UNIVERSIDAD AUTÓNOMA DE GUERRERO

Facultad de Matemáticas

Maestría en Matemáticas Aplicadas

Un Modelo Epidémico de Orden Fraccionario para Ébola

T E S I S

PARA OBTENER EL GRADO DE:

Maestro en Matemáticas Aplicadas

PRESENTA:

Luis Reyes Velázquez

DIRECTOR DE TESIS:

Francisco J. Ariza Hernandez

Martin P. Árciga Alejandre

Abril de 2018.

Dedicatoria

Dedico esta tesis a todos aquellos que no creyeron en mí, a aquellos que esperaban mi fracaso en cada paso que daba hacia la culminación de mis estudios, a aquellos que nunca esperaban que lograra terminar la maestría, a todos aquellos que apostaban a que me rendiría a medio camino, a todos los que supusieron que no lo lograría, a todos ellos les dedico esta tesis.

Agradecimientos

A mi madre, por todos los sacrificios que hizo para darme la oportunidad de estudiar.

A mi padre, por todas sus valiosas enseñanzas.

A mi hermano, porque siempre estuvo conmigo en los momentos más difíciles.

A mis amigos y compañeros, por todos sus consejos y su apoyo incondicional.

A mis asesores, por creer en mi y permitirme trabajar a su lado.

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

Resumen

En este trabajo de tesis se estudia un modelo epidémico tipo SEIR (Susceptible, Expuesto, Infectado, Recuperados) de orden fraccionario para describir el comportamiento de la enfermedad producida por el virus de Ébola. El objetivo es realizar una estimación de parámetros del modelo en estudio desde un punto de vista Bayesiano. Para tal fin, se usan datos reales de la epidemia de ébola ocurrida en los paises: Guinea, Liberia y Sierra Leona, en el continente africano. Se usa un esquema numérico para solucionar el sistema de ecuaciones diferenciales y se plantean distribuciones iniciales de cada uno de ellos. Las inferencias acerca de los parámetros de interés son obtenidas a partir de un esquema de muestreo Monte Carlo vía Cadenas de Markov de la distribución posterior. Esta implementación se realiza con la ayuda del programa Metropolis-Hastings vía t-walk dentro del paquete estadístico R.

Índice general

Índice de figuras

Índice de tablas (opcional)

Capítulo 1

Introducción

1.1. Antecedentes

El Ébola es un virus letal, cuyo origen no es muy claro, pero fue descubierto en 1976 por Peter Piot quien ayudó a contener la primera epidemia de ébola registrada ese mismo año en el primer brote registrado en Nzara, Maridi y alrededores (Sudán) entre Junio y Noviembre de 1976. Hasta 2012, cerca de 200 casos y 1600 muertes fueron registrados debido al virus del ébola. En el brote del 2014, Guinea, Liberia, Nigeria, Senegal, y Sierra Leona tenemos aproximadamente 15935 casos y 5689 muertes.

Existen actualmente cinco virus del Ébola según el Comité Internacional de Taxonomía de Virus: Virus Ébola (EBOV), virus Sudán (SUDV), virus Reston (RESTV), virus Taï Forest (TAFV) y virus Bundibugyo (BDBV). Cuatro de estos virus (excepto RESTV) son conocidos por causar la enfermedad del virus del Ébola en seres humanos. Hasta 2014, el SUDV estuvo presente en 784 casos, el virus BDBV en 226 casos, una persona fue infectada por el virus TAFV, y los casos restantes (aproximadamente 2169) se debieron a virus EBOV. El brote 2014 está relacionado con el virus EBOV.

Un virus no puede sobrevivir tanto tiempo, ni replicarse en ningún caso, a no ser que se agazape en el interior de un ser vivo. Eso significa que necesita un huésped, un tipo de animal, planta, hongo o microbio cuyo organismo se convierte en su entorno primario y de

cuya maquinaria celular se aprovecha para reproducirse. Algunos virus dañinos habitan en animales no humanos y solo ocasionalmente pasan a las personas. Causan enfermedades que los científicos llaman zoonosis. El ébola es una zoonosis, una zoonosis especialmente grave y sorprendente: mata a muchas de sus víctimas humanas en cuestión de días, deja a otros al borde de la muerte y acto seguido desaparece. (17).

¿Dónde se esconde, callado e inadvertido, entre brote y brote? No entre chimpancés o gorilas; estudios de campo han demostrado que también ellos son a menudo víctimas del Ébola. No solo se han registrado elevadas mortandades de chimpancés y gorilas en el mismo marco temporal y espacial que los brotes de ébola en humanos, sino que alguna vez se han detectado indicios del virus en sus despojos. De hecho, una de las vías por las que el Ébola llega a los humanos es la ingestión de carne de primates antropomorfos. Tiene pues que ocultarse en algún otro lugar.

1.1.1. Planteamiento del problema

En este trabajo, se plantea un modelo general para los problemas inversos que serán estudiados posteriormente.

Modelo general:

1. Sistema dinámico:

$$\begin{cases} \frac{d\mathbf{X}_i(t)}{dt} = H(\mathbf{X}_i(t), t, \boldsymbol{\theta}); \quad i = 1, \cdots, n \\ \mathbf{X}_i(t_0) = \mathbf{X}_0 \end{cases} \tag{1.1}$$

2. Ecuación de observación:

$$y_i = h(\mathbf{X}(t_i)) + \epsilon_i, \quad i = 1, \cdots, n \tag{1.2}$$

Asumiendo un proceso de observaciones $\mathbf{y} = (y_1, y_2, \ldots, y_n)^t$ a un tiempo discreto $t_1, t_2, \cdots, t_n$, donde y_i corresponde al i-ésimo valor observado bajo incertidumbre, $h : \mathbb{R}^p \to \mathbb{R}^k$ es la función de observación, $\mathbf{X}(t)$ corresponde a la solución del sistema (1.1). Se desea estimar las cantidades $\theta = (\mathbf{X}_0, \xi, \nu)$, el cual es el vector de parámetros conocidos, $\theta \in \Theta \subset \mathbb{R}^d$, esto con la finalidad de caracterizar dicho sistema. La condición de Lipschitz para la función

$H : \mathbb{R}^p \times [0, T] \times \Theta \to \mathbb{R}^p$ asegura la existencia y unicidad de la solución del problema de valor inicial (1.1). Existen varios tipos de funciones de observación h que pueden ser consideradas, por ejemplo modelar una sola componente del p-vector $(X)(t)$ o una combinación lineal de ellas, (Capistrán et. al., 2013). Aquí, se considera un problema de observaciones de 4-dimensiones. El parámetro ν representa las cantidades de interes contenidas en el error de medición para las observaciones al tiempo t_i, los cuales son considerados como variables aleatorias independientes e idénticamente distribuidas $(i.i.d)$ de una distribución normal con media cero y varianza constante σ^2, denodado por $\epsilon \sim \mathcal{N}(0, \sigma^2)$. La presente tesis está estructurada como sigue: en el Capítulo 2, se presenta una breve introducción a los problemas inversos que serán estudiados mediante la teoría de la inversión estadística usando la estadística Bayesiana. En el Capítulo 3, se aborda el problema inverso para un modelo SEIR (susceptible, expuesto,infectado, recuperado) fraccionario. Y finalmente en el Capítulo 4, se estudia el problema inverso para un modelo de crecimiento de la población infectada fraccionario; y una aplicación a ejemplos con datos reales.

1.2. Objetivos

1.2.1. Objetivo general

Estimar, desde el punto de vista Bayesiano, de los parámetros involucrados en una modelo tipo SEIR fraccionario para describir el número de enfermos en una población infectada por el virus de Ébola.

1.2.2. Objetivos especificos

- Implementar un método numérico para obtener aproximaciones a la solución del sistema SEIR fraccionario.
- Estimar la función de verosimilitud de los datos a partir de la aproximación numérica
- Calcular las distribuciones posteriores marginales de los parámetros de interés en el modelo, usando métodos MCMC.

Capítulo 2

Estadística Bayesiana en problemas inversos

Los problemas inversos son definidos, como el propio término lo indica, como el inverso de problemas directos. Es evidente que esta definición es vacía a menos que se defina el concepto de problemas directos. Los problemas inversos se encuentran típicamente en situaciones en las que se hacen observaciones indirectas de una cantidad de interés.

Las teorías físicas permiten hacer predicciones: dada una descripción completa de un sistema físico, se puede predecir el resultado de algunas mediciones. Este problema de predecir el resultado de mediciones se llama problema de modelización, problema de simulación, o problema directo, (Tarantola, 2005). El problema inverso consiste en utilizar el resultado existente de algunas mediciones para inferir valores de los parámetros que caracterizan el sistema.

Mientras que un problema directo tiene una única solución, un problema inverso podría tener múltiples soluciones, lo que se convierte en un problema inestable y mal planteado. Se dice que un problema inverso,

$$\mathbf{y} = h(\xi) + \epsilon$$

esta bien planteado si:

1. Existe una solución para alguna $\mathbf{y}$ en el espacio de espacios observados.

2. La solución es única.

3. La solución depende sensiblemente de las condiciones iniciales.

Debido a esto, en problemas inversos, se necesita tener información inicial disponible acerca de las cantidades que son de interés. Además, se necesita tener cuidado en la representación de la *incertidumbre* en los datos.

La teoría más general (y simple), es obtenida cuando se usa un punto de vista probabilístico, donde la información disponible de una cantidad es representada por una distribución de probabilidad. Esta teoría explica como la información inicial es transformada en una distribución posterior de probabilidad, mediante un modelo y los resultados actuales de observaciones (con sus incertidumbres).

Con el fin de cuantificar la incertidumbre en las estimaciones de las cantidades de interés, se formula un modelo estadístico de la forma:

$$\mathbf{Y} = h(t, \theta_0) + \epsilon,$$

donde θ_0 son los verdaderos valores hipóteticos de los parámetros desconocidos, y ϵ es un vector aleatorio que representa el error de medición para las variables medidas al tiempo t. Las propiedades para los errores son definidos como:

$$\begin{aligned} E(\epsilon) &= 0, \\ Var(\epsilon) &= V_0 = \text{diag}(\sigma_{0,1}^2, \sigma_{0,2}^2, \ldots, \sigma_{0,m}^2), \\ Cov(\epsilon_i, \epsilon_i) &= Cov(\epsilon_i, \epsilon_j) = 0, \qquad i \neq j. \end{aligned}$$

En la ecuación de observación (1.2) se tiene que ϵ_i tiene distribución Normal, por lo tanto las observaciones y_i también tendrán una distribución similar, la cual es denotada como $y_i \leq \mathcal{N}(h(\mathbf{X}(t_i)), \sigma^2)$. Si existen n observaciones (*i.i.d.*), $y_1, y_2, \ldots, y_n$, con ésta distribución conocida, entonces la función de verosimilitud para las observaciones dadas es:

$$L(\mathbf{y}|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - h(X(t_i)))^2 \right\}.$$

Cualquier decisión estadística de los datos, como la estimación, predicción o selección del modelo se basa en ésta función, la cual, para un conjunto de observaciones $\mathbf{y} = (y_1, y_2, \ldots, y_n)^t$ puede ser expresada de la forma

$$L(\mathbf{y}|\theta) = \sigma^{-n}(2\pi)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - h(X(t_i)))^2 \right\}. \tag{2.1}$$

Note que (2.1) involucra el cálculo de $X(t)$, una solución de (1.1).
Añadiendo la información inicial acerca de $\boldsymbol{\theta}$, el problema inverso, entonces se puede establecer como un problema de combinar toda esta información.
La *teoría de la inversión estadística* resuelve losproblemas inversos sistematiamente en una manera tal que toda la información disponible es propiamente incorporada en el modelo.

