Instituto Tecnológico Autónomo de México

Procesos de Duración Marcados para el estudio de trayectorias de padecimientos crónico-degenerativos

Tesis

que para obtener el título

Licenciada en Actuaría

presenta

ADRIANA PÉREZ ARCINIEGA SOBERÓN

Asesor:
dr. juan carlos martínez ovando

México, D.F. 2017

Çon fundamento en los artículos 21 y 27 de la Ley Federal del Derecho de Autor y como titular de los derechos moral y patrimonial de la obra titulada "**TÍTULO DE LA TESIS**", otorgo de manera gratuita y permanente al Instituto Tecnológico Autónomo de México y a la Biblioteca Raúl Bailléres Jr., la autorización para que fijen la obra en cualquier medio, incluido el electrónico, y la divulguen entre sus usuarios, profesores, estudiantes o terceras personas, sin que pueda percibir por tal divulgación una contraprestación".

AUTOR

Fecha

Firma

Índice general

Agradecimientos

Muchas gracias a todos!

Capítulo 1

Introducción

1.0.1. Introducción

Introducción general a la tesis

1.0.2. Definición de Variables Latentes

$$P(x) = \int p(x|\theta) f(d\theta)$$

θ es una variable no observable o latente.

Capítulo 2

Procesos de Duración Marcada

2.1. Introducción

En este capítulo se tratarán los fundamentos de los modelos de duración y de duración marcada y su aplicación al tema de descripción de evolución de una enfermedad y su tratamiento. Además de discutir algunas nociones de dependencia estocástica relevantes, tales como independencia estocástica, intercambiabilidad y estacionareidad que son vitales para realizar inferencia y predicción de los datos.

Después de que el proceso puntual sea definido más adelante, se introducirá el concepto de variables latentes para poder construir las distribuciones de las variables observables, concepto que igualmente será definido en una sección posterior. Para un individuo particular, estas variables conectan las observaciones a través del tiempo creando el modelo general de probabilidad que permite hacer inferencia y predicción sobre futuras observaciones, a nivel individual. Tomando el modelo gene-

ral de probabilidad, se elegirán las distribuciones que mejor se adaptan al comportamiento de las variables.

Es importante remarcar que el desarrollo de los procesos puntuales ha estado unido al desarrollo de la estadística actuarial y de seguros, Daley and Vere-Jones (2003) se refiere a las tablas de mortalidad como el primer estudio de procesos de intervalos, entre otros, los procesos de renovación [1], los procesos Markov y semi-Markov. Por lo que el uso de estos procesos como una forma de tarificar es solamente otra colaboración en la larga lista de estas dos disciplinas.

2.2. Definiciones y notación

Para el objeto de este estudio consideramos el caso donde está disponible la información desagregada de un conjunto de individuos respecto a las transiciones de etapas de tratamiento para una enfermedad crónico-degenerativa, cada etapa con la información sobre sus respectivas duraciones y costos. A su vez, cada uno de los individuos tiene asociadas covariables sociodemográficas, socioeconómicas y médicas particulares; la inclusión de estas covariables en el modelo es posible pero no está dentro del alcance de esta tesis. Aunque las covariables asociadas a los individuos inciden en la trayectoria del padecimiento, estas relaciones son un tema a explorarse en un trabajo de investigación futuro.

De este modo, podríamos decir que tenemos $i = 1, ..., I$ individuos, donde I es el número total de individuos observados por un período de tiempo

[1] Según Daley and Vere-Jones (2003) es el estudio de la secuencia de intervalos entre reemplazos sucesivos de un componente que es susceptible a fallar y es reemplazado por un nuevo componente cada vez que ocurre un fallo.

con costos asociados a su padecimiento. El objetivo es no solo modelar y predecir la duración y el costo de las etapas de estos padecimientos por individuo, sino también, introducir la estructura de dependencia entre éstas con un significado intuitivo. De acuerdo a Daley and Vere-Jones (2003), las relaciones de dependencia entre las variables pueden llegar a ser muy complejas, por lo que la especificación del modelo puede se puede tornar significativamente complicada.

Entonces, supongamos que empezamos el estudio de un individuo i en el tiempo $t_{i0} = 0$, es decir, este es el tiempo en el que el individuo i entra al panel de estudio. El momento final en el que el individuo i está en el estudio es T_i, es decir que el individuo permanece en el estudio en el intervalo $(t_{i0}, T_i]$, como horizonte calendario. Esto no quiere decir que no puedan ocurrir observaciones posteriores a T_i, sin embargo, éstas ya no serán consideradas parte del estudio; a esto se le conoce como datos con censura por la derecha.

Según Paik Schoenberg (2000), un proceso puntual N es una medida de probabilidad definida sobre un espacio métrico separable S tomando valores en los enteros no negativos Z^+. Con esta definición general se puede decir que la medida $N(A)$ representa el número de puntos que cae en el subconjunto A de S, ésto puede interpretarse para dimensiones mayores en espacios abstractos. Sin embargo, si tomamos el conjunto S como una dimensión temporal podemos entender el caso particular de $N(t)$ como un proceso de conteo del número de puntos que ocurren antes del tiempo t.

Sea $t_{ij} \in (t_{i0}, T_i]$ el momento en el que ocurre el j-ésimo cambio de tratamiento del i-ésimo individuo, por lo que definimos la variable aleatoria $N(t_{ij})$ como el proceso de conteo que cuenta el número de cortes o

cambios en el intervalo del individuo i antes del cambio j; de este modo se puede expresar $N(T_i)$ como el número de cambios o de ocurrencias en todo el horizonte calendario del individuo i. Esto es una generalización para cuando no se conoce la información a utilizar o cuando el número de cambios en la trayectoria de un paciente también es una variable aleatoria. Sin embargo, en la información utilizada para la implementación de este trabajo, se trabaja con tiempos discretos y se conocen todos los cortes en los intervalos de trayectorias individuales.

Dado que la muestra consiste en microcostos a través del tiempo de un individuo i, decimos que a cada t_{ij} se le asocia la variable costo de tratamiento; es decir, a cada momento en que ocurre un cambio de tratamiento le corresponde un nuevo precio de tratamiento p_{ij}. De este modo, para cualquier i tenemos una sucesión de variables asociadas $\{t_{i1}, p_{i1}\}, \{t_{i2}, p_{i2}\}, ..., \{t_{ik}, p_{ik}\}$, donde k indica la variable de cambio de diagnóstico. De este modo la sucesión de variables es una colección aleatoria de puntos en un espacio con una marca asociada a cada punto, así ya se pueden modelar los datos como en un proceso puntual marcado.

Daley and Vere-Jones (2003) definen el proceso puntual marcado como un proceso localizado en un espacio métrico completamente separable [2] χ y las marcas en otro espacio métrico completamente separado κ, entonces $\{(\chi_i, \kappa_i)\}$ en $\chi \times \kappa$ es un proceso puntual marcado con la propiedad adicional de que el proceso que se desarrolla en una dimensión temporal, $N(t)$ es a su vez un proceso puntual.

Desde un punto de vista práctico, para un solo individuo, deseamos

[2] De acuerdo con Schervish (2012) un espacio métrico completo χ se dice que es separable si existe un subconjunto numerable D tal que los elementos de D pueden usarse para aproximar cualquier elemento de χ.

caracterizar;

$$P(t_{i1}, ..., t_{in_i}, p_{i1}, ..., p_{in_i} | N(T_i) = n_i) \tag{2.1}$$

Es decir, la función de distribución conjunta del tiempo de ocurrencia de los eventos y los precios asociados condicionados por la variable aleatoria del número de eventos en el intervalo $(t_{i0}, T_i]$. Sin embargo, dado que al usar las variables en sus valores absolutos estas pueden dar saltos muy altos entre si, es recomendable el uso de variables alternas.

De acuerdo a Daley and Vere-Jones (2003), al tener un proceso puntual o de conteo $N(\cdot)$, se puede establecer una relación entre el proceso de conteo y los intervalos si definimos $t_j(N) = inf\{t : N(t) \geq j\}$, por lo tanto, se define el siguiente intervalo $\tau_j(N) = t_j(N) - t_{j-1}(N)$, donde $\sum_{j=1}^{\infty} \tau_i = \infty$. Esta definición es análoga a aquella que define t_j como el j-ésimo cambio de tratamiento para un solo individuo y los intervalos como las duraciones; el concepto de intervalos se hace extensivo para los costos de la siguiente manera para un solo individuo i:

- $D_{ij} = t_{ij} - t_{ij-1}$, donde D_{ij} es la duración entre los tiempos de ocurrencia de cada individuo.
- $C_{ij} = p_{ij} - p_{ij-1}$, donde C_{ij} representa el costo, es decir, la diferencia entre los precios en cada tiempo de ocurrencia de cada individuo.

.

Es importante mencionar, como lo hacen Daley and Vere-Jones (2003), que hay una correspondencia de uno a uno entre las distribuciones de probabilidad del proceso puntual y las distribuciones de los intervalos, es

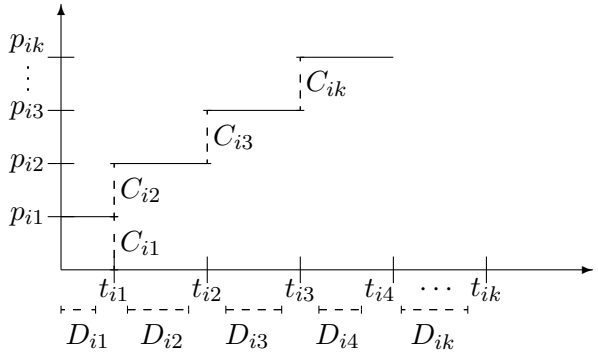

Figura 2.1: Trayectorias del individuo i.

por eso que el modelo inicial de probabilidad con las variables de tiempos y precios, les corresponde una medida de probabilidad distinta a aquella de las variables de duraciones y costos. De este modo,

$$Q(t_{i1}, ..., t_{in_i}, p_{i1}, ..., p_{in_i} | N(T_i) = n_i) \cong P(D_{i1}, ..., D_{in_i}, C_{i1}, ..., C_{in_i} | N(T_i) = n_i) \tag{2.2}$$

Esto quiere decir que calcular la función de distribución conjunta de los tiempos de ocurrencia y los precios asociados a éstos es análogo a a calcular la función de distribución conjunta de las duraciones y los costos asociados condicionados a la variable aleatoria del número de eventos en el intervalo de tiempo. Así pasamos de un proceso puntual marcado a uno de duración marcada.

De este modo, quedan definidas las variables aletorias del modelo que se pueden observar, siendo éstas D_{ik} para la duración k-ésima del i-ésimo individuo y C_{ik} para el costo k-ésimo del i-ésimo individuo y la distribución de probabilidad indicada para el mismo. Por otro lado, en la aplicación del modelo tenemos las variables aleatorias observables según

la información existente, las cuales tendrán la siguiente notación: d_{ik} como la k-ésima duración observada para el i-ésimo individuo y c_{ik} como el k-ésimo costo observado para el i-ésimo individuo. De igual modo, con las variables observadas sabemos que $N(T_i) = n_i$, éste valor es fijo pues ya fue observado y relativo a cada individuo.

Así pues, el modelo de probabilidad consiste en dos variables observables, duración y costo. La relación entre estas dos variables para un individuo i es una de las siguientes opciones,

Las variables son independientes entre sí, es decir,

$$\begin{aligned} P(d_{i1}, ..., d_{in_i}, c_{i1}, ..., c_{in_i} | N(T_i) = n_i) &= P(d_{i1}, ..., d_{in_i} | N(T_i) = n_i) \\ &\times \quad P(c_{i1}, ..., c_{in_i} | N(T_i) = n_i) \end{aligned}$$

O las variables de duración y costos no son independientes entre sí. En este caso, necesitamos determinar la estructura de dependencia entre las variables que puede expresarse de las siguientes dos formas,

Puede ser que las duraciones dependen de los costos, o bien, en el contexto del proceso puntual marcado, que los puntos dependen de las marcas. Es decir,

$$\begin{aligned} P((d_{i1}, c_{i1}), ..., (d_{in_i}, c_{in_i}) | N(T_i) = n_i) &= P(d_{i1}, ..., d_{in_i} | N(T_i) = n_i, c_{i1}, ..., c_{in_i}) \\ &\times \quad P(c_{i1}, ..., cin_i | N(T_i) = n_i) \end{aligned}$$

O bien, los costos dependen de las duraciones, esto quiere decir, que en el contexto del proceso puntual marcado, las marcas dependen de los puntos. Se observa una estructura de dependencia similar en el artículo

de Engle and Russell (1998), donde se observa tipo de estructura de dependencia en el contexto de transacciones financieras. Esta estructura de dependencia se puede expresar como,

$$\begin{aligned} P((d_{i1}, c_{i1}), ..., (d_{in_i}, c_{in_i})|N(T_i) = n_i) &= P(c_{i1}, ..., c_{ik}|N(T_i) = n_i, d_{i1}, ..., d_{in_i}) \\ &\times \quad P(d_{i1}, ..., d_{in_i}|N(T_i) = n_i) \end{aligned}$$

En este caso, los precios dependen del cambio de tratamiento, de manera análoga, la variable costo está asociada a la variable de duración. Es decir, que aunque la marca se localice en otro espacio métrico, ésta sigue anclada al proceso puntual primario, el que cuenta el número de ocurrencias en el espacio temporal.

Por la estructura de construcción del modelo este se puede pensar,para un solo momento k en el tiempo de un individuo i, como

$$P_{t,p}(t_{ik}, p_{ik}|N(t)) \cong P(d_{ik}, c_{ik}|N(t)) = P(c_{ik}|d_{ik})P(d_{ik}) \tag{2.3}$$

Este proceso es replicable para cada momento de la trayectora y cada individuo en la muestra.

2.3. Propieda de los Procesos de Duración Marcada

Una vez que hemos definido qué es el proceso de duración y de duración marcada y cómo es que los datos que tenemos para este estudio se adaptan a este modelo, necesitamos especificar las propiedades que van a hacer posible la inferencia y la predicción. Estas propiedades son la independencia, la intercambiabilidad y, principalmente, la estacionareidad.

Independencia

En una concepción tradicional, Resnick (1999) define la independencia de un número finito de eventos como:

Definicion 1. *Los eventos* X_1, ... , X_n *(n≥ 2) son independientes si*

$$P(\bigcap_{i\in I} X_i) = \prod_{i\in I} P(X_i), \qquad I \subset \{1, ..., n\}$$

Los eventos son independientes si la probabilidad de la intersección de estos eventos o la probabilidad conjunta de los eventos es igual a la multiplicación de la probabilidad de los mismos.

