\, \lambda_1, \ldots, \lambda_{j-1}) = \lambda_{j-1}$).

Finalmente, las *mixturas aleatorias y finitas* han sido recientemente utilizadas por diferentes autores como procesos iniciales. Gelfand y Mallick (1995) utilizan mixturas de densidades Beta y comentan la posible utilización de otras. Mukhopadhyay y Sinha (1995) proponen mixturas aleatorias de densidades Gamma.

El trabajo de Sinha y Dey (1997) supone una excelente puesta al día de la metodología bayesiana semiparamétrica en el análisis de supervivencia. Estos autores realizan una completa revisión de los distintos procesos estocásticos iniciales empleados para la modelización de la parte no paramétrica y comentan las soluciones bayesianas más utilizadas para diversos tipos de datos de vida.

Por su analogía con el planteamiento semiparamétrico de la función de azar que aquí se ha comentado, cabe citar el reciente trabajo de Kouassi y Singh (1997). En él se plantea modelizar la función de azar mediante un promedio ponderado de un azar no paramétrico y de un azar paramétrico. Realizan un tratamiento no bayesiano del problema de estimación del parámetro de la ponderación y proponen un estimador que converge a sus valores extremos (0 o 1) cuando prevalece alguno de los modelos, paramétrico o no paramétrico.

1.3 Modelo Gamma para incorporar covariables

En cualquier contexto de supervivencia, parece razonable pensar que las particulares características de cada individuo influyan en su tiempo de vida y que, por lo tanto, sea necesario incorporar covariables al estudio.

La forma habitual de incluir covariables en el análisis de datos de vida es utilizando modelos de regresión. Entre ellos, el modelo de regresión de Cox es sin duda el más popular, si bien es aplicable sólo en el contexto de azares proporcionales.

En este apartado presentamos un *modelo jerárquico Gamma* para el análisis bayesiano de datos de vida con covariables (Bermúdez y Beamonte, 1995).

1.3.1 Modelo jerárquico Gamma

Este modelo supone una generalización de uno anterior (Beamonte y Bermúdez, 1993) para permitir la entrada de covariables en el estudio. De este modo, suponemos que el tiempo de vida de cada individuo sigue el siguiente modelo jerárquico:

$$\begin{aligned} t &\sim Ga\left(t \mid \alpha, \beta\right) \\ (\alpha, \beta)' &\sim N_2\left((\log\alpha, \log\beta)' \mid B\mathbf{x}, H\right). \end{aligned}$$

Es decir, cada tiempo sigue una distribución Gamma de parámetros α y β. Estos parámetros son características propias de cada individuo y están relacionados con su vector de covariables, $\mathbf{x}$, a través de un mecanismo

aleatorio dependiente de ciertos hiperparámetros B y H. B es la matriz de coeficientes, H la matriz de precisión de la distribución Normal bivariante y ambos son comunes a todos los individuos.

Beamonte y Bermúdez (1995) realizan un análisis bayesiano del modelo utilizando una distribución inicial Normal-Wishart para los hiperparámetros y el muestreo de Gibbs para la obtención de una muestra de la distribución final.

Concretamente, si disponemos de n datos de vida tales que $\{t_1, \ldots, t_r\}$ son no censurados y $\{T_{r+1}, \ldots, T_n\}$ son *progresivamente censurados por la derechh* (ver apéndice A), entonces el vector paramétrico completo objeto del muestreo de Gibbs es el formado por los parámetros del modelo, $(\alpha_1, \beta_1, \ldots, \alpha_n, \beta_n)$, los hiperparámetros, B y H, y los tiempos no observados, $\{t_{r+1}, \ldots, t_n\}$, correspondientes a los datos censurados.

Las distribuciones condicionales completas, necesarias para el muestreo de Gibbs, resultan:

- *tiempo censurado no observado* $t_i, i = r+1, \ldots, n$.

$$f(t_i \,|\, T_i, \alpha_i, \beta_i) \propto Ga(t_i \,|\, \alpha_i, \beta_i) \text{ si } t_i > T_i, \tag{1.2}$$

  es decir, una distribución Gamma truncada.

- *hiperparámetros* B *y* H.

$$f(B, H \,|\, X, \alpha_1, \beta_1, \ldots, \alpha_n, \beta_n) \propto f(B, H) \cdot \\ \cdot f(\alpha_1, \beta_1, \ldots, \alpha_n, \beta_n \,|\, X, B, H), \tag{1.3}$$

  donde X es la matriz de covariables.

Utilizando una distribución inicial Normal-Wishart para (B, H), la correspondiente final también pertenece a la familia Normal-Wishart (consúltese, por ejemplo, Broemeling, 1985, pp. 378-379).

- *parámetro* α_i, $i = 1, \ldots, n$.

$$f(\alpha_i \,|\, t_i, \beta_i, \mathbf{x_i}, B, H) \propto \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} t_i^{\alpha_i} N\left(\log \alpha_i \,|\, \mu_1', \sigma_1'^2\right), \qquad (1.4)$$

con $\mu_1' = \mu_1 + \rho \frac{\mu_1}{\mu_2} (\log \beta_i - \mu_2)$ y $\sigma_1'^2 = \frac{1}{h_1}$, los momentos de la distribución Normal condicionada obtenida a partir de la Normal bivariante de media $\boldsymbol{\mu} = \binom{\mu_1}{\mu_2}$ y matriz de precisión $H = \binom{h_1 \; h_{12}}{h_{12} \; h_2} = \Sigma^{-1}$, donde $\Sigma = \binom{\sigma_1^2 \; \sigma_{12}}{\sigma_{12} \; \sigma_2^2}$ y $\sigma_{12} = \rho \sigma_1 \sigma_2$.

- *parámetro* β_i, $i = 1, \ldots, n$.

$$f(\beta_i \,|\, t_i, \alpha_i, \mathbf{x_i}, B, H) \propto \beta_i^{\alpha_i} \exp(-\beta_i t_i) N\left(\log \beta_i \,|\, \mu_2', \sigma_2'^2\right), \qquad (1.5)$$

siendo $\mu_2' = \mu_2 + \rho \frac{\mu_2}{\mu_1} (\log \alpha_i - \mu_1)$ y $\sigma_2'^2 = \frac{1}{h_2}$, la media y varianza de la distribución Normal condicionada correspondiente.

La implementación del algoritmo de Gibbs no resulta demasiado complicada, dado que se obtienen distribuciones analíticas en (1.2) y (1.3) y la simulación a partir de (1.4) y (1.5) puede realizarse mediante el método de aceptación-rechazo con funciones importantes log t-Student.

La posibilidad de realizar un análisis bayesiano del modelo utilizando técnicas MCMC, unido al hecho de que este modelo jerárquico Gamma incluya situaciones de supervivencia con azares no proporcionales, resalta su aplicabilidad práctica. En contrapartida, la función de azar Gamma es siempre monótona, no recogiendo azares con puntos extremos.

1.3.2 Función de azar Gamma

Definición 1.3.1 *La función de densidad Gamma con parámetros α y β, $Ga(\alpha, \beta)$, es:*

$$f(t|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} \exp(-\beta t),\ t > 0,\ \alpha, \beta > 0.$$

De la definición anterior se deduce inmediatamente la forma funcional del azar Gamma:

Proposición 1.3.1 *La función de azar Gamma con parámetros α y β es:*

$$h(t|\alpha,\beta) = \frac{t^{\alpha-1}\exp(-\beta t)}{\int_t^\infty s^{\alpha-1}\exp(-\beta s)\,ds},\ t > 0,\ \alpha, \beta > 0. \tag{1.6}$$

Proposición 1.3.2 *La función de azar correspondiente a una distribución $Ga(\alpha, \beta)$ coincide con la función de azar de una densidad $Ga(\alpha, 1)$ multiplicada por β.*

Demostración

Si consideramos la variable $T \sim Ga(\alpha, \beta)$ y realizamos un cambio de variable $T^* = \beta T$, entonces:

$$f_{T^*}(t^*) = \frac{1}{\beta} f_T\left(\frac{t^*}{\beta}\right) = \frac{1}{\beta} f_T(t)$$

$$S_{T^*}(t^*) = p(T^* > t^*) = p(\beta T > \beta t) = p(T > t) = S_T(t),$$

luego:

$$h_{T^*}(t^*) = \frac{f_{T^*}(t^*)}{S_{T^*}(t^*)} = \frac{1}{\beta} h_T(t) = \frac{1}{\beta} h_T\left(\frac{t^*}{\beta}\right).$$

■

Por lo tanto, el estudio de las propiedades funcionales de $h\,(t\,|\,\alpha,\beta)$ podemos reducirlo al caso $\beta = 1$ y considerar la función de t y α:

$$h(t,\alpha) = \frac{t^{\alpha-1}\exp(-t)}{\int_t^\infty s^{\alpha-1}\exp(-s)\,ds},\ t>0,\ \alpha>0,$$

de donde se tiene de forma inmediata la siguiente:

Proposición 1.3.3 $h(t,\alpha=1)=1,\ \forall t>0.$

Si $\alpha \neq 1$, se obtienen los siguientes resultados:

Proposición 1.3.4 *Para un α fijo, $h(t,\alpha)$ tiene una asíntota vertical en $t=0$ si $\alpha<1$ y $\lim_{t\to 0} h(t,\alpha)=0$ si $\alpha>1$.*

Demostración

Resulta inmediata pues:

$$\begin{aligned}\lim_{t\to 0} h(t,\alpha) &= \lim_{t\to 0} \frac{t^{\alpha-1}\exp(-t)}{\int_t^\infty s^{\alpha-1}\exp(-s)\,ds} = \frac{1}{\Gamma(\alpha)} \lim_{t\to 0} t^{\alpha-1} \\ &= \begin{cases} \infty & \text{si } \alpha<1 \\ 0 & \text{si } \alpha>1. \end{cases}\end{aligned}$$

■

Proposición 1.3.5 *Para un α fijo, $h(t,\alpha)$ tiene una asíntota horizontal en uno.*

Demostración

Integrando por partes, se tiene que:

$$\int_t^\infty s^\alpha \exp(-s)\, ds = t^\alpha \exp(-t) + \alpha \int_t^\infty s^{\alpha-1} \exp(-s)\, ds.$$

Esto es,

$$h^{-1}(t,\alpha+1) = 1 + \frac{\alpha}{t}\, h^{-1}(t,\alpha). \tag{1.7}$$

Por lo tanto, $\lim_{t\to\infty} h(t,\alpha+1) = 1$ si $\lim_{t\to\infty} h(t,\alpha) = 1$ y demostrando esto último $\forall \alpha < 1$, un simple argumento de inducción permitirá extrapolar dicha propiedad $\forall \alpha \in \mathbb{R}^+$.

Supongamos, pues, que $\alpha < 1$.

$$\int_t^\infty s^{\alpha-1} \exp(-s)\, ds < \int_t^\infty t^{\alpha-1} \exp(-s)\, ds = t^{\alpha-1} \exp(-t),$$

luego,

$$h(t,\alpha) > 1 \ \ \forall t > 0. \tag{1.8}$$

$$\begin{aligned}
\int_t^\infty s^{\alpha-1} \exp(-s)\, ds &= \int_t^\infty t^{\alpha-1} \exp\left[-s + (\alpha-1)\log\frac{s}{t}\right] ds \\
&> \int_t^\infty t^{\alpha-1} \exp\left[-s + (\alpha-1)\left(\frac{s}{t} - 1\right)\right] ds \\
&= t^{\alpha-1} \exp(1-\alpha) \int_t^\infty \exp\left[-s\left(1 + \frac{1-\alpha}{t}\right)\right] ds \\
&= t^{\alpha-1} \exp(1-\alpha) \frac{1}{1+\frac{1-\alpha}{t}} \exp\left[-t\left(1 + \frac{1-\alpha}{t}\right)\right] \\
&= \frac{1}{1+\frac{1-\alpha}{t}} t^{\alpha-1} \exp(-t),
\end{aligned}$$

por lo que,

$$h(t,\alpha) < 1 + \frac{1-\alpha}{t} \quad \forall t > 0. \tag{1.9}$$

De (1.8) y (1.9) obtenemos que $\lim_{t\to\infty} h(t,\alpha) = 1$. ■

Proposición 1.3.6 *Considerada $h(t,\alpha)$ como función de t, para un α fijo es monótona decreciente si $\alpha < 1$ y monótona creciente si $\alpha > 1$.*

Demostración

Haciendo el cambio de variable $u = s/t$ se tiene que:

$$h^{-1}(t,\alpha) = \exp(t) \int_1^\infty t\, u^{\alpha-1} \exp(-tu)\, du,$$

y derivando respecto a t y utilizando (1.7):

$$\begin{aligned} \frac{dh^{-1}(t,\alpha)}{dt} &= \exp(t) \left[\int_1^\infty t\, u^{\alpha-1} \exp(-tu)\ du + \right. \\ &\quad \left. + \int_1^\infty [u^{\alpha-1} \exp(-tu) - t\, u^\alpha \exp(-tu)]\, du \right] \\ &= \frac{t+1}{t} h^{-1}(t,\alpha) - h^{-1}(t,\alpha+1) \\ &= \frac{1}{th(t,\alpha)} [t + 1 - \alpha - th(t,\alpha)]. \end{aligned} \tag{1.10}$$

Si $\alpha < 1$, por (1.9) se cumple que $t + 1 - \alpha - th(t,\alpha) > 0, \forall t > 0$ y, por consiguiente, también es positiva la derivada anterior y $h(t,\alpha)$ es monótona decreciente $\forall t > 0$.

Si $\alpha > 1$:

$$\begin{aligned}\int_t^\infty s^{\alpha-1}\exp(-s)\,ds &= \int_t^\infty t^{\alpha-1}\exp\left[-s+(\alpha-1)\log\frac{s}{t}\right]\,ds \\ &< \int_t^\infty t^{\alpha-1}\exp\left[-s+(\alpha-1)\left(\frac{s}{t}-1\right)\right]ds \\ &= t^{\alpha-1}\exp(-t)\frac{t}{t+1-\alpha} \text{ si } t > \alpha-1,\end{aligned}$$

por lo tanto $1 < h(t,\alpha)\frac{t}{t+1-\alpha}$ si $t > \alpha - 1$, o lo que es lo mismo $t+1-\alpha-th(t,\alpha) < 0$ si $t > \alpha - 1$.

Si $t \leq \alpha - 1$ entonces $t+1-\alpha-th(t,\alpha) \leq -th(t,\alpha) < 0$, con lo que por la igualdad (1.10) se tiene que $\frac{dh^{-1}(t,\alpha)}{dt} < 0, \forall t > 0$ y $h(t,\alpha)$ es monótona creciente $\forall t > 0$. ■

En la figura 1.1 puede observarse la forma funcional del azar Gamma para $\beta = 1$ y distintos valores del parámetro α.

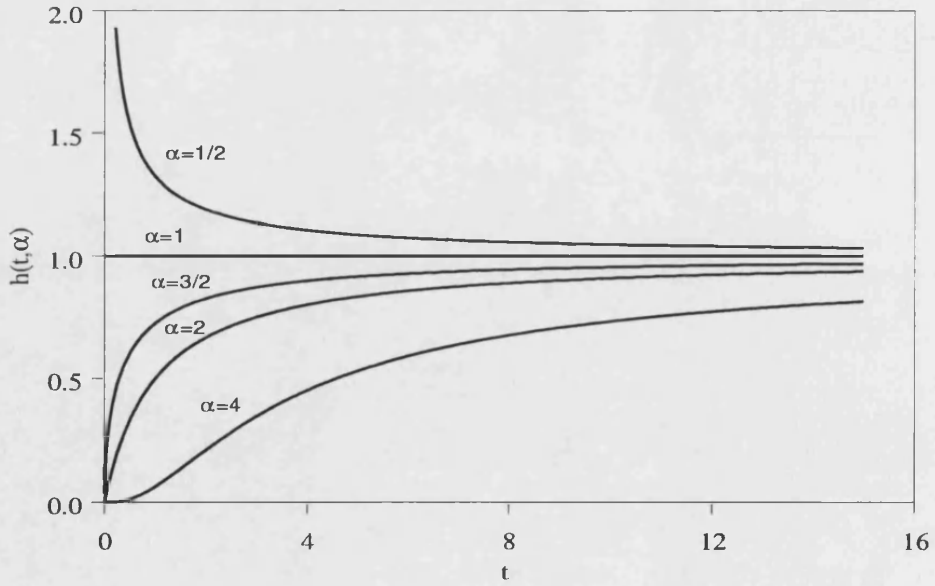

Figura 1.1: *Función de azar Gamma.*

1.4 Métodos MCMC

Desde el trabajo de Gelfand y Smith (1990) se ha producido un incremento espectacular en la utilización de los métodos bayesianos, pues los métodos MCMC permiten el análisis de modelos complejos y no son difíciles de aplicar. De este modo, ha sido posible utilizar modelos más realistas en todas las áreas de la estadística aplicada. Esto ha sido así también en el análisis bayesiano de datos de vida con covariables, donde ya es habitual la utilización de métodos MCMC para estimar alguna característica de la distribución de interés (ver, por ejemplo, Gilks et al., 1993; Best et al., 1996; Chib y Greenberg, 1996).

1.4.1 Preliminares

En cuanto se complican un poco los modelos estocásticos utilizados en el análisis de datos, buscando modelos más realistas, suele ocurrir que el conocimiento acerca de la distribución final se limite a su forma funcional hasta una constante de proporcionalidad nada sencilla de calcular. Los métodos MCMC constituyen, en este caso, no sólo la solución comúnmente empleada sino en muchas ocasiones la única, si bien estas técnicas también son perfectamente aplicables dentro de un contexto frecuentista en el caso de que la distribución a analizar presente cierto grado de dificultad.

La idea subyacente en la aplicación de los métodos MCMC es la obtención de una muestra de una determinada función de densidad de probabilidades $p(\theta|y)$, dada la dificultad o imposibilidad que conlleva el estudio analítico de la misma. Supóngase que estando interesado en la densidad $p(\theta|y)$, y reconociendo la dificultad inherente al análisis basado en la mis-

ma, uno puede aproximarla, con un alto grado de precisión, mediante una muestra obtenida a partir de $p(\theta \mid y)$. En este caso, el problema deja de ser la propia densidad y pasa a ser cómo y de qué forma obtener una muestra de $p(\theta \mid y)$. En realidad, se trata de generalizar el método científico obteniendo conclusiones estadísticas a partir de la experimentación y de los resultados empíricos basados en datos obtenidos mediante simulación.

Para la obtención de una muestra proveniente de $p(\theta \mid y)$ existen dos grandes tipos de métodos de simulación: los *métodos de simulación estática* y los *métodos de simulación dinámica.* La diferencia principal entre ellos es que los segundos utilizan algún tipo de iteración en cadenas de Markov y requieren convergencia en las mismas (ver apéndice B), mientras que los primeros llevan a cabo una simulación directa. En contrapartida, los métodos de simulación estática son de difícil aplicación en problemas con muchos parámetros.

En la siguiente sección se realiza una revisión de los métodos de simulación dinámica más empleados y en cualquier manual clásico de simulación (Knuth, 1981; Devroye, 1986; Ripley, 1987; por ejemplo) pueden consultarse algunos de los métodos directos más utilizados. Además de los métodos *de la transformada inversa* y *cociente de uniformes*, cabe citar expresamente el *método S.I.R.* (Rubin, 1988; Smith y Gelfand, 1992) y el *método de aceptación-rechazo* (Geweke, 1989; Zellner y Rossi, 1984). Éste requiere la especificación de una función de densidad de probabilidades, llamada *función importante*, similar a la densidad objetivo y fácilmente muestreable, además de una buena cota superior del cociente de ésta y la función importante. Cuando no es muy costoso el cumplimiento de estas dos premisas es realmente un método de simulación verdaderamente útil.

Una vez obtenida una muestra de la distribución de interés, $p(\theta \mid y)$, de

tamaño suficientemente grande, cualquier característica de la misma puede aproximarse utilizando la distribución empírica. Así por ejemplo, cualquier momento de la distribución $p(\theta \mid y)$ puede estimarse por Monte Carlo mediante el correspondiente momento muestral y proporcionar, a su vez, el error de estimación.

1.4.2 Métodos de simulación dinámica

La mayor parte de los métodos de simulación dinámica son particularizaciones del propuesto por Hastings (1970). En su trabajo generaliza un método propuesto por un grupo de físicos en la década de los cincuenta (Metropolis et al., 1953) que había pasado completamente desapercibido para la mayoría de los estadísticos. Básicamente, se trata de construir una cadena de Markov irreducible cuya distribución estacionaria sea la distribución de interés. Después de realizadas un gran número de iteraciones en la misma y tras una fase inicial transitoria, la cadena obtenida es una muestra proveniente de la distribución objetivo (Gilks et al., 1996; Draper, 1997; Gamerman, 1997).

Algoritmo de Hastings

Está basado en un algoritmo de aceptación-rechazo con función importante, $f(\theta \mid \theta_t)$, de modo que si en el instante t el estado de la cadena es θ_t, la probabilidad de cambiar a un nuevo estado, θ_{t+1}, en el instante $t+1$, es:

$$\alpha(\theta_t, \theta_{t+1}) = \min\left[\frac{p(\theta_{t+1} \mid y) f(\theta_t \mid \theta_{t+1})}{p(\theta_t \mid y) f(\theta_{t+1} \mid \theta_t)}, 1\right]. \tag{1.11}$$

El esquema algorítmico es el siguiente:

INICIALIZAR θ_0; $t \leftarrow 0$
REPETIR
generar $\theta^* \sim f(\theta \mid \theta_t), u \sim Un(0,1)$
si $u \leq \alpha(\theta_t, \theta^*)$ entonces $\theta_{t+1} \leftarrow \theta^*$
si no $\theta_{t+1} \leftarrow \theta_t$
$t \rightarrow (t+1)$.

Hastings (1970) demostró que, en el caso discreto, partiendo de cualquier valor inicial θ_0 y casi con cualquier función importante $f(\theta \mid \theta_t)$, la cadena así construida es ergódica y, por tanto, tiene distribución estacionaria (ver apéndice B) y la distribución estacionaria es precisamente $p(\theta \mid y)$. La extensión a cadenas de Markov con espacio de estados continuo puede consultarse en Tierney (1994).

Algoritmo de Metropolis

Casi dos décadas antes de la aparición del crucial trabajo de Hastings (1970), un grupo de físicos de Los Álamos (Nuevo Méjico) había propuesto un método de Monte Carlo para el cálculo de las ecuaciones de estado en sistemas de esferas rígidas (Metropolis et al., 1953). Como ya se ha comentado con anterioridad, este método es una particularización del algoritmo de Hastings al considerar tan sólo funciones importantes simétricas en el sentido de que $f(\theta^* \mid \theta_t) = f(\theta_t \mid \theta^*)$. De este modo, el cociente $\frac{f(\theta_t \mid \theta^*)}{f(\theta^* \mid \theta_t)}$ cancela en (1.11) y el algoritmo de Metropolis queda como sigue:

INICIALIZAR θ_0; $t \leftarrow 0$
REPETIR
generar $\theta^* \sim f(\theta \mid \theta_t)$ simétrica, $u \sim Un(0,1)$
si $u \leq \alpha(\theta_t, \theta^*)$ entonces $\theta_{t+1} \leftarrow \theta^*$
si no $\theta_{t+1} \leftarrow \theta_t$
$t \rightarrow (t+1)$,

donde:

$$\alpha(\theta_t, \theta^*) = \min\left[\frac{p(\theta^* \mid y)}{p(\theta_t \mid y)}, 1\right]$$

Algoritmo de Gibbs

El algoritmo de Gibbs (Geman y Geman, 1984; Gelfand y Smith, 1990; Casella y George,, 1992) es, sin duda, el método de simulación dinámica más utilizado en los últimos tiempos. También resulta un caso particular del método de Hastings, en el que se obtienen muestras de las distribuciones condicionales completas componente a componente. Si consideramos el vector paramétrico $\boldsymbol{\theta}$, la estrategia radica en descomponerlo en bloques y definir la función de transición de probabilidades como el producto de las densidades condicionales completas (la densidad condicional de cada bloque dados los datos y el resto de parámetros no incluidos en el bloque). El gran atractivo de este método es que es habitual en inferencia bayesiana encontrar en muchas aplicaciones distribuciones condicionales completas tratables (a partir de las cuales no resulta complicado simular) de una distibución objetivo intratable. El esquema algorítmico de este método es el siguiente:

INICIALIZAR

descomponer $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$

$\boldsymbol{\theta}^{(0)}$; $t \leftarrow 0$

REPETIR

generar $\theta_1^{(1)} \sim p\left(\theta_1 \,|\, y, \theta_2^{(0)}, \ldots, \theta_k^{(0)}\right)$

generar $\theta_2^{(1)} \sim p\left(\theta_2 \,|\, y, \theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_k^{(0)}\right)$

.

.

.

generar $\theta_{k-1}^{(1)} \sim p\left(\theta_{k-1} \,|\, y, \theta_1^{(1)}, \ldots, \theta_{k-2}^{(1)}, \theta_k^{(0)}\right)$

generar $\theta_k^{(1)} \sim p\left(\theta_k \,|\, y, \theta_1^{(1)}, \ldots, \theta_{k-1}^{(1)}\right)$

$t \rightarrow (t+1)$.

Bajo condiciones muy generales, la sucesión $\left(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}\right)$ construida de este modo es una realización de una cadena de Markov con $p(\boldsymbol{\theta} \,|\, y)$ como distribución estacionaria (Chan, 1993; Roberts y Polson, 1994; Tierney, 1994; Liu et al., 1995).

Algoritmo de Metropolis dentro de Gibbs

Aunque este método es tan sólo un caso particular del algoritmo de Gibbs en cuanto a la forma de muestrear de las distribuciones condicionales completas, es citado expresamente dada su posterior utilización en el presente trabajo. Muller (1991), propuso la utilización del método de Metropolis para simular a partir de las condicionales completas, cuando no es

posible un muestreo directo de las mismas. Es decir, se trata de utilizar una determinada función importante para el muestreo dentro de algunas etapas de actualización del algoritmo de Gibbs. Aunque, rigurosamente, esto supondría una iteración adicional dentro de dichas etapas hasta alcanzar convergencia en el método de Metropolis, debe notarse que con una única iteración se reproduce dicho algoritmo con un esquema de visitas de una sola vez a cada componente (Gilks et al., 1995).

1.4.3 Problemas prácticos en la implementación de métodos MCMC

Inherentes a la implementación de cualquier método MCMC (por ejemplo, el genérico método de Hastings), aparecen tres cuestiones de capital importancia: la elección del punto inicial, la elección de la función importante y cómo obtener la muestra final (número de iteraciones iniciales a desechar hasta la convergencia y qué iteraciones considerar) para aproximar $p(\theta \mid y)$. Dada su relevancia en los resultados finalmente alcanzados, en esta sección se pretende realizar una somera revisión de las respuestas más habituales dadas en la literatura a estas cuestiones. Si se desea un tratamiento más profundo y genérico de todo lo relacionado con la implementación de métodos MCMC pueden consultarse, por ejemplo, Smith y Roberts (1993), Gilks et al. (1996) y Roberts y Sahu (1997).

Elección del punto inicial θ_0

Una buena elección del punto inicial con el que comenzar las iteraciones en la cadena de Markov puede ser fundamental para que ésta alcance su distribución estacionaria rápidamente. La cadena debe mezclarse bien en el sentido de visitar frecuentemente cualquier región con alta densidad final.

La idea es escoger un punto inicial cercano al centro de la distribución objetivo, por lo que, cuando sea posible, un buen valor inicial es la moda de la distribución final. En cualquier caso, si la elección del punto inicial es única cabe la posibilidad de no visitar regiones de alta densidad en distribuciones multimodales. En efecto, partiendo de un punto inicial fijo podemos llegar a visitar la región de una de las modas de $p(\theta \mid y)$, siendo difícil (en el sentido de muy costosa en número de iteraciones) la convergencia a las demás modas. Una solución a este problema fue proporcionada por Gelman y Rubin (1992a, 1992b) al sugerir el empleo de varias cadenas de Markov con valores iniciales muy dispersos. En sus trabajos proponen comparar características de dichas cadenas, utilizando técnicas similares a las del ANOVA, para comprobar si todas ellas han convergido a la misma solución.

Como alternativa puede utilizarse el método propuesto por Geman y Geman (1984) para localizar todas las modas de la distribución que, aunque puede resultar muy lento en aplicaciones prácticas, tiene la atractiva característica de que resulta verdaderamente complicado encontrar un máximo local y permanecer en él. Geyer y Thompson (1995) desarrollaron otro método de búsqueda de posibles valores iniciales que daba lugar a que las cadenas se mezclaran rápidamente, permitiendo el análisis de problemas con muchos parámetros.

De todas formas, hay que resaltar que son atípicas las situaciones en las que la distribución final es multimodal. Generalmente, esto sólo ocurre cuando la distribución inicial es muy informativa contradiciendo la información de los datos. Por lo tanto, lo aconsejable puede ser intentar detectar multimodalidad y, si no se halla, utilizar una única cadena. Con varias cadenas se pierde efectividad al tener que desechar unas iteraciones iniciales en cada una de ellas (Geyer, 1992).

Elección de la función importante $f(\theta \mid \theta_t)$

Es precisamente por el propio desconocimiento que se tiene de la distribución objetivo que la elección de una adecuada función importante es, quizá, la cuestión más oscura y de difícil respuesta de las tres planteadas. Es conocido que generalmente la mayoría de los métodos MCMC garantizan convergencia a la distribución estacionaria en un número finito de iteraciones, independientemente de la función importante elegida. No obstante, ésta puede ser extremadamente lenta para algunas elecciones arbitrarias haciendo inviable la aplicación del algoritmo.

Obviamente, buenas elecciones para $f(\theta \mid \theta_t)$ son aquellas densidades similares a $p(\theta \mid y)$ que conduzcan a cadenas de Markov que se mezclen bien. Desgraciadamente, no existe una propuesta concreta que globalice su utilización en distintas aplicaciones. A modo de indicación, y como estrategia genérica que suele proporcionar buenos resultados, cabe apuntar un par de ideas.

