ESTIMACIÓN BAYESIANA DEL MODELO POISSON OCULTO DE MARKOV PARA UNA SERIE TEMPORAL DE NÚMERO DE TERREMOTOS *

Rafael Eduardo Díaz *Universidad Santo Tomás*

Resumen
Los modelos de Markov son una técnica de modelización altamente flexible que en los últimos 30 años, se han utilizado en diversos campos, generalmente para el reconocimiento de patrones, se han utilizado en traducción automática, criptoanálisis, bioinformática, genética, finazas entre otros. En esta revison se muestran los elementos esencial del PHMM, se presentan algunos algoritmos de estimación, propios de los métodos MCMC utilizados regularmente en la práctica, como el algoritmo EM, las recursiones forward - Backward, haciendo especial énfasis en el muestreador de Gibbs ampliamente utilizado en estadística Bayesiana. Finalmente se realiza un ejercicio de aplicación a datos reales donde se sugiere el uso de un modelo Poisson Oculto de Markov para el ajuste a la serie de datos sobre el número de terremotos ocurridos en el mundo con magnitud 7 o mayor desde el año de 1900 al 2006. Se muestra el ajuste del modelo, la estimación de los parámetros, la convergencia de la cadena y la validación de los supuestos.

Palabras Clave: Modelos ocultos de Markov, recursion forward-backward, muestreador de Gibbs.

1. Introducción

Un modelo de Markov oculto (HMM) es un modelo de mixto cuya distribución es una cadena de Markov de estado finito. El uso de los modelos de Ocultos de Markov se remonta a finales de la década de 1960 donde fue introducido por (Baum et al., 1970); a mediados de los años de 1970 aparecen las primeras aplicaciones en el reconocimiento de la voz. Más tarde en los $80's$ los HMMs son utilizados para el análisis de secuencias genéticas (Churchill, 1989), y en muchas otras aplicaciones, relacionadas con la bioinformática. Pero no es sino hasta los últimos 30 años donde se ha masificado en otras áreas, como la economía (Hamilton, 1989), el análisis de imágenes (Romberg, Choi and Baraniuk, 2001) y en general en el reconocimiento de patrones.

En este documento se presenta un análisis sobre un tipo especial de HMMs, es el Modelo Poisson Oculto de Markov (PHMM), originalmente desarrollado y utilizado en el campo de biométrica, (ver Albert and Chib, 1993; también Leroux and Puterman, 1992). Contrario al proceso de Poisson el cual se caracteriza por tener intensidad constante a lo largo del tiempo es decir media igual a su varianza, en los datos de conteo con varianza mayor a su media es decir sobre dispersión se considera que la intensidad Poisson no es constante, sin embargo tienen una distribución dada, en este caso podemos modelar los datos de conteo sobre dispersos con una mixtura de modelos asumiendo dependencia entre estos, pero con independencia entre las observaciones, es decir un Modelo Poisson Oculto de Mrkov (PHMMs).

En este artículo sugerimos el uso de un PHMM, para modelar el número de terremotos mayores (magnitud 7 o mayor en la escala de Richter) ocurridos en el mundo desde el año de 1900-2006. Asumimos un proceso estocástico de tiempo discreto $\{(X_t, Y_t)\}_{t\in\mathbb{N}}$ donde $\{X_t\}_{t\in\mathbb{N}}$ son los

*Rafael Eduardo Díaz. email: rafael.diaz@usantotomas.edu.co

estados-finitos ocultos o no observados de la cadena de Markov y $\{Y_t\}_{t\in\mathbb{N}}$ es la secuencia de años t del numero de terremos tal que Y_t dado un estado X_t es, para cada t, una variable aleatoria Poisson, cuyos parámetros dependen del estado de X_t.

2. Modelos Ocultos de Markov

Los HMM nos permiten modelar la dinámica de un sistema (oculto), al cual no podemos acceder (observar) de forma directa; por el contrario de forma indirecta mediante la observación de eventos externos, suponemos que están correlacionados con dicho sistema y su estado. En las *cadenas de Markov* [1], las señales observadas corresponden a los estados del modelo, mientras que en en los *modelos ocultos de Markov* no se conoce la secuencia que de estados por la que pasa el modelo, sino una función probabilística de ella. Existen diversas razones por los cuales el sistema no es accesible de forma directa, como la imposibilidad física o la presencia de ruido en la medición.(Rabiner, 1990)

De forma general definimos un HMM, como un modelo probabilístico, utilizado para representar la probabilidad conjunta de un conjunto de variables aleatorias (Bilmes, 1998). En este conjunto de variables aleatorias distinguimos dos tipos. El primero corresponde a los posibles eventos o símbolos observables Y_t, que pueden presentarse al realizar una observación indirecta del sistema oculto. El segundo corresponde al estado en el cual se encuentra el sistema oculto X_t durante una observación.