Análogamente, podemos hacer la definición de independencia para el proceso de duración marcada. Recordemos que tenemos la función de probabilidad conjunta de las duraciones y los costos, por lo que la independencia en el proceso es:

$$\begin{aligned} P(d_{i1}, c_{i1}, ..., d_{in_i}, c_{in_i} | N(T_i) = n_i) &= \prod_{j=1}^{N(T_i)} P(d_{ij}, c_{ij}) \\ &= \prod_{j=1}^{n_i} P(d_{ij}, c_{ij}) \end{aligned}$$

En este caso, la única diferencia reside en el hecho de que el número de funciones de probabilidad a multiplicar es a su vez una variable aleatoria, la cual se encarga de contar los cambios en el costo de tratamiento en el tiempo. En la implementación ya sabemos el valor observado de la variable aleatoria de conteo como n_i, como se puede ver en la segunda igualdad. El supuesto de independencia es útil para la inferencia de futuras observaciones.

Intercambiabilidad

Otra propiedad muy importante para la inferencia y predicción de variables en un proceso de duración marcada es la intercambiabilidad que, de acuerdo a Hahn and Zhang (2012), se define como:

Definicion 2. *Una sucesión de variables numerable o contable* $(X_j)_{j=1}^{\infty}$ *es intercambiable si para cada operador de permutación* $\sigma(\cdot)$ *de* $\{1, ..., n\}$, *para toda* $n < \infty$

$$P(X_1, X_2, ..., X_n) = P(X_{\sigma(1)}, X_{\sigma(2)}, ..., X_{\sigma(n)})$$

Si la sucesión de variables es independiente e idénticamente distribuida entonces es intercambiable. El concepto de intercambiabilidad está muy relacionado con la independencia, pues la independencia es un caso particular de la intercambiabilidad.

Para poder entender mejor la propiedad podemos citar el Teorema de Fenetti(1937) que nos dice que una sucesión infinita de variables aleatorias intercambiables $\bar{X} = (X_1, X_2, ...)$ es una mezcla de variables condicionalmente independientes e idénticamente distribuidas (i.i.d). Esto es, que existe un espacio de probabilidad (U, Θ) tal que

$$P(\bar{X} \in B) = \int_U P(\bar{X}(u) \in B)\Theta(du)$$

donde $\bar{X}(u) = (X_1(u), X_2(u), ...)$ es una secuencia de variables aleatorias i.i.d. y $\Theta(\cdot)$ es una medida de probabilidad.

Esto se puede adaptar al proceso de duración marcada correspondiente

a este análisis de la siguiente manera, tomando el Teorema de Fenetti

$$\begin{aligned} P(d_{i1}, c_{i1}, ..., d_{in_i}, c_{in_i} | N(T_i) = n_i) &= \int_\Theta \prod_{j=1}^{N(T_i)} P(d_{ij}, c_{ij}|\theta)\pi(\theta)d(\theta) \\ &= \int_\Theta \prod_{j=1}^{n_i} P(d_{ij}, c_{ij}|\theta)\pi(\theta)d(\theta) \end{aligned}$$

donde θ es una variable aleatoria no observable y $\pi(\theta)$ es una medida de probabilidad común a todas las variables aleatorias. Es decir, que a lo postulado en el apartado de independencia le agregamos la variable no observable con su respectiva medida de probabilidad, sobre cuyo espacio de probabilidad está definida la integral. La variable no observable común a todas las variables aleatorias es un tema que se posteriormente se desarrollará con mayor profundidad. De igual modo que en la contextualización de la propiedad de independencia, en este caso, conocemos el valor observado de $N(T_i)$ como n_i fijo y único para cada individuo.

Estacionareidad

Una vez que han sido definidas la independencia y la intercambiabilidad faltaría definir la estacionareidad para poder hacer predicciones sobre futuras observaciones, cuando el orden de las observaciones es relevante en el modelo.

De manera intuitiva, podemos definir la estacionareidad en un proceso de duración cuando la función de probabilidad conjunta del proceso no cambia cuando es ésta es desplazada en el tiempo, lo cual indicaría que lo importante es la longitud de los intervalos, no la localización de los mismos. Sin embargo, de una manera más técnica, Daley and Vere-Jones (2003) definen la estacionareidad en un proceso como:

Definicion 3. *Un proceso puntual es estacionario por intervalos cuando para cada* $r = 1, 2, ...$ *y todos los enteros* $i_i, ..., i_r$ *, la distribución conjunta de* $\{\tau_{i_{1+k}}, ..., \tau_{i_{r+k}}\}$ *no depende de* k *(*$k = 0, \pm 1, ...$*).*

Esto implicaría que el orden de las observaciones importa y que las observaciones pasadas ayudan a construir la variable aleatoria. Es decir que con una sucesión de variables de duraciones para un solo individuo $\bar{D} = (D_1, ..., D_n)$ tendríamos que,

$$\begin{aligned} P(D_n, ..., D_1) &= P(D_n|D_{n-1}, ..., D_1) \times P(D_{n-1}|D_{n-2}, ..., D_1) \times \cdots \\ &\times \quad P(D_2|D_1) \times P(D_1) \end{aligned}$$

Así si la variable aleatoria depende de su historia, podríamos entonces predecir observaciones futuras. Es decir, que para toda $s \geq 0$

$$\begin{aligned} P(D_{n+1}, D_n, ..., D_1) &= P(D_{n+1}|D_n, ..., D_1) \\ &= P(D_{n+s+1}|D_{n+s}, ..., D_{1+s}) \end{aligned}$$

De este modo, para el proceso de duración marcada la estacionareidad, junto con la noción desarrollada en la ecuación (3) sobre la relación de las marcas con el proceso de duración, se podría plantear, para un individuo

i, como

$$
\begin{aligned}
P(d_{i1}, c_{i1}, ..., d_{in_i}, c_{in_i} | N(T_i) = n_i) &= P(d_{i1}, c_{i1}) \prod_{j=2}^{N(T_i)} P(d_{ij}, c_{ij} | d_{ij-1}, c_{ij-1}) \\
&= P(c_{i1}|d_{i1}) P(d_{i1}) \times \\
&\times \prod_{j=2}^{N(T_i)} P(d_{ij}|d_{ij-1}) P(c_{ij}|d_{ij}, c_{ij-1}) \\
&= P(c_{i1}|d_{i1}) P(d_{i1}) \times \\
&\times \prod_{j=2}^{n_i} P(d_{ij}|d_{ij-1}) P(c_{ij}|d_{ij}, c_{ij-1})
\end{aligned}
$$

Lo que quiere decir que la función conjunta de probabilidad se puede definir con base a observaciones pasadas. De igual modo que con las propiedades pasadas, en la implementación conocemos el valor observado de $N(T_i)$ como n_i fijo y único para cada indiviudo.

Citando a Daley and Vere-Jones (2003) el proceso puntual marcado es estacionario si la estructura de probabilidad del proceso no cambia a pesar de los cambios que puedan existir en el espacio métrico χ, es decir, en el espacio métrico del proceso de conteo primario o el proceso de conteo que se desarrolla en la dimensión temporal. De acuerdo a lo desarrollado en la sección anterior se concluye que tanto el proceso primario como el Proceso Puntual Marcado, ambos son estacionarios, lo que permitiría la inferencia sobre el mismo.

Una vez que nuestro modelo de duración marcada cumple las propiedades descritas en esta sección podemos empezar a hacer inferencia sobre las variables y predecir las observaciones futuras. En la siguiente sección,

desarrollaremos un modelo complementario de variables latentes que terminaría de conectar la idea de la variable no observable presentada en el concepto de intercambiabilidad con el resto de la sucesión.

2.4. Construcción de Procesos vía Variables Latentes

Como está formulado al final de la sección pasada, el modelo general de probabilidad para un solo individuo i bajo estacionareidad con valor observado, único y fijo para la variable $N(T_i)$, lo escribimos como

$$P((d_{ij}, c_{ij})_{j=1}^{n_i}) = P(c_{i1}|d_{i1})P(d_{i1}) \prod_{j=2}^{n_i} P(d_{ij}|d_{ij-1})P(c_{ij}|d_{ij}, c_{ij-1})$$

Se puede observar en el modelo general que, gracias a la estacionareidad del mismo, es necesario conocer las distribuciones de una variable en un momento j condicionada a la misma variable en el momento $j-1$. Es decir, se necesitan conocer las distribuciones $f(d_{ij}|d_{ij-1})$ y $f(c_{ij}|c_{ij-1}, d_{ij})$. El método de para conocer estas distribuciones es análogo entre ambas, por lo que se desarrollará el proceso en la distribución de las duraciones para después extenderla a aquella de los costos.

En este modelo se tienen las variables aleatorias observables que son las duraciones y los costos, pues son las que se pueden obtener de la información disponible. Para lograr construir estas distribuciones se supone que no se conoce la relación entre las variables observables, en este caso las variables observables se refieren a las duraciones, a través del tiempo. Para efecto de la construcción del modelo se supone también que existe una variable no observable, o latente, que conecta a las observaciones.

2.4.1. Construcción del modelo estacionario tipo Gibbs

Tomando como base el procedimiento propuesto por Pitt et al. (2002), la distribución $f(d_{ij}|d_{ij-1})$ se puede reescribir como $f_{Y|Z}(y|z)$ siendo $Y = d_{ij}$ y $Z = d_{ij-1}$. Esta distribución a su vez se puede reescribir como

$$f_Y(y) = \int f_{Y|Z}(y|z) f_Y(z) dz$$

Para lograr la conexión entre las variables observables se necesita agregar otra variable, que le llamaremos latente, a esta distribución por lo que consideramos ahora las distribuciones de trancisión como

$$f_{Y|Z}(y|z) = \int f_{Y|\Theta}(y|\theta) f_{\Theta|Z}(\theta|z) d\lambda(\theta)$$

donde $\lambda(\theta)$ es una medida de probabilidad correspondiente a la variable latente.

Regresando a la notación original, la anterior distribución se reescribe como,

$$f_{D_{ij}|D_{ij-1}}(d_{ij}|d_{ij-1}) = \int f_{D_{ij}|\Theta}(d_{ij}|\theta) f_{\Theta|D_{ij-1}}(\theta|d_{ij-1}) d\lambda(\theta)$$

El punto crucial para asegurar la distribución de trancisión es construir una distribución conjunta $f_{D_{ij},\Theta}(d_{ij}, \theta)$, tal que las densidades condicionales sean $f_{D_{ij}|\Theta}(d_{ij}|\theta)$ y $f_{\Theta|D_{ij-1}}(\theta|d_{ij-1})$ y con una distribución marginal $f_{D_{ij}}(d_{ij})$. Es decir, el proceso de trancisión se logra mediante el proceso de la variable latente, que aunque no es observable, este es conocido.

2.4.2. Modelo Oculto de Markov

Según lo desarrollado en la sección anterior, vemos que para conocer la distribución de una observación condicionada a la observación anterior se necesita introducir una variable no observable o latente. La introducción de esta variable se puede hacer mediante el Modelo Oculto de Markov, que de acuerdo con las nociones expuestas por Ghahramani (2001), es un modelo donde la variable observable d_{ij}, en este caso la j-ésima duración del individuo i y la variable no observable θ_{ij} son independientes y la variable no observable cumple la propiedad markoviana. Es decir, que dado el valor de θ_{ij-1}, el estado actual θ_{ij} es independiente de todos aquellos estados previos a $j-1$. La función de distribución conjunta sería

$$P((d_{ij},\theta_{ij})_{j=1}^{n_i}) = P(d_{i1}|\theta_{i1})P(\theta_{i1})\prod_{j=2}^{n_i} P(d_{ij}|\theta_{ij})P(\theta_{ij}|\theta_{ij-1})$$

Es decir,

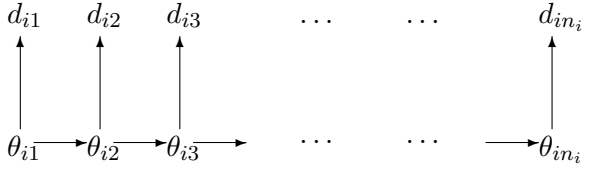

Figura 2.2: Representación gráfica de un Modelo Oculto de Markov.

El Modelo Oculto de Markov es un modelo muy flexible, según Ghahramani (2001), puede ser utilizado para modelar cualquier distribución con un número infinito de componentes y con las adecuaciones correspondientes se pueden modelar un sinnúmero de de problemas dinámicos no-lineales; esta misma flexibilidad es lo que le resta fiabilidad a la infe-

rencia. Aunado a esto, el modelo se basa en la relación existente entre los estados en el tiempo del parámetro, es decir, $P(\theta_{ij}|\theta_{ij-1})$, esta relación no es conocida en el problema de estudio del presente trabajo por lo que el Modelo Oculto de Markov no es el que mejor se adapta.

2.4.3. Modelos Dinámicos Lineales

Dado que el Modelo Oculto de Markov descrito en la sección anterior no es el adecuado para el caso de estudio de este trabajo, pero se necesita un modelo que integre variables latentes para explicar las relaciones entre las variables observables. ASí pues, se toma como base los Modelos Dinámicos Lineales descritos por Harrison and West (1999) que son parecidos en el planetamieno a los Modelos Ocultos de Markov. De este modo, el proceso de la variable latente para el caso de las duraciones de un individuo i, sería

Ecuación de Observación: $d_{ij} = \theta_{ij} + \epsilon_{ij}$

Ecuación de Sistema: $\theta_{ij} = \theta_{ij-1} + \nu_{ij}$

Donde los errores ϵ_{ij} y ν_{ij} son mutuamente independientes. El modelo dinámico lineal es análogo para las variables de duración y costos. Además de que no se conoce la distribución marginal de las observaciones.