En primer lugar, resulta conveniente elegir una función importante, $f(\theta \mid \theta_t)$, de la que sea sencillo y rápido simular y con una variabilidad mayor que la de la distribución final (Tierney, 1994). En segundo lugar, la elección puede realizarse de modo que la esperanza del nuevo estado θ^*, dado que ha sido aceptado el movimiento desde el estado θ_t, sea precisamente θ_t, esto es, $E_f(\theta^* \mid \theta_t) = \theta_t$, (Gilks et al., 1996).

Obtención de la muestra final

La obtención de una muestra de la distribución final con la que aproximar cualquier característica desconocida de la misma requiere, básicamente, la especificación del número de iteraciones iniciales a desechar (correspondientes a un período transitorio donde se efectúan movimientos en la cadena

para converger a la distribución estacionaria) y el número de iteraciones finales de la cadena de Markov a considerar (para ello pueden efectuarse saltos en la misma con el fin de disminuir la correlación y facilitar su almacenamiento).

Es posible que el problema del diagnóstico de la convergencia sea el tema de estudio más tratado por los investigadores de los métodos MCMC. No obstante, a pesar de que se hallan propuestos multitud de métodos teóricos de diagnóstico en la literatura, todavía no existe uno de ellos genérico y de fácil utilización en la mayoría de aplicaciones prácticas. Debido a ello, los más informales de monitorización gráfica de algún determinado cuantil o de la autocorrelación (Gelfand et al., 1990; Gelfand y Smith, 1990), a la vez que de muy sencilla y rápida implementación, siguen siendo los más utilizados. No se pretende en esta sección realizar una exhaustiva revisión de los métodos de diagnosis de convergencia (Cowles y Carlin, 1996; Brooks y Roberts, 1997), sino tan sólo citar brevemente los posteriormente utilizados a lo largo del trabajo.

Aparte de la propia traza de la serie temporal correspondiente a algún parámetro determinado, pueden ser de gran utilidad las gráficas de la *función de autocorrelación (ACF)* y de la *función de autocorrelación parcial (ACPF)* (véase, por ejemplo, Box y Jenkins, 1976). Es bastante frecuente en la aplicación de los métodos MCMC, encontrar que la cadena $\{\theta_1, \ldots, \theta_n\}$ tiene un comportamiento muy similar al de un *proceso autorregresivo de orden 1, AR(1)*. En ese caso, está cuantificado el error stándard de la media muestral $\overline{\theta}$:

$$S.E.\left(\overline{\theta}\right) = \frac{s}{n}\sqrt{\frac{1+\hat{\rho}_1}{1-\hat{\rho}_1}},$$

siendo $\hat{\rho}_1$ la correlación serial y s la desviación típica de la serie. Por lo tanto, la longitud final de la cadena es un problema de tamaño muestral

que puede resolverse utilizando técnicas AR(1).

La función de autocorrelación correspondiente a un proceso AR(1) es decreciente geométricamente en $\hat{\rho}_1$ positivo y la función de autocorrelación parcial sólo tiene su primer valor distinto de cero (es igual a $\hat{\rho}_1$). Procesos autorregresivos de órdenes superiores pueden ser rápidamente diagnosticados con la función de autocorrelación parcial.

Geweke (1992) propuso un sencillo método de diagnóstico de la convergencia de una cadena de Markov basado en técnicas de análisis espectral de series temporales. Si suponemos que $\{\theta_1, \ldots, \theta_n\}$ es la cadena obtenida después de n iteraciones, su idea era comparar una subcadena inicial, $\{\theta_i : i = 1, \ldots, n_a\}$, con una subcadena final, $\{\theta_i : i = n^*, \ldots, n\}$, en orden a establecer si no hay diferencias entre las medias espectrales. Para ello, si los radios n_a/n y n_b/n, donde $n_b = n - n^* + 1$, son fijos con

$$\frac{n_a + n_b}{n} < 1$$

y si la cadena de Markov es estacionaria, entonces:

$$Z_n = \frac{\overline{\theta_a} - \overline{\theta_b}}{\sqrt{\frac{1}{n_a}\hat{S}^a_\theta(0) + \frac{1}{n_b}\hat{S}^b_\theta(0)}} \xrightarrow{d} N(0,1) \text{ cuando } n \to \infty.$$

$\hat{S}^a_\theta(0)$ y $\hat{S}^b_\theta(0)$ son estimadores espectrales de la varianza. Geweke (1992) recomienda utilizar $n_a = n/10$ y $n_b = n/2$.

Ha de hacerse notar que este test proporciona una condición necesaria para la convergencia, pero no suficiente. De modo que en la práctica habitual es útil en el sentido de informarnos de cuándo la cadena no ha alcanzado la convergencia, no de cuándo la alcanzará.

Raftery y Lewis (1992) propusieron un método para calcular el número de iteraciones necesarias para la estimación de algún cuantil de la distribu-

ción final. Este método permite obtener el número de iteraciones iniciales a desechar y el tamaño final de la cadena a considerar, así como el *adelgazamiento* de la misma (número de saltos entre iteraciones).

Ha de fijarse el orden del cuantil (0.025 y 0.975 son los habituales), la precisión deseada para la estimación (por ejemplo, 0.005) y la probabilidad de obtener la precisión requerida. Raftery y Lewis (1992) proponen sustituir la cadena de Markov por un proceso binario, de modo que el análisis del mismo permite extraer las conclusiones acerca de la estacionariedad de la cadena.

El llamado adelgazamiento de la cadena de Markov tiene, fundamentalmente dos objetivos. Un primero es el de disminuir la autocorrelación para obtener muestras casi independientes y un segundo (y, quizá, más importante) es el de permitir el almacenamiento informático de la cadena. Cuando hay bastantes parámetros involucrados en el análisis es habitual necesitar bastantes iteraciones en la cadena de Markov (quizá varios miles). El almacenar una iteración de cada 25, por ejemplo, puede suponer en ese caso un ahorro de varios megabytes a costa de una pérdida mínima de información.

Existen en la actualidad distintas herramientas informáticas para la utilización de los métodos MCMC. Además de los conocidos lenguajes de programación para la confección personal de los propios programas, es cita obligada el paquete estadístico S-PLUS (Becker et al., 1988), que dada su agradable sintaxis y su naturaleza interactiva constituye un idóneo entorno para el trabajo estadístico, si bien algunas implementaciones MCMC requieren otras alternativas. El programa BUGS (*bayesian inference using Gibbs sampling*), debido a Spiegelhalter et al. (1995), constituye una adecuada herramienta de trabajo en la implementación del muestreo de Gibbs. Para el análisis de la convergencia de cadenas de Markov, Best et al. (1995)

han desarrollado un conjunto de subrutinas en S-PLUS muy agradables de trabajar y que se encuentran integradas en una aplicación llamada CODA (*convergence diagnosis and output analysis software for Gibbs sampling output*). Su utilización es bastante sencilla y permite aplicar un buen número de métodos de convergencia. Tan sólo requiere presentar las iteraciones de la cadena de Markov en un fichero con un formato específico.

Capítulo 2

Modelo semiparamétrico aditivo

2.1 Introducción

En este capítulo se presenta un modelo semiparamétrico para el análisis de datos de supervivencia definido a partir de su función de azar. La intuitiva y sencilla interpretación de esta función como una tasa instantánea de fallo motiva la modelización de este tipo de datos mediante la misma en lugar de utilizar la función de densidad. Consideramos la función de azar como la suma de dos funciones, una relativa a una distribución paramétrica concreta y otra, no paramétrica. Con ello se pretende diversificar la forma del azar y no restringirnos a las habituales de las distribuciones de supervivencia conocidas. Se trata de un modelo en poblaciones en el que la parte paramétrica es peculiar de cada individuo y en la que incorporamos sus características propias en forma de vector de covariables, mientras que el azar no paramétrico poligonal es común a todos ellos. Estudiamos con mayor detalle, en la parte final del capítulo, una particularización del modelo anterior en la que la distribución paramétrica es Gamma. El realizar el vínculo entre las dos partes mediante una adición simplifica notablemente el posterior análisis bayesiano del modelo.

2.2 Presentación del modelo

Supongamos realizada una partición del eje temporal $[0, +\infty)$, $I_1, I_2, \ldots, I_{g+1}$, donde $I_j = [a_{j-1}, a_j)$ para $j = 1, \ldots, g+1$, con $a_0 = 0$ y $a_{g+1} = +\infty$. El resto de valores, $a_1, \ldots, a_g$, constituyen lo que vamos a denominar *divisiones no triviales* del eje temporal. Definimos $\tau(t)$, la función poligonal positiva:

$$\tau(t) = \sum_{j=1}^{g+1} 1_{I_j} \frac{t\,(\tau_{j+1} - \tau_j) + \tau_j\, a_j - a_{j-1}\,\tau_{j+1}}{a_j - a_{j-1}},\ t > 0, \qquad (2.1)$$

con $\tau_j > 0$, $j = 1, \ldots, g+2$, $\tau_{g+2} = \tau_{g+1}$ y 1_{I_j}, la función indicatriz en el intervalo I_j.

De esta forma, $\tau(t)$ es una función no negativa definida sobre la semirecta real positiva, por lo que cumple todas las condiciones de una función de azar.

La definición de $\tau(t)$ es similar a todas las funciones de azar no paramétricas utilizadas en el análisis de supervivencia, pero en vez de considerarla escalonada la hemos construido poligonal para asegurar continuidad.

Definición 2.2.1 *El modelo semiparamétrico aditivo es un modelo de supervivencia en el que la función de azar se descompone del siguiente modo:*

$$h(t) = \vartheta(t) + \tau(t),\ t > 0, \qquad (2.2)$$

donde $\vartheta(t)$ es la función de azar correspondiente a una distribución paramétrica conocida, excepto por el valor concreto de su vector paramétrico $\boldsymbol{\theta}$, y $\tau(t)$ es una función de azar poligonal.

De este modo, los parámetros correspondientes a este modelo son los de la distribución paramétrica, $\boldsymbol{\theta}$, y los relativos al azar poligonal, $\tau_1, \ldots, \tau_{g+1}$.

Definición 2.2.2 *Decimos que un modelo de supervivencia es en poblaciones y semiparamétrico aditivo si se trata de un modelo jerárquico, dado por el azar (2.2), y en el que los parámetros $\tau_1, \ldots, \tau_{g+1}$ son comunes a todos los individuos, pero el vector $\boldsymbol{\theta}$, considerado tras las transformaciones oportunas como perteneciente a todo $\mathbb{R}^p$ y cuya primera componente es el logaritmo de la media (o de alguna medida de localización) de la parte paramétrica, es tal que:*

$$\begin{aligned} \boldsymbol{\theta_2} &\sim N_{p-1}\left(\boldsymbol{\theta_2} \,|\, \boldsymbol{\nu}, \Sigma\right) \\ \theta_1 \,|\, \boldsymbol{\theta_2} &\sim N\left(\theta_1 \,|\, \mathbf{b}'\mathbf{x}, \sigma_\theta^2\right), \end{aligned}$$

donde $\nu, \Sigma, \mathbf{b}$ y σ_θ^2 son los hiperparámetros del modelo y $\mathbf{x}$ es un vector de covariables asociado a cada individuo.

Este modelo en poblaciones semiparamétrico aditivo guarda ciertas analogías con el modelo de Cox al considerar ambos dos partes, una paramétrica y otra no paramétrica, en la función de azar e incorporar las covariables vía la parte paramétrica. Sin embargo, la principal diferencia entre los dos, aparte de que el propuesto se trata de un modelo en poblaciones, es que el considerar el vínculo entre las dos partes mediante una adición en lugar de una multiplicación, simplifica bastante todo el aparato matemático del modelo y permite el análisis bayesiano del mismo.

Una interesante propiedad de las funciones de supervivencia correspondientes a la parte paramétrica y no paramétrica del modelo semiparamétrico aditivo se recoge en el siguiente resultado.

Proposición 2.2.1 *La función de supervivencia del modelo semiparamétrico aditivo puede expresarse como el producto de la supervivencia paramétrica*

y la supervivencia poligonal:

$$S(t) = S_0(t)\, S_T(t),\ t > 0,$$

Demostración

Dado que $S(t) = \exp[-H(t)]$, $t > 0$, y que la función de azar acumulado correspondiente al modelo semiparamétrico aditivo es:

$$H(t) = \int_0^t \tau(s)\, ds = \int_0^t (\vartheta(s) + \tau(s))\, ds = H_0(t) + H_T(t),\ t > 0,$$

con $H_0(t)$ y $H_T(t)$ las funciones de azar acumulado correspondientes a las partes no paramétrica y paramétrica del modelo, respectivamente. De forma inmediata resulta que $S(t) = S_0(t)\, S_T(t)$, $t > 0$. ■

El modelo semiparamétrico aditivo puede considerarse como un caso particular de modelo de *riesgos competitivos* en el que existen dos riesgos independientes. En efecto, pues la función de supervivencia del mínimo de dos variables aleatorias independientes es el producto de las funciones de supervivencia.

En situaciones *puras* de riesgos competitivos, aquellas en las que se estudian diversas causas concretas que pueden causar el fallo del individuo, no suele ser adecuado suponer independencia. Por ello, no proponemos este modelo como modelo de riesgos competitivos, pero esta interpretación del mismo puede ayudar a determinar si es adecuado para la modelización de un problema concreto.

Con el fin de poder contemplar bajo nuestro modelo situaciones de supervivencia con azares no proporcionales, consideramos distinto vector paramétrico del azar poligonal para grupos diferentes de individuos. De este

modo, si tenemos n datos de supervivencia divididos en M grupos, el vector paramétrico completo es:

$$\left(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}, \boldsymbol{\tau}^{(1)}, \ldots, \boldsymbol{\tau}^{(M)}, \nu, \Sigma, \mathbf{b}, \sigma_\theta^2\right),$$

donde $\boldsymbol{\tau}^{(j)} = \left(\tau_1^{(j)}, \ldots, \tau_{g+1}^{(j)}\right)$, $j = 1, \ldots, M$.

Cabe resaltar que cualquier covariable cualitativa puede entrar en el modelo vía la parte paramétrica (como un conjunto de covariables *dummy*) o como grupos de individuos con distintos azares poligonales. Las covariables continuas se incluyen todas ellas mediante el azar paramétrico. Esta es la práctica habitual en análisis de supervivencia, donde a los grupos con azares no paramétricos distintos se les denomina *estratos*.

Algunos ejemplos de modelos de supervivencia que pueden ser contemplados dentro del modelo semiparamétrico aditivo que proponemos incluyen un modelo *Exponencial-poligonal aditivo* cuando el azar paramétrico es el correspondiente a la distribución Exponencial, $\vartheta(t) = \beta$, $\beta > 0$. Modelizamos $\log\beta \sim N\left(\log\beta \,|\, \mathbf{b}'\mathbf{x}, \sigma_\beta^2\right)$ y resulta un modelo similar al de Cox, pero tratándose de un modelo en poblaciones y con azares aditivos (entre la parte paramétrica y no paramétrica) en vez de multiplicativos. Su función de azar es:

$$h(t) = \tau(t) + \exp\left(\mathbf{b}'\mathbf{x} + \sigma_\beta \epsilon\right)$$

donde $\epsilon \sim N(0,1)$ es una covariable no observada asociada a cada individuo.

Mayor interés tiene el llamado modelo *Gamma-poligonal aditivo*, que es el que desarrollamos con detalle en este trabajo, y que se tiene cuando la distribución paramétrica de (2.2) es la distribución Gamma.

2.2.1 Modelo Gamma-poligonal aditivo

Si utilizamos la distribución Gamma para modelizar la parte paramétrica del azar (2.2) tenemos el siguiente modelo.

Definición 2.2.3 *Se define el modelo Gamma-poligonal aditivo como aquel modelo con función de azar (2.2), donde $\vartheta(t)$ es la función de azar de una distribución $Ga(\alpha, \beta)$ y donde modelizamos la relación estocástica de los parámetros mediante:*

$$\begin{aligned} \beta &\sim N\left(\log\beta \,|\, \mu_\beta, \sigma_\beta^2\right) \\ \alpha \,|\, \beta &\sim N\left(\log\frac{\alpha}{\beta} \,|\, \mathbf{b}'\mathbf{x}, \sigma_\alpha^2\right). \end{aligned}$$

La distribución Gamma supone una generalización de la distribución Exponencial, por lo que la función de azar del modelo Gamma-poligonal aditivo tiene una mayor riqueza en la forma funcional y puede adecuarse mejor a diferentes situaciones de supervivencia.

Dado el carácter de riesgos competitivos que tienen los modelos semiparamétricos aditivos, si uno de los dos riesgos es mucho mayor que el otro, en la práctica, el modelo se comportará como si sólo existiera ese riesgo. En efecto, en esas situaciones una de las funciones de supervivencia permanecerá en valores próximos a uno, mientras que la otra tomará valores significativamente menores que uno haciendo que su producto (función de supervivencia del modelo semiparamétrico aditivo) sea prácticamente igual a la segunda. Esto es así también en el modelo Gamma-poligonal aditivo como a continuación mostramos en algunos ejemplos.

En la figura 2.1 se representan dos funciones de azar poligonales $\tau_1(t)$ y $\tau_2(t)$. Puede observarse que la segunda de ellas alcanza valores más grandes

dando lugar, consecuentemente, a supervivencias menores.

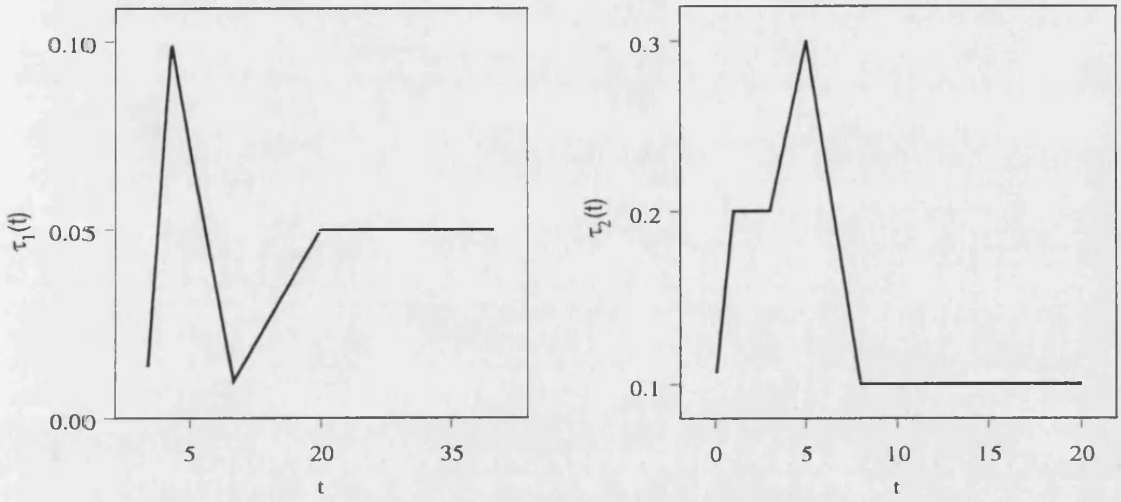

Figura 2.1: *Funciones de azar poligonales.*

En la figura 2.2 hemos representado la parte paramétrica Gamma de la función de azar de un individuo concreto para el que $\alpha = 665$ y $\beta = 20$.

La figura 2.3 recoge la parte paramétrica de la función de supervivencia del individuo anterior, la supervivencia poligonal correspondiente a las dos funciones $\tau_1(t)$ y $\tau_2(t)$, y la función de supervivencia de dicho individuo con ambos modelos Gamma-poligonal aditivos.

Utilizando en el modelo Gamma-poligonal aditivo la función $\tau_1(t)$, se observa en la figura 2.3 que las dos partes, paramétrica y no paramétrica, influyen en la supervivencia del individuo a pesar de la diferencia entre sus funciones de azar. Dado que al principio la función de azar de la parte paramétrica es mucho más pequeña que la de la poligonal, con tiempos pequeños la función de supervivencia del individuo es prácticamente la supervivencia poligonal, haciendo que la función de supervivencia decrezca rápidamente

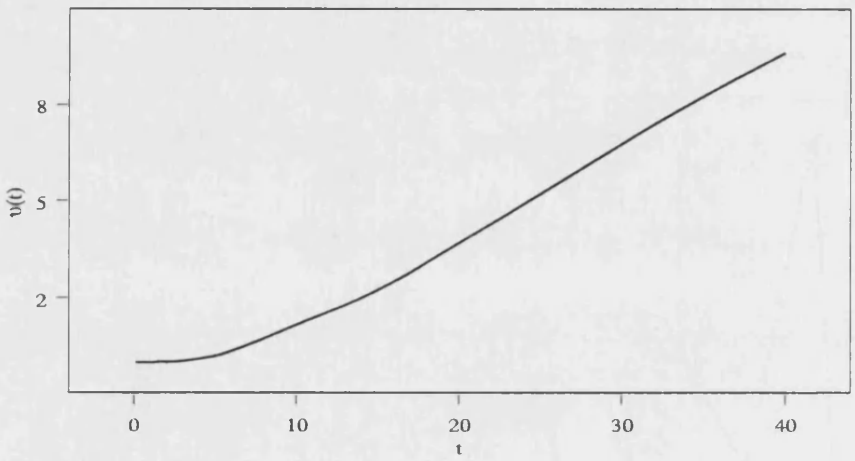

Figura 2.2: *Parte paramétrica de la función de azar del modelo Gamma-poligonal aditivo para un individuo con parámetros* $\alpha = 665$ *y* $\beta = 20$.

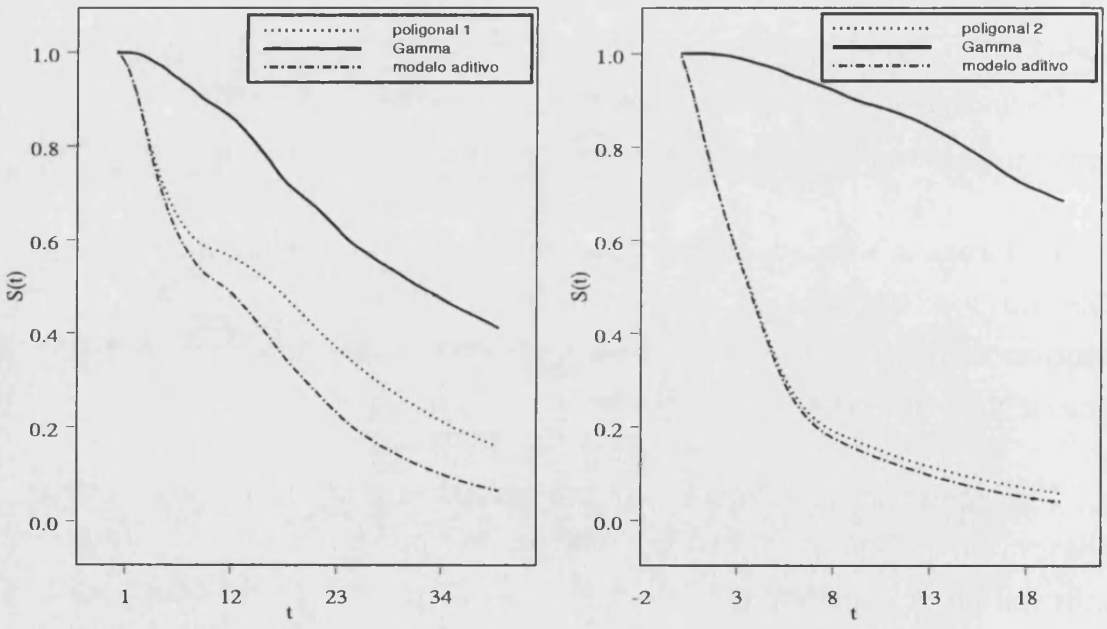

Figura 2.3: *Funcion de supervivencia del modelo Gamma-poligonal aditivo para un individuo con parámetros* $\alpha = 665$ *y* $\beta = 20$.

para posteriormente verse influida por la parte paramétrica. Si la función de azar poligonal es $\tau_2(t)$, los valores tan elevados de la misma desde el principio del rango temporal provocan el fallo del individuo con rapidez, dando lugar a una función de supervivencia para él casi coincidente con la función de supervivencia poligonal.

Otra característica importante del modelo Gamma-poligonal aditivo es que sus funciones de azar están estocásticamente ordenadas en términos de individuos medios. Esto es, aunque al tratarse de un modelo en poblaciones es posible el solapamiento de las funciones de azar para individuos concretos se cumple el siguiente resultado.

Proposición 2.2.2 *Dos funciones de azar del modelo Gamma-poligonal aditivo, $h_1(t)$ y $h_2(t)$, correspondientes a dos individuos con el mismo valor para el parámetro β están estocásticamente ordenadas. Esto es, se verifica que $h_1(t) > h_2(t)\ \forall t > 0$ si y sólo si $\alpha_1 < \alpha_2$.*

Demostración

La ordenación de los azares del modelo Gamma-poligonal aditivo viene dada por la propia ordenación de las funciones de azar de la parte Gamma (Bermúdez y Beamonte, 1997), al no depender del individuo la función de azar poligonal y, por tanto, ser idéntica para los dos individuos.

Supongamos que (α_1, β) y (α_2, β) son los parámetros Gamma correspondientes a las funciones de azar $\vartheta_1(t)$ y $\vartheta_2(t)$. El logaritmo de la función de azar Gamma (1.6) es:

$$\log \vartheta(t) = (\alpha - 1)\log t - \beta t - \log \int_t^{+\infty} x^{\alpha-1} \exp(-\beta x)\, dx.$$

Y se tiene que:

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\log\vartheta(t) = \log t - \frac{\int_t^{+\infty} x^{\alpha-1}\log x\exp(-\beta x)\,dx}{\int_t^{+\infty} x^{\alpha-1}\exp(-\beta x)\,dx} < 0,$$

pues $\log x > \log t\ \forall x > t$. Por lo tanto, $\vartheta(t|\alpha,\beta)$ es estrictamente decreciente en α, $\forall\beta > 0$ y se cumple que $\vartheta_1(t) > \vartheta_2(t), \forall t > 0$ si y sólo si $\alpha_1 < \alpha_2$. ■

De forma inmediata se tiene también el siguiente resultado.

Proposición 2.2.3 *Dos funciones de supervivencia, $S_1(t)$ y $S_2(t)$, del modelo Gamma-poligonal están estocásticamente ordenadas.*

Demostración

Resulta inmediata dada la relación que liga a las funciones de supervivencia y azar:

$$S(t) = \exp\left[-\int_0^{+\infty} h(x)\,dx\right].$$

■

Así pues, como el parámetro β de todos los individuos proviene de la misma distribución, el individuo medio tendrá el mismo valor del parámetro β, mientras que el parámetro α dependerá de sus covariables. La ordenación de los parámetros α determinará la consiguiente ordenación en azares y supervivencias.

2.3 Análisis del modelo

La función de densidad del modelo semiparamétrico aditivo es:

$$f(t) = S(t)\, h(t) = S_0(t)\, S_T(t)\, [\vartheta(t) + \tau(t)]\,, \; t > 0, \qquad (2.3)$$

donde $S_0(t)$ es la función de supervivencia poligonal, $S_T(t)$ es la correspondiente función de supervivencia paramétrica y en el último término tenemos la suma de azares paramétrico y poligonal.

Debido precisamente a esta descomposición del azar, la función de verosimilitud presentará problemas técnicos similares a los que aparecen en el análisis de mixturas. Concretamente, aparecerán dichos problemas para los tiempos no censurados, mientras que la información correspondiente a los tiempos de supervivencia censurados se incorporará de forma sencilla a la función de verosimilitud a partir de la función de supervivencia dada en la proposición 2.2.1.

Supuesta obtenida una muestra aleatoria a partir de (2.3), en la que posiblemente existan tiempos censurados, y que realizamos un análisis bayesiano del modelo, a continuación desarrollamos las correspondientes funciones de verosimilitud, inicial y final.

2.3.1 Función de verosimilitud

Denominamos $t_1^{(j)}, \ldots, t_{n^{(j)}}^{(j)}$, a los $n^{(j)}$ tiempos de supervivencia del grupo j-ésimo, $j = 1, \ldots, M$, donde los $r^{(j)}$ primeros son no censurados y los $n^{(j)} - r^{(j)}$ restantes son tiempos censurados.

La función de verosimilitud del modelo es:

$$f\left(t_1^{(1)},\ldots,t_{n^{(1)}}^{(1)},\ldots,t_1^{(M)},\ldots,t_{n^{(M)}}^{(M)} \mid \right.$$
$$\left.\boldsymbol{\theta^{(1)}},\ldots,\boldsymbol{\theta^{(n)}},\boldsymbol{\tau^{(1)}},\ldots,\boldsymbol{\tau^{(M)}},\mu_\beta,\sigma_\beta^2,\mathbf{b},\sigma_\alpha^2,\sigma_\epsilon^{2^{(1)}},\ldots,\sigma_\epsilon^{2^{(M)}}\right)$$
$$=\prod_{j=1}^{M}\left[\prod_{i=1}^{r^{(j)}} f\left(t_i^{(j)}\right)\right]\left[\prod_{i=r^{(j)}+1}^{n^{(j)}} S\left(t_i^{(j)}\right)\right]=\prod_{j=1}^{M}\left[\prod_{i=1}^{n^{(j)}} S\left(t_i^{(j)}\right)\right]\left[\prod_{i=1}^{r^{(j)}} h\left(t_i^{(j)}\right)\right],$$

donde utilizando la proposición 2.2.1 podemos descomponer $S\left(t_i^{(j)}\right)$ en el producto de las supervivencias paramétrica y poligonal, y por (2.2), $h\left(t_i^{(j)}\right)$ es la suma de los azares paramétrico y poligonal.

