Centro de Investigación en Matemáticas, A.C.

CIMAT

Inferencia de modelos epidemiológicos compartimentales en redes sociales.

T E S I S

Que para obtener el grado de
Maestría en Ciencias
con especialidad en
Probabilidad y Estadística

P r e s e n t a
Rocío Maribel Ávila Ayala

Directora de tesis:
Dra. L. Leticia Ramírez Ramírez

Guanajuato, Gto. Noviembre de 2016

Centro de Investigación en Matemáticas, A.C.

Acta de Examen de Grado

Acta No.: 111

Libro No.: 002

Foja No.: 111

En la Ciudad de Guanajuato, Gto., siendo las 16:00 horas del día 14 de noviembre del año 2016, se reunieron los miembros del jurado integrado por los señores:

DR. JOSÉ ANDRÉS CHRISTEN GRACIA (CIMAT)
DR. MARCOS AURELIO CAPISTRÁN OCAMPO (CIMAT)
DRA. LILIA LETICIA RAMÍREZ RAMÍREZ (CIMAT)

bajo la presidencia del primero y con carácter de secretario el segundo, para proceder a efectuar el examen que para obtener el grado de

MAESTRO EN CIENCIAS
CON ESPECIALIDAD EN PROBABILIDAD Y ESTADÍSTICA

Sustenta

ROCÍO MARIBEL ÁVILA AYALA

en cumplimiento con lo establecido en los reglamentos y lineamientos de estudios de posgrado del Centro de Investigación en Matemáticas, A.C., mediante la presentación de la tesis

"INFERENCIA DE MODELOS EPIDEMIOLÓGICOS COMPARTIMENTALES EN REDES SOCIALES"

Los miembros del jurado examinaron alternadamente al (la) sustentante y después de deliberar entre sí resolvieron declararlo (a):

Aprobada

DR. JOSÉ ANDRÉS CHRISTEN GRACIA
Presidente

DR. MARCOS AURELIO CAPISTRÁN OCAMPO
Secretario

DRA. LILIA LETICIA RAMÍREZ RAMÍREZ
Vocal

CIMAT
DIRECCIÓN GENERAL

DR. JOSÉ ANTONIO STEPHAN DE LA PEÑA MENA
Director General

Resumen

En el contexto de epidemiología, resulta sumamente importante el uso de modelos matemáticos para describir la dispersión de un agente infeccioso en una población con fines de aprender sobre las características infecciosas del agente, o bien, de predecir escenarios futuros para un brote.

Una forma de modelar matemáticamente el proceso de dispersión de un agente infeccioso, es mediante modelos epidemiológicos compartimentales, los cuales consisten en clasificar a los individuos de una población en distintas categorías de acuerdo a su estado respecto a la enfermedad. Estos modelos dependen de parámetros que regulan el contagio y la recuperación de los individuos, por ejemplo, y que son determinantes en el desarrollo y magnitud de un brote.

Este trabajo tiene como objetivo revisar y proponer, en algunos casos, metodologías para realizar inferencia estadística, desde el enfoque bayesiano, de los parámetros que rigen este tipo de modelos. Los modelos compartimentales se consideran, a su vez, bajo dos diferentes tipos de regímenes. El primero se denomina "ley de acción de masas" y corresponde al escenario en donde el número de contactos de cada individuo en la población es muy similar. El segundo tipo incorpora una red de contactos y puede considerar poblaciones con contactos individuales que son muy diferentes entre sí.

Agradecimientos

A Dios, por permitirme concluir este proyecto. A mis padres y mis hermanos, por su amor y apoyo incondicional. Son el impulso que me ayuda a salir adelante. ¡Los amo!.

Al CIMAT y todos mis profesores. Gracias por ayudarme a crecer tanto intelectual como personalmente, y por su interés en la formación integral de todos los alumnos del posgrado. Al CONACyT por la beca No. 392688, la cual permitió mi traslado y estancia en Guanajuato para la realización de mis estudios de maestría.

A mi asesora, Lety Ramirez, por su orientación, paciencia y gran apoyo en la realización de este trabajo, y por su entusiasmo en mi formación y crecimiento. A mis sinodales, el Dr. Andrés Christen y el Dr. Marcos Capistrán, por su ayuda invaluable en la revisión de mi tesis.

Al Dr. Arturo Erdely por haber despertado mi interés en la estadística años atrás, por su amistad y sus sabios consejos.

A Fer, Nat, Diana y Armando, por cuidar de nuestra amistad y conservarla a pesar de la distancia y por apoyarme emocionalmente siempre que lo necesité; y a mis amigos de la maestría, en especial Irving, Cris y Jesús, por siempre ayudarme y acompañarme en las buenas y en las malas. ¡Los quiero mucho!

Índice general

Introducción

Un paso importante en el método científico es la experimentación, sin embargo, en el contexto de epidemiología no es concebible (ni sería ético) experimentar sobre una población para ver cómo se dispersa un agente infeccioso. Esto hace que en esta área sea sumamente importante el planteamiento de modelos matemáticos que intenten describir la evolución de una enfermedad y que al mismo tiempo puedan incorporar resultados de laboratorio o estimaciones originadas de brotes anteriores. Al tener un modelo bien determinado, es posible plantearse distintos escenarios bajo los cuales se desarrolle la dipersión de la enfermedad, y establecer así políticas públicas que intenten erradicar o disminuir la magnitud de un brote.

Una gran parte de los modelos epidemiológicos considera que un individuo puede transitar por diversos estatus al ser infectado durante un brote. Esta clasificación individual lleva naturalmente a dividir a la población de estudio en grupos o categorías disjuntas de acuerdo a su estado respecto a la enfermedad. Este tipo de modelos epidemiológicos se denominan modelos compartimentales.

Por otro lado, los modelos epidemiológicos pueden dividirse en dos grandes categorías: deterministas o estocásticos. En un modelo determinista se considera que se tiene control o conocimiento absoluto de los factores que intervienen en el estudio del proceso o fenómeno y por tanto se pueden predecir con exactitud sus efectos. En un modelo estocástico no es posible controlar todos los factores que intervienen y en consecuencia no se tienen resultados únicos. Debido a esta variabilidad en los modelos estocásticos, los resultados están acompañados con distribuciones o afirmaciones probabilísticas. Por ejemplo, si modelamos la propagación de una enfermedad usando un modelo compartimental en una población compuesta de N individuos, un modelo determinista nos permite obtener para un tiempo fijo t el número (o proporción) de individuos en cada uno de los compartimentos, mientras que usando un modelo estocástico se obtendría la probabilidad de que se tengan (i_0, i_1) individuos infectados (con $i_0 < i_1$).

Supóngase que se tiene un modelo compartimental con K categorías mutuamente excluyentes y que la salida de un grupo implica la entrada inmediata a uno de los otros. Siguiendo la notación de Haran (2009), para un tiempo fijo t, se puede definir un sistema vector-valuado $\mathbf{X}(t) = (X_1(t), \ldots, X_K(t))$, donde $X_i(t)$ es un conteo que corresponde al número de individuos en el i-ésimo compartimento, o un número real en el intervalo $[0, 1]$, cuando lo que se describe es la proporción de

individuos en cada uno de los grupos. Asociado a estos estados, se puede definir un proceso de flujo para cada par de compartimentos i, j. Denotemos $\mathcal{N}_{ij}(t_2) - \mathcal{N}_{ij}(t_1)$ el número de transiciones del compartimento i al j entre los tiempos t_1 y t_2, con $t_1 < t_2$. Este flujo está asociado con las tasas de transferencia entre compartimentos $\nu_{ij} \geq 0$, $i, j \in \{1, \ldots, K\}$.

En los modelos epidemiológicos compartimentales existe un parámetro umbral que determina si el número de infectados decrece rápidamente hasta desaparecer, o si la enfermedad se propaga en una gran parte de la población y se presenta un brote. Este parámetro umbral se conoce como R_0 y es una función de las tasas de transferencia entre los compartimentos.

Dado un conjunto de observaciones, por ejemplo pensemos en reportes de nuevos infectados cada cierto intervalo de tiempo, y un modelo epidemiológico que describa adecuadamente los datos observados (puede consultarse Herrera Reyes, 2010 para mayor detalle sobre el tema de selección de modelos), es relevante poder realizar inferencia sobre los parámetros que rigen el modelo para los datos, ya que esto permitiría un mejor entendimiento del comportamiento de la epidemia, y además el establecimiento de medidas que ayuden a controlar el brote, evitando un impacto mayor en la salud de la población y en la economía.

El objetivo principal de este trabajo es presentar un proceso de inferencia en modelos epidemiológicos compartimentales planteados bajo escenarios distintos. Se supondrá que los datos observados son el número de nuevos infectados reportados en un intervalo de tiempo y se trabajará con modelos de tipo bayesiano.

El trabajo se divide en cuatro capítulos. En el Capítulo 1 se presentan las bases teóricas necesarias para el desarrollo de la inferencia de los modelos epidemiológicos que se presentan. Estos fundamentos se enfocan principalmente en métodos de simulación e inferencia bayesiana.

En el Capítulo 2 se revisa el planteamiento de los modelos compartimentales desde el enfoque determinista, el cual supone una población grande donde las interacciones entre los individuos son homogéneas y se rigen bajo la ley de acción de masas, por lo que el número de contactos de los individuos de la población es muy similar, y el sistema de ecuaciones diferenciales asociado al modelo tiene una solución única para un tiempo fijo. También se presenta una forma de realizar inferencia en este tipo de modelos y algunos ejemplos de su implementación.

Posteriormente, en el Capítulo 3 se añade un componente aleatorio relacionado con los tiempos en que los individuos permanecen en cada compartimento. De esta forma, los modelos compartimentales que aquí se consideran son cadenas de Markov de saltos puros. En este caso también se plantea la metodología de inferencia y se presenta un ejemplo.

Por último, en el Capítulo 4 se plantea el modelo estocástico en una red que representa las interacciones entre los individuos de una población. Este escenario es más realista que los anteriores cuando la población es pequeña o el agente infeccioso se dispersa siguiendo contactos que varían considerablemente entre los individuos de la población. Se presenta una adaptación del método de inferencia en el modelo estocástico y un ejemplo en el modelo SIR.

La implementación de los métodos de inferencia se programó usando el software estadístico `R` (R Development Core Team, 2008) y el código de programación que se utiliza a lo largo de este trabajo puede consultarse en `http://leticiaramirez2.net/supplementary_material.html`.

Capítulo 1

Fundamentos teóricos

En este capítulo se abordan las bases de los elementos estadísticos utilizados para el desarrollo de los procesos de inferencia que se abordarán en los capítulos siguientes.

En la Sección 1.1 se explica brevemente el paradigma bayesiano, y cómo este esquema de modelación incorpora, además de la información proporcionada por los datos (reflejada en la verosimilitud), información adicional obtenida de expertos sobre el tema que puede ser valiosa y aportar ventajas al considerarla en el modelo. Es importante dar un breve resumen sobre este tópico, ya que la inferencia en los modelos compartimentales desarrollada más adelante se basa en modelos bayesianos.

Posteriormente, en la Seccion 1.2 se describen algunos algoritmos MCMC que permiten simular de una densidad posterior en ocasiones que se dificulta su obtención de manera explícita. En particular, se hace especial énfasis en el algoritmo Metrópolis-Hastings, puesto que el *t-walk* y el algoritmo ABC-MCMC son casos particulares de éste que se usarán para la simulación de la densidad posterior de los parámetros de los modelos compartimentales abordados en los capítulos posteriores.

1.1. Inferencia Bayesiana

El primer paso a realizar en la modelación estadística paramétrica, es elegir una familia paramétrica $\mathcal{P} = \{f(y \mid \theta) : \ \theta \in \Theta\}$ que logre describir el comportamiento probabilístico del fenómeno aleatorio de estudio, pero aún queda la incertidumbre acerca de cuál θ elegir para que el modelo esté definido de forma explícita. En contraste con el enfoque clásico o frecuentista, en estadística bayesiana se modela a θ, el parámetro del modelo probabilístico general $\pi(y \mid \theta)$, como una variable (o vector) aleatoria cuya ***distribución de probabilidad a priori o inicial*** $\pi(\theta)$ está basada en información previa. Elegir tal distribución es todo un tema, existen varios métodos para capturar la información subjetiva en una distribución de probabilidad para θ y también hay formas de expresar en dicha

distribución la incertidumbre acerca del parámetro mediante una distribución a priori poco informativa[1].

Ya que se cuenta con la *distribución a priori*, se procede a obtener la muestra, y la distribución inicial se actualiza con las observaciones $\mathbf{x} = (x_1, \ldots, x_n)$ conforme a la Regla de Bayes para obtener una ***distribución posterior o final***:

$$\pi(\theta \,|\, \mathbf{y}) = \frac{f(\mathbf{y} \,|\, \theta)\pi(\theta)}{\int_\Theta f(\mathbf{y} \,|\, \vartheta)\pi(\vartheta)\, d\vartheta} \tag{1.1}$$

Obsérvese que la distribución $\pi(\theta \,|\, \mathbf{y})$ contempla tanto la información muestral como la información incorporada por $\pi(\theta)$. El denominador de (1.1) es una constante que típicamente en dimensiones paramétricas mayores es difícil de calcular de manera explícita, sin embargo existen métodos para aproximarlo numéricamente.

La característica primordial que hace interesante al enfoque bayesiano es que éste, además de usar información muestral, hace posible incorporar de manera consistente al modelo información subjetiva *(a priori)* derivada de las creencias previas a la realización del experimento (experiencia de expertos, información histórica, etc.) y aún con muestras pequeñas proporciona inferencias aceptables (Box y Tiao, 2011; Bernardo y Smith, 2001).

En la metodología bayesiana, una vez hallada la distribución posterior ya es posible plantear cualquier problema de inferencia: estimación puntual, estimación por intervalos, contrastes de hipótesis, etc. (Bernardo y Smith, 2001; Blasco, 2005).

1.2. Algoritmos MCMC

Los algoritmos MCMC son métodos de simulación basados en una Cadena de Markov; es decir, las variables generadas serán dependientes. De esta forma se deberán sacrificar observaciones intermedias para obtener una muestra pseudoindependiente. La ventaja de este tipo de algoritmos radica en que una cadena de Markov puede tener propiedades de convergencia que facilitan la simulación de una función objetivo π arbitraria a partir de una distribución adicional q de la cual es sencillo simular.

El objetivo es simular variables aleatorias $X_1, \ldots, X_n$ que se distribuyan aproximadamente como π sin simular directamente de esta distribución. Los métodos

[1] Para mayor detalle acerca de distribuciones a priori, véase Robert (2007).

MCMC logran esto usando una cadena de Markov ergódica que tenga como distribución estacionaria π.

Existen diversos esquemas que producen kerneles de transición válidos asociados con distribuciones estacionarias arbitrarias, pero se conserva el esquema general que consiste en proponer un punto inicial $x^{(0)}$ dentro del soporte de π y generar una cadena $\left\{X^{(t)}\right\}$ usando una densidad propuesta con distribución estacionaria π.

1.2.1. Algoritmo Metrópolis-Hastings

Este algoritmo debe su nombre a Nicholas Metrópolis, que publicó un artículo en conjunto con otros autores, donde se proponía por primera vez el algoritmo en el caso de propuestas simétricas (Metropolis et al., 1953). Posteriormente, Hastings (1970) extendió el algoritmo al caso más general.

El algoritmo Metrópolis-Hastings hace uso de una densidad condicional $q(y \mid x)$ definida respecto a la medida dominante del modelo (Robert y Casella, 2013), y puede ser implementado en la práctica cuando es sencillo simular de $q(\cdot \mid x)$, ya sea que se pueda simular explícitamente de ella o de algo proporcional (salvo una constante multiplicativa que no dependa de x); o bien, si es simétrica, es decir, que $q(x \mid y) = q(y \mid x)$.

El algoritmo M-H asociado con la densidad objetivo π y la densidad condicional q (a la cual se le llama densidad instrumental, propuesta o kernel de transición), produce una cadena de Markov $\left\{X^{(t)}\right\}$ mediante las transiciones mostradas en el Algoritmo 1.

A $\alpha(x \mid y)$ se le llama probabilidad de aceptación de Metrópolis-Hastings y a $\{\pi(y)q(x \mid y)\} \,/\, \{\pi(x)q(y \mid x)\}$ se le llama razón de Metrópolis-Hastings. Para revisar las condiciones que se requieren para la convergencia del método y mayor detalle sobre las propiedades asintóticas ver la Sección 7.3.2 de Robert y Casella (2013).

El Algoritmo 1 siempre acepta valores de y_t para los cuales el cociente $\pi(y_t)/q(y_t \mid x^{(t)})$ se incrementa comparado con el valor anterior $\pi(x^{(t)})/q(x^{(t)} \mid y_t)$. En el caso en que la propuesta q es simétrica, este término se cancela en la razón de M-H y se acepta un punto de acuerdo al valor del cociente $\pi(y_t)/\pi(x^{(t)})$. Este método también es capaz de aceptar puntos para los cuales la razón descrita anteriormente no se

Algoritmo 1: Metrópolis-Hastings.

Dado un punto $x^{(t)}$,

1. Generar $Y_t \sim q(y \,|\, x^{(t)})$

2. Hacer

$$X^{(t+1)} = \begin{cases} Y_t & \text{con probabilidad } \alpha\left(x^{(t)}, Y_t\right), \\ x^{(t)} & \text{con probabilidad } 1 - \alpha\left(x^{(t)}, Y_t\right). \end{cases}$$

donde

$$\alpha\,(x, y) = \text{mín}\left\{\frac{\pi(y)}{\pi(x)}\,\frac{q(x \,|\, y)}{q(y \,|\, x)}, 1\right\}$$

incrementa, pero con probabilidad menor a uno.

La elección de la densidad propuesta q es crucial para la rápida convergencia de este algoritmo. Naturalmente, la densidad propuesta que haría que el método convergiera más rápidamente sería $q(y \,|\, x) = \pi(y)$, en cuyo caso la probabilidad de aceptación de M-H sería 1 y la convergencia se daría de manera inmediata, sin embargo, en el contexto de métodos MCMC suponemos que no se puede muestrear de π directamente. En el caso más general, es común proponer una caminata aleatoria para mover a los puntos (Random Walk Metrópolis-Hastings), en la cual la propuesta está dada por $Y_t = x^{(t-1)} + Z_t$, donde los incrementos $\{Z_t\}$ son v.a.i.i.d. provenientes de una distribución simétrica determinada, por ejemplo, $N(0, \sigma^2 I)$. En este caso, el problema se enfoca en cómo escalar la propuesta, que en el caso de la Normal, equivale a elegir σ^2 (a la cual se le llama tamaño de paso). Si σ^2 se elige muy pequeña, la cadena se moverá muy lentamente y producirá puntos muy correlaccionados, pero si es demasiado grande, las propuestas se rechazarán con facilidad.

Típicamente, la búsqueda de los parámetros de escalamiento para la propuesta se hace manualmente, a prueba y error; sin embargo, esto puede resultar tedioso y complicado, sobre todo en altas dimensiones. Un enfoque alternativo son los métodos MCMC *adaptativos*, los cuales hacen a la computadora *aprender* sobre mejores valores sobre los parámetros mientras el algoritmo corre. Esto se logra actualizando los parámetros de escalamiento de la densidad propuesta en cada iteración, a fin de encontrar un valor más adecuado. Se requieren ciertas condicio-

nes para garantizar la convergencia, las cuales pueden consultarse en Haario et al. (2001) y Andrieu y Thoms (2008).

