

# COMPARACIÓN DE CURVAS DE SUPERVIVENCIA GAMMA

EDUARDO BEAMONTE\* y JOSÉ D. BERMÚDEZ‡

*A Gamma Hierarchical Model is used for the Bayesian Analysis of Survival Data with covariates. We center our interest on the comparison of two groups of individuals using Monte-Carlo methods. Concretely, we use the Gibbs sampling to obtain a sample from the posterior distribution. The paper ends with the analysis of two data sets, one of them simulated and the other real.*

## Comparison of Gamma Survival Curves.

**Key words:** Censored data, Covariates, Gamma Hierarchical Model, Markov-Chain Monte-Carlo Methods, Prediction.

## 1. INTRODUCCIÓN

En un trabajo anterior (Bermúdez y Beamonte, 1993) propusimos un modelo Gamma con parámetros comunes a todos los individuos, para datos de supervivencia que podían estar progresivamente censurados por la derecha. Esto es, las observaciones realizadas en algunos individuos no muestran su tiempo total de supervivencia, tan sólo se sabe que ese tiempo es mayor que el observado, conocido como tiempo de censura, con un mecanismo de censura independiente del mecanismo de muerte.

---

\*Eduardo Beamonte. Departamento de Economía Aplicada. Universitat de València. Avda. Blasco Ibáñez, 30. 46010 València.

‡José D. Bermúdez. Departamento de Estadística e Investigación Operativa. Universitat de València. C/ Dr. Moliner, 50. 46100 Burjasot. València.

Trabajo parcialmente subvencionado por la Generalitat Valenciana con cargo al proyecto GV-1081/93.

—Article rebut l'agost de 1994.

—Acceptat l'abril de 1995.

Analizábamos el modelo desde una perspectiva bayesiana, estudiando la distribución final mediante una muestra obtenida a partir de cadenas de Markov generadas por Monte-Carlo.

En este trabajo generalizamos el modelo anterior para disminuir su rigidez y poder incorporar covariables. Así, suponemos que el tiempo de supervivencia de cada individuo no tiene una distribución común a todos ellos, sino que esa distribución depende de ciertas características del individuo. Para ello proponemos el siguiente modelo jerárquico:

$$t \sim \text{Ga}(t|\alpha, \beta)$$

$$(\log \alpha, \log \beta)' \sim \text{NM}_2((\log \alpha, \log \beta)'|Bx, H),$$

siendo  $t$  el tiempo de supervivencia y  $x$  el vector de covariables de dicho individuo, que habitualmente incorporará un primer elemento constante e igual a 1, como toda matriz de diseño en los modelos lineales generalizados. Es decir, cada tiempo de supervivencia es  $\text{Ga}(\alpha, \beta)$ , pero los parámetros  $\alpha$  y  $\beta$  son características propias no observables del individuo, que están relacionados con su vector de covariables a través de un mecanismo aleatorio que, a su vez, depende de ciertos hiperparámetros  $B$  y  $H$  comunes a todos los individuos. Así,  $\log \alpha$  y  $\log \beta$  siguen una distribución Normal bivalente con media  $Bx$  y matriz de precisión  $H$ .

En este trabajo estamos especialmente interesados en la comparación de dos grupos de individuos. Para ello basta con incluir una única covariable dicotómica, aparte de la covariable constante; así,  $x' = (1, 1)$  si el individuo pertenece al primer grupo y  $x' = (1, 0)$  en otro caso. El objetivo primordial será hacer inferencias sobre los hiperparámetros relacionados con esa covariable dicotómica.

La notable complicación de la distribución final (independientemente de la inicial utilizada y acentuada por la presencia habitual de datos censurados) estimula a realizar su estudio a través de una muestra generada a partir de la misma. En esa línea fue pionero el trabajo de Gelfand y Smith (1990).

Uno de los métodos iterativos para la obtención de una muestra de la distribución final más utilizados en los últimos tiempos es el muestreo de Gibbs (ver, por ejemplo, Smith y Roberts, 1993, y referencias allí citadas). Básicamente, consiste en construir una función de transición que defina una cadena de Markov irreducible y para la que la distribución final sea estacionaria. Este proceso exige que sea fácil muestrear a partir de las distribuciones condicionales completas, pero eso suele ser trivial si el vector paramétrico se descompone de forma que tales distribuciones sean univariantes.

En el apartado 2 se analiza el modelo propuesto, detallando la aplicación del muestreo de Gibbs en el estudio de la distribución final. Este procedimiento se utiliza en el apartado 3 para el análisis de dos bancos de datos que incluyen una covariable

dicotómica; uno de ellos simulados y el otro basado en datos reales. Por último, el apartado 4 recoge los comentarios y conclusiones finales.