Proposición 2.3.1 *Dado el tiempo de supervivencia del individuo i-ésimo del grupo l, $t_i^{(l)}$, $i=1,\ldots,n^{(l)}$, $l=1,\ldots,M$, su azar acumulado poligonal es una combinación lineal positiva de los parámetros del azar poligonal del grupo, $\tau_1^{(l)},\ldots,\tau_{g+1}^{(l)}$.*

Demostración

Integrando la función de azar $\tau(s)$ entre 0 y $t_i^{(l)}$, el azar acumulado poligonal resulta:

$$H_0\left(t_i^{(l)}\right) = \frac{1}{2}\sum_{j=1}^{g+1} 1_{I_j}\left[\left(t_i^{(l)}-a_{j-1}\right)\left(\tau(t_i^{(l)})+\tau_j^{(l)}\right)+\right.$$
$$\left.\sum_{k=1}^{j-1}(a_k-a_{k-1})\left(\tau_{k+1}^{(l)}+\tau_k^{(l)}\right)\right],$$

que teniendo en cuenta la expresión (2.1) del azar poligonal podemos

reescribir del siguiente modo:

$$
\begin{aligned}
H_0\left(t_i^{(l)}\right) &= \frac{1}{2}\sum_{j=1}^{g+1} 1_{I_j}\left[\left(t_i^{(l)}-a_{j-1}\right)\cdot\right.\\
&\cdot\left(\frac{t_i^{(l)}\left(\tau_{j+1}^{(l)}-\tau_j^{(l)}\right)+\tau_j^{(l)}\,a_j-a_{j-1}\,\tau_{j+1}^{(l)}}{a_j-a_{j-1}}+\tau_j^{(l)}\right)+\\
&\left.+\sum_{k=1}^{j-1}(a_k-a_{k-1})\,\tau_k^{(l)}+\sum_{k=2}^{j}(a_{k-1}-a_{k-2})\,\tau_k^{(l)}\right]\\
&=\frac{1}{2}\sum_{j=1}^{g+1} 1_{I_j}\left[\left(t_i^{(l)}-a_{j-1}+\frac{\left(t_i^{(l)}-a_{j-1}\right)\left(a_j-t_i^{(l)}\right)}{a_j-a_{j-1}}\right)\tau_j^{(l)}+\right.\\
&+\frac{\left(t_i^{(l)}-a_{j-1}\right)^2}{a_j-a_{j-1}}\tau_{j+1}^{(l)}+1_D\,a_1\,\tau_1^{(l)}+1_D\,(a_{j-1}-a_{j-2})\,\tau_j^{(l)}+\\
&\left.+\sum_{k=2}^{j-1}(a_{k-1}-a_{k-2})\,\tau_k^{(l)}+\sum_{k=2}^{j-1}(a_k-a_{k-1})\,\tau_k^{(l)}\right],
\end{aligned}
$$

donde $D=\bigcup_{j=2}^{g+1} I_j=\mathbb{R}^+-I_1$, y consecuentemente:

$$
H_0\left(t_i^{(l)}\right)=\frac{1}{2}\sum_{j=1}^{g+1} 1_{I_j}\sum_{k=1}^{j+1} B_k\,\tau_k^{(l)}, \tag{2.4}
$$

con

$$
\begin{aligned}
B_1 &= 1_D\,a_1\\
B_k &= a_k-a_{k-2},\ k=2,\ldots,j-1\\
B_j &= 1_D\,(a_{j-1}-a_{j-2})+\frac{\left(t_i^{(l)}-a_{j-1}\right)\left(a_j-a_{j-1}+a_j-t_i^{(l)}\right)}{a_j-a_{j-1}}\\
B_{j+1} &= \frac{\left(t_i^{(l)}-a_{j-1}\right)^2}{a_j-a_{j-1}},
\end{aligned} \tag{2.5}
$$

todas ellas cantidades positivas. ■

Proposición 2.3.2 *Dado* $l \in \{1,\ldots,M\}$,

$$\prod_{i=1}^{n^{(l)}} S_0\left(t_i^{(l)}\right) = \exp\left(-\sum_{j=1}^{g+1} c_j^{(l)}\,\tau_j^{(l)}\right),$$

donde

$$\begin{aligned}
c_1^{(l)} &= \frac{1}{2}\left[\left(n^{(l)} - n_1^{(l)}\right)a_1 - \frac{S_1^{*(l)}}{a_1}\right] + S_1^{(l)} \\
c_j^{(l)} &= \frac{1}{2}\left[\left(\sum_{i=j+1}^{g+1} n_i^{(l)}\right)(a_j - a_{j-2}) - n_j^{(l)}a_{j-2} + \right. \\
&\quad + \frac{a_j\left(2S_j^{(l)} - n_j^{(l)}a_{j-1}\right) - S_j^{*(l)}}{a_j - a_{j-1}} + \\
&\quad \left. + \frac{a_{j-2}\left(n_{j-1}^{(l)}a_{j-2} - 2S_{j-1}^{(l)}\right) + S_{j-1}^{*(l)}}{a_{j-1} - a_{j-2}}\right], \; j = 2,\ldots,g \\
c_{g+1}^{(l)} &= \frac{1}{2}\left[2S_{g+1}^{(l)} - n_{g+1}^{(l)}(a_g + a_{g-1}) + \right. \\
&\quad \left. + \frac{a_{g-1}\left(n_g^{(l)}a_{g-1} - 2S_g^{(l)}\right) + S_g^{*(l)}}{a_g - a_{g-1}}\right],
\end{aligned} \tag{2.6}$$

son estadísticos que contienen toda la información sobre $\tau^{(l)}$ *recogida en el producto de las supervivencias poligonales.*

Demostración

Sumando en (2.4) para todos los individuos del grupo, obtenemos:

$$\sum_{i=1}^{n^{(l)}} H_0\left(t_i^{(l)}\right) = \frac{1}{2}\sum_{i=1}^{n^{(l)}}\sum_{j=1}^{g+1} 1_{I_j}\sum_{k=1}^{j+1} B_k\,\tau_k^{(l)}, \tag{2.7}$$

donde los valores de B_k vienen dados por (2.5).

Dado $j = 1, \ldots, g$, denotamos por $n_j^{(l)}$ al número de individuos del grupo l-ésimo cuyo tiempo de supervivencia pertenece a I_j, por $S_j^{(l)}$ a la suma de dichos tiempos y por $S_j^{*(l)}$ a la suma de los cuadrados de los tiempos de supervivencia pertenecientes a I_j. Asímismo, suponemos, sin pérdida de generalidad, que los tiempos de supervivencia están ordenados de menor a mayor.

Desarrollando (2.7) y formulando dicha expresión como combinación lineal de $\tau_j^{(l)}$, $j = 1, \ldots, g+1$, se tiene que:

$$\begin{aligned}
\sum_{i=1}^{n^{(l)}} H_0\left(t_i^{(l)}\right) = \frac{1}{2}\Bigg\{ & \left[\left(n^{(l)} - n_1^{(l)}\right) a_1 + \sum_{i=1}^{n_1^{(l)}} \frac{t_i^{(l)}\left(2a_1 - t_i^{(l)}\right)}{a_1}\right] \tau_1^{(l)} + \\
& + \sum_{j=2}^{g} \Bigg[\left(\sum_{i=j+1}^{g+1} n_i^{(l)} \right) (a_j - a_{j-2}) + n_j^{(l)} (a_{j-1} - a_{j-2}) + \\
& + \sum_{i=1}^{n_j^{(l)}} \frac{\left(t^{(l)}_{n_{j-1}^{(l)}+i} - a_{j-1}\right)\left(a_j - a_{j-1} + a_j - t^{(l)}_{n_{j-1}^{(l)}+i}\right)}{a_j - a_{j-1}} + \\
& + \sum_{i=1}^{n_{j-1}^{(l)}} \frac{\left(t^{(l)}_{n_{j-2}^{(l)}+i} - a_{j-2}\right)^2}{a_{j-1} - a_{j-2}} \Bigg] \tau_j^{(l)} + \Big[n_{g+1}^{(l)} (a_g - a_{g-2}) + \\
& + \sum_{i=1}^{n_{g+1}^{(l)}} \frac{\left(t^{(l)}_{n_g^{(l)}+i} - a_g\right)\left(a_{g+1} - a_g + a_{g+1} - t^{(l)}_{n_g^{(l)}+i}\right)}{a_{g+1} - a_g} + \\
& + \sum_{i=1}^{n_g^{(l)}} \frac{\left(t^{(l)}_{n_{g-1}^{(l)}+i} - a_{g-1}\right)^2}{a_g - a_{g-1}} + \sum_{i=1}^{n_{g+1}^{(l)}} \frac{\left(t^{(l)}_{n_g^{(l)}+i} - a_g\right)^2}{a_{g+1} - a_g} \Bigg] \tau_{g+1}^{(l)} \Bigg\}. \qquad (2.8)
\end{aligned}$$

Teniendo en cuenta que

$$\prod_{i=1}^{n^{(l)}} S_0\left(t_i^{(l)}\right) = \exp\left(-\sum_{i=1}^{n^{(l)}} H_0\left(t_i^{(l)}\right)\right),$$

obtenemos inmediatamente a partir de (2.8) el valor de $c_1^{(l)}$ para el coeficiente de $\tau_1^{(l)}$ en la combinación lineal anterior.

El coeficiente de $\tau_j^{(l)}$, $j = 2, \ldots, g$, en dicha combinación lineal resulta:

$$\frac{1}{2}\left\{\left(\sum_{i=j+1}^{g+1} n_i^{(l)}\right)(a_j - a_{j-2}) + n_j^{(l)}(a_{j-1} - a_{j-2}) + \right.$$
$$+\frac{2a_j S_j^{(l)} - S_j^{*(l)} - 2n_j^{(l)} a_j a_{j-1} + n_j^{(l)} a_{j-1}^2}{a_j - a_{j-1}} +$$
$$\left.\frac{S_{j-1}^{*(l)} - 2a_{j-2} S_{j-1}^{(l)} + n_{j-1}^{(l)} a_{j-2}^2}{a_{j-1} - a_{j-2}}\right\},$$

que es igual al valor de $c_j^{(l)}$, $j = 2, \ldots, g$.

Finalmente, podemos escribir el coeficiente de $\tau_{g+1}^{(l)}$ como:

$$\frac{1}{2}\left\{n_{g+1}^{(l)}(a_g - a_{g-1}) + \frac{2a_{g+1}S_{g+1}^{(l)} - S_{g+1}^{*(l)} - 2n_{g+1}^{(l)} a_{g+1} a_g + n_{g+1}^{(l)}) a_g^2}{a_{g+1} - a_g} + \right.$$
$$\left. + \frac{S_g^{*(l)} - 2a_{g-1} S_g^{(l)} + n_g^{(l)} a_{g-1}^2}{a_g - a_{g-1}} + \frac{S_{g+1}^{*(l)} - 2a_g S_{g+1}^{(l)} + n_{g+1}^{(l)} a_g^2}{a_{g+1} - a_g}\right\},$$

expresión que, tras simplificar, iguala al valor de $c_{g+1}^{(l)}$. ■

Una propiedad de los coeficientes $c_j^{(l)}$, $j = 1, \ldots, g+1$, se recoge en la siguiente proposición.

Proposición 2.3.3 *Los coeficientes $c_j^{(l)}$ de la expresión (2.6) son positivos $\forall j = 1, \ldots, g+1$, $\forall l = 1, \ldots, M$.*

Demostración

Resulta evidente, pues hemos expresado en la proposición 2.3.1 el azar acumulado poligonal de un tiempo cualquiera como una combinación lineal positiva de los parámetros del azar poligonal del grupo. Por lo tanto, al sumar para todos los tiempos de supervivencia obtendremos coeficientes también positivos. ■

De aplicación directa de la proposición 2.3.2 se tiene el siguiente resultado.

Proposición 2.3.4 *La función de verosimilitud del modelo puede factorizarse como:*

$$f\left(t_1^{(1)},\ldots,t_{n^{(1)}}^{(1)},\ldots,t_1^{(M)},\ldots,t_{n^{(M)}}^{(M)} \mid \right.$$
$$\left. \alpha_1,\beta_1,\ldots,\alpha_n,\beta_n,\boldsymbol{\tau}^{(1)},\ldots,\boldsymbol{\tau}^{(M)},\mu_\beta,\sigma_\beta^2,\mathbf{b},\sigma_\alpha^2,\sigma_\epsilon^{2^{(1)}},\ldots,\sigma_\epsilon^{2^{(M)}}\right)$$
$$=\prod_{j=1}^{M}\exp\left(-\sum_{i=1}^{g+1}c_i^{(j)}\tau_i^{(j)}\right)\left[\prod_{i=1}^{n^{(j)}}S_T\left(t_i^{(j)}\right)\right]\left[\prod_{i=1}^{r^{(j)}}\left(\vartheta\left(t_i^{(j)}\right)+\tau\left(t_i^{(j)}\right)\right)\right].$$

2.3.2 Distribución inicial y distribución final

Para el análisis bayesiano del modelo Gamma-poligonal aditivo proponemos utilizar las distribuciones iniciales conjugadas habituales. En concreto, asumimos una distribución Normal para $\mu_\beta \mid \sigma_\beta^2$ y una distribución Gamma-inversa para σ_β^2,

$$\mu_\beta \mid \sigma_\beta^2 \sim N\left(\mu_\beta \mid m_b, \sigma_\beta^2 v_b^2\right),\ \sigma_\beta^2 \sim Ga\left(1/\sigma_\beta^2 \mid a_b, b_b\right).$$

También utilizamos una distribución Normal-Gamma inversa para $\mathbf{b}$ y σ_α^2,

$$\mathbf{b} \mid \sigma_\alpha^2 \sim N_k\left(\mathbf{b} \mid \mathbf{m}, \sigma_\alpha^2 A\right), \ \sigma_\alpha^2 \sim Ga\left(1/\sigma_\alpha^2 \mid a_a, b_a\right).$$

Para los parámetros correspondientes al azar poligonal, comunes a todos los individuos, proponemos un proceso autocorrelado de primer orden ya utilizado en una situación similar por Gamerman (1991).

$$\begin{aligned} \tau_{i+1} &= \tau_i \exp(\epsilon_i) \\ \epsilon_i &\sim N\left(\epsilon_i \mid 0, \sigma_\epsilon^2\right), \ i = 1, \ldots, g, \end{aligned}$$

donde los ϵ_i son independientes, $i = 1, \ldots, g$, e independientes de los $\tau's$ anteriores.

Para no tener una información inicial incompleta, especificamos una distribución inicial marginal para τ_1,

$$\tau_1 \sim Ga\left(\tau_1 \mid \alpha_\tau, \beta_\tau\right).$$

La distribución inicial considerada para el hiperparámetro σ_ϵ^2 del azar poligonal es:

$$\sigma_\epsilon^2 \sim Ga\left(1/\sigma_\epsilon^2 \mid a_\epsilon, b_\epsilon\right).$$

Todos estos grupos de parámetros los suponemos independientes a priori, por lo que la distribución inicial completa es el producto de esas distribuciones. La distribución final será proporcional al producto de la inicial anterior por la función de verosimilitud obtenida en la proposición 2.3.4. Así, el análisis de la distribución final sólo parece posible abordarlo mediante simulación, en concreto utilizando métodos MCMC. Por ello realizamos

el estudio bayesiano de este modelo Gamma-poligonal mediante una muestra aleatoria obtenida por Monte Carlo a partir de la distribución final, tal y como se desarrolla en el siguiente apartado.

2.4 Estudio de la distribución final mediante simulación

La implementación del algoritmo de Gibbs para la obtención de una muestra aleatoria a partir de la distribución final del vector paramétrico completo requiere, básicamente, la especificación de un punto inicial de la MCMC con el que comenzar el proceso iterativo y saber muestrear de las distribuciones condicionales completas. Para ello, resulta fundamental realizar una adecuada descomposición del vector paramétrico completo.

Teniendo presente la conveniencia de realizar la descomposición en el menor número de bloques, en este caso, utilizamos la siguiente: $(\mu_\beta, \sigma_\beta^2)$, $(\mathbf{b}, \sigma_a^2)$, $\left(\alpha_i^{(j)}, \beta_i^{(j)}\right)$, $i = 1, \ldots, n^{(j)}$, $j = 1, \ldots, M$, y $\left(\tau^{(j)}, \sigma_\epsilon^{2(j)}\right)$, $j = 1, \ldots, M$. Denotamos por $X_{n\times k}$ a la matriz de covariables de todos los individuos de los M grupos y, a continuación, obtenemos las distribuciones condicionales completas para esos grupos de parámetros.

Proposición 2.4.1 *La distribución condicional completa de* $(\mu_\beta, \sigma_\beta^2)$ *es una densidad Normal-Gamma inversa con parámetros*

$$\begin{aligned} \mu &= \frac{m_b + v_b^2 \sum_{j=1}^{M} \sum_{i=1}^{n^{(j)}} \log \beta_i^{(j)}}{1 + v_b^2} \\ \sigma &= \frac{\sigma_\beta^2 v_b^2}{1 + v_b^2}, \end{aligned}$$

para la distribución Normal y

$$\begin{aligned} \alpha &= a_b + \frac{n+1}{2} \\ \beta &= b_b, \end{aligned}$$

para la densidad Gamma.

Demostración

Eliminando en la distribución final todos los factores en los que no aparezca μ_β ni σ_β^2:

$$f\left(\mu_\beta, \sigma_\beta^2 \mid t_1^{(1)}, \ldots, t_{n^{(1)}}^{(1)}, \ldots, t_1^{(M)}, \ldots, t_{n^{(M)}}^{(M)}, X,\right.$$
$$\alpha_1^{(1)}, \beta_1^{(1)}, \ldots, \alpha_{n^{(1)}}^{(1)}, \beta_{n^{(1)}}^{(1)}, \ldots, \alpha_1^{(M)}, \beta_1^{(M)}, \ldots, \alpha_{n^{(M)}}^{(M)}, \beta_{n^{(M)}}^{(M)},$$
$$\left.\boldsymbol{\tau}^{(1)}, \ldots, \boldsymbol{\tau}^{(M)}, \mathbf{b}, \sigma_\alpha^2, \sigma_\epsilon^{2(1)}, \ldots, \sigma_\epsilon^{2(M)}\right)$$
$$\propto f\left(\mu_\beta \mid \sigma_\beta^2\right) f\left(\sigma_\beta^2\right) f\left(\beta_1^{(1)}, \ldots, \beta_{n^{(1)}}^{(1)}, \ldots, \beta_1^{(M)}, \ldots, \beta_{n^{(M)}}^{(M)} \mid \mu_\beta, \sigma_\beta^2\right)$$
$$\propto \frac{1}{\sigma_\beta} \exp\left[-\frac{1}{2v_b^2\sigma_\beta^2}(\mu_\beta - m_b)^2\right] \left(\frac{1}{\sigma_\beta^2}\right)^{a_b - 1} \exp\left(-\frac{b_b}{\sigma_\beta^2}\right) \cdot$$
$$\cdot \prod_{j=1}^{M} \prod_{i=1}^{n^{(j)}} \exp\left[-\frac{1}{2\sigma_\beta^2}\left(\log \beta_i^{(j)} - \mu_\beta\right)^2\right],$$

que resulta ser la situación habitual del proceso de aprendizaje bayesiano con datos normales (los logaritmos de los $\beta's$) con inicial conjugada habitual. Utilizando el resultado del teorema 1 (DeGroot, 1970, p. 169) se demuestra la proposición. ■

Proposición 2.4.2 *La distribución condicional completa de* $(\mathbf{b}, \sigma_\alpha^2)$ *es una densidad Normal-Gamma inversa con vector de valores medios y matriz de varianzas-covarianzas para la distribución Normal:*

$$\mu = \hat{\mathbf{b}}\left(A^{-1} + X'X\right)^{-1}\left(A^{-1}\mathbf{m} + X'\mathbf{y}\right)$$
$$\Sigma = \frac{1}{\sigma_\alpha^2}\left(A^{-1} + X'X\right)^{-1},$$

y parámetros para la densidad Gamma:

$$\alpha = a_a + \frac{n}{2}$$
$$\beta = b_a + \frac{1}{2}\left[\left(\mathbf{y} - X\hat{\mathbf{b}}\right)'\mathbf{y} + \left(\mathbf{m} - \hat{\mathbf{b}}\right)'A^{-1}\mathbf{m}\right],$$

donde $\mathbf{y} = \left(\log\frac{\alpha_1^{(1)}}{\beta_1^{(1)}}, \ldots, \log\frac{\alpha_{n^{(1)}}^{(1)}}{\beta_{n^{(1)}}^{(1)}}, \ldots, \log\frac{\alpha_1^{(M)}}{\beta_1^{(M)}}, \ldots, \log\frac{\alpha_{n^{(M)}}^{(M)}}{\beta_{n^{(M)}}^{(M)}}\right)'$.

Demostración

De igual manera que en la proposición anterior:

$$f\left(\mathbf{b}, \sigma_\alpha^2 \mid t_1^{(1)}, \ldots, t_{n^{(1)}}^{(1)}, \ldots, t_1^{(M)}, \ldots, t_{n^{(M)}}^{(M)}, X, \right.$$
$$\left.\alpha_1^{(1)}, \beta_1^{(1)}, \ldots, \alpha_{n^{(1)}}^{(1)}, \beta_{n^{(1)}}^{(1)}, \ldots, \alpha_1^{(M)}, \beta_1^{(M)}, \ldots, \alpha_{n^{(M)}}^{(M)}, \beta_{n^{(M)}}^{(M)}, \right.$$
$$\left.\tau^{(1)}, \ldots, \tau^{(M)}, \mu_\beta, \sigma_\beta^2, \sigma_\epsilon^{2(1)}, \ldots, \sigma_\epsilon^{2(M)}\right)$$
$$\propto f(\mathbf{b} \mid \sigma_\alpha^2) f(\sigma_\alpha^2) \prod_{j=1}^{M} \prod_{i=1}^{n^{(j)}} f\left(\alpha_i^{(j)} \mid \beta_i^{(j)}, X, \mathbf{b}, \sigma_\alpha^2\right)$$
$$\propto \exp\left[-\frac{1}{2}(\mathbf{b}-\mathbf{m})' \frac{1}{\sigma_\alpha^2} A^{-1} (\mathbf{b}-\mathbf{m})\right] \left(\frac{1}{\sigma_\alpha^2}\right)^{a_a - 1} \cdot$$
$$\cdot \exp\left(-\frac{b_a}{\sigma_\alpha^2}\right) \exp\left[-\frac{1}{2\sigma_\alpha^2} \sum_{j=1}^{M} \sum_{i=1}^{n^{(j)}} \left(\log \alpha_i^{(j)} - \mathbf{b}' \mathbf{x}_i^{(j)} - \log \beta_i^{(j)}\right)^2\right],$$

que coincide con el proceso de aprendizaje bayesiano habitual con datos (logaritmos de los $\alpha's$ y logaritmos de los $\beta's$) obtenidos según el modelo de regresión lineal múltiple normal homocedástica y las distribuciones conjugadas habituales sobre los parámetros de la regresión. Utilizando las fórmulas asociadas a ese proceso de aprendizaje (ver, por ejemplo, DeGroot, 1970, pp. 249-252) se demuestra la proposición.

■

Proposición 2.4.3 *La distribución condicional completa de* $\left(\alpha_i^{(j)}, \beta_i^{(j)}\right)$, $i = 1, \ldots, r^{(j)}$, $j = 1, \ldots, M$, *par paramétrico de la distribución Gamma*

correspondiente a un individuo con tiempo no censurado, es proporcional a:

$$S_T\left(t_i^{(j)} \mid \alpha_i^{(j)}, \beta_i^{(j)}\right)\left[\vartheta\left(t_i^{(j)} \mid \alpha_i^{(j)}, \beta_i^{(j)}\right) + \tau\left(t_i^{(j)} \mid \boldsymbol{\tau}^{(j)}\right)\right] \cdot$$
$$\cdot N\left(\log \alpha_i^{(j)} \mid \mathbf{b}'\mathbf{x}_i^{(j)} + \log \beta_i^{(j)}, \sigma_\alpha^2\right) N\left(\log \beta_i^{(j)} \mid \mu_\beta, \sigma_\beta^2\right).$$

Demostración

En la distribución final, los parámetros $\alpha_i^{(j)}$ y $\beta_i^{(j)}$ sólo aparecen en los factores asociados al individuo i-ésimo del grupo j, luego:

$$\begin{aligned}
&f\left(\alpha_i^{(j)}, \beta_i^{(j)} \mid \alpha_1^{(1)}, \beta_1^{(1)}, \ldots, \alpha_{n^{(1)}}^{(1)}, \beta_{n^{(1)}}^{(1)}, \ldots, \alpha_1^{(j)},\right. \\
&\beta_1^{(j)}, \ldots, \alpha_{i-1}^{(j)}, \beta_{i-1}^{(j)}, \alpha_{i+1}^{(j)}, \beta_{i+1}^{(j)}, \ldots, \alpha_{n^{(j)}}^{(j)}, \beta_{n^{(j)}}^{(j)}, \ldots, \alpha_1^{(M)}, \\
&\beta_1^{(M)}, \ldots, \alpha_{n^{(M)}}^{(M)}, \beta_{n^{(M)}}^{(M)}, t_1^{(1)}, \ldots, t_{n^{(1)}}^{(1)}, \ldots, t_1^{(M)}, \ldots, t_{n^{(M)}}^{(M)}, \\
&\left. X, \boldsymbol{\tau}^{(1)}, \ldots, \boldsymbol{\tau}^{(M)}, \mu_\beta, \sigma_\beta^2, \mathbf{b}, \sigma_\alpha^2, \sigma_\epsilon^{2^{(1)}}, \ldots, \sigma_\epsilon^{2^{(M)}}\right) \\
&\propto S_T\left(t_i^{(j)} \mid \alpha_i^{(j)}, \beta_i^{(j)}\right) h\left(t_i^{(j)} \mid \alpha_i^{(j)}, \beta_i^{(j)}, \boldsymbol{\tau}^{(j)}\right) \cdot \\
&\cdot f\left(\alpha_i^{(j)} \mid \beta_i^{(j)}, \mathbf{x}_i^{(j)}, \mathbf{b}, \sigma_\alpha^2\right) f\left(\beta_i^{(j)} \mid \mu_\beta, \sigma_\beta^2\right) \\
&\propto S_T\left(t_i^{(j)} \mid \alpha_i^{(j)}, \beta_i^{(j)}\right)\left[\vartheta\left(t_i^{(j)} \mid \alpha_i^{(j)}, \beta_i^{(j)}\right) + \tau\left(t_i^{(j)} \mid \boldsymbol{\tau}^{(j)}\right)\right] \cdot \\
&\cdot N\left(\log \alpha_i^{(j)} \mid \mathbf{b}'\mathbf{x}_i^{(j)} + \log \beta_i^{(j)}, \sigma_\alpha^2\right) N\left(\log \beta_i^{(j)} \mid \mu_\beta, \sigma_\beta^2\right).
\end{aligned}$$

■

Proposición 2.4.4 *La distribución condicional completa de* $\left(\alpha_i^{(j)}, \beta_i^{(j)}\right)$, $i = r^{(j)} + 1, \ldots, n^{(j)}$, $j = 1, \ldots, M$, *par de parámetros de la distribución Gamma correspondiente a un individuo con tiempo censurado, es proporcional a:*

$$S_T\left(t_i^{(j)} \mid \alpha_i^{(j)}, \beta_i^{(j)}\right) N\left(\log \alpha_i^{(j)} \mid \mathbf{b}'\mathbf{x}_i^{(j)} + \log \beta_i^{(j)}, \sigma_\alpha^2\right) N\left(\log \beta_i^{(j)} \mid \mu_\beta, \sigma_\beta^2\right).$$

Demostración

De forma similar a la proposición anterior:

$$
\begin{aligned}
& f\left(\alpha_i^{(j)}, \beta_i^{(j)} \mid \alpha_1^{(1)}, \beta_1^{(1)}, \ldots, \alpha_{n^{(1)}}^{(1)}, \beta_{n^{(1)}}^{(1)}, \ldots, \alpha_1^{(j)},\right. \\
& \beta_1^{(j)}, \ldots, \alpha_{i-1}^{(j)}, \beta_{i-1}^{(j)}, \alpha_{i+1}^{(j)}, \beta_{i+1}^{(j)}, \ldots, \alpha_{n^{(j)}}^{(j)}, \beta_{n^{(j)}}^{(j)}, \ldots, \alpha_1^{(M)}, \\
& \beta_1^{(M)}, \ldots, \alpha_{n^{(M)}}^{(M)}, \beta_{n^{(M)}}^{(M)}, t_1^{(1)}, \ldots, t_{n^{(1)}}^{(1)}, \ldots, t_1^{(M)}, \ldots, t_{n^{(M)}}^{(M)}, \\
& \left. X, \boldsymbol{\tau}^{(1)}, \ldots, \boldsymbol{\tau}^{(M)}, \mu_\beta, \sigma_\beta^2, \mathbf{b}, \sigma_\alpha^2, \sigma_\epsilon^{2^{(1)}}, \ldots, \sigma_\epsilon^{2^{(M)}}\right) \\
& \propto S_T\left(t_i^{(j)} \mid \alpha_i^{(j)}, \beta_i^{(j)}\right) f\left(\alpha_i^{(j)} \mid \beta_i^{(j)}, \mathbf{x_i^{(j)}}, \mathbf{b}, \sigma_\alpha^2\right) f\left(\beta_i^{(j)} \mid \mu_\beta, \sigma_\beta^2\right) \\
& \propto S_T\left(t_i^{(j)} \mid \alpha_i^{(j)}, \beta_i^{(j)}\right) \cdot \\
& \cdot N\left(\log \alpha_i^{(j)} \mid \mathbf{b}' \mathbf{x_i^{(j)}} + \log \beta_i^{(j)}, \sigma_\alpha^2\right) N\left(\log \beta_i^{(j)} \mid \mu_\beta, \sigma_\beta^2\right).
\end{aligned}
$$

■

Proposición 2.4.5 *La distribución condicional completa de los parámetros del azar poligonal* $\left(\boldsymbol{\tau}^{(j)}, \sigma_\epsilon^{2^{(j)}}\right)$, $j = 1, \ldots, M$, *resulta proporcional a:*

$$
\begin{aligned}
& \exp\left(-\sum_{i=1}^{g+1} c_i^{(j)} \tau_i^{(j)}\right) \left[\prod_{j=1}^{M} \prod_{i=1}^{r^{(j)}} h\left(t_i^{(j)} \mid \alpha_i^{(j)}, \beta_i^{(j)}, \boldsymbol{\tau}^{(j)}\right)\right] \cdot \\
& Ga\left(\frac{1}{\sigma_\epsilon^{2^{(j)}}} \mid \frac{g}{2} + a_\epsilon, b_\epsilon + \frac{1}{2} \sum_{i=1}^{g} \left(\log \frac{\tau_{i+1}^{(j)}}{\tau_i^{(j)}}\right)^2\right) \cdot Ga\left(\tau_1^{(j)} \mid \alpha_\tau^{(j)}, \beta_\tau^{(j)}\right).
\end{aligned}
$$