1.2.2. El t-walk

El t-walk es un algoritmo propuesto por Christen et al. (2010) que permite simular de una función objetivo continua arbitraria. Es un método MCMC que se basa en dos puntos seleccionados de manera independiente dentro del espacio muestral y propone nuevos puntos que se aceptan o rechazan con probabilidad dada por el cociente de Metrópolis-Hastings, por lo que se puede probar su convergencia bajo las condiciones habituales. A diferencia de los métodos adaptativos, en este caso la propuesta es fija, y produce un algoritmo invariante a la escala que se considere, y aproximadamente invariante a transformaciones afines del espacio paramétrico.

La ventaja del t-walk es que no requiere parámetros de escalamiento ni adaptaciones de la propuesta, haciéndolo muy versátil para poder simular de una distribución objetivo continua. Se basa en un *kernel híbrido*, es decir, en una mezcla de kerneles estándar de Metrópolis-Hastings y únicamente requiere la evaluación de la distribución objetivo en el punto actual y el anterior, así como la evaluación de la propuesta.

Supóngase que se desea simular de la distribución objetivo $\pi(x)$, con $x \in \mathcal{X}$ y $\mathcal{X} \subset \mathbb{R}^d$. Se define una nueva función objetivo f sobre $\mathcal{X} \times \mathcal{X}$ como $f(x, x') = \pi(x)\pi(x')$. El algoritmo parte de un par de puntos dentro del soporte de la distribución objetivo y en cada paso mueve uno de ellos con igual probabilidad, mediante la propuesta que genera las siguientes transiciones:

$$(y, y') = \begin{cases} (x, h\,(x', x)) & \text{con probabilidad } 0.5, \\ (h\,(x, x')\,, x') & \text{con probabilidad } 0.5 \end{cases} \tag{1.2}$$

donde $h(x, x')$ es una variable aleatoria utilizada para formar la propuesta. Obsérvese que no se están considerando dos cadenas paralelas en $\mathcal{X}$, sólamente se construye una cadena en $\mathcal{X} \times \mathcal{X}$. Se seleccionará al azar una de cuatro distintas propuestas (movimiento de caminata, transversal, de vuelo y de salto, ver Christen et al., 2010), cada una de ellas caracterizada por una función h. En primer lugar se elegirá una de las opciones en (1.2), y posteriormente se simulará la propuesta de h.

Sea $g\,(\cdot \mid x, x')$ la función de densidad de $h(x, x')$. Bajo un esquema de Metrópolis-

Hastings, el cociente de aceptación puede calcularse como

$$\frac{\pi\left(y'\right)}{\pi\left(y'\right)}\frac{g\left(x' \,|\, y', x\right)}{g\left(y' \,|\, x', x\right)}$$

para el primer caso de (1.2) y

$$\frac{\pi\left(y\right)}{\pi\left(y\right)}\frac{g\left(x \,|\, y, x'\right)}{g\left(y \,|\, x, x'\right)}$$

para el segundo caso.

Un método MCMC bien calibrado (es decir, donde los parámetros de la propuesta son elegidos de manera óptima) puede converger más rápidamente que el t-walk, sin embargo, a pesar de que un MCMC de Metrópolis-Hastings es un método muy flexible y general, calibrarlo puede resultar muy complicado, sobre todo en altas dimensiones. El t-walk brinda un método alternativo y muy general, al requerir solamente dos puntos dentro del soporte y el logaritmo de la función objetivo como entrada. Este método está implementado en Python y en R (R Development Core Team, 2008) en una paquetería llamada `t-walk`, la cual es utilizada en este trabajo.

1.2.3. Inferencia sin verosimilitud (ABC)

Considérese un modelo bayesiano donde $\pi(\theta)$ denota la densidad a priori del vector de parámetros $\theta \in \Theta$. Supongamos que se tienen observaciones $\mathbf{y} = y_1, \ldots y_n \in \mathcal{D}$ del modelo $f\left(\mathbf{y} \,|\, \theta\right)$. Desde el enfoque bayesiano, la inferencia se basea en la distribución posterior

$$\pi\left(\theta \,|\, \mathbf{y}\right) \propto f\left(\mathbf{y} \,|\, \theta\right) \pi\left(\theta\right).$$

Sin embargo, en ocasiones la verosimilitud de los datos es muy complicada, su evaluación puede resultar muy costosa computacionalmente o simplemente imposible de calcular; esto dificulta el uso de los algoritmos estándar para simular de $\pi\left(\theta \,|\, \mathbf{y}\right)$. Los métodos ABC (Marin et al., 2012; Del Moral et al., 2012), son una alternativa que únicamente requiere que sea posible simular pseudo-observaciones $\mathbf{x}$ del modelo $f\left(\cdot \,|\, \theta\right)$, es decir, basta con conocer el modelo estocástico a partir del cual se generan los datos para poder simular de una aproximación de la distribución posterior.

En Rubin et al. (1984) se afirma que la estadística bayesiana y los métodos Monte Carlo son muy útiles para probar el ajuste de una gran variedad de modelos posibles a un conjunto de datos. Más aún, en este paper se plantea el primer

algoritmo ABC, el cual se basa en un muestreo exacto de aceptación y rechazo (ver Robert y Casella, 2013).

De acuerdo con Del Moral et al. (2012), el esquema general del ABC consiste en muestrear de una aproximación π_h de la posterior, definida sobre $\Theta \times \mathcal{D}$ como:

$$\pi_\varepsilon\left(\theta, \mathbf{x} \,|\, \mathbf{y}\right) = \frac{\pi\left(\theta\right) f\left(\mathbf{x} \,|\, \theta\right) \mathbb{1}_{A_{\varepsilon,\mathbf{y}}}(\mathbf{x})}{\int_{A_{\varepsilon,\mathbf{y}} \times \Theta} \pi\left(\theta\right) f\left(\mathbf{z} \,|\, \theta\right) d\mathbf{z}\, d\theta},$$

donde $\varepsilon > 0$ es un nivel de tolerancia, $\mathbb{1}_B(\cdot)$ es la función indicadora de un conjunto dado B y $\mathbf{x} \in \mathcal{D}$ corresponde a un conjunto de pseudo-observaciones del modelo, las cuales son *cercanas* en algún sentido a las observaciones reales $\mathbf{y}$. Dicho conjunto se define formalmente por

$$A_{\varepsilon,\mathbf{y}} = \left\{\mathbf{z} \in \mathcal{D} : \rho\left(\eta(\mathbf{z}), \eta(\mathbf{y})\right) < \varepsilon\right\}. \tag{1.3}$$

La función η es un estadístico que resume la información de la muestra y la función ρ es una medida de distancia.

Se han propuesto diversas variaciones a la aproximación por ABC de la posterior presentada anteriormente. En Marin et al. (2012) se describen algunas de estas modificaciones a detalle. La que se utilizará en el desarrollo de este trabajo, es una generalización propuesta por Wilkinson (2013) y consiste en sustituir la función indicadora de (1.3) por una función suavizadora de tipo kernel (la indicadora corresponde a un kernel uniforme). Mediante este algoritmo se propone realizar inferencia sobre una aproximación controlada de la distribución objetivo, haciendo uso de la convolución de esta última con una función kernel arbitraria:

$$\pi_h^*\left(\theta, \mathbf{x} \,|\, \mathbf{y}\right) = \frac{\pi(\theta)\, f(\mathbf{x} \,|\, \theta)\, K_h(\mathbf{y} - \mathbf{x})}{\int \pi(\theta)\, f(\mathbf{z} \,|\, \theta)\, K_h(\mathbf{y} - \mathbf{z})\, d\mathbf{z}\, d\theta}, \tag{1.4}$$

donde K_h es un estimador de densidad kernel parametrizado por el ancho de banda h. Wilkinson (2013) estipula que si el modelo planteado tiene un término de error, y se estima la distribución de dicho error mediante K_h, esto produce un algoritmo ABC que permite simular exactamente de la posterior que depende del error de las variables.

En la práctica, el algoritmo de Wilkinson se puede modificar calculando el estimador de densidad tipo kernel sobre un estadístico que mida la discrepancia

entre las observaciones reales ($\mathbf{y}$) y las simulaciones ($\mathbf{x}$) obtenidas dado un valor fijo θ_0 para el parámetro. En primer lugar debe plantearse un estadístico $\eta : \mathcal{D} \to S$ que resuma la información de los datos, y posteriormente la función $\rho : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ que proporciona una medida de distancia o disimilaridad. Una posible alternativa es usar un estadístico $t : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ que no dependa de cantidades desconocidas, tal que $t(\mathbf{x}, \mathbf{y}) = \rho\left(\eta(\mathbf{x}), \eta(\mathbf{y})\right)$ (Braunack-Mayer, 2013; Durán Aguilar, 2014). De esta forma, se modifica la expresión (1.4) como sigue:

$$\pi_h^*\left(\theta, \mathbf{x} \,|\, \mathbf{y}\right) = \frac{\pi(\theta)\, f(\mathbf{x} \,|\, \theta)\, K_h(t(\mathbf{x}, \mathbf{y}))}{\int \pi(\theta)\, f(\mathbf{z} \,|\, \theta)\, K_h(t(\mathbf{z}, \mathbf{y}))\, d\mathbf{z}\, d\theta}, \tag{1.5}$$

El estadístico t debe resumir de alguna forma la discrepancia entre los datos observados y los simulados. Una vez propuesto el estadístico, se debe cuantificar entre las distintas propuestas, es decir, servirá para comparar entre simulaciones $\mathbf{x}_1$ y $\mathbf{x}_2$ provenientes de los parámetros $\theta_1 \in \Theta$ y $\theta_2 \in \Theta$, respectivamente. Esta cuantificación se logrará mediante la estimación de densidad tipo Kernel, la cual es una técnica no paramétrica cuyos fundamentos se resumen en la siguiente sección.

Si se propone un estadístico y un modelo para el kernel con su ancho de banda, y además se es capaz de simular pseudo-observaciones $\mathbf{x}$ del modelo, ya es posible simular de la densidad π_h haciendo uso de métodos MCMC para que no sea necesario el cálculo del denominador de (1.5).

Estimación de densidades por Kernel

Sean $X_1, \ldots, X_n$ variables aleatorias independientes e idénticamente distribuidas como f, todas definidas sobre $\mathbb{R}$. Se usará el caso unidimensional, ya que el estadístico que se utilizará es univariado. El estimador por kernel de la densidad f en el punto $x \in \mathbb{R}$ se denota y define

$$\hat{f}_h\left(x\,;\, X_1, \ldots, X_n\right) := \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) \tag{1.6}$$

donde $K : \mathbb{R} \to \mathbb{R}$ es una función integrable tal que $\int K(u)du = 1$ y se denomina función kernel. La densidad normal estándar, por ejemplo, cumple con las condiciones y en este contexto se le llama kernel normal. También se define $K_h(u) = (1/h)K(u/h)$. Al parámetro h se le conoce como *ancho de banda* del estimador (1.6).

En Tsybakov (2009) se demuestran algunas propiedades asintóticas de este estimador. Por ejemplo:

- Cuando el ancho de banda h es tal que $nh \to \infty$ cuando $n \to \infty$, la varianza del estimador (1.6) tiende a cero cuando $n \to \infty$.
- Bajo ciertas condiciones puede encontrarse una cota para el sesgo de (1.6) para cualquier n fijo y puede demostrarse que se trata de un estimador asintóticamente insesgado.

Al usar este tipo de estimadores se debe fijar una función kernel, pero ello no es tan determinante en la rapidez de la convergencia como lo es la selección del ancho de banda h. Este hecho ha sido estudiado ampliamente y se han propuesto distintas soluciones para encontrar el ancho de banda óptimo en algún sentido (Turlach et al., 1993).

Estimación por ABC

Recordemos que el problema original consiste en muestrear de la densidad posterior $\pi(\theta \,|\, \mathbf{y})$, y a partir de pseudo-observaciones $\mathbf{x}$ del modelo y (1.5) se puede simular de $\pi_h^*(\theta, \mathbf{x} \,|\, \mathbf{y})$. De esta forma puede plantearse una aproximación de la posterior que depende de la función kernel K a elegir y su ancho de banda h, como sigue:

$$\begin{aligned}
\pi_h(\theta \,|\, \mathbf{y}) &= \int_{\mathcal{D}} \pi_h^*(\theta, \mathbf{x} \,|\, \mathbf{y}) \; d\mathbf{x} \\
&= \int_{\mathcal{D}} \frac{\pi(\theta)\, f(\mathbf{x} \,|\, \theta)\, K_h(t(\mathbf{x}, \mathbf{y}))}{\int \pi(\theta)\, f(\mathbf{z} \,|\, \theta)\, K_h(t(\mathbf{z}, \mathbf{y}))\, d\mathbf{z}\, d\theta} \, d\mathbf{x} \\
&= c \int_{\mathcal{D}} \pi(\theta)\, f(\mathbf{x} \,|\, \theta)\, K_h(t(\mathbf{x}, \mathbf{y}))\, d\mathbf{x}. \qquad (1.7)
\end{aligned}$$

Cuando el ancho de banda h tiende a cero y bajo ciertas condiciones mínimas Tsybakov (2009), se sigue que

$$\int_{\mathcal{D}} f(\mathbf{x} \,|\, \theta)\, K_h(t(\mathbf{x}, \mathbf{y}))\, d\mathbf{x} \to f(\mathbf{y} \,|\, \theta).$$

Por lo tanto, cuando $h \to 0$,

$$\pi_h(\theta \,|\, \mathbf{y}) \to c f(\mathbf{y} \,|\, \theta)\, \pi(\theta) = c\pi(\theta \,|\, \mathbf{y}),$$

donde $c \in \mathbb{R}$ es la constante de proporcionalidad que resulta de (1.7). Lo anterior implica que para anchos de banda muy pequeños, la aproximación π_h será muy cercana a la distribución posterior de la que se pretende simular.

MCMC sin verosimilitud

En Marjoram et al. (2003) se propone un método de simulación MCMC de una distribución posterior considerando que se desconoce el modelo a partir del cual fueron generados los datos, pero puede simularse de éste. Este algoritmo se conoce como ABC-MCMC, ya que, a pesar de la ausencia de la verosimilitud, se genera una cadena de Markov que converge a la aproximación de la distribución objetivo, como se explica en la Sección 1.2.

Para el desarrollo de este algoritmo se utilizará la aproximación por kernel de la distribución que considera pseudo-observaciones simuladas dada por (1.5). Nótese que la aproximación π_h de la posterior puede escribirse como:

$$\pi_h(\theta \,|\, \mathbf{y}) = \frac{\pi_h^*(\theta, \mathbf{x} \,|\, \mathbf{y})}{f(\mathbf{x} \,|\, \theta)}. \tag{1.8}$$

Se utilizará esta expresión en el cálculo de la probabilidad de aceptación de las propuestas. El algoritmo de simulación de una distribución posterior sin verosimilitud por MCMC se muestra a continuación:

Algoritmo 2: ABC-MCMC.

Dado un parámetro $\theta^{(t)}$ y una simulación $\mathbf{x}_{\theta^{(t)}}$ del modelo $f\left(\cdot \,|\, \theta^{(t)}\right)$,

1. Generar ν_t de la densidad propuesta $q(\cdot \,|\, \theta^{(t)})$

2. Simular pseudo-observaciones $\mathbf{x}_{\nu_t}$ provenientes del modelo $f\left(\cdot \,|\, \nu_t\right)$.

3. Hacer

$$\theta^{(t+1)} = \begin{cases} \nu_t & \text{con probabilidad } \alpha\left(\theta^{(t)}, \nu_t\right), \\ \theta^{(t)} & \text{con probabilidad } 1 - \alpha\left(\theta^{(t)}, \nu_t\right). \end{cases}$$

donde

$$\alpha\left(a, b\right) = \text{mín}\left\{\frac{\pi_h(b \,|\, \mathbf{y})}{\pi_h(a \,|\, \mathbf{y})} \frac{q(a \,|\, b)}{q(b \,|\, a)}, 1\right\}$$

4. Si se acepta ν_t, guardar $\mathbf{x}_{\theta^{(t+1)}} = \mathbf{x}_{\nu_t}$; en caso contrario, $\mathbf{x}_{\theta^{(t+1)}} = \mathbf{x}_{\theta^{(t)}}$.

Se calculará la probabilidad de aceptación del punto b estando en a. Sean $\mathbf{x}_a$ y $\mathbf{x}_b$ pseudo-observaciones del modelo con parámetros a y b, respectivamente.

Sustituyendo (1.8) en la razón de Metrópolis-Hastings,

$$\frac{\pi_h(b \,|\, \mathbf{y})}{\pi_h(a \,|\, \mathbf{y})} \frac{q(a \,|\, b)}{q(b \,|\, a)} = \frac{\dfrac{\pi_h^*(b, \mathbf{x}_b \,|\, \mathbf{y})}{f(\mathbf{x}_b \,|\, b)}}{\dfrac{\pi_h^*(a, \mathbf{x}_a \,|\, \mathbf{y})}{f(\mathbf{x}_a \,|\, a)}} \frac{q(a \,|\, b)}{q(b \,|\, a)},$$

y de (1.5) se sigue que

$$= \frac{\dfrac{\pi(b) f\left(\mathbf{x}_b \,|\, b\right) K_h\left(t(\mathbf{x}_b, \mathbf{y})\right)}{f(\mathbf{x}_b \,|\, b)}}{\dfrac{\pi(a) f\left(\mathbf{x}_a \,|\, a\right) K_h\left(t(\mathbf{x}_a, \mathbf{y})\right)}{f(\mathbf{x}_a \,|\, a)}} \frac{q(a \,|\, b)}{q(b \,|\, a)}.$$

Así, la probabilidad de aceptar el punto b estando en a resulta

$$\alpha(a, b) = \text{mín}\left\{\frac{\pi(b) K_h\left(t(\mathbf{x}_b, \mathbf{y})\right) q(a \,|\, b)}{\pi(a) K_h\left(t(\mathbf{x}_a, \mathbf{y})\right) q(b \,|\, a)}, 1\right\}, \tag{1.9}$$

y no depende de la distribución desconocida de los datos. Si se tiene una propuesta q simétrica (que será nuestro caso en las simulaciones a desarrollar), la expresión (1.9) se simplifica aún más, ya que se elimina este término en la razón de M-H.

Capítulo 2

Modelos compartimentales deterministas

En este capítulo se explica el planteamiento de un modelo compartimental determinista, así como los principales supuestos que se consideran en este caso. En la Sección 2.1 se describe la metodología para realizar inferencia paramétrica en un modelo de esta clase. Se trabaja sobre dos casos particulares, el modelo SIR y el SEIR deterministas, en las Secciones 2.2 y 2.3, respectivamente. En ambos modelos se presenta un ejemplo de la implementación del proceso de inferencia.

En un modelo determinista, las tasas ν_{ij} se expresan matemáticamente como la velocidad de cambio del volumen del compartimento i en favor o en dirección al j. Así, el modelo se formula como un sistema de ecuaciones diferenciales ordinarias.

Existen condiciones adicionales, como efectos demográficos, efectos de vacunación, etc. que pueden ser sumamente relevantes en el proceso infeccioso. Aunque varios de estos cambios pueden introducirse a su vez como nuevas categorías de los individuos en el modelo compartimental, en este trabajo suponemos que la evolución del brote epidémico es rápida (semanas) y que cambios demográficos pueden ser omitidos. En este sentido asumimos que la población es cerrada. Esto es, libre de nacimientos, muerte natural de los individuos de la población y migración, de manera que los parámetros del modelo serán únicamente aquellos que determinan el tiempo que los individuos pasan en cada compartimento.