## 2. ANÁLISIS DEL MODELO JERÁRQUICO GAMMA

Denotamos por  $\{t_1, \dots, t_r\}$  a los tiempos de supervivencia correspondientes a los datos no censurados y por  $\{T_{r+1}, \dots, T_n\}$  a los tiempos de censura correspondientes a los datos censurados. El vector paramétrico completo objeto del muestreo de Gibbs es el formado por los parámetros del modelo,  $(\alpha_1, \beta_1, \dots, \alpha_n, \beta_n)$ , los hiperparámetros,  $(B$  y  $H)$ , y los tiempos de supervivencia no observados,  $(t_{r+1}, \dots, t_n)$ , sujetos a las restricciones  $t_i > T_i$ ,  $i = r+1, \dots, n$ . La distribución final viene dada por:

$$f(\alpha_1, \beta_1, \dots, \alpha_n, \beta_n, B, H, t_{r+1}, \dots, t_n | t_1, \dots, t_r, T_{r+1}, \dots, T_n, X),$$

siendo  $X_{n \times k}$  la matriz de covariables; esto es, una matriz cuya fila  $i$ -ésima coincide con el vector de covariables del  $i$ -ésimo individuo.

A partir de esa distribución podemos obtener la distribución marginal de interés:

$$f(B, H | t_1, \dots, t_r, T_{r+1}, \dots, T_n, X).$$

Para utilizar el muestreo de Gibbs es necesario obtener las distribuciones condicionales completas. La correspondiente a un tiempo censurado no observado es:

$$(1) \quad \begin{aligned} f(t_i | t_1, \dots, t_{i-1}, t_{i+1}, \dots, t_n, T_{r+1}, \dots, T_n, X, \alpha_1, \beta_1, \dots, \alpha_n, \beta_n, B, H) = \\ = f(t_i | T_i, \alpha_i, \beta_i), \quad i = r+1, \dots, n, \end{aligned}$$

siendo  $f(t_i | T_i, \alpha_i, \beta_i)$  una distribución Gamma truncada. Específicamente, Bermúdez y Beamonte (1993),  $f(t_i | T_i, \alpha_i, \beta_i) \propto \text{Ga}(t_i | \alpha_i, \beta_i)$  si  $t_i > T_i$ .

La distribución condicional completa de los hiperparámetros es:

$$(2) \quad \begin{aligned} f(B, H | t_1, \dots, t_n, T_{r+1}, \dots, T_n, X, \alpha_1, \beta_1, \dots, \alpha_n, \beta_n) = f(B, H | X, \alpha_1, \beta_1, \dots, \alpha_n, \beta_n) \propto \\ \propto f(B, H) f(\alpha_1, \beta_1, \dots, \alpha_n, \beta_n | X, B, H), \end{aligned}$$

que se reduce a un problema habitual de regresión bivalente normal homocedástica: si se utiliza una distribución inicial  $f(B, H)$  perteneciente a la familia Normal-Wishart, la distribución final también es Normal-Wishart cuya expresión puede consultarse, por ejemplo, en Broemeling (1985, capítulo 8, pp. 378-379).

La distribución condicional completa del parámetro  $\alpha_i$ ,  $i = 1, \dots, n$ , es:

$$\begin{aligned}
 f(\alpha_i | \alpha_1, \dots, \alpha_{i-1}, \alpha_{i+1}, \dots, \alpha_n, \beta_1, \dots, \beta_n, t_1, \dots, t_n, T_{r+1}, \dots, T_n, \mathbf{X}, \mathbf{B}, \mathbf{H}) &= \\
 &= f(\alpha_i | t_i, \beta_i, x_i, \mathbf{B}, \mathbf{H}) \propto f(t_i | \alpha_i, \beta_i) f(\alpha_i | \beta_i, x_i, \mathbf{B}, \mathbf{H}) = \\
 (3) \quad &= \text{Ga}(t_i | \alpha_i, \beta_i) \text{NM}(\log \alpha_i | \beta_i, x_i, \mathbf{B}, \mathbf{H}) \propto \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} t_i^{\alpha_i} \text{NM}(\log \alpha_i | \mu'_1, \sigma_1'^2),
 \end{aligned}$$

con  $\mu'_1 = \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (\ln \beta_i - \mu_2)$  y  $\sigma_1'^2 = \frac{1}{h_{11}}$ , los momentos de la distribución Normal condicionada obtenida a partir de la Normal bivalente con vector de medias  $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \mathbf{B}x_i$  y matriz de precisión  $\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} \\ h_{12} & h_{22} \end{bmatrix} = \Sigma^{-1}$ , con  $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$  y  $\sigma_{12} = \rho \sigma_1 \sigma_2$ .

Y, finalmente, la distribución completa del parámetro  $\beta_i$ ,  $i = 1, \dots, n$ , es:

$$\begin{aligned}
 f(\beta_i | \beta_1, \dots, \beta_{i-1}, \beta_{i+1}, \dots, \beta_n, \alpha_1, \dots, \alpha_n, t_1, \dots, t_n, T_{r+1}, \dots, T_n, \mathbf{X}, \mathbf{B}, \mathbf{H}) &= \\
 &= f(\beta_i | t_i, \alpha_i, x_i, \mathbf{B}, \mathbf{H}) \propto f(t_i | \alpha_i, \beta_i) f(\beta_i | \alpha_i, x_i, \mathbf{B}, \mathbf{H}) = \\
 (4) \quad &= \text{Ga}(t_i | \alpha_i, \beta_i) \text{NM}(\log \beta_i | \alpha_i, x_i, \mathbf{B}, \mathbf{H}) \propto \beta_i^{\alpha_i} \exp(-\beta_i t_i) \text{NM}(\log \beta_i | \mu'_2, \sigma_2'^2),
 \end{aligned}$$

siendo  $\mu'_2 = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (\ln \alpha_i - \mu_1)$  y  $\sigma_2'^2 = \frac{1}{h_{22}}$ , la media y varianza de la distribución Normal condicionada correspondiente.