Siguendo con el argumento propuesto por Pitt et al. (2002), necesitamos una variable latente sobre la cual se pueda hacer inferencia. Haciendo el símil con los modelos dinámicos lineales, incorporando la variable latente para la variable de duración,

Ecuación de Observación: $d_{ij}|\theta_{ij} \sim f_{d|\Theta}(\cdot|\theta_{ij})$

Ecuación de Sistema: $\theta_{ij}|d_{ij-1} \sim f_{\Theta|d}(\cdot|d_{ij-1})$

De este modo, podemos definir la probabilidad marginal como,

$$P(d_{ij}) = \int p(d_{ij}|d_{ij-1})p(d_{ij-1}) = \int p(d_{ij}|\theta_{ij})p(\theta_{ij}|d_{ij-1})d\theta_{ij}$$

Una vez que tenemos el modelo para la variable de las duraciones, lo hacemos extensivo para los costos. De este modo, el proceso de variables latentes para un individuo i quedaría,

Ecuación de Observación: $c_{ij}|\gamma_{ij} \sim f_{c|\gamma}(\cdot|\gamma_{ij})$

Ecuación de Sistema: $\gamma_{ij}|c_{ij-1} \sim f_{\gamma|c}(\cdot|c_{ij-1})$

De manera análoga, podemos definir la probabilidad marginal de los costos como,

$$P(c_{ij}) = \int p(c_{ij}|c_{ij-1}, d_{ij})p(c_{ij-1}) = \int p(c_{ij}|\gamma_{ij}, d_{ij})p(\gamma_{ij}|c_{ij-1}, d_{ij})d\gamma_{ij}$$

Dado que el Modelo Oculto de Markov no necesariamente es estacionario afectando así la capacidad inferencial del mismo, el modelo que se construye a través de las estructuras de dependencia con la variable latente que se acabande definir, es estacionario por construcción y es capaz de modelar la observación actual en relación a la observación anterior. Tomando solamente la variable duración del individuo i junto con su parámetro θ, la representación gráfica de ambos modelos se compara del siguiente modo,

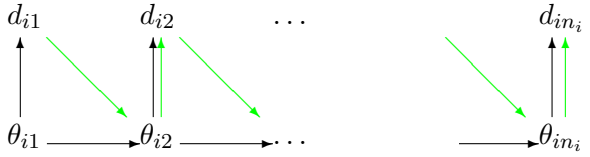

Figura 2.3: Comparación entre un Modelo Oculto de Markov y un modelo de variables latentes

Las flechas negras representan un Modelo Oculto de Markov, donde la relación con el estado anterior se presenta solamente en los parámetros $\theta_{ij}|\theta_{ij-1}$; mientras que las flechas verdes representan el modelo construido donde los observaciones están condicionadas al parámetro y el parámetro está condicionado por la observación anterior. La diferencia entre estos modelos reside, en su mayor parte, en que el modelo de variables latentes inducido por Pitt et al. (2002) es estacionario por construcción, además de que se pueden definir las distribuciones condicionales y marginales; mientras que en el modelo oculto de markov no.

Ahora bien, utilizando las relaciones entre las variables observables y latentes, el proceso de trancisión para una observación en el modelo general de probabilidad para un individuo i con procesos de variables latentes se escribe asegurando la estacionariedad desde la construcción,

$$\begin{aligned}
P(c_{ij}, d_{ij}|c_{ij-1}, d_{ij-1}) &= \int P(c_{ij}, d_{ij}|\gamma, \theta) P(\gamma, \theta|c_{ij-1}, d_{ij-1}) \\
&= \int\int F(c_{ij}, d_{ij}|\gamma, \theta)\pi(\gamma|d_{ij}, c_{ij-1})\pi(\theta|d_{ij-1}) \\
&= \int\int F(c_{ij}|\gamma, d_{ij})F(d_{ij}|\theta)\pi(\gamma|d_{ij}, c_{ij-1})\pi(\theta|d_{ij-1}) \\
&= \int F(c_{ij}|\gamma, d_{ij})\pi(\gamma|d_{ij}, c_{ij-1}) \int F(d_{ij}|\theta)\pi(\theta|d_{ij-1}) \\
&= P(c_{ij}|d_{ij}, c_{ij-1})P(d_{ij}|d_{ij-1})
\end{aligned}$$

De este modo, el modelo general de la probabilidad para un individuo i con la variable aleatoria $N(T_i)$ observada con valor n_i fijo y único para cada individuo, es

$$P((d_{ij}, c_{ij})_{j=1}^{n_i} | N(T_i) = n_i) \quad = \quad P(c_{i1}|d_{i1})P(d_{i1}) \prod_{j=2}^{n_i} P(c_{ij}|d_{ij}, c_{ij-1})P(d_{ij}|d_{ij-1})$$

Este es el modelo general de probabilidad del proceso puntual marcado de las duraciones y costos de las enfermedades crónico degenerativas. Al introducir las variables latentes se construyen funciones de trancisión más sólidas, tomando en cuenta los parámetros que influyen en el mismo proceso.

Como se ve en el modelo general de probabilidad, la estructura de dependencia con las variables latentes y como, aunque los valores de estas variables no son observables, son influenciados por los valores de las variables observables. Esta estructura de dependencia, según Gelman et al. (2014), puede denorminarse como un modelo jerárquico con resultados observables condicionados a ciertos parámetros, los cuales, a su vez, están dados como variables aleatorias definidos a su vez con otros parámetros. Es decir, para un individuo i,

.

.

.

Es importante notar que la distribución de las variables latentes es arbitrario, por lo que con este modelo general de probabilidad resta de-

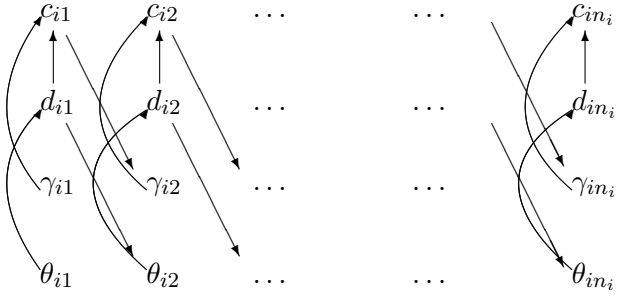

Figura 2.4: Representación gráfica de la estructura jerárquica del modelo general de probabilidad.

terminar las distribuciones que mejor describan las características de la relación entre las variables latentes. En la siguiente sección se relizará una comparación entre las distintas distribuciones que trabajarían mejor con las características particulares de los datos para su modelado.

2.5. Construcción del Modelo de Marcas

Con el modelo general de probabilidad formulado en la sección anterior se sientan las bases para hacer inferencia sobre duración y costos de padecimientos crónico degenerativos. Igualmente, la determinación de las distribuciones de las variables latentes son cruciales para la predicción sobre futuras observaciones.

Según Fader and Hardie (2013) la distribución Gamma es muy útil para modelar gastos en el campo actuarial, particularmente los siniestros. Esto es debido a que esta distribución tiene su soporte en los reales positivos, el sesgo a la derecha y las propiedades de las convoluciones que siguen los patrones de gasto si los siniestros se distribuyen Gam-

ma. Aunado a esto, en una óptica bayesiana, la distribución Gamma tiene muchas distribuciones conjugadas como la distribución Poisson, la distribución exponencial, la distribución normal con media conocida, la distribución Pareto entre otras; que ayudan a la linealidad del modelo general de probabilidad.

Las duraciones se modelan mediante un modelo Gamma-Gamma, mientras que los costos con un modelo Gamma Inversa-Weibull. El modelo Gamma-Gamma de las duraciones se deriva del propio proceso de conteo que es una distribución conjugada.

$$P(d_{ij}|d_{ij-1}) = \int P(d_{ij}|\theta_{ij})P(\theta_{ij}|d_{ij-1})d\theta_{ij} \quad i \neq 1$$

donde,

$$P(d|\theta) \sim Gamma(d|\alpha_d, \theta)$$
$$P(\theta) \sim Gamma(\theta|\alpha_\theta, \beta_\theta)$$

De manera análoga, para los costos el modelo es Gamma Inversa-Weibull, la distribución Weibull es usada frecuentemente en el campo actuarial para modelar los siniestros, esto es porque al ser una distribución relacionada con la distribución exponencial, es útil también en la modelación de valores extremos. Por estas características, la distribución Weibull es la que modelará los costos en este trabajo, así se asegura que si alguno de los costos del modelo llegara a diferir mucho del resto, el modelo tenga la flexibilidad para incluir los datos. Tomando en cuenta que los costos se modelan con la distribución Weibull, por conveniencia, la variable

latente se distribuye Gamma Inversa.

$$P(c_{ij}|c_{ij-1}) = \int P(c_{ij}|\gamma_{ij}, d_{ij})P(\gamma_{ij}|d_{ij}, c_{ij-1})d\gamma_{ij} \quad i \neq 1$$

donde,

$$P(c|\gamma, d) \sim Weibull(c|d, \gamma)$$
$$P(\gamma) \sim Inv - Gamma(\gamma|\alpha_\gamma, \beta_\gamma)$$

Con el modelo general de probabilidad y las distribuciones designadas ya es posible hacer inferencia sobre observaciones futuras, como se demostrará en el siguiente capítulo.

2.6. Discusión

Los beneficios que presentan los modelos de procesos puntuales marcados con variables latentes para la estimación de costos totales de enfermedades crónico-degenerativas son la flexibilidad que brindan para modelar las estructuras de dependencia entre las marcas y el proceso puntual. Es decir, que es un modelo construido de manera específica para ajustarse a las trayectorias de cada paciente con un padecimiento crónico degenerativo.

La especificidad del modelo aunado a la estacionareidad asegurada por construcción, se puede asegurar la inferencia de observaciones futuras. Aunque fuera del alcance de este proyecto de investigación, la inferencia es una herramienta poderosa para un estudio riguroso de tarificación en el sector asegurador para seguros de gastos médicos mayores. La verosimilitud del modelo y la manera de hacer inferencia se desarrollarán

a profundidad en el siguiente capítulo, de acuerdo a las distribuciones designadas.

Capítulo 3

Inferencia y Predicción

3.1. Introducción

En el capítulo anterior se describió el modelo general de probabilidad que describe el proceso de duraciones y costos en un padecimiento crónico degenerativo, por lo que el siguiente paso sería hacer inferencia sobre el mismo. En este tipo de modelos, el objetivo de hacer inferencia es predecir futuras observaciones en base a los datos ya observados. En este capítulo se sentarán las bases para realizar esta inferencia.

El primer paso para realizar inferencia es la construcción de la función de verosimilitud, en este caso extendida a las variables latentes y a los parámetros correspondientes a sus distribuciones de todos los individuos de la población. Una vez que se determinaron las funciones de verosimilitud, se analizan los métodos de estimación que se podrían usar para la predicción de futuras observaciones.

3.2. Verosimilitud Extendida

Como especificado en la sección anterior, una vez que el modelo de probabilidad describe de manera precisa los datos del problema podemos empezar a hacer inferencia sobre observaciones futuras. La base sobre la que se puede hacer predicción en base a los datos ya observados es la función de verosimilitud definida como la función de de distribución conjunta de los datos.

Así, para un modelo de probabilidad para un individuo i como el expuesto en el capítulo anterior, donde los datos de duraciones con su respectiva variable latente se se caracterizan mediante $p(d_{ij}|\theta)$; de acuerdo a Held and Sabanés Bové (2014), la función de verosimilitud $V(\theta)$ se define como la función masa o la función de densidad de los datos observados d_i, entendidos en función del parámetro desconocido o latente θ.

Es decir, que en este caso las variables observables se definen en función de las variables latentes, estas a su vez se describen en función de sus parámetros. De esto se desprende la noción de verosimilitud extendida para incluir las variables latentes. Como se especifica en Pitt et al. (2002), la construcción de la función de verosimilitud resulta sencilla, incluso intuitiva. Sin embargo, la estimación de los parámetros mediante máxima verosimilitud no es tan sencilla pues no tiene una solución que se pueda expresar de manera analítica cerrada. Usando esta construcción, se escribe una función de verosimilitud para el modelo general de probabilidad de duraciones y costos de un solo individuo

$$
\begin{aligned}
V(\{\theta_j\},\{\gamma_j\},\{d_j\},\{c_j\}) &= f(d_1|\theta_1)f(c_1|d_1,\gamma_1)f(\gamma_1)f(\theta_1)\times \\
&\times \prod_{j=2}^{N(t)} f(d_j|\theta_j)f(\theta_j|d_{j-1})f(c_j|d_j,\gamma_j)f(\gamma_j|c_{j-1})
\end{aligned}
$$

Para poder calcular la función de verosimilitud que permite hacer inferencia, es necesario conocer las distribuciones de las variables latentes con base en las observaciones anteriores para ambas variables observables, duraciones y costos.

Para la primera variable observable se toma en cuenta la relación

$$f_{\theta|d}(\theta|d) \propto f_{d|\theta}(d|\theta)f(\theta)$$

y que $d|\theta \sim Gamma(d|\alpha_d,\theta)$ y $\theta \sim Gamma(\theta|\alpha_\theta,\beta_\theta)$.

$$
\begin{aligned}
f_{\theta|d}(\theta|d) &\propto \frac{\theta^{\alpha_d}}{\Gamma(\alpha_d)} \quad d^{\alpha_d-1}e^{-\{\theta d\}} \times \frac{\beta_\theta^{\alpha_\theta}}{\Gamma(\alpha_\theta)} \quad \theta^{\alpha_\theta-1}e^{-\{-\beta_\theta\theta\}} \\
&= \frac{\beta_\theta^{\alpha_\theta}}{\Gamma(\alpha_d)\Gamma(\alpha_\theta)} \quad d^{\alpha_d-1} \quad \theta^{\alpha_d+\alpha_\theta-1} \quad e^{\{-\theta(d+\beta_\theta)\}} \\
&\propto \theta^{\alpha_d+\alpha_\theta-1} \quad e^{\{-\theta(d+\beta_\theta)\}} \\
&\Rightarrow \theta|d \sim Gamma(\alpha_d+\alpha_\theta, d+\beta_\theta)
\end{aligned}
$$

Para la variable de duraciones, la distribución de la variable latente que depende de la observación se puede expresar de una manera analítica cerrada como la distribución Gamma. Análogamente, para la variable de

costos se vuelve a tomar en cuenta la misma relación y las distribuciones

$$c|d,\gamma \sim Weibull(c|d,\gamma) \qquad \gamma \sim InvGamma(\gamma|\alpha_\gamma,\beta_\gamma)$$

$$\begin{aligned} f_{\gamma|d,c}(\gamma|d,c) &\propto \frac{d}{\gamma^d} \quad c^{d-1}e^{\{-(\frac{c}{\gamma})^d\}} \times \frac{\beta_\gamma^{\alpha_\gamma}}{\Gamma(\alpha_\gamma)} \quad (\frac{1}{\gamma})^{\alpha_\gamma+1} \quad e^{\{-(\frac{\beta_\gamma}{\gamma})\}} \\ &= \frac{d\beta_\gamma^{\alpha_\gamma}c^{d-1}}{\Gamma(\alpha_\gamma)} \quad (\frac{1}{\gamma})^{d+\alpha_\gamma+1} \quad e^{-((\frac{\beta_\gamma}{\gamma})+(\frac{c}{\gamma})^d)} \\ &\propto \left(\frac{1}{\gamma}\right)^{d+\alpha_\gamma+1} \quad e^{-((\frac{\beta_\gamma}{\gamma})+(\frac{c}{\gamma})^d)} \end{aligned}$$

Para la variable de costos, la distribución de la variable latente según la observación anterior no tiene una forma analítica cerrada como distribución, sin embargo, el kernel se puede simular con un slice sampler; este método se explicará con detalle en el apéndice.