2.1. Inversión estadística

Esta teoría reformula los problemas inversos como problemas de inferencia estadística mediante la estadística Bayesiana, en la cual todas las cantidades son modeladas como variables aleatorias (*v.a.'s*). La aleatoriedad, la cual refleja la incertidumbre ante el observador es codificada en distribuciones de probabilidad de las cantidades de interés. En esta sección, se explica la aproximación Bayesiana hacia problemas inversos. Se tienen cantidades observables directamente y otras que no pueden ser observadas. En los problemas inversos, algunas de las cantidades no observables son de mayor interés. Estas cantidades dependen una de la otra a través de modelos. El objetivo de la inversión estadística, es extraer información y evaluar la incertidumbre acerca de las variables basadas en todo el conocimiento disponible de un proceso de observación. La aproximación de la inversión estadística está basada en los siguientes principios:

1. Todas las variables incluidas en el modelo son modeladas como *v.a.'s*.
2. La aleatoriedad describe nuestro grado de información concerniente a las observaciones.
3. El grado de información es codificado mediante distribuciones de probabilidad.
4. La solución del problema inverso es la distribución posterior de probabilidad de las cantidades de interés.

Esta distribución describe el grado de confianza acerca de la cantidad de interés $\boldsymbol{\theta}$ después de realizadas las observaciones $\mathbf{y}$.
En la mayoría de los problemas inversos, uno de los desafíos es la enorme dimensión del problema, y en consecuencia tal distribución posterior también es de dimensión alta; para esto es posible utilizar métodos de simulación mediante Cadenas de Markov (descritos en la sección 2.3) para explorar esta distribución.

2.2. Fundamentos de estadística Bayesiana

La estadística Bayesiana es un enfoque particular de la estadística en general. El término "Bayesiana" hace referencia a Thomas Bayes (1702-1761), un matemático británico y ministro presbiteriano, cuya obra más conocida es el teorema que lleva su nombre, "*El teorema de Bayes*", el cual se refiere a la probabilidad de un evento condicionado por la ocurrencia de otro evento.

Más específicamente con su teorema se resuelve el problema conocido como "probabilidad inversa". Se trata de probabilidad "inversa" en el sentido de que la "directa" sería la probabilidad de observar algo supuesto que rigen ciertas condiciones.

Desde el punto de vista de la aproximación Bayesiana, existe una relación entre probabilidad e información, en donde el teorema de Bayes proporciona un modo natural de actualización de las creencias a cerca de una cantidad de interés cuando nueva información es incorporada al modelo. Tal proceso, es la base de la inferencia Bayesiana.

Dos aspectos importantes que caracterizan el enfoque de la aproximación Bayesiana son:

a) Considerar que todas las formas de incertidumbre son expresadas en términos de una medida de probabilidad. Se piensa que la probabilidad es una medida de lo que se sabe acerca de un evento.

b) Las probabilidades de un evento son actualizadas mediante evidencias.

El teorema de Bayes es la herramienta que permite explicar este enfoque. Este teorema será de utilidad para la solución a los problemas inversos que se tratarán en este trabajo.

2.2.1. El teorema de Bayes

Suponiendo que todas las variables aleatorias son absolutamente continuas, esto es, sus distribuciones de probabilidad pueden ser expresadas en términos de densidades de probabilidad. Como en el caso de problemas inversos clásicos, suponiendo que se están observando cantidades

$$\mathbf{y} \in \mathbb{R}^k,$$

para obtener información de alguna otra cantidad

$$\boldsymbol{\theta} \in \mathbb{R}^d.$$

Con el fin de relacionar estas dos cantidades, se necesita un modelo para su dependencia. Este modelo puede ser inapropiado y puede contener parámetros que no son conocidos. Además, las observaciones $\mathbf{y}$ siempre contienen error. Desde la perspectiva de la inversión estadística, se llama a Y como la *v.a.* observable (mediciones) y su realización, $\mathbf{y}$, en el proceso de medición. La v.a. no observable $\boldsymbol{\theta}$ que es de nuestro interés, es llamada *desconocida*. Aquellas variables que no son observables ni de interés, son llamadas *ruído*. Asumiendo que antes de realizar las mediciones de $\mathbf{y}$, se tiene alguna información acerca de la *v.a.* $\boldsymbol{\theta}$. Del teorema de Bayes, ésta información puede ser introducida por una densidad de probabilidad inicial $p(\boldsymbol{\theta})$, la cual expresa que conocimiento se tiene acerca de la cantidad de interés. Asumiendo que después de analizar las observaciones $\mathbf{y}$ e incorporando la información disponible acerca de $\boldsymbol{\theta}$, la densidad de probabilidad conjunta de $\boldsymbol{\theta}$ y $\mathbf{y}$ es denotada por $p(\boldsymbol{\theta}, \mathbf{y})$. La función de densidad marginal de $\boldsymbol{\theta}$ debe ser

$$\int_{\mathbb{R}^k} \pi(\boldsymbol{\theta}, \mathbf{y}) d\mathbf{y} = p(\boldsymbol{\theta}).$$

Finalmente, si los datos $\mathbf{y}$ son dados. La distribución de probabilidad condicional

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta}, \mathbf{y})}{\pi(\mathbf{y})},$$

es llamada la *distribución posterior* de $\boldsymbol{\theta}$. Se resume lo mencionado en el siguiente teorema, el cuál puede ser referido como:

Teorema 2.1 (teorema de Bayes) *Suponiendo que la variable $\boldsymbol{\theta} \in \mathbb{R}^d$ tiene una densidad de probabilidad inicial $P(\boldsymbol{\theta})$ y los datos consisten de valores observados $\mathbf{y}$ de una v.a. observada $Y \in \mathbb{R}^k$ tal que $\pi(\mathbf{y}) > 0$. Entonces la distribución de probabilidad de $\boldsymbol{\theta}$, dados los datos $\mathbf{y}$ es:*

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{\pi(\mathbf{y})} \tag{2.2}$$

El objetivo de la aproximación Bayesiana, es suministrar una metodología para estudiar adecuadamente la información mediante análisis de datos y usar la teoría de la decisión para actuar de la mejor manera. Dicha metodología Bayesiana puede ser idealizada por tres aspectos importantes:

1. establecer un modelo completo de probabilidad. Una distribución conjunta de probabilidad para todas las cantidades observables y no observables en un problema. El modelo

debe ser consistente con el conocimiento acerca del problema en estudio y el proceso de colección de datos.

2. Condicionamiento de los datos observados: calcular e interpretar la distribución posterior apropiada. La distribución condicional de probabilidad de las cantidades no observables de interés dados los datos observados.

3. Evaluar el ajuste del modelo y las implicaciones de la distribución posterior resultante.

A continuación se presentarán los elementos para la construcción de un modelo estadístico Bayesiano mediante la inferencia Bayesiana.

2.2.2. Inferencia Bayesiana

Los métodos bayesianos han sido generalizados particularmente porque son útiles para la solución de problemas en la toma de secisiones.
La estadística Bayesiana proporciona cantidades conocidas y desconocidas, lo cual permite incorporar los datos de los que se tiene conocimiento dentro de la estimación de los parámetros dados inicialmente, logrando así un proceso de estimación más rico en información, realizando inferencias sobre aquellas cantidades desconocidas.
Una vez que un modelo es definido, el objetivo principal de la estadística Bayesiana es realizar inferencias acerca de los paramétros desconocidos $\boldsymbol{\theta}$. Es posible utilizar información previa acerca de los valores de éstos para incorporarlos al análisis de los datos. En el contexto de la teoría de la inversión estadística, la solución aun problema inverso es la distribución posterior de las cantidades de interés, $\boldsymbol{\theta}$, dado que toda la información disponible se ha incorporado en el modelo, (Kaipio y somerslao, 2006). Así, por el teorema de Bayes, la distribución posterior de los parámetros de interes está dada por

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{L(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{m_{\mathbf{Y}}(\mathbf{y})}, \tag{2.3}$$

donde $p(\boldsymbol{\theta})$ es la distribución inicial para $\boldsymbol{\theta}$ y

$$m_{\mathbf{Y}}(\mathbf{y}) = \int_{\Theta} L(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta},$$

es una constante de normalización, también llamada *verosimilitud marginal* (ML) de los datos $\mathbf{y}$, Θ denota el espacio paramétrico de $\boldsymbol{\theta}$. Suponiendo que $\boldsymbol{\theta}$ es una variable aleatoria

que tiene una distribución a priori denotada por $p(\boldsymbol{\theta})$, la información concerniente a $p(\boldsymbol{\theta})$ está basada en la distribución posterior, la cual es obtenida por el teorema de Bayes. Tal distribución de $p(\boldsymbol{\theta})$ en presencia de las observaciones $\mathbf{y}$ ésta dada por

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{L(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\Theta} L(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{2.4}$$

Es claro que $\pi(\boldsymbol{\theta}|\mathbf{y})$ es proporcional a la función de verosimilitud multiplicada por la distribución a priori,

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto L(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{2.5}$$

y así, ésta involucra una contribución de los datos observados a través de $L(\mathbf{y}|\boldsymbol{\theta})$, y una contribución de la información a priori cuantificada a través de $p(\boldsymbol{\theta})$. El teorema de Bayes actualiza el conocimiento de $\boldsymbol{\theta}$ extrayendo información útil de interés, la cual está contenida en el conjunto de observaciones $\mathbf{y}$.

El uso de la inferencia Bayesiana tiene mucha ventaja en la aplicación de fenomenos físicos, pero presenta un gran esfuerzo al momento de seleccionar la distribución apriori adecuada, además de la complejidad en los cálculos de la distribución posterior.

Una de las limitaciones de las inferencia Bayesiana se concentra en la presentación explícita de la distribución posterior de las cantidades de interés, ya que resulta analíticamente intratable y las distribuciones posteriores marginales de los parámetros son complicadas. Sin embargo, en recientes años los estadísticos han desarrollado métodos de muestreo Monte Carlo vía Cadenas de Markov (MCMC) para simular tales distribuciones. Los métodos más comunes son Metroplis-Hastings, (Metropolis et al.,1953;Chib y Greenberg,1995) y el muestro Gibbs, (Gelfand y smith,1990; Casella y George, 1992).