2.1. Metodología de inferencia en modelos deterministas

Como se mencionó anteriormente, en un modelo compartimental determinista, la dinámica de la dispersión del agente infeccioso está dada por un sistema de ecuaciones diferenciales ordinarias. Una forma natural de abordar el problema de inferencia en este modelo epidemiológico si se tiene información a priori además de los datos observados, es concibiéndolo como un problema inverso (Stuart, 2010).

Se denomina *problema inverso* al proceso de calcular, a partir de un conjunto de observaciones, los factores causales que las originaron (Aster et al., 2011). La relevancia del estudio de problemas inversos radica en que permite obtener información acerca de parámetros que no se pueden observar directamente. Tienen un gran rango de aplicaciones en óptica, acústica, teoría de la comunicación, procesamiento de señales, etc.

El problema inverso se puede conceptualizar como:

$$\text{Datos} \longrightarrow \text{Parámetros del modelo;}$$

se considera *inverso* al problema directo, el cual relaciona los parámetros del modelo con los datos que se observan, a partir del modelo subyacente.

$$\text{Parámetros del modelo} \longrightarrow \text{Datos.}$$

En el contexto de un modelo epidemiológico compartimental, supongamos que se tiene:

a) el modelo de ecuaciones diferenciales que explica la dinámica del sistema epidemiológico,

b) un vector fijo $\boldsymbol{\theta}$ de parámetros del modelo, y

c) un conjunto de condiciones iniciales $\mathbf{X}_{\boldsymbol{\theta}}(t_0) = \mathbf{X}_0$;

y que se desea obtener $\mathbf{X}_{\boldsymbol{\theta}}(t)$ para $t > t_0$. A esto se le llama *problema directo* o *forward map*, y como se mencionó anteriormente no tiene una solución analítica cerrada pero ésta puede aproximarse usando los métodos numéricos empleados en la función `lsoda` del paquete `deSolve` de `R` (Soetaert et al., 2010).

Por otro lado, si se tiene:

a) un modelo de ecuaciones diferenciales que describe la dinámica del sistema epidemiológico, y

b) un vector $\mathbf{y} = (y_1, \ldots, y_n)$ de observaciones del proceso $\mathbf{X}_{\boldsymbol{\theta}}$ (o una función del mismo) en los tiempos $t_1, \ldots, t_n$, respectivamente,

y se desea inferir acerca del vector de parámetros $\boldsymbol{\theta}$ a partir del cual fueron generados los datos observados; entonces se trata del *problema inverso*.

El tipo de datos que se tienen generalmente en el monitoreo de dispersión de enfermedades, corresponden a los reportes con el número de nuevos infectados en intervalos de tiempo definidos, por ejemplo diarios o semanales. Supongamos que se tienen los datos $\mathbf{y} = (y_1, \ldots, y_n)$, donde y_ℓ es el reporte de nuevos infectados en el intervalo $(t_{\ell-1}, t_\ell]$, para $\ell = 1, \ldots, n$, con la convención de que $t_0 = 0$. De acuerdo al modelo determinista, el número de nuevos infectados en el ℓ-ésimo periodo (al cual se denotará $y^*_\ell(\theta)$), está relacionado con las tasas de transferencia entre los compartimentos, por lo que existe $h : [0,1]^K \to \mathbb{R}$, donde K es el número total de compartimentos, tal que

$$y^*_\ell(\theta) = \int_{t_{\ell-1}}^{t_\ell} h(X_{\boldsymbol{\theta}}(t))dt, \quad \forall \ell = 1, \ldots, n. \tag{2.1}$$

La función h se identifica con la *fuerza de infección*, la cual corresponde a la tasa con la que entran nuevos individuos infectados al compartimento I, y se usa la notación $y^*_\ell(\theta)$ para denotar al número exacto de nuevos infectados en el intervalo $(t_{\ell-1}, t_\ell]$ de acuerdo al sistema de ecuaciones diferenciales asociado.

Idealmente, y de acuerdo al modelo compartimental determinista en el cual se esté trabajando, se podría obtener el número de nuevos infectados en cada intervalo, $y^*_\ell(\theta)$, mediante la ecuación (2.1), sin embargo, los reportes que se reciben, y_ℓ, no son datos perfectos, y en este sentido puede asociarse un error de medición a dichos reportes con un comportamiento aleatorio. De esta forma tendría sentido suponer que el número de reportes es una variable aleatoria Y_ℓ tal que tiene una distribución F cuya media es el número de nuevos infectados en el mismo intervalo de tiempo, $y^*_\ell(\theta)$. Sea $f_\ell(\cdot \mid \boldsymbol{\theta})$ la función de densidad de probabilidades asociada a cada Y_ℓ para $\ell = 1, \ldots, n$. Si se tiene un vector $\mathbf{y} = (y_1, \ldots, y_n)$ de observaciones de los reportes en cada intervalo de tiempo correspondiente, la función de verosimilitud del vector de parámetros satisface

$$L(\boldsymbol{\theta} \mid \mathbf{y}) \propto \prod_{i=1}^{n} f_\ell(y_\ell \mid \boldsymbol{\theta}).$$

Generalmente, gracias a estudios de laboratorio u observaciones de otros brotes, se tiene algún tipo de información acerca de la biología y desarrollo del agente infeccioso. Esta información se liga a los parámetros del modelo. Es deseable que el modelo epidémico considere esta información además de las observaciones del número de casos del brote. Dicha información puede incorporarse mediante una densidad *a priori* $\pi(\boldsymbol{\theta})$. De acuerdo con la Regla de Bayes, la densidad posterior del vector de parámetros se obtendría como

$$\pi\left(\boldsymbol{\theta}\,|\,\mathbf{y}\right)=\frac{L\left(\boldsymbol{\theta}\,|\,\mathbf{y}\right)\pi\left(\boldsymbol{\theta}\right)}{\pi(\mathbf{y})}=\frac{L\left(\boldsymbol{\theta}\,|\,\mathbf{y}\right)\pi\left(\boldsymbol{\theta}\right)}{\int L\left(\boldsymbol{\vartheta}\,|\,\mathbf{y}\right)\pi\left(\boldsymbol{\vartheta}\right)d\boldsymbol{\vartheta}}. \tag{2.2}$$

La densidad posterior es la herramienta fundamental en inferencia bayesiana (Bernardo y Smith, 2001; Bolstad, 2013), ya que a partir de ella podemos obtener estimadores puntuales para los parámetros, las densidades marginales e incluso intervalos de probabilidad.

En ocasiones no es posible obtener la densidad posterior de forma explícita porque la constante $\pi\left(\mathbf{y}\right)$ involucrada en el denominador de (2.2) puede requerir el cálculo de integrales múltiples complicadas o la integral no tiene una solución analítica cerrada.

Como se vio en la Sección 1.2 , los algoritmos Markov Chain MonteCarlo (MCMC) permiten simular de la densidad posterior a pesar de que no se tenga dicha constante. La inferencia de los modelos SIR y SEIR deterministas se propone implementando el método t-walk descrito en la Sección 1.2.2. En ambos casos se presentan ejemplos cuyo código puede consultarse en `http://leticiaramirez2.net/supplementary_material.html`.

2.2. Modelo SIR determinista

Uno de los modelos epidemiológicos más sencillos es el SIR. Éste describe la dinámica de enfermedades en que los individuos susceptibles son infectados, pero posteriormente desarrollan una inmunidad a la enfermedad o mueren. En la práctica ha sido utilizado para modelar enfermedades comunes en la niñez como sarampión, varicela y paperas, que son originadas por virus y de los cuales se suele desarrollar inmunidad.

El modelo SIR consta entonces de tres compartimentos: susceptibles (S), infectados (I) y removidos (R), y considera que el agente infeccioso se transmite por contacto entre susceptibles e infectados. La dinámica de la enfermedad sigue el esquema mostrado en la Figura 2.1. Los individuos nacen dentro de la clase susceptible, por lo que si el virus es nuevo, el total de personas pertenecen a esta clase. Un individuo susceptible nunca ha tenido contacto con la enfermedad y se contagia por interacción con un infectado, después de lo cual pasa a la clase I, donde permanece durante el periodo de infección. Luego de este periodo pasa a

la clase R, donde adquiere inmunidad de por vida, o al menos hasta después de terminada la evolución completa del brote infeccioso en la población.

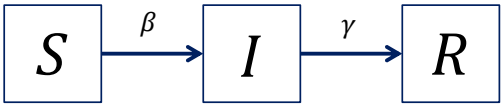

Figura 2.1: Dinámica del modelo SIR.

Para evitar soluciones triviales es necesario establecer un conjunto de condiciones iniciales para el sistema dinámico al tiempo t_0. Denotemos $S(t)$, $I(t)$ y $R(t)$ al número de individuos susceptibles, infectados y removidos al tiempo t, respectivamente. Típicamente se toman como condiciones iniciales $t_0 = 0$ y $S(t_0) > 0$, $I(t_0) > 0$ para que pueda haber contagio. Además, como se supone que el brote inicia en t_0, suele tomarse $R(t_0) = 0$.

Adicionalmente a los supuestos anteriores, el modelo SIR determinista propuesto por Kermack y McKendrick (1927) asume una población grande, homogénea y distribuida uniformemente, de manera que cualesquiera dos individuos tengan la misma probabilidad de contacto.

Como se mencionó anteriormente, en el SIR determinista se supondrá una población de tamaño $N = N(t)$ constante para cualquier $t \geq t_0$. De tal forma que $S(t) + I(t) + R(t) = N$, $\forall t \geq t_0$.

Al ser una población cerrada, los parámetros involucrados en el modelo son la tasa de contagio $\beta > 0$ y la tasa de recuperación o remoción $\gamma > 0$.

El modelo de Kermack y McKendrick supone que los individuos infecciosos dejan esta clase a una tasa γI por unidad de tiempo, lo cual realmente significa que el periodo de infección tiene una densidad exponencial con media $1/\gamma$ (Brauer, 2008). Es fácil corroborar el resultado anterior. Considérese el conjunto de individuos infectados a un tiempo fijo, y sea $u(s)$ el número de estos individuos que siguen infectados s unidades de tiempo después de haber sido contagiados. Si una fracción α deja la clase de infectados en una unidad de tiempo, se tiene

$$u' = -\alpha u,$$

y la solución a esta ecuación diferencial está dada por

$$u(s) = u(0)e^{-\alpha s}.$$

Así, la fracción de individuos infecciosos que lo siguen siendo s unidades de tiempo después de haber sido infectados es $e^{-\alpha s}$, por lo que la longitud del periodo infeccioso se distribuye exponencial con media

$$\int_0^\infty e^{-\alpha s} ds = \frac{1}{\alpha}.$$

Bajo los supuestos anteriores, el sistema de ecuaciones diferenciales que describe la dinámica del modelo SIR determinista es de la forma:

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta SI}{N} \\ \frac{dI}{dt} &= \frac{\beta SI}{N} - \gamma I \\ \frac{dR}{dt} &= \gamma I. \end{aligned} \tag{2.3}$$

Las relaciones (2.3) corresponden a un sistema de ecuaciones diferenciales ordinarias no lineales, que no admite una solución analítica explícita. Sin embargo, existen métodos numéricos muy precisos para aproximarse a dicha solución. En este caso se usará la función `lsoda` dentro del paquete `deSolve` de `R` (Soetaert et al., 2010), la cual está diseñada para incluir métodos rígidos y no rígidos para ecuaciones diferenciales ordinarias de primer orden. La función `lsoda` brinda una interfaz al solver de ecuaciones diferenciales ordinarias de `FORTRAN` que lleva el mismo nombre, diseñado por Hindmarsh (1983) y Petzold (1983).

La dinámica del sistema (2.3) está determinada por el número reproductivo básico $R_0 = \beta/\gamma$, que de acuerdo con Allen (2008) puede ser interepretado como el número promedio de casos nuevos que producirá una persona infectada, en una población completamente susceptible.

El parámetro R_0 induce un efecto umbral determinante en el comportamiento a largo plazo de la dinámica de la enfermedad (Anderson et al., 1992; Hethcote, 2000). Su incidencia en esto se observa en el siguiente resultado, extraido de Hethcote (1976), y que se deriva del análisis del sistema de ecuaciones diferenciales (2.3).

Supóngase una población cerrada y sea $S(t)$, $I(t)$, $R(t)$ una solución al sistema (2.3).

i) Si $(R_0\, S(0))/N > 1$*, entonces hay un incremento inicial en el número de individuos infectados* $i(t)$ *(epidemia).*

ii) Si $(R_0\, S(0))/N \leq 1$*, entonces* $I(t)$ *decrece monótonamente a cero (equilibrio libre de enfermedad).*

A la cantidad $(R_0\, S(0))/N$ se le llama *número de reemplazo inicial* y puede interpretarse como el número promedio de infecciones secundarias producidas por un individuo infectado al comienzo de la epidemia, y a lo largo de su periodo infeccioso. Cuando el número de susceptibles es cercano al total de personas en la población ($S(0) \approx N$) , entonces el número de reemplazo inicial y el número reproductivo básico coinciden. Obsérvese que en el caso *ii)* del resultado anterior, la enfermedad comienza a desaparecer de la población desde su inicio, mientras que en el caso *i)*, cuando el número de reemplazo inicial es mayor que 1, se inicia con un incremento en el número de infectados hasta que el número de susceptibles decrece al punto que $S(t) < N/R_0$. Después de este periodo, el número de infecciosos decrece.

De lo anterior se deduce que en una población cerrada, un patógeno puede evolucionar en una epidemia si hay una proporción de susceptibles mayor que $1/R_0$; hecho que permitiría establecer políticas de vacunación adecuadas para reducir la proporción de susceptibles hasta ser menor a $1/R_0$ y así evitar el desarrollo del brote. A esta fracción se le conoce como inmunidad del grupo (herd immunity).

En resumen, para poder analizar el comportamiento de la dinámica de la dispersión de un agente infeccioso en una población, es de suma importancia el estudio de los parámetros del modelo. A continuación se describe un método para hacer inferencia estadística de los parámetros en un modelo SIR determinista, suponiendo una población cerrada.

2.2.1. Inferencia en el SIR determinista

Supongamos que $y_1, \ldots, y_n$ son reportes de nuevos individuos infectados en los intervalos de tiempo $(t_0, t_1], \ldots, (t_{n-1}, t_n]$, respectivamente, provenientes de un modelo SIR con vector de parámetros $\boldsymbol{\theta_0} = (\beta_0, \gamma_0)$ desconocido. Sea N el tamaño de la población. Estamos interesados en el número de individuos que pasaron del compartimento S al I entre los tiempos $t_{\ell-1}$ y t_ℓ para cada $\ell = 1, \ldots, n$. Usando la notación introducida anteriormente, esto es $\mathcal{N}_{SI}(t_\ell) - \mathcal{N}_{SI}(t_{\ell-1})$. De acuerdo con el sistema (2.3), los individuos dejan la clase susceptible a tasa $\beta SI/N$, por

lo que el número de reportes en el intervalo $(t_{\ell-1}, t_\ell]$ puede obtenerse como

$$y^*_\ell(\theta) = \mathcal{N}_{SI}(t_\ell) - \mathcal{N}_{SI}(t_{\ell-1}) = \int_{t_{\ell-1}}^{t_\ell} \frac{\beta S(t) I(t) dt}{N} dt, \quad \text{para } \ell = 1, \ldots, n, \tag{2.4}$$

donde $y^*_\ell(\theta)$ es el número de nuevos infectados de acuerdo al sistema (2.3).

Podemos ver al número de reportes de nuevos infectados en cada intervalo de tiempo como una variable aleatoria de conteo, donde la media es $y^*_\ell(\theta)$. Se propone un modelo Poisson, ya que no sólo modela una v.a. entera, sino que es consistente con los tiempos de infección exponenciales asociados Sistema (2.3). Supondremos entonces, para $\ell = 1, \ldots, n$, que la relación entre el número de nuevos casos reportados en el l-ésimo periodo (Y_ℓ) y el número de de nuevos infectados para el mismo periodo es

$$Y_\ell \sim Pois\left(y^*_\ell(\theta)\right). \tag{2.5}$$

Es importante notar que dada la solución del sistema (2.3), las Y_ℓ son independientes, mas no idénticamente distribuidas. La función de verosimilitud de $\boldsymbol{\theta}$ basada en las n observaciones $y_1, ...y_n$, es

$$L\left(\boldsymbol{\theta} \,|\, \mathbf{y}\right) \propto \prod_{\ell=1}^{n} e^{-y^*_\ell(\theta)} \left(y^*_\ell(\theta)\right)^{y_\ell}. \tag{2.6}$$

Ejemplo

Para ilustrar el procedimiento de inferencia, supongamos como densidad a priori para $\boldsymbol{\theta}$ un producto de densidades Gamma correspondientes a cada uno de los parámetros, es decir,

$$\begin{aligned} \pi\left(\boldsymbol{\theta}\right) &= \pi\left(\beta \,|\, a_1, b_1\right) \pi\left(\gamma \,|\, a_2, b_2\right) \\ &= \frac{b_1^{a_1}}{\Gamma(a_1)} \beta^{a_1 - 1} e^{-b_1 \beta} \frac{b_2^{a_2}}{\Gamma(a_2)} \gamma^{a_2 - 1} e^{-b_2 \gamma}, \end{aligned}$$

donde $a_i > 0$ y $b_i > 0$ para $i = 1, 2$, son parámetros que describen el conocimiento previo de los parámetros del modelo mediante una densidad Gamma. Por ejemplo, si se toma una $\text{Gam}(2, 2)$ para β y una $\text{Gam}(1.5, 1.5)$ para γ como densidades marginales, éstas se ven como en la Figura 2.2.

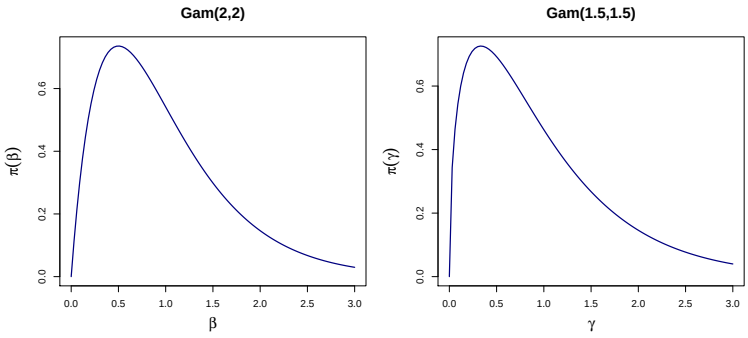

Figura 2.2: Ejemplo de densidades a priori para β y γ

Tomando una muestra simulada de 30 reportes semanales de nuevos infectados con distribución (2.5), y las densidades a priori de la Figura 2.2, puede obtenerse la densidad posterior como en (2.2). Las observaciones se simularon tomando una población de tamaño $N = 500$ y parámetros reales $\beta = 0.55$ y $\gamma = 0.25$. Los contornos de la densidad posterior se muestran en la Figura 2.3. El vector correspondiente a los parámetros con los que fueron simulados los datos se señala con un punto blanco y se observa que es cercano a los contornos centrales.