Las variables aleatorias de observación, puede ser bien discretas $Y = y \in S = \{1, 2, ..., N\}$, o continuas. La medida de probabilidad en cada caso estará definida, bien por una función de masa de probabilidad (pmf) o por una función de densidad de probabilidad (pdf) de tipo gaussiano generalmente (Rabiner, 1990). Las variables aleatorias de estado oculto son discretas y finitas , pero variantes como los HMM infinitos (Beal, Ghahramani and Rasmussen, 2002), permiten superar esta restricción. Con base en estos dos tipos, son construidas secuencias de variables aleatorias tanto de observaciones, como de estados ocultos del sistema. De esta forma, el par (Y_t, X_t), representa la posible historia dinámica del sistema oculto. La probabilidad de una determinada secuencia de estados ocultos es calculada empleando probabilidades de transición entre los estados, siguiendo un proceso de Markov (Ephraim and Merhav, 2006). En este proceso la probabilidad de transición de estado, asume invariancia en el tiempo y dependencia frente a los m estados anteriores. Para el caso $m = 1$, tenemos un HMM de primer orden y su proceso es descrito mediante cadenas de Markov condicionales. Los HMM de primer orden son tradicionalmente los más usados. La razón de esto, deriva en la simplificación de los cálculos, mediante el empleo de técnicas de programación dinámica, explotadas por algoritmos como *forward*, *backward*, *Viterbi* y *Baum-Welch*. Definimos un HMM de primer orden denotado como μ, mediante la tripla $\{\pi, \Gamma, B\}$

- π es el vector de probabilidades iniciales. π_i es $P(X_1 = S_i)$.
- Γ: es la matriz de probabilidades de transición. Con $\gamma_{i,j}$: es $P(X_{t+1} = S_j | X_t = s_i)$
- B es la matriz de probabilidad de difusión Σ. Donde $b_{j,k} = P(Y_t = y | X_t = s_j)$.

Se denota una secuencia de estados $X = (X_1, ..., X_{T+1})$ donde $X_t : S \to \{1, ..., N\}$ y una secuencia de observaciones $Y_t = (y_1, ..., y_T)$ con $y_T \in \Sigma$.

[1] Un proceso de Markov es un proceso estocástico que sirve para representar secuencias de variables aleatorias no independientes entre sí. Es decir, donde la probabilidad del siguiente estado sobre una secuencia completa depende de estados previos al estado actual Prof. Gloria Inés Alvarez V.. Definición: Sea $X = (X_1, ..., X_t)$una secuencia de variables aleatorias que toman valores en un conjunto finito $S = (s_1, ..., s_N)$ que se llama espacio de estados. Entonces, las propiedades de Markov son: 1. Horizonte Limitado: $P(X_{t+1} = s_k | X_1, ., X_t) = P(X_{t+1} = s_k | X_t)$. 2. Invariante en el Tiempo: $P(X_{t+1} = s_k | X_1, ..., X_t) = P(X_2 = s_k | X_1)$.