Diagnósticos de convergencia

Suponiendo que se desea generar una muestra de tamaño N de la distribución $\pi(\boldsymbol{\theta}|\mathbf{y})$. si para cada uno de N valores iniciales $\boldsymbol{\theta}_1^{(0)}, \ldots, \boldsymbol{\theta}_N^{(0)}$ se realiza alguno de los algoritmos discutidos en esta sección, entonces, después de un cierto número de iteraciones T suficientemente grande, los valores $\boldsymbol{\theta}_1^{(T)}, \ldots, \boldsymbol{\theta}_N^{(T)}$ pueden considerarse como una muestra de tamaño N de la distribución posterior de $\boldsymbol{\theta}$. Alternativamente se puede generar una sola Cadena y tomar valores $\boldsymbol{\theta}^{(T+K)}, \ldots, \boldsymbol{\theta}^{(T+NK)}$ como una muestra de $\pi(\boldsymbol{\theta}|\mathbf{y})$, donde K se elige de manera que la correlación entre las observaciones sea pequeña. Un método MCMC crea una muestra de la

distribución posterior, y por lo general se desea saber si la muestra es suficientemente cercana a la distribución $\pi(\boldsymbol{\theta}|\mathbf{y})$ para ser usada en el análisis de los datos. Existen varias pruebas para verificar la convergencia de la cadena de Markov, ambas son visuales y estadísticas. Alguna de estas pruebas de inspección visual es la siguiente:

Traza de la cadena

La traza es una gráfica del número de iteración contra el valor de la estimación del parámetro en cada iteración. Se puede ver si la Cadena se detiene en determinadas zonas del espacio de parámetros, lo que indica una mala convergencia. Algunas de las pruebas estadísticas para evaluar la convergencia se mencionan a continuación:

Diagnóstico de secuencia múltiple de Gelman y Rubin.

Para cada parámetro se deben realizar los siguientes pasos:

1. Correr $m \geq 2$ cadenas de longitud $2N$ de valores iniciales sobredispersados.
2. Descartar las primeras N muestras en cada cadena.
3. Calcular la varianza dentro y entre la cadena.

   a) Varianza dentro de la cadena:

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

   donde $s_j^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\theta_{ij} - \bar{\theta}_j)^2$ es simplemente la formula para la varianza de la j-ésima cadena. W es la media de las varianzas de cada cadena.

   b) Varianza entre la cadena:

   $$B = \frac{N}{m-1}\sum_{j=1}^{m}(\bar{\theta}_j - \bar{\bar{\theta}})^2,$$

   donde $\bar{\bar{\theta}} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta}_j$. Esto es la varianza de las medias de la cadena multiplicada por N debido a que cada cadena está basada en N muestras.

4. Calcular la varianza estimada del parámetro como una suma ponderada de la varianza dentro y entre la Cadena.

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

5. Calcular el factor potencial de reducción de escala. En Gelman et al. (2014), se puede ver que dicho factor es calculado como:

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

donde $\hat{R}$, se considera alto si es mayor que 1.1 ó 1.2, entonces se debería correr las cadenas más tiempo para mejorar la convergencia a la distribución estacionaria. Si se tiene más de un parámetro, entonces se necesita calcular dicho factor para cada parámetro.

Diagnóstico de Geweke.

Este diagnóstico toma dos partes (usualmente el 10 % de la primer mitad, y la segunda mitad) de la cadena de Markov. Suponiendo que la segunda mitad de la cadena ha convergido a la distribución estacionaria. Se realiza una prueba de comparación de medias de ambas partes. Si la media del primer 10 % de las muestras no es significativamente diferente del último 50 %, entonces se concluye que estas dos partes provienen de la misma distribución, y por consiguiente la cadena de Markov ha convergido. El estadístico de prueba es un Z-score estándar con errores estándar ajustados por autocorrelación, (Geweke et al., 1991).

Diagnóstico de Heidelberg y Welch.

Este diagnóstico calcula un estadístico de prueba (basado en el estadístico de prueba de Mises Cramer-von) para aceptar o rechazar la hipótesis nula de que la cadena de Markov es de una distribución estacionaria. El diagnóstico consiste de dos partes:

1. Primera parte:

   a) Generar una cadena de N iteraciones y definir un nivel de significación α.

b) Calcular el estadístico de prueba en toda la cadena. Aceptar o rechazar la hipótesis nula de que la cadena es de una distribución estacionaria.

c) Si la hipótesis nula es rechazada, descartar el primer 10 % de la cadena. Calcular el estadístico de prueba y aceptar o rechazar la hipótesis nula.

d) Si la hipótesis nula es rechazada, descartar el próximo 10 % de la cadena y calcular el estadístico de prueba.

e) Repetir hasta que la hipótesis nula sea aceptada o el 50 % de la cadena sea descartada. Si el estadístico todavía rechaza la hipótesis nula, entonces la cadena falla y se necesita un número mayor de iteraciones.

2. Segunda parte:

   Si la cadena pasa la primera parte del diagnóstico, entonces ésta toma la parte de la cadena no descartada de la primera parte para probar la segunda parte. La prueba half-width calcula un intervalo de confianza del 95 % para la media, utilizando la porción de la cadena que pasó la prueba de estacionariedad (primera parte). La mitad del ancho de este intervalo es comparado con la estimación de la media. Si la razón entre la mitad del ancho y la media es inferior a algún $\epsilon > 0$, entonces la prueba half-width es aceptada. De lo contrario, la longitud de la muestra no es lo suficientemente grande para estimar la media con suficiente precisión.

2.2.3. Implementación de software del MCMC vía t-walk

En la actualidad, muchos de los algoritmos MCMC ya se han implementado en programas de computadora, tales como WinBUGS (Spiegelhalter, D. J. et al, 2003), JAGS (Plummer, 2012), Stan (Stan Development Team, 2014) y t-walk, (Christen y Fox, 2010). Todos estos programas proporcionan paquetes para el modelado Bayesiano mediante la simulación de la distribución posterior dada la información existente y un modelo específico.
En esta tesis, se utiliza el algoritmo de Metropolis-Hastings vía t-walk para obtener muestras de las distribuciones posteriores marginales de interés.

Algoritmo de Metropolis-Hastings

El algoritmo básico propuesto por Metropolis-Hastings, el cual fue desarrollado por Metropolis, Rosenbluth y Teller en 1953 y subsecuentemente generalizado por Hastings en 1970 (Chib y Greenberg, 1995), constituye un método para simular distribuciones multivariadas. El principio de los algoritmos MCMC es que dado un valor inicial $x^{(0)}$, la cadena $(X^{(t)})$ es generada utilizando un kernel de transición con distribución estacionaria, $\pi(\cdot)$, la cual garantiza la convergencia en distribución de $(X^{(t)})$ a $\pi(\cdot)$. Dado que la cadena es ergódica, el valor inicial $x^{(0)}$ en principio no es importante.

El algoritmo Metropolis-Hastings comienza muestreando de una densidad de candidatos y una distribución objetivo, pero como se esta considerando cadenas de Markov, la densidad depende del estado actual del proceso, (Lee, 2012). Denotando a la densidad de candidatos por $q(\phi|\theta)$ y suponiendo que $\sum_{\phi} q(\phi|\theta) = 1$. Si resulta que la densidad $q(y|x)$ siempre es la misma, entonces se necesita otra opción. Más sin embargo, si se encuentra que

$$\pi(\theta)q(\phi|\theta) > \pi(\phi)q(\theta|\phi),$$

entonces parece que el proceso se mueve de θ a ϕ, y de ϕ a θ. Se puede reducir el número de movimientos de θ a ϕ introduciendo una probabilidad $\alpha(\phi\theta)$, llamada probabilidad de aceptación. Con el fin de alcanzar el tiempo de reversalidad, se toma $\alpha(\phi|\theta)$ tal que la ecuación

$$\pi(\theta)q(\phi|\theta) = \pi(\phi)q(\theta|\phi)$$

se asegura, y consecuentemente,

$$\alpha(\phi|\theta) = \frac{\pi(\phi)q(\theta|\phi)}{\pi(\theta)q(\phi|\theta)}.$$

No se quiere reducir el número de movimientos de ϕ a θ en tal caso, así, se toma $\alpha(\theta|\phi) = 1$, y similarmente $\alpha(\phi|\theta) = 1$ en el caso donde la igualdad se invierte, entonces se tiene

$$\pi(\theta)q(\phi|\theta) < \pi(\phi)q(\theta|\phi).$$

Es claro que una fórmula general es

$$\alpha(\phi|\theta) = min\left\{\frac{\pi(\phi)q(\theta|\phi)}{\pi(\theta)q(\phi|\theta)}, 1\right\},$$

así, la probabilidad de ir de un estado θ a un estado ϕ es $p^*(\phi|\theta) = q(\phi|\theta)\alpha(\phi|\theta)$, mientras que la probabilidad de que la cadena permanezca en el estado θ es

$$r(\theta) = 1 - \sum_{\phi} q(\phi|\theta)\alpha(\phi|\theta).$$

La matriz de transición de probabilidades es dada por

$$p(\phi|\theta) = p^*(\phi|\theta) + r(\theta)\delta(\phi|\theta) = q(\phi|\theta)\alpha(\phi|\theta) + \left(1 - \sum_{\phi} q(\phi|\theta)\alpha(\phi|\theta)\right)\delta(\phi|\theta).$$

Teniendo en cuenta que es necesario conocer la densidad objetivo $\phi(\theta)$ hasta un multiple constante, porque esta aparece en el numerador y denominador de la expresión para $\alpha(\phi|\theta)$. Además, si la densidad generadora de candidatos $q(\phi 1\theta)$ es simétrica, se tiene que $q(\phi|\theta) = q(\theta|\phi)$, y $\alpha(\phi|\theta)$ se reduce a

$$\alpha(\phi|\theta) = min\left\{\frac{\pi(\phi)}{\pi(\theta)}, 1\right\}.$$

Se puede resumir el algoritmo de Metropolis-Hastings como sigue:

1. Muestrear un candidato θ^* de una distribución propuesta $q(\theta^*|\theta^{(t-1)})$.

2. Calcular
$$\alpha = min\left\{\frac{p(\theta^*)q(\theta^{(t-1)}|\theta^*)}{p(\theta^{(t-1)})q(\theta^*|\theta^{(t-1)})}, 1\right\}.$$

3. Generar un valor $U \sim \mathcal{U}(0, 1)$.

4. Si $U \leq \alpha$, se define $\theta^{(t)} = \theta^*$; en otro caso se define $\theta^{(t)} = \theta^{(t-1)}$.

5. Se devuelve la secuencia $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(n)}\}$.

Más tarde, en Geman y Geman (1984), se presenta un método de simulación, que tambien genera una cadena de Markov y que después pasa a ser conocido en la literatura como muestreador de Gibbs. El algoritmo de Metropolis-Hastings y el muestreador de Gibbs forman los dos esquemas básicos de la metodología MCMC a partir de las cuales se han creado otros con fines más específicos y distintas propiedades.