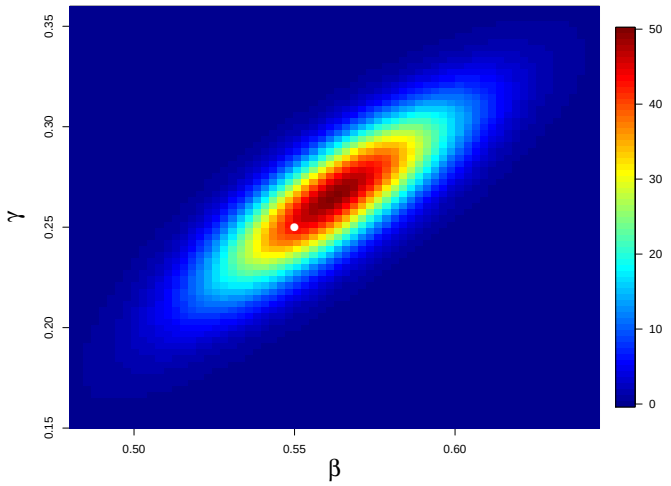

Figura 2.3: Contornos de la densidad posterior para (β, γ)

Considerando como función objetivo la densidad posterior mencionada anteriormente, se simuló una cadena de Markov de tamaño 10,000 usando el algoritmo MCMC t-walk (Christen et al., 2010). Dicha cadena, por construcción, tiene como distribución estacionaria la posterior de la cual se deseaba simular.

Se simuló el punto inicial a partir de la densidad a priori y para monitorear la convergencia de la cadena se observa la evolución de la logdensidad posterior, la cual comienza a estabilizarse cuando llega al soporte de la densidad deseada. Al periodo en que tarda la cadena en converger a su soporte se le llama *burn-in*. Visualmente podemos evaluar la convergencia de la cadena al soporte de la densidad posterior, monitoreando la evaluación de la logposterior en cada punto de la cadena, lo cual se observa en la Figura 2.4. A partir de lo anterior, parece razonable considerar un periodo de *burn-in* de 30 iteraciones.

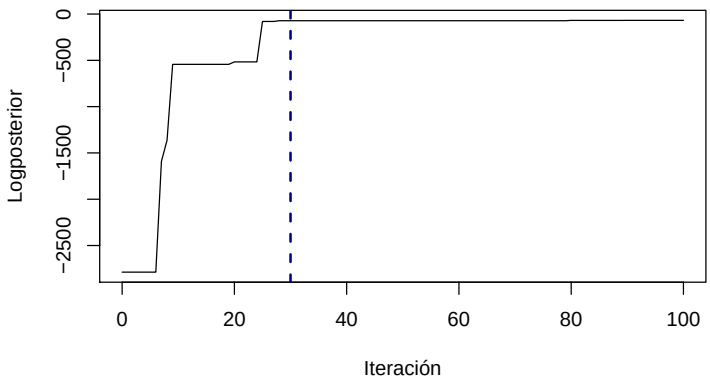

Figura 2.4: Evolución de la logdensidad posterior evaluada en los primeros 100 puntos de la cadena generada.

Se desean simulaciones pseudo-independientes de la densidad posterior, pero al haber simulado de una cadena de Markov, es natural que ésta presente dependencias entre las observaciones. Para lograr independencia entre los puntos de la muestra se toman puntos de la cadena cada cierto intervalo de tiempo. Al número que indica cada cuántas observaciones se considerará un punto dentro de la muestra le llamaremos *rezago*, y una aproximación de éste se puede obtener mediante un índice llamado Integrated Autocorrelation Time o IAT (Ver Roberts et al., 2001). El IAT compara la autocorrelación de una muestra independiente contra la autocorrelación de la cadena (Geyer, 1992). En este trabajo se tomará como indicador del rezago al promedio de los IAT's obtenidos para cada uno de los parámetros individualmente. En el ejemplo explorado se obtuvo un IAT cercano a 24, por lo que se consideró un rezago de 24 puntos.

Al eliminar el periodo de *burn-in* y considerar el rezago correspondiente, obtenemos una muestra efectiva de 415 puntos a partir de la cadena de longitud

10, 000. En la Figura 2.5 se grafican los puntos de la muestra simulada sobre los contornos de la posterior.

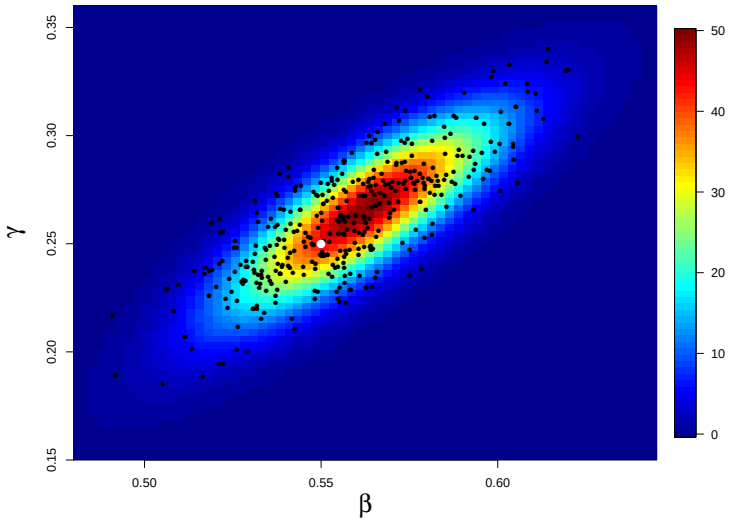

Figura 2.5: Simulaciones de la densidad posterior usando t-walk.

En la Figura 2.6 se muestran los histogramas de las simulaciones de la densidad posterior de los parámetros, individualmente. Se observa que la moda de las densidades marginales estimadas es cercana a los verdaderos valores de los parámetros, y que su dispersión es pequeña.

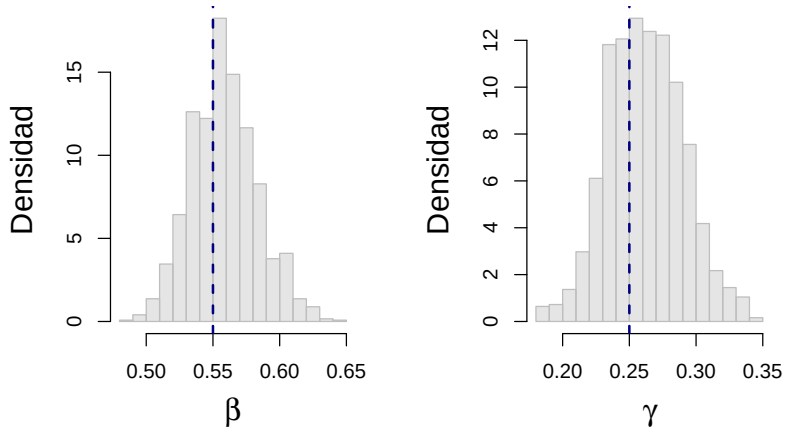

Figura 2.6: Histogramas muestrales de la densidad posterior.

Por último, a partir de la muestra de la densidad posterior simulada, se obtienen intervalos de 95 % de probabilidad y la mediana muestral como estimador puntual del vector de parámetros. Dichos resultados se muestran en la Tabla 2.1 y

se observan intervalos estrechos que contienen al verdadero valor de los parámetros a partir de los cuales fueron simulados los datos.

Cuantil	β	γ
2.5 %	0.5116	0.2070
50 %	0.5572	0.2617
97.5 %	0.6084	0.3211
Real	0.55	0.25

Tabla 2.1: Estimadores puntuales de los parámetros e intervalos de probabilidad 95 % de los parámetros del modelo SIR determinista (a prioris Gamma).

También se simuló la densidad posterior utilizando como densidad a priori para $\boldsymbol{\theta}$ un producto de densidades $Unif(0, 4)$ independientes, y los estimadores puntuales e intervalos de probabilidad resultan similares al caso anterior, como se muestra en la Tabla 2.2.

Cuantil	β	γ
2.5 %	0.5101	0.2087
50 %	0.5554	0.2601
97.5 %	0.6082	0.3161
Real	0.55	0.25

Tabla 2.2: Estimadores puntuales de los parámetros e intervalos de probabilidad 95 % de los parámetros del modelo SIR determinista (a prioris Uniformes).

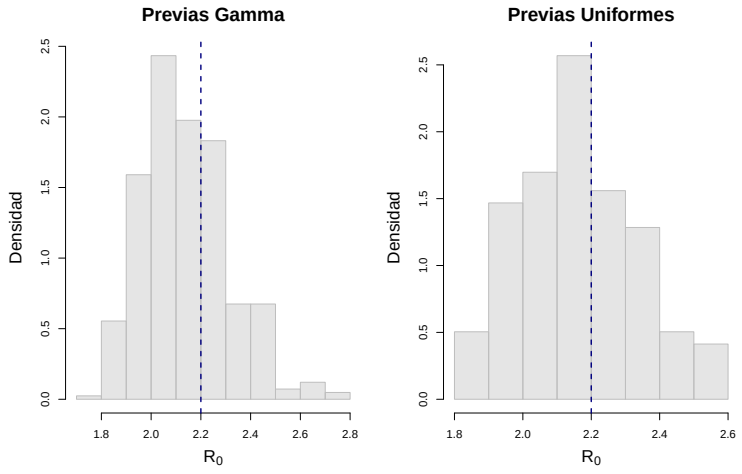

Figura 2.7: Histogramas de simulaciones de la densidad posterior de R_0 con densidades a priori Gamma (izquierda) y Uniformes (derecha).

A partir de las simulaciones de la densidad posterior de los parámetros, es inmediato obtener simulaciones de la posterior de R_0 calculando el cociente β/γ. En la Figura 2.7 se muestran los histogramas obtenidos usando previas Gamma y Uniformes.

La densidad de R_0 parece no ser simétrica, ya que ambos histogramas se encuentran ligeramente sesgados hacia la izquierda. Esto también se comprueba observando los intervalos de probabilidad 95 % obtenidos a partir de los cuantiles de la posterior (Tabla 2.3).

En ambos casos se obtienen conclusiones similares. El valor real de R_0 cae dentro del intervalo y es muy cercano a la mediana. Además el soporte de la densidad posterior de R_0 no incluye al 1, lo cual significa que existe potencial de que se presente un brote (epidemia).

Cuantil	p. Gamma	p. Uniformes
2.5 %	1.8511	1.8543
50 %	2.1295	2.1407
97.5 %	2.5351	2.5296
Real	2.2	2.2

Tabla 2.3: Estimadores puntuales e intervalos de probabilidad 95 % de R_0.

2.3. Modelo SEIR determinista

El modelo SEIR se utiliza para algunas enfermedades donde hay un periodo en que el individuo ya ha sido infectado pero aún no es capaz de ser infeccioso. A esta condición le llamaremos *estado de exposición* (del inglés *exposed*) y al tiempo en que el individuo permanece en este compartimento se le denomina periodo de latencia. Si el periodo de latencia es corto, es común que se omita en el modelo (Roberts y Heesterbeek, 2003).

A diferencia del modelo SIR, los individuos susceptibles que son infectados pasan a la clase de expuestos E, y después del periodo de latencia pasan a la clase I y adquieren la capacidad de infectar a otros individuos. La dinámica del flujo de individuos entre los compartimentos de este modelo se muestra en la Figura 2.8. Los supuestos del modelo SIR se mantienen para este caso, con la precisión de que ahora la enfermedad es transmitida solamente por contacto de susceptibles (S) con infecciosos (I).

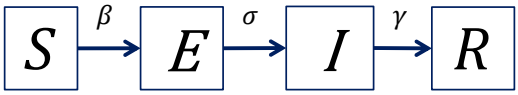

Figura 2.8: Dinámica del modelo SEIR.

Sean $S(t)$, $E(t)$, $I(t)$, $R(t)$ el número de individuos susceptibles, expuestos, infecciosos y removidos, respectivamente, en una población cerrada de tamaño N; de tal forma que $S(t) + E(t) + I(t) + R(t) = N \, \forall t > t_0$. Consideremos una tasa de contagio $\beta > 0$ y una tasa de recuperación $\gamma > 0$, y ahora incorporemos al modelo un periodo de latencia (o de exposición) distribuido exponencialmente con media $1/\sigma$, con $\sigma > 0$ (Brauer, 2008). De esta forma, la dinámica del modelo SEIR determinista puede ser representada por el sistema de ecuaciones ordinarias (2.7).

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta SI}{N} \\ \frac{dE}{dt} &= \frac{\beta SI}{N} - \sigma E \\ \frac{dI}{dt} &= \sigma E - \gamma I \\ \frac{dR}{dt} &= \gamma I. \end{aligned} \tag{2.7}$$

El número reproductivo básico asociado a este modelo sigue siendo $R_0 = \beta/\gamma$. De manera similar al SIR, este parámetro determina el comportamiento del tamaño del brote infeccioso en la población. Según Hethcote (2000), cuando $R_0 < 1$ a largo plazo se tiene el equilibrio libre de enfermedad, y en caso contrario se tiene una epidemia, es decir, el sistema converge al equilibrio endémico.

Modelar una enfermedad con un SEIR en ocasiones puede resultar complicado, sobre todo en la parte de obtención de los datos, ya que generalmente los periodos de exposición de los individuos no son observables y no coinciden con los periodos en que se presentan los síntomas de la enfermedad (periodo de incubación). Sin embargo, pueden incorporar en el modelo datos de otras fuentes como estudios de laboratorio o información de virus que biologos, patólogos o epidemiólogos consideren similares. Esta información, desde el enfoque Bayesiano, se incorpora como información a priori expresada como una función de densidad.

2.3.1. Inferencia en el SEIR determinista

En el caso del modelo SEIR, la inferencia se complica un poco si no se tienen los tiempos de exposición de los individuos, lo cual sucede generalmente. Sin embargo, si se tiene información sobre σ (que es usualmente el caso si se cuentan con estudios previos sobre el agente infeccioso) y se adopta el enfoque bayesiano utilizado para el SIR, se pueden obtener resultados muy competitivos.

Tomando como referencia el Sistema de Ecuaciones (2.7), en este caso la tasa que representa el flujo de nuevos individuos infectados es σE, y únicamente depende del número de personas que se encuentran en el periodo de latencia de la enfermedad. Así, el número de nuevos infectados $y_\ell^*(\theta)$ en el intervalo de tiempo $(t_{\ell-1}, t_\ell]$, para $t_{\ell-1} < t_\ell$, se puede obtener como

$$y_\ell^*(\theta) = \mathcal{N}_{EI}\left(t_{\ell-1}\right) - \mathcal{N}_{EI}\left(t_\ell\right) = \int_{t_{\ell-1}}^{t_\ell} \sigma E(t) dt. \tag{2.8}$$

Sean $y_1, \ldots, y_n$ datos observados correspondientes a reportes de nuevos infectados en los intervalos de tiempo $(t_0, t_1], \ldots, (t_{n-1}, t_n]$. Al igual que en el modelo SIR, modelamos la incertidumbre acerca de los reportes con respecto al número real de nuevos infectados con una distribución de probabilidad Poisson, con el argumento de que los tiempos que tardan los individuos en pasar de la clase E a la clase I son exponenciales (Brauer, 2008). De esta forma consideraremos

$$y_\ell \sim Pois\left(y_\ell^*(\theta)\right), \quad \text{para } \ell = 1, \ldots, n. \tag{2.9}$$

En la Figura 2.9 se muestran simulaciones de reportes en 30 intervalos de tiempo unitarios, bajo el modelo SEIR determinista con vector de parámetros $(0.55, 0.25, 0.4)$, y se señala con puntos continuos la media de cada densidad Poisson asociada.

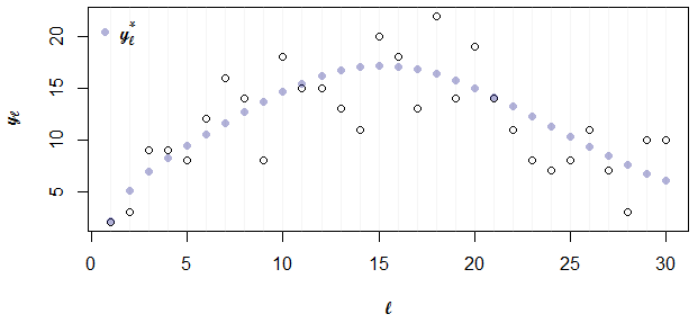

Figura 2.9: Simulaciones de reportes de nuevos infectados para el modelo SEIR con $\beta_0 = 0.55$, $\gamma_0 = 0.25$ y $\sigma_0 = 0.4$

La función de verosimilitud del vector de parámetros $\boldsymbol{\theta} = (\beta, \gamma, \sigma)$, dados los reportes correspondientes está dada por:

$$L\left(\boldsymbol{\theta} \mid \mathbf{y}\right) \propto \prod_{\ell=1}^{n} e^{-y_\ell^*(\theta)} \left(y_\ell^*(\theta)\right)^{y_\ell},$$

donde

$$y_\ell^*(\theta) = \int_{t_{\ell-1}}^{t_\ell} \sigma E(t) dt.$$

Ahora supongamos como densidad a priori para $\boldsymbol{\theta}$, un producto de densidades Gamma independientes para los tres parámetros, de la siguiente forma,

$$\begin{aligned}\pi\left(\boldsymbol{\theta}\right) &= \pi\left(\beta \mid a_1, b_1\right) \pi\left(\gamma \mid a_2, b_2\right) \pi\left(\sigma \mid a_3, b_3\right) \\ &= \frac{b_1^{a_1}}{\Gamma(a_1)} \beta^{a_1-1} e^{-b_1\beta} \frac{b_2^{a_2}}{\Gamma(a_2)} \gamma^{a_2-1} e^{-b_2\gamma} \frac{b_3^{a_3}}{\Gamma(a_3)} \sigma^{a_3-1} e^{-b_3\gamma},\end{aligned}$$

donde deben elegirse a_i, b_i para $i = 1, 2, 3$ tal que reflejen el conocimiento previo que se tiene del vector de parámetros. Si se considera una densidad $Gam(2, 2)$ para β, una $Gam(1.5, 1.5)$ para γ y una $Gam(1.5, 1)$ para σ, las densidades marginales se pueden observar en la Figura 2.10.

Ejemplo

Considerando esta última densidad a priori para los parámetros y los datos simulados que se muestran en la Figura 2.9, puede obtenerse la densidad posterior mediante (2.2), donde la constante que aparece en el denominador puede ser aproximada por integración numérica, pero no será necesaria ya que los métodos MCMC nos permiten simular de la posterior sin hacer uso de esa constante.

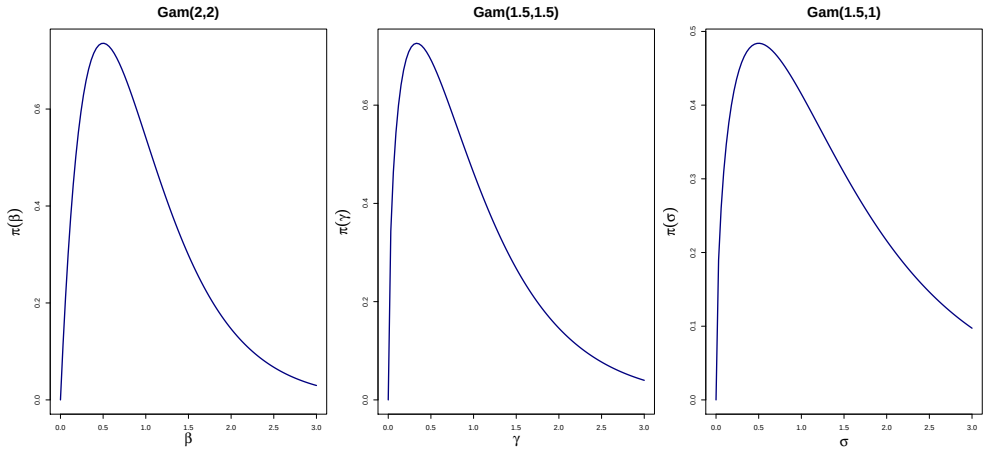

Figura 2.10: Ejemplo de densidades a priori para β, γ y σ.