Una vez que queda definidas las distribuciones de las variables latentes con base en las observaciones anteriores es importante notar que los párametros de las distribuciones $(\alpha_d, \alpha_\theta, \beta_\theta, \alpha_\gamma, \beta_\gamma)$, por constucción, no dependen de la realización; por lo que al modelo jerárquico establecido en la Figura 2.4 se agrega otro nivel. En la siguiente figura se muestra una representación del nuevo modelo jerárquico, donde los parámetros definen a las variables latentes y éstas a su vez, mediante las relaciones que ya establecimos, definen a las observaciones. Esta figura representa la estructura jerárquica del modelo de probabilidad del individuo i.

Es importante mencionar que las distribuciones desarrolladas de duraciones y costos, además de aquellas correspondientes a las variables latentes y a los parámetros, que no tienen forma analítica cerrada son

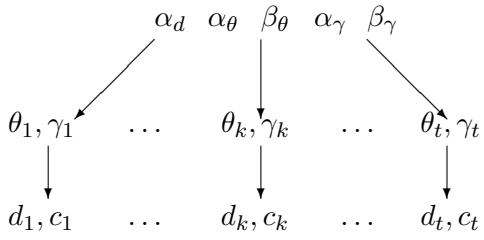

Figura 3.1: Modelo jerárquico de los parámetros, variables latentes y observaciones del individuo i.

funciones log-cóncavas. Esta propiedad se explorará con mayor detalle en la siguiente sección.

3.3. Log-concavidad en las funciones de distribución

Las funciones log-cóncavas son, de acuerdo con Bagnoli and Bergstrom (2005), funciones que se grafican con una curva cóncava en los números reales positivos y cuyo logaritmo es también una función cóncava; o bien, según An (1996) un vector aleatorio se distribuye de manera log-cóncava si es logaritmo de la distribución de densidad es cóncavo en su soporte. Es decir, que el vector de variables aleatorias de las duraciones del individuo i, D_i, está distribuido de manera log-cóncava si para cada D_{i1}, D_{i2} en el espacio de los reales positivos y cualquiera $\lambda \in [0, 1]$,

$$f(\lambda D_{i1} + (1 - \lambda) D_{i2}) \geq [f(D_{i1})]^{\lambda} [f(D_{i2})]^{1-\lambda}$$

Esta misma propiedad se hace extensiva para las variables de costos y los parámetros de las distribuciones.

De acuerdo a lo descrito en el capítulo anterior referente a las duraciones siguiendo un modelo Gamma- Gamma y los costos un modelo Gamma Inversa-Weibull, según Bagnoli and Bergstrom (2005) y An (1996), la distribución Weibull cumple con las características de log-concavidad si su parámetro de forma es mayor o igual a uno, para el caso de la distribución de los costos este parámetro corresponde a la duración, que por definición es mayor a uno. También la distribución Gamma de $\alpha_d, \alpha_\theta, \alpha_\gamma$ debe tener el parámetro de forma mayor a uno, como sucede para la distribución Weibull, esta condición se cumple por construcción. Una de las propiedades de las distribuciones log-cóncavas es que cualquier transformación lineal de una variable aleatoria no afecta su propiedad de log-concavidad. En el contexto de este trabajo de investigación, es importante la proposición que estipula An (1996) para las distribuciones multivariadas, donde si las variables aleatorias son independientes y con distribuciones log-cóncavas, la densidad conjunta también es log-cóncava.

Esto quiere decir que si podemos demostrar log-concavidad para una de las distribuciones, esta propiedad se extiende a la distribución conjunta. De este modo aseguramos la log-concavidad para estas distribuciones que no tienen una fórmula analítica cerrada y que se muestrearan de manera previa a la estimación de parámetros para el modelo general de probabilidad.

3.4. Inferencia en modelos con variables latentes.

Una vez que se especifican las distribuciones de los parámetros a estimar y las propiedades de los mismos, se necesitan métodos que los

puedan estimar. Pitt et al. (2002) especifica que la estimación de máxima verosimilitud puede resolverse mediante el algoritmo EM, aunque también, debido a que las densidades son dos condicionales de la densidad conjunta puede ligarse con el Muestreador Gibbs.

El algoritmo EM es un algoritmo para calcular el estimador de máxima verosimilitud o el EML mediante iteraciones, cada iteración consiste en un paso donde se calcula la esperanza y en otro se maximiza la misma, de ahí el nombre de EM. Este algoritmo se explica con más detalle en el apéndice II.

A pesar de reconocer la utilidad del algoritmo EM, el método numérico que se utilizará para la estimación en este trabajo será el Muestreador de Gibbs, el cual se explicará con más detalle en el apéndice. Debido a que el modelo general de probabilidad ha sido construido mediante el concepto bayesiano de las distribuciones conjugadas.

Capítulo 4

Inferencia Bayesiana

4.1. Introducción

En los capítulos anteriores se especificaron el modelo general de probabilidad y las distribuciones de las variables sobre las que se busca hacer inferencia. Existen varios métodos de estimación que se podrían utilizar en la inferencia de este trabajo de investigación, en este capítulo se pretende desarrollar la justificación para la utilización de un enfoque bayesiano para la estimación.

En la primera sección se da una breve introducción al paradigma bayesiano en general, para después desarrollar sobre el muestreador de Gibbs como método de estimación para los parámetros y variables latentes. De este modo se puede completar la parte teórica de este trabajo de investigación.

4.2. Paradigma Bayesiano

El paradigma bayesiano se refiere a una manera de hacer inferencia basado en el trabajo del inglés Thomas Bayes. En este paradigma se establece que la hipótesis se va actualizando de acuerdo a la nueva información relevante. Según Gelman et al. (2014), una de las principales razones para el pensamiento bayesiano es que facilita la interpretación basada en el sentido común de conclusiones estadísticas.

De acuerdo con Gelman et al. (2014) la inferencia bayesiana se hace en base en una evaluación retrospectiva del procedimiento utilizado para estimar el parámetro sobre la distribución de todas las posibles observaciones. Es decir, que mediante la regla de Bayes se describe la relación entre la asiganción previa de la probabilidad y la reasignación de esta misma condicionada a los datos observados. El paradigma está basado en la regla o teorema de Bayes.

Una definición de la probabilidad condicional de y dado x será la división de la función conjunta de probabilidad entre la función de probabilidad de x. Es decir,

$$p(y|x) = \frac{p(x,y)}{p(x)}$$

O en otras palabras, la probabilidad de y *suceda* dado x es la probabilidad de que *sucedan* ambos eventos relativo a que x *suceda* en absoluto.

Tomando en cuenta esta definición de la probabilidad condicional, el teorema de Bayes, se define como

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

El Teorema de Bayes resulta muy útil cuando el modelo de probabilidad se basa en variables observables y variables latentes, D y θ respectivamente. De este modo, el modelo se escribe como

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

Donde los elementos de esta ecuación significan

- $p(\theta|D)$ = la distribución posterior, es decir, la probabilidad de θ tomando en cuenta las observaciones.
- $p(D|\theta)$ = la verosimilitud, es decir, la probabilidad de los datos generados por el modelo con el parámetro θ.
- $p(\theta)$ = la distribución previa, es decir, la probabilidad de θ sin tomar en cuenta las observaciones D.
- $p(D)$ = la distribución de las observaciones, es decir, la probabilidad total de las observaciones, ponderadas por todos los valores que puede tomar el parámetro de acuerdo al peso que se le asigna.

Esta misma ecuación se puede reescribir como,

$$\begin{aligned} p(\theta|D) &= \frac{p(D|\theta)p(\theta)}{p(D)} \\ &\propto p(D|\theta)p(\theta) \\ &\propto \textit{verosimilitud} \times \textit{inicial}. \end{aligned}$$

La distribución posterior está en función del parámetro, por lo que la distribución posterior es proporcional a la multiplicación de la función de verosimilitud por la distribución inicial del parámetro. En otras palabras, la distribución previa del parámetro es la información a priori del mismo, que se va actualizando con las observaciones, tomando la información relevante al parámetro. Este mismo principio puede extenderse

para varios parámetros.

Como describe Gelman et al. (2014), esta lógica es similar para hacer inferencias sobre futuras observaciones. Una vez que se tienen todas las observaciones $D = (d_1, ..., d_n)$ se quiere inferir la siguiente observación d_{n+1}. La distribución de esta observación se llama la distribución posterior predictiva, posterior porque toma la información de las observaciones pasadas y predictiva porque predice la siguiente observación,

$$\begin{aligned} p(d_{n+1}|D) &= \int p(d_{n+1}, \theta|D)d\theta \\ &= \int p(d_{n+1}|\theta, D)p(\theta|D)d\theta \\ &= \int p(d_{n+1}|\theta)p(\theta|D)d\theta. \end{aligned}$$

En la segunda y tercera línea de la ecuación se muestra la distribución posterior predictiva como un promedio de las distribuciones predictivas condicionales de la distribución posterior del parámetro θ. En la última línea se asume la independencia condicional de D y d_{n+1} dado θ.

Tomando estas mismas ideas en un conexto más general, como el planteado por Goldstein (2013), donde la incertidumbre espistémica se expresa a través la distribuión previa del parámetro y la incertidumbre aleatoria como la función de verosimilitud dada por las observaciones. Por lo que se podría decir que la distribución posterior, es una mezcla entre ambas incertidumbres. En el análisis estadístico, se necesitan construir modelos basados en los valores desconocidos de la distribución de las observaciones, es decir, de incertidumbre espistémica pero tomando valores de un modelo paramétrico o de incertidumbre aleatoria; esto se logra mediante el concepto de intercambiabilidad.

Como fue definido en el capítulo dos, a través del Teorema de Finetti; la intercambiabilidad es una propiedad de la cual se deriva que sin importar la reordenación de las observaciones, éstas tienen la misma probabilidad de ocurrir. Las implicaciones del teorema de intercambiabilidad, de acuerdo con Goldstein (2013), son sorprendentemente simples. Debido a la simetría con la que se aplica el concepto de intercambiabilidad sobre las observaciones, es como si se estuviera muestreando independientemente del modelo de los valores desconocidos de las observaciones con una distribución previa sobre el parámetro, retomando los conceptos de incertidumbre epistémica y aleatoria. De este modo, mediante la observación de las variables observables y la aplicación del paradigma bayesiano, se reduce la incertidumbre al actualizar la información sobre las variables y facilitando la inferencia.

Otra de las mayores implicaciones del teorema de intercambiabilidad es muy intuitivo, pues la distribución de las observaciones no es mas que el resultado de todas las posibles observaciones futuras; y la división entre los componentes epistémicos y aleatorios de incertidumbre de esta estructura es nuestro propio juicio sobre dichas observaciones futuras. De igual modo, esto da una entrada natural a la inferencia.

Con el enfoque del análisis bayesiano los problemas clásicos de inferencia como la estimación puntual, estimación por regiones y contraste de hipótesis pueden resolverse de esta manera. Además, los estimadores obtenidos no solo suelen coincidir con los estimadores clásicos en algunos casos, sino en otros casos de hecho los mejoran.

En el desarrollo de este trabajo de investigación, veremos que la estimación del modelo general de probabilidad no se puede realizar a través

de métodos analíticos cerrados, por lo que se sugiere la utilización de métodos numéricos. El método númerico a utilizar, por la naturaleza del estudio, será el Muestreador de Gibbs, que se describe con más detalle en el apéndice.

Como mencionado en el capítulo anterior, la estimación de parámetros del modelo general de probabilidad que se utilizará en este trabajo no se puede hacer a través de métodos analíticos tradicionales, por lo que se utilizarán métodos numéricos. Existen varios métodos, entre estos se encuentran el algoritmo EM y el Muestreador de Gibbs, el algoritmo EM será más desarrolado en la sección anterior y aunque útil para el análisis del modelo presentado, el Muestreador de Gibbs tiene una interpretación más simple.

De acuerdo con Gelman et al. (2014), la simulación a través de las cadenas de Markov también llamadas Cadenas de Markov vía simulación Monte Carlo (MCMC, por sus siglas en inglés) consiste en construir una Cadena de Markov cuya distribución estacionaria (límite) sea una distribución de la cual se pretenda simular. Una de estas simulaciones MCMC es el Muestreador de Gibbs.

De este modo, se establece no solo las bases para la inferencia y predicción de futuras observaciones en base a la verosimilud extendida y la resolución de su función sino también la implementación numérica de la misma. Una vez que se definieron estas herramientas, lo que resta es la adaptación del modelo a la ilustración con los datos, esto se explorará más adelante.

4.3. Desarrollo del algoritmo para el modelo de duración marcada

Una vez que se han enunciado las nociones del paradigma bayesiano de manera general, se necesitan desarrollar las distribuciones particulares al modelo de duración marcada. En el capítulo anterior se definieron las distribuciones de las variables latentes condicionadas a las observaciones. Así, la distribución posterior de la variable latente θ condicionada a las duraciones observadas es,

$$f_{\theta|d}(\theta|d) \propto \theta^{\alpha_d+\alpha_\theta-1} e^{\{-\theta(d+\beta_\theta)\}}$$

De igual modo, la distribución de la variable latente γ condicionada a las duraciones y costos observados es,

$$f_{\gamma|d,c}(\gamma|d,c) \propto \left(\frac{1}{\gamma}\right)^{d+\alpha_\gamma+1} \quad e^{-\{(\frac{\beta_\gamma}{\gamma})+(\frac{c}{\gamma})^d\}}$$

De este modo, las distribuciones que resultan de los cálculos anteriores son la llave que se necesita para empezar a hacer inferencia, tomando estas distribuciones se redefine la verosimilitud de un solo individuo i como,

$$\begin{aligned}
&V(\{\alpha_d,\alpha_\theta,\beta_\theta,\alpha_\gamma,\beta_\gamma\},\{\theta_i,\gamma_i\}_{j=1}^{N(T_i)}|\{d_j,c_j\}_{i=1}^{I}) = \\
&\prod_{j=1}^{N(T_i)} Gamma(d_j|\alpha_d,\theta_j)Gamma(\theta_j|\alpha_d+\alpha_\theta,d_{j-1}+\beta_\theta) \times \\
&Weibull(c_j|d_j,\gamma_j)\left(\frac{1}{\gamma}\right)^{d_{ij-1}+\alpha_\gamma+1} e^{\{-(\frac{\beta_\gamma}{\gamma}+(\frac{c_{j-1}}{\gamma})^{d_{ij}})\}} \times \\
&\times\pi(\alpha_d)\pi(\alpha_\theta)\pi(\beta_\theta)\pi(\alpha_\gamma)\pi(\beta_\gamma)
\end{aligned}$$