Algoritmo de t-walk

Desarrollamos una nueva muestra MCMC de propósito general para distribuciones arbitrarias continuas que no requieren ajuste. Llamamos a este MCMC el t-walk. El t-walk mantiene dos puntos independientes en el espacio de muestra, y todos los movimientos se basan en propuestas que luego se aceptan con una probabilidad de aceptación estándar de Metropolis-Hastings en el espacio del producto. Por lo tanto, el t-walk es probablemente convergente bajo los requisitos suaves habituales. Restringimos las distribuciones de propuestas, o "movimientos", a aquellas que producen un algoritmo que es invariable a escala, y aproximadamente invariante para afinar las transformaciones del espacio de estado. Por lo tanto, la ampliación de las propuestas, y efectivamente también la transformación de coordenadas, que podrían ser utilizados para aumentar la eficiencia de la toma de muestras, no son necesarios ya que el t-walk es idéntico en cualquier versión a escala de la distribución de destino. Se dan cuatro movimientos que dan como resultado un algoritmo de muestreo efectivo, (Cristen y Fox, 2010).

Usamos el dispositivo simple de actualizar solo un subconjunto de coordenadas al azar en cada paso para permitir la aplicación del t-walk a problemas de alta dimensión. En una serie de problemas de prueba a través de dimensiones nos encontramos con que el t -walk es solamente un pequeño factor menos eficiente que los algoritmos afinados de manera óptima, pero supera significativamente del paseo aleatorio muestreadores generales MH que no están sintonizados para problemas específicos. Además, el t-walk sigue siendo efectiva para las distribuciones objetivo para las que no existe una transformación afín óptima, como aquellas en las que la estructura de correlación es muy diferente en diferentes regiones del espacio de estados.

Para una función objetivo (distribución posterior) $\pi(x)$, $x \in \mathcal{X}$ ($\mathcal{X}$ es de dimensión n y es un subconjunto de $\mathbb{R}^n$), donde creamos la nueva función objetivo $f(x, x^{'}) = \pi(x)\pi(x^{'})$ en el espacio producto correspondiente $\mathcal{X} \times \mathcal{X}$. Mientras que una propuesta general tiene la forma:

$$q\{(y, y^{'})|(x, x^{'})\}.$$

consideramos las dos propuestas restringidas, con igual probabilidad:

$$(y, y^{'}) = \begin{cases} (x, h(x^{'}, x)) \\ (h(x, x^{'}), x^{'}). \end{cases} \tag{2.6}$$

Donde $h(x, x^{'})$ es una variable aleatoria utilizada para formar la propuesta. Es decir, cambiamos solo x por $x^{'}$ en cada paso. Sin embargo, tenemos en cuenta que no estamos considerando dos cadenas paralelas independientes en cada $\mathcal{X}$; en cambio, todo el proceso se encuentra en $\mathcal{X} \times \mathcal{X}$. Seleccionaremos al azar entre cuatro propuestas diferentes, que se definirán a continuación, cada una caracterizada por una función particular $h(\cdot, \cdot)$. Primero elegiremos una opción en (2.6) y segundo, crear la propuesta $(y, y^{'})$ simulando de la función h correspondiente.

Dentro de un esquema de Metropolis-Hastings, necesitamos calcular la tasa de aceptación correspondiente. Denotando la función de densidad de $h(x, x^{'})$ por $g(\cdot|(x, x^{'}))$, esta relación es igual a:

$$\frac{\pi(y^{'})}{\pi(x^{'})} \frac{g(x^{'}|y^{'}, x)}{g(y^{'}|x^{'}, x)}.$$

Para el primer caso en las ecuaciones (2.6) y :

$$\frac{\pi(y)}{\pi(x)} \frac{g(x|y, x^{'})}{g(y|x, x^{'})}.$$

Para el segundo caso tenga en cuenta que la restricción a la propuesta (2.6) implica que solo un único requiere evaluación de la densidad objetivo, en cualquier caso.

Es sencillo mostrar que si la variable aleatoria h es invariante para las transformaciones afines, es decir, $h(\phi x, \phi x^{'})$ para cualquier transformación afín ϕ, entonces también lo son las propuestas (2.6) y la muestra de MCMC resultante. El diseño de un algoritmo de muestreo invariable descansa en la pregunta de si es posible encontrar una o más variables aleatorias h que den un algoritmo de muestreo efectivo. Hemos encontrado que las cuatro opciones para h, dadas a continuación, dan una mezcla adecuada a través de un amplio rango de distribuciones objetivo de dimensión moderada.

En muchas aplicaciones, particularmente con correlaciones débiles, encontramos que la mezcla de la cadena se logra principalmente mediante una caminata aleatoria escalada a la que nos referimos como el movimiento de t-walk.

El movimiento de t-walk, está definida por la función:

$$h_w(x, x^{'})_j = \begin{cases} (x_j + (x_j - x_j^{'}), \ I_j = 1 \\ x_j \qquad\qquad\qquad\quad I_j = 0. \end{cases}$$

para $j = 1, 2, \cdots, n$.

Capítulo 3

Modelo SEIR fraccionario

El proceso de modelización en epidemiología tiene, en el fondo, la misma filosofía y objetivos subyacentes que el modelado ecológico. Ambos intentos comparten el objetivo último de intentar comprender la prevalencia y distribución de una especie, junto con los factores que determinan la incidencia, la dispersión y la persistencia. Mientras que en la ecología la abundancia precisa de una especie suele ser de gran interés, establecer o predecir el número exacto de, por ejemplo, partículas de virus en una población (o incluso dentro de un individuo) es a la vez desalentador e inviable. En cambio, los modeladores se concentran en la tarea más simple de clasificar a las personas en la población "*de acogida*" de acuerdo con su estado de infección. Como tales, estos modelos epidemiológicos pueden compararse con los modelos de metapoblación utilizados en ecología, donde cada huésped individual se considera un parche de recursos para el patógeno, con transmisión y recuperación análoga a la dispersión y la extinción.

Introducimos brevemente un refinamiento al modelo SIR para tener en cuenta el período latente. El proceso de transmisión a menudo ocurre debido a una inoculación inicial con un número muy pequeño de unidades patógenas. Luego se produce un período de tiempo durante el cual el patógeno se reproduce rápidamente dentro del huésped, relativamente no cuestionado por el sistema inmunitario. Durante esta etapa, la abundancia de patógenos es demasiado baja para la transmisión activa a otros hospedadores susceptibles, y sin embargo, el patógeno está presente. Por lo tanto, el huésped no puede clasificarse como susceptible, infeccioso o recuperado; tenemos que introducir una nueva categoría para estas personas que

están infectadas pero aún no son infecciosas. Estos individuos se conocen como Expuestos y están representados por la variable E en los modelos **SEIR**. [Modeling Infectious Diseases in Human and Animals, Matt J. Keeling and Pejman Rohani; (2008) Princeton University Press; USA].

Denotaremos por *S(t), E(t), I(t)* y *R(t)* a los susceptibles, expuestos, infecciosos y eliminados en el tiempo t, respectivamente. También asumiremos que

$$N = S(t) + E(t) + I(t) + R(t)$$

es la población (constante) estudiado en el modelo. Note que N incluye los individuos eliminados *R(t)*.

Suponiendo que la duración promedio del período latente está dada por $1/\sigma$, la ecuación SEIR es:

$$\begin{cases} S'(t) = -\frac{\beta S(t)(qE(t)+I(t))}{N}, \\ E'(t) = \frac{\beta S(t)(qE(t)+I(t))}{N} - \delta E(t) \\ I'(t) = \delta E(t) - \gamma I(t) \\ R'(t) = \gamma I(t) \end{cases} \tag{3.1}$$

Este modelo toma en consideración el número de personas infectadas debido al contacto directo con una persona infectada y el número de personas infectadas por contacto directo con una persona latente: $\beta S(I + qE)/N$. Los individuos en estado latente eventualmente muestran los síntomas de la enfermedad, y pasan a la fase infecciosa, esto es denotado por δE. Como antes, se considera que la muerte y la recuperación son las mismas, ya que no ha habido un caso en el que una persona que sobrevivió a ébola contraiga la enfermedad nuevamente.

Modelo Fraccionario

La derivada de *Riemann-Liouville* de orden α está definida como

$$D^\alpha f(t) = D^1 I^{1-\alpha} f(t) = \frac{1}{\Gamma(1-\alpha)} \frac{d}{dt} \int_0^t (t-s)^\alpha f(s) ds$$

Podemos escribir el sistema (3.2) en términos de una ecuación diferencial fraccionaria como:

$$\begin{cases} D^{\alpha}S(t) = -\frac{\beta S(t)(qE(t)+I(t))}{N}, \\ D^{\alpha}E(t) = \frac{\beta S(t)(qE(t)+I(t))}{N} - \delta E(t) \\ D^{\alpha}I(t) = \delta E(t) - \gamma I(t) \\ D^{\alpha}R(t) = \gamma I(t) \end{cases} \tag{3.2}$$

donde los parámetros β, γ, δ, y q tienen el mismo significado como en el modelo (3.1), N es el total de la población y $\alpha \in (0, 1)$ es el orden de la derivada.

3.1. Solución numérica

Introduciremos un algoritmo para la solución numérica de ecuaciones diferenciales no lineales de orden fraccional α, donde el énfasis principal esta en el caso $0 < \alpha < 1$. Los métodos de solución estándar para las ecuaciones lineales generalmente fallan en el caso no lineal. Así, ahora presentamos un nuevo algoritmo, llamado FracPECE, que nos permite resolver las ecuaciones diferenciales y así analizar el modelo de manera eficiente.

El método de FracPECE es un método eficiente para la solución numérica de ecuaciones diferenciales de orden fraccionario. Como el nombre FracPECE indica, el algoritmo se basa en el PECE clásico (Predecir, Evaluar, Corregir, Evaluar) el cual es modificado para ser capaz de manejar los operadores fraccionarios diferenciales. En la derivación del algoritmo hemos tomado especial cuidado en el hecho de que el modelo no consiste únicamente en ecuaciones diferenciales fraccionales, sino que también contienen ecuaciones diferenciales de primer orden.

La derivada de Riemann-Liouville del orden $\alpha > 0$ de una función f con respecto al punto t_0 se denota y define por

$$D_{t_0}^{\alpha} f(t) = \frac{1}{\Gamma(m-\alpha)} \frac{\mathrm{d}^m}{\mathrm{d}t^m} \int_{t_0}^{t} f(u)(t-u)^{m-\alpha-1} \mathrm{d}u$$

donde m es el entero definido por la relación $m - 1 < \alpha < m$, si α es un número natural, entonces recuperamos la definición clásica de derivada.