3. Modelo Poisson Oculto de Markov

Los Modelos Poisson Ocultos de Markov son un tipo especial de Modelos Ocultos de Markov (HMMs), que son procesos estocásticos de tiempo discreto $\{(X_t; Y_t)\}_{t\in N}$ tal que X_t es una cadena de Markov de estado finito no observable y $\{Y_t\}_{t\in\mathbb{N}}$ es una secuencia de variables aleatorias dependientes sobre $\{X_t\}_{t\in\mathbb{N}}$. Esta dependencia se modela suponiendo que la distribución condicional de cada observación Y_t, dada la secuencia X_t, Depende únicamente del proceso actual no observado X_t (condición de dependencia contemporánea); además, dado $\{X_t\}_{t\in\mathbb{N}}$, $\{Y_t\}_{t\in\mathbb{N}}$ es una secuencia de variables aleatorias condicionalmente independientes (condición de independencia condicional). Si asumimos que, para cada t, Y_t dado un estado de X_t es una variable aleatoria de Poisson, tenemos los denominados modelos Poisson ocultos de Markov. En este caso, X_t determina el parámetro Poisson utilizado para generar Y_t. Introducción de algunas nociones y suposiciones. Asumimos que el proceso no observado $\{X_t\}_{t\in\mathbb{N}}$ es una cadena de Markov discreta, homogénea, aperiódica e irreducible en un espacio de estados finitos $S_X = 1, 2, ., m$ (para más detalles sobre cadenas de Markov, mire, por ejemplo, Grimmett and Stirzaker, 2001); denotamos con γ_{ij} la probabilidad de transición del estado i, en el tiempo $t-1$, al estado j, en el tiempo t (para algún estado i, j y para algún tiempo t), i.e.: $\gamma_{ij} = P(X_t = j|X_{t-1} = i)$. Sea $\Gamma = [\gamma_{ij}]$ la matriz de transición de probabilidades $m \times m$, con $\sum_{j\in S_X} \gamma_{ij} = 1$, con algún $i \in S_X$. La distribución marginal de X_1 es la distribución inicial denotada por $\delta = (\delta_1, ..., \delta_m)$, con $\delta_i = P(X_1 = i)$, para algún $i = 1, 2, ..., m$, y $\sum_{j\in S_X} \delta_i = 1$; es una consecuencia inmediata del supuesto sobre la cadena de Markov $\{X_t\}_{t\in\mathbb{N}}$,δ es la distribución estacionaria y la igualdad $\delta' = \delta'\Gamma$ se mantiene; i.e. la parte izquierda δ, es el vector propio de la matriz Γ, asociada con el valor propio 1, que existe siempre ya que Γ es una matriz estocástica (ver Guttorp and Minin, 1995, p.19). Consideremos ahora la secuencia observada $\{Y_t\}_{t\in\mathbb{N}}$. En un PHMMs, cualquier variable observada Y_t condicionada sobre X_t es Poisson para cualquier t; cuando X_t es un estado $i(i \in S_X; t \in \mathbb{N})$, entonces la distribución condicional de Y_t es una variable aleatoria con parámetro λ_i; para algún $y \in \mathbb{N}$, las probabilidades de los estados dependientes están dadas por:

$$p(y_i) = P(Y_t = y|X_t = i) = e^{-\lambda_i}\frac{\lambda_i^y}{y!}$$

Con $\sum_{y\in\mathbb{N}} p(y_i) = 1$ para cada $i \in S_X$. Dado que $\{X_t\}_{t\in\mathbb{N}}$ es un proceso fuertemente estacionario también el proceso observado $\{Y_t\}$ es fuertemente estacionario; por lo tanto, para cada t de Y_t, tiene la misma distribución marginal:

$$P(Y_t = y) = \sum_{i\in S_X} P(Y_t = y, X_t = i) = \sum_{i\in S_X} P(y_t = y|X_t = i) = \sum_{i\in S_X} \delta_i p(y_i)$$

Que es una mixtura finita de distribuciones de Poisson. Además, se puede demostrar fácilmente que el valor esperado de Y_t, para cada t, viene dado por:

$$E(Y_t) = \sum_{i\in S_X} \delta_i\lambda_i$$

Finalmente, observamos que, las variables Y_t están sobredispersadas, tal que la varianza es mayor que la media; De hecho, se tiene que: $V(Y_t) = \lambda' D\lambda + \delta'\lambda - (\delta'\lambda)^2 > E(Y_t) = \delta'\lambda$, para algún t, con $\lambda = (\lambda_1, ..., \lambda_m)'$ y $D = diag(\delta)$. (ver MacDonald and Zucchini, 2009)

3.1. Los tres problemas fundamentales en HMM

Existen tres problemas básicos al emplear HMM (Rabiner, 1990), los cuales son:

1. Problema de la evaluación:dada una secuencia de observaciones $Y = (y_1, ..., y_N)$ y un HMM $\mu = (\pi, \Gamma, B)$, determinar $P(Y|\mu)$. Los algoritmos: *forward* o *backward* son comúnmente utilizados en su solución.

2. Problema de la decodificación: dada una secuencia $Y = (y_1, ..., y_N)$, y un HMM $\mu = (\pi, \Gamma, B)$, encontrar la secuencia de estados ocultos $X^m = \{x_1^m, ..., x_N^m\}$, tal que:

$$X^m = \underset{X^i}{max} P(X^i|\mu, Y)$$

   Su solución se obtiene mediante el algoritmo de Viterbi [@viterbi1967].

3. Problema del aprendizaje: dada una secuencia de observaciones $Y = (y_1, ..., y_N)$ determinar los parámetros del modelo $\mu^* = (\pi, \Gamma, B)$, tal que:

$$\mu^* = \underset{\mu^i \in \Omega}{max} P(Y|\mu^i)$$

   Donde Ω corresponde al espacio de parámetros en la topología del HMM particular. El algoritmo empleado tradicionalmente para su solución es el algoritmo *Baum-Welch* o también conocido como *forward-backward*.

