Universidad Autónoma de Guerrero

Unidad Académica de Matemáticas

Maestría en Matemáticas Aplicadas

Modelación empírica de crecimiento de tumores usando estadística bayesiana

T E S I S

PARA OBTENER EL GRADO DE:

Maestría en Matemáticas Aplicadas

PRESENTA:

Lic. Carlos David Marquez Landa

DIRECTOR DE TESIS:

Dr. Flaviano Godínez Jaimes

Dr. Ramón Reyes Carreto

Universidad Autónoma de Guerrero

Unidad Académica de Matemáticas

Maestría en Matemáticas Aplicadas

Modelación empírica de crecimiento de tumores usando estadística bayesiana

T E S I S

PARA OBTENER EL GRADO DE:

Maestría en Matemáticas Aplicadas

PRESENTA:

Lic. Carlos David Marquez Landa

DIRECTOR DE TESIS:

Dr. Flaviano Godínez Jaimes

Dr. Ramón Reyes Carreto

Dedicatoria

Dedico esta tesis a mi FAMILIA, mis padres Petra Landa y Carlos Marquez, mis hermanos Luis y Karla, mis pequeños sobrinos Yoselyn y Alexander, así como a todos mis tios y primos que de alguna manera me apoyaron.

A mis compañeros de la Maestría Juan Carlos, Uriel, Luis Xavier y Victor por su ayuda, consejos y apoyo durante este viaje.

A mis Profesores quienes nunca desistieron en la enseñanza que me dieron y continuaron depositando su confianza.

A mis grandes amigos Alejandra, Michelle, Jesus y sobre todo dedicada a mi Amor Isabel.

Recuerda, todos tropezamos, todos y cada uno de nosotros. Por eso es un consuelo ir de la mano de alguien. (Emily kimbrough)

Agradecimientos

A mis Padres, por todos los sacrificios que hicieron para darme la oportunidad de estudiar.

A mis hermanos que aún con la distancia siento su apoyo.

A mis amigos y compañeros con los que compartí salon y juntos obtuvimos conocimientos.

A mi al núcleo académico de la Maestría de Matemáticas Aplicadas que me dieron la oportunidad de ser parte de este posgrado, en especial al Dr Flaviano Godínez Jaimes por dejarme trabajar a su lado y por su paciencia.

Al Consejo Nacional de Ciencia y Tecnología, por haberme brindado el apoyo económico necesario para realizar mis estudios de posgrado.

Y todos los que directa o indirectamente me apoyaron en este proceso.

¡GRACIAS!

Resumen

A pesar de la complejidad interna, la cinética del crecimiento tumoral sigue leyes relativamente simples que se pueden expresar mediante modelos matemáticos. Un modelo empírico general para estos fenómenos es el Modelo Biparamétrico Generalizado (MBG) que depende de cuatro parámetros $a, b, \alpha, \beta > 0$

$$V' = aV^{\alpha} - bV^{\beta}.$$

El modelo supone que la tasa de cambio del volumen es una diferencia entre la tasa de proliferación y la tasa de degradación. El MBG incluye a tres de los modelos más usados: el Logístico, Gompertz y Bertalanffy. Los datos del crecimiento de tumores pueden presentar el problema de varianza no constante, es decir, la variabilidad incrementa con el tiempo. En este trabajo se generaron datos mediante la solución del Modelo Logístico Generalizado (MLG) que depende de tres parámetros $a, b > 0$ y $\beta > 1$

$$V' = aV - bV^{\beta},$$

agregando un error aleatorio $\varepsilon \sim N(0, \sigma^2)$ y para generar la heterocedasticidad se multiplicó el error por el tiempo elevado a una potencia ν. Se generaron dos casos; el Caso 1 con $\beta = 1.4$ y el Caso 2 con $\beta = 1.6$, en ambos casos $a = 0.5$ y $b = 0.04$. En estos escenarios se propusieron estimadores bayesianos para los parámetros de los tres modelos clásicos (Logístico, Gompertz y Bertalanffy) y el MLG. Los estimadores bayesianos se obtuvieron considerando distribuciones *a priori* no informativas especificas como la Gamma y la Uniforme, además se compararon mediante criterios de información, suma de cuadrados de predicción, error de predicción y factor bayes. En el Caso 1 el MLG fue el mejor de acuerdo a los criterios de información y factor bayes. En el Caso 2 el modelo Logístico fue el mejor

según los criterios de información, suma de cuadrados de predicción, error de predicción y factor bayes. En ambos casos los modelos clásicos bayesianos son mejores que el MLG de acuerdo a las sumas de cuadrados de predicción y al error de predicción.

Finalmente los estimadores bayesianos se utilizaron para analizar datos experimentales. Al comparar los modelos estimados con los criterios de información, suma de cuadrados de predicción, error de predicción y factor bayes, los modelos de Gompertz y de Bertalanffy fueron mejores que el modelo Logístico. Pero al considerar la predicción de observaciones futuras, el modelo Gompertz fue mejor.

Palabras claves: Crecimiento tumoral, modelos matemáticos, estimación bayesiana, distribuciones *a priori* no informativa.

Abstract

Despite internal complexity, tumor growth kinetics follows simple laws that can be expressed as mathematical models. A general empirical model for these phenomena is the Generalized Two-parameter Model (GTM) which depends on four parameters $a, b, \alpha, \beta > 0$

$$V' = aV^{\alpha} - bV^{\beta}.$$

The model assumes that the volumen change rate is a difference between the rate of proliferation and the rate of the degradation. The GTM includes three of the more used models: Logistic, Gompertz and Bertalanffy. The data tumor growth can present the problem of the not constant variance, that is the variability increases with the time. In this work the data are generated by the solution of the Generalized Logistic Model (GLM) which depends on the three parameters $a, b > 0$ and $\beta > 1$

$$V' = aV - bV^{\beta},$$

adding a random error $\varepsilon \sim N(0, \sigma^2)$ and to generate the heteroscedasticity the error was multiplied by the time raised to a power ν. Two cases were generated, the Case 1 with $\beta = 1.4$ and the Case 2 with $\beta = 1.6$, in both cases $a = 0.5$ and $b = 0.04$. In these scenarios bayesian estimators were proposed for the three classical models (Logistic, Gompertz and Bertalanffy) and GLM. The bayesian estimators were obtained considering the non informative *a priori* distributions as Gamma and Uniform, in addition they were compared using information criteria, sum of squares of prediction, prediction error and bayes factor. In Case 1 the GLM it was the best according to the information criteria and the bayes factor. In Case 2 the Logistic model was the best according to the information criteria, sum of squares of prediction, prediction error and bayes factor. In both cases the

bayesian estimation of the classical models were better than the GLM according to the sum of squares of prediction and prediction error.

Finally the bayesian estimators were used to analyze experimental data. Comparison of the models with the information criteria, sum of squares of prediction, prediction error and bayes factor the Gompertz and the Bertalanffy models were the better that the Logistic model. But prediction of future observations was better with Gompertz model.

Keyword: Tumor growth, mathematical models, bayesian estimation, non informative *a priori* distributions.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

En este capítulo se expone brevemente algunos de los trabajos realizados con anterioridad sobre la modelación de crecimiento de tumores y los modelos que se analizan en el presente trabajo de investigación. Además un breve panorama de lo que es el cáncer y las últimas estadísticas de esta enfermedad.

1.1. Antecedentes

El cáncer puede ser estudiado bajo un enfoque molecular, epidemiológico o matemático. Bajo el enfoque matemático, se han propuesto modelos que estudian el crecimiento del tumor respecto al tiempo. Se han realizado diversos trabajos sobre el modelado de crecimiento tumoral.

Freyer y Sutherland (1986) investigaron los efectos de la glucosa y oxígeno en el crecimiento de Esferoides Multicelulares[1]. Los esferoides inicialmente crecieron con una misma tasa exponencial en todas las condiciones de cultivo, con el volumen y número de células de los esferoide duplicandose en un tiempo de 20 a 24 horas. Las tasas de crecimiento disminuyeron a un tiempo mayor y el máximo de volumen y de células fueron proporcionales

[1]Los Esferoides Mulricelulares constituyen un modelo experimental, como una técnica de cultivo célular útil para el estudio la biología tumoral dada su gran similitud a los tumores in-vivo.

a las concentraciones de oxígeno y glucosa. Utilizaron el modelo de Gompertz (Gompertz, 1825) para ajustar diferentes cultivos con distintas concentraciones de oxígeno y glucosa.

Aunque el modelo de Gompertz fue muy conocido y ha sido utilizado para describir distintos tipos de crecimiento biológicos, no fue hasta que Marušić y Bajzer (1993) generalizaron el modelo propuesto por Gompertz a partir del la ecuación general de crecimiento biparamétrica, el cual llamaron modelo de Gompertz generalizado (Marušić y Bajzer, 1993). Además de proponer soluciones analíticas para la ecuación general biparamétrica, mostraron que la solución del modelo de Gompertz generalizado se puede ver como un caso límite de la solución de la ecuación general.

Marušić y Vuk-Pavlović (1993) compararon el poder de predicción de 4 modelos; el modelo de Gompertz, el modelo de Gompertz Generalizado, el modelo de Piantadosis (Piantadosi, 1985) y el modelo de autoestimulación (Marûšić et al., 1991), mediante un error relativo usado por Vaidya y Alexandro (1982) (citado por Marušić y Vuk-Pavlović, 1993). Los datos usados fueron los descritos por Freyer y Sutherland (1986). De esta comparación el modelo de Gompertz fue mejor al tener un poder de predicción mayor, seguido del modelo de Piantadosis.

Marušić et al. (1994) realizaron una comparación del crecimiento de tumores con 17 modelos clasificados como: modelos empíricos, funcionales y estructurales mediante datos experimentales. La comparación de los modelos se realizó para determinar que modelo fue mejor, ya sea para el ajuste de los datos o para predecir alguna observación. El modelo de Gompertz y el modelo de Piantadosis fueron los mas adecuados para describir el comportamiento de las observaciones, además mostraron que algunos modelos, como el Logístico y el Bertalanffy (von Bertalanffy, 1957) son inadecuados para describir el comportamiento de los datos.

Una modificación del error relativo utilizado por Vaidya y Alexandro, además de una puntuación de predicción, es utilizada por Benzekry et al. (2014) para el analísis de modelos clásicos en la descripción y predicción de crecimiento experimental de tumores.

En los trabajos ya mencionados, se utilizaron metodos de ajuste clásicos como mínimos cuadrados ordinarios y generalizados.

Cáncer

El cáncer es una enfermedad provocada por un grupo de células que se multiplican sin control y de manera autónoma, que invaden localmente y a distancia otros tejidos. Esto es resultado de la interacción de factores genéticos y externos (físicos y químicos).

La Organización Mundial de la Salud señala que en 2012 fallecieron 8.2 millones de personas que hubo 14 millones de casos nuevos y que este número aumentará a 22 millones en las siguientes dos décadas. El 70 % de todas las muertes por cáncer registradas en 2012 se produjeron en África, Asia, América Central y Sudamérica (Organización Mundial de la Salud, 2015).

En la región de las Américas (continente americano e islas del Caribe) fallecieron 1.3 millones de personas en el 2012, un 47 % de las cuales ocurrieron en América Latina y el Caribe. Los hombres fueron afectados principalmente por el cáncer de próstata, pulmón, colorrectal y estómago; y las mujeres por el de pulmón, mama, colorrectal y cervicouterino según la Organización Panamericana de la Salud (2015).

En México, la Unión Internacional Contra el Cáncer, menciona que el cáncer es la tercera causa de muerte y estima que cada año se detectan 128 mil casos nuevos. En México durante 2013, la morbilidad hospitalaria por tumores malignos (población que egresa de un hospital por dicha enfermedad) más alta tanto en mujeres como en hombres menores de 20 años, por cáncer en órganos hematopoyéticos[2] (Instituto Nacional de Estadística y Geografía, 2016).

Si bien la mayoria de los cánceres se pueden prevenir, esta enfermedad ha sido responsable de un número importante de muertes a nivel mundial. Esto se debe a que el cáncer es el estado final de un largo y complejo proceso evolutivo que incluye la regulación de la proliferación, el control del ciclo celular, el reclutamiento del estroma, la angiogénesis y escape de la vigilancia inmune.

La mayoria de los cánceres se debe a la acción de agentes externos que actuan sobre el organismo causando alteraciones célulares (Menchón, 2007). Algunas de las causas son:

[2]Son los encargados de la formación de células sanguíneas, estos órganos son bazo, ganglio linfático, timo, hígado y medula ósea.

herencia, virus, radiaciones, productos químicos, alteraciones inmunologicas, entre otros factores (Figura 1.1).

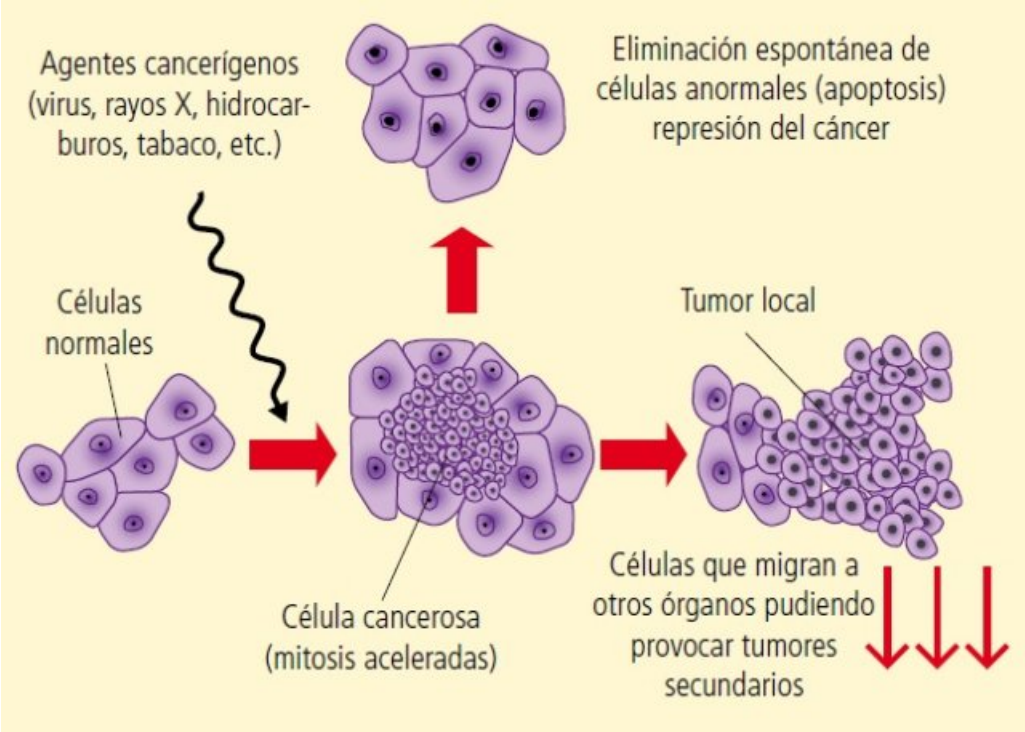

Figura 1.1: *Generación de un tumor*

En condiciones fisiólogicas se requiere de una estricta regulación genética para mantener un equilibrio celular de forma adecuada a las necesidades de cada tejido, esto se hace mediante la contraposición de dos procesos:

1. **La proliferación**, es el proceso simultáneo de diferenciación celular, particular en cada tejido. La formación por mitosis de nuevas células determina la tasa de proliferación.

2. **La apoptosis**, muerte celular programada, es un mecanismo de eliminación de células al final de su vida activa.

La alteración del equilibrio en condiciones patológicas (neoplasias) provoca un crecimiento incontrolado que puede deberse a una excesiva proliferación o una reducida muerte celular (Cabrera et al., 2008). A pesar de lo complejo que es el proceso, el crecimiento de tumores sigue leyes simples que se pueden expresar con modelos matemáticos.

Modelos matemáticos

Diversos modelos matemáticos de crecimiento tumoral se han desarrollado desde el siglo XIX. Se utilizan la mayoría de ellos para describir de manera cualitativa los estados tempranos (prevasculares) de crecimiento y estabilidad del tejido tumoral, bajo numerosas hipótesis y simplificaciones (Menchón, 2007).

Los modelos empíricos se basan en el principio de que la tasa de cambio del tamaño del tumor, V', es una diferencia entre la tasa de crecimiento efectiva $G(V)$ y la tasa de degradación efectiva $D(V)$, es decir:

$$V' = G(V) - D(V), \quad V(0) = V_0 \tag{1.1}$$

donde $G(V)$ y $D(V)$ son funciones positivas, crecientes y diferenciables (Bajzer et al., 1996). Basado en el principio de alometría, von Bertalanffy (1957) propuso funciones potencia que describen las tasas empíricas de crecimiento y degradación:

$$G(V) = aV^{\alpha}, \quad D(V) = bV^{\beta} \tag{1.2}$$

respectivamente. La ecuación de crecimiento se describe por el Modelo Biparamétrico Generalizado (MBG):

$$V' = aV^{\alpha} - bV^{\beta}, \quad V(0) = V_0 \tag{1.3}$$

con $a, b, \alpha, \beta > 0$, donde a y b son las tasas constantes de proliferación y degradación respectivamente. Para von Bertalanffy (1957), α y β son exponentes que indican que las tasas son proporcionales a alguna potencia del volumen V, mientras que Bajzer et al. (1996) indican que caracterizan el orden fractal de las tasas efectivas de forma análoga a la cinética de reacción de tipo fractal (Kolpeman, 1988, citado por Bajzer et al. 1996).

El volumen depende del tiempo (t) y se utilizará V o $V(t)$ sin distinción alguna, además, V_0 denotará el valor inicial, $V(0)$.

Si en el MBG $\alpha = \beta$, entonces se obtiene el modelo Gompertz generalizado (MGG) :

$$V' = aV^{\alpha} - bV^{\alpha} \ln V, \quad V(0) = V_0 \tag{1.4}$$

Las Ecuaciones (1.3) y (1.4) tienen una solución explícita, pero Marušić et al. (1994) recomiendan resolver las ecuaciones numéricamente, ya que la solución contiene funciones

especiales, como la función beta incompleta modificada y la función integral exponencial modificada (Marušić y Bajzer, 1993) .

Si en el MBG $\alpha = 1$, entonces se obtiene el modelo logístico generalizado (MLG)

$$V' = aV - bV^{\beta}, \quad \beta > 1, \quad V(0) = V_0$$

que tiene como solución:

$$V(t) = \left[\frac{b}{a} - \left(\frac{b}{a} - V_0^{1-\beta}\right) e^{-a(\beta-1)t}\right]^{1/(1-\beta)}, \quad \beta > 1. \tag{1.5}$$

Si en el MBG $\beta = 1$ se tiene la ecuación *tipo metabólico* de von Bertalanffy también conocida como la ecuación de Richards:

$$V' = aV^{\alpha} - bV, \quad \alpha < 1, \quad V(0) = V_0$$

que tiene como solución:

$$V(t) = \left[\frac{a}{b} - \left(\frac{a}{b} - V_0^{1-\alpha}\right) e^{-b(1-\alpha)t}\right]^{1/(1-\alpha)}, \quad \alpha < 1 \tag{1.6}$$

Un aspecto importante de los sistemas biológicos es saber su comportamiento a largo plazo. En este trabajo se denota como V_m al volumen máximo del tumor, es decir:

$$V_m = \lim_{t\to\infty} V$$

Además del MBG, otros modelos clásicos usados en crecimiento de tumores.

a) Si en el MBG $\alpha = 1$ y $\beta = 2$, entonces se obtiene el modelo logístico (ML):

$$V' = aV - bV^2, \quad V(0) = V_0$$

que tiene como solución:

$$V(t) = \frac{a}{b}\left[1 - \left(1 - \frac{a}{bV_0}\right) e^{-at}\right]^{-1}$$

con $V_m = \frac{a}{b}$, que es la razón entre la tasa constantes de proliferación y la tasa constante de degradación. En la Figura (1.2) se observa el comportamiento del modelo logístico con distintas tasas. Este tipo de curva *sigmoidal* describen distintos procesos naturales, las líneas horizontales denotan el volumen máximo.