La definición de la derivada fraccional y algunos resultados bien conocidos del cálculo fraccional nos dicen que podemos interpretar una ecuación diferencial fraccional

$$D_{t_0}^{\alpha}(y - y_0)(t) = f(t, y(t)), \quad y(t_0) = y_0 \tag{3.3}$$

Como una ecuación integral de la forma

$$\frac{1}{\Gamma(-\alpha)}\int_{t_0}^{t}\frac{y(u)-y_0}{(t-u)^{\alpha-1}}\mathrm{d} = f(t,y(t))\;\; y(t_0)=y_0$$

Alternativamente, podemos aplicar un operador integral fraccionario a la ecuación diferencial e incorporar las condiciones iniciales, convirtiendo así la ecuación en la ecuación equivalente

$$y(t) = y(t_0) + \frac{1}{\Gamma(\alpha)}\int_{t_0}^{t}(t-u)^{\alpha-1}f(u,y(u))\mathrm{d}u \tag{3.4}$$

Descripción del Algoritmo

La clave para la derivación del método es reemplazar la ecuación diferencial fraccional original (3.3) por la ecuación equivalente débilmente singular (3.4) e implementar un método de integración del producto para este último. Lo que hacemos es simplemente usar la fórmula de cuadratura trapezoidal del producto con los nodos $t_j \quad (j = 0, 1, ..., n+1)$, tomados con respecto a la función de peso $(tn+1-\cdot)^{\alpha-1}$, para reemplazar la integral. En otras palabras, aplicamos la aproximación

$$\int_{t_0}^{t_{n+1}}(t_{n+1}-u)^{\alpha-1}g(u)\mathrm{d}u \approx \int_{t_0}^{t_{n+1}}(t_{n+1}-u)^{\alpha-1}g_{n+1}(u)\mathrm{d}u \tag{3.5}$$

donde g_{n+1} es el interpolante lineal por partes para g cuyos nodos y nudos se eligen en el t_j, $j = 0, 1, 2, \cdots, n+1$. Un cálculo explícito produce que podemos escribir la integral en el lado derecho de la ecuación (3.5) como

$$\int_{t_0}^{t_{n+1}}(t_{n+1}-u)^{\alpha-1}g_{n+1}(u)\mathrm{d}u = \sum_{j=0}^{n+1}a_{j,n+1}g(t_j)$$

donde

$$a_{j,n+1} = \int_{t_0}^{t_{n+1}}(t_{n+1}-u)\,\phi_{j,n+1}(u)\mathrm{d}u \tag{3.6}$$

y

$$\phi_{j,n+1}(u) = \begin{cases} (u-t_{j-1})/(t_j-t_{j-1}) & \text{si } \; t_{j-1} < u < t_j; \\ (t_{t+j}-u)/(t_{j+1}-t_j) & \text{si } \; t_j < u < t_{j+1}; \\ 0 & \text{en otro caso} \end{cases}$$

en el caso de los nodos equi-espaciados $t_j = t_0 + jh$ con un h fijo, las relaciones de la ecuación (3.6) reducida

$$a_{j,n+1} = \begin{cases} \frac{h^\alpha}{\alpha(\alpha+1)} \left(n^{\alpha+1} - (n-\alpha)(n+1)^\alpha\right) & \text{si } j = 0, \\ \frac{h^\alpha}{\alpha(\alpha+1)} & \text{si } j = n+1, \end{cases}$$

mientras que para $1 < j < n$, tendremos

$$a_{j,n+1} = \frac{h^\alpha}{\alpha(\alpha+1)} \left((n-j+2)^{\alpha+1} + (n-j)^{\alpha+1}\right)$$

Esto nos proporciona nuestra fórmula correctora, es decir, la variante fraccionada del método Adams-Moulton de un solo paso, que es

$$y_{n+1} = y_0 + \frac{1}{\Gamma(\alpha)} \left(\sum_{j=0}^{n} a_{j,n+1} f(t_j, y_j) + a_{n+1,n+1} f\left(t^P_{n+1,y_{n+1}}\right) \right) \tag{3.7}$$

El problema restante es la determinación de la fórmula del predictor que necesitamos para calcular el valor y^P_{n+1}. La idea que usamos para generalizar el método de Adams-Bashforth de un solo paso es la misma que la descrita anteriormente para la técnica de Adams-Moulton: reemplazamos la integral en el lado derecho de la ecuación (3.4) por la regla del rectángulo del producto, es decir

$$\int_{t_0}^{t_{n+1}} (t_{n+1} - u)^{\alpha-1} g(u) \mathrm{d}u \approx \sum_{j=0}^{n} b_{j,n+1} g(t_j)$$

donde

$$b_{j,n+1} = \int_{t_j}^{t_{j+1}} (t_{n+1} - u)^{\beta-1} \mathrm{d}u = \frac{1}{\alpha} \left((t_{n+1} - t_j)^\alpha - (t_{n+1} - t_{j+1})^\alpha\right)$$

Nuevamente, en el caso equi-espaciado, tenemos la expresión más simple

$$b_{j,n+1} = \frac{h^\alpha}{\alpha} \left((n+1-j)^\alpha - (n-j)^\alpha\right)$$

Así, el predictor y^P_{n+1} es determinado por

$$y^P_{n+1} = y_0 + \frac{1}{\Gamma(\alpha)} \sum_{j=0}^{n} b_{j,n+1} f(t_j, y_j) \tag{3.8}$$

Esto completa la descripción de nuestro algoritmo básico, que es la versión fraccionada del método Adams-Bashforth-Moulton de un solo paso. Recapitulando, vemos que primero tenemos que calcular el predictor y^P_{n+1} de acuerdo a la ecuación (3.8), entonces evaluamos $f(t_{n+1}, y^P_{n+1})$, usamos esto para determinar el corrector y_{n+1} por medio de la ecuación (3.7), y finalmente evaluamos $f(t_{n+1}, y_{n+1})$ que luego se usa en la siguiente etapa de integración.

Capítulo 4

Análisis Bayesiano del modelo SEIR fraccionario.

El modelo a estudiar fue definido en (1.1) y (1.2), el cuál esta dado por:

1. Sistema dinámico:

$$\begin{cases} S^{\alpha}(t) = -\frac{\beta S(t)(qE(t)+I(t))}{N}, \\ E^{\alpha}(t) = \frac{\beta S(t)(qE(t)+I(t))}{N} - \delta E(t) \\ I^{\alpha}(t) = \delta E(t) - \gamma I(t) \\ R^{\alpha}(t) = \gamma I(t) \end{cases} \tag{4.1}$$

donde:

$\beta =$ es el producto de las tasas de contacto y la probabilidad de transmisión.

$\delta =$ es la razón de infección per-cápita.

$\gamma =$ es la razón de mortalidad per-cápita.

$q =$ es que una persona susceptible tenga mayor probabilidad de infectarse de una persona infecciosa que de una latente.

1. Una ecuación de observación:

$$\mathbf{y}_i = \mathbf{X}(t_i) + \boldsymbol{\epsilon}_i, \quad i \in 1, \ldots, n \tag{4.2}$$

donde:

$$\mathbf{y}_i = \begin{bmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \\ y_{4i} \end{bmatrix} \quad \mathbf{X}_i = \begin{bmatrix} S(t) \\ E(t) \\ I(t) \\ R(t) \end{bmatrix} \quad \epsilon_i = \begin{bmatrix} \epsilon_{1i} \\ \epsilon_{2i} \\ \epsilon_{3i} \\ \epsilon_{4i} \end{bmatrix} \quad \epsilon_{1i}, \epsilon_{2i}, \epsilon_{3i}, \epsilon_{4i} \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

Los parámetros de interes para el modelo SEIR son $\boldsymbol{\theta} = (\beta, q, \delta, \gamma)$

La función de verosimilitud como se definió en (2.1), para las observaciones $\mathbf{y}_i$ esta dada por:

$$\begin{aligned} L(\mathbf{y}|\boldsymbol{\theta}) &= \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi}} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{y}_i - \mathbf{x}_i)^t \Sigma^{-1} (\mathbf{y}_i - \mathbf{x}_i) \right] \right\} \\ &= \frac{1}{(2\pi)^{n/2}} |\Sigma|^{-n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{x}_i)^t \Sigma^{-1} (\mathbf{y}_i - \mathbf{x}_i) \right\} \end{aligned}$$

Para incorporar el conocimiento disponible acerca de $\boldsymbol{\theta}$, se asignan las siguientes distribuciones a priori´s no informativas:

$$p(\boldsymbol{\theta}) = p(\theta_1) \times p(\theta_2) \times p(\theta_3) \times p(\theta_4) \times p(\theta_5)$$

donde

$$p(\theta_i) = \frac{1}{\Gamma(a) b^a} \boldsymbol{\theta}_i^{a-1} \exp(-\boldsymbol{\theta}_i / b); 0 < \theta < \infty$$

$$p(\theta_5) = \frac{1}{b-a}, \quad a \leq \theta_5 \leq b$$

Un método MCMC será empleando, en este caso el método de Metropolis-Hastin vía t-walk en R para determinar $p(\boldsymbol{\theta}|\mathbf{y})$.

4.1. Resultados

4.1.1. Ejemplo datos reales

Para nuestro trabajo se tomaron datos publicados por la OMS del brote del virus de Ébola del año 2014 en los países de Guinea, Liberia y Sierra Leona. Se hizó una simulación en Metropolis-Hastings vía t-walk con datos genenerados por el método FracPECE. Tenemos los siguientes valores de los parámetros en el clasico (3.1) y el modelo fraccional (3.2)

$$TP = 18805278, \quad \delta = 1/12, \quad \gamma = 1/7$$

donde TP es la población total de las tres ciudades consideradas. Más aún, las condiciones iniciales siguientes son fijas en los experimentos numéricos realizados.

$$S(0) = TP\frac{m}{100} \quad E(0) = 0, \quad I(0) = 15, \quad R(0) = 0$$

La población susceptible es un porcentaje del total de la población. Note que en este caso $N = TP\frac{m}{100} + 15$. Los números iniciales de personas infectadas es fijado en 15 por la OMS siguiendo la enfermedad del ébola virus -datos actualizados el 27 de marzo de 2014, "hasta la fecha, 15 casos tienen pruebas positivas por prueba PCR para virus del Ébola, confirmado por la colaboración de los laboratorios Institut Pasteur Lyon, France, Institut Pasteur (IP) Dakar, Senegal y Bernhard-Nocht Institute of tropical Medicine Hamburg, Alemania. Estudios de laboratorios demostraron que el virus Ébola Zaire es el virus responsable por el brote".