Demostración

En este caso, la distribución condicional completa resulta ser:

$$
\begin{aligned}
&f\left(\boldsymbol{\tau}^{(j)}, \sigma_\epsilon^{2(j)} \mid t_1^{(1)}, \ldots, t_{n^{(1)}}^{(1)}, \ldots, t_1^{(NG)}, \ldots, t_{n^{(NG)}}^{(NG)}, X,\right. \\
&\alpha_1^{(1)}, \beta_1^{(1)}, \ldots, \alpha_{n^{(1)}}^{(1)}, \beta_{n^{(1)}}^{(1)}, \ldots, \alpha_1^{(NG)}, \beta_1^{(NG)}, \ldots, \alpha_{n^{(NG)}}^{(NG)}, \beta_{n^{(NG)}}^{(NG)}, \\
&\boldsymbol{\tau}^{(1)}, \sigma_\epsilon^{2(1)}, \ldots, \boldsymbol{\tau}^{(j-1)}, \sigma_\epsilon^{2(j-1)}, \boldsymbol{\tau}^{(j+1)}, \sigma_\epsilon^{2(j+1)}, \ldots, \boldsymbol{\tau}^{(NG)}, \\
&\left.\sigma_\epsilon^{2(NG)}, \mu_\beta, \sigma_\beta^2, \mathbf{b}, \sigma_\alpha^2\right) \propto f\left(\tau_{g+1}^{(j)} \mid \tau_g^{(j)}, \sigma_\epsilon^{2(j)}\right) \cdot \\
&\cdot f\left(\tau_g^{(j)} \mid \tau_{g-1}^{(j)}, \sigma_\epsilon^{2(j)}\right) \cdots f\left(\tau_2^{(j)} \mid \tau_1^{(j)}, \sigma_\epsilon^{2(j)}\right) f\left(\tau_1^{(j)}\right) \cdot \\
&\cdot f\left(\sigma_\epsilon^{2(j)}\right) \prod_{j=1}^{NG}\left[\prod_{i=1}^{n^{(j)}} S_0\left(t_i^{(j)} \mid \boldsymbol{\tau}^{(j)}\right)\right]\left[\prod_{i=1}^{r^{(j)}} h\left(t_i^{(j)} \mid \alpha_i^{(j)}, \beta_i^{(j)}, \boldsymbol{\tau}^{(j)}\right)\right] \\
&\propto \exp\left(-\sum_{i=1}^{g+1} c_i^{(j)} \tau_i^{(j)}\right)\left[\prod_{j=1}^{NG}\prod_{i=1}^{r^{(j)}} h\left(t_i^{(j)} \mid \alpha_i^{(j)}, \beta_i^{(j)}, \boldsymbol{\tau}^{(j)}\right)\right] \cdot \\
&\cdot\left(\sigma_\epsilon^{2(j)}\right)^{-\frac{1}{2}g} \exp\left[-\frac{1}{2\sigma_\epsilon^{2(j)}} \sum_{i=1}^{g}\left(\log \frac{\tau_{i+1}^{(j)}}{\tau_i^{(j)}}\right)^2\right] \cdot \\
&\cdot Ga\left(\tau_1^{(j)} \mid \alpha_\tau^{(j)}, \beta_\tau^{(j)}\right) Ga\left(\sigma_\epsilon^{2(j)} \mid a_\epsilon^{(j)}, b_\epsilon^{(j)}\right) \\
&\propto \exp\left(-\sum_{i=1}^{g+1} c_i^{(j)} \tau_i^{(j)}\right)\left[\prod_{j=1}^{NG}\prod_{i=1}^{r^{(j)}} h\left(t_i^{(j)} \mid \alpha_i^{(j)}, \beta_i^{(j)}, \boldsymbol{\tau}^{(j)}\right)\right] \cdot \\
&\cdot Ga\left(\frac{1}{\sigma_\epsilon^{2(j)}} \mid \frac{g}{2} + a_\epsilon, b_\epsilon + \frac{1}{2}\sum_{i=1}^{g}\left(\log \frac{\tau_{i+1}^{(j)}}{\tau_i^{(j)}}\right)^2\right) Ga\left(\tau_1^{(j)} \mid \alpha_\tau^{(j)}, \beta_\tau^{(j)}\right) .
\end{aligned}
$$

■

El algoritmo de Gibbs puede ser implementado partiendo de un punto inicial para los hiperparámetros $\mu_\beta, \sigma_\beta^2, \sigma_\alpha^2$, los coeficientes $\mathbf{b}$ y los parámetros del modelo (los $\alpha's$ y $\beta's$ de la distribución Gamma y los parámetros $\boldsymbol{\tau}$'s de los azares poligonales). A partir de ese punto se pueden simular los pares $\left(\alpha_i^{(j)}, \beta_i^{(j)}\right)$, $j = 1, \ldots, M$, $i = 1, \ldots, n^{(j)}$, utilizando las proposiciones

2.4.3 y 2.4.4. Posteriormente, simular $\left(\boldsymbol{\tau}^{(j)}, \sigma_{\epsilon}^{2(j)}\right)$, $j = 1, \ldots, M$, utilizando la proposición 2.4.5. A continuación, simular $\left(\mu_{\beta}, \sigma_{\beta}^{2}\right)$ según la proposición 2.4.1 y, por último, simular $(\mathbf{b}, \sigma_{\alpha}^{2})$ utilizando la proposición 2.4.2.

Repitiendo indefinidamente este proceso se puede obtener una realización de una cadena de Markov con la distribución final como distribución estacionaria.

Una vez observados un número suficientemente grande de pasos en esa cadena, podemos aproximar, por Monte Carlo, la distribución predictiva de un nuevo individuo con vector de covariables $\mathbf{x}$. Así, teniendo en cuenta la forma (2.3) de la densidad del modelo:

$$f(t \mid \mathbf{x}) \simeq \frac{1}{m} \sum_{j=1}^{m} S_0(t \mid \boldsymbol{\tau}_{(i)})\, S_T(t \mid \alpha_{(i)}, \beta_{(i)})\, \left[\vartheta(t \mid \alpha_{(i)}, \beta_{(i)}) + \tau(t \mid \boldsymbol{\tau}_{(i)})\right],$$

donde $\left\{\left(\alpha_{(i)}, \beta_{(i)}, \boldsymbol{\tau}_{(i)}\right),\ i = 1, \ldots, m\right\}$ es una muestra de los parámetros del modelo obtenida a partir de la muestra de los hiperparámetros $\left(\mu_{\beta}, \sigma_{\beta}^{2}, \mathbf{b}, \sigma_{\alpha}^{2}\right)$ generada por Gibbs.

De forma similar se puede obtener una aproximación a las funciones de supervivencia, azar y azar acumulado, así como a los momentos de la distribución predictiva.

Algunas de las distribuciones condicionales completas que se han de utilizar en este proceso son muy conocidas y resulta sencillo simular de ellas. De las otras condicionales se podría simular mediante el algoritmo de aceptación-rechazo, después de encontrar una aproximación aceptable, o mediante un algoritmo Metropolis-Hastings, lo que se conoce como Metropolis dentro de Gibbs, que es el procedimiento que proponemos y utilizamos en el siguiente capítulo.

Capítulo 3

Aplicación al análisis de datos de supervivencia

3.1 Introducción

En este capítulo se muestra cómo implementar los resultados teóricos obtenidos en el capítulo anterior, ejemplificándolo con el análisis de algunos bancos de datos simulados y reales. Con los datos simulados podemos comprobar la adecuación del modelo a distintas situaciones de supervivencia, mientras que el análisis de datos reales permite la comparación de resultados con otros estudios. Aquí presentamos el análisis de los datos de Stanford y de unos datos provenientes de una encuesta a licenciados en Matemáticas por la Universitat de València.

El aparato informático utilizado en este capítulo incluye unas aplicaciones en Fortran 77 para la implementación del método, el programa CODA (Best et al., 1995) para el análisis de la convergencia de la MCMC y el programa S-PLUS (Becker et al., 1988) para las gráficas que incorpora el presente capítulo.

Los tiempos de cómputo a los que se hace referencia en cada uno de los bancos de datos corresponden a tiempos de CPU en un PC Pentium II 266 Mhz.

3.2 Implementación del modelo Gamma-poligonal aditivo

A continuación se detalla la metodología genérica utilizada en el estudio de los bancos de datos y se comenta el soporte informático necesario para llevar a cabo dicho análisis.

Las distribuciones iniciales utilizadas fueron, en todos los casos, poco informativas, pero propias, con el fin de estudiar principalmente la información proporcionada por los datos. De este modo, las iniciales Gamma-inversas sobre las varianzas tienen parámetros $\alpha = \beta = 1$ (así, obtenemos varianzas iniciales para $\log\alpha, \log\beta$ y ϵ_i, $i = 1, \ldots, g$, de media 1, razonablemente grande), la distribución inicial log-Normal sobre μ_β, la media de $\log\beta$, es de media 0 y varianza 1, la Normal inicial sobre $\mathbf{b}$ tiene media $\mathbf{0}$ y matriz de varianzas-covarianzas I_k, la matriz identidad y, finalmente, la distribución inicial Gamma sobre $\tau_1^{(j)}$ tiene parámetros $\alpha = \beta = 2$ (media 1 y varianza 0.5 para la función de azar poligonal en cero), para todos los grupos $j = 1, \ldots, M$.

Partiendo de un punto inicial, implementamos el algoritmo de Gibbs simulando, en primer lugar, los parámetros α y β de la distribución Gamma de cada individuo, distinguiendo si son los correspondientes a un tiempo censurado o no. Esto es, simulamos $\left(\alpha_i^{(j)}, \beta_i^{(j)}\right)$, $j = 1, \ldots, M$, utilizando la proposición 2.4.3 para $i = 1, \ldots, r^{(j)}$ y con la proposición 2.4.4 para $i = r^{(j)} + 1, \ldots, n^{(j)}$. En ambos casos utilizamos el algoritmo de Metropolis con función importante:

$$N\left(\log\alpha_i^{(j)} \,|\, \mathbf{b}'\mathbf{x}_\mathbf{i}^{(\mathbf{j})} + \log\beta_i^{(j)}, \sigma_\alpha^2\right) N\left(\log\beta_i^{(j)} \,|\, \mu_\beta, \sigma_\beta^2\right).$$

Esta función importante resulta sencilla de muestrear y permite la simplificación de términos en el cociente del algoritmo de Metropolis.

A continuación, simulamos conjuntamente los parámetros de la función de azar poligonal, $\left(\boldsymbol{\tau}^{(j)}, \sigma_\epsilon^{2(j)}\right)$, $j = 1, \ldots, M$, utilizando la proposición 2.4.5 y mediante el método de Metropolis con el siguiente proceso de simulación:

- primeramente, generamos los hiperparámetros $\epsilon_i^{(j)}$, $i = 1, \ldots, g$, mediante un proceso de aprendizaje utilizando una distribución inicial $N\left(0, \sigma_\epsilon^{2(j)}\right)$ y el dato observado en la etapa anterior modelizado como $z_i^{(j)} \sim N\left(z_i^{(j)} | \log \frac{\tau_{i+1}^{(j)}}{\tau_i^{(j)}}, C\right)$. Así, $z_i^{(j)}$ está centrado en el valor del hiperparámetro $\epsilon_i^{(j)}$ en la etapa anterior con una varianza C que actúa como parámetro de sintonización para Metropolis. De este modo, simulamos $\epsilon_i^{(j)}$ a partir de:

$$N\left(\epsilon_i^{(j)} | \mu_i, \sigma^2\right),$$

con $\mu_i = \frac{z_i^{(j)}}{C}\sigma^2$ y $\sigma^2 = \left(\frac{1}{C} + \frac{1}{\sigma_\epsilon^{2(j)}}\right)^{-1}$, $i = 1, \ldots, g$.

- generamos $\tau_1^{(j)}$ de una distribución log-Normal centrada en el valor anterior de $\tau_1^{(j)}$ en la MCMC.
- calculamos los parámetros $\tau_{i+1}^{(j)} = \tau_i^{(j)} \exp\left(\epsilon_i^{(j)}\right)$, $i = 1, \ldots, g$.
- generamos $\sigma_\epsilon^{2(j)}$ a partir de:

$$Ga\left(\frac{1}{\sigma_\epsilon^{2(j)}} | \frac{g}{2} + a_\epsilon, b_\epsilon + \frac{1}{2}\sum_{i=1}^{g}\left(\log \frac{\tau_{i+1}^{(j)}}{\tau_i^{(j)}}\right)^2\right).$$

Posteriormente, simulamos los hiperparámetros a partir de distribuciones Normal-Gamma inversas utilizando las proposiciones 2.4.1 y 2.4.2.

Construimos el punto inicial del algoritmo de Gibbs tomando para los hiperparámetros valores iniciales $\mathbf{b} = \mathbf{0}$, $\mu_\beta = 0$ y valores unidad para

las varianzas. Para los parámetros del modelo obtenemos el punto inicial automáticamente a partir de los valores dados a los hiperparámetros del siguiente modo:

$$\begin{aligned} \alpha_i^{(j)} &= \exp\left(\mathbf{b}'\mathbf{x}_\mathbf{i}^{(\mathbf{j})} + \mu_\beta\right) = 1 \\ \beta_i^{(j)} &= \exp(\mu_\beta) = 1 \\ \tau^{(j)} &= d^{(j)}\,\mathbf{1},\ j = 1, \ldots, M, \end{aligned}$$

donde $d^{(j)} = \frac{n^{(j)}}{\sum_{i=1}^{n^{(j)}} t_i^{(j)}}$, $j = 1, \ldots, M$, es la estimación máximo verosímil de una función de azar constante.

Este criterio de elección del punto inicial resulta sencillo y ha proporcionado buenos resultados en todos los bancos de datos analizados.

De este modo, no resulta excesivamente complicada la implementación del método de Metropolis dentro de Gibbs para la obtención de muestras de la distribución final, tal y como se explica en el apartado anterior, si bien cabe hacer algunas consideraciones puntuales acerca de la implementación del método de análisis de nuestro modelo.

En las etapas donde la simulación se realiza a partir de distribuciones analíticas Normal-Gamma inversas se utiliza el algoritmo polar (Box y Muller, 1958) para simular de la distribución Normal y para simular de la distribución Gamma, el algoritmo de Best en combinación con el teorema de Stuart (Devroye, 1986, pp. 182 y 410).

En las etapas donde la simulación se realiza por Metropolis, para la comparación de las funciones de densidad en los nuevos parámetros generados y los anteriores en la MCMC, se requiere la evaluación de una parte de la función de verosimilitud. En concreto, no resulta trivial la evaluación de la función de supervivencia Gamma dado que se precisa la integral Gamma

incompleta. Ésta puede obtenerse utilizando el algoritmo AS32 (Bhattacharjee, 1970), basado en un desarrollo en serie de Pearson y en fracciones continuas. No obstante, para valores grandes del parámetro α puede dar errores numéricos en casos muy atípicos. Para evitar esto, lo que hacemos es modificar dicho algoritmo para calcular:

$$\log \int_0^{+\infty} Ga(\alpha, \beta),$$

pues es lo único que se necesita para la supervivencia Gamma. De este modo hemos conseguido evitar todos los errores numéricos comentados con anterioridad.

Por otra parte, también se precisa la evaluación de la función Gamma, tanto para el cálculo de la supervivencia Gamma como para el de la función de azar. Utilizamos el algoritmo ACM291 (Pike y Hill, 1966) para calcular el logaritmo de la función Gamma y obtenemos el logaritmo de la función de azar como diferencia del logaritmo de la función de densidad Gamma y el logaritmo de la supervivencia. De esta forma evitamos también errores numéricos.

Como apoyo informático a la implementación del método de análisis, en primer lugar se confeccionó un programa llamado *lectura*, en Fortran 77, para la lectura de datos. Bastante genérico, tan sólo requiere leer los mismos de un fichero en un determinado formato. Asímismo, el programa calcula todos los estadísticos que son comunes a las distintas etapas del análisis, entre los que cabe destacar los c_j, $j = 1, \ldots, g+1$, dados en la proposición 2.3.2 y escribe toda esa información en un nuevo fichero para su posterior utilización.

Para la obtención de la MCMC utilizando el algoritmo de Gibbs se desarrolló, también en Fortran 77, otro programa llamado *mcmc* que ofrece

la posibilidad al usuario de obtener una primera aproximación a la función de azar poligonal de cada uno de los grupos o bien generar la cadena de Markov para su análisis con CODA y para la obtención de predicciones. En la primera opción, se trata de dividir el eje temporal en un número suficientemente grande de puntos para aproximar el azar poligonal con la media de los correspondientes parámetros $\tau's$ del grupo generados en la MCMC. Así, la función de azar poligonal tiene un gran número de segmentos de definición y puede, por tanto, recoger perfectamente la forma del azar base. Una vez observada la forma del azar base, puede buscarse una poligonal más sencilla que también la aproxime suficientemente y que tenga menos parámetros. Para ello sólo hay que elegir adecuadamente los puntos no triviales del eje. La segunda opción del programa genera una cadena de Markov con los parámetros finalmente considerados, así como dos ficheros, *coda.ind* y *coda.out*, para el análisis de la convergencia con CODA. Los parámetros de la MCMC se escriben en un fichero con un formato adecuado para su utilización con el programa de predicción, mientras que los ficheros *coda.ind* y *coda.out* contienen la misma información en el formato *ad hoc* para la aplicación de CODA.

El tercer y último de los programas desarrollados, *predic*, calcula funciones y medidas predictivas para nuevos individuos. Proporciona las coordenadas de los puntos para el dibujo de las funciones de azar, densidad y supervivencia predictivas para valores concretos de las covariables, con y sin bandas de confianza para las mismas, y calcula la media y varianza de la distribución predictiva. Las bandas de confianza se han obtenido como intervalos de confianza puntuales para cada uno de los puntos de dibujo. Esto es, no son bandas de confianza sobre la distribución final, sino que como las funciones predictivas se calculan por Monte Carlo en cada uno de los puntos, podemos obtener, aplicando el teorema central del límite, el error Monte Carlo cometido y el correspondiente intervalo de confianza puntual

del 95% para el valor de la función predictiva en el punto de dibujo.

La metodología de análisis de un banco de datos cualquiera ha sido, genéricamente, la siguiente. Previa lectura y adecuación de los datos con el programa *lectura* y dividido el eje temporal en bastantes (de 10 a 20) intervalos con suficiente información (número de datos) en cada uno de ellos, se obtiene una aproximación a la función de azar poligonal con la opción correspondiente del programa *mcmc*. Se elige una partición definitiva del eje temporal teniendo en cuenta, básicamente, los cambios de monotonicidad de la aproximación anterior y de modo que no haya diferencias considerables en el número de tiempos de supervivencia en cada uno de los intervalos. Como tan sólo se trata de recoger la forma genérica del azar poligonal, unos pocos puntos son suficientes para realizar ya el análisis definitivo. Con esta partición, procesamos los datos y la nueva información con *lectura* y obtenemos con *mcmc* una cadena de Markov cuya convergencia analizamos con el programa CODA. Una vez alcanzada la convergencia y obtenida una muestra de la distribución final, utilizamos *predic* para la obtención de los puntos de dibujo de las funciones predictivas deseadas.

3.3 Datos simulados

3.3.1 Banco de datos 1

Este primer banco de datos consta de un único grupo de 100 individuos con dos covariables. Consideramos otra covariable adicional x_1, constante e igual a 1, para que aparezca el término independiente en la combinación lineal, mientras que x_2 y x_3 miden características propias de cada individuo y fueron simuladas a partir de distribuciones $N(0,1)$ independientes.

Los parámetros de este modelo correspondientes a la función de azar poligonal son:

$$\boldsymbol{\tau}^{(1)} = (0.1, 0.2, 0.2, 0.3, 0.1)',$$

con $\mathbf{a} = (1, 3, 5, 8)'$ como divisiones no triviales del eje temporal y los relativos a las covariables:

$$\mathbf{b} = (7.5, -0.5, 1)',\ \sigma_\alpha^2 = 1,$$
$$\mu_\beta = 0,\ \sigma_\beta^2 = 1.$$

Generamos 100 tiempos de supervivencia a partir de este modelo utilizando un mecanismo de censura progresiva por la derecha. Consideramos el tiempo de entrada en el estudio proveniente de una distribución uniforme entre 0 y un tiempo final fijo y establecimos la censura cuando el tiempo generado más el tiempo de entrada en el estudio fuera mayor que el tiempo final fijo. De este modo, obtuvimos para este banco de datos un 16% de tiempos censurados.

Los parámetros $\mathbf{b}$ y μ_β, en conjunción con las covariables, dan lugar a elevados valores para la media de $\log\alpha$. Además, como β está alrededor

de uno, esto quiere decir que la función de azar Gamma de este modelo es muy pequeña en comparación con el azar poligonal y que, por tanto, las covariables apenas influyen. Por ello, realizamos un primer análisis del modelo sin incorporar éstas y teniendo en cuenta, por tanto, sólo el azar poligonal.

Tras observar una rápida convergencia del vector paramétrico $\tau^{(1)}$, desechamos 100000 pasos iniciales y efectuando saltos de 10 pasos obtuvimos una MCMC de tamaño 1000 con la que predecir supervivencias. Es de resaltar la buena convergencia de los cinco parámetros del azar poligonal y lo bien que se mezcla la MCMC en este modelo. En la figura 3.1 se representa la función de supervivencia del modelo correcto, calculada en $x_2 = 0$ y $x_3 = 0$ (cualquier otro par de valores razonables para las covariables proporcionan una curva de supervivencia indistinguible de la aquí representada) y la supervivencia predicha por nuestro análisis. Las bandas de confianza presentan el intervalo de confianza puntual al 95%. El error Monte Carlo cometido en la predicción es tan pequeño que dichas bandas de confianza se superponen con la estimación de la función de supervivencia.

En la tabla 3.1 se proporciona la media y desviación típica de las funciones anteriores.

	modelo	predictiva
media	4.21	5.04
desviación	4.08	5.35

Tabla 3.1: *Media y desviación típica del modelo correcto y de la densidad predictiva sin utilizar covariables. Banco de datos simulados 1.*

Como en ellas puede observarse, obtenemos una predicción para la función de supervivencia muy parecida a la verdadera, con una media ligeramente superior y una mayor varianza, como suele ocurrir en toda distri-

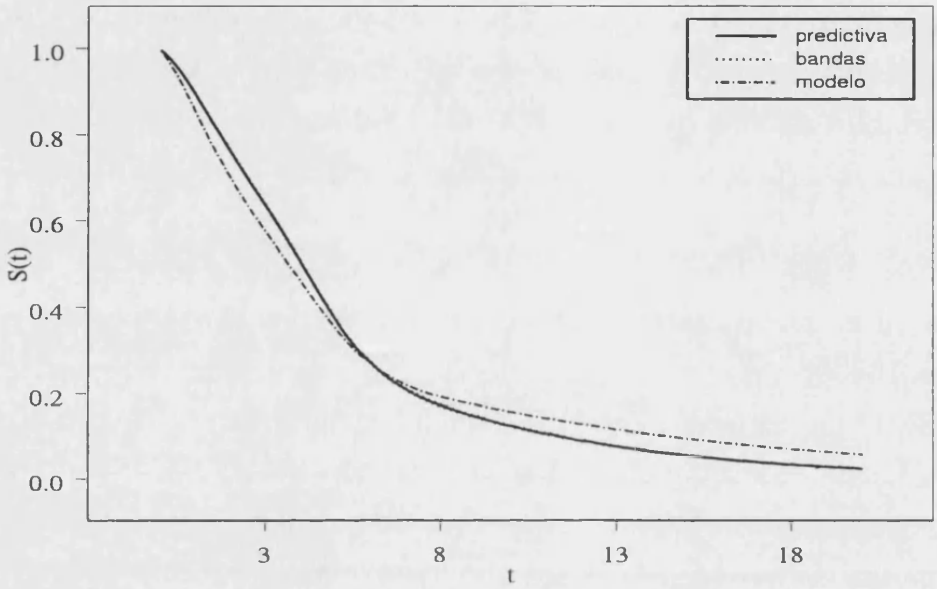

Figura 3.1: *Función de supervivencia del modelo correcto y su predicción sin utilizar covariables. Banco de datos simulados 1.*

bución predictiva. También se observa en la figura 3.1 que las dos curvas de supervivencia intersectan en un punto cercano a la media, siendo hasta entonces superior la supervivencia predictiva. Esto quiere decir que las dos colas de la función de densidad predictiva son más pesadas que las del modelo.

Realizamos a continuación el análisis del modelo con las covariables, partiendo del punto inicial $\mathbf{b} = \mathbf{0}$, $\sigma_\alpha^2 = 1, \mu_\beta = 0$, $\sigma_\beta^2 = 1$ y $\boldsymbol{\tau}^{(1)} = d^{(1)}\,\mathbf{1}$, donde $d^{(1)} = 0.19$ es la estimación máximo verosímil de una función de azar constante. Utilizando la misma partición temporal de la simulación de los datos, generamos 1000000 pasos en la MCMC (17 minutos). En la figura 3.2 puede observarse la evolución de la cadena, cada 100 pasos, para distintos parámetros de la misma.

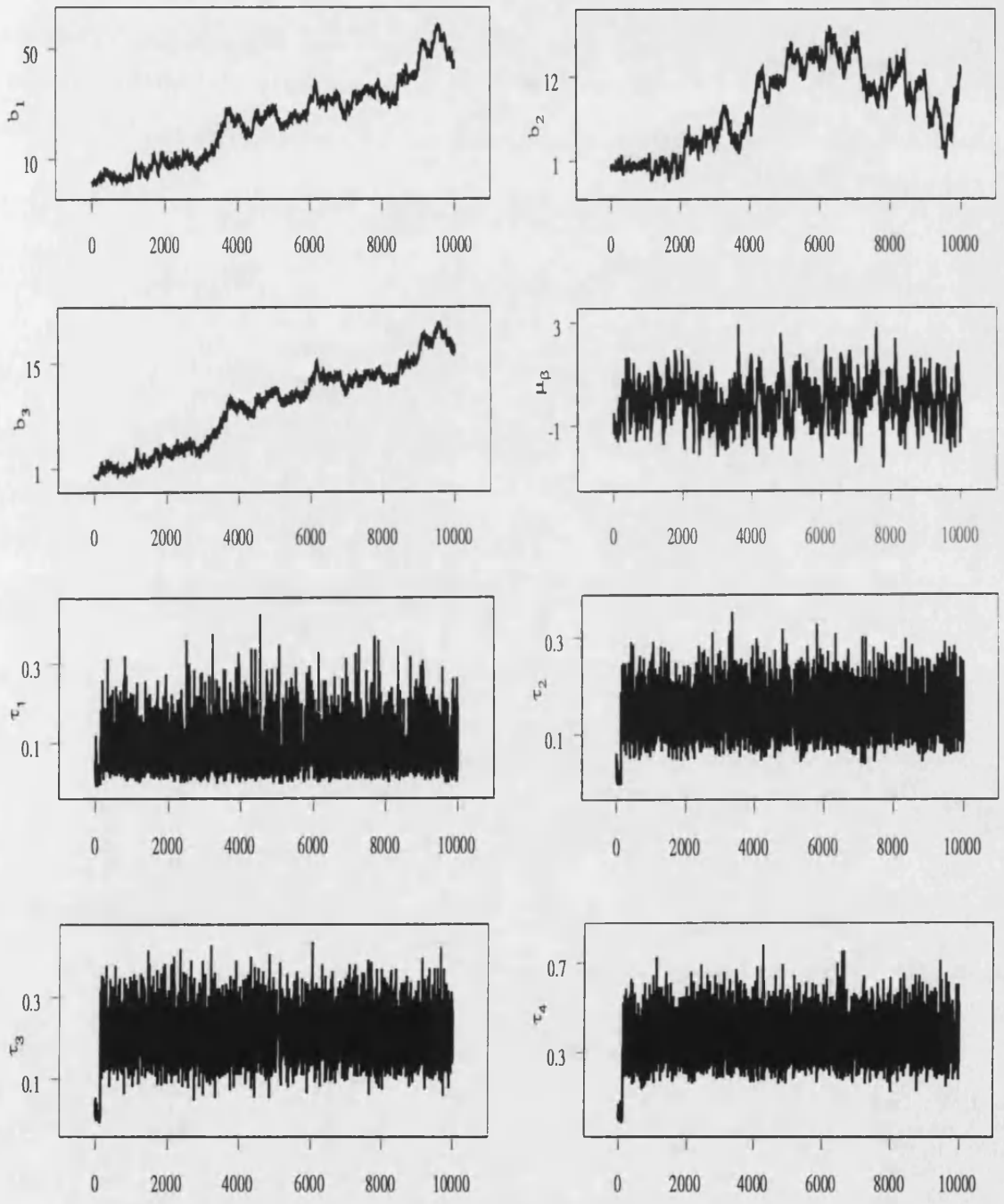

Figura 3.2: *Evolución de algunos parámetros de la MCMC con todas las covariables en el estudio. Banco de datos simulados 1.*

El análisis de la convergencia de la MCMC con el programa CODA se realizó sobre la segunda mitad de la cadena. De este modo, habiendo desechado 500000 pasos iniciales y registrando 1 de cada 100 hasta obtener una muestra de tamaño 5000, obtuvimos los resultados que se recogen en las figuras 3.3 y 3.4.