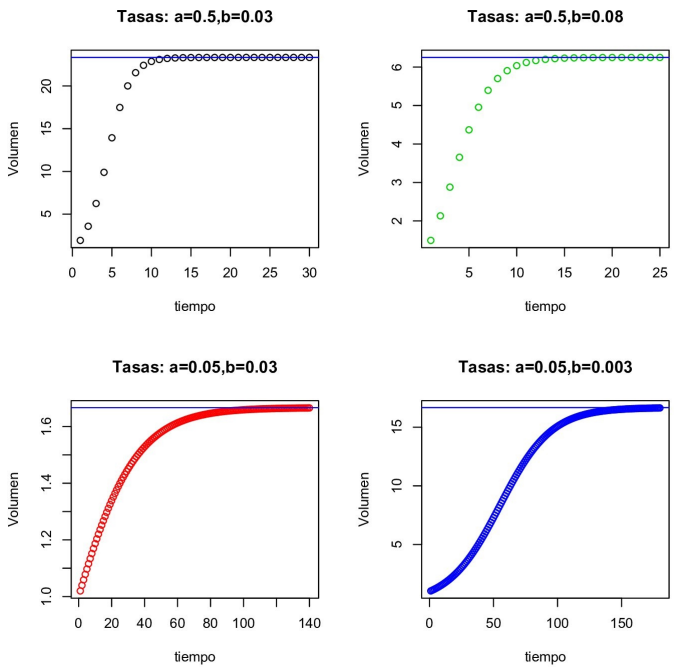

Figura 1.2: *Curvas logísticas con distintas tasas constantes de proliferación y degradación.*

b) Si en el MGG $\alpha = 1$, entonces se obtiene el modelo de Gompertz (MG):

$$V' = aV - bV \ln V, \quad V(0) = V_0$$

que tiene como solución:

$$V(t) = \exp\left[\frac{a}{b} - \frac{a}{b}\left(1 - \frac{b}{a}\ln V_0\right)e^{-bt}\right]$$

en este caso $V_m = \exp\left(\frac{a}{b}\right)$. Al igual que el ML, en el límite interviene la razón entre la tasa de crecimiento y degradación. Pero el MG se aproxima al límite en un tiempo (t) mayor. Por ejemplo si $a = 0.5$ y $b = 0.08$, cuando $t = 100$, volumen está muy próximo al límite $\exp\left(\frac{0.5}{0.08}\right) = 518.012$ (Figura 1.3), mientras que con las mismas tasas constantes el mínimo para aproximarse al límite en el ML es en $t = 20$.

c) Si en el MBG $\alpha = 2/3$ y $\beta = 1$, entonces se obtiene el modelo de Bertalanffy (MB):

$$V' = aV^{2/3} - bV, \quad V(0) = V_0$$

que tiene como solución:

$$V(t) = \left[\frac{a}{b} + e^{-bt/3}\left(V_0^{1/3} - \frac{a}{b}\right)\right]^3$$

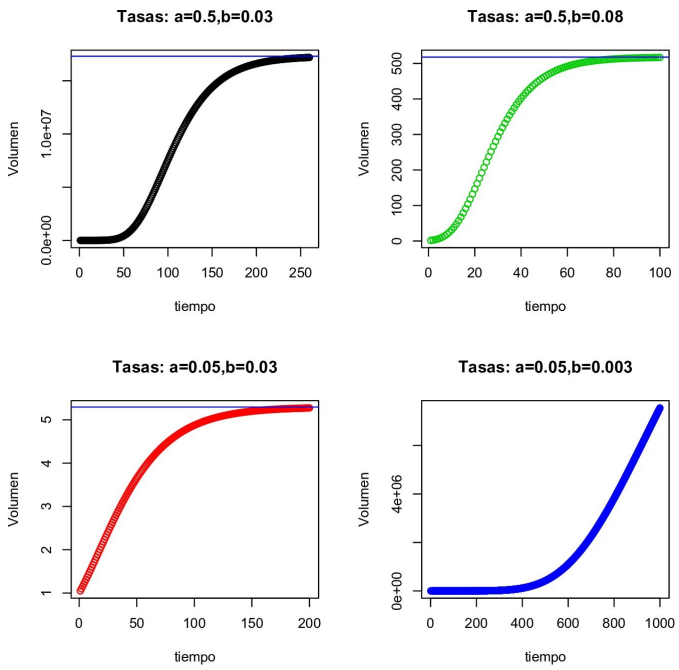

Figura 1.3: *Curvas de Gompertz con distintas tasas constantes de proliferación y degradación.*

con $V_m = \left(\frac{a}{b}\right)^3$. Esta vez el modelo se aproxima al límite en un tiempo mayor al de ML y MG. En la Figura (1.4) se observa que si $a = 0.5$ y $b = 0.08$ el mínimo tiempo para *alcanzar* el límite es en $t = 200$. Mucho más grande que el mínimo para el ML y casi el doble para el MG, por lo menos para las tasas propuestas.

Como se observa la representación gráfica del volumen en función del tiempo se traduce a una curva de forma tipicamente sigmoidal en la cual es facil distinguir varias fases: 1) Fase inicial, en la que se observa un incremento exponencial del volumen; 2) Fase de retraso, caracterizada por una expansión lineal; 3) Funto de inflexión, que se caracteriza al momento en que la curva cambia de ser cóncava a ser convexa y 4) Fase de aplanamiento, que se prolonga hasta el final del tiempo de observación (Torres, 1990).

La modelación del crecimiento de tumores se ha abordado desde un enfoque estadístico, estimando los parámetros mediante mínimos cuadrados ponderados como se mencionó anteriormente.

Existen dos aspectos que complican la estimación de los parámetros:

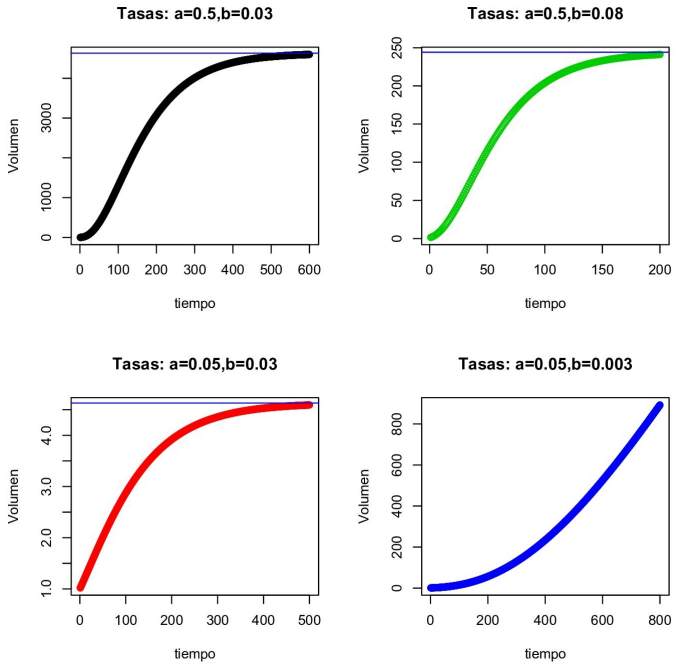

Figura 1.4: *Curvas de Bertalanffy con distintas tasas constantes de proliferación y degradación.*

1. No se cumple el supuesto de homogeneidad de varianzas, lo cual es común en la formulación de muchos modelos (Figura 1.5).

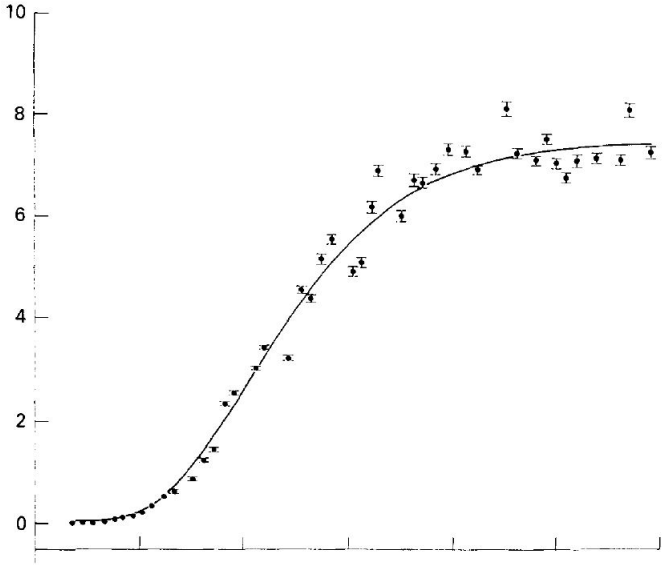

Figura 1.5: *Datos utilizados por Marušić (1994).*

2. No se plantean umbrales que permitan decidir cuando ligeras variaciones en los valores de los parámetros estimados $\widehat{a}$, $\widehat{b}$, dan evidencia de los valores que definen los distintos casos del MBG.

El primer problema puede ser resuelto usando métodos de estimación diferentes al clásico, como es la estimación Bayesiana, que además usa información previa acerca del fenómeno estudiado. El segundo problema puede ser resuelto proponiendo intervalos creíbles y pruebas de hipótesis bayesianas para a y b.

1.2. Objetivos

1. Proponer modelos bayesianos para describir y predecir el crecimiento de tumores con el problema de heterocedasticidad..

2. Comparar y evaluar los modelos bayesianos mediante criterios de tipo predictivo, de bondad de ajuste y el factor bayes.

Capítulo 2

Marco teórico

En este capítulo se describe el enfoque bayesiano, su inferencia, así como los métodos de comparación de modelos. Las definiciones en este capitulo fueron tomadas en su mayoria de los libros; *Bayesian modeling using winbugs de Ntzoufras (2009)* y *Bayesian data analysis de Gelman et al. (2014)*.

2.1. Análisis bayesiano

El uso de modelos bayesianos es una alternativa a problemas resueltos bajo el enfoque estadístico clásico. El término *bayesiano* hace referencia a Thomas Bayes (1702 - 1761), matemático británico que demostró un caso particular del teorema que ahora lleva su nombre, este teorema es la base del enfoque bayesiano. La estadística bayesiana proporciona un completo paradigma de inferencia estadística y teoría de la decisión con incertidumbre. El teorema de Bayes proporciona una forma adecuada para incluir información inicial de los parámetros. En contraste con el enfoque clásico, la inferencia bayesiana introduce como parte del modelo una distribución *a priori* $f(\theta)$, que expresa el conocimiento o ignorancia acerca del parámetro θ.

2.2. Inferencia Bayesiana

Sea $\boldsymbol{y} = (y_1, \ldots, y_n)$ un vector de observaciones cuya distribución de probabilidad $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ depende de un vector de parámetros de dimensión k, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. Además, $\boldsymbol{\theta}$ tiene una función densidad $f(\boldsymbol{\theta})$. Dados los datos observados $\boldsymbol{y}$, junto con el teorema de Bayes, la distribución condicional de $\boldsymbol{\theta}$ es

$$f(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})}{f(\boldsymbol{y})} \propto f(\boldsymbol{y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}) \tag{2.1}$$

donde $f(\boldsymbol{\theta})$ es la distribución *a priori* de $\boldsymbol{\theta}$, $f(\boldsymbol{\theta} \mid \boldsymbol{y})$ es la distribución *a posteriori* de $\boldsymbol{\theta}$ dado $\boldsymbol{y}$, y $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ es la función de verosimilitud,

Definición 2.2.1 *Sea una muestra aleatoria $y_1, \ldots, y_n$, de tamaño n, la función de verosimilitud*

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i \mid \boldsymbol{\theta}) \tag{2.2}$$

Desde el punto de vista bayesiano $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ contiene la información disponible proporcionada por la muestra observada.

Al hacer inferencia sobre algún parámetro $\boldsymbol{\theta}$, generalmente se cuenta con información acerca de su valor, antes de ver los datos. Esto se representa en la distribución *a priori*

Definición 2.2.2 *La distribución a priori $f(\boldsymbol{\theta})$ mide el grado de conocimiento inicial que se tiene de los parámetros en estudio.*

La distribución *a priori* es una distribución conjunta en caso de que θ dimensión mayor a 1. Esto complica la busqueda de esta distribución. Comúnmente se considera la distribución *a priori* como el producto de las distribuciones marginales, es decir, se supone que los parámetros θ_i y θ_j, son independientes con $i \neq j$. En ciertos problemas, el conocimiento inicial sobre el parámetro $\boldsymbol{\theta}$ puede ser muy débil o vago, esto lleva a un tipo de distribuciones *a priori*, llamadas no informativas.

Definición 2.2.3 *Una distribución de $\boldsymbol{\theta}$ es no informativa si no contiene información sobre $\boldsymbol{\theta}$, es decir, no establece si unos valores de $\boldsymbol{\theta}$ son más favorables que otros.*

Las distribuciones no informativas más usuales son la distribución uniforme, la distribución normal difusa y la distribución gamma con parámetros de forma y escala muy cercanos a cero.

Una vez obtenidos los datos, la distribución *a priori* y la verosimilitud se obtiene la distribución *a posteriori* dada por la Ecuación (2.1). Esta describe lo que se conoce del parámetro dados los datos y es la base de inferencia bayesiana. Algunas veces la distribución *a posteriori* es muy compleja para realizar inferencia ya que intervienen integrales multiples dificiles o imposibles de calcular, por ello se utilizan algoritmos computacionales conocidos como métodos de cadenas de Markov Monte Carlo.

2.2.1. Cadenas de Markov Monte Carlo

Los métodos de cadenas de Markov Monte Carlo (MCMC, por sus siglas en inglés) son algoritmos que permiten obtener una muestra de una distribución de probabilidad f sin necesidad de simular directamente la distribución. Esta técnica es popular desde la decada de los 90's. Metropolis propuso por primera vez un algoritmo usando cadenas de Markov, y lo utilizó para el caso específico de la distribución de Boltzmann (Metropolis et al., 1953, citado por Nzoufras, 2009).

Una cadena de Markov es un proceso estocástico $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(T)}\}$ tal que

$$f(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}, \ldots, \boldsymbol{\theta}^{(1)}) = f(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)})$$

es decir, la distribución de $\boldsymbol{\theta}$ en el estado $t+1$ dados todos los estados previos de $\boldsymbol{\theta}$ depende solo del estado anterior, $\boldsymbol{\theta}^{(t)}$. Además $f(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)})$ es independiente del tiempo t. Por último, cuando la cadena de Markov es irreducible, aperiódica y positiva recurrente; si $t \to \infty$ la distribución de $\boldsymbol{\theta}^{(t)}$ converge a una distribución estacionaria o de equilibrio, que es independiente del valor inicial de la cadena $\boldsymbol{\theta}^{(0)}$ (Gilks et al., 1996).

Con el fin de generar una muestra de $f(\boldsymbol{\theta})$ se debe contruir una cadena de Markov con dos propiedades; la primera es que $f(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)})$ debe ser fácil de generar y la segunda es que la distribución estacionaria debe ser la distribución *a posteriori* de interés. Suponiendo que hemos construido una cadena de Markov con estos requisitos, entonces el algoritmo para obtener la MCMC consiste en:

1. Seleccionar un valor inicial.
2. Generar T resultados hasta alcanzar la distribución estacionaria.
3. Monitorear la convergencia del algoritmo usando diagnósticos de convergencia. Si no se obtiene la convergencia, el diagnóstico lo detecta y se generan más observaciones.
4. Desechar las primeras B observaciones
5. Considerar $\{\boldsymbol{\theta}^{(B+1)}, \boldsymbol{\theta}^{(B+2)}, \ldots, \boldsymbol{\theta}^{(T)}\}$ como una muestra para el análisis posterior.
6. Graficar la distribución *a posteriori* (usualmente se pone atención en las distribuciones marginales univariadas)
7. Resumir la distribución *a posteriori* (media, mediana, desviación estándar, cuantiles)

Los métodos MCMC más populares son el algoritmo de Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) y el muestreo de Gibbs (Ntzoufras, 2009, pp 36-37).

2.2.1.1. Metropolis-Hastings

El algoritmo Metropolis-Hastings (M-H) es probablemente el método MCMC por excelencia. Fue nombrado en honor a los trabajos realizados por Nicholas Metropolis y W. Keith Hastings. Como ya se mencionó Metropolis lo utilizó para la distribución Boltzmann (Metropolis et al. 1953). Posteriormente Hastings presentó una versión para casos más generales (Hastings, 1970).

La idea básica del Metropolis-Hastings es construir una cadena de Markov estacionaria que converja a $f(\boldsymbol{\theta} \mid \boldsymbol{y})$. El componente principal del algoritmo es la distribución instrumental, $q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)})$, a partir de la cual se genera un resultado de $\boldsymbol{\theta}^{(t+1)}$ condicionado en $\boldsymbol{\theta}^{(t)}$. Para asegurar que el algoritmo converja a $f(\boldsymbol{\theta} \mid \boldsymbol{y})$, la cadena de Markov debe satisfacer la condición de reversibilidad

$$f(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y})q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) = f(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{y})q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t+1)}) \tag{2.3}$$

La condición de reversibilidad puede ser impuesta en (2.3) para inducir equilibrio en la ecuación:

$$f(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y})q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)})\alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)}) = p(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{y})q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t+1)}) \tag{2.4}$$

donde $\alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)})$ es la probabilidad de transición y está definida como:

$$\alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)}) = min\left[\frac{f(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{y})q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t+1)})}{f(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y})q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)})}, 1\right] \tag{2.5}$$

con $f(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y})q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) > 0$.

El algoritmo Metropolis-Hastings queda determinado como sigue:

1. Dado el valor inicial $\boldsymbol{\theta}^{(0)}$, muestrear el candidato aleatorio z de $q(\boldsymbol{\theta}^{(1)} \mid \boldsymbol{\theta}^{(0)})$ y u de $U(0, 1)$.

2. Si $u < \alpha(\boldsymbol{\theta}^0, \boldsymbol{\theta}^1)$, fijar $\boldsymbol{\theta}^1 = z$.

3. Si $u > \alpha(\boldsymbol{\theta}^0, \boldsymbol{\theta}^1)$, fijar $\boldsymbol{\theta}^1 = \boldsymbol{\theta}^0$.

4. Regresar al paso 1 y usar $\boldsymbol{\theta}^1$ para generar $\boldsymbol{\theta}^2$.

Se repite el algoritmo m veces hasta que se tenga una muestra grande, en Tanner (1996) se dan varios criterios de terminación.

2.2.1.2. Muestreo Gibbs

El muestreo Gibbs fue introducido por Geman y Geman (1984) y es uno de los más usados. Esta técnica consiste en generar variables aleatorias indirectamente de una distribución sin tener que calcular la función de densidad.

Sean $\boldsymbol{\theta}$ el vector que contiene todos los parámetros del modelo y $f(\boldsymbol{\theta} \mid \boldsymbol{y})$ la distribución objetivo, entonces para cada uno de los elementos de $\boldsymbol{\theta}$ las distribución condicional completa son:

$$f(\theta_1 \mid \theta_2, \theta_3, \ldots, \theta_k, \boldsymbol{y}) \ldots f(\theta_k \mid \theta_1, \theta_2, \ldots, \theta_{k-1}, \boldsymbol{y}).$$

Bajo el muestreo Gibbs, las distribuciones condicionales se usan para generar sucesiones de valores de los parámetros aleatorios univariados, para cada uno de los elementos en $\boldsymbol{\theta}$.

Dado un valor inicial $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_k^{(0)})$, el algoritmo simula una cadena de Markov en la que $\boldsymbol{\theta}^{(t+1)}$ se obtiene a partir de $\boldsymbol{\theta}^{(t)}$ de la siguiente manera:

Generar la observación $\theta_1^{(t+1)}$ de $f(\theta_1 \mid \theta_2^{(t)}, \theta_3^{(t)}, \ldots, \theta_k^{(t)}, \boldsymbol{y})$

Generar la observación $\theta_2^{(t+1)}$ de $f(\theta_2 \mid \theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_k^{(t)}, \boldsymbol{y})$

$$\vdots$$

Generar la observación $\theta_k^{(t+1)}$ de $f(\theta_k \mid \theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{k-1}^{(t+1)}, \boldsymbol{y})$.

La sucesión $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$, es una realización de una cadena de Markov cuya distribución de transición está dada por

$$f(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}, \boldsymbol{y}) = \prod_{i=1}^{k} f(\theta_i^{(t+1)} \mid \theta_1^{(t+1)}, \ldots, \theta_{i-1}^{(t+1)}, \theta_{i+1}^{(t)}, \ldots, \theta_k^{(t)}, \boldsymbol{y}).$$

2.2.1.3. Diagnósticos de convergencia

Al utilizar métodos con MCMC se espera la convergencia de las cadenas de Markov a una distribución estacionaria. Sin embargo, no hay garantía de que las cadenas converjan después de k muestras. Es imposible estar seguros, pero existen varias pruebas gráficas y estadísticas, para verificar si la o las cadenas convergen.

2.2.1.3.1 Diagnóstico visual

Una manera de identificar la convergencia de las cadenas es observar que tan bien se mezclan o se mueven alrededor del espacio de parámetros. Si las cadenas están tomando mucho tiempo para moverse en el espacio de parámetros, entonces necesitarán más tiempo para converger.

La Figura (2.1), muestra una buena mezcla (izquierda), lo que indica que la convergencia de la cadenas en el tiempo sugerido, en el gráfico de la derecha, indica el caso contrario, es decir, se lleva más tiempo para alcanzar la convergencia.

En el software estadístico *R* (R Core Team, 2015a), la función *traceplot* {*coda*} (Plummer et al., 2006), grafica el número de iteraciones contra el valor de la muestra de cada parámetro.