Se muestra a continuación el algoritmo FracPECE con los datos

Predice

$$y^P_{n+1} = (16924750, 0, 15, 0) + \frac{1}{\Gamma(\alpha)}\sum_{j=0}^{n} b_{j,n+1} f(t_j, y_j)$$

donde $f(t_j, y_j) = (-b1 * S * (q * E + I)/Nt, b1 * S * (q * E + I)/Nt - d * E, d * E - g * I, g * I)$

Evalúa

$$f(t_{n+1}, y^P_{n+1})$$

donde $y_{n+1} = \boldsymbol{x}(t_i; \boldsymbol{\theta}) + \epsilon_i; \quad \epsilon_i \underset{iid}{\sim} N(0, \sigma^2 I)$

Corrige

$$y_{n+1} = y_0 + \frac{1}{\Gamma(\alpha)} \left[\sum_{j=0}^{n} a_{j,n+1} f(t_j, y_j) + a_{n+1,n+1} f\left(t_{n+1,y_{n+1}^P} \right) \right]$$

Evalúa

$$f(t_{n+1}, y_{n+1})$$

Se hizo una simulación en Metropolis-Hastings vía t-walk con datos genenerados por el método FracPECE, en este caso se considero $m = 90$, $q = 0.058$ y $\alpha = 0.9$; y encontramos una estimación a los parámetros de $\theta = (\beta, q, \delta, \gamma)$, los cuales son

$$\theta = (0.21607132, 0.07399887, 0.12849042, 0.14079145)$$

A continuación se muestran los gráficos de los ajustes del fracPECE con los datos reales tomados de la página de la OMS.

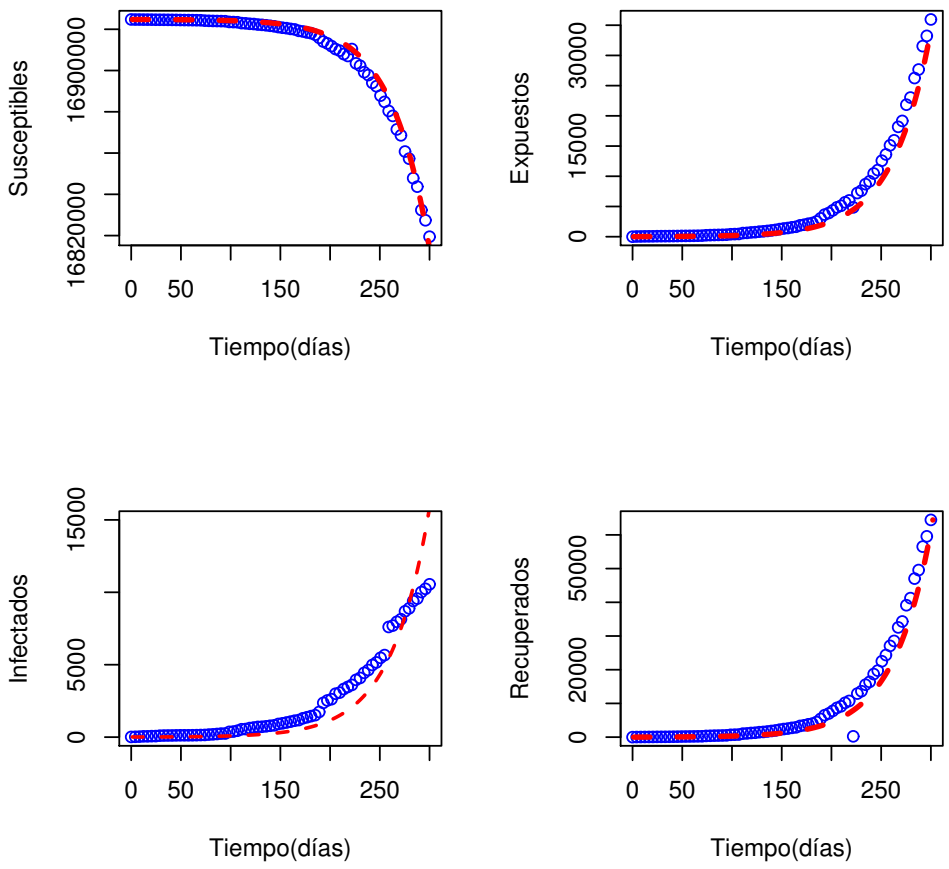

Figura 4.1: Comparación entre el modelo FracPECE y los datos de la OMS (Circulos azules son los datos reales y las líneas punteadas rojas son los ajustes por FracPECE.)

En la Figura 4.1 se muestran los ajustes de los datos reales vs datos FracPECE, se observa que las estimaciones proporcionadas por el FracPECE ajusta bien a los datos reales.

4.1.2. Ejemplo con datos simulados

En esta sección se muestran los resultados de las simulaciones hechas por el algoritmo de Metropolis-Hasting vía t-walk, en cada simulación de las cadenas se ejecuta el método FracPECE con 500 datos y se obtiene una simulación de cadena de 10,000 iteraciones con los parámetros $\beta = 0.2315$, $q = 0.058$, $\delta = .0833$, $\gamma = 0.1428$ y $\alpha = 0.9$.

Se muestra a continuación los histogramas de las 3 cadenas.

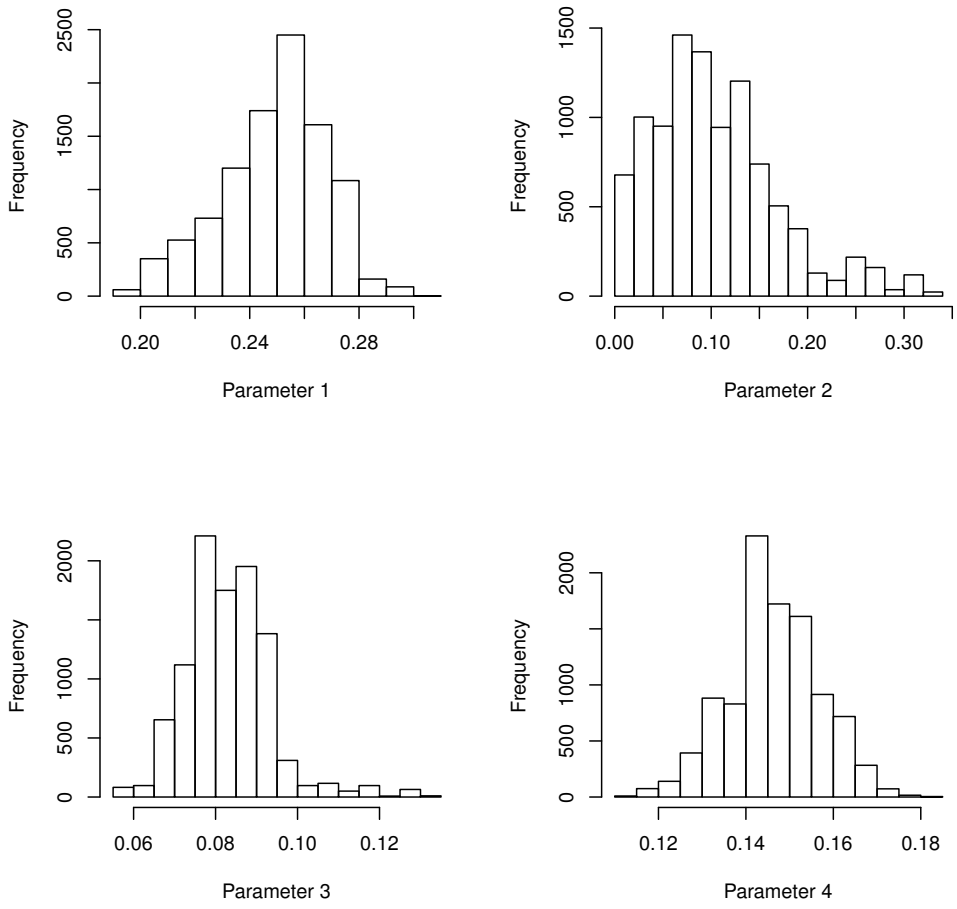

Figura 4.2: Histogramas de las distribuciones posteriores.

Se muestra el comportamiento de las trazas para las 3 cadenas generadas por el método de Metropolis-Hasting vía t-walk, y las trazas de las tres cadenas en un solo gráfico. (Ver figura 4.3)

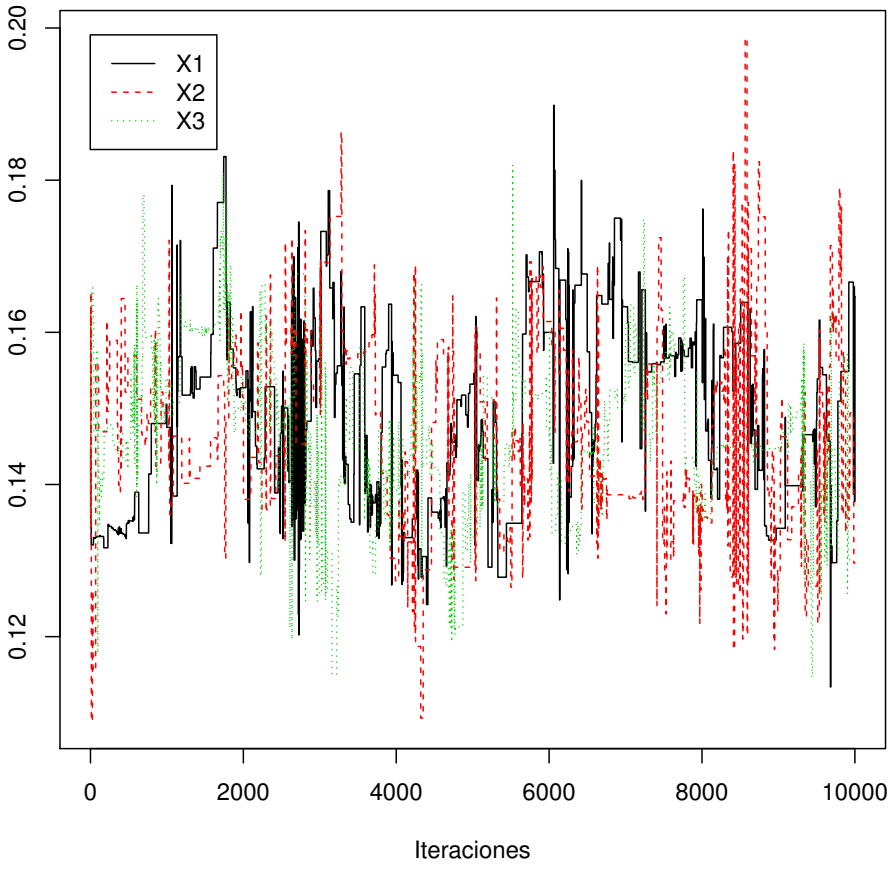

Figura 4.3: Traza de las cadenas.