Se simuló mediante el t-walk (Christen et al., 2010) una cadena de 10,000 puntos. Para monitorear la convergencia de la cadena al soporte de la posterior, en la Figura 2.11 se muestra la logdensidad posterior evaluada en los primeros 100 puntos de la cadena generada. Se tomó un periodo de *burn-in* de 60 iteraciones, que parece ser adecuado para que la logdensidad posterior se estabilice en un nivel alto.

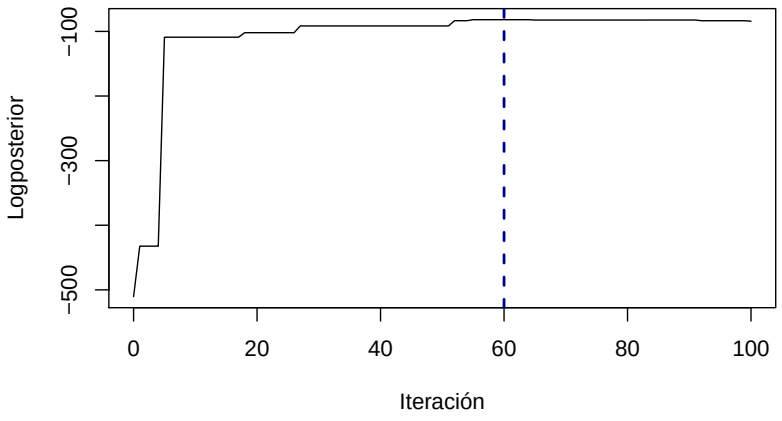

Figura 2.11: Logdensidad posterior para el modelo SEIR determinista evaluada en las primeras 100 iteraciones del t-walk.

En este caso, el IAT resultó de 11.68, por lo que se consideró un rezago de tamaño 12, lo cual resultó en una muestra efectiva de 2485 puntos muestreados de la posterior. En la Figura 2.12 se presentan los histogramas marginales de la

muestra generada. Se observa que la densidad de σ parece tener dos modas, aunque una de ellas muy pequeña, y que el verdadero valor de los parámetros (señalado con una recta punteada vertical) es cercano a las modas de cada histograma marginal.

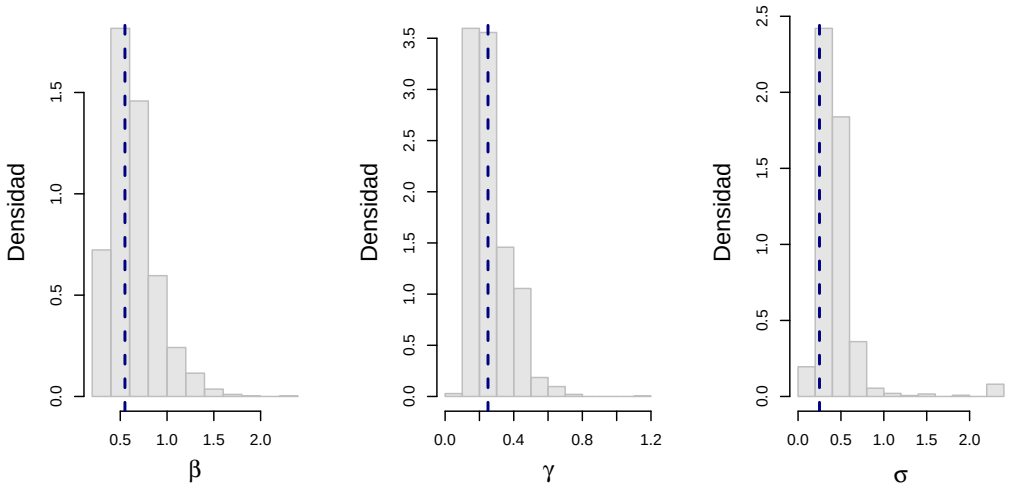

Figura 2.12: Histogramas marginales de la posterior para el SEIR determinista.

En la Tabla 2.4 se muestran intervalos de probabilidad 95 % obtenidos empíricamente con las simulaciones de la posterior. Se muestra también la mediana muestral, la cual podría considerarse como un estimador puntual de cada parámetro. Todos los intervalos contienen al verdadero valor de los parámetros a partir del cual fueron simulados los datos. Se observa también que el intervalo para σ resultó más amplio que los anteriores, debido a la otra pequeña moda que se aprecia en el histograma muestral.

Cuantil	β	γ	σ
2.5 %	0.3834	0.1591	0.1903
50 %	0.5939	0.2647	0.3948
97.5 %	1.1318	0.4843	0.7799
Real	0.55	0.25	0.4

Tabla 2.4: Estimaciones puntuales y por intervalos de 95 % de probabilidad para los parámetros del modelo SEIR determinista.

En este caso la densidad posterior marginal de R_0 tiene la forma que se observa en la Figura 2.13. La densidad es asimétrica con cola pesada a la derecha y se observa que el soporte no incluye al valor de 1, por lo que es evidente que en este modelo hay un brote.

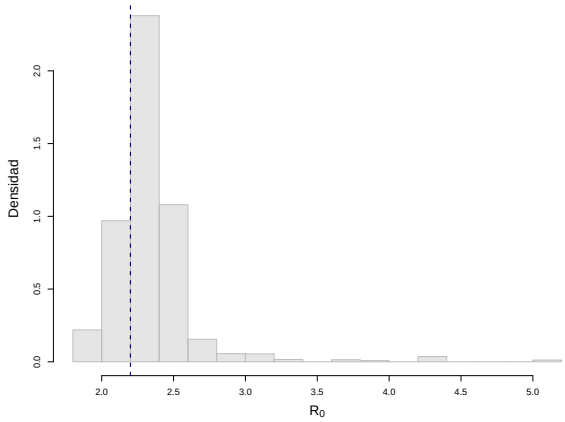

Figura 2.13: Histograma de simulaciones de la densidad posterior de R_0.

El intervalo de probabilidad 95 % para R_0 obtenido a partir de los cuantiles muestrales es $(1.8759, 3.1155]$, el cual incluye el valor real $R_0 = 2.2$ calculado a partir de los parámetros con que se simularon las observaciones. La mediana muestral, que puede verse como un estimador puntual, es de 2.3033, cercana al valor real.

Capítulo 3

Modelos compartimentales estocásticos

En este capítulo se introducen modelos que incorporan al modelo epidemiológico aleatoriedad a nivel poblacional, llamada estocasticidad demográfica. Esta consideración se refiere a que los tiempos de entrada y salida de los compartimentos tienen una distribución probabilística, lo cual se refleja en la aleatoriedad de la solución del sistema a un tiempo fijo t. Considerar la estocasticidad demográfica resulta sumamente relevante al trabajar con poblaciones pequeñas. En la Sección 3.1 se describe brevemente el proceso de inferencia en modelos compartimentales estocásticos. En la Sección 3.2 se describe el caso particular del modelo SIR, y en la Sección 3.3 se generaliza al SEIR, presentando un ejemplo de la metodología propuesta para este tipo de modelos.

Allen (2008) presenta brevemente tres métodos distintos de formular un modelo epidemiológico estocástico, los cuales se distinguen de acuerdo a los supuestos acerca del tiempo y el espacio de estados. En primer lugar se describe la modelación con una cadena de Markov a tiempo discreto (CMTD), donde el tiempo y el espacio de estados son discretos. Se muestra también un modelo usando una cadena de Markov a tiempo continuo (CTMC), donde el tiempo es continuo pero el espacio de estados es discreto. Por último, se aborda el problema mediante un sistema de ecuaciones diferenciales estocásticas (EDE) basado en un proceso de difusión, donde tanto el tiempo como el espacio de estados son continuos. En este trabajo se presenta el modelo de CMTC y la inferencia el modelos epidemiológicos compartimentales de este tipo.

La formulación de un modelo epidémico mediante una CMTC consiste en considerar un modelo compartimental cualquiera $\mathbf{X}(t) = (X_1(t), \ldots, X_u(t))$, definido sobre una escala de tiempo continua $t \in [0, \infty)$, donde los estados son variables aleatorias discretas, es decir, $X_i(t) \in \{0, 1, \ldots, N\}$, para todo $i = 1, \ldots, u$. Por la forma de las transiciones entre los posibles compartimentos, este modelo puede verse como un proceso de nacimiento y muerte $(u-1)$-variado, donde las tasas de nacimiento y muerte dependen directamente de las tasas de flujo entre compartimentos.

3.1. Inferencia en modelos epidemiológicos estocásticos

Se trabajará bajo el escenario de información parcial, ya que se supondrá que se tienen reportes de nuevos infectados agrupados cada cierto intervalo de tiempo. Es decir, no se tienen los tiempos exactos de ocurrencia de cada evento y se hace imposible el planteamiento explícito de la verosimilitud. Se utilizarán los métodos ABC vistos en la Sección 1.2.3 que nos permitan realizar inferencia bayesiana en ausencia de la función de verosimilitud. Este tipo de técnicas únicamente requieren la capacidad de simular del modelo del cual provienen los datos, lo cual se logrará con el algoritmo Gillespie que se detalla a continuación.

3.1.1. Ecuación Maestra y Algoritmo Gillespie

La ecuación forward de Kolmogorov asociada al modelo de CMTC se conoce como *ecuación maestra* o CME (Chemical Master Equation) por sus siglas en inglés, y representa la contraparte estocástica al sistema de ecuaciones diferenciales planteado en el Capítulo 2.

La CME es muy utilizada en los campos de física, química y genética para describir cómo evoluciona en el tiempo un sistema que puede tomar exactamente un estado entre una cantidad finita de ellos, en un tiempo determinado, y donde el cambio de un estado a otro se rige por una ley probabilística.

Usando la notación de reacciones químicas planteada en Boys et al. (2008), puede representase a un modelo que describe la evolución de un sistema con u *especies* $\mathbf{X} = (X_1, \ldots, X_u)$ y v *reacciones* posibles $R_1, \ldots, R_v$, como

$$R_k : p_{k1}X_1 + \ldots + p_{ku}X_u \xrightarrow{h_k} q_{k1}X_1 + \ldots + q_{ku}X_u,$$

con $k = 1, \ldots, v$; p_{kj} es la proporción que representa la cantidad de la especie j que entra en la reacción k y q_{kj} es aquella asociada con la cantidad de esta misma especie que se obtiene como producto en la reacción k.

Para un intervalo de tiempo Δt suficientemente pequeño, sólamente una de las v reacciones posibles ocurre. De esta forma, si sucede la reacción k, el j-ésimo reactante X_j cambia en $q_{kj} - p_{kj}$. Cada reacción R_k tiene asociada una

tasa $h_k(\mathbf{X}, \theta_k)$, la cual describe el efecto instantáneo de dicha reacción bajo la cinética de acción de masas. Los tiempos en que suceden las reacciones (a los cuales denotaremos T_k para $k = 1, \ldots, v$) se distribuyen exponencialmente, en consecuencia, el tiempo de la primera reacción (el mínimo entre los tiempos de cada una de las reacciones) es exponencial con parámetro

$$h_0(\mathbf{X}, \boldsymbol{\theta}) = \sum_{k=0}^{v} h_k\left(\mathbf{X}, \theta_k\right),$$

y la k-ésima reacción ocurre con probabilidad $h_k(\mathbf{X}, \theta_k)/h_0(\mathbf{X}, \boldsymbol{\theta})$. Esto constituye una cadena de Markov de saltos puros y hace que el proceso sea fácil de simular utilizando técnicas de simulación discreta. Este método se conoce como algoritmo Gillespie (Gillespie, 1977) y se muestra de manera abreviada a continuación.

Algoritmo 3: Gillespie.

---

1. Establecer valores iniciales $\boldsymbol{\theta}_0$ para el vector de parámetros, y el estado inicial del sistema $\mathbf{X}_0$.

2. Calcular el tiempo mínimo para el siguiente cambio de estado, donde

$$\text{mín}\left\{T_1, \ldots, T_v\right\} \sim exp\left(h_0(\mathbf{X}, \boldsymbol{\theta})\right)$$

3. Determinar el tipo de reacción que ocurre, donde la reacción k-ésima ocurre con probabilidad $h_k(\mathbf{X}, \theta_k)/h_0(\mathbf{X}, \boldsymbol{\theta})$.

4. Actualizar los valores del sistema a los correspondientes después de la reacción que ocurrió, e iterar hasta una condición de paro.

Este algoritmo permitirá simular las transiciones de los individuos entre los posibles estados de un modelo epidemiológico compartimental.

3.2. Modelo SIR estocástico

En el caso del modelo SIR estocástico, el sistema se puede monitorear con los estados $S(t)$ e $I(t)$, ya que $R(t)$ puede obtenerse directamente a partir de estos últimos al suponer la población constante a lo largo del estudio, $R(t) = N - S(t) - I(t)$.

Así, el proceso bivariado a monitorear es $\{S(t), I(t)\}$, el cual tiene asociado una función de probabilidad conjunta dada por $p_{(s,i)}(t) = \mathbb{P}[\{S(t), I(t)\} = (s, i)]$, donde (s, i) es un vector en el espacio de estados posibles del proceso $\{S(t), I(t)\}$.

Del proceso bivariado anterior se deduce el sistema de ecuaciones diferenciales *forward* de Kolmogorov o CME:

$$\begin{aligned}\frac{dp_{(s,i)}}{dt} &= \frac{\beta}{N}(s+1)(i-1)p_{(s+1,i-1)} + \gamma(i+1)p_{(s,i+1)} \\ &\quad - \left[\frac{\beta}{N}si + \gamma i\right] p_{(s,i)}.\end{aligned} \tag{3.1}$$

En este caso se tienen $u = 3$ compartimentos (especies), y los posibles cambios de estado (reacciones) se dan cuando hay una infección o una remoción ($v = 2$), las cuales ocurren a tasas h_1 y h_2, respectivamente:

$$\text{Infección } (R_1)\text{: } S + I \xrightarrow{h_1} 2I$$

$$\text{Remoción } (R_2)\text{: } I \xrightarrow{h_2} R.$$

Estos cambios pueden representarse matricialmente de la siguiente forma:

	R_1	R_2
S	-1	0
I	1	-1
R	0	1

Puede demostrarse (Capistrán et al., 2012) que el límite macroscópico de la CME coincide con el modelo SIR determinista, por lo que las tasas asociadas a las reacciones anteriores están dadas por:

$$h_1 = \frac{\beta SI}{N},$$

$$h_2 = \gamma I.$$

3.2.1. Inferencia en el SIR estocástico

Como se mencionó anteriormente, se considera una población cerrada, libre de inmigración y además se trabaja bajo el supuesto de información parcial, ya que tener el tiempo exacto de ocurrencia de cada uno de los eventos resulta poco

realista.

Al no ser capaces de plantear de manera explícita la función de verosimilitud, se utilizarán los métodos ABC (Sección 1.2.3). Durán Aguilar (2014) realiza un comparativo del proceso de inferencia usando ABC, considerando un estadístico univariado y uno multivariado, y distintas funciones para la función Kernel involucrada en (1.5). Concluye que la mejor combinación resulta al utilizar un estadístico univariado y un Kernel normal. En la sección 3.3.1 se retoma la descripción de la inferencia bajo el modelo estocástico más general SEIR.

3.3. Modelo SEIR estocástico

En el modelo SEIR estocástico aplican los mismos supuestos que para el SEIR determinista revisado en la Sección 2.3, simplemente se formula en términos de una cadena de Markov a tiempo continuo, donde es suficiente el monitoreo de las categorías $S(t)$, $E(t)$ e $I(t)$, ya que puede obtenerse $R(t) = N - S(t) - E(t) - I(t)$. Las posibles transiciones entre los $v = 4$ estados corresponden a nuevas exposiciones, infecciones y remociones del sistema, y se muestran a continuación.

$$\text{Exposición } (R_1)\text{: } S + I \longrightarrow E + I \text{ a tasa } \frac{\beta SI}{N}$$

$$\text{Infección } (R_2)\text{: } E \longrightarrow I \text{ a tasa } \sigma E$$

$$\text{Remoción } (R_3)\text{: } I \longrightarrow R \text{ a tasa } \gamma I.$$

La matriz que representa las reacciones de interés en este modelo es:

	R_1	R_2	R_3
S	-1	0	0
E	1	-1	0
I	0	1	-1
R	0	0	1

3.3.1. Inferencia en el SEIR estocástico

Se planteará el método de inferencia propuesto, que consiste en una generalización del presentado en Durán Aguilar (2014) en dos sentidos. El primero es que el modelo SEIR consta de un compartimento más que el SIR estocástico, modelando una epidemia donde existe un periodo de latencia; y el segundo en cuanto a los datos para realizar la inferencia, ya que en vez de tomar la cantidad exacta

de infectados en ciertos tiempos $t_1, \ldots, t_n$, se consideran los reportes de nuevos infectados agregados en ciertos intervalos de tiempo.

Supongamos que es sensato describir una epidemia mediante un modelo SEIR estocástico con parámetros β, γ y σ desconocidos, y que se tienen datos $\mathbf{y} = y_1, \ldots, y_n$, que corresponden a reportes diarios de nuevos infectados, cuando las tasas de propagación son por hora. Es decir, cada reporte corresponde a los nuevos infectados acumulados en los intervalos de 24 horas $I_1, \ldots, I_n$.

A partir de los datos $\mathbf{y}$, se desea hacer inferencia sobre los parámetros del modelo SEIR que generaron la epidemia. Esto se logrará mediante el algoritmo ABC-MCMC detallado en la Sección 1.2.3, el cual permite simular de una aproximación de la densidad posterior cuya precisión depende del ancho de banda h usado en el Kernel.

En primer lugar se deben inicializar los valores de los parámetros $\beta^{(0)}$, $\gamma^{(0)}$ y $\sigma^{(0)}$ y simular una trayectoria del modelo SEIR estocástico mediante el algoritmo Gillespie. Posteriormente se calculan los reportes agregados diariamente correspondientes a la trayectoria simulada, los cuales corresponderán a las pseudo-observaciones $\mathbf{x}$ involucradas en el Algoritmo 2.

Iterativamente, a partir de una densidad propuesta se genera un valor de los parámetros y un vector de pseudo-observaciones simuladas del modelo con dicho vector de parámetros. Se calcula la probabilidad de aceptar dicho punto mediante (1.9). De esta forma se obtiene una cadena de Markov que tiene como distribución límite una aproximación de la densidad objetivo. Eliminando el periodo de burn-in y considerando el rezago correspondiente, se puede obtener una muestra pseudo-independiente que se distribuye aproximadamente como la densidad posterior de la cual se deseaba simular.

Al tratarse de un modelo estocástico, dado un vector de parámetros fijo, las pseudo-observaciones simuladas a partir de éste tienen una variabilidad que no está siendo explícitamente considerada en el modelo, y se están aceptando únicamente trayectorias *cercanas* a la trayectoria observada. Esto puede hacer muy poco eficiente este método de inferencia sobre todo bajo el escenario en que exista una alta variabilidad del proceso (para cada valor de los parámetros). Considerar como pseudo-observaciones a un promedio de p trayectorias simuladas, o un intervalo de probabilidad, con cada vector de parámetros podría ayudar a delinear los

valores de los parámetros que originan resultados observados más rápidamente, sin embargo, esto incrementaría considerablemente el tiempo de cómputo en otro sentido, ya que en cada punto se requerirían múltiples simulaciones.

Ejemplo

Con fines de explorar el comportamiento del método de inferencia descrito en la Sección 3.3.1, se simuló la dispersión de un agente infeccioso en una población de tamaño $N = 500$ conformada inicialmente con el 5 % de los individuos infectados, con el 95 % restante de la población en el estado susceptible. Las curvas epidémicas correspondientes a cada compartimento del modelo SEIR en esta simulación se pueden observar en la Figura 3.1. Los parámetros a partir de los cuales se genera esta simulación son $\beta = 0.03$, $\gamma = 0.01$ y $\sigma = 0.016$. Dichos parámetros corresponden a las tasas por hora de exposición, infección y remoción, respectivamente.