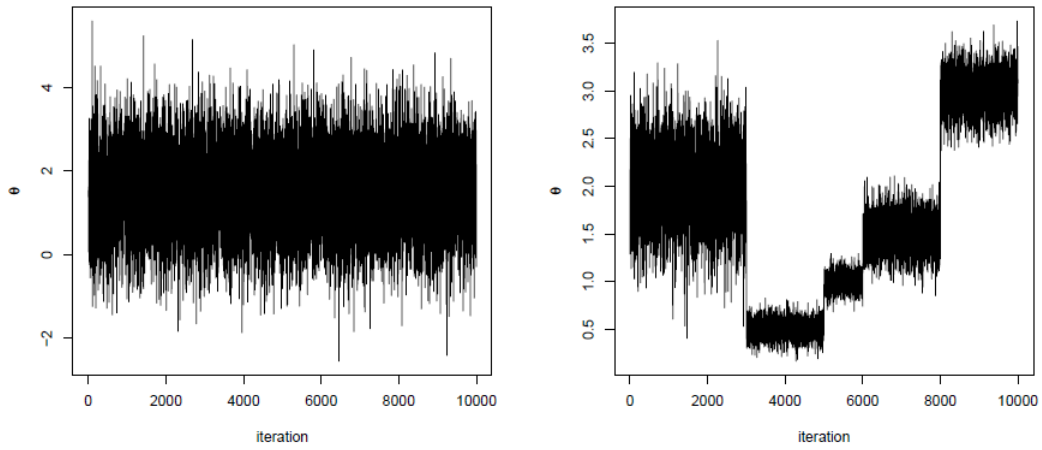

Figura 2.1: *Traza de un parámetro θ.*

2.2.1.3.2 Diagnóstico de Gelman-Rubin

Cuando se generan múltiples cadenas de Markov, cada una comenzando con valores iniciales diferentes, una diagnóstico de convergencia es el de Gelman-Rubin (Gelman y Rubin, 1992). La prueba compara la variabilidad entre las muestras y dentro de las mismas muestras. Para realizar este diagnóstico se siguen los siguientes pasos:

1) Simular $m \geq 2$ cadenas de longitud $2n$ con valores iniciales sobredispersos.

2) Descartar las primeras iteraciones en cada cadena, digamos las primeras n.

3) Se calcula la varianza dentro de las cadenas y entre las cadenas. La varianza dentro de la cadena se estima con:

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2$$

donde

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{ij} - \overline{\theta}_j)^2.$$

s_j^2 es la fórmula para la varianza de la j-ésima cadena. Entonces W es la media de las varianzas de las cadena.

La varianza entre las cadenas es:

$$B = \frac{n}{m-1}\sum_{j=1}^{m}\left(\overline{\theta}_j - \overline{\overline{\theta}}\right)^2 \tag{2.6}$$

donde

$$\overline{\overline{\theta}} = \frac{1}{m}\sum_{j=1}^{m}\overline{\theta}_j.$$

La expresión (2.6) es la varianza de las medias de las cadenas, esta es multiplicada por n debido a que cada cadena consta n valores.

4) Se estima la varianza de la distribución estacionaria como el promedio ponderado de W y B,

$$\widehat{Var}(\theta) = \left(1 - \frac{1}{n}\right)W + \frac{1}{n}B.$$

5) Por último se calcula el factor potencial de reducción de escala:

$$\widehat{R} = \sqrt{\frac{\widehat{Var}(\theta)}{W}}.$$

Hay convergencia de la cadena de Markov con el tamaño de la muestra usado si $\widehat{R} \approx 1$.

2.2.1.3.3 Diagnóstico de Geweke

Geweke (1991) sugirió un diagnóstico para verificar la convergencia de la media de cada parámetro por separado de los valores muestreados de una sola cadena. Para construir esta prueba, propuso la inspección de los valores simulados, obtenido por una MCMC. En ésta se aplica una simple prueba Z para comprobar si las medias estimadas, a partir de dos submuestras diferentes de la muestra total, son iguales. Usualmente se compara el 10 % inicial y el 50 % final de la muestra total.

Para la varianza asintótica de la media muestral, se usa una estimación de la densidad de la teoría espectral. Para un conjuto generado de valores $\theta_1^{(1)}, \ldots, \theta_1^{(T)}$ de un parámetro θ de interes, se calcula la media de la muestra $\overline{\theta}$, seguida por el cálculo de la distribución espectral $S_\theta(\omega)$ para series de tiempo. Entonces el error estándar de la media está dado por $\sqrt{S_\theta(0)/T}$ y por tanto el estadístico Z se define como

$$Z = \frac{\overline{\theta}^A - \overline{\theta}^B}{\sqrt{\frac{S_\theta^A(0)}{T_A} + \frac{S_\theta^B(0)}{T_B}}}.$$

Asintóticamente Z tiene distribución normal estandar, donde $\overline{\theta}^A$, $\overline{\theta}^B$ son las medias muestrales de las dos submuestras mencionadas; T_A y T_B son los tamaños de las submuestras; y $S_\theta^A(0)$, $S_\theta^B(0)$ son las varianzas de las submuetras.

Los parámetros con $|Z| \geq 2$ tienen medias diferentes de las primeras y últimas iteraciones, y por lo tanto indican no convergencia de la MCMC (Ntzoufras, 2009).

Si los diagnósticos no son favorables en algún caso en particular, se puede recurrir a las funciones *update* {*stats*} (R Core Team, 2015b) y *autojags* {*R2jags*} (Su y Yajima, 2015) de R, las cuales actualizan las cadenas hasta que converjan.

2.2.2. Estimadores bayesianos

Una vez obtenida la distribución *a posteriori*, la estimación de $\boldsymbol{\theta}$ se realiza considerando una función de pérdida $L(\widehat{\theta}, \theta)$ que cuantifica las posibles penalidades al estimar θ con $\widehat{\theta}$.

Definición 2.2.4 *Sean $f(\boldsymbol{\theta} \mid \boldsymbol{y})$ la distribución a posteriori y $L(\widehat{\theta}, \theta)$ una función de pérdida, entonces el estimador bayesiano está dado por*

$$\widehat{\theta} = \arg\min_{\widehat{\theta}} E[L(\theta, \widehat{\theta}) \mid \boldsymbol{y}]$$

donde

$$E[L(\theta, \widehat{\theta}) \mid \boldsymbol{y}] = \int_\Theta L(\widehat{\theta}, \theta) f(\boldsymbol{\theta} \mid \boldsymbol{y}).$$

La función de pérdida es a veces difícil de especificar, ya que la pérdida no siempre es fácilmente medible.

Algunas de las funciones de pérdidas más utilizadas son:

a) La función de pérdida cuadrática:

$$L(\widehat{\theta}, \theta) = (\widehat{\theta} - \theta)^2$$

produce como estimador bayesiano la media de la distribución *a posteriori*.

b) La función de pérdida lineal absoluta:

$$L(\widehat{\theta}, \theta) = |\widehat{\theta} - \theta|^2$$

produce como estimador bayesiano la mediana de la distribución *a posteriori*.

c) La función de pérdida todo/nada:

$$L(\widehat{\theta}, \theta) = \begin{cases} 0, si & \widehat{\theta} = \theta \\ 1, si & \widehat{\theta} \neq \theta. \end{cases}$$

produce como el estimador bayesiano la moda de la distribución *a posteriori*.

2.2.3. Intervalos creíbles

En la inferencia bayesiana, los intervalos creíbles son la contraparte de los intervalos de confianza en el análisis estadístico clásico.

Definición 2.2.5 *Un intervalo de credibilidad al* $100(1-\alpha)\,\%$ *para* θ *es un subconjunto* C *de* Θ *tal que*

$$1 - \alpha \leq P(C \mid \boldsymbol{y}) = \begin{cases} \int_C f(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta} & \textit{caso continuo} \\ \sum_{\theta \epsilon C} f(\boldsymbol{\theta} \mid \boldsymbol{y}) & \textit{caso discreto} \end{cases}$$

La distribución *a posteriori* $f(\boldsymbol{\theta} \mid \boldsymbol{y})$ permite calcular probabilidades en Θ, por lo tanto permite hablar de la probabilidad de que θ esté en C. Esta es la diferencia con los intervalos de confianza clásicos que pueden solamente ser interpretados en términos de probabilidad de cobertura. Al calcular un conjunto creíble para θ, una propiedad deseable es que la longitud del intervalo sea pequeña.

2.2.4. Comparación de modelos

El adelanto de la tecnología ha hecho posible resolver problemas complejos con diferentes modelos estadísticos ya sean clásicos o bayesianos. Lo que lleva a identificar que modelo es el apropiado. Enseguida se presentan medidas y criterios que ayudan a determinar que modelo es mejor en el sentido de ajuste o de predicción.

2.2.4.1. Criterios para la selección de modelos

Desde un punto de vista frecuentista la devianza es la diferencia en log-verosimilitudes entre el modelo saturado y el modelo ajustado. Por analogía, Dempster (1997) sugirió examinar la distribución *a posteriori* de la devianza clásica (Berg et al., 2004).

Definición 2.2.6 *La devianza para la selección de modelos bayesianos (Berg et al., 2004), se define como:*

$$D(\boldsymbol{\theta}_m, m) = -2\log[f(\boldsymbol{y} \mid \boldsymbol{\theta}_m, m)] + 2\log g(\boldsymbol{y}) \tag{2.7}$$

donde $\boldsymbol{y}$ *son los datos,* $\boldsymbol{\theta}_m$ *el vector de parámetros bajo el modelo* m*,* $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ *es función de verosimilitud y* $\log g(\boldsymbol{y})$ *denota un término de normalización totalmente especificado por los datos.*

Dempster propuso comparar la media posterior de la devianza. Siguiendo esta idea Spiegelhalter et al. (2002) propuso el criterio de información de la devianza (DIC por sus siglas en inglés) basado en la distribución *a posteriori* de $D(\boldsymbol{\theta})$.

Definición 2.2.7 *El DIC se define como:*

$$DIC(m) = D(\overline{\boldsymbol{\theta}}_m, m) + 2p_m$$

donde $\overline{\boldsymbol{\theta}}$ *es la media posterior del parámetro en el modelo* m*.*

La ventaja del DIC sobre otros criterios, en el caso de selección de modelos bayesianos, es que el DIC se calcula fácilmente a partir de las muestras generadas por una MCMC.

El criterio de información bayesiano (BIC por sus siglas en inglés) esta basado originalmente en el criterio que introdujo Schwarz (1978), dado por:

$$S_{01} = \log(f(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}}_{m_1}, m_1)) - \log(f(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}}_{m_0}, m_0)) - \frac{1}{2}(d_{m_1} - d_{m_0})\log(n) \tag{2.8}$$

donde n es el tamaño de muestra, $\widehat{\boldsymbol{\theta}}_m$ son las estimaciones de máxima verosimilitud del parámetro $\widehat{\boldsymbol{\theta}}_m$ del modelo m y d_m es la dimensión de $\widehat{\boldsymbol{\theta}}_m$.

Definición 2.2.8 *El BIC para el modelo* m *está definido como:*

$$BIC(m) = D(\widehat{\boldsymbol{\theta}}_m, m) + d_m \log(n) \tag{2.9}$$

donde $D(\widehat{\boldsymbol{\theta}}_m, m)$ es la devianza del modelo m definida en la Ecuación 2.7.

El criterio de información de Akaike (AIC, por sus siglas en inglés) fue introducido por Akaike (1973) como la distancia de Kullback-Leibler esperada entre un modelo verdadero y uno modelo estimado (Ntzoufras, 2009).

Definición 2.2.9 *El AIC para el modelo m está definido como*

$$AIC(m) = D(\widehat{\boldsymbol{\theta}}_m, m) + 2d_m \tag{2.10}$$

Al comparar modelos, se elige aquel que tenga el menor DIC, BIC o AIC. Esto esta argumentado con detalle en Ntzoufras (2009).

2.2.4.2. Factor Bayes

El factor bayes se usa para comparar modelos bayesianos. Supongamos un problema en el que se tiene que elegir entre dos posibles modelos m_1 y m_2 , una vez que se ha observado una muestra $\boldsymbol{y}$.

Definición 2.2.10 *El factor bayes (FB) se define como:*

$$FB = \frac{f(\boldsymbol{y} \mid m_1)}{f(\boldsymbol{y} \mid m_2)}$$

donde $f(\boldsymbol{y} \mid m)$ *es la verosimilitud marginal bajo el modelo* m, $m\epsilon\{m_1, m_2\}$ *que está dada por*

$$f(\boldsymbol{y} \mid m) = \int f(\boldsymbol{y} \mid \boldsymbol{\theta}_m) f(\boldsymbol{\theta}_m \mid m) d\boldsymbol{\theta}_m. \tag{2.11}$$

$f(\boldsymbol{y} \mid \boldsymbol{\theta}_m)$ es la verosimilitud bajo el modelo m con el vector de parámetros $\boldsymbol{\theta_m}$ y $f(\boldsymbol{\theta}_m \mid m)$ es la distribución *a priori* de $\boldsymbol{\theta}_m$ bajo el modelo m. Existen varios métodos para estimar (2.11), en este trabajo se realizó via MCMC y se detalla el procedimiento en la Subsección (2.2.4.2.1).

El FB es similar a las pruebas de la razón de verosimilitudes pero ahora, en lugar de maximizar la verosimilitud, el factor bayes realiza un promedio ponderado en la distribución de los parámetros. Una interpretación sugerida para el factor de Bayes es proporcionada por Kass y Raftery (1995) (Cuadro 2.1).

2.2.4.2.1 Estimación de la verosimilitud marginal

Un método para aproximar la verosimilitud marginal es utilizar los valores simulados, via MCMC, de la distribución *a posteriori* mediante la siguiente ecuación:

$$\int \frac{1}{f(\boldsymbol{y} \mid \boldsymbol{\theta}_m, m)} f(\boldsymbol{\theta}_m \mid \boldsymbol{y}, m) d\boldsymbol{\theta}_m = \int \frac{1}{f(\boldsymbol{y} \mid \boldsymbol{\theta}_m, m)} \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta}_m, m) f(\boldsymbol{\theta}_m \mid m)}{f(\boldsymbol{y} \mid m)} d\boldsymbol{\theta}_m = \frac{1}{f(\boldsymbol{y} \mid m)}.$$

Cuadro 2.1: *Interpretación del FB según Kasss and Raftery*

FB	Fuerza de evidencia a favor de m_1
$1-3$	Negativa (apoya a m_2)
$3-20$	Positiva
$20-150$	Fuerte
>150	Muy Fuerte

Newton y Raftery (1994) propusieron un estimador para la verosimilitud marginal. Este estimador, se obtiene en base a la media armónica de las verosimilitudes calculadas en cada paso de un algoritmo MCMC y está definido como:

$$\widehat{f}_1(\boldsymbol{y} \mid m) = \left(\frac{1}{T} \sum_{t=1}^{T} \left\{ f(\boldsymbol{y} \mid \boldsymbol{\theta}_m^{(t)}, m) \right\}^{-1} \right)^{-1}. \tag{2.12}$$

Gelfand y Dey (1994) generalizaron esta idea con:

$$\widehat{f}_2(\boldsymbol{y} \mid m) = \left(\frac{1}{T} \sum_{t=1}^{T} \frac{g(\boldsymbol{\theta}_m^{(t)})}{f(\boldsymbol{y} \mid \boldsymbol{\theta}_m^{(t)}, m) f(\boldsymbol{\theta}_m^{(t)} \mid m)} \right)^{-1} \tag{2.13}$$

donde $g(\boldsymbol{\theta})$ es una distribución con colas más delgadas que el producto de la distribución *a priori* y la verosimilitud. Una elección sugerida para $g(\boldsymbol{\theta})$ es una distribución multivariada normal o, t con media y varianza igual a la media y varianza *a posteriori*.

2.2.4.3. Suma de cuadrados de predicción

Otro criterio para comparar los modelos es cuantificando el error de predicción. Una forma de hacerlo consiste en obtener las diferencias entre los valores ajustados y los observados.

Definición 2.2.11 *La Suma de Cuadrados de Predicción (SCP) es dada por:*

$$SCP = \frac{1}{n} \sum_{t=1}^{n} (V_t - \widehat{V}_t)^2 \tag{2.14}$$

donde V_t es el valor observado al tiempo t, y $\widehat{V}_t$ el valor ajustado con el modelo al tiempo t, definido como:

$$\widehat{V}_t = M(\widehat{\boldsymbol{\theta}}, t)$$

Mientras más pequeño es la SCP mejor ajuste tendrá el modelo.

2.2.4.4. Métrica de predicción y puntuación de éxitos

Otra forma de cuantificar la predicción de un modelos, es mediante el error normalizado (EN). Benzekry et al. (2014), lo definen como:

$$EN_{n,d} = \left| \frac{V_{n+d} - \widehat{V}_{n+d}}{\widehat{\sigma}_t} \right| \tag{2.15}$$

donde n es número de datos que se utilizan para predecir a una profundidad d, es decir, para predecir el valor al tiempo $n + d$. Mientras que V_{n+d} es la observación y $\widehat{V}_{n+d}$ es el valor predicho al tiempo $n + d$.

La predicción de un punto en el tiempo se considera aceptable cuando $EN_{n,d}$ es menor que tres, que corresponde a que el error de predicción está dentro de tres desviaciones estándar del error de medición de los datos (Benzekry et al., 2014). Para ello se define una puntuación de predicción, que es una variante del propuesto por Benzekry:

$$S_{n,d} = \#\left\{EN_{n,d} \leq 3\right\} \tag{2.16}$$

donde $\#\{\cdot\}$ es la cardinalidad del conjuto. Además se considera el error relativo (ER) que es una modificación del propuesto por Vaidya y Alexandro (1982) definido por:

$$ER_{n,d} = \left| \frac{V_{n+d} - \widehat{V}_{n+d}}{V_{n+d}} \right|. \tag{2.17}$$

2.2.4.5. Predicción de una observación futura

Si se está interesado en la predicción de una observación futura correspondiente a un tiempo $n + d$.

La distribución predictiva *a posteriori* $f(\tilde{V} \mid V)$, puede ser representada por la mezcla de la muestra de la distribución $f(\tilde{V} \mid \boldsymbol{\theta})$, donde se promedian sobre la distribución *a posteriori* del parámetro $\boldsymbol{\theta}$:

$$f(\tilde{V} \mid V) = \int f(\tilde{V} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta} \mid V) d\boldsymbol{\theta}. \tag{2.18}$$

Supongase que $\tilde{V}_{n+d}$ es una observación futura al tiempo t_d. Una vez que se tiene la distribución *a posteriori* se calcula el estimador bayesiano $\widehat{\boldsymbol{\theta}}$ y se simula el valor de $\tilde{V}$ a partir de la muestra de la distribución de $\widehat{\boldsymbol{\theta}}$.

De la muestra simulada para la distribución *a posteriori* de $\boldsymbol{\theta}$ se obtienen bandas creibles, con un nivel de credibilidad de $1-\alpha$ (Albert, 2009).

Capítulo 3

Metodología

En este capítulo se exponen los modelos estudiados, la distribución *a priori* asi como las distribuciones *a posteriori* utilizadas en cada caso. También se presenta el estudio de simulación realizado en este trabajo.

3.1. Formulación de los modelos

En este trabajo se utilizan los modelos deterministicos, presentados en el Capítulo 1. El modelo logístico:

$$M_1(a,b,t) \quad = \frac{a}{b}\left[1-\left(1-\frac{a}{bV_0}\right)e^{-at}\right]^{-1}$$

el modelo de Gompertz:

$$M_2(a,b,t) \quad = \exp\left[\frac{a}{b}-\frac{a}{b}\left(1-\frac{b}{a}\ln V_0\right)e^{-bt}\right]$$

y el modelo de Bertalanffy:

$$M_3(a,b,t) = \left[\frac{a}{b}+e^{-bt/3}\left(V_0^{1/3}-\frac{a}{b}\right)\right]^3$$

con dos parámetros. Además el modelo logístico generalizado con tres parámetros:

$$M_4(a,b,\beta,t) = \left[\frac{b}{a}-\left(\frac{b}{a}-V_0^{1-\beta}\right)e^{-a(\beta-1)t}\right]^{1/(1-\beta)}, \quad \beta>1$$

Para inducir la heterocedasticidad se generan las observaciones de la siguiente forma:

$$V(t) = M_4(a, b, \beta, t) + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_t^2) \tag{3.1}$$

donde a y b son las tasas constantes de proliferación y degradación respectivamente, y $V(t)$ es volumen del tumor al tiempo t. Usualmente la inferencia de los parámetros se basa en el supuesto de que los errores $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ son independientes e idénticamente distribuidos normal con media cero y varianza σ^2. Sin embargo, algunos fenómenos de crecimiento presentan heterocedasticidad y se requiere introducir esta estructura en el término de error en el modelo. En este trabajo se expresa la varianza σ^2 en la forma:

$$\sigma_t^2 = \sigma^2 t^\nu$$

bajo estas condiciones, $\varepsilon_t \sim N(0, \sigma^2 t^\nu)$. Además, se puede demostrar que

$$V(t) \sim N(\ M(a, b, \beta, t)\ ,\ \sigma^2 t^\nu\ ).$$

3.2. Inferencia

En el resto del documento se usará la notación $V(t)$ o V_t de manera indistinta. La funcion de verosimilitud para V_t es

$$\begin{aligned} f(\boldsymbol{V} \mid \boldsymbol{t}, \boldsymbol{\theta^*}, \sigma^2, \nu) &= \prod_{t=1}^{n} \frac{1}{\sqrt{2\sigma^2 t^\nu \pi}} \exp\left[-\frac{1}{2} \left(\frac{V_t - M(\boldsymbol{\theta^*}, t)}{\sigma t^{\nu/2}} \right)^2 \right] \\ &= (2\pi\sigma^2)^{-n/2} \prod_{t=1}^{n} t^{-\nu/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{t=1}^{n} t^{-\nu} (V_t - M(\boldsymbol{\theta^*}, t))^2 \right] \end{aligned}$$

donde $\boldsymbol{t} = (1, 2, \ldots, n)'$, $\boldsymbol{V} = (V_1.V_2, \ldots, V_n)'$, $\boldsymbol{\theta^*} = (a, b, \beta)$ y $\tau = 1/\sigma^2$ es la precisión, porque en estadística bayesiana es preferible utilizar la precisión. En el caso de los modelos clásicos $\boldsymbol{\theta^*}$ no incluye a β. Para este trabajo, la precisión depende del tiempo:

$$\tau_t = \frac{1}{\sigma_t^2} = \frac{1}{\sigma^2 t^\nu} = \frac{\tau}{t^\nu}.$$

La distribución *a priori* conjunta como producto de las *a priori* no informativas para los parámetros a, b, β, τ, ν. Las distribuciones *a priori* para a, b y τ son Gamma y para los

parámetros ν y β son distribución Uniforme, es decir:

$$\begin{aligned} f(a,b,\beta,\tau,\nu) &= f(a)f(b)f(\beta)f(\tau)f(\nu) \\ &\propto a^{p_1-1}b^{p_2-1}\tau^{p_3-1}\exp(-q_1a-q_2b-q_3\tau) \end{aligned}$$

donde $p_1, p_2, p_3, q_1, q_2, q_3$, son hiperparámetros de las distribuciones *a priori* gamma. En el caso de los modelos clásicos, la distribución *a priori* conjunta no incluye a $f(\beta)$.