Una vez construidas las distribuciones condicionales completas, y partiendo de un punto inicial  $\alpha_1^{(0)}, \beta_1^{(0)}, \dots, \alpha_n^{(0)}, \beta_n^{(0)}, \mathbf{B}^{(0)}, \mathbf{H}^{(0)}$ , el algoritmo de Gibbs puede implementarse de la siguiente forma:

Completamos los tiempos censurados,  $T_{r+1}, \dots, T_n$ , generando los tiempos de supervivencia no observados  $(t_{r+1}^{(0)}, \dots, t_n^{(0)})$  a partir de (1) con  $(\alpha_i, \beta_i)' = (\alpha_i^{(0)}, \beta_i^{(0)})'$ ,  $i = r+1, \dots, n$ . A continuación, generamos  $\alpha_1^{(1)}$  a partir de (3), con  $t_j = t_j^{(0)}$ ,  $j = r+1, \dots, n$ ,  $\beta_1 = \beta_1^{(0)}$ ,  $\mathbf{B} = \mathbf{B}^{(0)}$  y  $\mathbf{H} = \mathbf{H}^{(0)}$ , mediante el método de aceptación-rechazo (ver, por ejemplo, Devroye, 1986, p. 40), utilizando como función importante la densidad log-Student t  $(\log \alpha | \mu'_1, \sigma_1'^2)$ . De forma similar, generamos  $\beta_1^{(1)}$  a partir de (4), con  $t_j = t_j^{(0)}$ ,  $j = r+1, \dots, n$ ,  $\alpha_1 = \alpha_1^{(1)}$ ,  $\mathbf{B} = \mathbf{B}^{(0)}$  y  $\mathbf{H} = \mathbf{H}^{(0)}$ , utilizando como función importante la densidad log-Student t  $(\log \beta | \mu'_2, \sigma_2'^2)$ . Del mismo modo, generamos  $\alpha_2^{(1)}$  a partir de (3), con  $t_j = t_j^{(0)}$ ,  $j = r+1, \dots, n$ ,  $\beta_2 = \beta_2^{(0)}$ ,  $\mathbf{B} = \mathbf{B}^{(0)}$  y  $\mathbf{H} = \mathbf{H}^{(0)}$ , generamos  $\beta_2^{(1)}$  a partir de (4), con  $t_j = t_j^{(0)}$ ,  $j = r+1, \dots, n$ ,  $\alpha_2 = \alpha_2^{(1)}$ ,  $\mathbf{B} = \mathbf{B}^{(0)}$  y  $\mathbf{H} = \mathbf{H}^{(0)}$ , y así sucesivamente. Finalmente, generamos los hiperparámetros  $\mathbf{B}^{(1)}$  y  $\mathbf{H}^{(1)}$ , a partir de la distribución Normal-Wishart (2), con  $(\alpha_i, \beta_i) = (\alpha_i^{(1)}, \beta_i^{(1)})$ ,  $i = 1, \dots, n$ .

Repetimos el bucle hasta alcanzar estacionariedad en las cadenas de Markov y, posteriormente, hasta construir una muestra de la distribución final. A partir de esa muestra de la distribución final puede estudiarse cualquier característica deseada.

Comprobamos gráficamente la estacionariedad de las cadenas de Markov, monitorizando la evolución de las mismas. Actualmente sigue abierta la investigación de métodos que calculen de forma automática el número de pasos iniciales a desechar en la cadena de Markov, no habiéndose propuesto todavía ninguno de fácil utilización en aplicaciones concretas. También es un tema actual de investigación el número de cadenas a utilizar para la obtención de la muestra, e incluso si utilizar las cadenas enteras o solamente un submuestreo de las mismas que disminuya la autocorrelación, ver, por ejemplo, Gelman y Rubin (1992), Geyer (1992), MacEachern y Berliner (1994) y referencias allí citadas.

Este esquema teórico del algoritmo de Gibbs es válido para cualquier vector de covariables  $x$ . No obstante, a partir de ahora vamos a centrarnos en el caso particular de un vector de covariables dicotómico.

El tiempo medio de supervivencia de un individuo de la primera población es:

$$\begin{aligned}\mu_1 &= E(t) = E(E(t|\alpha, \beta)) = E(\alpha/\beta) = E(\exp\{\log\alpha - \log\beta\}) = \\ &= \exp\{b_{11} + b_{12} - b_{21} - b_{22} + 1/2[\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2]^2\},\end{aligned}$$

y podemos construir una muestra de la distribución final de  $\mu_1$  a partir de la muestra  $\{B^{(i)}, H^{(i)}\}$ . Similarmente, el cociente de los tiempos medios de supervivencia de las dos poblaciones es:

$$\tau = \frac{\mu_1}{\mu_2} = \exp\{b_{12} - b_{22}\}.$$

Si la distribución de  $\tau$  está 'cerca' del 1, las dos poblaciones no serán diferentes en tiempos medios de supervivencia. De hecho, la hipótesis de igualdad de tiempos medios de supervivencia puede formularse como:  $b_{12} = b_{22}$ .