SUMMARY STATISTICS

Quantiles for each variable:

VARIABLE	2.5%	50%	97.5%
b1	22.50000	33.00000	56.50000
b2	3.47000	12.20000	17.20000
b3	9.89000	13.20000	19.30000
mub	-1.65000	-0.00410	1.54000
tao1	0.00393	0.06610	0.20700
tao2	0.07870	0.14200	0.23300
tao3	0.11700	0.20500	0.31900
tao4	0.22300	0.36300	0.53800
tao5	0.08120	0.15500	0.26500

Figura 3.3: *Intervalos intercuantílicos, obtenidos con CODA, de algunos parámetros de la MCMC considerando todas las covariables en el estudio. Banco de datos simulados 1.*

Los valores tan grandes obtenidos para las estimaciones de las componentes del vector paramétrico $\mathbf{b}$ dan lugar a valores muy elevados para la media de $\log \alpha$, indicando que la parte paramétrica de la función de azar apenas influye en el azar del modelo. Esto es, la suma de los dos azares, Gamma y poligonal, es prácticamente el poligonal, como era presumible. Además, ello puede conllevar una gran variabilidad en estos parámetros (dentro de un rango de valores grandes), con una autocorrelación muy al-

GEWEKE CONVERGENCE DIAGNOSTIC (Z-score):

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

VARIABLE	MCMC
b1	-28.200
b2	20.600
b3	-29.100
mub	0.309
tao1	-1.170
tao2	1.570
tao3	2.290
tao4	2.110
tao5	-2.420

RAFTERY AND LEWIS CONVERGENCE DIAGNOSTIC:

Quantile = 0.025
Accuracy = +/- 0.005
Probability = 0.95

VARIABLE	Thin (k)	Burn-in (M)	Total (N)	Lower bound (Nmin)	Dependence factor (I)
b1	3	144	142482	3746	38
b2	1	165	179043	3746	47.8
b3	3	42	41955	3746	11.2
mub	2	40	42556	3746	11.4
tao1	1	117	126817	3746	33.9
tao2	1	4	5038	3746	1.34
tao3	1	5	5771	3746	1.54
tao4	1	4	4713	3746	1.26
tao5	1	5	6078	3746	1.62

Figura 3.4: *Análisis de la convergencia, obtenida con CODA, de algunos parámetros de la MCMC considerando todas las covariables en el estudio. Banco de datos simulados 1.*

ta y provocar, por la falta de información que aportan los datos sobre los mismos, la no convergencia de los parámetros (ver figura 3.4). Un comportamiento similar hemos observado en otros estudios simulados en los que incluíamos covariables que no habían sido usadas en la simulación. En esos casos hemos observado dos tipos de comportamiento. Bien la distribución final sobre el coeficiente de la covariable se centraba en cero, con poca varianza, o bien tomaba valores muy elevados con mucha varianza, como en este estudio.

Posteriormente, realizamos la selección de variables no considerando x_2 y obtuvimos una MCMC de tamaño 10000 desechando 600000 pasos iniciales y efectuando saltos de 50 pasos. La figura 3.5 recoge la evolución de dicha cadena.

Algunos diagnósticos de la convergencia para los parámetros anteriores y los intervalos intercuantílicos proporcionados por CODA son mostrados en las figuras 3.6 y 3.7.

Tanto las gráficas de la evolución como los análisis de la convergencia realizados indican que se ha alcanzado la convergencia para la mayoría de los parámetros. Las gráficas de b_1 y b_3 de la figura 3.5, aunque centradas en 3 y 1 respectivamente, se muestran muy inestables a causa de la poca información que los datos aportan sobre estos parámetros. Los valores un poco más atípicos para la convergencia de b_1, b_3 y μ_β que se observan en la figura 3.6 pueden ser debidos a la gran autocorrelación que presentan. Cabe destacar que los intervalos de confianza del 95% recogen, todos ellos, a los verdaderos parámetros del azar poligonal. Siendo la componente poligonal la principal aportación a la función de azar del modelo, ello indica que las predicciones realizadas serán bastante buenas.

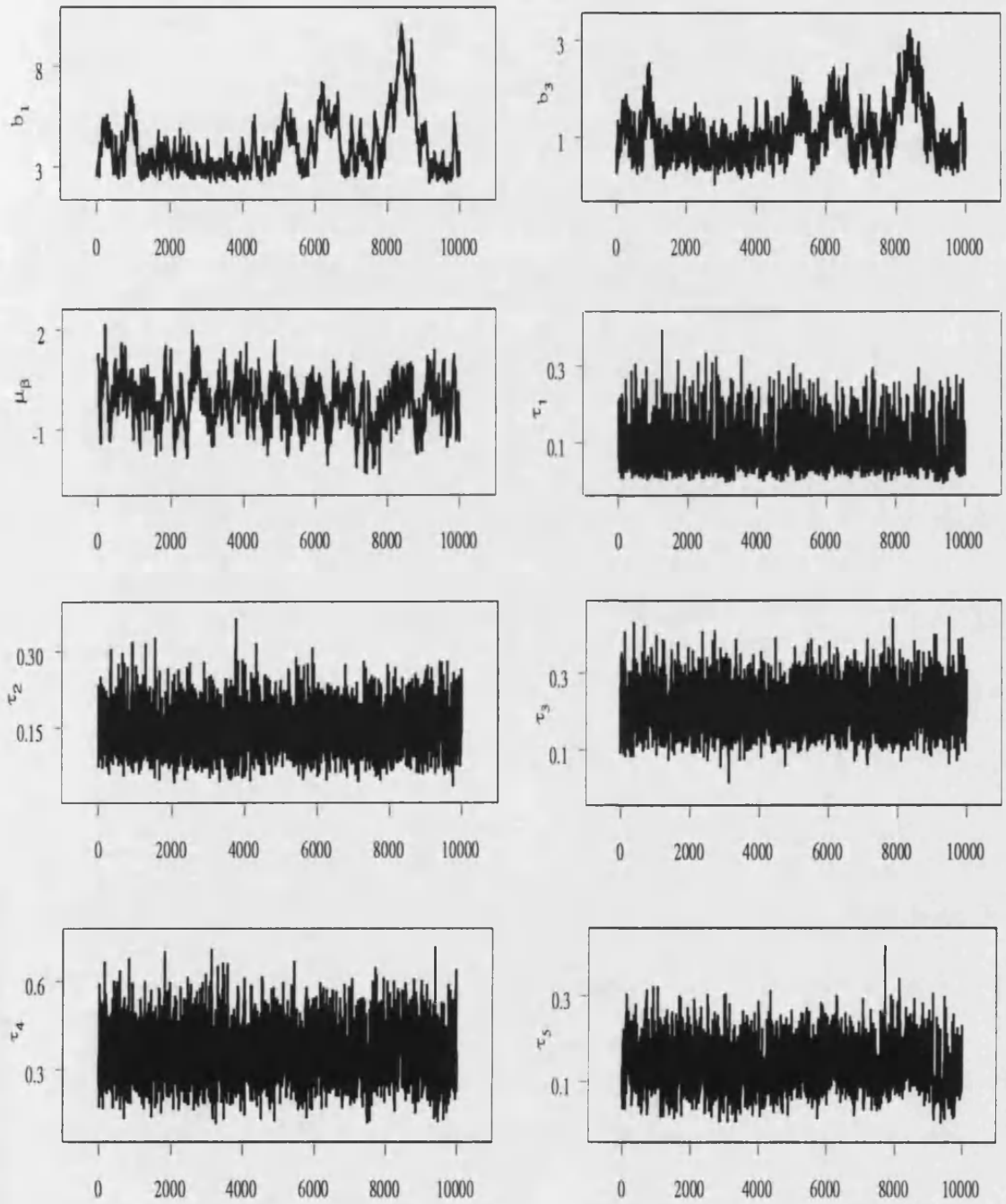

Figura 3.5: *Evolución de algunos parámetros de la MCMC obtenida con CODA, considerando únicamente x_1 y x_3 en el estudio. Banco de datos simulados 1.*

GEWEKE CONVERGENCE DIAGNOSTIC (Z-score):

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

VARIABLE	MCMC
b1	-2.440
b3	-2.140
mub	7.920
tao11	0.545
tao21	-0.247
tao31	0.362
tao41	0.971
tao51	-0.676

RAFTERY AND LEWIS CONVERGENCE DIAGNOSTIC:

Quantile = 0.025
Accuracy = +/- 0.005
Probability = 0.95

VARIABLE	Thin (k)	Burn-in (M)	Total (N)	Lower bound (Nmin)	Dependence factor (I)
b1	4	20	24244	3746	6.47
b3	2	10	11512	3746	3.07
mub	3	57	62919	3746	16.8
tao11	1	104	115797	3746	30.9
tao21	1	7	7967	3746	2.13
tao31	1	8	8356	3746	2.23
tao41	1	6	7130	3746	1.9
tao51	1	17	17908	3746	4.78

Figura 3.6: *Análisis de la convergencia de algunos parámetros de la MCMC obtenido con CODA, considerando únicamente x_1 y x_3 en el estudio. Banco de datos simulados 1.*

SUMMARY STATISTICS

Quantiles for each variable:

VARIABLE	2.5%	50%	97.5%
b1	2.50000	3.55000	7.74000
b3	0.41200	0.97200	2.42000
mub	-1.52000	-0.14700	1.15000
tao11	0.00442	0.06650	0.20700
tao21	0.07730	0.14200	0.23300
tao31	0.11500	0.20200	0.31800
tao41	0.20700	0.34700	0.53700
tao51	0.04140	0.13200	0.23600

Figura 3.7: *Intervalos intercuantílicos, obtenidos con CODA, de algunos parámetros de la MCMC considerando únicamente x_1 y x_3 en el estudio. Banco de datos simulados 1.*

En la tabla 3.2 aparece reflejada la media y desviación típica de la densidad del modelo y de la predictiva para algunos individuos, utilizando x_1 y x_3 como únicas covariables y considerando los 1000 últimos pasos de la MCMC anterior.

	modelo	x_3=-2	x_3=0	x_3=2
media	4.21	2.67	4.69	5.03
desviación	4.08	2.1	4.5	5.85

Tabla 3.2: *Media y desviación típica del modelo correcto y de la densidad predictiva para algunos individuos. Banco de datos simulados 1.*

En la figura 3.8 se representa la función de supervivencia predictiva de los individuos anteriores.

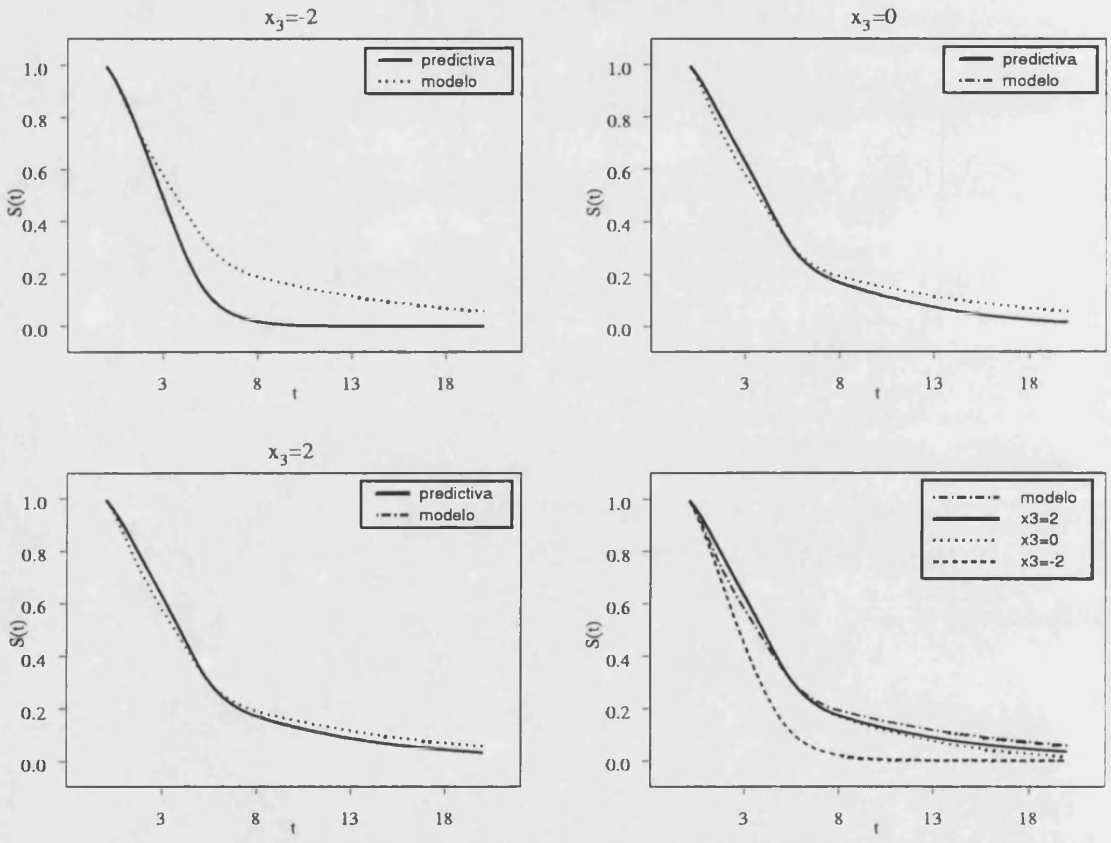

Figura 3.8: *Función de supervivencia del modelo correcto y su predicción para algunos individuos. Banco de datos simulados 1.*

Observamos en la figura 3.8 y en la tabla 3.2 que tan sólo valores pequeños de la covariable x_3 influyen en la supervivencia. Las funciones de supervivencia obtenidas para valores de dicha covariable por encima de la media son prácticamente indistinguibles y casi coincidentes con la supervivencia del modelo.

3.3.2 Banco de datos 2

Este banco de datos consta de dos grupos de 500 individuos cada uno. Las características particulares de cada individuo se recogen en las covariables x_2 y x_3, que fueron simuladas a partir de distribuciones $N(0,1)$ independientes, y al igual que en el banco de datos anterior, añadimos una covariable x_1, constante e igual a 1, para contemplar el término independiente de la combinación lineal.

Los parámetros relativos al azar poligonal de cada uno de los grupos son:

$$\tau^{(1)} = (0.01, 0.1, 0.06, 0.01, 0.05)'$$
$$\tau^{(2)} = (0.1, 0.05, 0.01, 0.05, 0.1)',$$

con el vector $\mathbf{a} = (3, 6, 10, 20)'$ como divisiones no triviales del eje temporal y los correspondientes a las covariables:

$$\mathbf{b} = (3.5, 1, 3)',\ \sigma_\alpha^2 = 1,$$
$$\mu_\beta = 3,\ \sigma_\beta^2 = 1.$$

Utilizando el mismo mecanismo de censura progresiva por la derecha comentado con anterioridad, obtuvimos un 22.6% de tiempos censurados en el primer grupo y un 13.8% en el segundo.

Considerando un punto inicial con el mismo criterio anterior, utilizamos $\mathbf{a} = (3, 6, 10, 20)'$ como partición temporal y generamos una MCMC desechando los 100000 primeros pasos y efectuando saltos de 50 pasos hasta un tamaño muestral igual a 10000. Esto supuso aproximadamente 480 minutos de tiempo CPU. Ha de hacerse notar que la diferencia en los tiempos de computación con el banco de datos anterior es debida al gran número de

parámetros que se maneja en éste. Como en total hay 1000 individuos, ello supone más de 2000 parámetros en la MCMC.

En la figura 3.9 puede observarse la evolución de la MCMC para algunos parámetros de la misma y la estimación Monte Carlo de la distribución marginal. Los parámetros del azar poligonal presentaban una rápida convergencia casi desde el principio de la cadena, por ello hemos representado los parámetros asociados al azar Gamma, que son los que presentan mayor autocorrelación.

Los intervalos intercuantílicos de los parámetros obtenidos por Monte Carlo con la cadena anterior se reflejan en la figura 3.10. Así como las estimaciones de los parámetros asociados a la función de azar poligonal son generalmente buenas y se obtienen con gran rapidez, parece necesaria mayor información (más datos) para la estimación de los parámetros relativos a las covariables. Como es habitual, los parámetros con mayor dificultad en ser estimados son los relacionados con las varianzas. En concreto, μ_β resulta difícil de estimar. Destacar, no obstante, que siendo la predicción el objetivo principal en el análisis de supervivencia, el problema de la estimación queda relegado a un segundo término, máxime cuando se obtienen predicciones razonablemente buenas con las estimaciones presentadas.

Los diagnósticos de convergencia proporcionados por CODA, utilizando los métodos de Geweke (figura 3.11) y de Raftery y Lewis (figura 3.12), indican convergencia para la mayoría de los parámetros, si bien la fuerte autocorrelación en los hiperparámetros de la distribución Gamma no permite obtener mejores resultados para ellos.

Finalmente, considerando todas las covariables en el modelo y utilizando los 100 últimos pasos no repetidos de la MCMC en cada uno de los grupos,

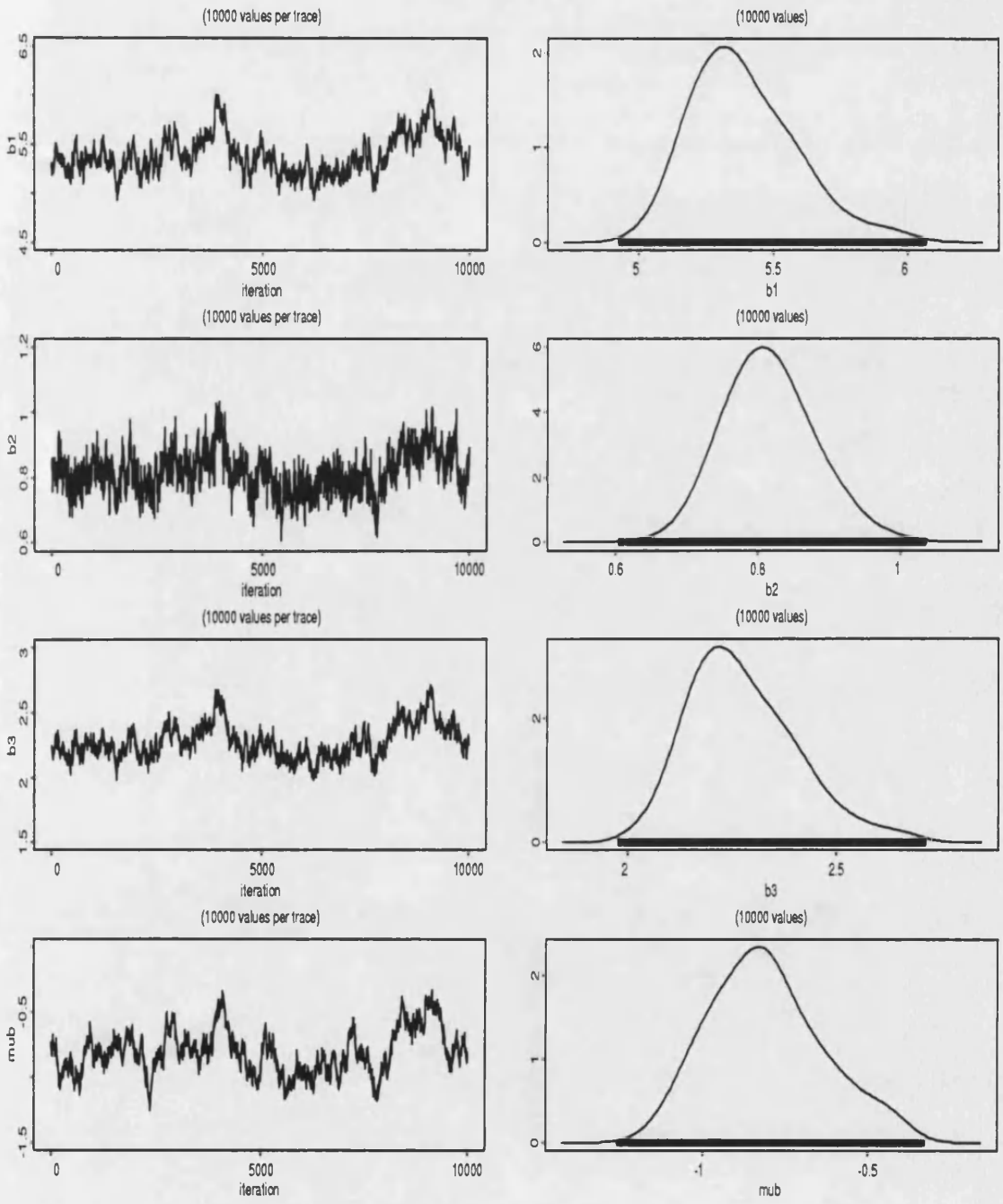

Figura 3.9: *Evolución y densidad predictiva marginal, obtenidas con CODA, de algunos parámetros de la MCMC. Banco de datos simulados 2.*

SUMMARY STATISTICS

Quantiles for each variable:

VARIABLE	2.5%	50%	97.5%
b1	5.08000	5.36000	5.87000
b2	0.70200	0.81200	0.94700
b3	2.07000	2.25000	2.58000
mub	-1.10000	-0.82000	-0.43500
tao11	0.04420	0.06680	0.09750
tao21	0.05900	0.07850	0.10200
tao31	0.02700	0.04330	0.06270
tao41	0.01020	0.02010	0.03190
tao51	0.02350	0.03310	0.04480
tao12	0.12200	0.16300	0.21200
tao22	0.06230	0.08700	0.11600
tao32	0.00664	0.02150	0.03990
tao42	0.04110	0.06160	0.08520
tao52	0.07760	0.10200	0.12700

Figura 3.10: *Intervalos intercuantílicos, obtenidos con CODA, de algunos parámetros de la MCMC. Banco de datos simulados 2.*

```
GEWEKE CONVERGENCE DIAGNOSTIC (Z-score):
========================================
      Fraction in 1st window = 0.1
      Fraction in 2nd window = 0.5
      -+----------+-------------+-
        | VARIABLE |     MCMC    |
        | ======== |     ====    |
        | b1       |   -3.750    |
        | b2       |   -0.628    |
        | b3       |   -4.590    |
        | mub      |   -5.210    |
        | tao11    |   -2.490    |
        | tao21    |   -0.677    |
        | tao31    |    1.980    |
        | tao41    |    0.399    |
        | tao51    |    0.908    |
        | tao12    |   -2.050    |
        | tao22    |    1.730    |
        | tao32    |   -1.200    |
        | tao42    |   -1.140    |
        | tao52    |   -3.160    |
      -+----------+-------------+-
```

Figura 3.11: *Análisis de la convergencia de algunos parámetros de la MCMC obtenido con CODA. Método de Geweke. Banco de datos simulados 2.*

RAFTERY AND LEWIS CONVERGENCE DIAGNOSTIC:

Quantile = 0.025
Accuracy = +/- 0.005
Probability = 0.95

VARIABLE	Thin (k)	Burn-in (M)	Total (N)	Lower bound (Nmin)	Dependence factor (I)
b1	4	72	71160	3746	19
b2	6	66	73998	3746	19.8
b3	4	104	109904	3746	29.3
mub	6	186	189438	3746	50.6
tao11	1	27	28890	3746	7.71
tao21	1	9	10461	3746	2.79
tao31	1	11	12334	3746	3.29
tao41	1	20	21652	3746	5.78
tao51	1	11	12284	3746	3.28
tao12	1	9	10461	3746	2.79
tao22	1	16	17317	3746	4.62
tao32	1	48	51638	3746	13.8
tao42	1	24	25731	3746	6.87
tao52	1	20	21937	3746	5.86

Figura 3.12: *Análisis de la convergencia de algunos parámetros de la MCMC obtenido con CODA. Método de Raftery y Lewis. Banco de datos simulados 2.*

obtuvimos funciones predictivas de algunos individuos concretos. Como alternativa a las funciones de supervivencia, en la figura 3.13 se representan las gráficas de las densidades predictivas y de las verdaderas.

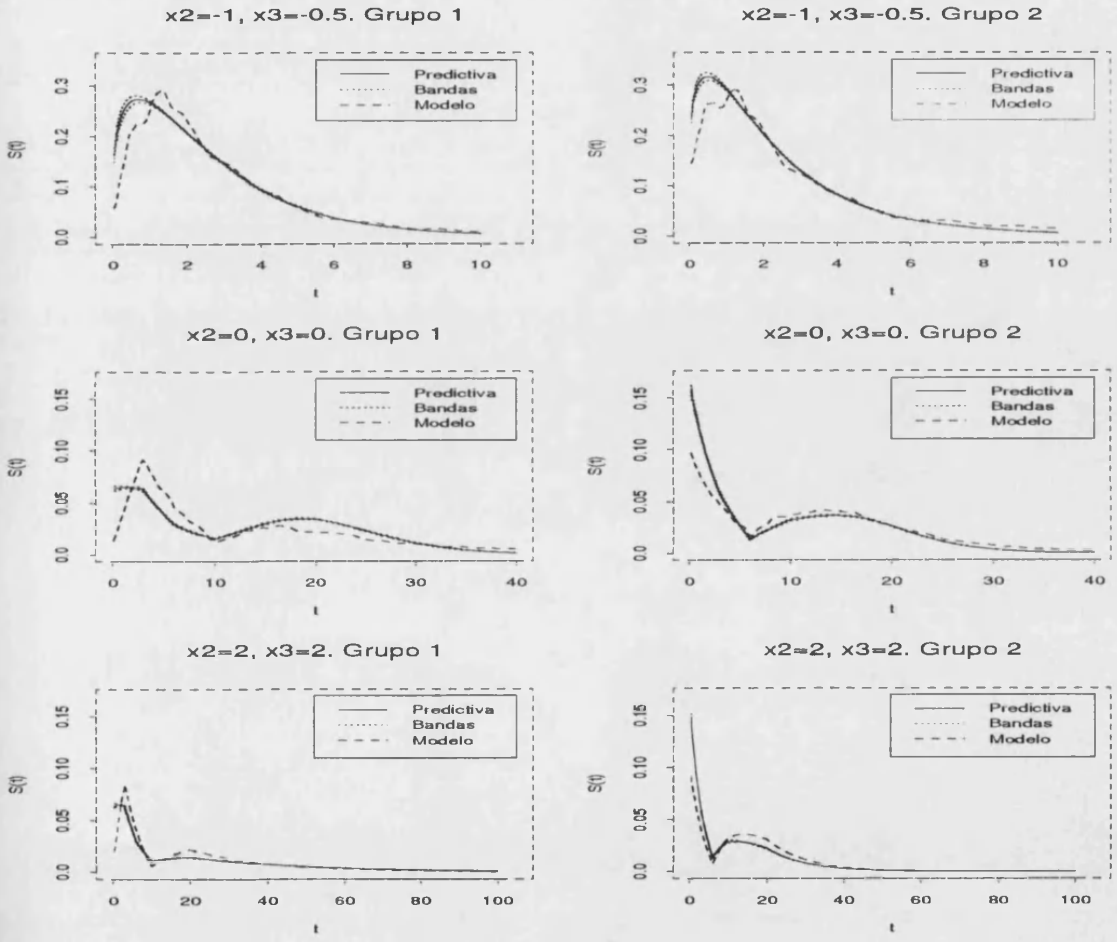

Figura 3.13: *Función de densidad del modelo correcto y su predicción para algunos individuos. Banco de datos simulados 2.*

Es de destacar el gran parecido entre la predicción y la realidad para todos los conjuntos de covariables estudiados. Asímismo, también es de resaltar que la predictiva suaviza la atípica densidad del modelo. Aunque se han utilizado tan sólo 100 valores muestrales, los errores Monte Carlo come-

tidos en las predicciones son, nuevamente, tan reducidos que las bandas de confianza al 95% se confunden prácticamente con las densidades predictivas.

La tabla 3.3 recoge las medias y desviaciones típicas de las distribuciones predictivas para dichos valores concretos de las covariables.

	x_2=-1, x_3=-0.5		x_2=0, x_3=0		x_2=2, x_3=2	
	Grupo 1	Grupo 2	Grupo 1	Grupo 2	Grupo 1	Grupo 2
modelo	m=3.35	m=3.28	m=15.16	m=12.88	m=20.33	m=15.49
	dt=3.59	dt=3.64	dt=14.22	dt=10.14	dt=19.41	dt=12.3
predictiva	m=2.59	m=2.35	m=13.13	m=9.93	m=24.51	m=11.94
	dt=2.42	dt=2.25	dt=10.08	dt=8.38	dt=29.65	dt=11.82

Tabla 3.3: *Media y desviación típica del modelo correcto y de la densidad predictiva para algunos individuos. Banco de datos simulados 2.*

3.4 Datos reales

3.4.1 Banco de datos de Stanford

El conocido banco de datos de transplantes de corazón de Stanford (Miller y Halpern, 1982) tiene registrados a 184 individuos (hasta el 1 de Febrero de 1984) con medidas de los días de supervivencia tras un transplante de corazón, la edad en años al transplantar y un coeficiente de disimilaridad entre paciente y donante. En este estudio hemos considerado un único grupo con los 156 pacientes para los que se disponía de toda la información anterior (ver apéndice C) y tres covariables: x_1, constante e igual a 1, x_2, la edad y x_3, el coeficiente de disimilaridad. En este banco de datos contamos con un 35.26% de tiempos censurados.

Debido a la gran variabilidad que presentan los tiempos de supervivencia y a los elevados valores que toman los mismos, realizamos un cambio (ya propuesto por numerosos autores en la literatura) a una escala logarítmica del tiempo.

Con posterioridad, obtuvimos una aproximación a la función de azar poligonal monitorizando las medias de los vectores $\boldsymbol{\tau}^{(1)}$ obtenidos por Monte Carlo. La figura 3.14 refleja dicha aproximación. Observados los cambios de monotonicidad en esos puntos, decidimos utilizar como divisiones no triviales del eje temporal $\mathbf{a} = (3, 4, 6)'$, tratando de reflejar en pocos puntos la forma de la función de azar poligonal. El último cambio en la monotonicidad alrededor de 7 no fue tenido en cuenta por la falta de información en el último intervalo.