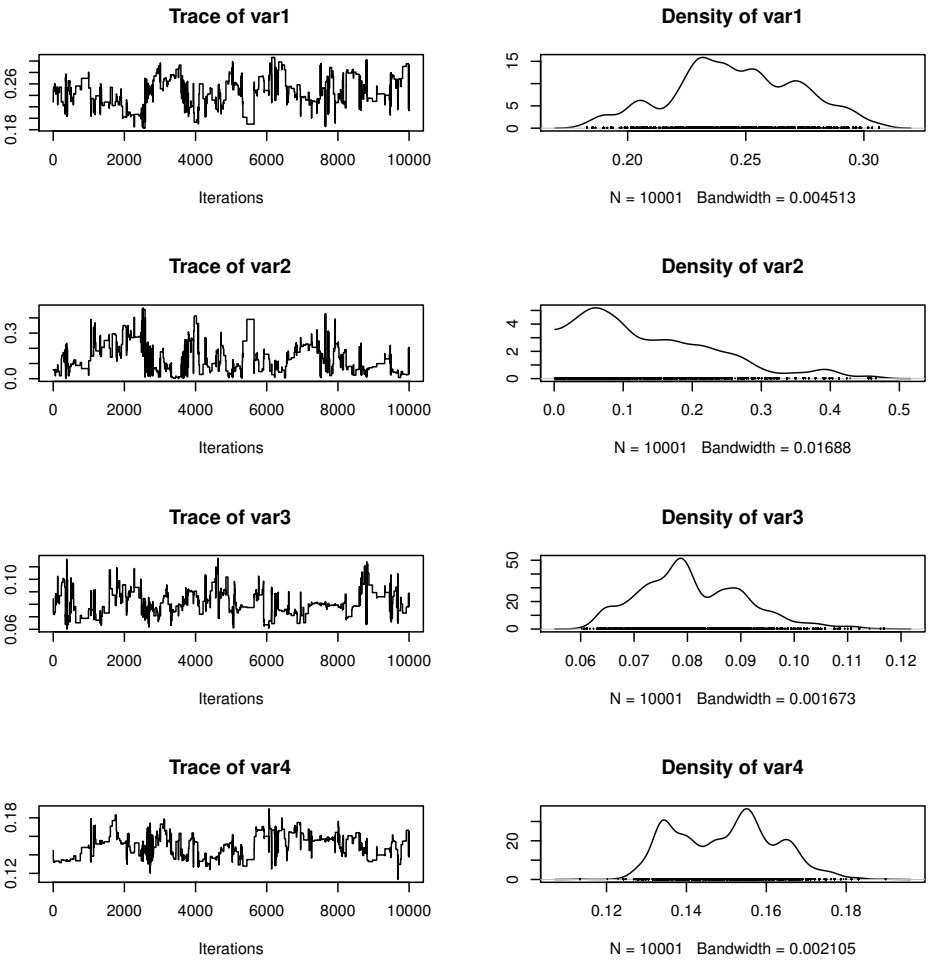

Figura 4.4: Trazas y densidades de β, q, δ y γ de la 1ra. cadena.

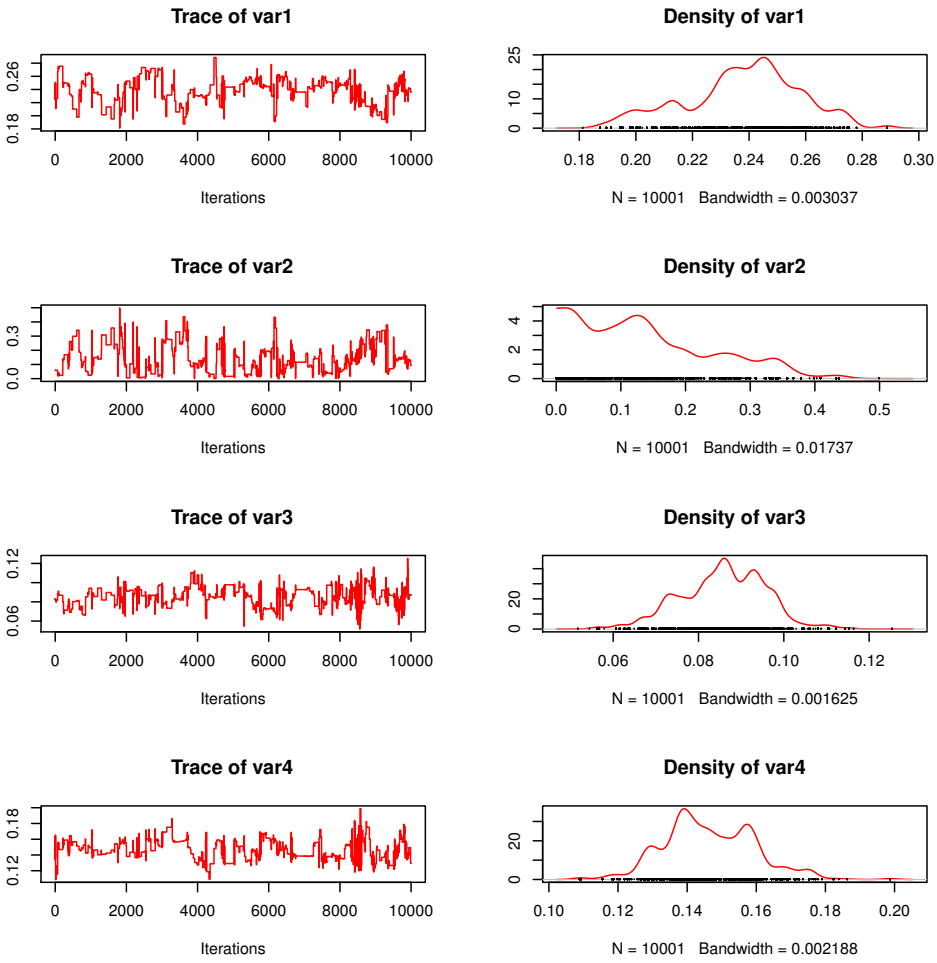

Figura 4.5: Trazas y densidades de β, q, δ y γ de la 2da. cadena.

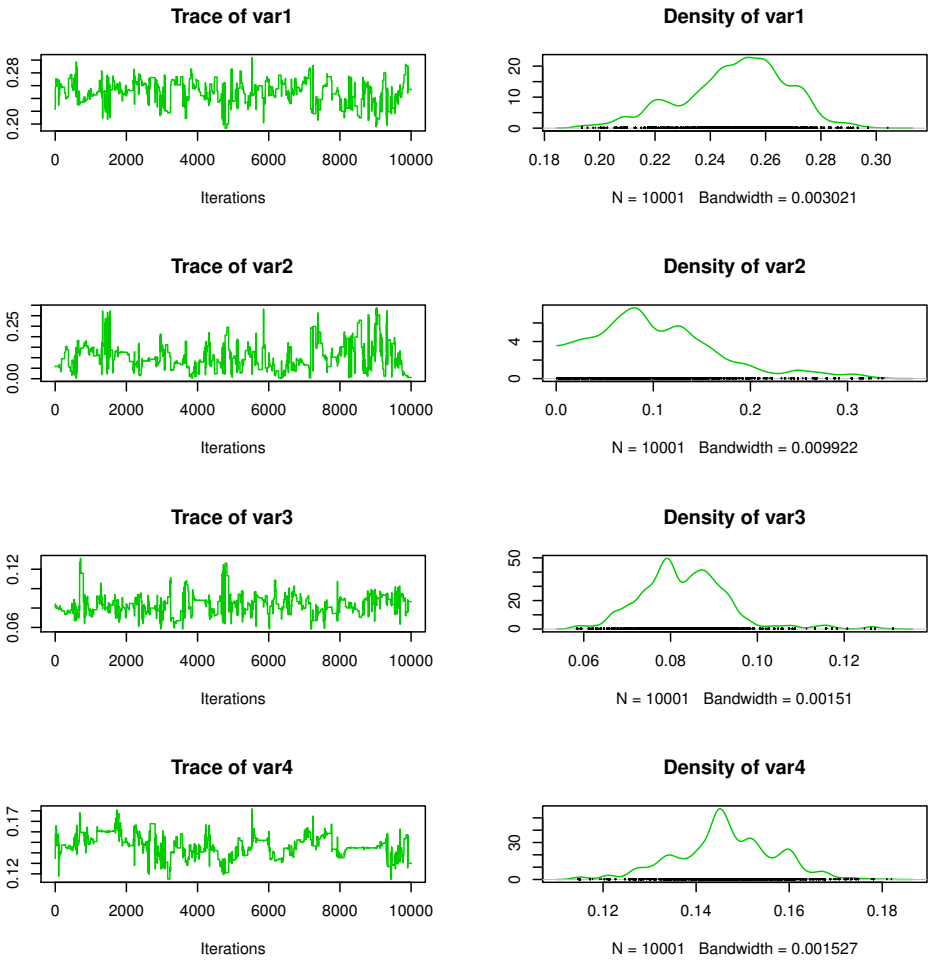

Figura 4.6: Trazas y densidades de β, q, δ y γ de la 3ra. cadena.

4.1.3. Diagnóstico de convergencia de Gelman y Rubin.

Dado que tenemos más de dos cadenas se aplica el diagnóstico de Gelman y Rubin, entonces se obtienen los resultados en la tabla 4.1. Se observa que el intervalo de confianza superior al 97.5 % y con 5000 iteraciones son suficientes para lograr convergencia para β, q, δ, γ. Y que las muestras de 5001-1000 de las tres cadenas pueden suponerse que surgen de las distribuciones marginales posteriores para cada variable.

Tabla 4.1: Resultados de Gelman y Rubin para las cadenas.

Variable	Point est.	Upper C.I.
β	1.09	1.26
q	1.03	1.09
δ	1.07	1.21
γ	1.10	1.30

Se genera un gráfico de Gelman y Rubin como se muestra en la figura 4.7. Estas secciones se producen al dividir la cadena para cada variable. El diagnóstico de Gelman y Rubin se calcula para cada segmento, y la mediana y el cuantil al 97.5 % para β δ y γ se estabilizan alrededor de 1, para los segmentos de la cadena que contienen las primeras 4500 iteraciones o más. Puesto que el diagnóstico se calcula a partir de la segunda mitad de cada cadena, se sugiere que la convergencia fue alcanzada después de 2250 iteraciones. Los factores de reducción estimados para q parece haber estabilizado alrededor de 1 para segmentos de cadena mayores de 5000 iteraciones.