Donde para las variables latentes asociada a las duraciones,

$$\begin{aligned}
\pi(\theta_{ij}|\alpha_d,\alpha_\theta,\beta_\theta) &\propto Gamma(d_{ij}|\alpha_d,\theta_{ij}) \times Gamma(\theta_{ij}|\alpha_d+\alpha_\theta, d_{ij-1}+\beta_\theta) \\
&= \frac{\theta_{ij}^{\alpha_d}}{\Gamma(\alpha_d)} d_{ij}^{\alpha_d-1} e^{\{-\theta_{ij}d_{ij}\}} \frac{(d_{ij-1}+\beta_\theta)^{\alpha_d+\alpha_\theta}}{\Gamma(\alpha_d+\alpha_\theta)} \theta_{ij}^{\alpha_d+\alpha_\theta} e^{\{d_{ij-1}+\beta_\theta\}} \\
&= \frac{d_{ij}^{\alpha_d-1}(d_{ij-1}+\beta_\theta)^{\alpha_d+\alpha_\theta}}{\Gamma(\alpha_d)\Gamma(\alpha_d+\alpha_\theta)} \theta_{ij}^{2\alpha_d+\alpha_\theta-1} e^{\{-\theta_{ij}(d_{ij-1}+d_{ij}+\beta_\theta)\}} \\
&\propto \theta_{ij}^{2\alpha_d+\alpha_\theta-1} e^{\{-\theta_{ij}(d_{ij-1}+d_{ij}+\beta_\theta)\}} \\
&\Rightarrow \theta_{ij} \sim Gamma(2\alpha_d+\alpha_\theta, d_{ij}+d_{ij-1}+\beta_\theta)
\end{aligned}$$

Y para la variable latente asociada a los costos,

$$\begin{aligned}
\pi(\gamma_{ij}|\alpha_\gamma,\beta_\gamma,d_{ij},c_{ij-1}) &\propto Weibull(c_{ij}|d_{ij},\gamma_{ij}) \times (\frac{1}{\gamma_{ij}})^{d_{ij}+\alpha_\gamma+1} e^{\{-(\frac{\beta_\gamma}{\gamma_{ij}}+(\frac{c_{ij-1}}{\gamma_{ij}})^{d_{ij}}\}} \\
&= \frac{d_{ij}}{\gamma_{ij}^{d_{ij}}} c_{ij}^{d_{ij-1}} e^{\{-(\frac{c_{ij}}{\gamma_{ij}})\}} (\frac{1}{\gamma_{ij}})^{d_{ij}+\alpha_\gamma+1} e^{\{-(\frac{\beta_\gamma}{\gamma_{ij}}+(\frac{c_{ij-1}}{\gamma_{ij}})^{d_{ij}})\}} \\
&= d_{ij} c_{ij}^{d_{ij}-1} (\frac{1}{\gamma_{ij}})^{2d_{ij}+\alpha_\gamma+1} e^{\{-(\frac{\beta_\gamma}{\gamma_{ij}}+(\frac{c_{ij-1}}{\gamma_{ij}})^{d_{ij}}+(\frac{c_{ij}}{\gamma_{ij}})^{d_{ij}})\}} \\
&\propto \left(\frac{1}{\gamma_j}\right)^{2d_j+\alpha_\gamma+1} e^{\{-(\frac{\beta_\gamma}{\gamma_j}+(\frac{c_{ij-1}+c_{ij}}{\gamma_{ij}})^{d_{ij}}\}}
\end{aligned}$$

De manera análoga, las distribuciones que corresponden a los parámetros para un solo individuo i. En primer lugar tenemos las distribuciones relativas a los parámetros del modelo Gamma-Gamma correspondiente

a la variable observable de las duraciones $\alpha_d, \alpha_\theta, \beta_\theta$,

$$\begin{aligned}
\pi(\alpha_d|...) &\propto \prod_{j=1}^{N(T_i)} Gamma(\theta_{ij}|\alpha_d+\alpha_\theta, d_{ij-1}+\beta_\theta) \times Gamma(\alpha_d|\alpha_0,\beta_0) \\
&= \prod_{j=1}^{N(T_i)} \frac{(d_{ij-1}+\beta_\theta)^{\alpha_d+\alpha_\theta}}{\Gamma(\alpha_d+\alpha_\theta)} \theta_{ij}^{\alpha_d+\alpha_\theta-1} e^{\{-\theta_{ij}(d_{ij-1}+\beta_\theta)\}} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \alpha_d^{\alpha_0-1} e^{\{-\alpha_d\beta_0\}} \\
&\propto \prod_{j=1}^{N(T_i)} \frac{(d_{ij-1}+\beta_\theta)^{\alpha_d+\alpha_\theta}}{\Gamma(\alpha_d+\alpha_\theta)} \theta_{ij}^{\alpha_d} \alpha_d^{\alpha_0-1} e^{\{-\alpha_d\beta_0\}}
\end{aligned}$$

$$\begin{aligned}
\pi(\alpha_\theta|...) &\propto \prod_{j=1}^{N(T_i)} Gamma(\theta_{ij}|\alpha_d+\alpha_\theta, d_{ij-1}+\beta_\theta) \times Gamma(\alpha_\theta|\alpha_0,\beta_0) \\
&= \prod_{j=1}^{N(T_i)} \frac{(d_{ij-1}+\beta_\theta)^{\alpha_d+\alpha_\theta}}{\Gamma(\alpha_d+\alpha_\theta)} \theta_{ij}^{\alpha_d+\alpha_\theta-1} e^{\{-\theta_{ij}(d_{ij-1}+\beta_\theta)\}} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \alpha_\theta^{\alpha_0-1} e^{\{-\alpha_\theta\beta_0\}} \\
&\propto \prod_{j=1}^{N(T_i)} \frac{(d_{ij-1}+\beta_\theta)^{\alpha_d+\alpha_\theta}}{\Gamma(\alpha_d+\alpha_\theta)} \theta_{ij}^{\alpha_\theta} \alpha_\theta^{\alpha_0-1} e^{\{-\alpha_\theta\beta_0\}}
\end{aligned}$$

$$
\begin{aligned}
\pi(\beta_\theta|...) &\propto \prod_{j=1}^{N(T_i)} Gamma(\theta_{ij}|\alpha_d+\alpha_\theta, d_{ij-1}+\beta_\theta) \times Gamma(\beta_\theta|\alpha_0,\beta_0) \\
&= \prod_{j=2}^{N(T_i)} \frac{(d_{ij-1}+\beta_\theta)^{\alpha_d+\alpha_\theta}}{\Gamma(\alpha_d+\alpha_\theta)} \theta_{ij}^{\alpha_d+\alpha_\theta-1} e^{\{-\theta_{ij}(d_{ij-1}+\beta_\theta)\}} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \beta_\theta^{\alpha_0-1} e^{\{-\beta_\theta\beta_0\}} \\
&\propto \prod_{j=2}^{N(T_i)} (d_{ij-1}+\beta_\theta)^{\alpha_d+\alpha_\theta} \beta_\theta^{\alpha_0-1} e^{\{-\beta_\theta(\theta_{ij}+\beta_0)\}}
\end{aligned}
$$

De igual modo, tenemos las distribuciones para los parámetros del modelo Gamma Inversa-Weibull que describe el comportamiento de los costos a través de la variable latente γ, $\alpha_\gamma, \beta_\gamma$,

$$
\begin{aligned}
\pi(\alpha_\gamma|...) &\propto \prod_{j=1}^{N(T_i)} \pi(\gamma_{ij}|\alpha_\gamma,\beta_\gamma,d_{ij},c_{ij-1}) \times Gamma(\alpha_\gamma|\alpha_0,\beta_0) \\
&= \prod_{j=1}^{N(T_i)} \left(\frac{1}{\gamma_{ij}}\right)^{d_{ij}+\alpha_\gamma+1} e^{\{-(\frac{\beta_\gamma}{\gamma_{ij}}+(\frac{c_{ij-1}}{\gamma_{ij}})^{d_{ij}})\}} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \alpha_\gamma^{\alpha_0-1} e^{\{-\alpha_\gamma\beta_0\}} \\
&\propto \prod_{j=1}^{N(T_i)} \left(\frac{1}{\gamma_{ij}}\right)^{\alpha_\gamma} \alpha_\gamma^{\alpha_0-1} e^{\{-\alpha_\gamma\beta_0\}}
\end{aligned}
$$

$$
\begin{aligned}
\pi(\beta_\gamma|...) &\propto \prod_{j=2}^{N(T_i)} \pi(\gamma_{ij}|\alpha_\gamma, \beta_\gamma, d_{ij}, c_{ij-1}) \times Gamma(\beta_\gamma|\alpha_0, \beta_0) \\
&= \prod_{j=2}^{N(T_i)} \left(\frac{1}{\gamma_{ij}}\right)^{d_{ij}+\alpha_\gamma+1} e^{\{-(\frac{\beta_\gamma}{\gamma_{ij}}+(\frac{c_{ij-1}}{\gamma_{ij}})^{d_{ij}})\}} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \beta_\gamma^{\alpha_0-1} e^{\{-\beta_\gamma\beta_0\}} \\
&\propto \beta_\gamma^{\alpha_0-1} e^{\{-\beta_\gamma(\frac{1}{\gamma_{ij}}+\beta_0)\}} \\
&\propto \beta_\gamma \sim Gamma(\alpha_0, \frac{1}{\gamma_i} + \beta_0)
\end{aligned}
$$

Los parámetros a estimar son aquellos correspondientes a las variables latentes, cuyo kernel se estima para cada momento j $(j = 1, ..., N(T_i))$, y a las variables observables del modelo de cada individup i. Las verosimilitudes mostradas para los parámetros deben ser extendidas para incluir toda la población muestreada. De este modo, desarrollamos las verosimilitudes extendidas para los parámetros que corresponden al modelo Gamma-Gamma que describen las duraciones y la variable latente

θ; $\alpha_d, \alpha_\theta, \beta_\theta$,

$$\begin{aligned}
\pi(\alpha_d|(\theta_{ij})_{i=1.j=1}^{I\ N(T_i)}) &\propto \prod_{i=1}^{I}\prod_{j=1}^{N(T_i)} \pi(\alpha_d|\theta_{ij},...) \\
&\propto \prod_{i=1}^{I}\prod_{j=1}^{N(T_i)} \frac{(d_{ij-1}+\beta_\theta)^{\alpha_d+\alpha_\theta}}{\Gamma(\alpha_d+\alpha_\theta)}\theta_{ij}^{\alpha_d}\alpha_d^{\alpha_0-1}e^{\{-\alpha_d\beta_0\}} \\
&\propto \left(\frac{\alpha_d^{\alpha_0-1}}{\Gamma(\alpha_d+\alpha_\theta)} \quad e^{\{-\alpha_d\beta_0\}}\right)^{\sum_{i=1}^{I}N(T_i)} \times \\
&\times \prod_{i=1}^{I}\prod_{j=2}^{N(T_i)} (d_{ij-1}+\beta_\theta)^{\alpha_\theta+\alpha_d} \quad \theta_{ij}^{\alpha_d}
\end{aligned}$$

$$\begin{aligned}
\pi(\alpha_\theta|(\theta_{ij})_{i=1.j=1}^{I\ N(T_i)}) &\propto \prod_{i=1}^{I}\prod_{j=1}^{N(T_i)} \pi(\alpha_\theta|\theta_{ij},...) \\
&\propto \prod_{i=1}^{I}\prod_{j=1}^{N(T_i)} \frac{(d_{ij-1}+\beta_\theta)^{\alpha_d+\alpha_\theta}}{\Gamma(\alpha_d+\alpha_\theta)}\theta_{ij}^{\alpha_\theta}\alpha_\theta^{\alpha_0-1}e^{\{-\alpha_\theta\beta_0\}} \\
&\propto \left(\frac{\alpha_\theta^{\alpha_0-1}}{\Gamma(\alpha_d+\alpha_\theta)} \quad e^{\{-\alpha_\theta\beta_0\}}\right)^{\sum_{i=1}^{I}N(T_i)} \times \\
&\times \prod_{i=1}^{I}\prod_{j=2}^{N(T_i)} (d_{ij-1}+\beta_\theta)^{\alpha_\theta+\alpha_d} \quad \theta_{ij}^{\alpha_\theta}
\end{aligned}$$

$$
\begin{aligned}
\pi(\beta_\theta|(\theta_{ij})_{i=1\cdot j=1}^{I\ N(T_i)}) &\propto \prod_{i=1}^{I}\prod_{j=1}^{N(T_i)} \pi(\beta_\theta|\theta_{ij},...) \\
&\propto \prod_{i=1}^{I}\prod_{j=1}^{N(T_i)} (d_{ij-1}+\beta_\theta)^{\alpha_d+\alpha_\theta}\beta_\theta^{\alpha_0-1}e^{\{-\beta_\theta(\theta_i j+\beta_0)\}} \\
&\propto (\beta_\theta^{\alpha_0-1})^{\sum_{i=1}^{I}N(T_i)} \quad e^{\{-\beta_\theta(\sum_{i=1}^{I}\sum_{j=1}^{N(T_i)}(\theta_{ij}+\beta_0))\}} \times \\
&\times \prod_{i=1}^{I}\prod_{j=2}^{N(T_i)} (d_{ij-1}+\beta_\theta)^{\alpha_\theta+\alpha_d}
\end{aligned}
$$

De igual modo, se describen las verosimilitudes extendidas para los parámetros del modelo Gamma Inversa-Weibull que describe los costos y su variable latente γ; $\alpha_\gamma, \beta_\gamma$,

$$
\begin{aligned}
\pi(\alpha_\gamma|(\gamma_{ij})_{i=1\cdot j=1}^{I\ N(T_i)}) &\propto \prod_{i=1}^{I}\prod_{j=1}^{N(T_i)} \pi(\alpha_\gamma|\gamma_{ij},...) \\
&\propto \prod_{i=1}^{I}\prod_{j=1}^{N(T_i)} \left(\frac{1}{\gamma_{ij}}\right)^{\alpha_\gamma} \alpha_\gamma^{\alpha_0-1}e^{\{-\alpha_\gamma\beta_0\}} \\
&\propto (\alpha_\gamma^{\alpha_0-1} \quad e^{\{-\alpha_\gamma\beta_0\}})^{\sum_{i=1}^{I}N(T_i)} \times \\
&\times \prod_{i=1}^{I}\prod_{j=1}^{N(T_i)} \left(\frac{1}{\gamma_{ij}}\right)^{\alpha_\gamma}
\end{aligned}
$$

$$\begin{aligned}
\pi(\beta_\gamma|(\theta_{ij})_{i=1\cdot j=1}^{I\ N(T_i)}) &\propto \prod_{i=1}^{I}\prod_{j=1}^{N(T_i)} \pi(\beta_\gamma|\gamma_{ij},...) \\
&\propto \prod_{i=1}^{I}\prod_{j=1}^{N(T_i)} \beta_\gamma^{\alpha_0-1} e^{\{-\beta_\gamma(\frac{1}{\gamma_{ij}}+\beta_0)\}} \\
&\propto (\beta_\gamma^{\alpha_0-1})^{\sum_{i=1}^{I} N(T_i)} \quad e^{\{-\beta_\gamma(\sum_{i=1}^{I}\sum_{j=1}^{N(T_i)} \frac{1}{\gamma_{ij}}+\beta_0)\}}
\end{aligned}$$

Es importante mencionar que las distribuciones resultantes de la verosimilitud extendida cumplen con las propiedades de log-concavidad descritas en el capítulo anterior. De este modo y de manera general para cada individuo i, tomando la verosimilitud extendida definida anteriormente se fijan los valores iniciales para los parámetros $\alpha_d^{(0)}, \alpha_\theta^{(0)}, \beta_\theta^{(0)}, \alpha_\gamma^{(0)}, \beta_\gamma^{(0)}$ y para las variables latentes $\{\theta_j^{(0)}\}_{j=1}^{N(T_i)}, \{\gamma_j^{(0)}\}_{j=1}^{N(T_i)}$ y para cada $k = 1, ..., N(T_i)$ tenemos la siguiente distribución que es proporcional a la verosimilitud,

$$\pi(\alpha_d^{(k)}, \alpha_\theta^{(k)}, \beta_\theta^{(k)}, \alpha_\gamma^{(k)}, \beta_\gamma^{(k)}|\{\theta_i^{(k-1)}\}_{i=2}^{N(t)}, \{\gamma_i^{(k-1)}\}_{i=2}^{N(t)}, \{d_i, c_i\}_{i=1}^{N(t)})$$

Este es el principio necesario para utilizar el Muestreador de Gibbs de modo que se estimen los parámetros en base a las variables latentes que a su vez se estiman en base a las observaciones para que con los parámetros estimados se estimen las variables latentes que ayuden a predecir futuras observaciones.