Partiendo del mismo punto inicial que en el banco de datos 1, obtuvimos

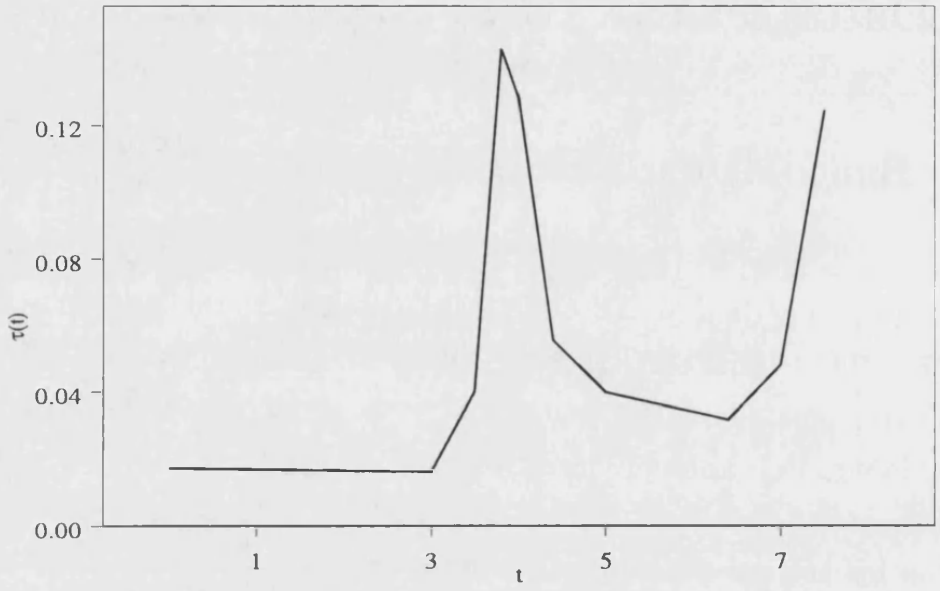

Figura 3.14: *Aproximación al azar poligonal. Banco de datos de Stanford.*

una MCMC desechando los 100000 primeros pasos y efectuando saltos de 10 pasos hasta completar la cadena de tamaño 10000 (33 minutos). En la figura 3.15 se representa la evolución de la MCMC para algunos parámetros de la misma y la estimación Monte Carlo de la distribución marginal.

Los resultados del análisis de la convergencia con CODA, representados en la figura 3.16, indican una rápida convergencia en la MCMC. Únicamente el parámetro μ_β falla en los diagnósticos.

Tal y como se aprecia en la figura 3.17, el cero está incluido claramente en el intervalo de confianza del 95% de b_3, por lo tanto eliminamos la covariable del coeficiente de disimilaridad del estudio y obtuvimos, con las restantes, una MCMC de tamaño 1000 con la que realizar las predicciones. En la figura 3.18 se representan las funciones de supervivencia predictivas para los valores $x_2 = 15$, $x_2 = 40$ y $x_2 = 60$ de la covariable edad. Las

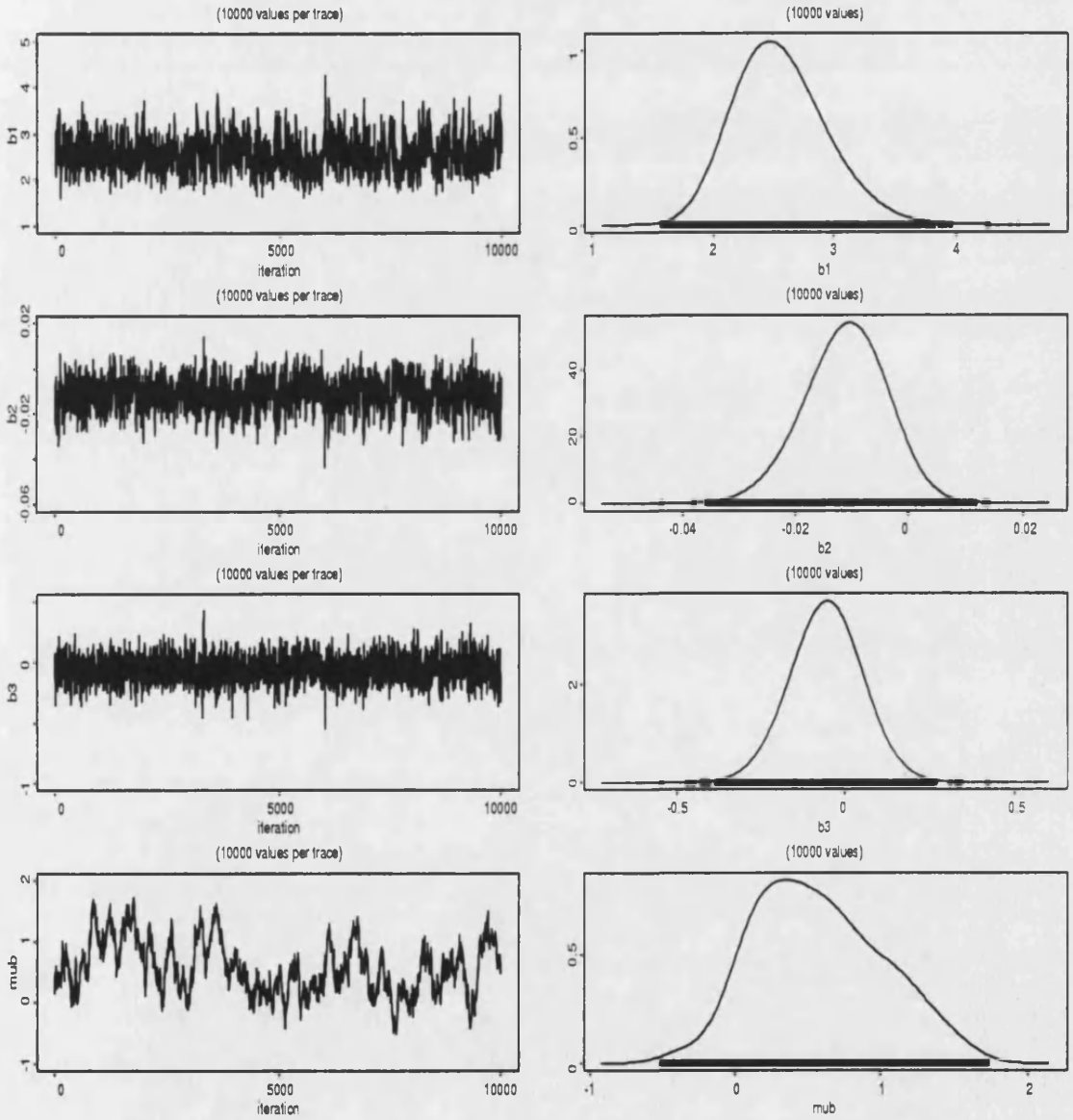

Figura 3.15: *Evolución y densidad predictiva marginal, obtenidas con CODA, de algunos parámetros de la MCMC. Banco de datos de Stanford.*

GEWEKE CONVERGENCE DIAGNOSTIC (Z-score):

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

VARIABLE	MCMC
b1	1.490
b2	-1.350
b3	-1.660
mub	4.910
tao11	0.188
tao21	2.040
tao31	2.500
tao41	0.127

RAFTERY AND LEWIS CONVERGENCE DIAGNOSTIC:

Quantile = 0.025
Accuracy = +/- 0.005
Probability = 0.95

VARIABLE	Thin (k)	Burn-in (M)	Total (N)	Lower bound (Nmin)	Dependence factor (I)
b1	2	10	12476	3746	3.33
b2	5	25	27115	3746	7.24
b3	2	10	11688	3746	3.12
mub	1	96	104591	3746	27.9
tao11	1	26	28317	3746	7.56
tao21	3	15	19893	3746	5.31
tao31	4	12	17440	3746	4.66
tao41	3	15	16953	3746	4.53

Figura 3.16: *Análisis de la convergencia de algunos parámetros de la MCMC obtenido con CODA. Banco de datos de Stanford.*

medias, obtenidas por Monte Carlo, de las densidades predictivas fueron, respectivamente, 450.34, 278.66 y 119.1. Puede observarse cómo se produce la lógica y esperada ordenación de las supervivencias en sentido inverso al orden de las edades.

SUMMARY STATISTICS

Quantiles for each variable:

VARIABLE	2.5%	50%	97.5%
b1	1.97000	2.52000	3.32000
b2	-0.02530	-0.01090	0.00070
b3	-0.24300	-0.05450	0.12600
mub	-0.14800	0.53000	1.46000
tao11	0.00305	0.01460	0.04140
tao21	0.00392	0.01820	0.04400
tao31	0.00470	0.11400	0.24000
tao41	0.00284	0.04330	0.17000

Figura 3.17: *Intervalos intercuantílicos, obtenidos con CODA, de algunos parámetros de la MCMC. Banco de datos de Stanford.*

La gráfica 3.19 puede ser de utilidad para la comparación de las supervivencias predichas por este modelo y las supervivencias predictivas obtenidas utilizando el modelo de Cox que, considerándose adecuado, es el habitualmente utilizado para el análisis de este banco de datos. Nuestro modelo proporciona resultados similares.

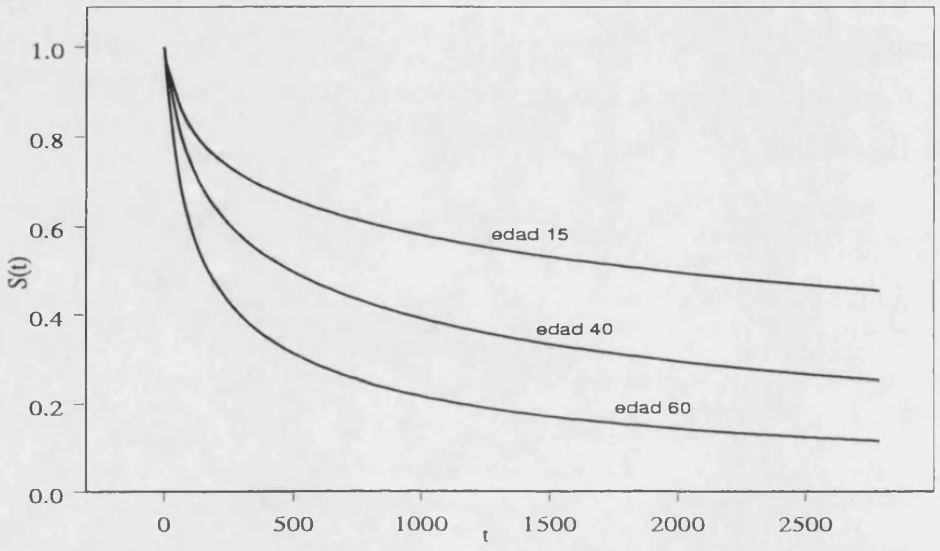

Figura 3.18: *Función de supervivencia predictiva para algunos individuos. Banco de datos de Stanford.*

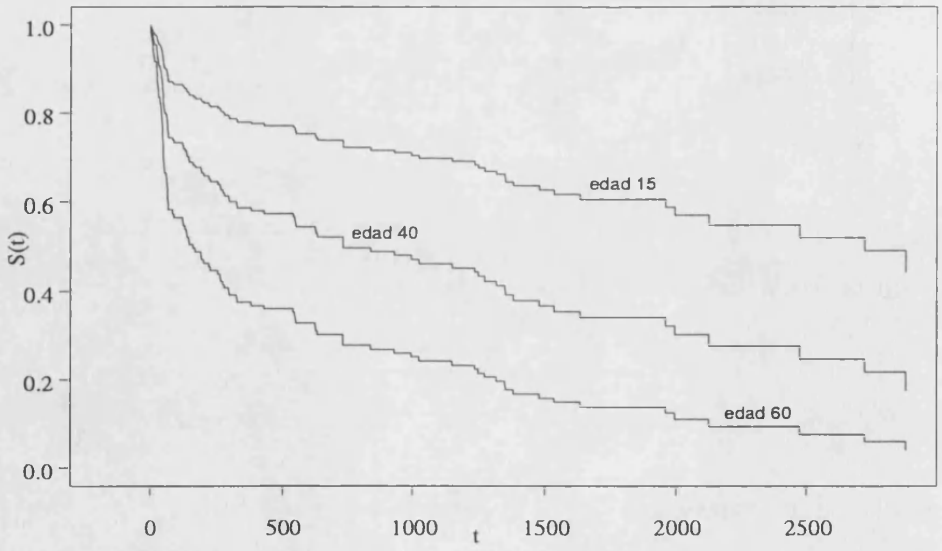

Figura 3.19: *Función de supervivencia predictiva de Cox para algunos individuos. Banco de datos de Stanford.*

3.4.2 Banco de datos de una encuesta a licenciados en Matemáticas

Los datos analizados en esta sección proceden de una encuesta dirigida a los licenciados en Ciencias Matemáticas por la Universitat de València. La labor de campo con los cuestionarios cumplimentados se cerró en Octubre de 1994 (Encuesta sobre la valoración de la adecuación de los estudios a la actividad profesional. Convenio de la Universitat de València con la Conselleria de Educació i Ciència).

En este estudio se consideran únicamente seis de los items de la encuesta (ver apéndice C): la nota media de licenciatura (aprobado, notable y sobresaliente), que utilizamos para formar tres estratos o grupos distintos; el tiempo transcurrido, en meses, desde la obtención de la licenciatura hasta la consecución del primer empleo, que va a ser nuestro tiempo de supervivencia; una variable indicadora de la censura (0, si no ha encontrado el primer empleo al cierre de la encuesta y 1, en caso contrario), el sexo (0, mujer y 1, hombre); el año de licenciatura y la actitud ante el hecho de volver a cursar los mismos estudios (0, no y 1, sí). Estos tres últimos items son los que hemos utilizado como covariables, añadiéndoles una primera covariable constante e igual a 1 para la inclusión del término independiente en la combinación lineal.

En el primer grupo hay 352 individuos con un 10.79% de censura, en el segundo 171 con un 5.85% y hay 36 individuos con nota media sobresaliente y uno solo en situación de desempleo, lo que significa una censura del 2.78% en este grupo.

En la figura 3.20 se representa la aproximación a la función de azar poligonal en cada grupo, obtenida con las medias Monte Carlo de los vectores

$\tau^{(j)}$, $j = 1, 2, 3$. En ella se observa un comportamiento cíclico para la función de azar que se suaviza a lo largo del tiempo. Resulta más acentuado en el grupo de nota media aprobado (tiempos de supervivencia más grandes), mientras que el de nota media sobresaliente (el de menor tarnaño y con los menores tiempos de supervivencia) apenas permite apreciar dicho comportamiento. Coincide este ciclo anual en el tiempo de desempleo de estos licenciados con el inicio de curso en la enseñanza secundaria. Siendo ésta su principal salida laboral, es en los meses de septiembre y octubre, coincidiendo con las contrataciones por parte de las instituciones públicas, cuando muchos de estos licenciados acceden a su primer empleo.

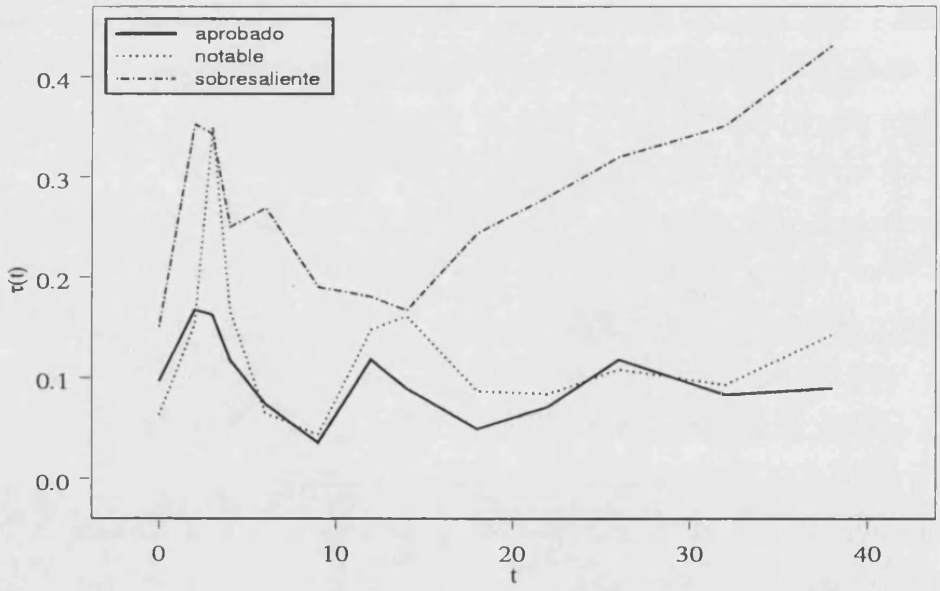

Figura 3.20: *Aproximación al azar poligonal de cada grupo de individuos. Banco de datos de la encuesta.*

Dada la estacionariedad observada en la figura 3.20, decidimos utilizar $\mathbf{a} = (2, 3, 4, 6, 12, 18, 24)'$ como partición del eje temporal. Los primeros intervalos de la partición son de amplitud uno pues una cantidad considerable de los tiempos de desempleo eran uno, dos o tres meses. Tomamos

$\tau^{(j)} = (0.02, 0.08, 0.06, 0.02, 0.01, 0.08, 0.02, 0.1)'$, $j = 1, 2, 3$, como punto inicial para los parámetros relativos al azar poligonal de los tres grupos y $\mathbf{b} = \mathbf{0}$, $\sigma_\alpha^2 = 1$, $\mu_\beta = 0$, y $\sigma_\beta^2 = 1$, como valores iniciales para los hiperparámetros de las covariables. Con estos valores iniciales de la MCMC, distintos a los utilizados con anterioridad y con los que pretendíamos recoger el marcado carácter temporal de los parámetros $\tau^{(j)}$, $j = 1, 2, 3$, así como acelerar la convergencia, generamos una MCMC desechando los 100000 primeros pasos y efectuando saltos de 50 pasos hasta un tamaño muestral igual a 10000 (360 minutos). Este elevado tiempo de cómputo es debido a que el banco de datos consta de 559 individuos, lo que supone más de 1100 parámetros.

En las figuras 3.21 y 3.22 se refleja la evolución de algunos parámetros de la MCMC y la estimación Monte Carlo de su distribución marginal. Los parámetros relativos al azar poligonal mostraban una muy rápida convergencia, si bien se producían bastantes repeticiones en las etapas de Metropolis. Debido a ello, todos los análisis de convergencia y evolución efectuados con CODA se realizaron utilizando tan sólo aquellos puntos de la MCMC donde no había repeticiones en los parámetros $\tau^{(j)}$, para $j = 2$, pues éste era el grupo que presentaba más repeticiones.

A continuación se proporcionan, en la figura 3.23, los intervalos de confianza del 95% para algunos de los parámetros de las covariables obtenidos con la MCMC anterior.

En la figura 3.24 se proporcionan los resultados del análisis de la convergencia de los parámetros de las covariables efectuado con CODA.

Como en ambos casos quedaba incluido el cero en el correspondiente intervalo de confianza del 95%, eliminamos secuencialmente las covariables

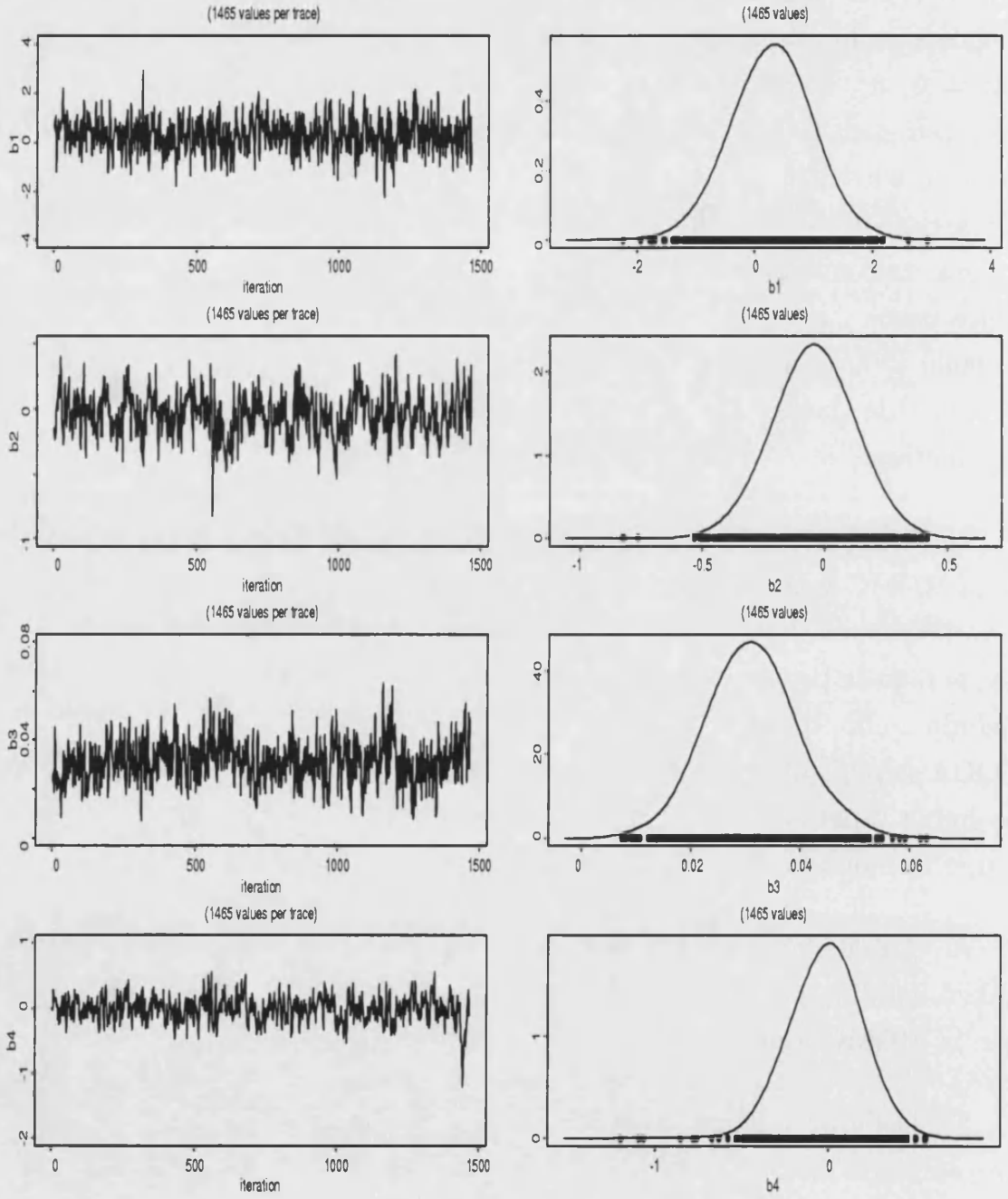

Figura 3.21: *Evolución y densidad predictiva marginal, obtenidas con CODA, de algunos parámetros de la MCMC asociados a las covariables. Banco de datos de la encuesta.*

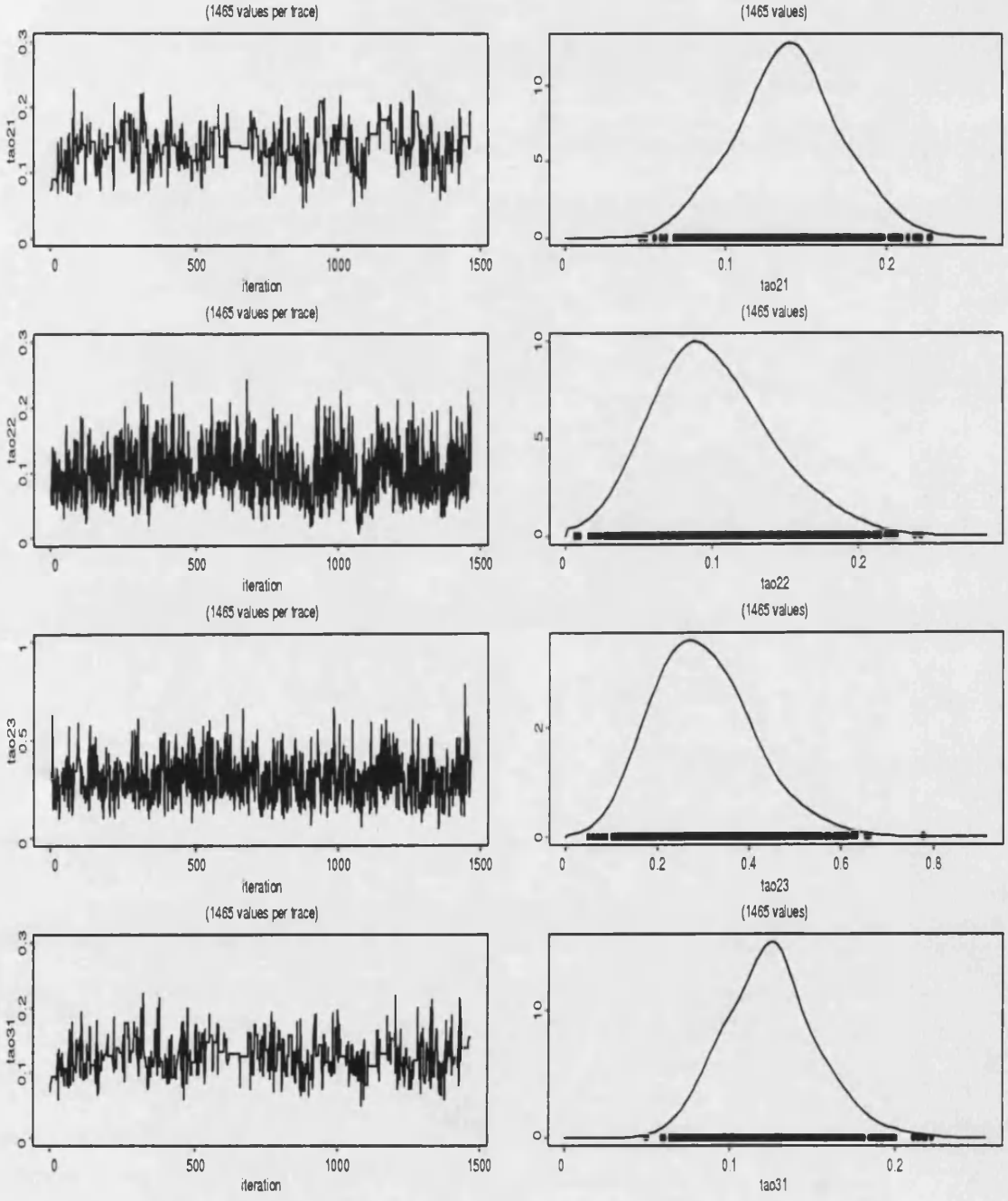

Figura 3.22: *Evolución y densidad predictiva marginal, obtenidas con CODA, de algunos parámetros de la MCMC relativos al azar poligonal. Banco de datos de la encuesta.*

SUMMARY STATISTICS

Quantiles for each variable:

VARIABLE	2.5%	50%	97.5%
b1	-0.98200	0.30800	1.60000
b2	-0.35300	-0.04550	0.26200
b3	0.01590	0.03120	0.04810
b4	-0.37900	-0.00820	0.34700
mub	-2.96000	-2.48000	-2.19000

Figura 3.23: *Intervalos intercuantílicos, obtenidos con CODA, de algunos parámetros de la MCMC asociados a las covariables. Banco de datos de la encuesta.*

sexo y actitud del modelo y obtuvimos una MCMC de tamaño 1000 para predecir supervivencias. En la figura 3.25 se representan las funciones de supervivencia predictivas para distintos valores de las covariables finalmente consideradas.

En la tabla 3.4 se proporcionan las medias y desviaciones típicas predictivas obtenidas por Monte Carlo para algunos individuos.

	APROBADO			NOTABLE			SOBRESALIENTE		
	1978	1988	1998	1978	1988	1998	1978	1988	1998
media	8.19	10.41	12.81	6.87	8.41	9.97	3.55	3.91	4.11
desviación	9.76	12.02	15.02	8.02	9.66	11.75	3.77	4.15	4.46

Tabla 3.4: *Media y desviación típica de la densidad predictiva para algunos individuos. Banco de datos de la encuesta.*

Por comparación, al igual que en el banco de datos anterior, proporcionamos algunas predicciones de supervivencias utilizando el modelo de

```
GEWEKE CONVERGENCE DIAGNOSTIC (Z-score):
========================================
      Fraction in 1st window = 0.1
      Fraction in 2nd window = 0.5
      -+----------+-------------+-
        | VARIABLE |     MCMC    |
        | ======== |     ====    |
        | b1       |    2.280    |
        | b2       |    1.260    |
        | b3       |   -3.530    |
        | b4       |    1.550    |
        | mub      |    2.000    |
      -+----------+-------------+-

RAFTERY AND LEWIS CONVERGENCE DIAGNOSTIC:
=========================================
                Quantile = 0.01
              Accuracy = +/- 0.005
               Probability = 0.9
-+----------+-------------------------------------------------------+-
  |          |  Thin    Burn-in  Total   Lower bound  Dependence     |
  | VARIABLE |  (k)       (M)    (N)       (Nmin)    factor (I)     |
  | ======== |  ====    =======  =====   ===========  ==========     |
  | b1       |  1       6        1871    1072         1.75           |
  | b2       |  1       14       4348    1072         4.06           |
  | b3       |  1       8        2532    1072         2.36           |
  | b4       |  1       17       5437    1072         5.07           |
  | mub      |  1       31       9796    1072         9.14           |
-+----------+-------------------------------------------------------+-
```

Figura 3.24: *Análisis de la convergencia, obtenida con CODA, de algunos parámetros de la MCMC asociados a las covariables. Banco de datos de la encuesta.*

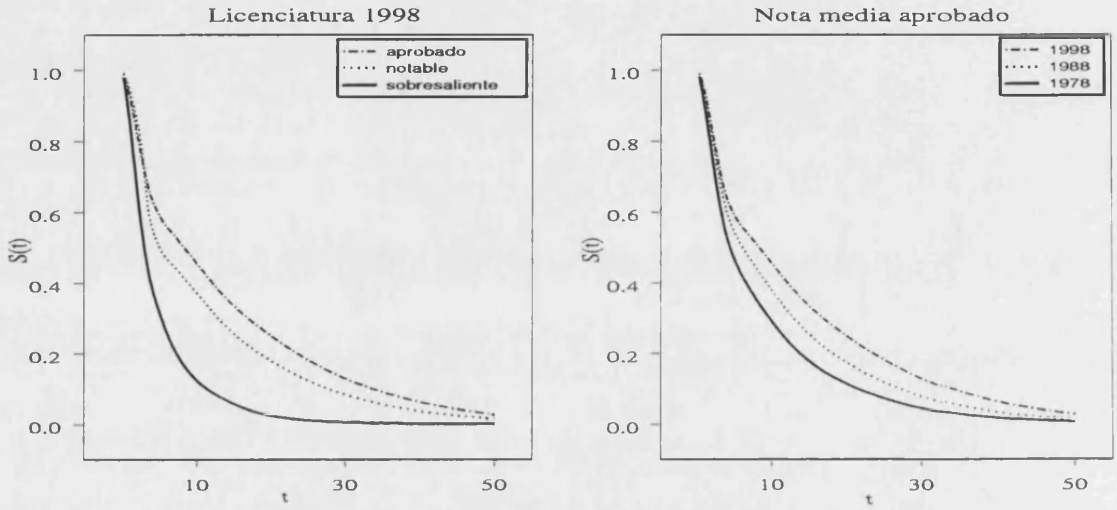

Figura 3.25: *Función de supervivencia predictiva para algunos individuos. Banco de datos de la encuesta.*

Cox. En la figura 3.26 pueden observarse las supervivencias predictivas de Cox para el año 1998 y para los tres grupos y en la figura 3.27 las correspondientes a la nota media de licenciatura aprobado para los años 1978, 1988 y 1998. En ambas, puede apreciarse el gran parecido entre ellas y las predicciones realizadas con nuestro modelo. Con ambos modellos, las supervivencias predichas para el año de licenciatura 1998 son muy similares para las notas medias aprobado y notable siendo menores para el sobresaliente, indicando la mayor accesibilidad al primer empleo para los licenciados con esa nota media. La supervivencia en función del año de licemciatura también aparece ordenada en sentido directo al mismo, proporcionando mayores tiempos de espera hasta el primer empleo para los licenciados de las últimas promociones.