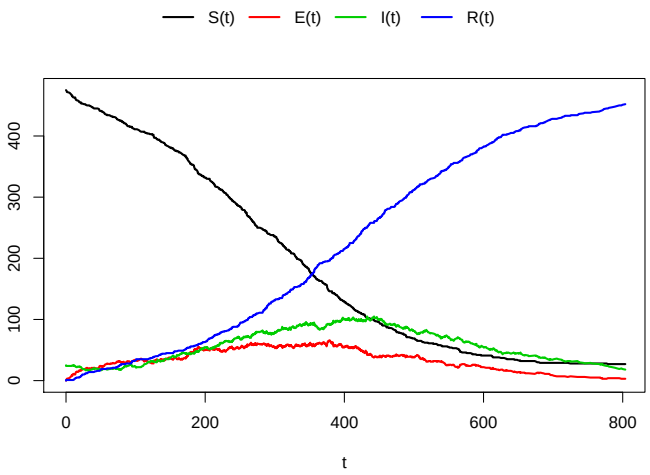

Figura 3.1: Curvas epidémicas correspondientes a la simulación base.

Posteriormente, se calculan los reportes de nuevos infectados agregados por día. Dichos reportes se considerarán como los datos observados a partir de los cuales se desea hacer inferencia de los parámetros. El número de reportes por día se muestra en la Figura 3.2.

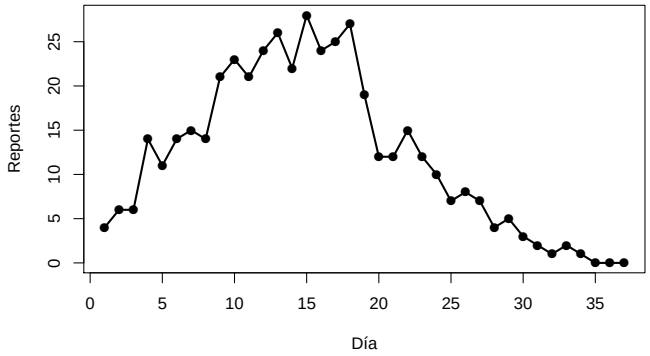

Figura 3.2: Número de reportes observados por día.

Se utilizaron densidades a priori Gamma(1,10) para todos los parámetros, suponiendo que éstas reflejan la información que se tiene de las tasas de transmisión. Se implementó el modelo descrito en la Sección 3.3.1 con 10 mil simulaciones, tomando como vector inicial de parámetros una simulación de la densidad a priori. Puede consultarse el detalle del código en `R` utilizado en http://leticiaramirez2.net/supplementary_material.html.

Para medir las diferencias entre los datos observados ($\mathbf{y}$) y cada vector de datos simulados ($\mathbf{x}$) se utilizó el siguiente estadístico univariado

$$t(\mathbf{x}, \mathbf{y}) = \sum_i \sqrt{\frac{(x_i - y_i)^2}{y_i + \varepsilon}},$$

donde $0 < \varepsilon << 1$ es un valor que permite que el cociente no se indetermine en días que el número de nuevos infectados observado es cero (en este caso se consideró $\varepsilon = 0.001$).

Para la estimación de densidades por Kernel se utilizó un Kernel Normal con ancho de banda proporcional al número de reportes observados, ya que en Durán Aguilar (2014) esta fue la combinación de estadístico y Kernel que hizo más eficientes las simulaciones de la densidad posterior.

Los reportes diarios generados con cada valor de los parámetros propuesto en el algoritmo ABC-MCMC, se muestran con líneas negras en la Figura 3.3. Las trayectorias correspondientes a valores de los parámetros aceptados por el algoritmo se grafican con color azul claro, y en color naranja se grafican como referencia

los reportes observados. Se puede apreciar que se aceptan curvas cercanas a la trayectoria observada.

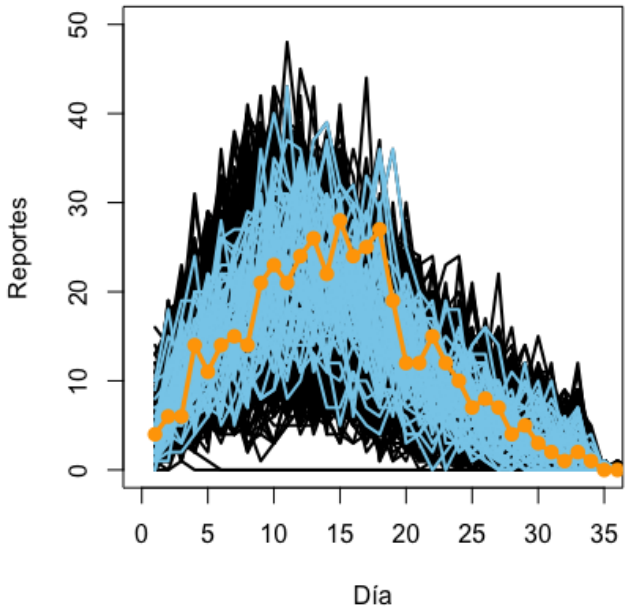

Figura 3.3: Reportes simulados.

Eliminando el periodo de burn-in y el rezago de la misma forma que se hizo en el análisis de las cadenas generadas en la Sección 2.2.1, se obtiene una muestra pseudo-independiente de 55 puntos. Las estimaciones puntuales y por intervalos de probabilidad se muestran en la Tabla 3.1. Se observa que el intervalo para β es el más amplio de los tres y la mediana es cercana al verdadero valor del parámetro. En la última columna se agregan los intervalos y estimación puntual para el parámetro umbral R_0. Todos los intervalos incluyen al parámetro a partir del cual se generaron las observaciones.

Cuantil	β	γ	σ	R_0
2.5 %	0.0157	0.0006	0.0041	1.467
50 %	0.0410	0.0161	0.0169	2.7119
97.5 %	0.1834	0.0685	0.0489	20.4394
Real	0.03	0.01	0.016	2.2

Tabla 3.1: Estimaciones puntuales y por intervalos de probabilidad de los parámetros del modelo SEIR estocástico.

En la Figura 3.4 se presentan los histogramas de las simulaciones de la densidad posterior de los parámetros. Se observa que las densidades de γ y σ presentan

una pequeña moda en un valor mayor al real. La densidad de R_0 tiene una cola pesada hacia la derecha, lo cual también puede observarse en el intervalo de 95 % de probabilidad. A pesar de lo anterior, la mediana de R_0 es cercana al valor real.

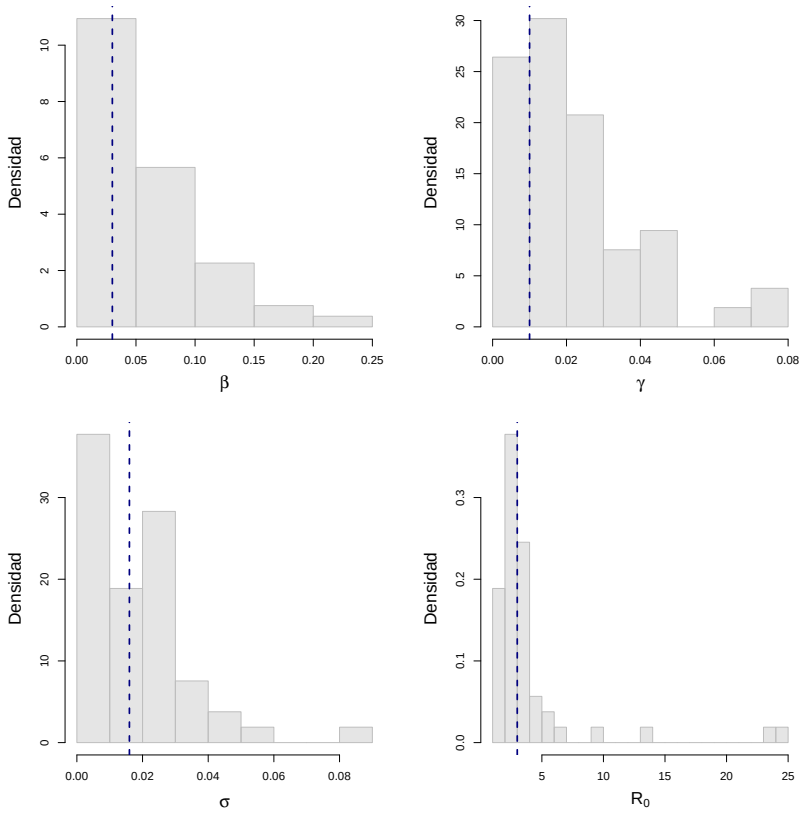

Figura 3.4: Histogramas de la densidad posterior de los parámetros del modelo SEIR estocástico.

Capítulo 4

Modelos compartimentales en redes de contactos

En este capítulo se incorporan al modelo estocástico visto en anteriormente las interacciones entre los individuos de la población de estudio, las cuales se supondrá que pueden modelarse con una red. En la Sección 4.1 se introducen los principales conceptos referentes a gráficas (o grafos), que serán de utilidad al plantear el modelo epidemiológico compartimental en una red. En la Sección 4.2 se describen dos principales aplicaciones de la estadística en datos de redes: el muestreo y la simulación de redes aleatorias. Esta última será de utilidad para el desarrollo del algoritmo ABC-MCMC en una red, que nos permitirá simular de la densidad posterior. Luego, en la Sección 4.3 se plantea el modelo SIR estocástico en una red y se describe cómo puede adaptarse el algoritmo de inferencia presentado anteriormente si se tienen datos en redes. Por último, se presenta un ejemplo de la implementación de la metodología de inferencia propuesta en el modelo SIR estocástico en una red social.

Los modelos epidemiológicos vistos hasta ahora suponen una población grande y homogénea, de tal modo que cada par de individuos tiene igual probabilidad de interactuar dado un intervalo de tiempo fijo. Éste es un supuesto muy común en algunos modelos epidemiológicos, incluyendo los presentados en los Capítulos 2 y 3. Sin embargo, este supuesto de homogeneidad resulta poco realista para algunas infecciones, ya que en la vida real existe una gran heterogeneidad en las tasas de contagio, no solo por la susceptibilidad de los individuos, sino porque existe un patrón poco homogéneo en la interacción entre individuos (Liljeros et al., 2001). Esta diferencia en el número de contactos puede llevar a algunos a tener poca exposición a la infección por los pocos contactos con quienes estan en contactos, mientras que otros pueden estar altamente conectados.

Hacer la extensión de los modelos anteriores, asumiendo que la población tiene una conectividad descrita por una red de contactos sobre la cual se propaga la enfermedad, resulta muy relevante, ya que la mayoría de los procesos de propagación de infecciones en la vida real muestran patrones de conectividad complejos,

dominados por heterogeneidades que pueden modelarse con distribuciones estadísticas de colas pesadas (Barabási, 2014; Pastor-Satorras et al., 2015).

La Teoría de Redes proporciona un marco teórico útil para modelar las interacciones entre los individuos y así plantear un escenario más realista de la dispersión de un agente infeccioso en una población pequeña. En este enfoque se modela la población como una estructura espacial donde los miembros de la misma son nodos de una red, y las aristas de la red representan interacciones entre los individuos que potencialmente pueden llevar a la transmisión de la enfermedad (Newman, 2010; Kolaczyk, 2009).

A la clase de redes que modelan interacciones entre los miembros de una población (entidades sociales) se le denomina redes sociales (Scott, 2000; Wasserman y Faust, 1994). El tipo de interacciones consideradas en una red depende de la naturaleza de las entidades sociales a analizar y el contexto del problema, por ejemplo, algunos tipos de interacción pueden ser la amistad entre individuos, la pertenencia a ciertos grupos políticos o intelectuales, o el intercambio de recursos. Dado esto, algunos ejemplos de redes sociales pueden ser amistades entre niños en una escuela, alianzas corporativas entre compañías, co-autoría en artículos de divulgación, o acuerdos y tratados entre países. El tipo de interacciones consideradas en el problema de nuestro interés (procesos epidémicos) dependerá de la forma en que el agente infeccioso es transmitido y la red social constará de las interacciones de ese tipo entre los individuos de una población objetivo.

Algunas referencias de modelación de epidemias en redes sociales son Pastor-Satorras y Vespignani (2001), quienes analizan datos reales de la esparción de un virus computacional utilizando un modelo de red libre de escala. Por otro lado, Keeling (1999) describe una metodología para modelar el comportamiento y las relaciones de los individuos en una red fija, la cual es aplicada a la dispersión de una enfermedad en una red determinada para encontrar umbrales importantes y algunas propiedades estadísticas. Newman (2002) describe la forma de obtener soluciones exactas del modelo SIR en distintos tipos de redes. Ramírez-Ramírez y Thompson (2013) también consideran la dispersión de un agente en una red aleatoria y estudian el tamaño final del brote, así como su variabilidad, calculando intervalos de probabilidad.

4.1. Conceptos y definiciones generales en teoría de gráficas

En esta sección se enunciarán algunos conceptos acerca de teoría de gráficas extraidos de Kolaczyk (2009) que serán utilizados más adelante en el desarrollo del modelo sobre el que se hará inferencia. Véase también Bollobás (1998), Diestel (2005) y Gross y Yellen (2005) para un mayor detalle acerca de teoría de gráficas.

Una gráfica $\mathcal{G} = (V, E)$ es una estructura matemática que consta de un conjunto finito de nodos o vértices V y un conjunto de aristas E. El conjunto de aristas está conformado por pares $\{u, v\}$ de vértices distintos $u, v \in V$. Una gráfica donde el orden de los vértices que conforman una arista es ordenado, es decir, si $\{u, v\}$ es distinto de $\{v, u\}$, se denomina *gráfica dirigida*. A las aristas de una gráfica dirigida se les llama *arcos* o *aristas dirigidas* y usualmente se representan con flechas. En el caso en que el orden de los vértices en una arista es irrelevante, se tiene una *gráfica no dirigida* y la representación común es como una recta que conecta un vértice con otro.

Una gráfica se denomina simple si no es dirigida y el conjunto de aristas conecta siempre dos diferentes vértices. Esto es, si la gráfica no tiene aristas que conecten a un nodo consigo mismo (bucles). Se denotará al número total de vértices en la gráfica como $N_V = |V|$ y al total de aristas como $N_E = |E|$. Por simplicidad se etiquetará a los vértices con los enteros $1, \ldots, N_V$.

La conectividad de la gráfica puede determinarse por las adyacencias que existen en la misma. Se dice que dos vértices $u, v \in V$ son *adyacentes* si existe una arista en E que conecte u con v. Se dice que una arista $e \in E$ es *incidente* en un vértice $v \in V$ si v es elemento del par de vértices a los que conecta e. De aquí surge la noción de *grado* d_v de un vértice, el cual se define como el número de aristas incidentes a dicho vértice. Por ejemplo, en la Figura 4.1 se muestra una gráfica no dirigida donde los nodos se representan con círculos y las aristas con rectas que conectan a los nodos. El vértice resaltado en rojo tiene grado 3, pues ese vértice tiene tres aristas incidentes.

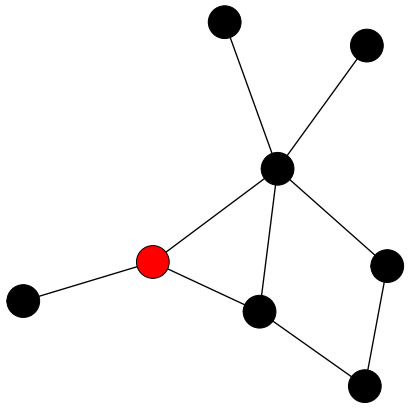

Figura 4.1: Gráfica no dirigida.

En el caso de gráficas dirigidas, se consideran el *grado interno* y el *grado externo* de los vértices, los cuales están dados por el número de arcos que apuntan hacia adentro y hacia afuera de un vértice, respectivamente. En adelante se hablará únicamente de gráficas no dirigidas, ya que éstas serán utilizadas para representar las redes sociales que se plantearán en la Sección 4.3.

El grado d_v de un vértice $v \in V$ nos brinda una cuantificación de la medida en la que v está conectado con los otros vértices de la gráfica. Para definir una medida de la conectividad total de la gráfica deben considerarse en conjunto los grados de cada uno de los vértices $\{d_1, \ldots, d_{N_v}\}$. Dada una gráfica $\mathcal{G}$ que representa a una red, defínase f_d como la fracción de vértices $v \in V$ con grado $d_v = d$. Al conjunto $\{f_d\}_{d \geq 0}$ se le llama *distribución de los grados de* $\mathcal{G}$, y equivale a los datos del histograma de $\{d_1, \ldots, d_{N_v}\}$. La distribución de los grados proporciona de manera resumida la conectividad en la gráfica, y es de mucha utilidad sobre todo en gráficas grandes, ya que proporciona la probabilidad de que un nodo de la red elegido al azar tenga exactamente k conexiones (o vecinos).

En algunas aplicaciones es útil hablar del concepto de movimiento sobre una gráfica. Un *camino* de longitud k en una gráfica es una secuencia alternada de vértices y aristas $\{v_0, e_0, v_1, e_1, v_2, \ldots, v_{k-1}, e_{k-1}, v_k\}$, la cual comienza y termina con vértices. Si la gráfica es no dirigida, los extremos de e_i son v_i y v_{i+1}, y en el caso de una gráfica dirigida, e_i es un arco que va de v_i a v_{i+1}. Un *camino simple* es un camino en el cual todas las aristas son distintas y una *trayectoria* es un camino simple en el cual todos los vértices (excepto posiblemente el primero y el último) son distintos.

Se dice que dos vértices están *conectados* si existe un camino que vaya de uno a otro, de lo contrario estarán desconectados. Dos vértices pueden estar conectados por varios caminos. El número de aristas dentro de un camino es su longitud. Así, los vértices adyacentes (vecinos inmediatos) están conectados por un camino de longitud 1, y los segundos vecinos por un camino de longitud 2. Si un camino empieza y termina en el mismo vértice se le llama ciclo.

La conectividad fundamental de una gráfica $\mathcal{G}$ puede capturarse en una matriz simétrica $\mathbf{A}$ de $N_v \times N_v$ llamada *matriz de adyacencias*, cuyas entradas

$$A_{ij} = \begin{cases} 1, & \text{si } \{i,j\} \in E, \\ 0, & \text{en otro caso,} \end{cases}$$

donde los elementos de V son etiquetados con los enteros $1, \ldots, N_v$ y se representa una arista $e \in E$ como un par no ordenado de vértices. Puesto en palabras, la matriz $\mathbf{A}$ es distinta de cero en aquellas entradas fila-columna que corresponden a vértices que están unidos por una arista en la gráfica $\mathcal{G}$, y toma el valor de cero en las entradas que no satisfagan esta condición.

Otra forma de representar matricialmente la estructura fundamental de $\mathcal{G}$ es mediante la *matriz de incidencias* $\mathbf{B}$, la cual es una matriz de $N_v \times N_e$ cuyas entradas

$$B_{ij} = \begin{cases} 1, & \text{si el vértice } i \text{ es incidente a la arista } j, \\ 0, & \text{en otro caso.} \end{cases}$$

Una gráfica $\mathcal{G}$ puede representarse mediante distintas estructuras de datos. Una de ellas es representar a la gráfica mediante la matriz de adyacencias $\mathbf{A}$ definida previamente, lo cual es una decisión que suena práctica, ya que las matrices son objetos fundamentales en la mayoría de los lenguajes de programación y softwares. Sin embargo, en ocasiones en que la gráfica es muy grande, y en particular cuando hay muchos conjuntos de vértices desconectados, la matriz $\mathbf{A}$ sería de gran dimensión y tendría muchos ceros. Una alternativa a este problema es representar a la gráfica con una colección de listas de adyacencias, en la cual cada elemento corresponde a un grupo de vértices conectados, o simplemente puede usarse una lista de aristas, una lista de dos columnas donde se listen los vértices de cada una de las aristas presentes en $\mathcal{G}$.