El estudio de la distribución bivalente  $\{b_{12}, b_{22}\}$  permite contrastar la hipótesis de igualdad de poblaciones, pues las dos poblaciones serán iguales si y sólo si  $b_{12} = b_{22} = 0$ .

Por último, la distribución predictiva podemos aproximarla, por Monte-Carlo, de la siguiente forma:

$$\begin{aligned}f(t|x, t_1, \dots, t_r, T_{r+1}, \dots, T_n, X) &= \int_0^\infty \int_0^\infty \text{Ga}(t|\alpha, \beta) \cdot \\ \cdot f(\alpha, \beta|x, t_1, \dots, t_r, T_{r+1}, \dots, T_n, X) d\alpha d\beta &\approx \frac{1}{m} \sum_{j=1}^m \text{Ga}(t|\alpha_{(j)}, \beta_{(j)}),\end{aligned}$$

con  $\{(\alpha_{(j)}, \beta_{(j)}), j = 1, \dots, m\}$ , una muestra obtenida a partir de:

$$f(\log \alpha, \log \beta | x, t_1, \dots, t_r, T_{r+1}, \dots, T_n, X) = \int \text{NM}((\log \alpha, \log \beta)' | Bx, H) \cdot f(B, H | t_1, \dots, t_r, T_{r+1}, \dots, T_n, X) dB dH \sim \frac{1}{N} \sum_{i=1}^N \text{NM}((\log \alpha, \log \beta)' | B^{(i)}x, H^{(i)}),$$

siendo  $\{(B^{(i)}, H^{(i)}), i = 1, \dots, N\}$  la muestra obtenida mediante Gibbs sampling.

En todas las implementaciones de este algoritmo hemos construido el punto inicial de la siguiente manera: utilizamos la media y la varianza de los tiempos de supervivencia no censurados de cada uno de los grupos para obtener los pares paramétricos iniciales  $(\alpha_i^{(0)}, \beta_i^{(0)})$ ,  $i = 1, \dots, n$ , (iguales para individuos del mismo grupo y que coinciden con los parámetros de una distribución Gamma cuya media y varianza igualan los correspondientes momentos muestrales de los tiempos de supervivencia no censurados) y una matriz  $B^{(0)}$  inicial, de modo que  $B^{(0)}x$  sea el vector columna formado por los logaritmos de los parámetros de la distribución Gamma anterior. Fijamos la matriz  $H^{(0)} = 20I_2$  (valor inicial que nos ha proporcionado buenos resultados en todas nuestras pruebas).

El algoritmo de Gibbs genera eventualmente una cadena de Markov estacionaria, con independencia del punto inicial elegido. Sin embargo, una elección razonable del mismo permite acelerar el proceso y conseguir estacionariedad en pocas etapas. El método aquí propuesto para la elección del punto inicial es sencillo, parece razonable y ha funcionado correctamente en todos los bancos de datos analizados.

### 3. EJEMPLOS NUMÉRICOS

El algoritmo expuesto en el apartado anterior lo empleamos para analizar una batería de bancos simulados a partir del modelo jerárquico, que nos permitían comparar los resultados obtenidos con los verdaderos valores paramétricos utilizados para su simulación. Todos esos estudios proporcionaron resultados similares. Aquí comentamos un banco de tamaño 200, que incluye 28 datos progresivamente censurados por la derecha.

También hemos utilizado este algoritmo en algún banco de datos reales. En este apartado incluimos el análisis de unos datos sobre remisión de leucemia pues son, posiblemente, los que se han utilizado con mayor frecuencia para comprobar el funcionamiento de métodos estadísticos de comparación de curvas de supervivencia.

En los dos ejemplos aquí desarrollados hemos utilizado una distribución inicial Normal-Wishart poco informativa para los hiperparámetros, en concreto:

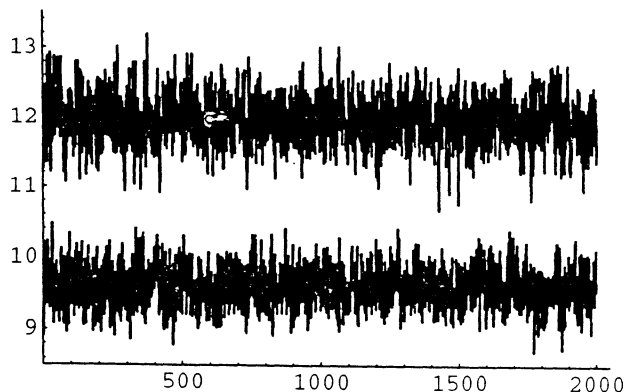
$$f(B, H) = \text{NMW}(B, H | M, A, T, \alpha) = \text{NM}(B^V | M^V, H \otimes A) W(H | T^{-1}, \alpha),$$

con  $M = 0_{2 \times 2}$ ,  $A = I_2$ ,  $T = (0.01)I_2$  y  $\alpha = 2$ .

### 3.1. Análisis de un banco de datos simulados

Construimos un banco de 200 datos simulados a partir del modelo jerárquico Gamma con hiperparámetros  $B = \begin{bmatrix} 4.6 & 0.2 \\ 2.3 & 0.0 \end{bmatrix}$  y  $H = \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}$ .