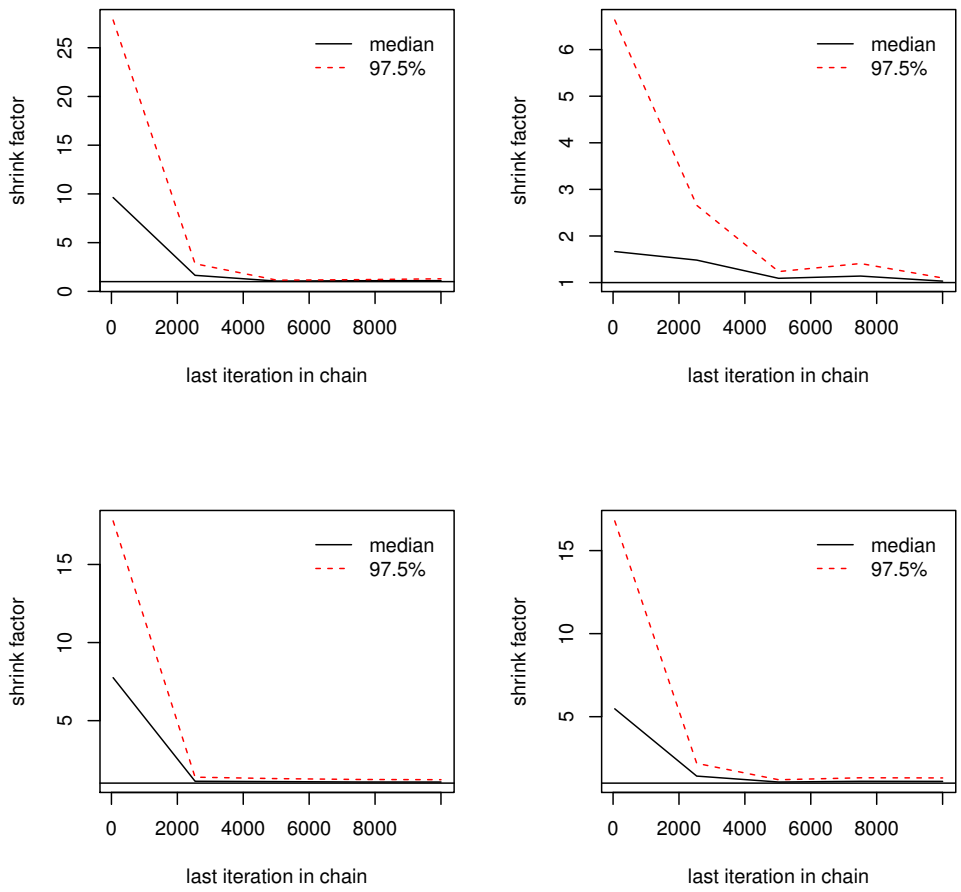

Figura 4.7: Factor de adelgazamiento de Gelman y Rubin para β, q, δ y γ

4.1.4. Diagnóstico de convergencia de Heidelberger y Welch.

La convergencia se logró inmediatamente en la primera ejecución para cada variable para la 1ra. cadena. El Halfwidth test indica que las iteraciones de 2-10001 deben proporcionar estimaciones de los medios posteriores para β, q, δ, γ que cumplen con el criterio de precisión, mientras que para β, δ y γ pasan el test de haflwidth test, las 10000 muestras no proporcionan una estimación suficientemente precisa para q. (Ver Tabla 4.2).

Tabla 4.2: Tabla de resultados de Heidelberger y Welch para la 1ra. cadena.

Variable	Stationarity test	start iteration	p-value
β	passed	1	0.144
q	passed	1	0.202
δ	passed	1	0.648
γ	passed	1	0.373
	Halfwidth test	Mean	Halfwidth
β	passed	0.2455	0.00893
q	failed	0.1373	0.03434
δ	passed	0.0809	0.00296
γ	passed	0.1498	0.00476

La convergencia se logró inmediatamente en la primera ejecución para cada variable para la 2da. cadena. El Halfwidth test indica que las iteraciones de 2-10000 deben proporcionar estimaciones de los medios posteriores para β, q, δ, γ que cumplen con el criterio de precisión, mientras que para β, δ y γ pasan el test de haflwidth test, las 10000 muestras no proporcionan una estimación suficientemente precisa para q. (Ver Tabla 4.3)

Tabla 4.3: Tabla de resultados de Heidelberger y Welch para la 2da. cadena.

Variable	Stationarity test	start iteration	p-value
β	passed	1	0.491
q	passed	1	0.432
δ	passed	1	0.500
γ	passed	1	0.568
	Halfwidth test	Mean	Halfwidth
β	passed	0.2378	0.00676
q	failed	0.1418	0.03192
δ	passed	0.0858	0.00258
γ	passed	0.1464	0.00404

La convergencia se logró inmediatamente en la primera ejecución para las variables β, q, δ y después de 2002 iteraciones para la variable γ de la 3ra. cadena. El Halfwidth test indica que las iteraciones de 2-10000 deben proporcionar estimaciones de los medios posteriores para β, q y δ, y 2003-1000 iteraciones deben proporcionar estimaciones posteriores para γ y que cumplen con el criterio de precisión, mientras que para β, δ y γ pasan el haflwidth test, las 10000 muestras no proporcionan una estimación suficientemente precisa para q. (Ver Tabla 4.4)

Tabla 4.4: Tabla de resultados de Heidelberger y Welch para la 3ra. cadena.

Variable	Stationarity test	start iteration	p-value
β	passed	1	0.257
q	passed	1	0.578
δ	passed	1	0.175
γ	passed	2002	0.616
	Halfwidth test	Mean	Halfwidth
β	passed	0.2491	0.00463
q	failed	0.1036	0.01681
δ	passed	0.0832	0.00223
γ	passed	0.1444	0.00374

En la Tabla 4.5, se muestra las estimaciones obtenidas de las cadenas.

Tabla 4.5: Estimación de las cadenas.

	β	q	δ	γ	α
Original	0.2315	0.058	0.0833	0.1428	0.9
Cadena 1	0.2455	0.1472	0.08092	0.1498	1.5047
Cadena 2	0.2378	0.1418	0.08579	0.1464	0.9434
Cadena 3	0.2491	0.1036	0.08320	0.1464	0.8504

Ahora se muestran los resultados con 10000 simulaciones realizadas con el Metropolis-Hasting y 500 datos sintéticos generados por el método FracPECE, con un burn-in de 1000 y un adelgazamiento de 5.

En la tabla 4.6, se muestra las medias obtenidas de las cadenas y se comparan con las medias originales que se utilizaron al generar las simulaciones.

Tabla 4.6: Estimacion de las cadenas.

	β	q	δ	γ	α
Original	0.2315	0.058	0.0833	0.1428	0.9
Cadena 1	0.2450	0.1432	0.0812	0.1511	1.5047
Cadena 2	0.2372	0.1417	0.0865	0.1461	0.9434
Cadena 3	0.2487	0.1033	0.0832	0.1469	0.8504

Capítulo 5

Conclusiones

En este trabajo de tesis se hizo la estimación, desde un punto de vista Bayesiano, de los parámetros involucrados en un modelo SEIR fraccionario que describe el número de enfermos e una población infectada por el virus del Ébola en los países de Guinea, Liberia y Sierra Leona en el continente africano con brote de virus Ébola en el año 2014. Se implemento el algoritmo para obtener simulaciones numéricas a la solución del sistema de ecuación no lineal tipo SEIR fraccionario, donde se observó un buen ajuste de los datos obtenidos por simulación a los datos dados por la Organización Mundial de la Salud (OMS). Se estimó las funciones de verosimilitud de los datos y las aprioris obtenemos las distribuciones marginales de cada parámetro vía MCMC en cada iteración del t-walk donde se incluye el método FracPECE para estimar las muestras bajo aproximaciones.

Referencias

[1] Francisco J. Ariza-Hernandez; Jorge Sanchez-Ortiz; Martin P. Arciga-Alejandre y Luis X. Vivas-Cruz. *Bayesian Analysis for a Fractional Population Growth Model*. Journal of Applied Mathematics,Volume 2017 (2017), Article ID 9654506, 9 pages. https://doi.org/10.1155/2017/9654506.

[2] Chib, S. y Greenberg, E. (1995). *Understanding the metropolis-hastings algorithm.* The american statistician, 49(4):327-335.

[3] Christen, J. A.; y Fox, C. (2010). *A general purpose sampling algorithm for continuous distributions (the t-walk)*. Bayesian Analysis, 5(2):263-282.

[4] Diethelm Kai y D. Freed, Alan. (1999). *The FracPECE Subroutine for the Numerical Solution of Differential Equations of Fractional Order*.

[5] George casella y Edward I. George, *Explaining the Gibbs Sampler*, The American Statistician, Vol. 46, No. 3. (Aug., 1992), pp. 167-174.

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

[7] Gelman, Andrew; Rubin, Donald B. *Inference from Iterative Simulation Using Multiple Sequences*. Statist. Sci. 7. (1992), no. 4, 457–472. doi:10.1214/ss/1177011136. https://projecteuclid.org/euclid.ss/1177011136.

[8] Ivan Area; Hanan Batarfi;Jorge Losada; Juan J. Nieto; Wafa Shammakh, y Ángela Torres; *On a fractional order Ebola epidemic model*; Advances in Difference Equations. Springer Open Journal.2015.

[9] 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.

[10] John Geweke. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. Staff Report 148, Federal Reserve Bank of Minneapolis.

[11] Heidelberger P.; Welch, Peter D. *Simulation run length control in the presence of an initial transient*. Vol. 31, No. 6, Simulation (Nov. - Dec., 1983), pp. 1109-1144.

[12] Jaime Astacio; DelMar Briere; Milton Guillén; Josué Martínez; Francisco Rodríguez y Noé Valenzuela-Campos. (1996). *Mathematical Models to Study the Outbreaks of Ebola.*Biometrics Unit Technical Reports; Number BU-1365-M, 18 Pags.

[13] Kai Diethelm; Neville J. Ford y Alan D. Freed. *A predictor-Corrrector Approach for the Numerical Solución of Fractional (2002). Differential Equation*. 29:3. https://doi.org/10.1023/A:1016592219341. Kluver Academic Publishers.

[14] Lee, P. M. (2012). *Bayesian statistics: an introduction.* Jhon Wiley & Sons.

[15] Matt J. Keeling; Pejman Rohani, *Modeling Infectious Diseases in Humans and Animals*. 2008. Princeton University Press. USA.

[16] Siddhartha Chib; Edward Greenberg, *Understanding the Metropolis-Hasting Algorithm*, The American Statistician, Vol.49, No. 4 (Nov., 1995), pp. 327-335.

[17] http://www.nationalgeographic.com.es/ciencia/grandes-reportajes/a-la-caza-del-asesino-2_9686/23.

[18] Tarantola, A. (2005). Inverse problem theory and methods for model parameter estimation. siam.

[19] George Casella; Roger L. Berger. (2001) *Statistical Inference*. Duxbury Press.

[20] kaipio, j.; y Somersalo, E. (2006). *Statistical and computational inverse problems*, volume 160. Springer Science & Business Media.

[21] Plummer, M. (2012). *Jags: A program for analysis of Bayesian graphical models using gibbs sampling.*

[22] Spiegelhalter, D. J. (2003). *WinBUGS version 1.4 User Manual*