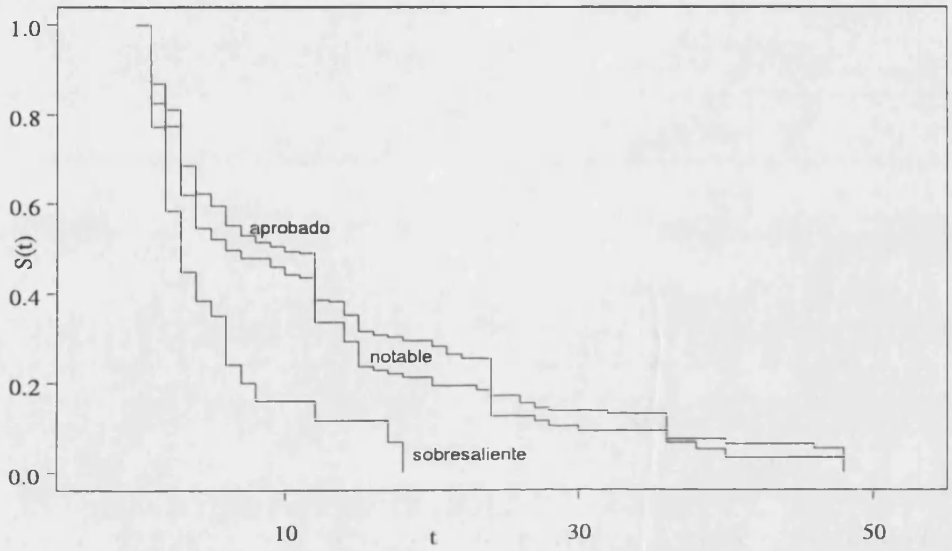

Figura 3.26: *Función de supervivencia predictiva de Cox para algunos individuos con año de licenciatura 1998. Banco de datos de la encuesta.*

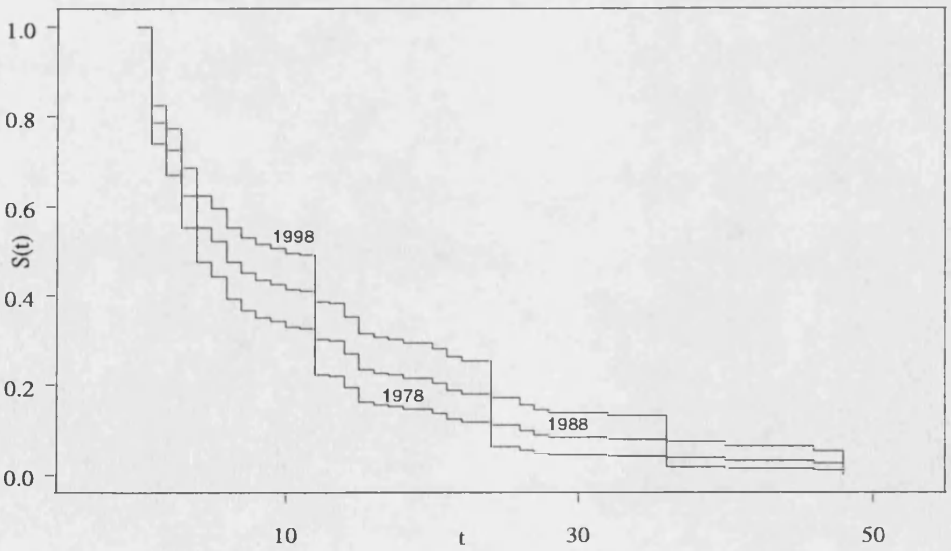

Figura 3.27: *Función de supervivencia predictiva de Cox para algunos individuos con nota mediia de licenciatura aprobado. Banco de datos de la encuesta.*

Capítulo 4

Conclusiones y futuras líneas de investigación

En este trabajo presentamos un modelo semiparamétrico aditivo como alternativa al modelo de Cox para el análisis bayesiano de datos de supervivencia. El modelo que desarrollamos permite la incorporación de covariables de un modo sencillo y admite un tratamiento bayesiano del problema de supervivencia con una matemática a desarrollar razonable.

Hemos analizado, dentro de los modelos semiparamétricos aditivos, con mayor detalle el modelo Gamma-poligonal aditivo, en el que la parte paramétrica corresponde a la distribución Gamma, pero otras distribuciones paramétricas pueden ser utilizadas en su lugar. La metodología de análisis a emplear sería la misma y en su implementación sólo se vería afectada la parte paramétrica del modelo.

Contrariamente a lo que es habitual en el análisis bayesiano de datos de supervivencia con covariables, la implementación del modelo Gamma-poligonal no ha resultado excesivamente complicada y utilizando técnicas de Monte Carlo permite el análisis, bajo la perspectiva bayesiana, de complejos bancos de datos, con gran cantidad de parámetros, en tiempos de cómputo razonables.

También ha resultado sencilla y fácilmente aplicable la técnica desarrollada para realizar la selección de variables.

El modelo Gamma-poligonal aditivo se adecúa bien a los modelos de azares proporcionales (aquellas situaciones de supervivencia donde funciona bien el modelo de Cox), habiendo proporcionado buenos resultados en todos los bancos de datos analizados con esa característica. También ha proporcionado generalmente buenos resultados con modelos de azares no proporcionales, donde el modelo de Cox no es adecuado, si bien la investigación en este sentido está abierta en orden a establecer algunas condiciones

de aplicabilidad de nuestro modelo.

Los bancos de datos presentados en la memoria constituyen una pequeña parte de una batería de bancos de datos analizados. Teniendo presente que el primordial objetivo en el análisis de supervivencia es la predicción para nuevos individuos, cabe destacar lo razonable de las predicciones realizadas en todos los bancos de datos estudiados. No sólo en los bancos de datos simulados, en los que se conocían los verdaderos parámetros del modelo y en los que los resultados confirmaron la adecuación del mismo, sino en estudios de datos reales donde las conclusiones alcanzadas fueron perfectamente extrapolables a las proporcionadas por otros autores.

Otra línea de investigación en la que cabe incidir en un futuro próximo es la de llevar a cabo una mayor profundización en aspectos teóricos del modelo, principalmente en el sentido de establecer relaciones con el modelo de azares proporcionales de Cox. Siendo éste el modelo comúnmente utilizado en la literatura para el tratamiento de tiempos de supervivencia con covariables resulta necesario concretar, con mayor rigurosidad, las analogías y diferencias entre ambos.

La utilización de una familia distinta a la Gamma para la parte paramétrica del modelo puede ampliar el ámbito de aplicación del mismo y también nos ocupa desde este momento.

La inclusión en el modelo de covariables dependientes del tiempo puede ser fundamental para completar la aplicabilidad del modelo y por ello ha de ser materia inmediata de investigación.

También queda planteado a corto plazo el desarrollo de alguna herramienta estadística, en forma de test o similar, para la comparación de estratos.

Asímismo, nos planteamos para un futuro inmediato un estudio más detallado de la convergencia en los procesos de simulación y el desarrollo de algún mecanismo automático de detección de la misma.

Apéndice A

Definiciones y propiedades en supervivencia

Con este apéndice tan sólo se pretende resumir un pequeño vocabulario específico de análisis de supervivencia. En cualquier manual básico pueden encontrarse estos y otros conceptos más detallados y profundizar en las propiedades de las funciones aquí definidas. Buenas referencias son las de Gross y Clark (1975), Johnson (1980), Kalbfleisch y Prentice (1980), Miller (1981), Cox y Oakes (1984), Parmar y Machin (1995), o Le (1997), entre otras.

Definición A.1 *Un dato de vida o tiempo de supervivencia es el tiempo transcurrido entre dos sucesos bien definidos.*

Definición A.2 *Un dato de vida o tiempo de supervivencia se dice censurado cuando alguno de los sucesos que lo definen no es observado.*

Existen diferentes tipos de censura dependiendo de cuál es el suceso no observado y de la forma de entrar los individuos en el estudio (ver, por

ejemplo, Lee, 1992; Collett, 1994; Klein y Moeschberger, 1997). Entre ellos, uno de los más frecuentes en la práctica habitual es el de la censura progresiva por la derecha, que se tiene cuando no todos los individuos entran en el estudio al mismo tiempo y es el segundo de los sucesos el que no se observa para algunos de ellos.

Denotamos por T a la variable aleatoria positiva que representa el tiempo de vida o supervivencia.

Definición A.3 *La función de supervivencia de T, denotada por $S(t)$, es la probabilidad de que el tiempo de vida sea superior a t,*

$$S(t) = p(T > t). \tag{A.1}$$

De la igualdad (A.1) se obtiene inmediatamente que $S(t) = 1 - F(t)$, donde $F(t)$ es la función de distribución de probabilidades de T.

En el caso de que T sea una variable aleatoria absolutamente continua, tenemos las siguientes definiciones:

Definición A.4 *La función de azar de T, $h(t)$, es la probabilidad de fallo (observación del segundo suceso que define el tiempo de vida) en un intervalo de tiempo infinitesimal dado que no había sido observado al principio del intervalo,*

$$h(t) = \lim_{\Delta t \to 0} \frac{p\,(t < T \leq t + \Delta | T > t)}{\Delta t}.$$

De ello se deduce que:

$$h(t) = \frac{f(t)}{S(t)},$$

donde $f(t)$ es la función de densidad de probabilidades de T.

Definición A.5 *La función de azar acumulado de T, $H(t)$, se define como:*

$$H(t) = \int_0^t h(s)\, ds.$$

Estos últimos conceptos, azar y azar acumulado, pueden establecerse, de un modo totalmente análogo y sin pérdida de generalidad, para el caso menos habitual de variables aleatorias discretas.

A continuación, enunciamos algunas de las propiedades básicas más importantes que relacionan las funciones definidas con anterioridad.

Propiedad A.1

$$f(t) = -\frac{d}{dt}S(t),\ t > 0.$$

Propiedad A.2

$$h(t) = -\frac{d}{dt}\log S(t),\ t > 0.$$

Propiedad A.3

$$S(t) = \exp\left[-H(t)\right],\ t > 0.$$

De modo que dada una cualquiera de las funciones de azar, azar acumulado, supervivencia o densidad, las otras pueden obtenerse fácilmente.

Apéndice B

Procesos estocásticos. Cadenas de Markov

El estudio teórico de los métodos MCMC exige un conocimiento adecuado de procesos estocásticos y, sobre todo, de cadenas de Markov. En este apéndice se recogen las definiciones y propiedades más importantes con ellos relacionadas. Para una consulta más detallada de los conceptos y propiedades aquí desarrollados, así como para la obtención de una visión global de los procesos estocásticos, puede verse, por ejemplo, Parzen (1972), Resnick (1992) y Meyn y Tweedie (1993).

Definición B.1 *Un proceso estocástico es una colección de variables aleatorias* $\{X(t), t \in T\}$ *definidas sobre un espacio probabilístico común e indexadas por* $t \in T$ *que describe la evolución de algún sistema.*

Con frecuencia $T = [0, \infty)$ y se dice que el proceso estocástico es de *tiempo continuo* o $T = \{0, 1, \dots\}$ y entonces se denomina de *tiempo discreto.*

Definición B.2 *Se denomina espacio de estados de un proceso estocástico*

al conjunto de posibles valores que pueden tomar las variables aleatorias que lo componen.

Definición B.3 *Una cadena de Markov es un proceso estocástico* $\{X(t), t \in T\}$ *de conjunto índice* T *discreto verificando que para un conjunto cualquiera de* n *instantes,* $t_1 < t_2 < \cdots < t_n$*, en el conjunto índice del proceso:*

$$p\left(X(t_n) \leq x_n \,|\, X(t_1) \leq x_1, \ldots, X(t_{n-1}) \leq x_{n-1}\right) = \\ = p\left(X(t_n) \leq x_n \,|\, X(t_{n-1}) \leq x_{n-1}\right) \;\; \forall (x_1, \ldots, x_n) \in \mathbb{R}^n, \quad \text{(B.1)}$$

siempre que esté definido el primer miembro de la ecuación (B.1).

Sea una cadena de Markov $\{X(n), n \geq 0\}$ con espacio de estados discreto S.

Definición B.4 *Se define la función de probabilidades de transición de una cadena de Markov como:*

$$p_{ij}(n, m) = p\left(X_m = j \,|\, X_n = i\right),$$

para cualquier par de instantes $m \geq n \geq 0$ *y cualquier par de estados* $i, j \in S$.

Definición B.5 *Una cadena de Markov se dice homogénea si las probabilidades de transición* $p_{ij}(n, m)$ *sólo dependen de la diferencia m-n.*

Definición B.6 *Se define la función de probabilidad de transición de* n *pasos de una cadena de Markov homogénea como:*

$$p_{ij}(n) = p\left(X_{n+t} = j \,|\, X_t = i\right) \;\; \forall t \geq 0.$$

Habitualmente se denota $p_{ij} = p_{ij}(1)$ a la función de probabilidad de transición de un paso.

Definición B.7 *El tiempo de espera de $B \subset S$ es:*

$$\tau_B = \inf\{n \geq 0 : X_n \in B\}.$$

Para subconjuntos $B \subset S$ que consten de un solo estado, denotaremos $\tau_j = \tau_{\{j\}}$.

Definición B.8 *Dados $i, j \in S$, se dice que i es accesible desde j, $i \to j$, si $\exists n \geq 0$ tal que $p_{ij}(n, 0) > 0$. Esto es, la cadena de Markov alcanza el estado j, desde el estado i, con probabilidad uno.*

Definición B.9 *Dos estados $i, j \in S$, se dice que comunican, $i \leftrightarrow j$, si $i \to j$ y $j \to i$.*

Propiedad B.1 *La comunicación entre estados es una relación de equivalencia, esto es:*

1. *$i \leftrightarrow j$ (reflexiva)*
2. *$i \leftrightarrow j$ sii $j \leftrightarrow i$ (simétrica)*
3. *$i \leftrightarrow j, j \leftrightarrow k$ entonces $i \leftrightarrow k$ (transitiva),*

y en consecuencia, crea una partición en el espacio de estados.

Definición B.10 *Un estado i es recurrente si la cadena retorna a i con probabilidad uno en un número finito de pasos. De otro modo, el estado es transitorio.*

Definición B.11 *Un estado es recurrente positivo si es recurrente y el número esperado de pasos hasta la recurrencia es finito.*

Definición B.12 *Se define el período de un estado recurrente i como:*

$$d(i) = mcd\{n \geq 1 : p_{ii}(n, 0) > 0\},$$

donde mcd es el máximo común divisor.

Definición B.13 *Un estado i se dice aperiódico si $d(i) = 1$. Si $d(i) > 1$, entonces es periódico de período i.*

Propiedad B.2 *La recurrencia, transitoriedad y el período de un estado son propiedades solidarias, en el sentido de transmitirse a todos los elementos de una clase de comunicación, es decir, si C es una clase de comunicación e $i \in C$ tiene alguna de estas propiedades, entonces también la tiene cualquier estado $j \in C$.*

Definición B.14 *Una cadena de Markov es aperiódica si todos sus estados son aperiódicos.*

Definición B.15 *Una cadena de Markov se dice recurrente si todos sus estados son recurrentes.*

Definición B.116 *Una cadena de Markov se dice recurrente positiva si todos sus estados son recurrentes positivos.*

Definición B.117 *Una cadena de Markov es irreducible si* $i \leftrightarrow j \ \forall i, j \in S$, *es decir,* S *consta de una sola clase de comunicación.*

Definición B.118 *Una distribución de probabilidad* $\pi = \{\pi_j,\ j \in S\}$ *es una distribución estacionaria para la cadena de Markov si:*

$$\pi_j = \sum_{k \in S} \pi_k \, p_{kj} \ \forall j \in S.$$

Definición B.119 *Se dice que una cadena de Markov con distribución estacionaria* $\pi = \{\{\pi_j,\ j \in S\}$ *es reversible si:*

$$\pi_j \, p_{ji}(n) = \pi_i \, p_{ij}(n) \ \ \forall i, j \in S, n \geq 0.$$

Definición B.120 *Una distribución de probabilidad* $\pi = \{\pi_j,\ j \in S\}$ *es una distribución límite para la cadena de Markov si:*

$$\lim_{n \to \infty} p_{ij}(n) = \pi_j \ \ \forall i, j \in S.$$

Definición B.121 *Se dice que una cadena de Markov es ergódica si es recurrente positiva, aperiódica e irreducible.*

Propiedad B.3 *Si una cadena de Markov es ergódica entonces existe una distribución estacionaria para ella.*

Propiedad B.4 *Una cadena de Markov irreducible y reversible es recurrente positiva.*

Propiedad B.5 *Dada una cadena de Markov irreducible, entonces existe una distribución estacionaria para ella si y sólo si la cadena es recurrente positiva.*

Propiedad B.6 *Dada una cadena de Markov irreducible y aperiódica, entonces existe la distribución límite si y sólo si la cadena es recurrente positiva.*

Apêndice C

Bancos de datos de Stanford y de una encuesta a licenciados en Matemáticas

Banco de datos de Stanford

t, tiempo de vida en días después del transplante cardíaco,
estado, 0, vivo y 1, muerto, al finalizar el estudio (1 de Febrero de 1984),
edad, edad en años en el momento del transplante,
T5, coeficiente de disimilaridad entre paciente y donante.

Fuente: Miller y Halpern (1982).

t	estado	edad	T5
15	1	54	1.11
3	1	40	1.66
46	1	42	0.61
623	1	51	1.32

t	estado	edad	T5
1778	0	27	0.70
1722	0	40	0.95
928	1	50	1.12
1718	0	39	1.77

(continúa)

t	estado	edad	T5
126	1	48	0.36
64	1	54	1.89
1350	1	54	0.87
23	1	56	2.05
279	1	49	1.12
1024	1	43	1.13
10	1	56	2.76
39	1	42	1.38
730	1	58	0.96
1961	1	33	1.06
136	1	52	1.62
1	1	54	0.47
836	1	44	1.58
60	1	64	0.69
3695	0	40	0.38
1996	1	49	0.91
47	1	62	0.87
54	1	49	2.09
2878	1	49	0.75
3410	0	45	0.98
44	1	36	0.0
994	1	48	0.81
51	1	47	1.38
1478	1	36	1.35
254	1	48	1.08
51	1	52	1.51
323	1	48	1.82
3021	0	38	0.98
66	1	49	0.66
2984	0	32	0.19
2723	1	32	1.93
550	1	48	0.12
66	1	51	1.12

t	estado	edad	T5
22	1	27	1.64
40	1	42	1.59
7	1	28	1.00
1638	0	48	0.43
1612	0	51	1.25
25	1	52	0.5
1534	1	44	1.71
1547	0	50	0.18
1271	1	32	1.05
44	1	46	1.71
1247	1	41	0.43
1232	1	18	0.70
191	1	42	1.74
1393	0	46	0.95
1378	0	41	1.65
1373	0	41	1.38
274	1	31	0.58
31	1	33	0.36
1341	0	50	1.13
42	1	19	0.63
381	1	45	0.98
1264	0	52	0.64
1262	0	34	1.68
1261	0	47	0.82
47	1	36	0.16
1193	0	24	1.15
626	1	53	1.74
48	1	51	0.99
1150	1	32	2.25
45	1	48	0.65
1116	0	14	0.54
1107	0	18	0.25
1102	0	39	1.35

(continúa)

t	estado	edad	T5
65	1	45	1.68
227	1	19	1.02
2805	0	48	1.20
25	1	53	1.68
631	1	26	1.46
2734	0	47	0.97
12	1	29	0.61
63	1	56	2.16
2474	1	52	1.70
1384	1	46	1.41
544	1	52	1.94
29	1	53	1.08
48	1	53	3.05
297	1	42	0.60
1318	1	48	1.44
1352	1	54	0.68
50	1	46	2.25
547	1	49	0.81
431	1	47	0.33
68	1	51	1.33
26	1	52	0.82
161	1	43	1.20
2313	0	26	0.46
1634	1	23	1.78
146	1	45	0.16
48	1	28	0.77
2127	1	35	0.67
263	1	49	0.48
2106	0	40	0.86
293	1	43	0.70
2025	0	30	1.44
2006	0	15	1.26
2000	0	45	1.46

t	estado	edad	T5
195	1	39	0.73
30	1	34	0.84
1040	0	43	0.50
993	0	30	0.95
729	1	49	1.10
202	1	48	1.24
841	0	48	0.86
265	1	49	1.22
1	1	21	0.47
793	0	19	1.98
328	1	34	1.02
781	0	20	1.12
752	0	43	1.50
738	0	41	0.53
86	1	12	1.26
132	1	46	1.09
663	0	36	0.47
660	0	42	0.75
221	1	35	1.04
90	1	38	1.00
619	0	47	0.90
618	0	50	0.82
576	0	53	2.25
36	1	45	0.20
549	0	40	2.53
548	0	30	0.47
541	0	47	0.43
169	1	51	1.89
122	1	51	1.33
468	0	24	1.39
464	0	38	2.07
10	1	13	1.49
406	0	39	1.18

(continúa)

t	estado	edad	T5
1995	0	47	1.65
1945	0	38	1.28
65	1	55	0.69
731	1	38	0.42
1866	0	49	0.51
538	1	49	2.76
1846	0	44	0.83
68	1	35	0.85

t	estado	edad	T5
391	0	27	1.17
50	1	50	0.50
139	1	51	0.96
322	0	36	1.73
292	0	43	1.40
278	0	41	0.98
145	1	50	0.96
176	0	29	1.72

Banco de datos de una encuesta a licenciados en Matemáticas

t, tiempo transcurrido en meses desde la obtención de la licenciatura hasta el primer empleo,
estado, 0, si no ha encontrado el primer empleo al final del estudio (Octubre de 1994) y 1, en otro caso,
sexo, 0, mujer y 1, hombre,
año, año de licenciatura,
nota, nota media de licenciatura: 1, aprobado, 2, notable y 3, sobresaliente,
act, actitud ante el hecho de volver a cursar la licenciatura: 0, no y 1, sí.

Fuente: Encuesta sobre la valoración de la adecuación de los estudios a la actividad profesional. Convenio de la Universitat de València con la Conselleria de Educació i Ciència. Responsable científico: D. Juan Ferrándiz Ferragud.

t	estado	sexo	año	nota	act
4	1	1	93	1	1
1	1	0	93	1	1
14	0	0	92	1	1
48	1	1	79	1	1
1	1	0	79	1	1
12	1	0	81	1	1
15	0	1	93	1	1
1	1	1	77	1	0
14	1	1	82	1	1
12	1	1	89	1	0
15	0	0	93	1	0
5	1	1	84	1	1
6	1	1	90	1	1
26	1	1	77	1	1
15	1	0	84	1	1
15	0	1	93	1	1
5	1	1	89	1	1
3	0	1	94	1	1
20	1	1	82	1	1
1	1	1	80	1	0
16	1	1	92	1	1
3	1	0	94	1	1
17	1	0	85	1	1
1	1	1	77	1	0
22	1	0	85	1	0
1	1	0	93	1	1
3	1	1	89	1	1
1	1	1	72	1	1
20	1	0	92	1	1
1	1	0	91	1	1
40	1	1	78	1	1
1	1	0	72	1	0
2	1	1	91	1	1
46	1	1	80	1	1
4	1	0	88	1	0
12	1	1	93	1	1
27	1	1	84	1	1
48	1	0	79	1	1
1	1	1	77	1	0
4	1	0	89	1	1
1	1	1	76	1	0
1	1	1	88	1	1
3	1	0	86	1	1
1	1	1	90	1	1
3	1	1	91	1	0
12	1	1	91	1	1
15	1	0	91	1	1
5	1	1	90	1	1
3	1	0	92	1	1
4	1	1	88	1	1
7	1	1	83	1	1
3	0	0	94	1	1
3	1	0	75	1	1
1	1	1	79	1	0
7	1	0	87	1	1
14	1	0	79	1	1

t	estado	sexo	año	nota	act
12	1	1	90	2	1
1	1	0	90	2	0
21	1	0	84	2	1
2	1	1	85	2	1
8	1	0	84	2	0
12	1	0	86	2	0
9	1	0	86	2	1
3	1	1	88	2	1
3	0	1	94	2	1
1	1	1	77	2	1
7	1	0	93	2	1
1	1	0	74	2	1
1	1	0	81	2	1
12	1	0	92	2	0
3	0	0	94	2	1
10	1	0	86	2	1
3	1	0	89	2	1
1	1	1	76	2	1
1	1	1	89	2	1
1	1	0	88	2	0
1	1	1	82	2	1
2	1	1	74	2	0
1	1	1	77	2	1
32	1	1	89	2	1
9	1	0	87	2	1
3	1	1	89	2	1
6	1	1	82	2	1
14	1	0	94	2	1
2	1	0	89	2	1
1	1	1	86	2	1
24	1	0	81	2	1
5	1	0	86	2	1
15	0	0	93	2	0
4	1	1	91	2	1
1	1	0	90	2	1
1	1	1	86	2	0
3	1	0	92	2	1
14	1	1	87	2	1
2	1	1	89	2	1
7	1	0	92	2	1
6	1	0	87	2	1
3	0	0	94	2	1
3	1	0	92	2	1
3	0	1	93	2	0
12	1	1	78	2	1
3	0	1	94	2	0
15	1	1	79	2	0
36	1	1	78	2	1
5	1	1	87	2	1
2	1	0	76	2	1
1	1	1	86	2	1
12	1	1	93	2	1
1	1	1	81	2	0
2	1	1	94	2	0
3	1	1	81	2	1
1	1	1	72	2	1

(continúa)

t	estado	sexo	ano	nota	act
7	1	0	88	1	1
1	1	1	90	1	1
15	1	0	91	1	1
1	1	1	76	1	0
3	1	0	88	1	1
3	0	0	94	1	1
3	1	0	92	1	1
15	0	0	93	1	0
2	1	1	92	1	0
3	1	1	87	1	1
1	1	1	72	1	1
1	1	0	93	1	1
12	1	1	84	1	1
12	1	1	77	1	0
4	1	1	92	1	1
1	1	1	77	1	0
3	0	0	94	1	1
1	1	1	87	1	1
6	1	1	85	1	1
6	1	1	75	1	1
1	1	0	88	1	1
27	0	0	92	1	1
4	1	1	80	1	0
3	0	1	94	1	0
15	1	0	86	1	0
3	1	1	77	1	0
5	1	1	76	1	1
15	1	0	80	1	1
4	1	1	79	1	1
18	1	1	86	1	1
1	1	0	93	1	1
6	1	1	78	1	1
3	0	1	94	1	1
15	1	0	91	1	1
12	1	1	87	1	1
1	1	1	79	1	1
21	1	0	83	1	1
12	1	0	84	1	1
1	1	1	90	1	0
12	1	0	86	1	0
2	1	0	89	1	1
1	1	1	90	1	1
2	1	0	89	1	1
1	1	0	91	1	1
12	1	1	90	1	1
26	1	0	81	1	0
24	1	1	80	1	1
3	1	1	71	1	1
15	0	0	93	1	1
4	1	1	76	1	1
24	1	1	82	1	1
6	1	1	78	1	1
1	1	1	92	1	1
24	1	1	92	1	1
21	1	0	81	1	0
36	1	1	77	1	1

t	estado	sexo	ano	nota	act
3	0	1	94	2	1
6	1	1	76	2	1
12	1	1	85	2	0
3	1	1	84	2	1
4	1	1	90	2	1
1	1	1	94	2	1
8	1	1	91	2	1
1	1	1	73	2	0
3	1	0	75	2	1
8	1	1	94	2	1
3	1	1	92	2	1
36	1	1	80	2	0
24	1	1	82	2	1
12	1	1	82	2	1
3	1	1	75	2	1
1	1	1	78	2	1
4	1	1	78	2	1
3	1	0	92	2	1
12	1	0	87	2	1
3	1	1	77	2	0
3	1	0	90	2	1
12	1	1	82	2	1
20	1	1	82	2	1
15	1	1	93	2	1
4	1	1	91	2	0
5	1	1	86	2	1
3	1	0	92	2	1
15	1	1	77	2	1
1	1	1	73	2	1
1	1	1	84	2	1
12	1	1	87	2	1
7	1	1	91	2	1
28	1	1	76	2	1
1	1	1	74	2	0
4	1	0	88	2	0
15	1	0	91	2	1
1	1	1	91	2	1
14	1	1	86	2	1
14	1	1	78	2	1
3	1	1	90	2	1
6	1	0	80	2	1
2	1	1	76	2	0
3	0	0	94	2	0
1	1	1	78	2	1
4	1	1	88	2	1
3	1	1	83	2	0
3	1	0	87	2	1
1	1	1	72	2	1
3	1	0	79	2	1
3	1	1	76	2	1
2	1	0	89	2	1
1	1	1	85	2	1
14	1	0	81	2	0
3	0	1	94	2	1
2	1	1	86	2	1
1	1	0	91	2	1