En particular, una red social puede representarse matemáticamente con una gráfica donde los vértices (llamados nodos en redes) corresponden a cada uno de los individuos en una población objetivo y las aristas representan las interacciones entre ellos.

4.2. Análisis estadístico de datos en redes

Nos referimos por *datos en redes* a las mediciones que son tomadas dentro de un sistema conceptualizado como una red. Es cuando además de la propia estructura de la red, se tienen medidas que son tomadas dentro de la misma. Estos datos pueden utilizarse para profundizar en las relaciones que existen entre los nodos y hacer predicciones acerca de fenómenos de interés que suceden en la población representada por la red. En una red puede asignarse a los nodos y a las aristas ciertos atributos que proporcionen un mayor entendimiento acerca de su estructura.

Los modelos estadísticos pueden ser útiles para simular redes o para hacer inferencia sobre parámetros de dichos modelos dada una red que representa la estructura de una población (Wasserman, 2013), ya sea tomando en cuenta atributos adicionales o no. A continuación se muestra el planteamiento de muestreo sobre una red y simulación en redes aleatorias. Este último procedimiento estadístico será útil para la inferencia de los parámetros del modelo SIR planteada en la Sección 4.3.1, ya que el algoritmo Gillespie requiere la capacidad de simular el modelo de interés.

4.2.1. Muestreo en redes

En una población típicamente se observa la información relacional de manera parcial y sólo se puede obtener una representación de un subconjunto del sistema complejo que se estudia. En ese caso la red resultante puede pensarse como una muestra de la red compleja subyacente y es posible usar los datos de la muestra para inferir propiedades de la red subyacente utilizando la teoría de muestreo estadístico. La idea de muestrear sobre una red, sin embargo, puede introducir potenciales complicaciones que se abordan en Kolaczyk (2009).

De manera formal, supóngase que el sistema de estudio (red) puede ser representado por una gráfica $\mathcal{G}$, a la cual le llamaremos *gráfica poblacional*. Además, supóngase que en vez de tener disponible toda la información de $\mathcal{G}$, se tienen mediciones que son parte de una muestra de nodos y aristas, que se pueden representar en una gráfica $\mathcal{G}^* = (V^*, E^*)$. A $\mathcal{G}^*$ le llamaremos *gráfica muestra*.

Algunos esquemas de muestreo útiles en la práctica se listan a continuación:

- **Muestreo por gráficas inducidas**. Se consideran *gráficas inducidas* a aquellas gráficas que resultan de tomar una muestra aleatoria de nodos en la

red y observar la subgráfica que inducen considerando sus aristas incidentes. Este esquema se utiliza con frecuencia en análisis de redes sociales como facebook o twitter, donde se toma al azar una muestra de individuos y posteriormente se investiga alguna medida de contacto entre ellos (amistad o cantidad de ”me gusta", por ejemplo).

- **Muestreo por gráficas incidentales**. En este caso se trabaja con *gráficas incidentales*, que consisten en seleccionar al azar aristas dentro de la red y posteriormente completar una subgráfica con los nodos en los que inciden dichas aristas.

- **Muestreo de bola de nieve**. El muestreo de bola de nieve consiste en tomar un nodo inicial y considerar todos sus vecinos inmediatos, y realizar este proceso iterativamente hasta una condición de paro. De esta forma el grupo muestral crece como una bola de nieve que va rodando cuesta abajo, por ello el nombre de la técnica de muestreo. Este tipo de muestreo se utiliza con frecuencia para estudios en *poblaciones ocultas*, como consumidores de drogas.

Pueden consultarse algunas ventajas y desventajas de utilizar estas y otras técnicas de muestreo adicionales en la Sección 5.3 de Kolaczyk (2009).

El objetivo de muestrear sobre una red compleja es generalmente estimar una característica de interés sobre la gráfica poblacional, ya sea una característica estructural como el número de aristas N_e o el grado promedio de los nodos; o bien, un resumen de los atributos de los nodos y vértices de la red, como la proporción de hombres con más amigas que amigos en una red social. Denotemos $\eta(\mathcal{G})$ a la característica de interés de la gráfica $\mathcal{G}$. A pesar de que a partir de una muestra no se podrá recuperar exactamente la característica, se desea obtener una estimación de $\eta(\mathcal{G})$, digamos $\hat{\eta}$ a partir de $\mathcal{G}^*$.

Intuitivamente se utilizaría el estimador *plug-in* $\hat{\eta} = \eta(\mathcal{G}^*)$, sin embargo, estos no son útiles para estimar características de una gráfica, ya que no se cumple el supuesto de independencia de las observaciones de la muestra. Algunas alternativas de estimación se presentan en Granovetter (1976) y Ahmed et al. (2014).

4.2.2. Simulación de redes aleatorias

En este trabajo se considera como *red aleatoria* a una gráfica no dirigida donde el grado de sus nodos sigue cierta distribución de probabilidad. La distribución de

los grados de los nodos debe ser una densidad discreta definida sobre los enteros no negativos.

Para simular una red aleatoria se utiliza el algoritmo Molloy-Reed (Molloy y Reed, 1995), el cual se explica brevemente a continuación. Supóngase que se desea simular una red aleatoria con N_v nodos. En primer lugar se generan los grados de los nodos de la red $\{d_1, \ldots, d_{N_v}\}$. El grado de cada nodo puede ser: a) fijo y especificado por el usuario, b) $n-1$ para generar una gráfica completa, o c) seguir cualquier distribución de probabilidad discreta y sobre los enteros no negativos. Las distribuciones implementadas son:

- Poisson,
- Poisson truncada (Modificación de la densidad Poisson con soporte $\{1, 2, \ldots\}$),
- Geométrica,
- Geométrica truncada (Modificación de la densidad Geométrica con soporte $\{1, 2, \ldots\}$),
- Binomial negativa,
- Polilogarítmica,
- Logarítmica,
- De ley de potencias (o powerlaw).

El siguiente paso en el algoritmo es generar la arista $\{u, v\}$, con $u, v \in V$ con probabilidad proporcional al producto de los grados d_u y d_v. Posteriormente se actualiza el grado de los vértices (restando uno a los nodos que se unieron), para considerar las conexiones que aún se pueden establecer, a éste nuevo grado se le llama *grado disponible*. Este proceso se continúa iterativamente, seleccionando en cada paso una arista con probabilidad proporcional al producto de los grados disponibles de los nodos que la conforman. El código de simulación en `R` se puede consultar en `http://leticiaramirez2.net/supplementary_material.html`.

Es importante mencionar que no toda sucesión de grados puede corresponder a una gráfica. La prueba de existencia de una gráfica asociada al conjunto $\{d_1, \ldots, d_{N_v}\}$ se puede realizar usando el teorema de Havel-Hakimi. Éste no solo nos dice si la sucesión puede ser asignada a una gráfica, sino que produce una. Sin embargo este teorema-algoritmo produce gráficas con la misma estructura y el algoritmo que se propone usar produce de manera aleatoria una grafica sobre

el conjunto de gráficas con la misma secuencia de grados.

Aunque bajo este esquema de generación de gráficas los grados de la gráfica resultante pueden no corresponder completamente con la secuencia original, esta diferencia tiende a cero cuando el número de nodos crece.

4.3. Modelo SIR en una red social

Se planteará el modelo SIR estocástico en una red social, donde cada nodo corresponde a un individuo y tiene un atributo asignado que describe su estado respecto a la enfermedad, es decir, el compartimento dentro del cual se encuentra dicho individuo.

Supóngase que se tiene una red $\mathcal{G}$ con N_v nodos etiquetados como $1, \ldots, N_v$. A continuación se muestra el pseudo-algoritmo para simular en $\mathcal{G}$ la dispersión de un agente infeccioso que se modela con un SIR estocástico de parámetros β y γ. Se supondrá que el número inicial de individuos infectados es i_0 y que al tiempo cero no existen aún individuos recuperados, por lo que el sistema inicial es $(N_v - i_0, i_0, 0)$. También se supondrá que un individuo infeccioso únicamente es capaz de infectar a sus vecinos inmediatos susceptibles.

El tiempo que los individuos pasan en el estado infeccioso puede ser fijo o tener una distribución de probabilidad de rango positivo, por ejemplo lognormal o exponencial. Nos enfocaremos en el caso exponencial, análogo al planteado en la Sección 3.2.

Algoritmo 4: Simulación del SIR estocástico en una red.

1. Elegir i_0 nodos de la red, ya sea de manera determinista o aleatoria, los cuales serán etiquetados como infecciosos al tiempo 0. Los demás nodos se etiquetan como susceptibles en esta etapa.

2. Determinar el tiempo del siguiente cambio, y a qué tipo de reacción corresponde (infección o recuperación) como se describe en el algoritmo Gillespie (Algoritmo 3) y actualizar el estado de cada nodo.

3. Iterar el proceso hasta un tiempo de observación máximo (si existe) o hasta que no haya más individuos infecciosos.

Se considera que las transmisiones de las aristas son independientes para cada conexión de un susceptible con un infeccioso. El código en `R` para generar estas simulaciones se muestra en http://leticiaramirez2.net/supplementary_material.html. El algoritmo para simular el agente infeccioso es análogo al SIMID (SIMulation of Infectious Diseases) implementado en Ramírez-Ramírez et al. (2013), y difieren en que con este último se pueden considerar en el modelo algunos esquemas de vacunación como políticas de control de la epidemia.

4.3.1. Inferencia del modelo SIR estocástico en una red

Supóngase que se modela la evolución de una epidemia con un SIR estocástico con tasas de transmisión y recuperación por hora β y γ, respectivamente, y que se tiene una red social que describe los contactos entre los individuos de una población, la cual puede ser representada con una gráfica. Además, supóngase que se cuenta con n reportes de nuevos infectados agregados diariamente $y_1, \ldots, y_n$.

A partir de los reportes anteriores se desea hacer inferencia acerca de los parámetros que rigen la evolución de la epidemia en cuestión. Al tener una forma de simular el modelo en la red para un vector fijo de parámetros (β, γ), puede utilizarse el Algoritmo 2 (ABC-MCMC) para obtener simulaciones de la densidad posterior de los parámetros. Es necesario agrupar en las simulaciones del modelo el número de reportes de nuevos infectados por día para hacer los datos simulados comparables con los observados.

Este procedimiento de inferencia puede generalizarse para cualquier modelo compartimental, únicamente se requiere poder simular dicho modelo en una red con un algoritmo análogo al Algoritmo 4. Por ejemplo, para el modelo SEIR únicamente habría que agregar el periodo de exposición, al cual se le puede asignar una distribución. Sin embargo, si únicamente se tiene información de los nuevos infectados del proceso como aquí se supone, si el número de parámetros del modelo aumenta, las estimaciones puntuales se vuelven menos precisas y los intervalos de probabilidad más amplios.

Ejemplos

Para ilustrar el algoritmo de inferencia del modelo SIR estocástico en una red, se simularon dos redes aleatorias con 500 nodos con las siguientes distribuciones de grados:

- Poisson(2.42)

- Polilogarítmica(0.1,2)

La particularidad de estas distribuciones es que con los parámetros anteriores tienen la misma media, pero tienen un comportamiento distinto. En las Figuras 4.2 y 4.3 se muestra el histograma de los grados observados en las redes simuladas a partir de la densidad poisson y la polilogarítmica, respectivamente. Se señala la función de masa de probabilidades teórica correspondiente a dichas densidades con puntos dentro de las mismas gráficas.

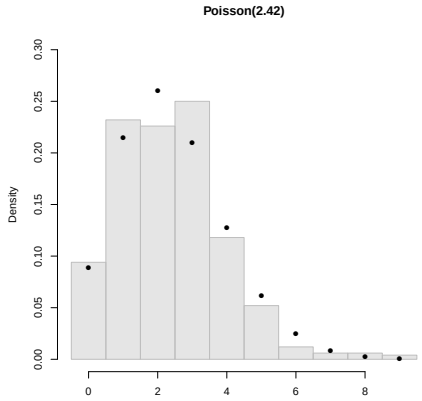

Figura 4.2: Distribución de los grados de la red Poisson simulada.

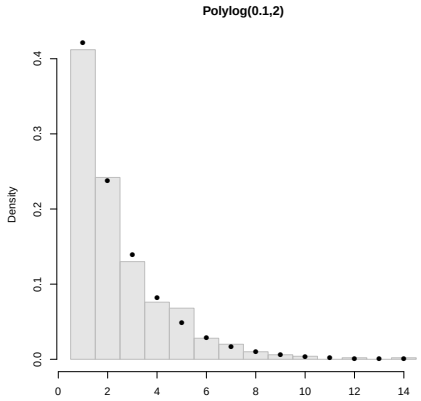

Figura 4.3: Distribución de los grados de la red Polilogarítmica simulada.

Supondremos que las redes simuladas modelan los contactos entre los individuos de una población hipotética. Fijando como estado inicial a dos individuos infectados, se simuló la dispersión de un agente infeccioso en la red con un modelo SIR estocástico, considerando $\beta = 0.03$ y $\gamma = 0.01$ como las tasas por hora de infección y recuperación, respectivamente. En las Figuras 4.4 y 4.5 se muestran las curvas de infecciosos para los tiempos en los que ocurrió un cambio en el proceso, las cuales presentan un comportamiento similar.

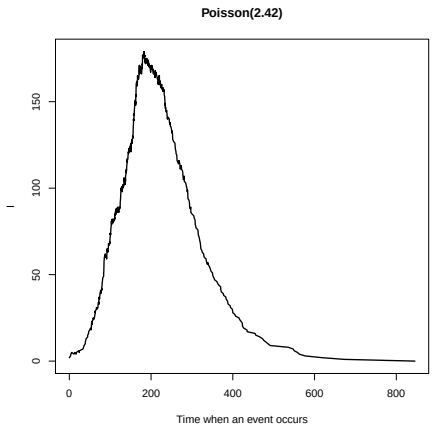

Figura 4.4: Curva de infecciosos (Poisson)

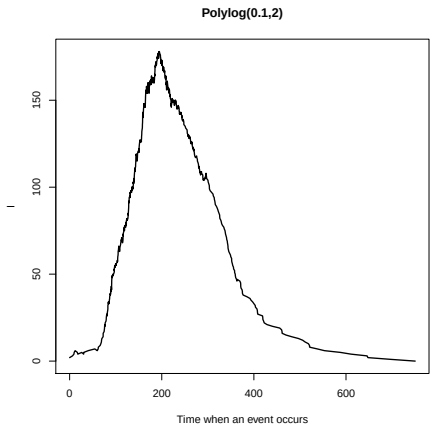

Figura 4.5: Curva de infecciosos (Polilogarítmica)

Bajo un escenario más realista, se considera que se tienen reportes diarios de nuevos infectados $\mathbf{y} = y_1, \ldots, y_n$. Es decir, se supondrá que no se cuenta con los tiempos exactos de cada cambio en el sistema, sino con el número de nuevos individuos infectados acumulados en periodos de 24 horas, con lo cual se obtienen 35 reportes que se observan en las gráficas 4.6 y 4.7.

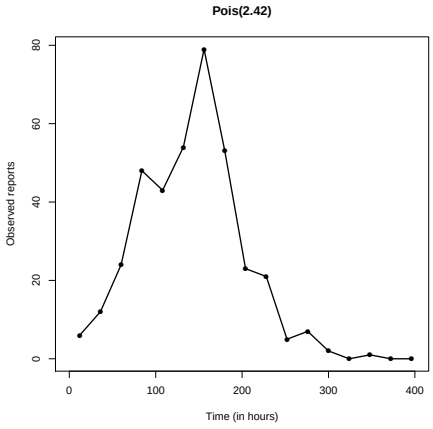

Figura 4.6: Reportes observados (Poisson)

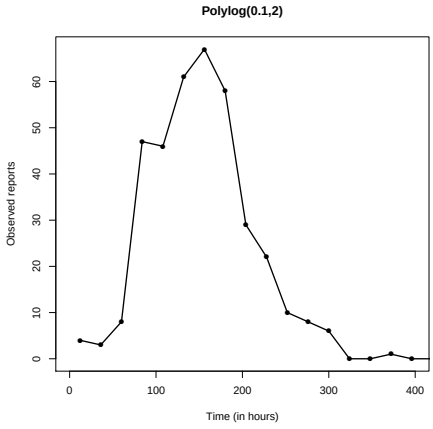

Figura 4.7: Reportes observados (Polilogarítmica)

Para medir la diferencia entre los datos observados $\mathbf{y}$ y una simulación $\mathbf{x}$ dada, se considera el siguiente estadístico univariado:

$$t(\mathbf{x}, \mathbf{y}) = \sum_i \sqrt{\frac{(x_i - y_i)^2}{y_i + \varepsilon}},$$

donde $0 < \varepsilon << 1$.

Para la estimación de la densidad del error de aproximación de la densidad posterior (ver Algoritmo 2) se considera un Kernel Normal con ancho de banda proporcional al número de reportes observados. Además, para la implementación del Algoritmo ABC-MCMC, se utilizó como densidad propuesta una mezcla de densidades normales, una con varianza más pequeña que la otra para permitir una mejor exploración del espacio paramétrico:

$$q(\theta' \mid \theta) = r\, N(\theta, \Sigma_1) + (1 - r)\, N(\theta, \Sigma_2),$$

donde $\Sigma_1 = 0.05Id$, $\Sigma_2 = 0.1Id$ y $0 < r < 1$.

Además se utilizaron densidades a priori Gamma(1,10) para ambos parámetros, que se supondrá que reflejan la información previa que se tiene acerca de las tasas de infección y recuperación del modelo SIR estocástico. Como se explicó en la Sección 1.2, el ABC-MCMC es un algoritmo de aceptación y rechazo, donde los puntos propuestos son aceptados con cierta probabilidad. En las Figuras 4.8 y 4.9 pueden observarse las curvas de individuos infecciosos que corresponden a parámetros rechazados por el modelo (negro), así como las que fueron aceptadas (azul claro), que son cercanas a la curva obtenida con los reportes observados (naranja).

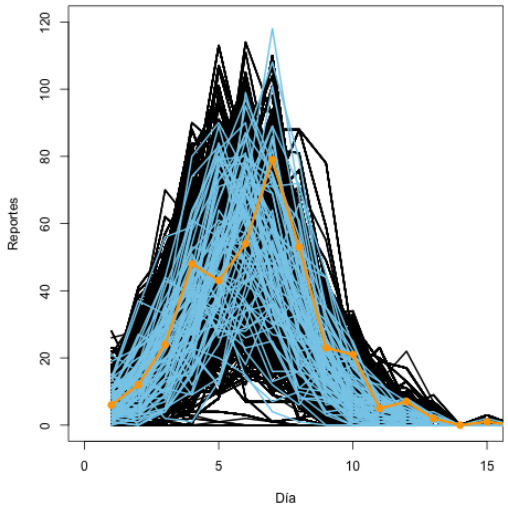

Figura 4.8: Curvas simuladas con parámetros aceptados (Poisson)

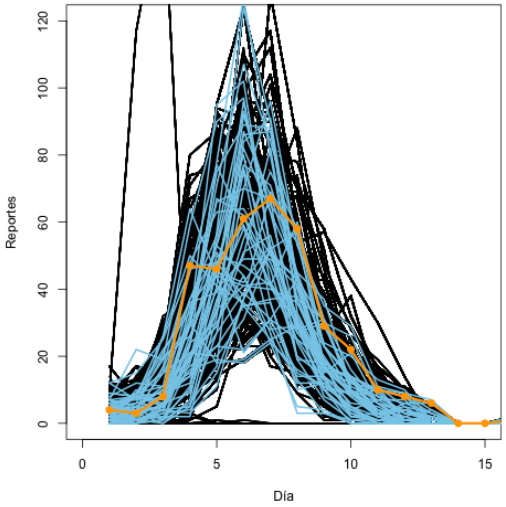

Figura 4.9: Curvas simuladas con parámetros aceptados (Polilogarítmica)

Después de analizar la convergencia de la cadena simulada y eliminar el burn-in, y considerando los rezagos correspondientes obtenidos mediante el IAT, se obtienen simulaciones de la densidad posterior que dan lugar a las estimaciones puntuales y por intervalos de 95 % de probabilidad que se muestran en las Tablas 4.1 y 4.2.