Tras realizar 20000 pasos en la cadena de Markov (19 minutos en una estación Sun SPARCclassic), partiendo de un punto inicial construido según el método propuesto en el apartado anterior, monitorizamos la evolución de las medias de los tiempos de supervivencia en cada uno de los grupos, representándolas cada 10 pasos en la figura 1. En dichas gráficas se observa estabilidad desde casi el principio, además de muy poca variabilidad en ambos grupos.



**Figura 1.**

Evolución de las medias de los tiempos de supervivencia en cada uno de los grupos del banco de datos simulados.

El estudio de la densidad predictiva lo realizamos desechando los 15000 primeros pasos de la cadena de Markov y, posteriormente, cogiendo uno de cada 10 hasta obtener una muestra de tamaño 500. Así, por Monte-Carlo obtuvimos las medias y

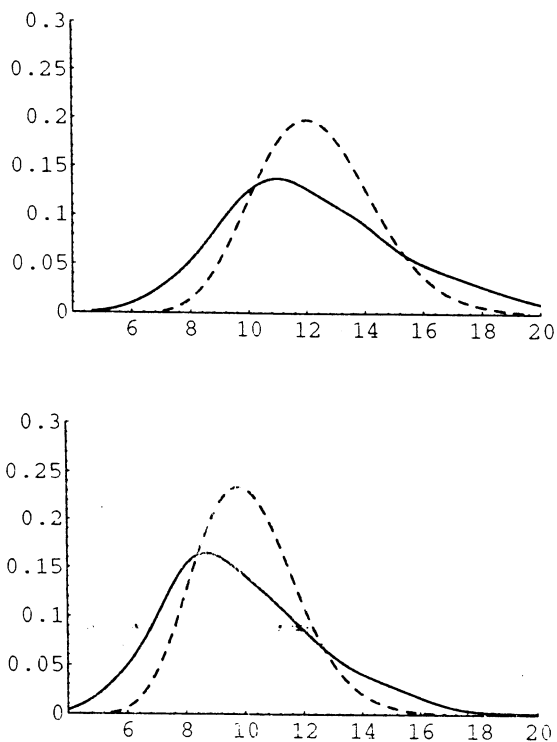
varianzas de las densidades predictivas para cada uno de los grupos. Se muestran en la tabla 1.

**Tabla 1**

*Media y varianza de la predictiva en cada uno de los grupos del banco de datos simulados*

	Grupo 1	Grupo 2
Media	12.3769	9.9653
Varianza	11.2328	7.7097

Las gráficas, para cada uno de los grupos, de las densidades predictivas (en trazo continuo) y de las densidades con las que realmente hemos simulado los datos (en trazo discontinuo), se recogen en la figura 2.

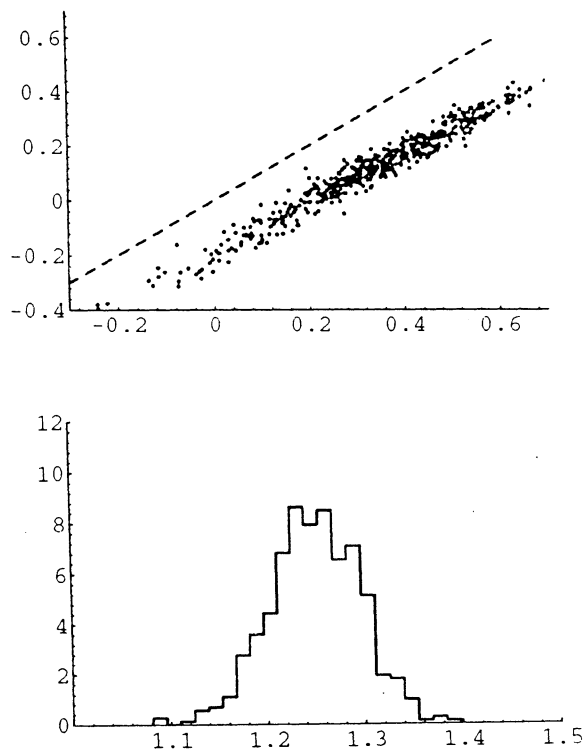


**Figura 2.**

Densidad predictiva y densidad correcta de cada uno de los grupos del banco de datos simulados.



Por último, hemos representado la nube de puntos de la distribución bivalente  $\{b_{12}, b_{22}\}$  y el histograma de  $\tau = \exp\{b_{12} - b_{22}\}$ . Estas gráficas se muestran en la figura 3 y permiten comparar los dos grupos como ya comentamos en el apartado 2. Observamos que la nube de puntos se encuentra por debajo de la diagonal y no se acerca al cero, por lo que podemos deducir la diferencia entre los dos grupos.



**Figura 3.**

Nube de puntos de  $\{b_{12}, b_{22}\}$  e histograma de  $\tau$  del banco de datos simulados.