4. Objetivos

- Obtener la distribución posterior de los parámetros del modelo.
- Estimar los parámetros del modelo, por estimación vía máxima verosimilitud (Algoritmo EM) algoritmo Baum-Welch y con técnicas bayesianas (muestreador de Gibbs).
- Utilizar diagnósticos para evaluar la validez del modelo y la convergencia MCMC.

5. Hipótesis a Verificar

En el presente trabajo, se pretende verificar que la serie de datos sobre el número de terremotos mayores se ajusta a una distribución poisson sobredispersada, con dependencia serial. Además se quiere la normalidad de los residuales. Si se cumplen las condiciones se ajustara un modelo poisson oculto de Markov (PHMMs).

6. Descripción de los datos

Los datos sobre el número de terremotos mayores (es decir de magnitud 7 o mayor en la escala de Richter) en el mundo, para los años 1900-2006 fueron obtenidos de del Servicio Geológico de los Estados Unidos o USGS (United States Geological Survey) Para esta serie, la aplicación de modelos estándar como modelos auto regresivos de media móvil (ARMA) sería inapropiado, ya que estos modelos se basan en la distribución normal. En cambio, se propone un modelo usual para los conteos con distribución de Poisson, pero, como se demostrará más adelante, la serie presenta una sobredispersión considerable con respecto a la distribución de Poisson, y fuerte dependencia serial positiva. Por lo tanto, un modelo que consiste en variables aleatorias independientes de Poisson; sería por dos razones inadecuado. Observando la tabla 1 sugiere que puede haber algunos períodos con una baja tasa de terremotos, y algunos con una tasa relativamente alta. Los HMMs, que permiten que la distribución de probabilidad de cada observación dependa del estado no observado (u oculto) de una cadena de Markov, puede acomodar tanto la sobredispersión como la dependencia en la serie.

```
dire <- "http://www.hmms-for-time-series.de/second/data/earthquakes.txt"
Terremotos <- read.table(dire)$V2;Terre <- read.table(dire)
stargazer::stargazer(rbind(Terremotos[1:20],Terremotos[21:40],Terremotos[41:60],
Terremotos[61:80],Terremotos[81:100],c(Terremotos[101:107],rep("",13))),
header = FALSE,column.sep.width="1pt",
title = "Número de terremotos (magnitud 7 o mayor) en el mundo, 1900 - 2006.")
```

Table 1: Número de terremotos (magnitud 7 o mayor) en el mundo, 1900 - 2006.

13	14	8	10	16	26	32	27	18	32	36	24	22	23	22	18	25	21	21	14
8	11	14	23	18	17	19	20	22	19	13	26	13	14	22	24	21	22	26	21
23	24	27	41	31	27	35	26	28	36	39	21	17	22	17	19	15	34	10	15
22	18	15	20	15	22	19	16	30	27	29	23	20	16	21	21	25	16	18	15
18	14	10	15	8	15	6	11	8	7	18	16	13	12	13	20	15	16	12	18
15	16	13	15	16	11	11													

6.1. Estadísticas de Resumen

A continuación se muestran algunas estadísticas descriptivas, sobre la serie de terremotos mundiales para los años 1900-2006.

```
colnames(Terre) <- c("A","Conteos");Terre$A <- as.character(Terre$A)
stargazer::stargazer(Terre, header = F,median = T, mean.sd = T,
          title = "Estadísticas de Resumen")
```

Table 2: Estadísticas de Resumen

Statistic	N	Mean	St. Dev.	Min	Median	Max
Conteos	107	19.364	7.181	6	18	41

Como vemos, el número mínimo de terremotos de magnitud siete o mayor ocurrido hasta el momento es de 6, que corresponde al año 1986, mientras que el máximo de terremotos registrados fue de 41 en el año 1943, por otra parte la media para terremotos mayores por año corresponde a 19 aproximadamente. Como se observa en la tabla 1, estos terremotos se han mantenido constantes esto puede ocasionar bastante preocupación a nivel mundial, especialmente en América que cuenta con fallas como la de San Andrés y las zonas costeras al Océano Pacifico donde se pueden ocasionar Tsunamis. La serie de tiempo no parece mostrar una tendencia clara, sin embargo los picos más alto parecen estar entre los años 1940 a 1955.

```
par(mar=c(2,2,1,.5)+.5, mgp=c(1.6,.6,0))
plot(ts(data = Terremotos, start = 1900),xlab="",type='o', col=4,
ylab = "Número", main = "Serie de tiempo terremotos")
rect(par("usr")[1], par("usr")[3],par("usr")[2],par("usr")[4],col=gray(.9,.9),
     border='white');grid(lty=1, col='white')
lines(ts(data = Terremotos, start = 1900), type='o', col=4)
```