La distribución *a posteriori* para el vector de parámetros $\boldsymbol{\theta} = (a, b, \beta, \nu, \tau)$ es:

$$\begin{aligned} f(\boldsymbol{\theta} \mid \boldsymbol{t}, \boldsymbol{V}) \propto a^{p_1-1}b^{p_2-1}\tau^{p_3-1+n/2}\prod_{t=1}^{n} t^{-\nu/2} \\ \times \exp\left[-\frac{\tau}{2}\sum_{t=1}^{n} t^{-\nu}(V_t - M(\boldsymbol{\theta^*}, t))^2 - (q_1a + q_2b + q_3\tau)\right]. \end{aligned} \tag{3.2}$$

Como estimador bayesiano de $\boldsymbol{\theta}$ se tomó la media posterior. Es decir se utilizó la función de pérdida cuadrática:

$$L(\widehat{\theta}, \theta) = (\widehat{\theta} - \theta)^2.$$

Nuevamente para los modelos clásicos, $\boldsymbol{\theta}$ no incluye a β al igual que la *a priori* conjunta.

3.3. Estudio de simulación

Como la distribución *a posteriori* (3.2) no es una distribución conocida, no es sencillo obtener teóricamente los estimadores bayesianos es decir la media de la distribución *a posteriori*. Por ello se realiza un estudio de simulación para aproximar la distribución *a posteriori*, bajo distintas condiciones.

3.3.1. Datos

Por la dificultad de conseguir datos reales se generaron dos casos mediante el siguiente mecanismo:

Caso 1:

$$V_t^1 = \left[\frac{b}{a} - \left(\frac{b}{a} - V_0^{1-\beta_1}\right)e^{-a(\beta_1-1)t}\right]^{1/(1-\beta_1)} + \varepsilon_t^1 t^2$$

con $\varepsilon_t^1 \sim N(0, 0.01)$, $a = 0.5$, $b = 0.04$, $\beta_1 = 1.4$ y $t = 1, \ldots, 40$.

Caso 2:

$$V_t^2 = \left[\frac{b}{a} - \left(\frac{b}{a} - V_0^{1-\beta_2}\right) e^{-a(\beta_2 - 1)t}\right]^{1/(1-\beta_2)} + \varepsilon_t^2 t$$

con $\varepsilon_i^2 \sim N(0, 0.16)$, $a = 0.5$, $b = 0.04$, $\beta_2 = 1.6$ y $t = 1, \ldots, 40$.

En el Caso 1 los datos favorecen al modelo de Gompertz porque $\beta_1 = 1.4$ que es más cercano a $\beta = 1$ que a $\beta = 2$. Por otro lado, en el Caso 2 los datos favorecen al modelo Logístico, porque $\beta_2 = 1.6$ que es más cercano a $\beta = 2$ (Figura 3.1). En ambos casos el valor inicial fue $V_0 = 1$.

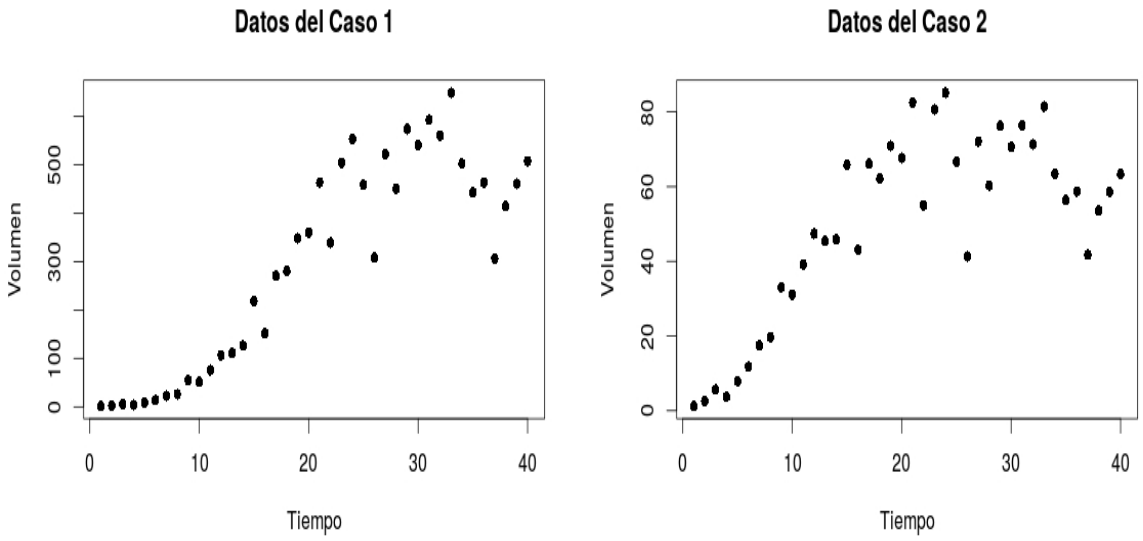

Figura 3.1: *Ejemplo de datos generados para los dos casos.*

3.3.2. Estimación

La distribución *a posteriori* (3.2) se obtuvo mediante métodos Monte Carlo via Cadenas de Markov (MCMC). Se utilizó JAGS[1] (Plummer, 2003), mediante los paquetes *rjags* (Plummer, 2015), *R2jags* (Su y Yajima, 2015) y *R2WinBUGS* (Sturtz et al., 2005) de *R* (R Core Team, 2015a). Se especificó un modelo en codigo BUGS. Las *a priori* fueron para a una $Gamma(0.001, 0.001)$, τ una $Gamma(0.001, 0.001)$, b una $Gamma(0.01, 0.01)$, ν una $U(0, 3)$ y para β una $U(1.1, 9)$, se inicia en 1.1 porque $\beta > 1$ en el MLG.

[1]Software para el análisis estadístico de modelos Bayesianos mediante MCMC

Se generaron 3 cadenas de Markov para cada modelo, con 500000 iteraciones de las cuales 250000 iteraciones se descartaron, con un adelgazamiento de 50. Para a, b y β se tomaron los estimadores calculados mediante mínimos cuadrados ponderados como valores iniciales, que se obtuvieron con la función *nlsLM* {*minpack.lm*} (Elzhov et al., 2013) en R, mientras que para los parámetros ν y τ, los valores iniciales fueron $\nu^0 = \{0.3, 0.5, 0.8\}$ y $\tau^0 = 1$.

En cada iteración se guardó el estimador de cada parámetro, los intervalos creibles del 95 %. Para verificar la convergencia, se usó el factor potencial de reducción de escala (*Rhat* denotado así por el software R) de cada estimador en cada modelo.

3.3.3. Comparación

Los FB se usaron para calcular el porcentaje de evidencia a favor del modelo m_1 contra el modelo m_2.

Otra forma de comparar los modelos es mediante el $EN_{n,d}$, el $ER_{n,d}$ y la $S_{n,d}$. Se calcularon con los primeros 30 volumenes, llamado grupo de *aprendizaje*. Este grupo se utilizó para el ajuste y los restantes se utilizaron para el calculo de estas medidas. Los errores normalizados en cada iteración se utilizaron para calcular $S_{n,d}$ para cada modelo, despues se calculó la media de las puntuaciones, y similarmente para los errores relativos.

Para calcular de los criterios de información, se utilizó la devianza producida por la función *jags* y con ella se obtuvo el AIC, BIC y DIC.

La SCP fueron calculados mediante la expresión (2.15).

Capítulo 4

Resultados

En este capítulo se muestran los resultados obtenidos de la simulación y la aplicación con datos experimentales. Se considera la notación ML, MG, MB y MLG como los estimadores bayesianos para los modelos Logístico, Gompertz, Bertalanffy y Logístico Generalizado.

4.1. Simulación

Estimación y convergencia

Para los dos conjuntos de datos generados, el ML y el MG tuvieron una buena convergencia, ya que los *Rhat* son iguales a 1.00, mientras que en el MB y el MLG, un 5 % y 4 % respectivamente fueron mayores a 1.5. Para el Caso 1, el mayor fue de 3.01 para el MB y 8.7 para el MLG. Para el Caso 2, el MB tuvo una buena convergencia con un $Rhat = 1.1$ como máximo y para el MLG el 2 % de los *Rhat* fue malo con un máximo de 4.51.

Criterios de información

En el Cuadro (4.1) los criterios de información indican que el MLG para el Caso 1 es el mejor, ya que tiene los menores valores de los criterios ($DIC = 262$, $AIC = 266$ y

Cuadro 4.1: *Estimadores, intervalos creíbles y criterios de información para el Caso 1.*

Modelos	Parámetros	Estimador bayesiano	95 % IC	DIC	AIC	BIC
ML	$a = 0.50$	0.3772	[0.3609, 0.3940]	270	273	304
	$b = 0.04$	0.0008	[0.0007, 0.0009]			
	$\nu = 4.00$	2.8173	[2.3735, 2.9948]			
	$\sigma^2 = 0.01$	0.4813	[0.1713, 1.5814]			
MG	$a = 0.50$	0.6418	[0.5984, 0.6868]	272	275	305
	$b = 0.04$	0.0950	[0.0860, 0.1043]			
	$\nu = 4.00$	2.7852	[2.2917, 2.9934]			
	$\sigma^2 = 0.01$	0.5680	[0.1836, 2.0086]			
MB	$a = 0.50$	0.9979	[0.7970, 1.2338]	298	298	328
	$b = 0.04$	0.0425	[0.0009, 0.0937]			
	$\nu = 4.00$	2.6956	[2.0235, 2.9905]			
	$\sigma^2 = 0.01$	1.8390	[0.4133, 8.1935]			
MLG	$a = 0.50$	0.4952	[0.3841, 0.6982]	262	266	304
	$b = 0.04$	0.0514	[0.0055, 0.1844]			
	$\beta = 1.40$	1.5314	[1.2459, 2.0976]			
	$\nu = 4.00$	2.8423	[2.4518, 2.9956]			
	$\sigma^2 = 0.01$	0.4788	[0.1148, 1.8035]			

$BIC = 304$), seguido del ML y MG. En el Caso 2, los criterios de información indican que el ML ($DIC = 186$, $AIC = 190$ y $BIC = 220$) es el mejor modelo, seguido del MLG y MG. En ambos casos el MB es el que tiene mayores valores de los criterios (Cuadro 4.2).

Factor bayes

Los porcentajes de casos en que la evidencia estuvo a favor de un modelo m_1 (renglón) vs m_2 (columna) mediante el factor bayes se muestran en el Cuadro (4.3). Nuevamente el MLG es mejor comparado con los modelos clásicos, para el Caso 1. Por el contrario en el Caso 2, el MLG es malo en comparación con los modelos clásicos. En ambos casos, el ML

es mejor que el MG. Además el MB es malo comparado con el Logístico y Gompertz.

Cuadro 4.2: *Estimadores, intervalos creíbles y criterios de información para el Caso 2.*

Modelos	Parámetros	Estimador bayesiano	95 % IC	DIC	AIC	BIC
ML	$a = 0.50$	0.4135	[0.3880, 0.4399]	186	190	220
	$b = 0.04$	0.0064	[0.0057, 0.0073]			
	$\nu = 2.00$	1.9510	[1.2194, 2.5635]			
	$\sigma^2 = 0.16$	0.4790	[0.0570, 2.2317]			
MG	$a = 0.50$	0.6758	[0.6129, 0.7458]	194	198	228
	$b = 0.04$	0.1556	[0.1377, 0.1752]			
	$\nu = 2.00$	1.7444	[0.9379, 2.4517]			
	$\sigma^2 = 0.16$	1.1699	[0.1044, 5.7580]			
MB	$a = 0.50$	1.2796	[1.0597, 1.5223]	206	210	240
	$b = 0.04$	0.2872	[0.2180, 0.3606]			
	$\nu = 2.00$	1.3856	[0.5269, 2.2186]			
	$\sigma^2 = 0.16$	4.0952	[0.2924,19.8376]			
MLG	$a = 0.50$	0.4925	[0.3700, 0.7411]	187	190	228
	$b = 0.04$	0.0530	[0.0030, 0.2205]			
	$\beta = 1.60$	1.9337	[1.3864, 3.2411]			
	$\nu = 2.00$	1.9809	[1.2472, 2.5892]			
	$\sigma^2 = 0.16$	0.4363	[0.0505, 2.0686]			

Cuadro 4.3: *Porcentaje de casos en que el factores bayes indica evidencia a favor*

	Caso 1			Caso 2		
vs	ML	MG	MB	ML	MG	MB
MG	36.1 %			13.1 %		
MB	10.3 %	0.0 %		4.9 %	0.0 %	
MLG	86.8 %	86.5 %	86.8 %	25.8 %	25.5 %	25.6 %

Suma de cuadrados de predicción

Los resultados de las SCP's, indican que el peor modelo para el ajuste es el MLG (Figura 4.1). Las medias de las SCP's para el Caso 1, fueron de 2053 para el ML, 2029 para el MG, 4142 para el MB y 30023 para el MLG, lo que indica que el ML es el mejor. Para el Caso 2 nuevamente el MLG es el peor (SCP=986) contra el ML con $SCP = 50$, MG con $SCP = 58$ y MB $SCP = 72$, que indica que el ML es ligeramente mas adecuado al ajustar los datos seguido del MG.

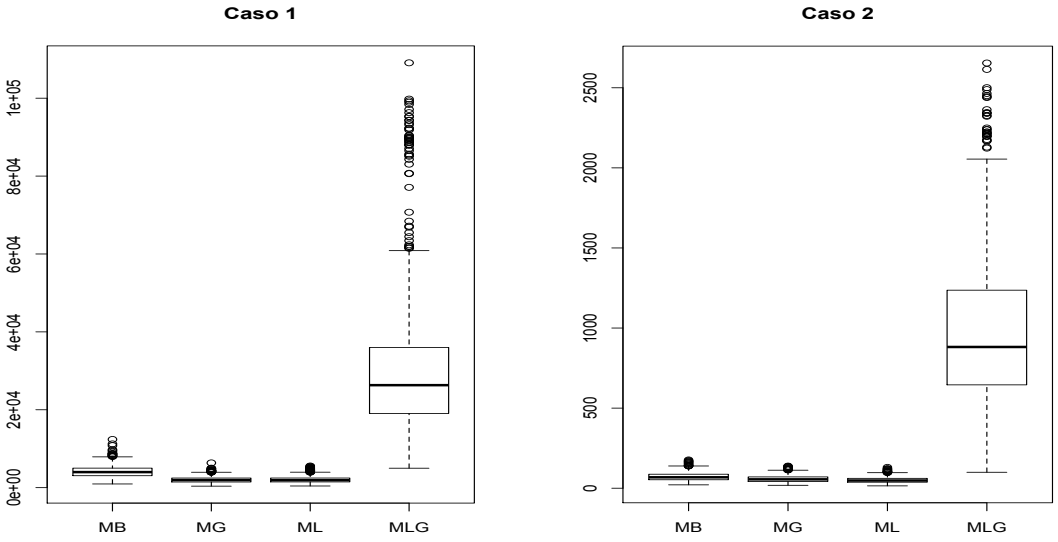

Figura 4.1: *Gráfica de la SCP de la simulación*

Predicción

Para la $S_{n,d}$, la interpretación es sencilla, por ejemplo el ML obtuvo una puntuación de 0.956 y 0.998 en el Caso 1 y Caso 2 respectivamente, esto quiere decir que con el ML se obtuvo el 95.6 % y 99.8 % de aciertos. El modelo que predice mejor, considerando $S_{n,d}$ para el Caso 1, es el ML y el peor es el MLG con solo 35.3 % de aciertos. Para el Caso 2, los estimadores bayesianos de los modelos clásicos son muy buenos en la predicción teniendo más del 99 % de aciertos. El MLG vuelve a ser el menos predictivo, porque tiene el 77 % de aciertos.

Cuadro 4.4: *Medias de los errores relativos y puntuación de predicción*

Modelo	Caso 1		Caso 2	
	$S_{30,10}$	$ER_{30,10}$	$S_{30,10}$	$ER_{30,10}$
ML	**0.956**	**0.248**	**0.998**	**0.185**
MG	0.896	0.416	**0.997**	0.240
MB	0.618	1.005	**0.992**	0.306
MLG	0.353	0.591	0.779	0.597

Con respecto al error relativo, entre más pequeño mejor es la predicción. Para el Caso 1 el ML es el mejor modelo al tener un error relativo de 0.248, le sigue el MG y el MLG. Para el Caso 2, nuevamente el ML es el mejor modelo con un error relativo de 0.185, en este caso el peor modelo es el MLG.

4.2. Ejemplo

En esta sección se presenta los resultados obtenidos con datos experimentales. Los datos fueron tomados de Torres (1990) quien realizó un cultivo de esferoides multicelulares (EM) en base a la línea célular MCF-7 procedente de un cáncer de mama humano. Con la finalidad de observar cambios en el desarrollo de los EM, realizó una división en su cultivo, una parte se cultivaron en presencia de agentes hormonales (estrógenos). La acción positiva de los estrógenos que ejercieron en el desarrollo de los EM, se manifestó principalmente en el desarrollo estructural y los cambios proliferativos (Torres, 1990). Además realizó técnicas de quimioterapia y radioterapia sobre los EM, entre otro tipos de análisis químicos.

También ajustó los modelos empíricos clásicos y el modelo exponencial a los datos obtenidos. El ajuste que realizó fue individual y en grupos. Los resultados fueron que el MG en primer lugar y el MB en segundo lugar predicen con notable exactitud los cambios de volumen que experimentan los EM.

Datos

El crecimiento de los esferoides se evaluó a los tiempos indicados en cada experimento

mediante la medida de 2 diámetros ortogonales. Debido a que la organización de las células cuando crecen en suspensión adoptan una forma semejante al de una esfera, el cálculo del volumen fue con la fórmula:

$$V = \frac{Dd^2\pi}{6}$$

donde D es el diámetro mayor y d el diámetro menor. Se utilizaron 3 EM que se cultivaron de manera individual con las mismas condiciones de cultivo (Tablas 6, 7 y 8 de Torres, 1990). Estos se presentan en la Figura (4.2) donde se observa la presencia de heterocedasticidad. El volumen de cada esferoide al tiempo t para el individuo i, $i = 1, 2, 3$, es denotado por V_t^i. La unidad de medida utilizada es el micrómetro cúbico.