Para comprobar la consistencia del estudio independientemente del punto inicial elegido, analizamos el mismo banco de datos partiendo de distintos puntos iniciales. Los resultados obtenidos fueron muy similares a los presentados. Por ejemplo, partiendo de un punto inicial que coincidía con los verdaderos valores paramétricos (esto es posible dado que los datos son simulados), observamos una evolución de las medias de los tiempos de supervivencia en cada uno de los grupos análoga a la de la figura 1 y realizando el estudio de la densidad predictiva siguiendo el mismo criterio, obtuvimos los siguientes estadísticos:

**Tabla 2**

*Media y varianza de la predictiva en cada uno de los grupos del banco de datos simulados, utilizando puntos iniciales iguales a los verdaderos valores paramétricos*

	Grupo 1	Grupo 2
Media	12.4865	10.0578
Varianza	10.5993	6.9834

### 3.2. Análisis de datos de leucemia

El banco de datos sometido a análisis consiste en 42 tiempos de remisión de la leucemia, medidos en semanas, en dos grupos con el mismo número de pacientes (Freireich, E.S. *et al.*, 1963). Al primero de ellos se le somete a tratamiento con 6-mercatopurina (6-MP), mientras que al segundo se le administra un placebo. Se recogen los datos en la tabla 3, donde el asterisco indica observación censurada.

**Tabla 3**

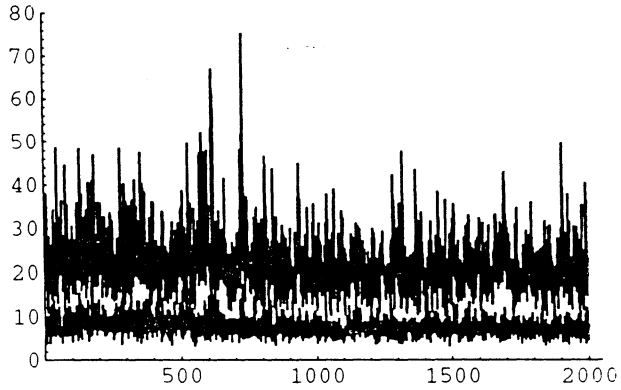
*Tiempos de remisión (en semanas) de leucemia en dos grupos de pacientes*

6-MP	6	6	6*	7	9*	10	10*	11*	13	16	17*	19*	20*	22	23	25*	32*	32*	34*	35*	
Placebo	1	1	2	2	3	4	4	5	5	8	8	8	8	11	11	12	12	15	17	22	23

Estos datos han sido profusamente analizados por diferentes autores para estudiar nuevos métodos y tests de comparación de dos distribuciones de supervivencia. Así, Lawless (1982), previa constatación gráfica de que los tiempos de remisión en cada uno de los grupos podían provenir de distribuciones Weibull, obtuvo los estimadores máximo verosímiles de los parámetros y concluyó que ambos grupos eran distintos en tiempos medios de supervivencia utilizando intervalos de confianza. Asimismo, Lee (1992) utilizó el test F de Cox para distribuciones exponenciales para demostrar que los grupos no podían ser considerados iguales.

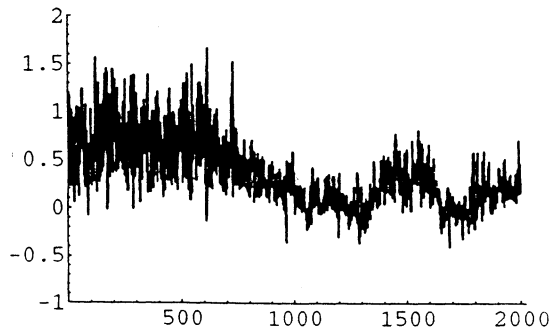
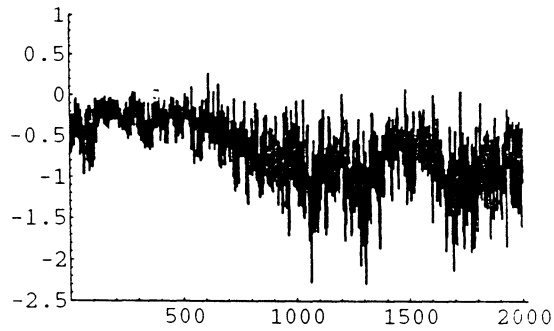
Al igual que en el subapartado anterior, construimos el punto de partida del algoritmo de Gibbs y monitorizamos la evolución de las medias de los tiempos de supervivencia en cada uno de los grupos, representándolas cada 20 pasos en la figura 4. También se observa estabilidad desde casi el principio, aunque mayor variabilidad en el primer grupo.

La evolución de los hiperparámetros B es bastante similar, mostrando estabilidad casi desde el principio de la cadena de Markov. De un modo totalmente análogo a la gráfica anterior, se muestra en la figura 5 la evolución de  $b_{12}$  y  $b_{22}$ , que son los dos parámetros que permiten diferenciar los grupos.



**Figura 4.**

Evolución de las medias de los tiempos de supervivencia en cada uno de los grupos del banco de datos de leucemia.



**Figura 5.**

Evolución de los hiperparámetros  $b_{12}$  y  $b_{22}$  en el banco de datos de leucemia.