Una vez que se definen las distribuciones particulares del modelo y los métodos numéricos con los que se realizará la inferencia, en el siguiente capítulo se implementará con una base de datos específica. Es importante

mencionar que, independientemente de la base de datos con la que se trabaje, las distribuciones y la forma del algoritmo es la correspondiente al modelo general de probabilidad.

Capítulo 5

Ilustración del Modelo con Datos

5.1. Introducción

Una vez que se han definido el modelo general de probabilidad y los algoritmos relativos a ese modelo de probabilidad en los capítulos anteriores se define el modelo general de probabilidad y cómo este puede hacer inferencia sobre observaciones; se puede pensar en una implementación de este modelo.

El modelo está definido para describir padecimientos crónico degenerativas, sus duraciones y los costos asociados a las mismas. En este capítulo, este modelo será implementado al padecimiento específico de diabetes mellitus tipo II con datos provenientes del Servicio Nacional de Salud en el Reino Unido, o NHS por sus siglas en inglés, en los años 2016-2017.

5.2. Diabetes Mellitus en el mundo: Costos y Prevalencia

En el panorama de salud mundial actual, las enfermedades crónicas han tomado un papel preponderante al ser padecimientos que no tienen cura, que se pueden prolongar de manera indefinida y requieren monitoreo y atención constante. Uno de los principales problemas de estos padecimientos consiste en la gran carga económica que estos presentan, junto con sus complicaciones; tanto a pacientes como a los proveedores de servicios de salud. Además de la decreciente calidad de vida que experimenta el paciente, aunado a la cantidad de procedimientos médicos que va necesitando y que muchas veces, su proveedor de servicios de salud ya no es capaz de suministrar.

Uno de los padecimientos con más prevalencia y que es más propenso a complicaciones es la diabetes mellitus. La insulina es la hormona que regula el azúcar en la sangre, entonces la diabetes mellitus es un padecimiento en el cual el páncreas no produce o produce poca insulina, o bien, las células del cuerpo no responden de manera normal a la insulina que se produce. Un elevado nivel de glucosa en la sangre puede, a largo plazo, derivar en una serie de complicaciones oftalmológicas, renales, cardiacas y de circulación.

La OrganizacionMundialdelaSalud (2016) (OMS) reconoce dos tipos de diabetes: Tipo I y Tipo II. La diabetes mellitus Tipo I (DM-TI) consiste en la falta de producción de insulina en el cuerpo, normalmente se diagnostica en edades tempranas y sus causas aún son desconocidas, por lo que no existe ninguna clase de tratamiento preventivo. La diabetes mellitus Tipo II (DM-TII) se debe a que el cuerpo no puede procesar correctamente la insulina, esto debido la combinación de sobrepeso y fal-

ta de actividad física; antes este padecimiento era exclusivo de la edad adulta, sin embargo, ahora se observa también en niños. La mayoría de enfermos son de tipo dos según el Atlas de la Diabetes publicado por Federación Internacional de la Diabetes (2015), donde se reporta que entre el 87 % y el 91 % de los enfermos de diabetes tiene DM-TII; esto provoca que no solo sea mayor la carga económica asociada a ellos, sino que además son más propensas a desarrollar complicaciones aunados a la falta de autocuidado del paciente.

Según la OMS, la prevalencia mundial de esta enfermedad casi se duplicó en las personas mayores de 18 años, de 4.7 % en 1980 a 8.5 % en 2014; además de provocar más de 2.2 millones de muertes tan solo en el año 2016 y ser una de las mayores causas de ceguera, fallas renales y amputación de extremidades. En las estadísticas reportadas por la OMS, la diabetes es la cuarta enfermedad que causa más muertes entre las no transmisibles mediante algún agente infeccioso, con un 6.01 % de la mortalidad.

De manera particular, se puede ver la prevalencia de la diabetes en los distintos países. En el Reino Unido la prevalencia en el año 2010/2011 fue de 3,818,545 personas de las cuales el 89.6 % se refieren a un diagnóstico de DM-TII; y según el estudio realizado por Hex et al. (2012) en el año 2035/2036 esta prevalencia aumentará a 6,289,925 mientras que las proporciones entre tipos de diabetes se mantienen. De acuerdo a datos de la OMS, de los 172 países en los que se analizan las estadísticas de mortalidad debido a diabetes mellitus, el Reino Unido ocupa el lugar 167 con 4.2 muertes por 100,000 personas. A pesar de que sus índices de mortalidad son relativamente bajos, comparados con el resto de los países, el problema con la creciente prevalencia de la enfermedad radica también en las complicaciones asociadas a la misma.

De acuerdo con Bolaños et al. (2010) para el año 2030 habrán alrededor de 366 millones de pacientes diagnósticados con diabetes en el mundo, de los cuales el 70 % se encontrarán en países de ingresos de medios a bajos, entre ellos los países de Latinoamérica. En específico en México se observan resultados preocupantes, en la Encuesta Nacional de Salud y Nutrición (ENSANUT) (Gutiérrez et al. (2012)) de 2012 la prevalencia de diabetes se reporta de 9.2 %, lo cual es un aumento considerable con lo reportado en la ENSANUT 2006 (7 %) y la ENSANUT 2000 (5.8 %). Viendo estos resultados, el aceleramiento en las tasas de obesidad y los padecimientos asociados el Instituto Nacional de Salud Pública decidió hacer otra encuesta intermedia a la que denominó ENSANUT de Medio Camino, la cual se realizó en el año 2016 con un resultado en la prevalencia de 9.4 %, el cual confirma la tendencia creciente de la misma. En las estadísticas de mortalidad de la OMS, México ocupa el décimo lugar con 90.5 muertes derivadas del diagnóstico de diabetes mellitus por 100,000 personas y las estadísticas nacionales muestran que la diabetes mellitus es la tercera causa de mortalidad. En nuestro país, este padecimiento toma un cariz más apremiante debido a la cantidad de gente diagnosticada y el porcentaje de mortalidad asociado.

A pesar que se preveé que la mayoría de los enfermos de diabetes se encuentren en países en vías de desarrollo, también hay que tomar en cuenta que dado los factores de riesgo que facilitan el desarrollo de DM-TII, la más común; son un estilo de vida sedentario y obesidad. Estos factores normalmente se pueden encontrar en países con economías desarrolladas, tales como EE.UU. Tomando como referencia el reporte elaborado en 2014 por el Centro para el Control y Prevención de Enfermedades (CDC (2014), por sus siglas en inglés) la prevalencia de diabetes es de 9.3 % de la población, es decir, 29.1 millones de personas; aunque

el 27.8 % de estas personas aún no estén diagnosticadas. Por otro lado, el estudio de AmericanDiabetesAssociation et al. (2013) (ADA, por sus siglas en inglés) estima que la prevalencia de este padecimiento para el año 2012 sea de 7 %, que aunque difiere de lo reportado por la CDC, sigue siendo consistente con las mismas proyecciones. Se estima que cada año se diagnostican 1.4 millones de pacientes con diabetes. Este padecimiento está listado como la séptima causa de muerte en EE.UU., ya sea como causa principal del deceso o como causa subyacente. En relación a lo reportado por la OMS, EE.UU. se coloca en la posición 124 con 13.40 muertes entre 100,000 personas.

Independientemente del país del que se hable, se puede concluir que la diabetes mellitus, particularmente la Tipo II, es uno de los más grandes retos de salud pública que se enfrentan actualmente. Lo amenazante del padecimiento no es sólo el crecimiento constante en sus tasas de prevalencia sino también las complicaciones que éste presenta y que conducen, eventualmente, a la muerte. Como consecuencia, se tiene que el gasto asociado a este padecimiento y a sus complicaciones se vuelve cada vez más grande y muchas veces imposible de sostener por las instituciones proveedoras de servicios de salud.

En el último Atlas de la Diabetes elaborado por la Federación Internacional de la Diabetes (2015) (IDF, por sus siglas en inglés), se estima que el gasto asociado a la diabetes mellitus es de USD 673,000 millones, es decir, el 12 % del total del gasto mundial en salud. Esto quiere decir que en la mayoría de los países gastan entre un 5 % y 20 % de su presupuesto de salud, exclusivamente en el tratamiento de la diabetes mellitus. Esto debido a que los gastos en servicios de salud promedios de las personas con diabetes es entre dos y tres veces mayores que aquellos de las personas sin diabetes. En cara a estos datos, se vuelven prioritaras

las estrategias de prevención y reducción de costos. Es importante destacar que los costos asociados a cualquier padecimiento, particularmente a la diabetes, deben tomarse en cuenta los costos directos del padecimiento (diagnóstico inicial, consultas, medicamentos, hospitalizaciones), los costos de las complicaciones (consultas, medicamentos, hospitalizaciones) y los costos indirectos (costos en productividad, muerte prematura, ausentismo, etc.).

En el análisis realizado por Hex et al. (2012) sobre la prevalencia y costos de la diabetes mellitus en el Reino Unido se estima un costo total de la diabetes en el año 2010/2011 de £23,631 millones, con una proyección de costos para el año 2035/2036 de £39,753 millones. El gasto de 2010/2011 es el 10 % del presupuesto asignado al Servicio Nacional de Salud (NHS, por sus siglas en inglés); mientras que la proyección de 2035/2036 llega hasta el 17 % de este mismo presupuesto. Lo que se muestra revelador de este estudio, en el cálculo de los costos de 2010/2011, son las diferencias entre los costos directos, de complicaciones e indirectos. Los costos directos son solo el 9 % de los costos totales, los costos de complicaciones 33 % mientras que lo más oneroso son los costos indirectos con una participación del 59 %; estas proporciones se mantienen con una leve variación para las proyecciones de 2035/2036. Igualmente se mantiene el supuesto que la DM-TII es mucho más costosa, pues el gasto total calculado en este tipo específico de diabetes oscila entre el 89 % y 92 %. Esto quiere decir que aunque el padecimiento en sÃ no tenga la mayor parte asociada de la carga económica, es la causa subyacente de muchas complicaciones y diagnósticos más costosos.

La región de Norteamérica y el Caribe tienen la mayor prevalencia de diabetes con un 12.9 % de la población adulta afectada, esto de acuerdo a la Federación Internacional de la Diabetes (2015). En esta región el

gasto total asociado está entre los USD 348,000 y USD 610,000 millones, este gasto representa más de la mitad del presupuesto mundial para la diabetes. Particularmente en México, según el estudio de Barcelo et al. (2003), México es el país con los costos anuales más elevados en comparación con el resto de Latinoamérica. Al igual, que el análisis realizado para el Reino Unido, el costo anual de México de USD 15,118 millones, este se puede dividir entre USD 1,974 millones de costos directos (diabetes y complicaciones) y USD 13,144 millones de costos indirectos. Estos costos asociados fueron calculados con una prevalencia del 4.1 %, la cual está desactualizada, por lo que se podría asumir que el gasto actual está por arriba. De nuevo y de manera más actual, el minucioso reporte de FUNSALUD realizado por Barraza-Lloréns et al. (2015), estima que los costos asociados, particularmente para la DM-TII, sean de $362,800 millones de pesos en el año base, de los cuales el 49 % se refieren a costos directos y el 51 % a costos indirectos. Para el año 2018, este mismo estudio proyecta un gasto total de $506,000 millones de pesos; el aumento se explica con las proyecciones demográficas y de prevalencia estimada. Este estudio confirma que la DM-TII es mucho más cara y común que la Tipo I y pone en evidencia la importancia de mejor planeación financiera para hacer frente a esta epidemia de salud pública.

Según la FID, del gasto asociado a la región de Norteamérica, los EE.UU. representan la mayor parte del mismo con un gasto estimado en 2015 de USD 320,000 millones. De igual modo, el estudio de la AmericanDiabetesAssociation et al. (2013) estima que el gasto asociado con la diabetes mellitus en el 2012 es de USD 245,000 millones, de este monto el 72 % son costos directos del padecimiento y sus complicaciones y el 28 % restante se refiere a los costos indirectos; a diferencia del Reino Unido o México donde los costos indirectos sobrepasaban los costos directos. Los costos en salud de las personas diagnosticadas con diabetes son 2.3 veces

más altos que los de aquellos no diagnosticados y según la Agencia para la Investigación y Calidad para el Servicios de Salud (AHRQ (2016), por sus siglas en inglés), la diabetes mellitus es el cuarto padecimiento más costoso en los EE.UU. Además, de acuerdo al estudio de Zhuo et al. (2014), entre más joven sea el paciente a la edad de diagnóstico mucho mayor será el costo de salud acumulado. Debido a que EE.UU. es uno de los países con mayor prevalencia y costos asociados a la diabetes mellitus en el mundo, la correcta cuantificación de los mismos es crucial para la correcta planeación financiera de los proveedores de salud.