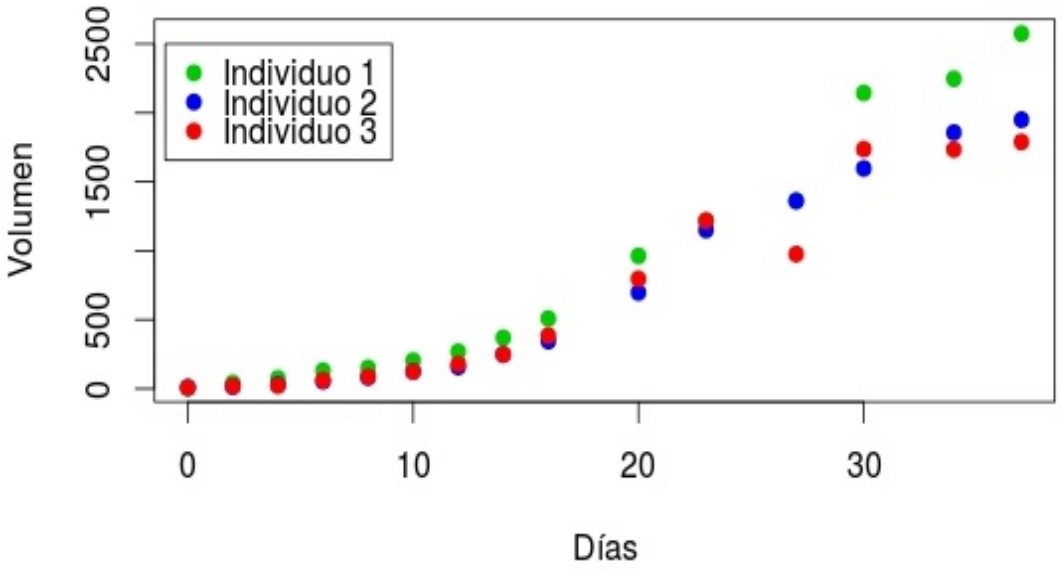

Figura 4.2: *Datos experimentales*

Ajuste y comparación

Se realizaron dos ajustes a los datos experimentales. En el primer ajuste se usan los 42 datos para obtener los estimadores bayesianos. En el segundo ajuste, se usan los primeros 10 volumenes de cada EM, por lo tanto se usa un total de 30 volumenes para obtener los modelos estimados y despues predecir las observaciones a los tiempos restantes. A este conjunto de 30 volumenes se denominó *grupo de aprendizaje*.

Se ajustaron los ML, MG y MB mediante la función *jags*. Las distribuciones *a priori* para a, b y τ fue una $Gamma(0.001, 0.001)$ y para ν fue $U(0, 3)$. Se generaron 3 cadenas de Markov para cada modelo, con 200000 iteraciones de las cuales 100000 se descartaron, con un adelgazamiento de 100. Los estimadores calculados por mínimos cuadrados ponderados se tomaron como valores iniciales para los parámetros a y b, mientras que para los parámetros ν y τ, los valores iniciales fueron $\nu^0 = \{0.3, 0.5, 0.8\}$ y $\tau^0 = 1$.

Las cadenas de Markov tuvieron una buena convergencia, el *Rhat* fue igual 1.00 en los tres modelos. Además el comportamiento de las trazas y los diagnósticos de convergencia de Gelman-Rubin y Geweke indican una buena convergencia de las cadenas (Anexo 1). Los criterios de información no dan evidencia a favor de un único modelo. Los modelos de Gompertz y Bertalanffy son mejores comparados con el modelo Logístico (Cuadro 4.5).

Cuadro 4.5: *Estimadores, intervalos creíbles y criterios de información, para los 42 datos experimentales.*

Modelos	Parámetros	Estimador bayesiano	95 % IC	DIC	AIC	BIC
ML	a	0.3748	[0.3469,0.4071]	544	548	572
	b	0.0002	[0.0002, 0.0003]			
	ν	1.9116	[1.3769, 2.4148]			
	σ^2	213.8785	[45.3060, 745.4202]			
MG	a	0.7463	[0.6925, 0.8037]	515	519	543
	b	0.0946	[0.0863, 0.1036]			
	ν	1.8003	[1.3308, 2.2255]			
	σ^2	133.4637	[36.2044, 410.0013]			
MB	a	1.6736	[1.4879, 1.8654]	514	518	542
	b	0.0847	[0.0591, 0.1100]			
	ν	2.0706	[1.6006, 2.5032]			
	σ^2	65.7788	[17.3304, 204.7728]			

Para el ajuste con 30 observaciones, las cadenas de Markov tuvieron una buena convergencia según los diagnósticos utilizados (Anexo 2). Los estimadores son un poco mayores

a los del ajuste con los 42 datos en cada modelo. Por ejemplo, en el ML el estimador de a para el ajuste con los 42 datos es 0.3748 y para el ajuste con 30 observaciones es de 0.4108, el estimador de b para el ajuste con los 42 datos es de 0.0002 mientras que para el ajuste con 30 observaciones es de 0.0003 (Cuadro 4.6). También se observa que los criterios de información indican que el MB y MG son mejores pues tienen el menor DIC, AIC y BIC.

Al usar las 42 observaciones y de acuerdo a la SCP, el MG es el que mejor ajuste tiene ($SCP = 26326.63$), seguido del MB ($SCP = 30742.87$). Sin embargo, en el ajuste con las 30 observaciones el MB es el mejor con un SCP de 3345.07 (Cuadro 4.7).

Para el ajuste con los 42 datos, el factor bayes indica que el MG y el MB son mejores que el ML y el MB ligeramente mejor que el MG ($FB = 1.8846$) (Cuadro 4.8). Para el ajuste con 30 observaciones, el factor bayes indica que el MG y el MB son mejores que el ML y el MB es definitivamente mejor que el MG (Cuadro 4.8).

Cuadro 4.6: *Estimadores, intervalos creíbles y criterios de información, para los 30 datos experimentales.*

Modelos	Parámetros	Estimador bayesiano	95 % IC	DIC	AIC	BIC
ML	a	0.4108	[0.3819, 0.4430]	355	359	383
	b	0.0003	[0.0002, 0.0005]			
	ν	1.1985	[0.3201, 2.0287]			
	σ^2	810.07	[82.1700, 3940.0076]			
MG	a	0.7005	[0.6252, 0.7802]	337	341	365
	b	0.0852	[0.0706, 0.1002]			
	ν	0.9540	[0.1760, 1.7047]			
	σ^2	697.45	[87.2251, 2897.5767]			
MB	a	1.2476	[1.2169, 1.2764]	328	333	357
	b	0.000001	[0, 0.00001]			
	ν	1.0670	[0.3190, 1.7663]			
	σ^2	404.3791	[57.4677, 1742.0702]			

Cuadro 4.7: *Las SCP para el ejemplo*

Modelo	Ajuste con 42 observaciones	Ajuste con 30 observaciones
ML	50203.73	7326.64
MG	26326.63	3888.62
MB	30742.87	3345.07

Cuadro 4.8: *Factor bayes.*

	Ajuste con los 42 observaciones			Ajuste con 30 observaciones		
vs	ML	MG	MB	ML	MG	MB
MG	172645.1	1		89.4	1	
MB	325372.3	1.9	1	3533670.8	1691.3	1

Predicción

Se tomaron los modelos estimados con las 30 observaciones para predecir observaciones a un tiempo mayor, se utilizaron los 4 últimos volumenes de cada EM para medir la predicción, calculandose los $EN_{10,d}$ y $ER_{10,d}$ con $d = 1, 2, 3, 4$.
En los tres modelos los EN's fueron todos mayores a 3, por lo tanto $S_{10,4} = 0$. Los valores de los $ER_{10,4}$ indican que el MG es mejor, seguido del ML y MB. Las dos medidas, el $EN_{10,4}$ y $ER_{10,4}$,indican que el peor modelo para predecir es el MB.

Cuadro 4.9: *Errores relativos y puntuación de predicción*

Modelo	$S_{10,4}$	$ER_{10,4}$
ML	0	0.9114
MG	0	0.8975
MB	0	1.1021

La distribución predictiva *a posteriori* se usó para obtener prediciones futuras e intervalos creíbles. En el ajuste con los 42 observaciones, para los tres modelos las observaciones están contenidas en su mayoria en los IC's. Sin embargo son mejores los resultados de los modelos de Gompertz y Bertalanffy por tener una varianza más pequeña, es decir hay más precisión (Figura 4.3).

Al realizar el ajuste con 30 observaciones, nuevamente se observa un buen ajuste para los tres modelos al estar contenidas las observaciones en los IC's. El MB es el mejor modelo al tener intervalos creibles con varianza pequeña, seguido del MG (Figura 4.4).

También, se presentan las curvas predictivas de los tres modelos, Logístico, Gompertz y Bertalanffy. El MB es el peor modelo al intentar predecir, ya que ningúna de las observaciones a predecir esta contenida en el IC. Si bien en el ML algunas observaciones a predecir están contenidas en los IC, el MG es el mejor al contener las observaciones a predecir en su mayoria.

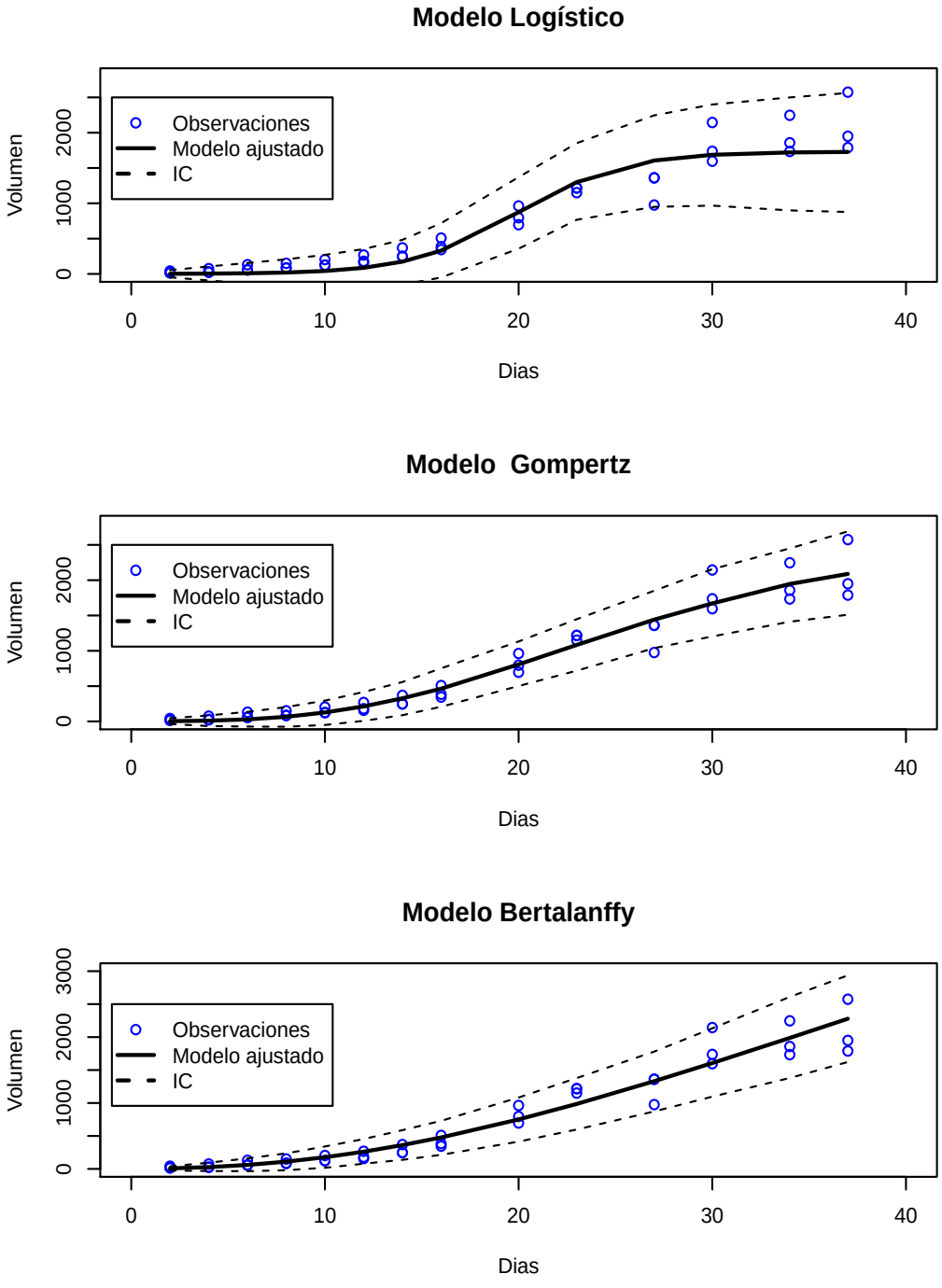

Figura 4.3: *Modelos ajustados e intervalos creibles*

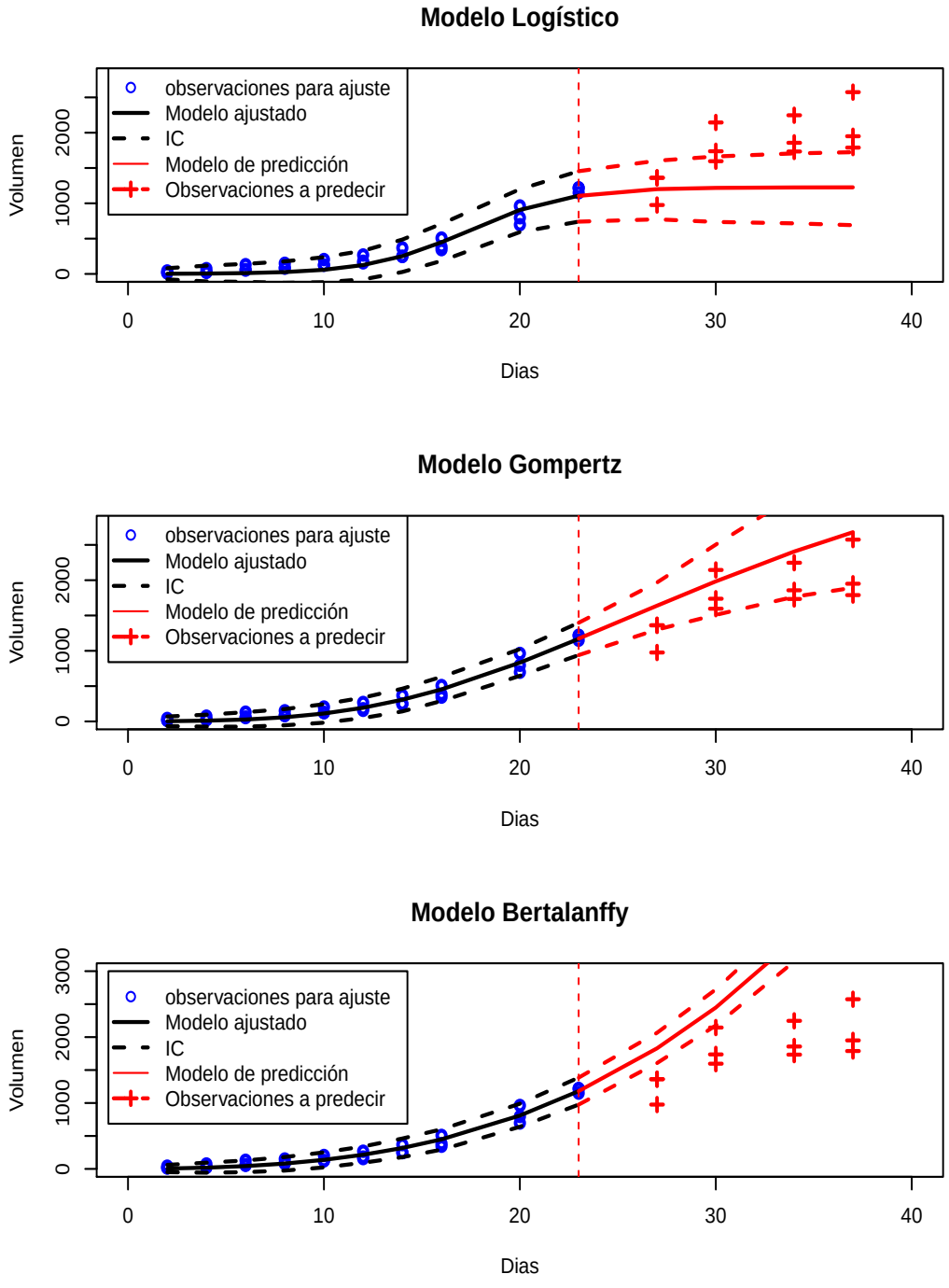

Figura 4.4: *Modelos ajustados e intervalos creibles*

Capítulo 5

Conclusiones

En la literatura se encontró que los modelos empíricos presentados en este trabajo han sido utilizados para describir el crecimiento tumoral utilizando estimadores clásicos como mínimos cuadrados ordinarios y generalizados, incluso se ha atacado el problema de heterocedasticidad usando mínimos cuadrados ponderados. Hasta el momento no se habían utilizado estimadores bayesianos para el análisis de estos modelos.

En este trabajo se usó la estimación bayesiana en los modelos Logístico, Gompertz, Bertalanffy y Logístico Generalizado que se usan para describir el comportamiento de crecimiento de tumores bajo el problema de varianza no constante. La distribución *a priori conjunta* fue el producto de distribuciones no informativas, tales como la distribucion gamma y la distribucion uniforme. Los modelos se compararon mediante la suma de cuadrados de predicción y factor bayes, criterios de informacion como el AIC, BIC y DIC, error de prediccion relativo y normalizado.

La estadística bayesiana, además de proporcionar métodos para la estimación de parámetros, también proporciona herramientas para decidir que modelo es el mejor. En el Caso 1 el mejor modelo según los criterios de información y factor bayes fue el modelo Logístico Generalizado, seguido del modelo Logístico que fue el mejor según los errores de predicción y sumas de cuadrados de predicción. En el Caso 2 los criterios de información, el factor bayes y los errores de predicción indican que el mejor es el modelo Logístico. Como los

conjuntos de datos fueron generados mediante la solución del modelo Logístico Generalizado, se esperaba que éste fuera el mejor en ambos caso; sin embargo, esto no sucede en el Caso 2.

Al ajustar los modelos bayesianos con a datos experimentales, los modelos de Bertalanffy y Gompertz fueron los más adecuados de acuerdo a los criterios de información, suma de cuadrados de predicción y factor bayes para describir el comportamiento de los datos. Sin embargo, en la predicción de observaciones futuras el modelo de Gompertz fue el mejor.

Referencias

Akaike, H. (1973). lnformation theory and an extension of the maximum likelihood principle. *Proceeding of the Second International Symposium on Information Theory*.

Albert, J. (2009). *Bayesian computation with R*. Springer Science & Business Media.

Bajzer, Z., Marušić, M., y Vuk-Pavlović, S. (1996). Conceptual frameworks for mathematical modeling of tumor growth dynamics. *Mathematical and computer modelling*, 23(6):31–46.

Benzekry, S., Lamont, C., Beheshti, A., Tracz, A., Ebos, J. M., Hlatky, L., y Hahnfeldt, P. (2014). Classical mathematical models for description and prediction of experimental tumor growth. *PLoS Comput Biol*, 10(8):e1003800.

Berg, A., Meyer, R., y Yu, J. (2004). Deviance information criterion for comparing stochastic volatility models. *Journal of Business and Economic Statistics*, 22(1):107–120.

Cabrera, L., Herráez, J., Ángel, M., Cabrera, J. L., y Sánchez, Á. H. (2008). *Texto ilustrado de biología molecular e ingeniería genética: conceptos, técnicas y aplicaciones en ciencias de la salud*. Number Sirsi) a456227.

Dempster, A. P. (1997). The direct use of likelihood for significance testing. *Statistics and Computing*, 7(4):247–252.

Elzhov, T. V., Mullen, K. M., Spiess, A.-N., y Bolker, B. (2013). *minpack.lm: R interface to the Levenberg-Marquardt nonlinear least-squares algorithm found in MINPACK, plus support for bounds*. R package version 1.1-8.

Freyer, J. P. y Sutherland, R. M. (1986). Regulation of growth saturation and development of necrosis in emt6/ro multicellular spheroids by the glucose and oxygen supply. *Cancer research*, 46(7):3504–3512.

Gelfand, A. E. y Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 501–514.

Gelman, A., Carlin, J. B., Stern, H. S., y Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Taylor & Francis.

Gelman, A. y Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472.

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

Geweke, J. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA.

Gilks, W. R., Richardson, S., y Spiegelhalter, D. J. (1996). Introducing markov chain monte carlo. *Markov chain Monte Carlo in practice*, 1:19.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of London*, 115.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Instituto Nacional de Estadística y Geografía (2016). Estadísticas a propósito del día mundial contra el cáncer (4 de febrero). `http://www.inegi.org.mx/saladeprensa/aproposito/2016/cancer2016_0.pdf`.

Kass, R. E. y Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430):773–795.

Kopelman, R. (1988). Fractal reaction kinetics. *Science*, 241(4873):1620–1626.

Marušić, M. y Bajzer, Z. (1993). Generalized two-parameter equation of growth. *Journal of mathematical analysis and applications*, 179(2):446–462.

Marušić, M., Bajzer, Z., Freyer, J., y Vuk-Pavlović, S. (1991). Modeling autostimulation of growth in multicellular tumor spheroids. *International journal of bio-medical computing*, 29(2):149–158.

Marušić, M., Bajzer, Ž., Freyer, J., y Vuk-Pavlović, S. (1994). Analysis of growth of multicellular tumour spheroids by mathematical models. *Cell proliferation*, 27(2):73–94.

Marušić, M. y Vuk-Pavlović, S. (1993). Prediction power of mathematical models for tumor growth. *Journal of Biological Systems*, 1(01):69–78.