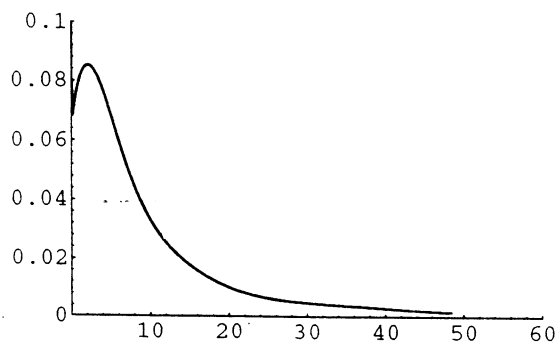
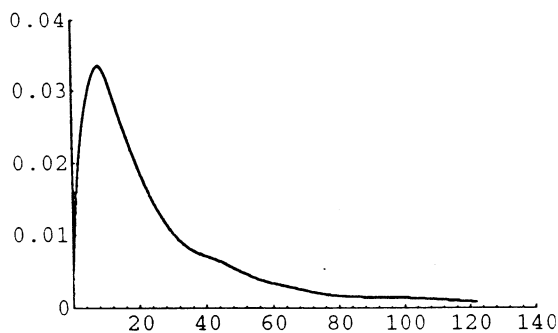
El estudio de la densidad predictiva lo realizamos desechando los 20000 primeros pasos de la cadena de Markov y, posteriormente, cogiendo uno de cada 20 hasta obtener una muestra de tamaño 500. Así, por Monte-Carlo obtuvimos las medias y varianzas de las densidades predictivas para cada uno de los grupos. Se muestran en la tabla 4.

**Tabla 4**

*Media y varianza de la predictiva en cada uno de los grupos en el banco de datos de leucemia*

	Grupo 1	Grupo 2
Media	35.6264	13.3486
Varianza	2550.9143	692.7137

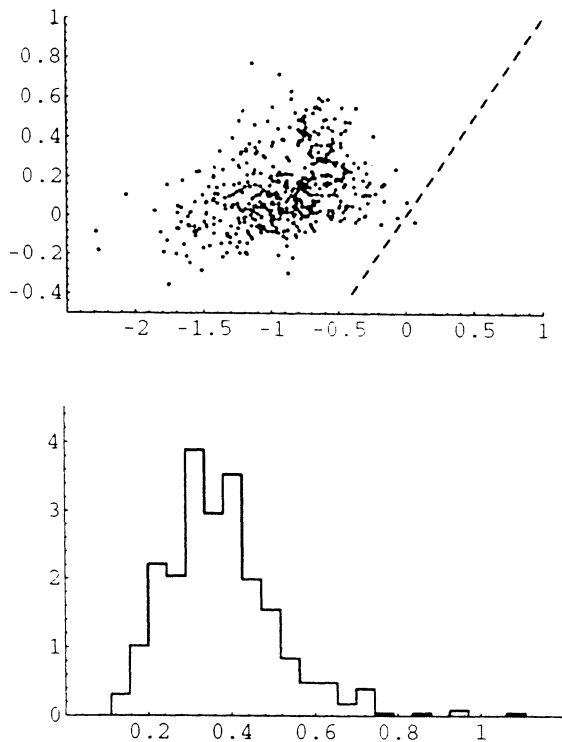
Representamos, en la figura 6, las gráficas de las densidades predictivas para cada uno de los dos grupos.



**Figura 6.**

Densidad predictiva de cada uno de los grupos en el banco de datos de leucemia.

Finalmente, en la figura 7 representamos la nube de puntos de la distribución bivalente  $\{b_{12}, b_{22}\}$  y el histograma de  $\tau = \exp\{b_{12} - b_{22}\}$  para ambos grupos. Observamos que la nube de puntos se encuentra por encima de la diagonal y no se acerca al cero, por lo que podemos deducir la diferencia entre los dos grupos.



**Figura 7.**

Nube de puntos de  $\{b_{12}, b_{22}\}$  e histograma de  $\tau$  en el banco de datos de leucemia.

#### 4. COMENTARIOS FINALES

El análisis de los dos ejemplos numéricos estudiados en el apartado anterior muestra claramente que existen diferencias entre los dos grupos de individuos. La respuesta bayesiana adecuada al problema de comparación de tiempos medios de supervivencia en las dos poblaciones vendría dada por el cálculo de algún intervalo de confianza

sobre la distribución final del parámetro  $\tau = \exp\{b_{12} - b_{22}\}$ . Como ya se comentó en el apartado 2, si existe igualdad de tiempos medios en las dos poblaciones  $\tau = 1$ . Similarmente, si existe igualdad entre las dos poblaciones,  $b_{12} = b_{22} = 0$ ; por tanto, para responder a la pregunta de igualdad de poblaciones habría que construir alguna región de confianza de la distribución final conjunta de  $(b_{12}, b_{22})$ .

Como una primera solución, hemos obtenido un intervalo de confianza 0.95 para  $\tau$  utilizando el histograma normalizado de la figura 7 como aproximación de su distribución final. Siguiendo las indicaciones de Wei y Tanner (1990), obtuvimos el intervalo (0.1525, 0.5637) que claramente no incluye el valor  $\tau=1$ .

Por otro lado, en este trabajo hemos exigido que la distribución inicial sobre los hiperparámetros B y H pertenezca a la familia conjugada. Esto no supone una restricción importante: si se desea utilizar otra inicial el método de aceptación-rechazo permite obtener una muestra de la distribución final deseada a partir de una muestra construida según el procedimiento descrito en el apartado 2, empleando una distribución inicial Normal-Wishart que no sea muy distinta de la inicial que se desea utilizar.