Como podemos ver por las cifras reportadas, el problema del gasto asociado a la diabetes mellitus es uno de los más grandes retos que existen en el horizonte de la salud pública. Uno de los mecanismos principales para disminuir la prevalencia y por ende los gastos, se reducen a la prevención y a la procuración de un estilo de vida más sano; sobre todo para la DM-TII. Sin embargo, aunado a las recomendaciones para el paciente también se necesitan nuevas maneras de hacer estos gastos más efectivos mediante la correcta localización del rubro que genera más presión en el gasto y el uso más eficiente de los recursos existentes, para así poder hacerlos más accesibles tanto a los pacientes como a las instituciones proveedoras de salud.

Como se menciona en Cichon et al. (1999) las herramientas estadísticas son de gran ayuda para entender la realidad actual y poder elaborar proyecciones futuras, especialmente en el campo de la medición de costos en el sector salud. Son muchas las preguntas que rodean la diabetes mellitus, acerca de su prevalencia, su morbilidad y sus costos. En este trabajo de investigación se busca hacer uso de las herramientas estadísticas mencionadas en los capítulos anteriores para contestar las preguntas sobre el costo que puede tener la enfermedad según la trayectoria del pa-

decimiento. Es importante mencionar que para efectos del análisis que se realizará en esta implementación del modelo general de probabilidad, solo se tomarán en consideración los costos directos generados por la enfermedad no aquellos relativos a las complicaciones.

5.3. Escasez de datos

Como mencionado en la sección anterior, se cuenta con mucha información sobre la prevalencia y los costos de la diabetes mellitus tipo-II en el mundo. La mayoría de esta información es obtenida de distintas fuentes aunque se refieran al mismo país, en el caso particular de México, se han realizado varios estudios tratando de estimar la prevalencia y los costos de este padecimiento; varios de ellos analizados en la sección anterior.

De acuerdo con Barquera et al. (2013), los mayores registros sobre las trayectorias de las enfermedades y los gastos asociados a los tratamientos son elaborados por las distintas instituciones públicas de salud como el IMSS, ISSSTE y la Secretaría de Salud, que son exclusivas para uso interno administrativo por lo que no se ha podido elaborar una base de datos consistente y precisa con la que se pueda realizar un análisis epidemiológico y estadístico. Aunque se cuenta con la ENSANUT como un estudio general para la prevalencia y costos y con el INEGI para las estadísticas de mortalidad desagregadas por causa, éstas no cuentan con la información individual de los pacientes y sus trayectorias. Es por esta razón que este trabajo de investigación no se pudo realizar con información sobre pacientes de México.

El problema de la falta de bases de datos con las trayectorias de los pacientes de cualquier padecimiento no solo afecta a México. También a países como Estados Unidos de Norteamérica donde existen tantas

instancias que estudian el padecimiento de diabetes como la AHRQ, la Asociación Americana de la Diabetes, el Centro para el Control y Prevención de Enfermedades; que resulta inexistente una base consolidada de datos con la información de los diagnósticos y tratamientos de los pacientes. Incluso, la OMS solo cuenta con la información general de prevalencia, costos y mortalidad con ayuda de la Federación Internacional de la Diabetes.

Sin embargo, el NHS de Reino Unido ha hecho un esfuerzo por elaborar este tipo de bases de datos con las trayectorias completas de los pacientes. Es por este motivo que en este trabajo de investigación se utilizará esta base de datos para la implementación del modelo.

Esta base de datos comprende las trayectorias de padecimiento a través del tiempo de 500 pacientes en un arreglo de datos panel, donde se combina una dimensión transversal para los individuos $(i = 1, ..., I)$ con una dimensión longitudinal o temporal que describe la trayectoria del padecimiento del individuo $(j = 1, ..., T_i)$. En la próxima sección se describirá con más detalle las características de este panel de datos.

5.4. Descripción de Datos

Como mencionado en la sección anterior, la base de datos describe las trayectorias individuales de 500 pacientes a través del tiempo. Las trayectorias de cada paciente tienen a su vez dos dimensiones, la dimensión del tratamiento o prescipción y el costo de cada una de las prescripciones.

Es importante mencionar que la dimensión de la prescripción solo hace referencia al diagnóstico inicial de DM-TII; es decir, que no se refiere a tratamientos por complicaciones de la enfermedad como pueden ser pa-

decimientos cardíacos, nefrológicos, circulatorios, etc; como mencionado anteriormente en este capítulo. Por lo tanto, para esta base, existen doce tipos de tratamientos recetados para las distintas etapas del padecimiento descritas en la siguiente tabla,

Código de tratamiento	**Tratamiento**
1	Total primary care (all BNF)
2	Drugs used for Diabetes (BNF 6.1)
3	Insulins (BNF 6.1.1)
4	Short-acting insulins (BNF 6.1.1.1)
5	Intermediate and long-acting insulins (BNF 6.1.1.2)
6	Antidiabetic drugs (BNF 6.1.2)
7	Sulfonylureas (BNF 6.1.2.1)
8	Biguanides (BNF 6.1.2.2)
9	Other antidiabetic drugs (BNF 6.1.2.3)
10	Diagnostic and monitoring devices (BNF 6.1.6)
11	Human analogue insulins
12	Other insulins

Cuadro 5.1: Códigos de Tratamientos

Cada uno de estos tratamientos tiene asociado un precio. Para la representación de estos precios se tomará el precio efectivo por dosis, pues un tratamiento puede requerir más de una dosis en cada aplicación o toma. Los precios considerados por tratamiento son los precios reales; esto quiere decir que se toma como a$\tilde{n}$ base a 2016/2017 excluyendo de esta manera el efecto de la inflación. También cabe mencionar que los precios están expresados en libras esterlinas (£), por lo que se convierten a pesos a una tipo de cambio de 23.0625 pesos por libra. De este modo,

Código de tratamiento	**Tratamiento**	**Precio (GBP/Unidad)**	**Precio (MXN/ Unidad)**
1	Total primary care (all BNF)	8.16	188.19
2	Drugs used for Diabetes (BNF 6.1)	18.91	436.19
3	Insulins (BNF 6.1.1)	49.33	1,137.72
4	Short-acting insulins (BNF 6.1.1.1)	46.57	1,074.04
5	Intermediate and long-acting insulins (BNF 6.1.1.2)	50.62	1,167.33
6	Antidiabetic drugs (BNF 6.1.2)	11.86	273.41
7	Sulfonylureas (BNF 6.1.2.1)	3.19	73.58
8	Biguanides (BNF 6.1.2.2)	4.53	104.44
9	Other antidiabetic drugs (BNF 6.1.2.3)	39.38	908.09
10	Diagnostic and monitoring devices (BNF 6.1.6)	25.93	597.96
11	Human analogue insulins	53.10	1,224.64
12	Other insulins	33.49	772.45

Cuadro 5.2: Precios por Tratamiento

Una vez que están definidos los tratamientos y los costos que caracterizan las trayectorias individuales de los padecimientos de cada paciente, se puede empezar a realizar un análisis exploratorio y descriptivo de los datos.

5.5. Análisis Descriptivo

5.6. Resultados

Descripción de los resultados inferenciales y predictivos.

Capítulo 6

Conclusiones

6.1. Qué me deja este trabajo?

6.2. Visión Crítica Autónoma

6.3. Trabajo Futuro

Como este trabajo sienta las bases para nuevos modelos de tarificación.

Bibliografía

AHRQ (2016). Medical expenditure panel survey, meps. `https://meps.ahrq.gov/mepsweb/data_stats/quick_tables_search.jsp?component=2&subcomponent=1`. Accesed: 30-01-2017.

AmericanDiabetesAssociation et al. (2013). Economic costs of diabetes in the us in 2012. *Diabetes care*, 36(4):1033–1046.

An, M. Y. (1996). Log-concave probability distributions: Theory and statistical testing.

Bagnoli, M. and Bergstrom, T. (2005). Log-concave probability and its applications. *Economic theory*, 26(2):445–469.

Barcelo, A., Aedo, C., Rajpathak, S., and Robles, S. (2003). The cost of diabetes in latin america and the caribbean. *Bulletin of the world health organization*, 81(1):19–27.

Barquera, S., Campos-Nonato, I., Aguilar-Salinas, C., Lopez-Ridaura, R., Arredondo, A., and Rivera-Dommarco, J. (2013). Diabetes in mexico: cost and management of diabetes and its complications and challenges for health policy. *Globalization and health*, 9(1):3.

Barraza-Lloréns, M., Guajardo-Barrón, V., Picó, J., García, R., Hernández, C., Mora, F., Athié, J., Crable, E., and Urtiz, A. (2015). Carga

económica de la diabetes mellitus en méxico, 2013. *México, DF: Funsalud.*

Bolaños, R. d. l. Á. R., Shigematsu, L. M. R., Ruíz, J. A. J., Márquez, S. A. J., and Ávila, M. H. (2010). Costos directos de atención médica en pacientes con diabetes mellitus tipo 2 en méxico: análisis de microcosteo. *Rev Panam Salud Publica*, 28(6):412–20.

Carter, C. K. and Kohn, R. (1996). Markov chain monte carlo in conditionally gaussian state space models. *Biometrika*, 83(3):589–601.

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

CDC (2014). National diabetes statistics report, 2014. `https://www.cdc.gov/diabetes/pdfs/data/2014-report-estimates-of-diabetes-and-its-burden-in-the-united-states.pdf`. Accesed: 30-01-2017.

Cichon, M., Newbrander, W., Yamabana, H., Weber, A., Normand, C., Dror, D., and Preker, A. (1999). *Modelling in health care finance: A compendium of quantitative techniques for health care financing.* International Labour Organization.

Daley, D. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods.* 2nd edition edition.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Engle, R. F. and Russell, J. R. (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, pages 1127–1162.

Fader, P. S. and Hardie, B. G. (2013). The gamma-gamma model of monetary value.

Federación Internacional de la Diabetes, B. (2015). Atlas idf diabetes, 2015. *Available from:[Last accessed: Enero 2017]*.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Ghahramani, Z. (2001). An introduction to hidden markov models and bayesian networks. *International journal of pattern recognition and artificial intelligence*, 15(01):9–42.

Goldstein, M. (2013). Observables and models: Exchangeability and the inductive argument. *Bayesian Theory and Its Applications*, pages 3–18.

Gutiérrez, J., Rivera-Dommarco, J., Shamah-Levy, T., Villalpando-Hernández, S., Franco, A., Cuevas-Nasu, L., et al. (2012). Encuesta nacional de salud y nutrición. ensanut 2012 resultados nacionales. cuernavaca, méxico: Instituto nacional de salud pública (mx), 2012 [accedido en enero 2017].

Hahn, M. G. and Zhang, G. (2012). Exchangeable random variables. *High Dimensional Probability*, 43:111.

Harrison, J. and West, M. (1999). *Bayesian forecasting & dynamic models*. Springer New York.

Held, L. and Sabanés Bové, D. (2014). *Applied statistical inference*, volume 10. Springer.

Hex, N., Bartlett, C., Wright, D., Taylor, M., and Varley, D. (2012). Estimating the current and future costs of type 1 and type 2 diabetes in the uk, including direct health costs and indirect societal and productivity costs. *Diabetic Medicine*, 29(7):855–862.

Neal, R. M. (2003). Slice sampling. *Annals of statistics*, pages 705–741.

OrganizacionMundialdelaSalud (2016). Diabetes. fact sheet. `http://www.who.int/mediacentre/factsheets/fs312/en/`. Accesed: 20-01-2017.

Paik Schoenberg, F. (2000). *Introduction to Point Processes*.

Pitt, M. K., Chatfield, C., and Walker, S. G. (2002). Constructing first order stationary autoregressive models via latent processes. *Scandinavian Journal of Statistics*, 29(4):657–663.

Resnick, S. I. (1999). *A Probability Path*. Birkhauser.

Schervish, M. J. (2012). *Theory of statistics*. Springer Science & Business Media.

Zhuo, X., Zhang, P., Barker, L., Albright, A., Thompson, T. J., and Gregg, E. (2014). The lifetime cost of diabetes and its implications for diabetes prevention. *Diabetes care*, 37(9):2557–2564.

Appendices

.1. Slice Sampler

En la función de verosimilitus definida en la sección anterior, las distribuciones de la variable latente γ_i y de los parámetros $\alpha_d, \alpha_\theta, \alpha_\gamma$ no se pueden expresar de una forma analítica cerrada. Es por esto que antes de empezar a estimar la función de verosimilitud es necesario muestrear primero estas distribuciones para lograr que los métodos de estimación funcionen de la mejor manera. Este primer acercamiento se hace mediante el Slice Sampler.

De acuerdo con Neal (2003) existen varias aplicaciones para esta manera de muestrear, desde la manera más simple con una distribución univariada como alternativa al Muestreador de Gibbs hasta la más compleja donde se pueden adaptar relaciones de dependencia entre las variables, permitiendo mayor flexibilidad al Muestreador de Gibbs. Esta es la clase de Slice Sampler que se utilizará en este trabajo de investigación.

La idea detrás del concepto del Slice Sampler es que si queremos muestrear la distribución de una variable x que toma valores en el conjunto $\mathbb{R}^n$, cuya densidad sea proporcional a alguna función $f(x)$, entonces se muestrea de forma uniforme en la región por debajo de la curva de $f(x)$ con dimensión $(n+1)$. Esta idea se puede concretar introduciendo una variable real auxiliar y y definiendo la distribución conjunta de x y y que es uniforme sobre la región debajo de la curva de $f(x)$, tal que, $U = \{(x, y) : 0 < y < f(x)\}$. Es decir, que para muestrear x, muestreamos conjuntamente (x, y) para después ignorar y.

Sin embargo, dado que generar puntos independientes muestreados de la distribución U no siempre es sencillo, se puede definir una Cadena de Markov que converja a esta distribución uniforme. Esto se hace median-

te el muestreo alternado de la distribución condicional de $y|x$ la cual es uniforme en el intervalo $(0, f(x))$ y de la distribución condicional de $x|y$ la cual también es uniforme sobre la región $S = \{x_y < f(x)\}$, la cual es nombrada como la *rebanada* definida por y. El procedimiento para construir la Cadena de Markov para una distribución univariada $f(x)$, tomando el valor inicial x_0, de acuerdo con Neal (2003), es:

- Se extrae un valor real y uniforme en $(0, f(x))$, definiendo una *rebanada* horizontal $S = \{x : y < f(x)\}$. Es importante notar que x_0 está siempre dentro de S.
- Se encuentra un intervalo $I = (L, R)$ en la vecindad de x_0 que contenga toda, o la mayor parte, de la *rebanada*.
- Se extrae un nuevo punto, x_1, de la parte de la *rebanada* que está dentro del intervalo I.