Menchón, S. A. (2007). *Modelado de las diversas etapas del crecimiento del cáncer y de algunas terapias antitumorales*. Tesis doctoral, Universidad Nacional de Cordoba.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., y Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Newton, M. A. y Raftery, A. E. (1994). Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–48.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*, volume 698. John Wiley & Sons.

Organización Mundial de la Salud (2015). Nota descriptiva Num 297. `http://www.who.int/mediacentre/factsheets/fs297/es/`.

Organización Panamericana de la Salud (2015). Cáncer. `http://www.paho.org/hq/index.php?option=com_content&view=category&id=1866&layout=blog&Itemid=3904&lang=es`.

Piantadosi, S. (1985). A model of growth with first-order birth and death rates. *Computers and biomedical research*, 18(3):220–232.

Plummer, M. (2003). JAGS: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, page 125. Technische Universit at Wien Wien, Austria.

Plummer, M. (2015). *rjags: Bayesian Graphical Models using MCMC*. R package version 3-15.

Plummer, M., Best, N., Cowles, K., y Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11.

R Core Team (2015a). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team (2015b). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., y Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Sturtz, S., Ligges, U., y Gelman, A. (2005). R2winbugs: A package for running winbugs from r. *Journal of Statistical Software*, 12(3):1–16.

Su, Y.-S. y Yajima, M. (2015). *R2jags: Using R to Run 'JAGS'*. R package version 0.5-7.

Tanner, M. A. (1996). *Tools for statistical inference*, volume 3. Springer.

Torres, M. V. (1990). *Modelos tumorales en oncología: los esferoides multicelulares en el estudio del cáncer hormonodependiente.* Tesis doctoral, Universidad de Granada.

Vaidya, V. G. y Alexandro, F. J. (1982). Evaluation of some mathematical models for tumor growth. *International journal of bio-medical computing*, 13(1):19–35.

von Bertalanffy, L. (1957). Quantitative laws in metabolism and growth. *Quarterly Review of Biology*, pages 217–231.

Anexo 1

En esta sección se muestran las trazas y las distribuciones *a posteriori* de los parámetros para cada modelo, para el ajuste donde se utilizaron los 42 volumenes de los datos experimentales. También se presentan los gráficos de los criterios de convergencia de Gelman-Rubin y Geweke.

Trazas y distribuciones *a posteriori*

Se observa que en los tres modelos las trazas se mezclan bien. Además las distribuciones *a priori* no informativas propuestas son una buena elección, esto porque generan distribuciones *a posteriori* cerradas.