(continúa)

t	estado	sexo	ano	nota	act
6	1	0	92	1	1
24	1	0	92	1	1
1	1	1	80	1	1
2	1	1	84	1	1
2	1	0	92	1	1
12	1	0	91	1	0
2	1	0	86	1	1
3	1	1	91	1	1
11	1	1	92	1	1
15	0	0	93	1	1
21	1	0	92	1	1
3	1	0	91	1	1
3	0	0	94	1	1
6	1	1	86	1	0
24	1	1	76	1	0
24	1	0	85	1	1
3	0	0	94	1	1
7	1	1	80	1	0
10	1	0	80	1	1
3	1	1	89	1	1
3	0	0	94	1	1
7	1	1	85	1	1
1	1	0	72	1	1
4	1	0	89	1	1
15	1	1	76	1	1
22	1	1	77	1	1
28	1	1	83	1	1
1	1	0	90	1	1
14	1	1	84	1	1
3	1	1	79	1	1
24	1	1	83	1	0
3	1	1	72	1	1
15	0	0	93	1	0
3	1	0	92	1	1
15	0	0	93	1	1
15	1	0	81	1	1
3	1	1	92	1	1
1	1	1	77	1	0
1	1	1	76	1	1
15	0	0	93	1	1
4	1	0	89	1	1
24	1	0	85	1	1
1	1	1	78	1	1
4	1	0	91	1	1
6	1	1	77	1	1
27	0	0	92	1	1
3	0	0	94	1	0
12	1	1	92	1	1
12	1	1	88	1	0
27	1	0	92	1	1
4	1	1	84	1	1
12	1	0	87	1	1
3	0	0	94	1	1
15	0	0	93	1	1
1	1	0	87	1	0
2	1	0	94	1	1

t	estado	sexo	ano	nota	act
7	1	1	92	2	1
4	1	1	75	2	1
1	1	0	90	2	1
1	1	1	78	2	1
48	1	0	81	2	1
12	1	1	82	2	0
16	1	1	74	2	1
9	1	1	91	2	1
15	1	1	82	2	1
40	1	1	77	2	1
10	1	1	83	2	1
3	1	1	76	2	1
12	1	1	92	2	0
2	1	0	75	2	0
3	1	1	87	2	1
24	1	1	87	2	0
20	1	1	82	2	1
3	1	1	84	2	1
6	1	1	88	2	1
3	0	1	94	2	1
3	1	0	88	2	1
38	1	1	75	2	0
1	1	1	74	2	1
1	1	1	88	2	0
3	1	0	87	2	1
4	1	0	87	2	0
3	1	1	78	2	1
12	1	1	76	2	1
30	1	0	92	2	1
2	1	1	79	2	1
3	1	0	90	2	0
3	1	1	88	2	1
3	1	1	94	2	1
5	1	1	80	2	1
24	1	1	90	2	1
24	1	1	85	2	1
4	1	1	85	2	1
5	1	1	85	2	1
3	1	1	90	2	1
1	1	0	76	2	1
23	1	1	84	2	1
1	1	0	85	2	1
12	1	0	93	2	1
14	1	1	85	2	1
36	1	1	79	2	0
15	1	1	77	2	0
3	1	1	85	2	1
1	1	0	75	2	1
14	1	1	88	2	0
1	1	1	75	2	1
3	1	1	76	2	0
4	1	1	79	2	1
12	1	0	80	2	1
2	1	1	92	2	1
12	1	0	85	2	1
1	1	1	72	2	1

(continúa)

t	estado	sexo	ano	nota	act
1	1	0	88	1	1
12	1	1	93	1	1
4	1	1	89	1	1
1	1	1	90	1	1
20	1	1	82	1	1
36	1	1	79	1	1
15	0	1	93	1	1
3	1	0	87	1	1
5	1	0	82	1	0
39	0	0	91	1	1
1	1	1	72	1	1
10	1	1	85	1	1
12	1	1	75	1	1
1	1	1	72	1	1
2	1	1	76	1	1
12	1	1	85	1	1
24	1	1	82	1	1
16	1	1	82	1	0
24	1	0	91	1	1
6	1	1	80	1	1
8	1	0	92	1	1
5	1	1	88	1	1
6	1	1	88	1	1
4	1	1	77	1	1
1	1	1	93	1	1
2	1	1	90	1	1
2	1	0	88	1	0
12	1	1	83	1	1
3	1	1	91	1	1
12	1	0	85	1	0
1	1	1	79	1	0
2	1	1	77	1	0
1	1	0	75	1	1
3	1	1	87	1	1
12	1	1	83	1	1
3	0	0	94	1	1
1	1	1	75	1	1
12	1	1	87	1	1
3	0	1	94	1	1
2	1	1	87	1	0
1	1	1	89	1	1
12	1	1	86	1	1
1	1	0	83	1	1
9	1	1	94	1	0
1	1	1	75	1	1
1	1	0	73	1	1
6	1	1	87	1	1
15	0	0	93	1	0
7	1	1	94	1	1
18	1	0	85	1	1
48	1	1	80	1	1
12	1	1	93	1	1
3	1	0	79	1	0
1	1	1	85	1	1
14	1	1	81	1	1
5	1	1	90	1	1

t	estado	sexo	ano	nota	act
24	1	1	80	2	0
1	1	0	80	2	1
14	1	1	75	2	0
12	1	1	92	2	1
1	1	1	77	2	1
12	1	1	80	2	1
3	0	0	94	2	1
12	1	0	92	2	1
4	1	1	88	2	1
3	1	0	77	2	1
3	1	1	92	2	1
2	1	1	89	2	0
17	1	1	84	2	1
12	1	0	84	2	1
2	1	0	74	2	1
2	1	0	89	2	1
6	1	0	81	2	0
3	0	1	94	2	1
3	1	0	73	2	1
1	1	0	93	2	1
10	1	1	89	2	1
24	1	1	89	2	1
1	1	0	92	2	1
1	1	1	71	2	1
4	1	1	91	2	1
1	1	0	88	2	1
1	1	1	74	2	0
2	1	1	76	2	1
3	1	1	90	2	1
4	1	1	79	2	1
1	1	0	87	2	1
3	0	1	94	2	1
6	1	1	75	2	1
3	1	1	78	2	1
3	0	0	94	2	1
5	1	0	93	2	1
15	1	0	86	2	0
3	1	0	90	2	1
27	1	1	82	2	1
4	1	1	87	2	0
36	1	1	81	2	1
4	1	1	87	2	1
10	1	0	85	2	1
3	1	1	90	2	1
24	1	1	76	2	1
2	1	0	76	2	1
3	1	0	73	2	1
2	1	0	75	2	1
3	1	1	78	2	0
3	1	0	89	2	1
18	1	1	85	2	1
3	0	1	94	2	1
1	1	1	91	2	1
3	1	1	78	2	1
11	1	1	78	2	1
3	1	0	90	2	1

(continúa)

t	estado	sexo	ano	nota	act
36	1	1	79	1	1
96	1	0	82	1	1
1	1	0	86	1	0
14	1	1	91	1	1
2	1	0	90	1	1
12	1	1	81	1	1
4	1	0	77	1	1
1	1	1	73	1	0
1	1	1	74	1	1
36	1	1	89	1	1
24	1	1	82	1	1
14	1	1	84	1	1
36	1	0	87	1	0
4	1	0	89	1	1
14	1	0	83	1	1
15	1	1	86	1	1
4	1	0	79	1	1
1	1	1	76	1	1
3	1	0	89	1	1
1	1	0	85	1	1
4	1	0	87	1	1
12	1	1	87	1	1
1	1	0	87	1	1
1	1	0	87	1	1
6	1	1	92	1	1
13	1	1	80	1	1
26	1	1	90	1	1
24	1	0	80	1	0
1	1	0	87	1	0
3	0	0	94	1	1
8	1	0	91	1	1
4	1	1	90	1	1
24	1	1	82	1	1
1	1	1	90	1	1
12	1	0	84	1	1
3	1	1	82	1	1
2	1	0	93	1	0
3	1	1	92	1	1
1	1	1	71	1	1
1	1	0	91	1	1
1	1	1	91	1	1
3	0	1	94	1	1
2	1	0	91	1	1
3	1	0	88	1	1
15	1	1	85	1	1
4	1	1	85	1	1
24	1	1	79	1	1
24	1	0	76	1	1
10	1	0	79	1	1
12	1	0	86	1	0
1	1	0	90	1	0
36	1	0	80	1	1
1	1	1	75	1	1
5	1	1	74	1	1
3	1	0	89	1	1
1	1	0	89	3	1

t	estado	sexo	ano	nota	act
1	1	1	87	2	1
2	1	1	80	2	1
3	1	1	74	2	1
12	1	1	77	2	1
3	1	0	81	2	1
4	1	1	84	2	1
1	1	1	77	2	1
3	1	1	88	2	1
12	1	1	81	2	1
15	1	1	85	2	0
3	1	1	91	2	1
1	1	0	78	2	1
12	0	0	94	2	1
7	1	0	91	2	1
3	1	0	88	2	0
3	1	1	76	2	1
1	1	1	82	2	1
9	1	1	91	2	1
9	1	1	80	2	1
3	0	0	94	2	1
1	1	1	75	3	0
1	1	1	76	3	1
1	1	1	73	3	1
2	1	1	92	3	1
3	1	0	71	3	1
1	1	0	81	3	1
1	1	0	90	3	1
1	1	0	92	3	1
1	1	1	73	3	1
2	1	0	84	3	1
2	1	1	77	3	1
17	1	1	92	3	1
2	1	1	71	3	1
6	1	1	85	3	1
1	1	1	76	3	1
4	1	1	93	3	1
3	1	0	85	3	1
3	1	0	85	3	1
6	1	1	88	3	1
18	1	1	85	3	1
6	1	0	93	3	1
5	1	0	78	3	1
2	1	0	87	3	1
7	1	1	91	3	1
2	1	1	89	3	1
1	1	1	77	3	0
3	0	1	94	3	1
2	1	1	73	3	1
12	1	1	81	3	0
8	1	1	94	3	1
2	1	1	88	3	0
1	1	1	84	3	1
3	1	1	75	3	1
4	1	1	74	3	1
3	1	1	78	3	1

Índice de Figuras

1.1 *Función de azar Gamma.* 20

2.1 *Funciones de azar poligonales.* 43

2.2 *Parte paramétrica de la función de azar del modelo Gamma-poligonal aditivo para un individuo con parámetros* $\alpha = 665$ *y* $\beta = 20$. 44

2.3 *Funcion de supervivencia del modelo Gamma-poligonal aditivo para un individuo con parámetros* $\alpha = 665$ *y* $\beta = 20$. 44

3.1 *Función de supervivencia del modelo correcto y su predicción sin utilizar covariables. Banco de datos simulados 1.* 74

3.2 *Evolución de algunos parámetros de la MCMC con todas las covariables en el estudio. Banco de datos simulados 1.* 75

3.3 *Intervalos intercuantílicos, obtenidos con CODA, de algunos parámetros de la MCMC considerando todas las covariables en el estudio. Banco de datos simulados 1.* 76

3.4 *Análisis de la convergencia, obtenida con CODA, de algunos parámetros de la MCMC considerando todas las covariables en el estudio. Banco de datos simulados 1.* 77

3.5 *Evolución de algunos parámetros de la MCMC obtenida con CODA, considerando únicamente x_1 y x_3 en el estudio. Banco de datos simulados 1.* 79

3.6 *Análisis de la convergencia de algunos parámetros de la MCMC obtenido con CODA, considerando únicamente x_1 y x_3 en el estudio. Banco de datos simulados 1.* 80

3.7 *Intervalos intercuantílicos, obtenidos con CODA, de algunos parámetros de la MCMC considerando únicamente x_1 y x_3 en el estudio. Banco de datos simulados 1.* 81

3.8 *Función de supervivencia del modelo correcto y su predicción para algunos individuos. Banco de datos simulados 1.* 82

3.9 *Evolución y densidad predictiva marginal, obtenidas con CODA, de algunos parámetros de la MCMC. Banco de datos simulados 2.* 85

3.10 *Intervalos intercuantílicos, obtenidos con CODA, de algunos parámetros de la MCMC. Banco de datos simulados 2.* 86

3.11 *Análisis de la convergencia de algunos parámetros de la MCMC obtenido con CODA. Método de Geweke. Banco de datos simulados 2.* 87

3.12 *Análisis de la convergencia de algunos parámetros de la MCMC obtenido con CODA. Método de Raftery y Lewis. Banco de datos simulados 2.* 88

3.13 *Función de densidad del modelo correcto y su predicción para algunos individuos. Banco de datos simulados 2.* 89

3.14 *Aproximación al azar poligonal. Banco de datos de Stanford.* . . 92

3.15 *Evolución y densidad predictiva marginal, obtenidas con CODA, de algunos parámetros de la MCMC. Banco de datos de Stanford.* 93

3.16 *Análisis de la convergencia de algunos parámetros de la MCMC obtenido con CODA. Banco de datos de Stanford.* 94

3.17 *Intervalos intercuantílicos, obtenidos con CODA, de algunos parámetros de la MCMC. Banco de datos de Stanford.* 95

3.18 *Función de supervivencia predictiva para algunos individuos. Banco de datos de Stanford.* 96

3.19 *Función de supervivencia predictiva de Cox para algunos individuos. Banco de datos de Stanford.* 96

3.20 *Aproximación al azar poligonal de cada grupo de individuos. Banco de datos de la encuesta.* 98

3.21 *Evolución y densidad predictiva marginal, obtenidas con CODA, de algunos parámetros de la MCMC asociados a las covariables. Banco de datos de la encuesta.* 100

3.22 *Evolución y densidad predictiva marginal, obtenidas con CODA, de algunos parámetros de la MCMC relativos al azar poligonal. Banco de datos de la encuesta.* 101

3.23 *Intervalos intercuantílicos, obtenidos con CODA, de algunos parámetros de la MCMC asociados a las covariables. Banco de datos de la encuesta.* 102

3.24 *Análisis de la convergencia, obtenida con CODA, de algunos parámetros de la MCMC asociados a las covariables. Banco de datos de la encuesta.* 103

3.25 *Función de supervivencia predictiva para algunos individuos. Banco de datos de la encuesta.* 104

3.26 *Función de supervivencia predictiva de Cox para algunos individuos con año de licenciatura 1998. Banco de datos de la encuesta.* 105

3.27 *Función de supervivencia predictiva de Cox para algunos individuos con nota media de licenciatura aprobado. Banco de datos de la encuesta.* 105

Índice de Tablas

3.1 *Media y desviación típica del modelo correcto y de la densidad predictiva sin utilizar covariables. Banco de datos simulados 1.* . 73

3.2 *Media y desviación típica del modelo correcto y de la densidad predictiva para algunos individuos. Banco de datos simulados 1.* . 81

3.3 *Media y desviación típica del modelo correcto y de la densidad predictiva para algunos individuos. Banco de datos simulados 2.* . 90

3.4 *Media y desviación típica de la densidad predictiva para algunos individuos. Banco de datos de la encuesta.* 102

Bibliografía

Andersen, P. K., Borgan, O., Gill, R. D. y Keiding, N. (1993). *Statistical models based on counting processes*. New York: Springer-Verlag.

Arjas, E. y Gasbarra, D. (1996). Bayesian inference of survival probabilities, under stochastic ordering constraints. *Journal of the American Statistical Association*, **91**, pp. 1101–1109.

Barlow, R. E. y Proschan, F. (1996). *Mathematical theory of reliability*. Philadelphia: S.I.A.M.

Beamonte, E. y Bermúdez, J. D. (1993). Comparación de curvas de supervivencia Gamma. *Qüestiió*, **19**, pp. 171–186.

Becker, R. A., Chambers, J. M. y Wilks, A. R. (1988). *The new S language*. Pacific Grove: Wadsworth & Brooks/Cole Advanced Books & Software.

Beenstock, M. (1996). Training and the time to find a job in Israel. *Applied Economics*, **28**, pp. 935–946.

Berkson, J. y Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**, pp. 501–515.

Berliner, L. M. y Hill, B. M. (1988). Bayesian nonparametric survival analysis (with comments). *Journal of the American Statistical Association*, **83**, pp. 772–784.

Bermúdez, J. D. y Beamonte, E. (1995). Análisis bayesiano de datos de supervivencia Gamma utilizando muestreo de Gibbs. *Estadística Española*, **35**, pp. 629–644.

Bermúdez, J. D. y Beamonte, E. (1997). Comparación de curvas de supervivencia Gamma estocásticamente ordenadas. *Technical report*, Departamento de Estadística e Investigación Operativa. Universitat de València.

Best, N. G., Cowles, M. K. y Vines, S. K. (1995). *CODA manual version 0.30*. Cambridge: MRC Biostatistics Unit.

Best, N. G., Spiegelhalter, D. J., Thomas, A. y Brayne, C. E. G. (1996). Bayesian analysis of realistically complex models. *Journal of the Royal Statistical Society A*, **159**, pp. 323–342.

Bhattacharjee, G. P. (1970). Algorithm AS32. The incomplete Gamma integral. *Applied Statistics*, **19**, pp. 285–287.

Box, G. E. P. y Jenkins, G. M. (1976). *Time series analysis: forecasting and control*. San Francisco: Holden-Day.

Box, G. E. P. y Muller, M. E. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, **29**, pp. 610–611.

Broemeling, L. D. (1985). *Bayesian analysis of linear models*. New York: Marcel Dekker.

Brooks, S. P. y Roberts, G. O. (1997). Assessing convergence of Markov chain Monte Carlo algorithms. *Technical report*, School of Mathematics. University of Bristol.

Buckley, J. y James, I. (1979). Linear regression with censored data. *Biometrika*, **66**, pp. 429–436.

Burridge, J. (1981). Empirical Bayes analysis of survival time data. *Journal of the Royal Statistical Society B*, **43**, pp. 65–75.

Casella, G. y George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, **46**, pp. 167–174.

Chan, K. S. (1993). Asymptotic behaviour of the Gibbs sampler. *Journal of the American Statistical Association*, **88**, pp. 320–328.

Chen, W. C., Hill, B. M., Greenhouse, J. B. y Fayos, J. V. (1985). Bayesian analysis of survival curves for cancer patients following treatment (with discussion). En *Bayesian Statistics 2 (Bernardo, J.M., DeGroot, M.H., Lindley, D.V. y Smith, A.F.M. eds.)*, pp. 299–328. Amsterdam: Elsevier Science Publishers B.V.

Chib, S. y Greenberg, E. (1996). Markov chain Monte Carlo simulation methods in Econometrics. *Econometric Theory*, **12**, pp. 409–431.

Christensen, R. y Johnson, W. (1988). Modelling accelerated failure time with a Dirichlet process. *Biometrika*, **75**, pp. 693–704.

Collett, D. (1994). *Modelling survival data in medical research.* London: Chapman & Hall.

Cowles, M. K. y Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, **91**, pp. 883–904.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B*, **34**, pp. 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, pp. 269–276.

Cox, D. R. y Oakes, D. A. (1984). *Analysis of survival data*. London: Chapman & Hall.

De Gruttola, V. y Lagakos, S. W. (1989). Analysis of doubly censored survival data, with application to AIDS. *Biometrics*, **45**, pp. 1–11.

DeGroot, M. H. (1970). *Optimal statistical decisions*. New York: McGraw-Hill.

Dellaportas, P. y Smith, A. F. M. (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Applied Statistics*, **42**, pp. 443–460.

Devroye, L. (1986). *Non-uniform random variate generation*. New York: Springer-Verlag.

Draper, D. (1997). *Bayesian hierarchical modeling (draft 2)*. Pendiente de publicación.

Fahrmeir, L. (1994). Dinamic modelling and penalized likelihood estimation for discrete time survival data. *Biometrika*, **81**, pp. 317–330.

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

Ferguson, T. S. y Phadia, E. G. (1979). Bayesian nonparametric estimation based on censored data. *Annals of Statistics*, **7**, pp. 163–186.

Follmann, D. A., Goldberg, M. S. y May, L. (1990). Personal characteristics, unemployment insurance, and the duration of unemployment. *Journal of Econometrics*, **45**, pp. 351–366.

Gamerman, D. (1991). Dynamic bayesian models for survival data. *Applied Statistics*, **40**, pp. 63–79.

Gamerman, D. (1997). *Markov chain Monte Carlo. Stochastic simulation for bayesian inference*. London: Chapman & Hall.

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

Gelfand, A. E. y Mallick, B. K. (1995). Bayesian analysis of proportional-hazards models built from monotone functions. *Biometrics*, **51**, pp. 843–852.

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

Gelman, A. y Rubin, D. B. (1992a). A single series from the Gibbs sampler provides a false sense of security. En *Bayesian Statistics 4 (Bernardo, J.M., Berger, J.O., Dawid, A.P. y Smith, A.F.M. eds.)*, pp. 625–631. Oxford: Oxford University Press.

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

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

Geweke, J. (1989). Bayesian inference in econometrics models using Monte Carlo integration. *Econometrica*, **57**, pp. 1317–1339.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (with discussion). En *Bayesian Statistics 4 (Bernardo, J.M., Berger, J.O., Dawid, A.P. y Smith, A.F.M. eds.)*, pp. 169–193. Oxford: Oxford University Press.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science*, **7**, pp. 473–511.

Geyer, C. J. y Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association*, **90**, pp. 909–920.

Gilks, W. R., Best, N. G. y Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling with Gibbs sampling. *Applied Statistics*, **44**, pp. 455–472.

Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D. y Kirby, A. J. (1993). Modelling complexity: applications of Gibbs sampling in medicine. *Journal of the Royal Statistical Society B*, **55**, pp. 39–52.

Gilks, W. R., Richardson, S. y Spiegelhalter, D. J., eds. (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Gómez, G. y Lagakos, S. (1994). Estimation of the infection time and latency distribution of AIDS with doubly censored data. *Biometrics*, **50**, pp. 204–212.

Gross, A. J. y Clark, V. A. (1975). *Survival distributions: reliability applications in the biomedical sciences*. New York: Willey.

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

Hjort, N. L. (1990). Nonparametric bayesian estimators based on Beta processes in models for life testing data. *Annals of Statistics*, **18**, pp. 1259–1294.

Hu, X. J., Lawless, J. F. y Suzuki, K. (1998). Nonparametric estimation of a lifetime distribution when censoring times are missing. *Technometrics*, **40**, pp. 3–13.

Jaggia, S. y Thosar, S. (1995). Contested tender offers: an estimate of the hazard function. *Journal of Business & Economic Statistics*, **13**, pp. 113–119.

Johnson, R. C. E. (1980). *Survival models and data analysis*. New York: John Willey & Sons.

Kalbfleisch, J. D. (1978). Nonparametric bayesian analysis of survival time data. *Journal of the Royal Statistical Society B*, **40**, pp. 214–221.

Kalbfleisch, J. D. y Prentice, R. L. (1980). *The statistical analysis of failure time data*. New York: Willey.

Kaplan, E. L. y Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, pp. 457–481.

Kim, M. Y., De Gruttola, V. G. y Lagakos, S. W. (1993). Analyzing doubly censored data with covariates. *Biometrics*, **49**, pp. 13–22.

Klein, J. P. y Moeschberger, M. L. (1997). *Survival analysis: techniques for censored and truncated data*. New York: Springer-Verlag.

Knuth, D. E. (1981). *The art of computer programming. Volume 2: seminumerical algorithms*. Reading: Addison Wesley.

Kouassi, D. A. y Singh, J. (1997). A semiparametric approach to hazard estimation with randomly censored observations. *Journal of the American Statistical Association*, **92**, pp. 1351–1355.

Koul, H., Susarla, V. y Van Ryzin, J. (1981). Regression analysis with randomly right censored data. *Annals of Statistics*, **9**, pp. 1276–1288.

Le, C. T. (1997). *Applied survival analysis*. New York: John Willey & Sons.

Lee, E. T. (1992). *Statistical methods for survival data analysis*. New York: John Willey & Sons.

Leonard, T. (1997). Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society B*, **40**, pp. 113–146.

Lindsey, J. K. (1997). Parametric multiplicative intensities models fitted to bus motor failure data. *Applied Statistics*, **46**, pp. 245–252.

Liu, J., Wong, W. H. y Kong, A. (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans. *Journal of the Royal Statistical Society B*, **57**, pp. 157–169.

Mallick, B. K. y Gelfand, A. E. (1994). Generalized linear models with unknown link functions. *Biometrika*, **81**, pp. 237–245.

Martz, H. F. y Waller, R. A. (1982). *Bayesian reliability analysis*. New York: John Willey & Sons.

Meeker, W. Q. y LuValle, M. J. (1995). An accelerated life test model based on reliability kinetics. *Technometrics*, **37**, pp. 133–146.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. y Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, pp. 1087–1092.

Meyn, S. P. y Tweedie, R. L. (1993). *Markov chains and stochastic stability*. New York: Springer-Verlag.

Miller, R. (1981). *Survival analysis*. New York: Wiley.

Miller, R. y Halpern, J. (1982). Regression with censored data. *Biometrika*, **69**, pp. 521–531.

Miller, R. G. (1976). Least squares regression with censored data. *Biometrika*, **63**, pp. 449–464.

Morales, D., Pardo, L. y Quesada, V. (1991). Bayesian survival estimation for incomplete data when the life distribution is proportionally related to the censoring time distribution. *Communications in Statistics-Theory and Methods*, **20**, pp. 831–850.

Mukhopadhyay, S. y Sinha, D. (1995). Bayesian analysis of interval-censored data using random mixture process. *Technical report*, Department of Statistics. University of Connecticut.

Muller, P. (1991). Metropolis based posterior integration schemes. *Technical report*, Department of Statistics. University of Purdue.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, **84**, pp. 487–493.

Oakes, D. (1992). Frailty models for multiple event times. En *Survival analysis: state of the art (Klein, J.P. y Goel, K. eds.)*, pp. 371–379. Dordrecht: Kluwer Academic Publishers.

Oakes, D. (1994). Use of frailty models for multivariate survival data. En *Proceedings of the XVIIth International Biometrics Conference.* Hamilton, Ontario.

Parmar, M. K. B. y Machin, D. (1995). *Survival analysis: a practical approach.* Madrid: Paraninfo.

Parzen, E. (1972). *Procesos estocásticos.* Chichester: John Willey & Sons.

Pike, M. C. y Hill, I. D. (1966). Algorithm 291. Logarithm of the Gamma function. *Communications of the ACM*, **9**, pp. 684.

Pradel, R., Hines, J. E., Lebreton, J. D. y Nichols, J. D. (1997). Capture-recapture survival models taking account of transients. *Biometrics*, **53**, pp. 60–72.

Quantin, C., Moreau, T., Asselain, B., Maccario, J. y Lellouch, J. (1996). A regression survival model for testing the proportional hazards hipothesis. *Biometrics*, **52**, pp. 874–885.

Raftery, A. E. y Lewis, S. (1992). How many iterations in the Gibbs sampler? En *Bayesian Statistics 4 (Bernardo, J.M., Berger, J.O., Dawid, A.P. y Smith, A.F.M. eds.)*, pp. 763–773. Oxford: Oxford University Press.

Resnick, S. I. (1992). *Adventures in stochastic processes.* Boston: Birkhäuser.

Ripley, B. D. (1987). *Stochastic simulation.* New York: John Willey & Sons.

Roberts, G. O. y Polson, N. G. (1994). On the geometric convergence of the Gibbs sampler. *Journal of the Royal Statistical Society B*, **56**, pp. 377–384.

Roberts, G. O. y Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameretization for the Gibbs sampler. *Journal of the Royal Statistical Society B*, **59**, pp. 291–317.

Rubin, D. B. (1988). Using the S.I.R. algorithm to simulate posterior distributions (with discussion). En *Bayesian Statistics 3 (Bernardo, J.M., DeGroot, M.H., Lindley, D.V. y Smith, A.F.M. eds.)*, pp. 395–402. Oxford: Oxford University Press.

Sinha, D. y Dey, D. K. (1997). Semiparametric bayesian analysis of survival data. *Journal of the American Statistical Association*, **92**, pp. 1195–1212.

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

Smith, A. F. M. y Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society B*, **55**, pp. 3–23.

Spiegelhalter, D. J., Thomas, A., Best, N. G. y Gilks, W. R. (1995). *BUGS: Bayesian inference using Gibbs sampling, version 0.50*. Cambridge: MRC Biostatistics Unit.

Susarla, V. y Van Ryzin, J. (1976). Nonparametric estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, **71**, pp. 897–202.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, **22**, pp. 1701–1762.

Tsai, W., Goedert, J. J., Orazem, J., Landesman, S. H., Rubinstein, A., Willoughby, A. y Gail, M. H. (1994). A nonparametric analysis of

the transmission rate of human immunodeficience virus from mother to infant. *Biometrics*, **50**, pp. 1015–1028.

Van der Laan, M. J., Jewell, N. P. y Peterson, D. R. (1997). Efficient estimation of the lifetime and dissease onset distribution. *Biometrika*, **84**, pp. 539–554.

Vaupel, J. W., Manton, K. G. y Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dinamics of mortality. *Demography*, **16**, pp. 439–454.

Wang, C. Y., Hsu, L., Feng, Z. D. y Prentice, R. L. (1997). Regression calibration in failure time regression. *Biometrics*, **53**, pp. 131–145.

Wei, L. J., Lin, D. Y. y Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. *Journal of the American Statistical Association*, **84**, pp. 1065–1073.

Whitehead, J. (1980). Fitting Cox's regression model to survival data using GLIM. *Applied Statistics*, **29**, pp. 268–275.

Yue, H. y Chan, K. S. (1997). A dinamic frailty model for multivariate survival data. *Biometrics*, **53**, pp. 785–793.

Zellner, A. y Rossi, P. (1984). Bayesian analysis of dichotomous quantal response models. *Journal of Econometrics*, **25**, pp. 365–393.