En el primer paso del procedimiento se elige la variable auxiliar que es característica al Slice Sampler, pues este valor no es necesario de una iteración de la Cadena de Markov a la siguiente; mientras que los siguientes dos pasos pueden ser implementados de muchas maneras en tanto que la Cadena de Markov resultante permita que la distribución definida $f(x)$ permanezca invariante. El siguiente problema que se enfrenta es delimitar el intervalo, pues necesita ser lo suficientemente grande para que el nuevo punto (x_1) esté lo más lejos del anterior (x_0) dentro de la misma *rebanada*, pero el intervalo tampoco puede salirse de la misma pues eso volvería ineficiente al muestreo.

Para garantizar convergencia en la distribución $f(x)$ y su invarianza, la Cadena de Markov debe ser ergódica. Según Neal (2003), para demostrar que la función de distribución permanece invariante, suponemos que el estado inicial x_0 se distribuye $f(x)$, en el primer paso del procedimiento

la distribución conjunta es con las variables x_0 y y, por lo que al actualizar x_0 a x_1 la función de distribución conjunta se mantiene invariante, ignorando así la variable y. La distribución de x_1 es la distribución marginal de la conjunta, es decir, la función $f(x)$ definida. De este modo, lo único que se necesita demostrar que la selección de x_0 y x_1 en los siguientes dos pasos del procedimiento deja a la distribución conjunta de x y y invariante, y con la distribución condicional sobre $S = \{x : y < f(x)\}$, es decir, la *rebanada* definida por y. Esta invarianza se puede demostrar si la probabilidad de que x_1 sea el próximo estado dado que el estado actual es x_0 es igual a la probabilidad de que x_0 sea el próximo estado dado que el estado actual es x_1, para cualquier x_0 y x_1 en S.

Este es el procedimiento que aplica para cuando la distribución es univariada, sin embargo, si esta llegara a ser una distribución multivariada ($x = (x_1, ..., x_n)$) existen dos caminos para hacer el Slice Sampler: el primero es muestrear para cada variable por separado; para esto es necesario poder calcular la función $f_i(x_i)$ y que esta sea proporcional a $p(x_i|\{x_j\}_{j\neq i})$, donde $\{x_j\}_{j\neq i}$ son los valores de las variables. El otro camino es seguir el mismo procedimiento que se definió para las distribuciones univariadas, solamente que en el segundo paso se reemplaza el intervalo con un hiperrectángulo $H = \{x : L_i < x_i < R_i, \quad i = 1, ..., n\}$ donde L_i y R_i definen la extensión del hiperrectángulo a lo largo del eje para la variable x_i.

En el caso de las distribuciones de los parámetros $\alpha_d, \alpha_\theta, \alpha_\gamma$ y de la variable latente γ_i son distribuciones con una sola variable, por lo que el Slice Sampler a utilizar para estimar es el de una sola variable. Ahora, una vez estimado estas distribuciones es necesario utilizar el Muestreador de Gibbs que se desarrolla en la siguiente sección para la estimación del modelo general de probabilidad.

.2. El Algoritmo EM

El algoritmo EM es un algoritmo para calcular el estimador de máxima verosimilitud, que de acuerdo con Held and Sabanés Bové (2014), se define como

Definicion 4. *El Estimador de Máxima Verosimilitud (EMV)* $\hat{\theta}_{MV}$ *del parámetro* θ *se obtiene maximizando la función de verosimilitud.*

$$\hat{\theta}_{MV} = max_{\theta \in \Theta} L(\theta)$$

Según Dempster et al. (1977) el algoritmo EM calcula el EML mediante iteraciones, cada iteración consiste en un paso dónde se calcula la esperanza y en otro se maximiza la misma, de ahí el nombre de EM. Este algoritmo se relaciona con las variables latentes suponiendo dos variables x y y las cuales se relacionan $x \to y(x)$, donde y son los datos observables.

De este modo, análogamente a lo expresado en el capítulo anterior por Pitt et al. (2002), se proponen las siguientes funciones de densidad $f(x|\phi)$ y $g(y|\phi)$; en las cuales, de acuerdo a Dempster et al. (1977) los datos completos (variables latentes) $f(x|\cdot)$ se relacionan con los datos incompletos (variables observadas) $g(y|\cdot)$ mediante

$$g(y|\phi) = \int_{\chi(y)} f(x|\phi)dx$$

El algoritmo EM se dedica a encontrar un valor de ϕ que maximice $g(y|\phi)$ dada la y observada usando la familia asociada de $f(x|\phi)$. Una de las caracterizaciones más simples supone $\phi^{(p)}$ es el valor actual de ϕ después de p iteraciones y $t(x)$ como el estadístico suficientes de los datos completos, es decir, el estimador de la variable latente; por lo que

la siguiente iteración se puede desglosar en los siguientes dos pasos:

- Paso E: Estimar los estadísticos suficientes de los datos completos.

$$t^{(p)} = E[t(x)|y, \phi^{(p)}]$$

- Paso M: Determinar $\phi^{(p+1)}$ como solución a la ecuación

$$E[t(x)|\phi] = t^{(p)}$$

Es decir, que si suponemos que $t^{(p)}$ es el estadístico suficiente calculado de x observada en la distribución $f(x|\phi)$ entonces la ecuación definida en el Paso M se define como el EMV. Este concepto se hace general al definir la siguiente función

$$Q(\phi'|\phi) = E[logf(x|\phi')|y, \phi]$$

Esta función se asume que existe para toda pareja (ϕ', ϕ). Se define la iteración EM para $\phi^{(p)} \rightarrow \phi^{(p+1)}$,

- Calcular $Q(\phi, \phi^{(p)})$.
- Determinar $\phi^{(p+1)}$ tal que maximice $Q(\phi, \phi^{(p)})$.

La idea central es tomar una ϕ' que maximice $logf(x|\phi)$, dado que esta distribución y su correspondiente logaritmo no necesariamente se conoce, se puede maximizar los datos observados y $\phi^{(p)}$.

El algoritmo EM es muy útil pues por su estructura iterativa puede dar resultados a modelos de probabilidad muy complejos, además de que al igual que el Muestreador de Gibbs, utiliza una estructura subyacente o de variables latentes.

.3. El Muestreador de Gibbs

.3.1. El Muestreador de Gibbs general

Aunque el muestreador de Gibbs pueda ser útil en la visión clásica de la estadística, normalmente el muestreador Gibbs se asocia con la estadística bayesiana, como es el caso de este trabajo. Según Casella and George (1992) este algoritmo es una técnica que genera variables aleatorias indirectamente de distribuciones marginales sin tener que calcular la densidad, debido a que se basa en las propiedades principales de las Cadenas de Markov como la estacionareidad para simplificar cálculos y tener estimados más precisos.

Siguiendo la ilustración de Casella and George (1992), supongamos que tenemos una distribución conjunta $f(\theta, y_1, y_2, ..., y_p)$

$$f(\theta) = \int \cdots \int f(\theta, y_1, y_2, ..., y_p) dy_1, dy_2, ..., dy_p$$

Si el interés se encuentra en la marginal $f(\theta)$ y ésta es demasiado complicada para calcularse directamente, con el Muestreador de Gibb se puede generar una muestra $\theta_1, ..., \theta_m \sim f(\theta)$ sin la necesidad de calcular la distribución marginal. Esto permite obtener información de la misma con alto grado de precisión.

Para ejemplificar mejor el mecanismo del muestreador de Gibbs se toman dos variables aleatorias (Θ, Y). El algoritmo genera una muestra de $f(\theta)$ muestreando de las distribuciones condicionales $f(\theta|y)$ y $f(y|\theta)$ que son la que normalmente se conocen en los modelos estadísticos. Esta muestra se obtiene mediante, lo que Casella and George (1992) nombra como, una secuencia de Gibbs $(Y'_0, \theta'_0, Y'_1, \theta'_1, ..., Y'_k, \theta'_k)$ que de manera iterativa genera variables aleatorias a partir de valores iniciales especifi-

cados ($Y'_0 = y'_0$).

El proceso iterativo es como sigue

$$\theta'_j \sim f(\theta|Y'_j = y'_j)$$
$$Y'_{j+1} \sim f(y|\theta'_j = \theta'_j)$$

Si la muestra es suficientemente grande, es decir, que si $k \to \infty$ la distribución de θ'_k convergerá con la verdadera distribución marginal de θ.

El muestreador de Gibbs puede pensarse como una implementación práctica del concepto de que solo conociendo las distribuciones marginales se puede determinar la distribición conjunta. Esto sería cierto en la mayoría de los casos bivariados, el procedamiento no es tan directo para los casos multivariados.

De acuerdo con Casella and George (1992) para el caso bivariado, suponemos dos variables aleatorias θ, Y, de las cuales se conocen sus distribuciones condicionales $f_{\Theta|Y}(\theta|y)$ y $f_{Y|\Theta}(y|\theta)$. A partir de estas podríamos calcular la función marginal de θ y la distribución conjunta de ambas variables, mediante el siguiente argumento:

$$f_\theta(\theta) = \int f_{\theta Y}(\theta, y)dy$$

donde la distribución conjunta es aún desconocida, tomando el hecho que $f_{\theta Y}(\theta, y) = f_{\theta|Y}(\theta|y)f_Y(y)$ tendríamos que,

$$f_\theta(\theta) = \int f_{\theta|Y}(\theta|y)f_Y(y)dy$$

Asimismo, si sustituimos la distribución marginal de y ($f_Y(y)$) con el mismo argumento utilizado para la distribución marginal de θ, se tiene que

$$\begin{aligned} f_\theta(\theta) &= \int f_{\theta|Y}(\theta|y) f_{Y|\theta}(y|t) f_\theta(t) dt dy \\ &= \int [\int f_{\theta|Y}(\theta|y) f_{Y|\theta}(y|t) dy] f_\theta(t) dt \end{aligned}$$

Esta ecuación es una forma limitada de la iteración de Gibbs, ilustrando como las distribuciones condicionales producen una distribución marginal. Aunque la distribución conjunta de las variables determinan las distribuciones condicionales y marginales, no siempre las condicionales determinen de manera tan directa la distribución marginal. Esto es cierto no solo para los casos bivariados, sino que se extiende a los multivariados.

En cuantas más variables existan, el problema se vuelve más complejo pues la relación entre las condicionales, marginales y conjuntas se vuelve más intrincada. Por ejemplo, la relación $condicional \times marginal = conjunta$ no se sontiene para todas las condicionales y marginales. Pero se pueden hacer varios conjuntos de variables y construir las ecuaciones integrales para calcular la distribución marginal de interés.

Para casos multivariados Casella and George (1992) supone las variables aleatorias X, Y, Z con interés en la distribución $f_X(x)$. Para esto, se toman las variables (Y, Z) como una sola variable, lo que resultaría en

$$f_X(x) = \int [\int \int f_{X|YX}(x|y,z) f_{YZ|X}(y,z|t) dy dz] f_X(t) dt$$

De esta manera, muestreando iterativamente de $f_{X|YZ}$ y $f_{YZ|X}$ resultarían en una serie de variables aleatorias que convergen en $f_X(x)$. Por otro lado, el Muestreador de Gibb muestrearía iterativamente las distri-

buciones $f_{X|YZ}, f_{Y|XZ}, f_{Z|X}$, de tal modo que en la j-ésima iteración se tendría,

$$\begin{aligned} X'_j &\sim f(x|Y'_j = y'_j, Z'_j = z'_j) \\ Y'_{j+1} &\sim f(y|X'_j = x'_j, Z'_j = z'_j) \\ Z'_{j+1} &\sim f(z|X'_j = x'_j, Y'_{j+1} = y'_{j+1}) \end{aligned}$$

Este esquema de iteraciones nos produce una secuencia de Gibbs,

$$Y'_0, Z'_0, X'_0, Y'_1, Z'_1, X'_1, ...$$

con la misma propiedad de convergencia que en el caso bivariado, ente más grande es la k, $X'_k = x'_k$ es un punto de la distribución marginal $f(x)$.

De este modo queda evidenciada la utilidad del Muestreador de Gibbs en el ahorro de cálculos y la precisión de sus resultados. Como mencionado en la sección anterior, esta técnica inferencial es muy útil tanto en la estadística bayesiana como en la clásica, en la primera para calcular la distribución posterior y en la última, para calcular la función de verosimilitud. Según Gelman et al. (2014), la clave del éxito de este método es la iteración en la cual las distribuciones aproximadas mejoran hasta converger en la distribución deseada.

.3.2. El Muestreador de Gibbs para modelos espacio-estado

El Muestreador de Gibbs tradicional puede ser un poco limitado en lo que se refiere a su aplicación en un problema que se desarrolla a través del tiempo; sin embargo, este se puede adaptar a los requerimientos par-

ticulares del caso. Carter and Kohn (1996) exponen un caso particular en el contexto de Modelos Espacio-Estado Gaussianos, que aunque no es exactamente el modelo planteado en este trabajo, algunas de las ideas expuestas pueden ser extendidas.

El modelo planteado por Carter and Kohn (1996) es,

$$y_i = h_i' x_i + \gamma_i e_i; \qquad x_i = F_i x_{i-1} + \Gamma_i u_i$$

Donde las observaciones y_i son escalares y x_i es el vector de estados de dimensión $m \times 1$. Los errores e_i y u_i son independientes y se distribuyen $N(0, \sigma^2)$ y $N(0, \tau^2 I_m)$. Los coeficientes $h_i, \gamma_i, F_i, \Gamma_i$ son determinados por la variable discreta K_i. Usando la notación para el vector de observaciones de $Y := (y_1, ..., y_n)'$, el vector total de estados $X := (x_1', ..., x_n')'$, $K := (K_1, ..., K_n)$. Sea $g_i := h_i' x_i$, por lo que $G := (g_1, ..., g_n)'$. Asumiendo que σ^2, τ^2 y K son independientes; las distribuciones priori es Gamma Inversa y la distribución priori de K es una Cadena de Markov con probabilidades de transición conocidas. También se asume que dado K_1 y τ^2, la distribución de x_1 es normal.

Se propone el siguiente muestreador para estimar X, K, σ^2 y τ^2, mediante la generación de las siguientes distribuciones condicionales,

1. $p(\tau^2|Y, G, K, \sigma^2)$ que se puede reescribir como $p(\tau^2|G, K)$.
2. $p(K_i|Y, K_{j\neq i}, \sigma^2, \tau^2)$ para $i = 1, ..., n$.
3. $p(X|Y, K, \sigma^2, \tau^2)$.
4. $p(\sigma^2|Y, X, K, \tau^2)$ que se puede reescribir como $p(\sigma^2|Y, G, K)$.

Este muestreador que se propone, a diferencia del Muestreador de Gibbs, la variable K_i es generada sin estar condicionada a la variable de estados

X. Es decir, que se pueden hacer modificaciones a los muestreadores de tal manera que se adapten al modelo a través del tiempo manteniendo la estructura MCMC.