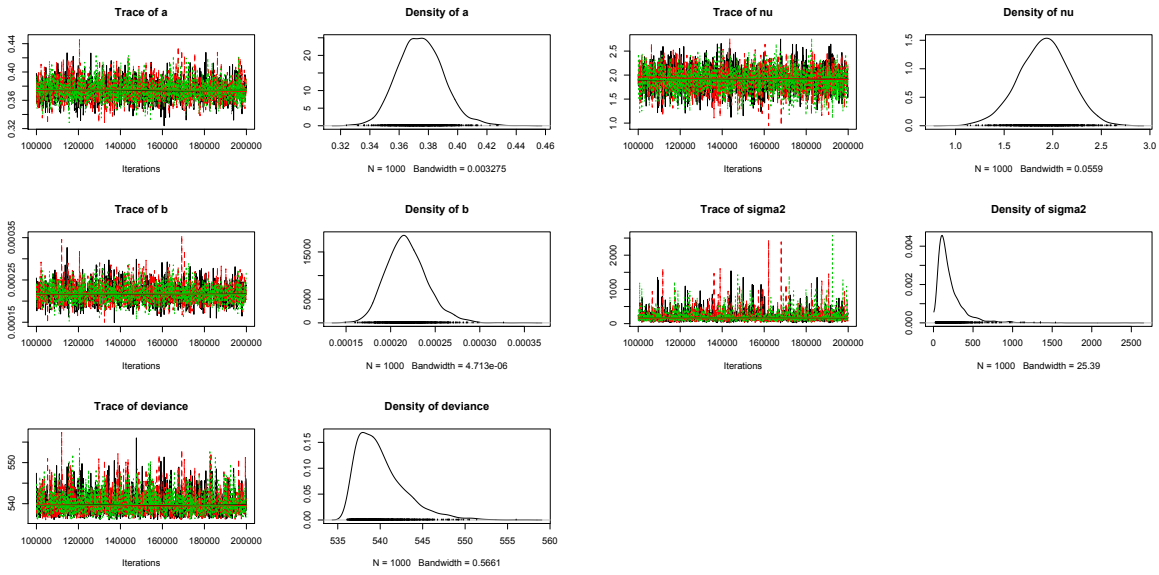

Figura 1: *Trazas y densidades del Modelo Logístico.*

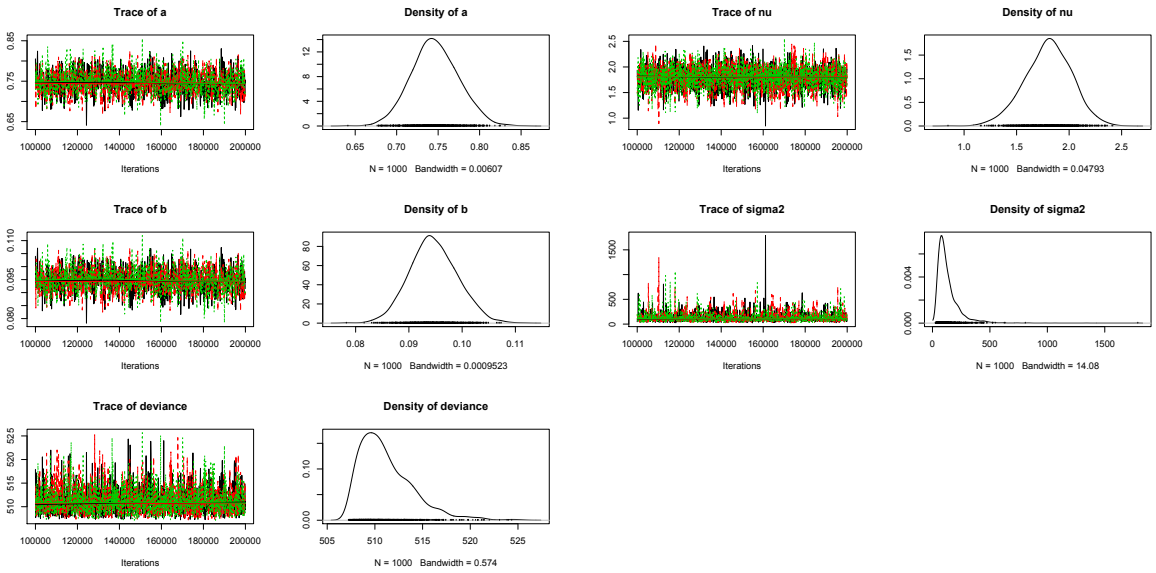

Figura 2: *Trazas y densidades del Modelo de Gompertz.*

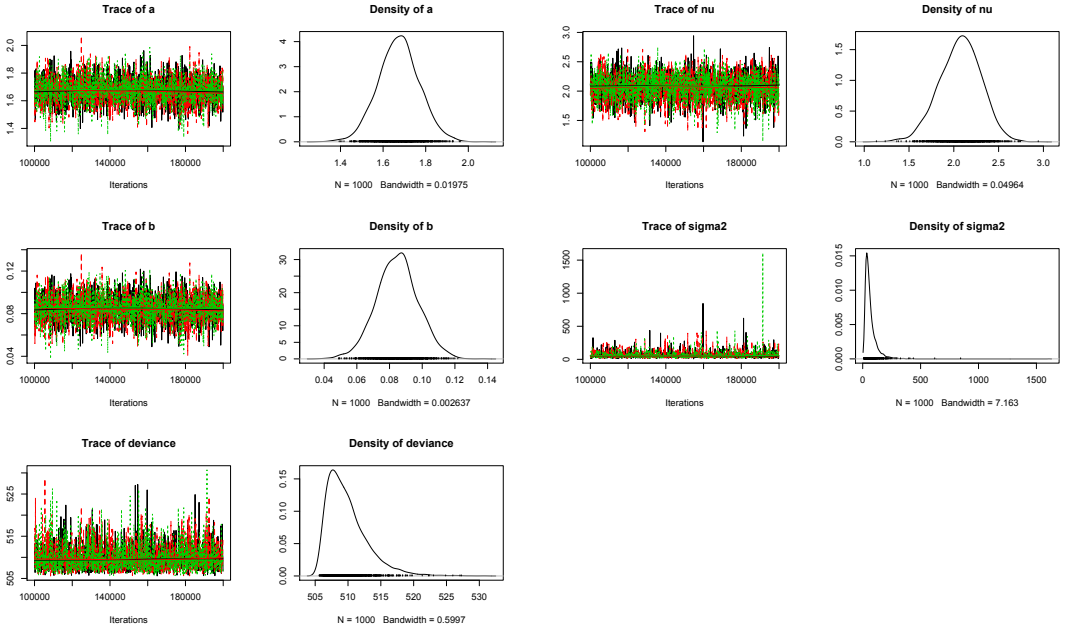

Figura 3: *Trazas y densidades del Modelo de Bertalanffy.*

Diagnóstico de Gelman-Rubin

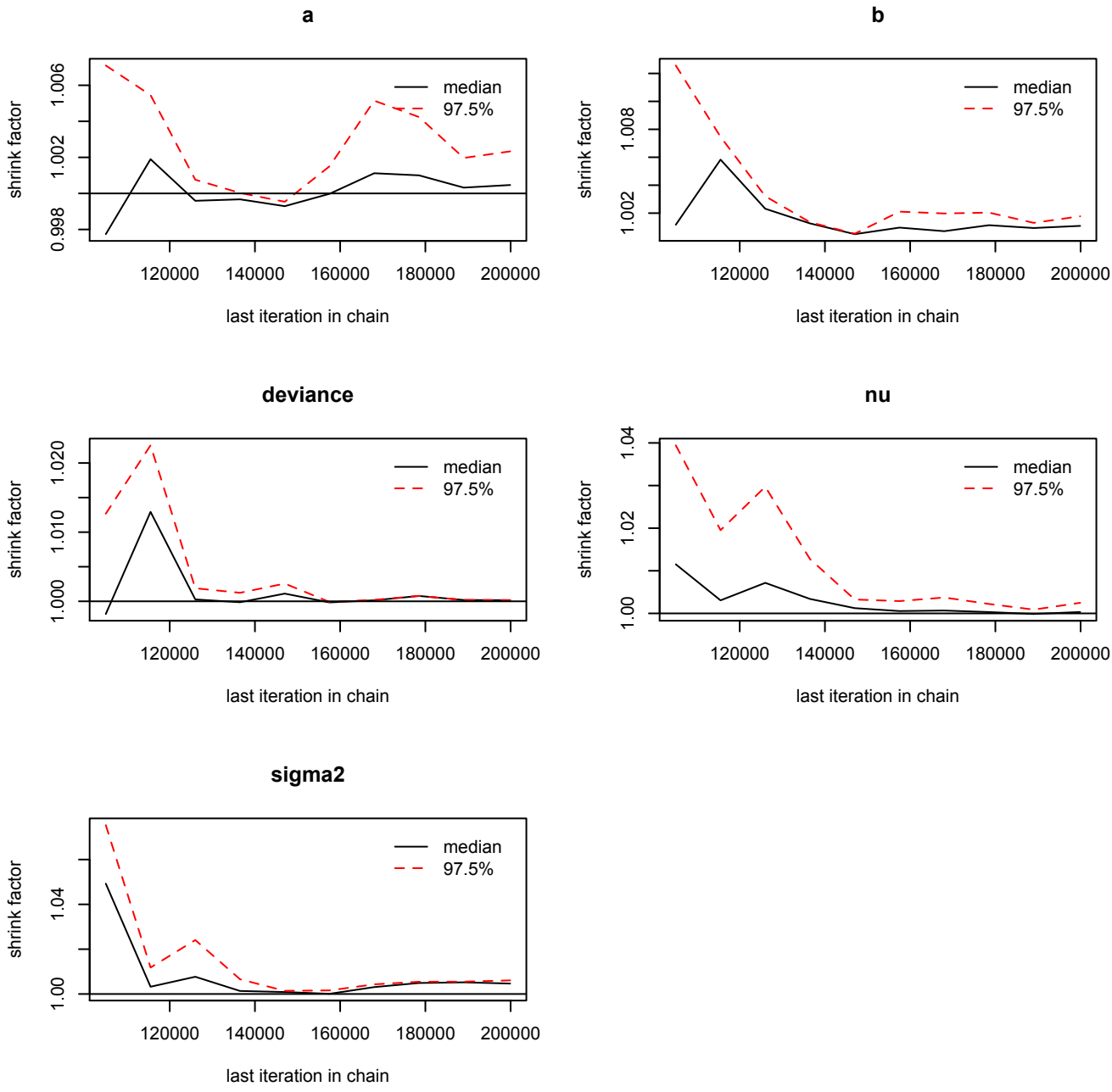

Figura 4: *Gráfica del diagnóstico de Gelman-Rubin del Modelo Logístico.*

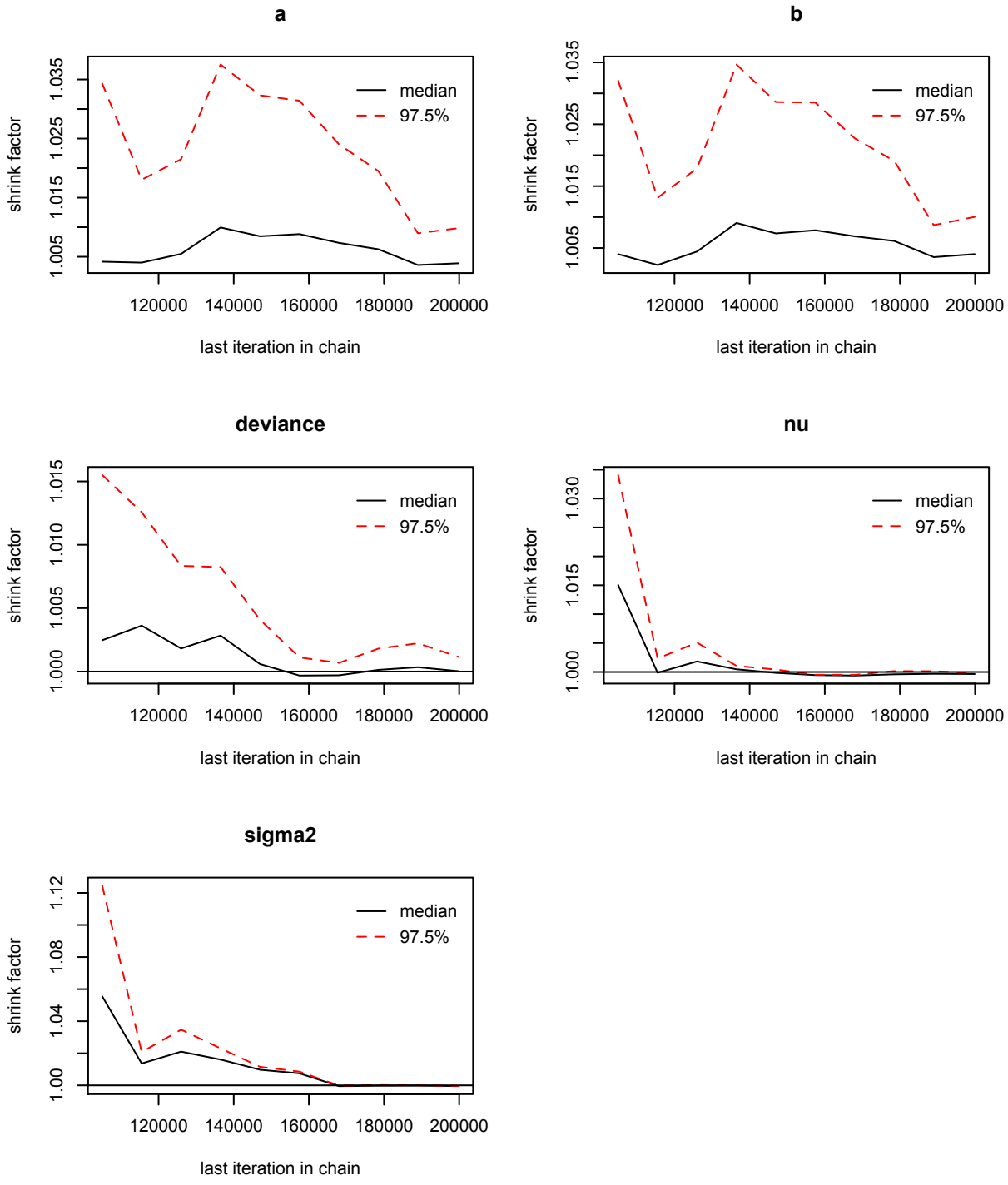

Figura 5: *Gráfica del diagnóstico de Gelman-Rubin del Modelo de Gompertz.*

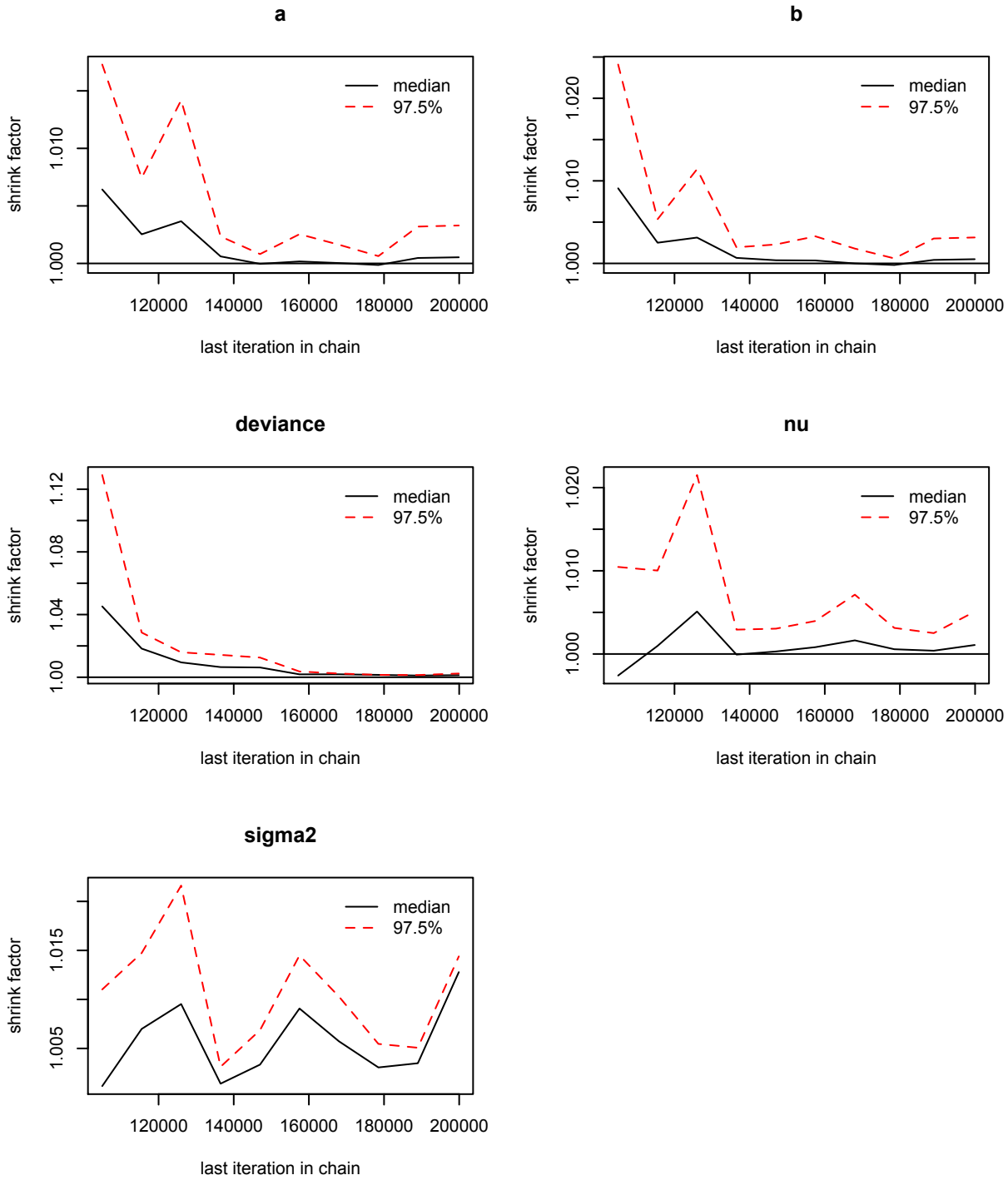

Figura 6: *Gráfica del diagnóstico de Gelman-Rubin del Modelo de Bertalanffy.*

Diagnóstico de Geweke

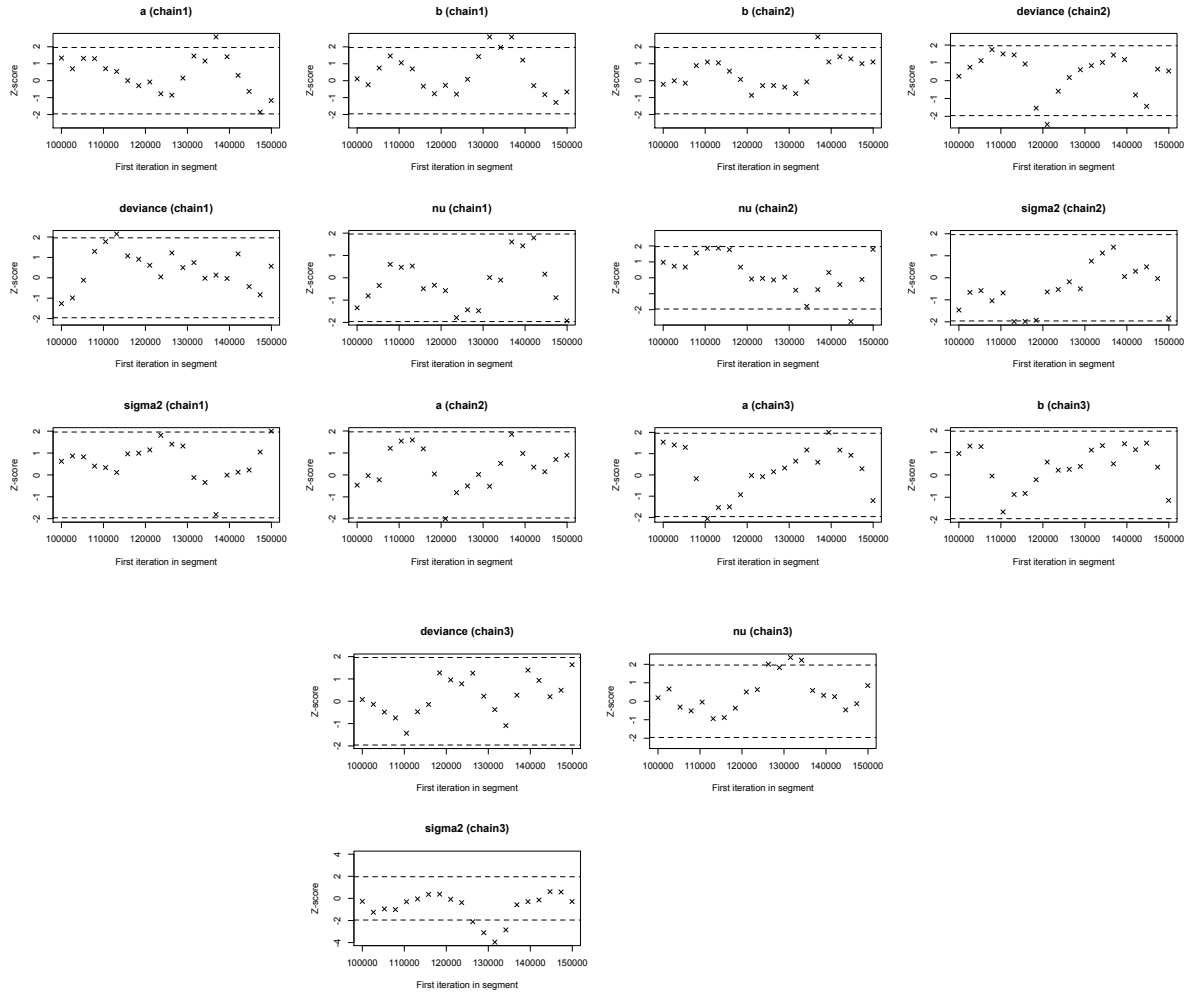

Figura 7: *Gráfica del diagnóstico de Geweke del Modelo Logístico.*

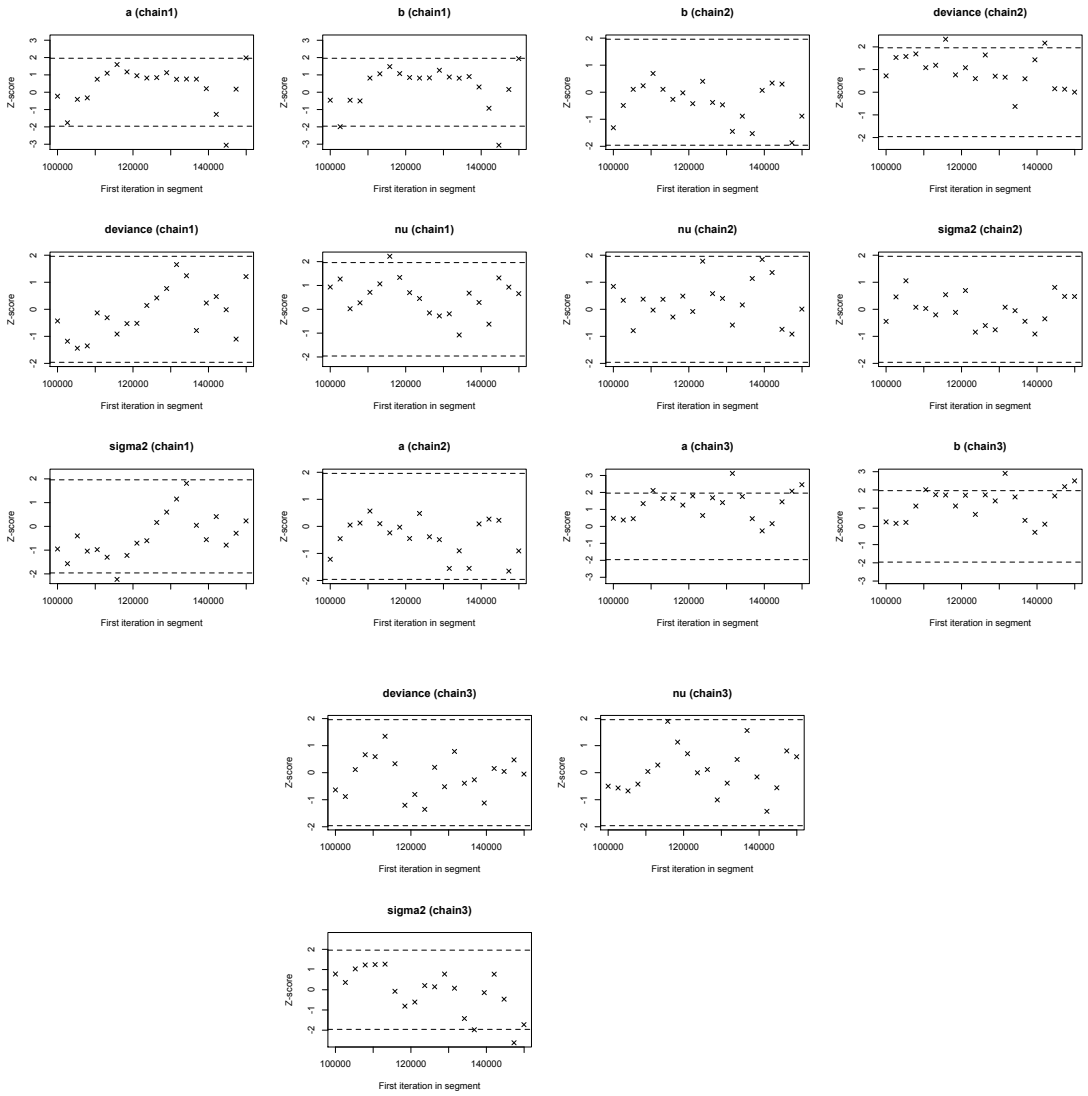

Figura 8: *Gráfica del diagnóstico de Geweke del Modelo de Gompertz.*

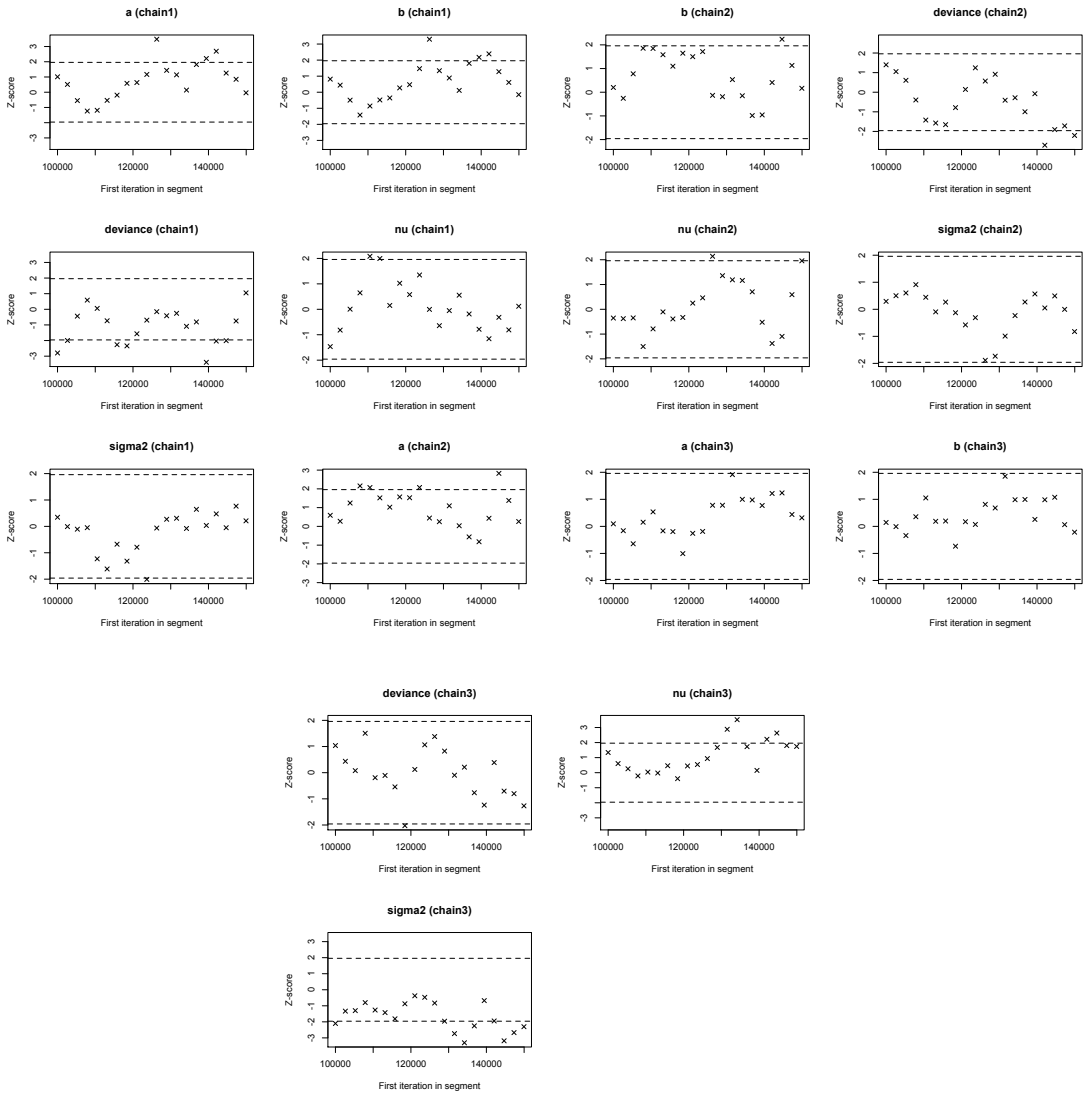

Figura 9: *Gráfica del diagnóstico de Geweke del Modelo de Bertalanffy.*

Anexo 2

En esta sección se muestran las trazas y las distribuciones *a posteriori* de los parámetros para cada modelo, para el ajuste donde se utilizaron los primero 10 volumenes por individuo de los datos experimentales. También se presentan los gráficos de los criterios de convergencia de Gelman-Rubin y Geweke.

Trazas y distribuciones *a posteriori*

Se observa que en los tres modelos las trazas se mezclan bien. Además las distribuciones *a priori* no informativas propuestas son una buena elección, esto porque generan distribuciones *a posteriori* cerrada.