Los comentarios de dos evaluadores anónimos sobre una versión anterior de este trabajo, han permitido mejorar notablemente la claridad en la exposición del mismo.

## REFERENCIAS

- [1] **Bermúdez, J.D. y Beamonte, E.** (1993). "Análisis bayesiano de datos de supervivencia Gamma utilizando muestreo de Gibbs". *Estadística Española*, **35**, 629–644.
- [2] **Broemeling, L.D.** (1985). *Bayesian analysis of linear models*. New York: Marcel Dekker.
- [3] **Devroye, L.** (1986). *Non-uniform random variate generation*. New York: Springer-Verlag.
- [4] **Freireich, E.J., Gehan, E.A., Frei, E. et al.** (1963). "The effect of 6-Mercaptopurine on the duration of steroid-induced remissions in acute leukemia: a model for evaluation of other potential useful therapy". *Blood*, **21**, 699–716.
- [5] **Gelfand, A.E. and Smith, A.F.M.** (1990). "Sampling-based approaches to calculating marginal densities". *J. Amer. Statist. Assoc.*, **85**, 398–409.
- [6] **Gelman, A. and Rubin, D.B.** (1992). "Inference from iterative simulation (with discussion)". *Statistical Science*, **7**, 457–511.
- [7] **Geyer, C.J.** (1992). "Practical Markov chain Monte Carlo (with discussion)". *Statistical Science*, **7**, 473–511.

- [8] **Lawless, J.F.** (1982). *Statistical models and methods for lifetime data*. New York: John Wiley & Sons.
- [9] **Lee, E.T.** (1992). *Statistical methods for survival data analysis*. 2nd. edition. New York: John Wiley & Sons.
- [10] **MacEachern, S.N.** and **Berliner, L.M.** (1994). "Subsampling the Gibbs sampler". *The American Statistician*, **48**, 188–190.
- [11] **Smith, A.F.M.** and **Roberts, G.O.** (1993). "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods". *J. Roy. Statist. Soc. B*, **55**, 3–23.
- [12] **Wei, G.C.G.** and **Tanner, M.A.** (1990). "Calculating the content and boundary of the highest posterior density region via data augmentation". *Biometrika*, **77**, 649–652.

## ENGLISH SUMMARY:

### COMPARISON OF GAMMA SURVIVAL CURVES

Eduardo Beamonte y José D. Bermúdez

In this paper we propose a hierarchical model for the bayesian analysis of survival data, progressively censored on the right, with covariates. This is a generalization of a previous work (Bermúdez and Beamonte, 1993) where each survival time is assumed to be distributed as a  $Ga(\alpha, \beta)$  with  $\log \alpha, \log \beta$  distributed as a bivariate Normal with mean  $Bx$  and precision matrix  $H$ , where  $x$  is the vector of covariates for each individual.

We are specially interested on the comparison of two groups of individuals. For that we considerate a dichotomous covariate, which identifies the group membership of each individual, together with a constant covariate equal to one. Our main objective is obtain inferences on the hiperparameters related with the dichotomous covariate.

The great complexity of the posterior distribution, independently of the prior distribution used and emphasised by the typical presence of censored data, drive us to study it through Monte Carlo methods (in this line was pioneer the work of Gelfand and Smith, 1990). In order to do this, we use the Gibbs sampling, one of the most

important iterative methods of obtaining a sample from the posterior distribution (see Smith and Roberts, 1993, and references there). It requires that the sampling from all of the complete conditional distributions be easy, but that is our case.

If we denote the survival times corresponding to no censored data by  $(t_1, \dots, t_r)$  and the censorship times corresponding to censored data by  $(T_{r+1}, \dots, T_n)$ , then the parametric vector object of the Gibbs sampling is composed by the parameters of the model  $(\alpha_1, \beta_1, \dots, \alpha_n, \beta_n)$ , the hiperparameters  $(B$  and  $H)$ , and the nonobserved survival times  $(t_{r+1}, \dots, t_n)$ , subjets to the restrictions  $t_i > T_i, \quad i = r + 1, \dots, n$ . So the complete conditional distribution of a nonobserved survival time is the truncated Gamma (1) and the corresponding to the hiperparameters is the Normal-Wishart (2) (if we use a distribution belonging to that family for the prior  $f(B, H)$ ). In both cases is relatively easy sampling from them. We solve the problem of sampling from the complete conditional distribution of  $\alpha_i$  (3) and the corresponding to  $\beta_i$  (4), using the acceptance-rejection method with importance functions belonging to the log-t-student family.

We implement the algorithm of the Gibbs sampling with an initial point based on the first moments of the exact survival times and we check grafically the convergence of the Markov chains. Once a sample of the posterior distribution is obtained, we can calculate some of its characteristics by Monte Carlo methods and even to obtain the predictive distribution, of special interest in most applications.

The paper ends with the analysis of two data sets, one of them simulated and the other real. The first one, together with some others which we have analysed, permits to compare the results obtained with the true parameters. The real data consist in 42 remission times of leukemia (Freireich *et al.*, 1963), that has been analysed by many other authors. We apply our method and obtain the predictive distribution for each group. We conclude that the two groups are differents by obtaining a confidence interval of the parameter of interest, agreeing with previous works.