Cuantil	β	γ
2.5 %	0.0286	0.0072
50 %	0.0440	0.0214
97.5 %	0.0681	0.0520
Real	0.03	0.01

Tabla 4.1: Estimaciones puntuales y por intervalos de probabilidad (Red Poisson)

Cuantil	β	γ
2.5 %	0.0234	0.0055
50 %	0.0346	0.0244
97.5 %	0.0560	0.0562
Real	0.03	0.01

Tabla 4.2: Estimaciones puntuales y por intervalos de probabilidad (Red Polilogarítmica)

Se observa que las amplitudes de los intervalos son similares en ambas redes, y que los intervalos contienen a los parámetros con que se simularon los datos $(0.03, 0.01)$. Los histogramas de la densidad posterior de los parámetros se muestran en las Figuras 4.10 (red Poisson) y 4.11 (red Polilogarítmica). En este caso que se agrega la estructura de las conexiones entre los individuos de la población, se observa que las modas de los histogramas están a la derecha del valor real en todos los casos, sin embargo, la robustez de la mediana ayuda a que este estimador puntual no se aleje mucho del valor real a partir del cual se simularon los datos.

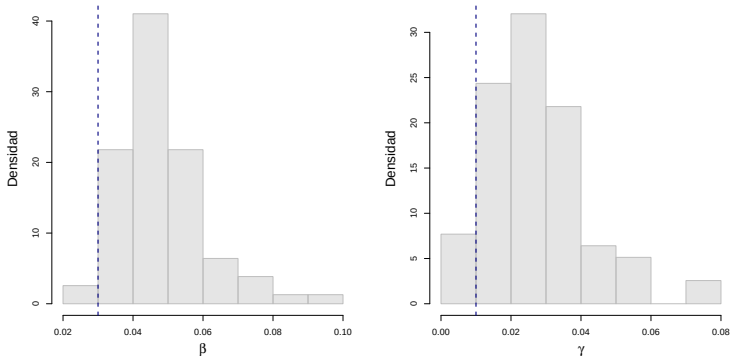

Figura 4.10: Histogramas de la densidad posterior de los parámetros (Red Poisson).

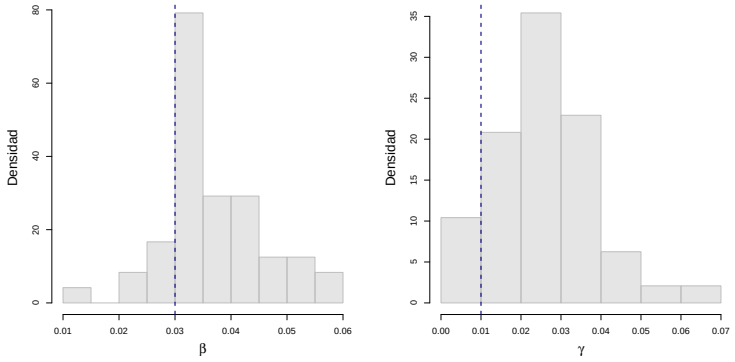

Figura 4.11: Histogramas de la densidad posterior de los parámetros (Red Polilogarítmica).

En el caso de inferencia en una red, el IAT detecta la dependencia en la cadena con mayor facilidad y es necesario considerar un rezago mayor que en el caso donde se supone que los individuos interactúan bajo la ley de acción de masas. Esto produce un tamaño de muestra menor y en consecuencia es necesario realizar más simulaciones que en los modelos anteriores.

Conclusiones y trabajo futuro

Un modelo epidemiológico que sea capaz de describir la evolución de una enfermedad en una población puede ayudar no solo a planear sobre los recursos de salud necesarios para atender a los futuros casos, sino a evaluar la eficacia de las medidas sanitarias que buscan combatir la dispersión de la enfermedad. Los modelos compartimentales estudiados en este trabajo dependen de una serie de parámetros relacionados con las características del agente y algunas características de la transmisión en la población. La estimación de estos parámetros es relevante, ya que cualidades como la dimensión de un brote y la duración de una epidemia están asociadas con dichos parámetros.

Se estudiaron dos puntos de vista para plantear los modelos, el determinista y el estocástico. Un modelo determinista supone una población de gran magnitud, y en donde cualquier par de individuos tiene la misma probabilidad de contacto, y el sistema tenga una solución única para cualquier tiempo $t > 0$. Bajo el en foque estocástico (estocasticidad demográfica), se asigna a los tiempos que los individuos pasan en cada compartimento una distribución de probabilidad, lo cual se refleja en la aleatoriedad de la solución del sistema en un tiempo fijo.

El planteamiento de un modelo epidemiológico estocástico se puede hacer de distintas maneras, de acuerdo a los supuestos que se establezcan sobre el tiempo y el espacio de estados del sistema. En este caso se supuso que el sistema se desarrollaba en una escala de tiempo continua, pero que los posibles estados del sistema eran variables aleatorias discretas. Si se asigna una distribución exponencial a los tiempos de cambio en el sistema, se trata de un Proceso Poisson no homogéneo, donde las tasas al tiempo $t > 0$ dependen del número de individuos que en ese instante se encuentren en cada uno de los compartimentos.

Adicionalmente, se agregó al modelo una estructura relacional de los individuos mediante una red social, de tal forma que un individuo únicamente fuera capaz de contagiar a sus vecinos inmediatos. Este planteamiento permite un enfoque más realista, sin embargo las simulaciones y la inferencia se vuelve más pesada computacionalmente.

El objetivo principal de este trabajo fue realizar inferencia sobre los parámetros de un modelo epidemiológico compartimental bajo los modelos mencionados anteriormente. Se supuso un escenario apegado a la realidad, en el cual se contaba

con una serie de reportes de nuevos infectados $y_1, \ldots, y_n$, a partir de los cuales se pretendía realizar inferencia sobre los parámetros del modelo. En todos los casos se asume que los reportes son el número de nuevos infectados en un lapso de tiempo. En el caso particular del modelo determinista además se considera que las observaciones están sujetas a ruido. Este ruido puede tener múltiples fuentes pero el total de estas variaciones la modelamos con una distribución de probabilidad, con media igual al número observado de reportes de nuevos infectados.

En todos los casos se plantearon esquemas y modelos bayesianos de inferencia estadística, ya que en el contexto de epidemiología generalmente se cuenta con información previa acerca de los parámetros (tasas de transferencia entre los compartimentos). La información previa puede ser obtenida analizando enfermedades similares, por ejemplo. Si dicha información puede traducirse en términos de una distribución de probabilidad a priori para los parámetros, la distribución posterior nos proporciona información del vector de parámetros del modelo donde contribuyen tanto la información previa del femómeno, como la verosimilitud de los datos observados.

A partir de la densidad posterior es posible obtener estimaciones puntuales e intervalos de confianza para los parámetros, y los algoritmos MCMC nos hacen más sencilla la simulación de dicha densidad.

En el modelo determinista se simuló de la densidad posterior usando el *t - walk*, un algoritmo de Metrópolis-Hastings desarrollado por Christen et al. (2010) e implementado en varios softwares para análisis estadístico. Dicho algoritmo facilita la simulación de cualquier distribución objetivo y no depende de parámetros de *tuning* para hacer converger a la cadena simulada.

En el modelo estocástico, los tiempos en que los individuos cambian de compartimento se modelan como variables aleatorias con una distribución que depende del estado actual del sistema. Así, para poder plantear la verosimilitud de los datos es necesario contar con todos los tiempos de cambio del sistema y en el caso estudiado únicamente se tenían reportes agregados. Para el planteamiento de la densidad posterior exacta es necesaria la verosimilitud, sin embargo, el algoritmo ABC-MCMC permite simular de una aproximación de la posterior a pesar de no contar con este término. Este algoritmo funciona si es posible simular del modelo dado un vector de parámetros fijo, lo cual se logra con el algoritmo Gillespie.

Se desarrollaron algunos ejemplos con el modelo SIR y el SEIR para explorar y mostrar a detalle cómo se desarrolla el proceso de inferencia en ambos modelos,

además de plantear el modelo estocástico en una red. Algunas características observadas al desarrollar el proceso de inferencia en cada uno de los modelos fueron las siguientes:

- En el modelo determinista las simulaciones se demoraron más de lo esperado, ya que para cada punto involucrado en el algoritmo debe aproximarse numéricamente la integral de la tasa de nuevos infectados.
- En el modelo estocástico se simula de una aproximación de la verosimilitud, cuya precisión depende del ancho de banda elegido para el Kernel involucrado en el algoritmo ABC-MCMC. El algoritmo desarrollado en este trabajo podría mejorarse buscando un método para elegir el ancho de banda óptimo que nos produzca la precisión deseada.
- El algoritmo ABC-MCMC nos permitió simular de la densidad posterior de los parámetros en el modelo estocástico a pesar de que no se tenía la información completa de los tiempos de cambio del sistema. El proceso de inferencia al considerar una red para modelar la estructura de dependencia de la población se vuelve computacionalmente más intensivo, ya que el IAT es mayor y en consecuencia se debe considerar un rezago más grande en la cadena simulada.
- Comparando los ejemplos implementados para inferencia en el modelo SIR y el SEIR, se observó que el tiempo de cómputo se incrementó al considerar una mayor cantidad de parámetros. En general, si la estructura del modelo es más compleja, considerar el número de reportes de nuevos infectados brinda menos información y la inferencia se vuelve menos precisa.

Una posible deficiencia del modelo de inferencia sobre redes grandes es la eficiencia de las simulaciones. En estudios futuros esto podría mejorarse utilizando algún método de procesamiento en paralelo, o migrando el código a algún otro lenguaje de programación más eficiente como C o Python. Otra alternativa al tratarse de una red es considerar un horizonte menor al final de la epidemia para realizar la inferencia, ya que, a pesar de tener menos información, esto podría reflejarse en un ahorro considerable de tiempo de cómputo.

Varias generalizaciones puede sugerirse a partir de este trabajo. Algunas relacionadas con el modelo, tal como considerar escenarios más realistas, como un horizonte de tiempo más amplio donde se permita la migración de los individuos. Otras mejoras puede apuntar a la eficiencia de los métodos de cómputo, como ya se mencionó. Otra dirección a explorar es sobre variantes de los métodos de inferencia usados y el análisis de sus propiedades.

Bibliografía

Ahmed, N. K., Neville, J., y Kompella, R. Network sampling: From static to streaming graphs. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(2):7, 2014. (Citado en página 51.)

Allen, L. J. An introduction to stochastic epidemic models. In *Mathematical epidemiology*, pages 81–130. Springer, 2008. (Citado en páginas 21 y 35.)

Anderson, R. M., May, R. M., y Anderson, B. *Infectious diseases of humans: dynamics and control*, volume 28. Wiley Online Library, 1992. (Citado en página 21.)

Andrieu, C. y Thoms, J. A tutorial on adaptive mcmc. *Statistics and Computing*, 18(4):343–373, 2008. (Citado en página 8.)

Aster, R. C., Borchers, B., y Thurber, C. H. *Parameter estimation and inverse problems*, volume 90. Academic Press, 2011. (Citado en página 17.)

Barabási, A.-L. Network science book. *Network Science*, 2014. (Citado en página 46.)

Bernardo, J. M. y Smith, A. F. Bayesian theory, 2001. (Citado en páginas 5 y 19.)

Blasco, A. Bayesian statistic course. 2005. (Citado en página 5.)

Bollobás, B. Random graphs. In *Modern Graph Theory*, pages 215–252. Springer, 1998. (Citado en página 47.)

Bolstad, W. M. *Introduction to Bayesian statistics*. John Wiley & Sons, 2013. (Citado en página 19.)

Box, G. E. y Tiao, G. C. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 2011. (Citado en página 5.)

Boys, R. J., Wilkinson, D. J., y Kirkwood, T. B. Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing*, 18(2):125–135, 2008. (Citado en página 36.)

Brauer, F. Compartmental models in epidemiology. In *Mathematical epidemiology*, pages 19–79. Springer, 2008. (Citado en páginas 20, 29 y 30.)

Braunack-Mayer, L. Approximate bayesian computation and summary statistic selection in epidemic models. 2013. (Citado en página 11.)

Capistrán, M. A., Christen, J. A., y Velasco-Hernández, J. X. Towards uncertainty quantification and inference in the stochastic sir epidemic model. *Mathematical biosciences*, 240(2):250–259, 2012. (Citado en página 38.)

Christen, J. A., Fox, C., et al. A general purpose sampling algorithm for continuous distributions (the t-walk). *Bayesian Analysis*, 5(2):263–281, 2010. (Citado en páginas 8, 24, 32 y 61.)

Del Moral, P., Doucet, A., y Jasra, A. An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and Computing*, 22(5):1009–1020, 2012. (Citado en páginas 9 y 10.)

Diestel, R. Graph theory, vol. 173 of. *Graduate Texts in Mathematics*, 2005. (Citado en página 47.)

Durán Aguilar, J. Inferencia bayesiana en el modelo SIR. Tesis de maestría, CIMAT, 2014. (Citado en páginas 11, 39 y 42.)

Geyer, C. J. Practical markov chain monte carlo. *Statistical Science*, pages 473–483, 1992. (Citado en página 25.)

Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977. (Citado en página 37.)

Granovetter, M. Network sampling: Some first steps. *American Journal of Sociology*, pages 1287–1303, 1976. (Citado en página 51.)

Gross, J. L. y Yellen, J. *Graph theory and its applications*. CRC press, 2005. (Citado en página 47.)

Haario, H., Saksman, E., y Tamminen, J. An adaptive metropolis algorithm. *Bernoulli*, pages 223–242, 2001. (Citado en página 8.)

Haran, M. An introduction to models for disease dynamics. *Spatial Epidemiology, SAMSI*, 2009. (Citado en página 1.)

Hastings, W. K. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. (Citado en página 6.)

Herrera Reyes, A. Una aproximación integral al modelamiento de la epidemiología del virus sincicial respiratorio. Tesis de licenciatura, Universidad de Guanajuato, 2010. (Citado en página 2.)

Hethcote, H. W. Qualitative analyses of communicable disease models. *Mathematical Biosciences*, 28(3):335–356, 1976. (Citado en página 21.)

Hethcote, H. W. The mathematics of infectious diseases. *SIAM review*, 42(4): 599–653, 2000. (Citado en páginas 21 y 29.)

Hindmarsh, A. C. Odepack, a systematized collection of ode solvers, rs stepleman et al.(eds.), north-holland, amsterdam,(vol. 1 of), pp. 55-64. *IMACS transactions on scientific computation*, 1:55–64, 1983. (Citado en página 21.)

Keeling, M. J. The effects of local spatial structure on epidemiological invasions. *Proceedings of the Royal Society of London B: Biological Sciences*, 266(1421): 859–867, 1999. (Citado en página 46.)

Kermack, W. O. y McKendrick, A. G. A contribution to the mathematical theory of epidemics. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 115, pages 700–721. The Royal Society, 1927. (Citado en página 20.)

Kolaczyk, E. D. *Statistical Analysis of Network Data: Methods and Models*. Springer Publishing Company, Incorporated, 1st edition, 2009. ISBN 038788145X, 9780387881454. (Citado en páginas 46, 47, 50 y 51.)

Liljeros, F., Edling, C. R., Amaral, L. A. N., Stanley, H. E., y Åberg, Y. The web of human sexual contacts. *Nature*, 411(6840):907–908, 2001. (Citado en página 45.)

Marin, J.-M., Pudlo, P., Robert, C. P., y Ryder, R. J. Approximate bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012. (Citado en páginas 9 y 10.)

Marjoram, P., Molitor, J., Plagnol, V., y Tavaré, S. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26): 15324–15328, 2003. (Citado en página 13.)

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

Molloy, M. y Reed, B. A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2-3):161–180, 1995. (Citado en página 52.)

Newman, M. *Networks: an introduction*. Oxford university press, 2010. (Citado en página 46.)

Newman, M. E. Spread of epidemic disease on networks. *Physical review E*, 66 (1):016128, 2002. (Citado en página 46.)

Pastor-Satorras, R. y Vespignani, A. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001. (Citado en página 46.)

Pastor-Satorras, R., Castellano, C., Van Mieghem, P., y Vespignani, A. Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925, 2015. (Citado en página 46.)

Petzold, L. Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations. *SIAM journal on scientific and statistical computing*, 4(1):136–148, 1983. (Citado en página 21.)

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL http://www.R-project.org. ISBN 3-900051-07-0. (Citado en páginas 3 y 9.)

Ramírez-Ramírez, L. L. y Thompson, M. E. Applications of the variance of final outbreak size for disease spreading in networks. *Methodology and Computing in Applied Probability*, 16(4):839–862, 2013. ISSN 1573-7713. doi: 10.1007/s11009-013-9325-z. URL http://dx.doi.org/10.1007/s11009-013-9325-z. (Citado en página 46.)

Ramírez-Ramírez, L. L., Gel, Y. R., Thompson, M., de Villa, E., y McPherson, M. A new surveillance and spatio-temporal visualization tool simid: Simulation of infectious diseases using random networks and gis. *Computer methods and programs in biomedicine*, 110(3):455–470, 2013. (Citado en página 54.)

Robert, C. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer, 2 edition, 2007. (Citado en página 5.)

Robert, C. y Casella, G. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013. (Citado en páginas 6 y 10.)

Roberts, G. O., Rosenthal, J. S., et al. Optimal scaling for various metropolis-hastings algorithms. *Statistical science*, 16(4):351–367, 2001. (Citado en página 25.)

Roberts, M. y Heesterbeek, J. Mathematical models in epidemiology. *Mathematical models*, 2003. (Citado en página 28.)

Rubin, D. B. et al. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172, 1984. (Citado en página 9.)

Scott, J. Social network analysis: A handbook. 2000. (Citado en página 46.)

Soetaert, K., Petzoldt, T., y Setzer, R. W. Solving differential equations in r: package desolve. *Journal of Statistical Software*, 33, 2010. (Citado en páginas 17 y 21.)

Stuart, A. M. Inverse problems: a bayesian perspective. *Acta Numerica*, 19: 451–559, 2010. (Citado en página 16.)

Tsybakov, A. B. *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, New York, 2009. (Citado en página 12.)

Turlach, B. A. et al. *Bandwidth selection in kernel density estimation: A review.* Université catholique de Louvain, 1993. (Citado en página 12.)

Wasserman, L. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013. (Citado en página 50.)

Wasserman, S. y Faust, K. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994. (Citado en página 46.)

Wilkinson, R. D. Approximate bayesian computation (abc) gives exact results under the assumption of model error. *Statistical applications in genetics and molecular biology*, 12(2):129–141, 2013. (Citado en página 10.)