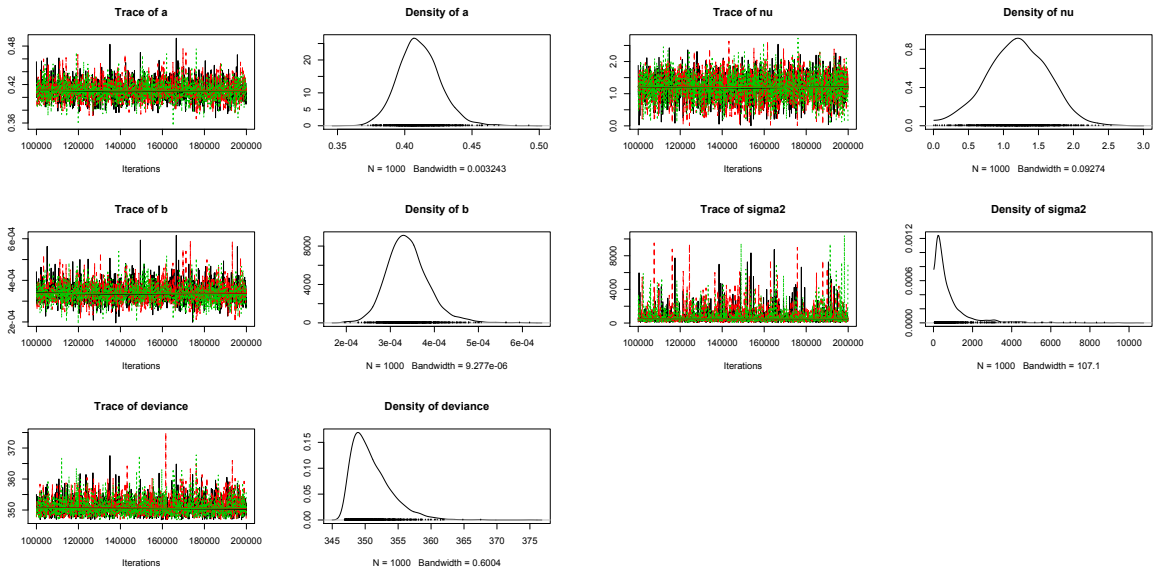

Figura 10: *Trazas y densidades del Modelo Logístico.*

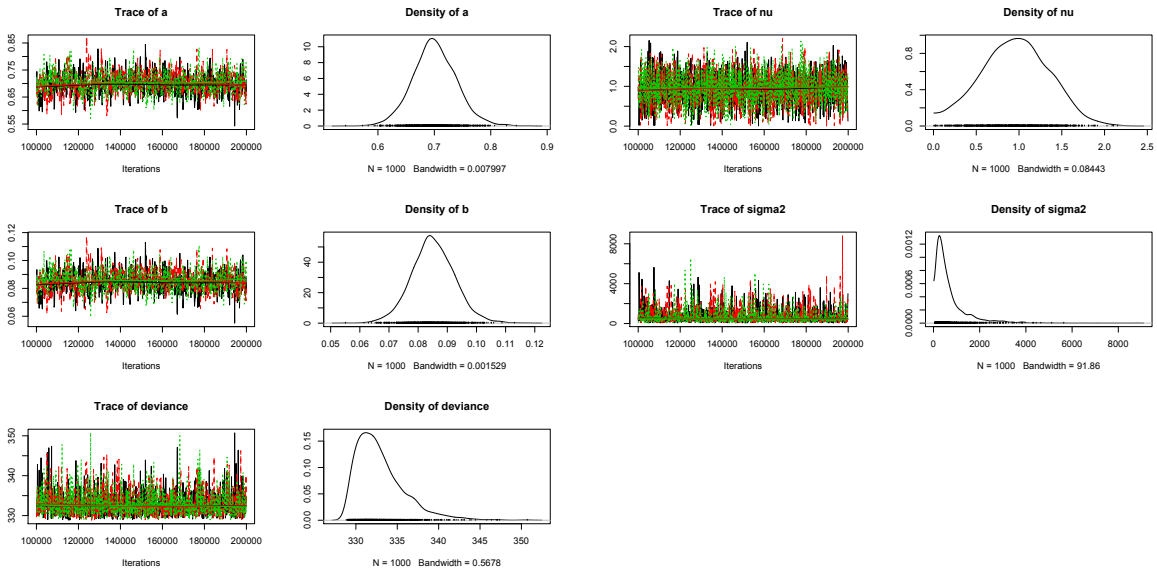

Figura 11: *Trazas y densidades del Modelo de Gompertz.*

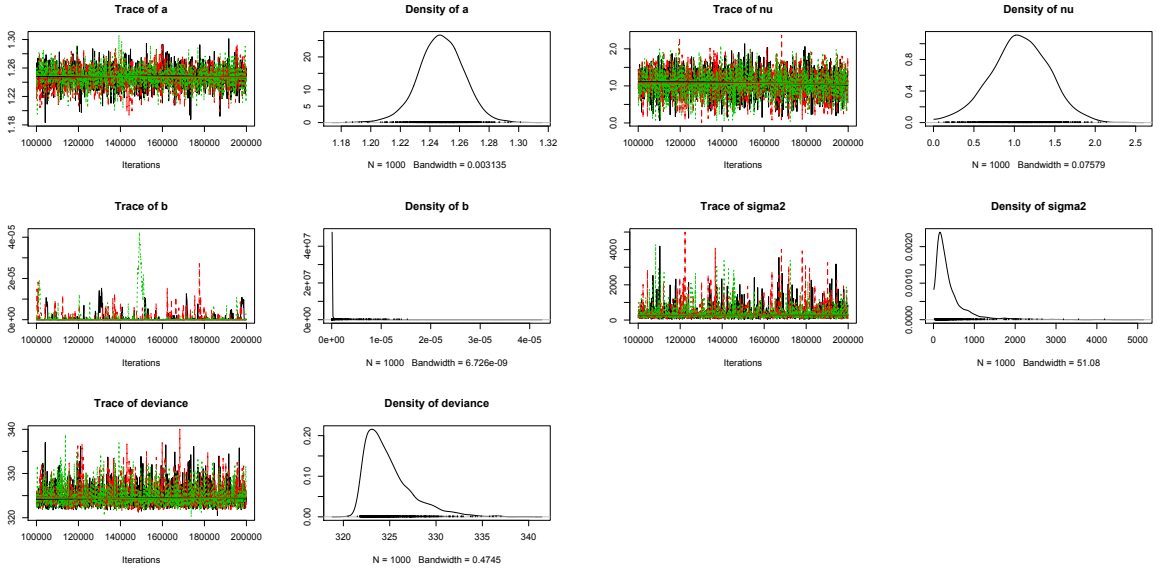

Figura 12: *Trazas y densidades del Modelo de Bertalanffy.*

Diagnóstico de Gelman-Rubin

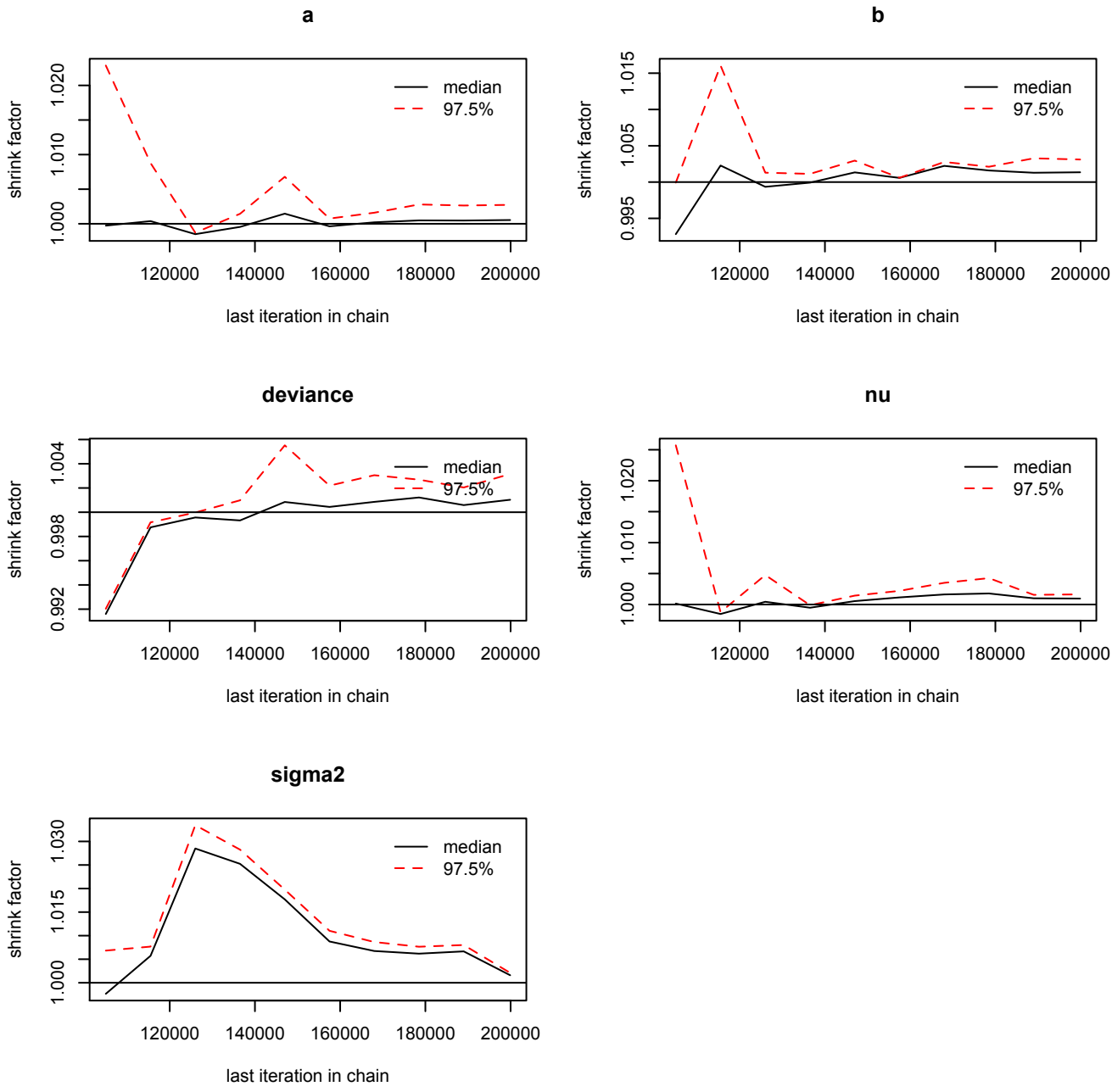

Figura 13: *Gráfica del diagnóstico de Gelman-Rubin del Modelo Logístico.*

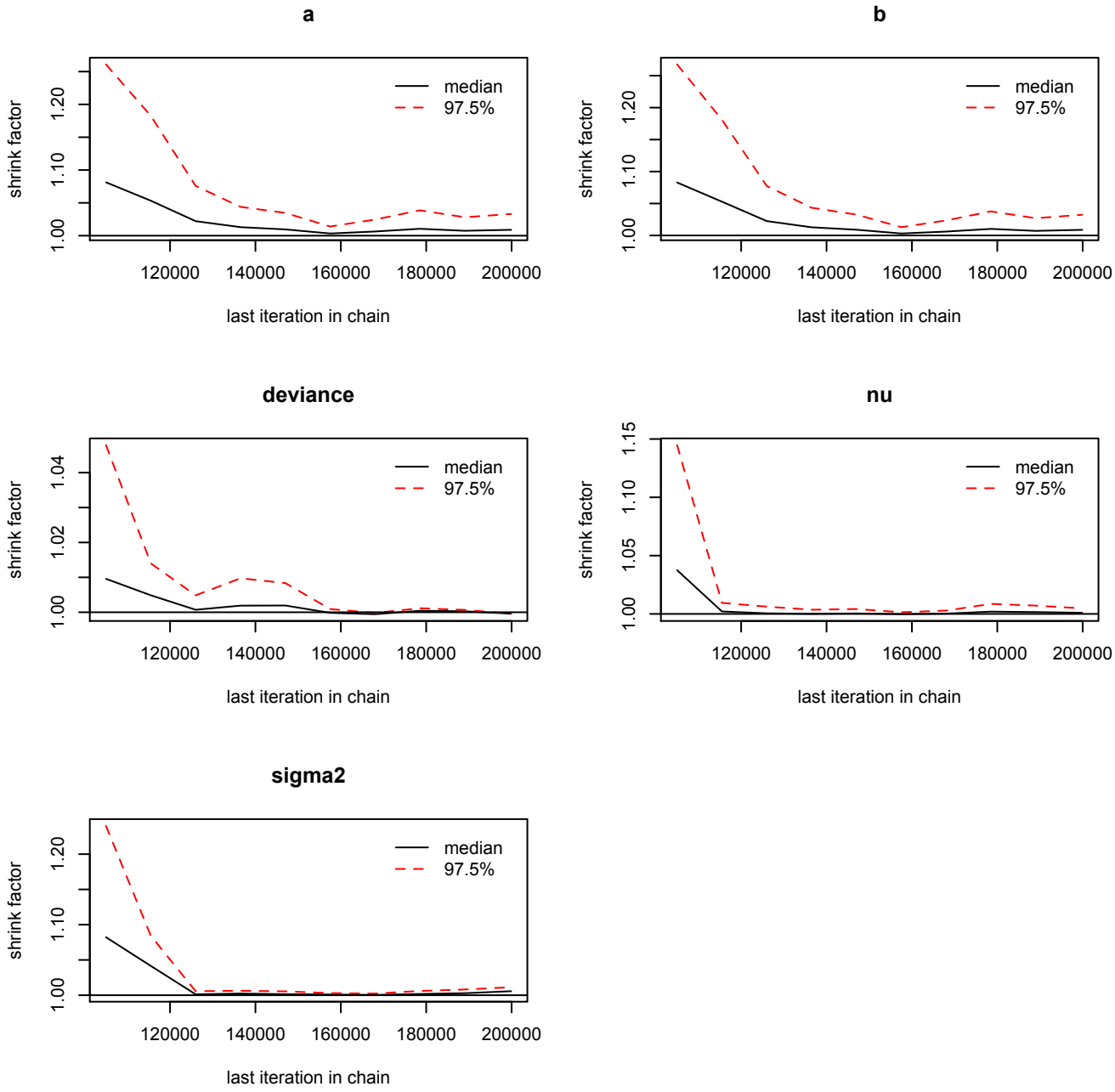

Figura 14: *Gráfica del diagnóstico de Gelman-Rubin del Modelo de Gompertz.*

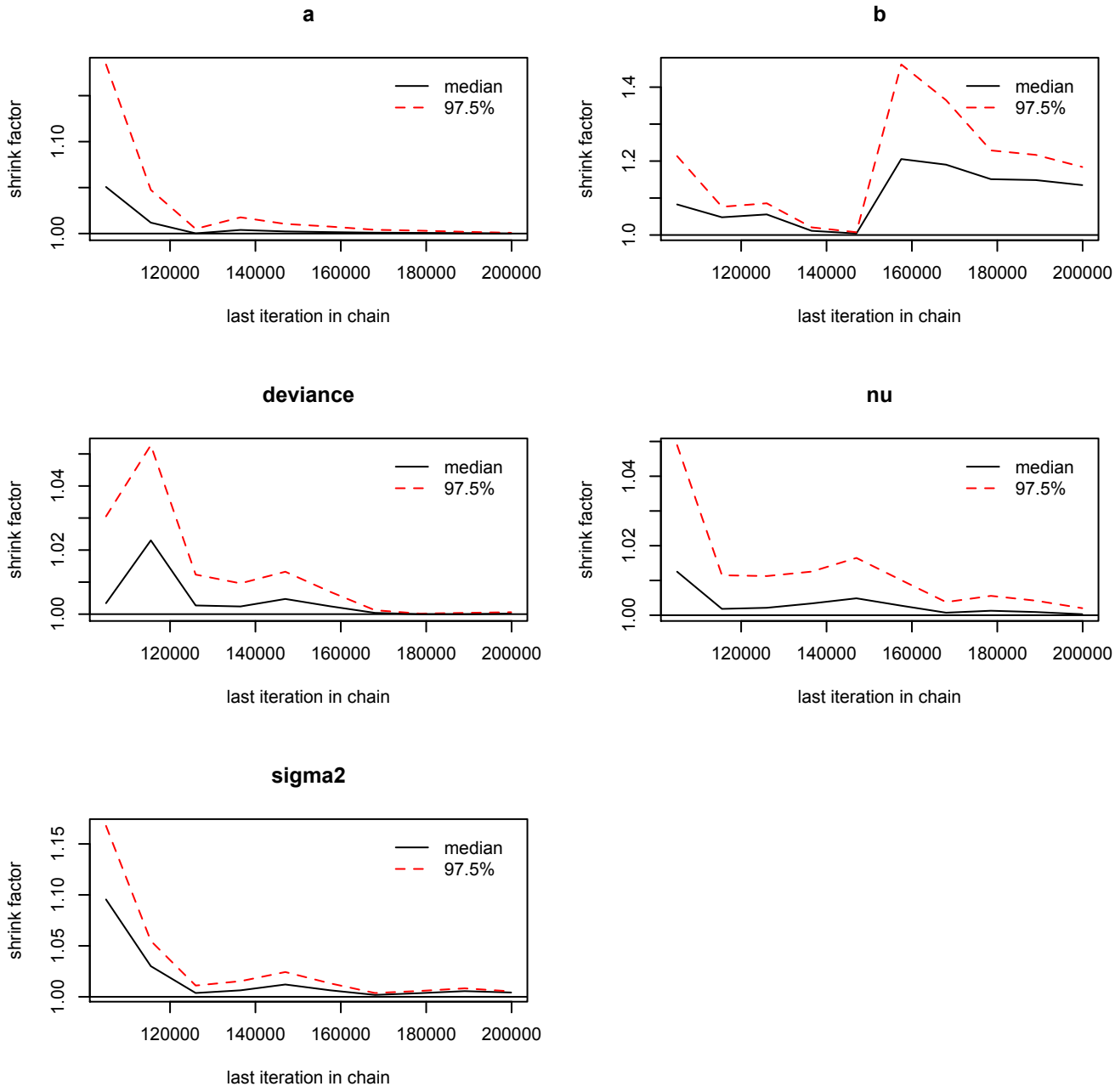

Figura 15: *Gráfica del diagnóstico de Gelman-Rubin del Modelo de Bertalanffy.*

Diagnóstico de Geweke

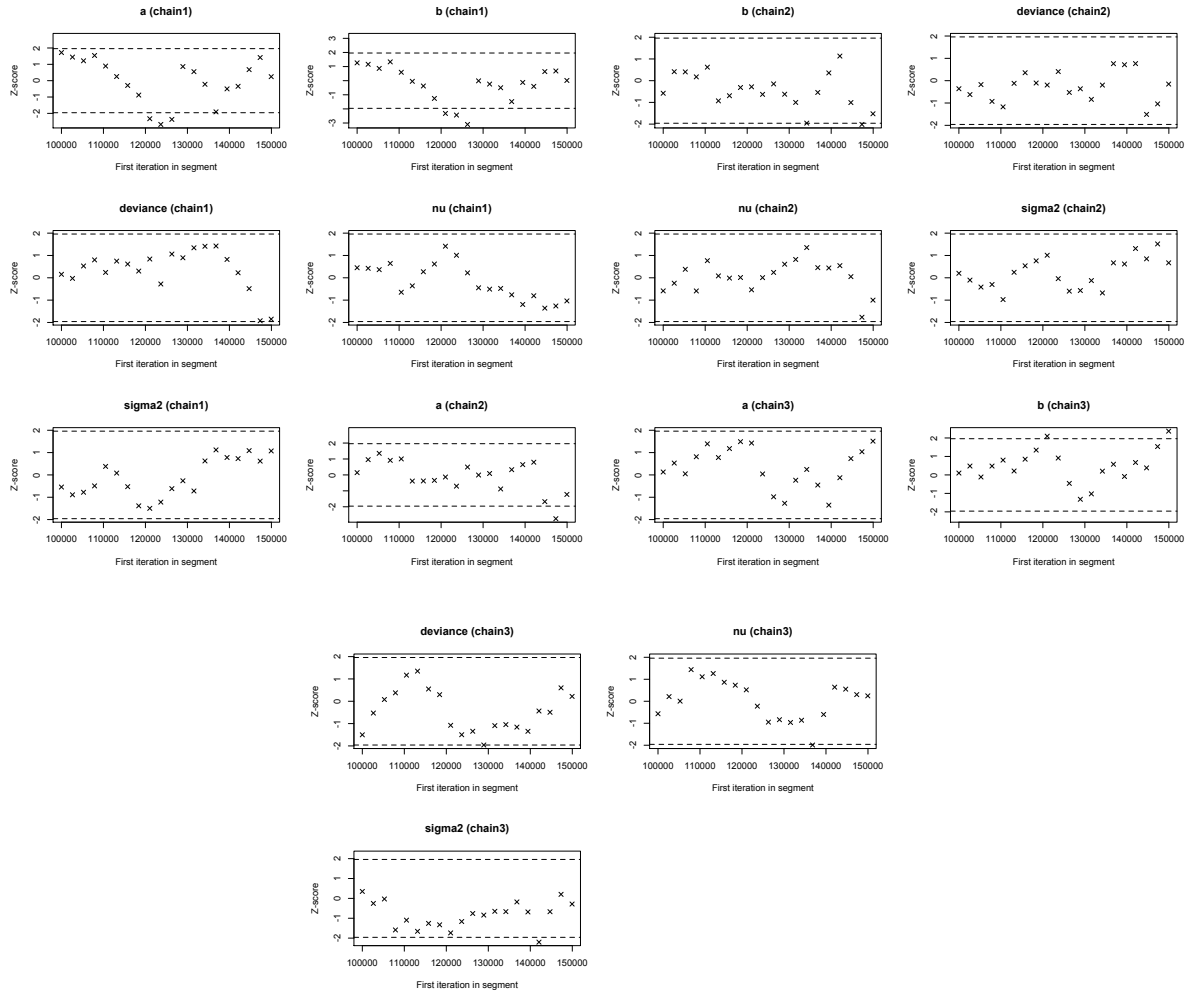

Figura 16: *Gráfica del diagnóstico de Geweke del Modelo Logístico.*

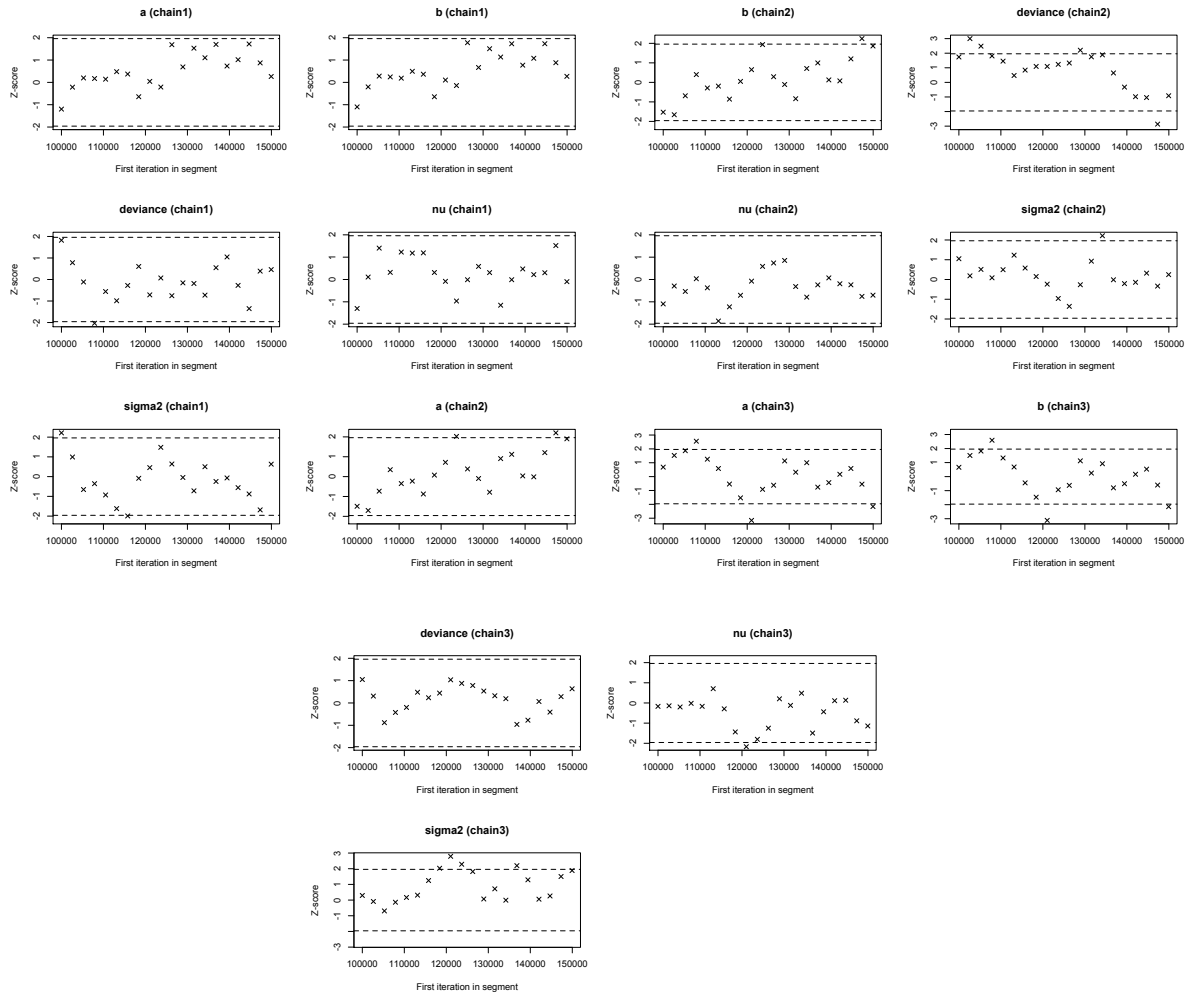

Figura 17: *Gráfica del diagnóstico de Geweke del Modelo de Gompertz.*

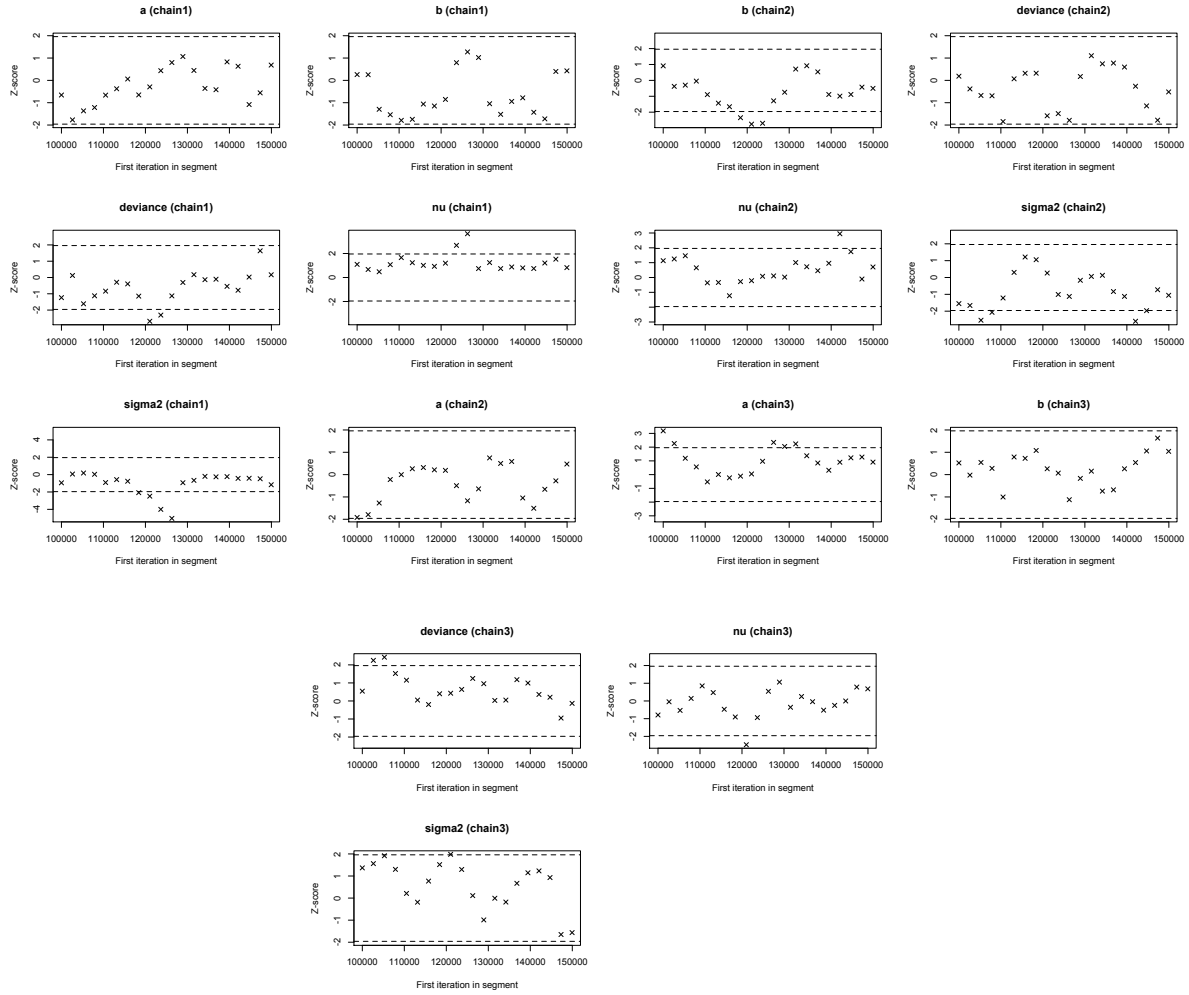

Figura 18: *Gráfica del diagnóstico de Geweke del Modelo de Bertalanffy.*