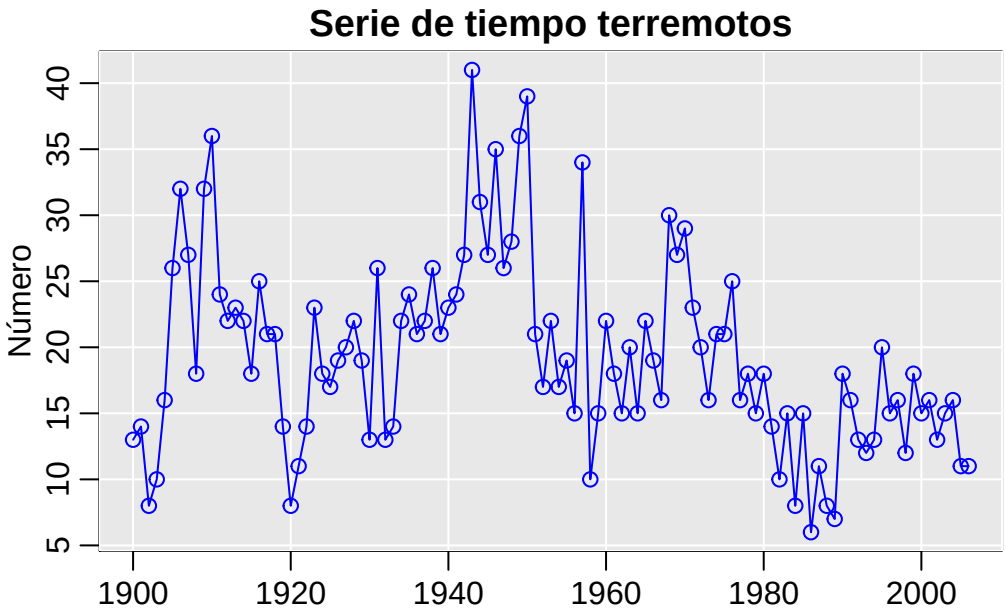


Figure 1: Número de terremotos (magnitud 7 o mayor) en el mundo, 1900 - 2006.

En la figura 2 la función de autocorrelación muestral (ACF), indica dependencia de los datos por lo tanto una mixtura de modelos independiente no servirá para la serie de terremotos pues, por definición, no permite la dependencia en las observaciones de la serie. Una manera de permitir la dependencia en las observaciones de la serie es debilitar el supuesto de que los parámetros del proceso son independientes en la serie. Una forma sencilla y matemáticamente conveniente de hacerlo es asumir que es una cadena de Markov. El modelo resultante de las observaciones se denomina un modelo Poisson oculto de Markov (PHMMs).

```
par(mar=c(2,2,1,.5)+.5, mgp=c(1.6,.6,0))
acf(ts(data = Terremotos, start = 1900),xlab="",main="")
```

Como se había mencionado anteriormente, los datos parecen ajustarse una distribución poisson con media $\bar{x} \approx 19.4$ y varianza muestral $s^2 \approx 51.6$, por lo que no se cumple la propiedad de esta distribución de que su media es igual a su varianza, indica la sobre dispersión de los datos, además de algunas barras que sobresalen más de lo normal. Una posible metodología seria ajustar una distribución bimodal o un modelo mixto, que pueda constituir grupos no observados de la población. Es decir una mixtura de distribuciones Poisson con medias $\lambda_1, ..., \lambda_m$. La escogencia de la media, estará dada por un proceso aleatorio, en el cual la media λ_i es seleccionada con probabilidad δ_i, con $i = 1, ., m$.

```
par(mar=c(2,2,1,.5)+.5, mgp=c(1.6,.6,0))
plot(prop.table(table(Terremotos)), xlim = c(0,45),xlab = "",ylab = "frecuencias",
main = "Gráfico de Frecuencias con ajuste Poisson")
rect(par("usr")[1], par("usr")[3],par("usr")[2],par("usr")[4],col=gray(.9,.9),border='white')
grid(lty=1, col='white')
points(0:45,dpois(0:45,19), pch = 20)
lines(prop.table(table(Terremotos)))
```

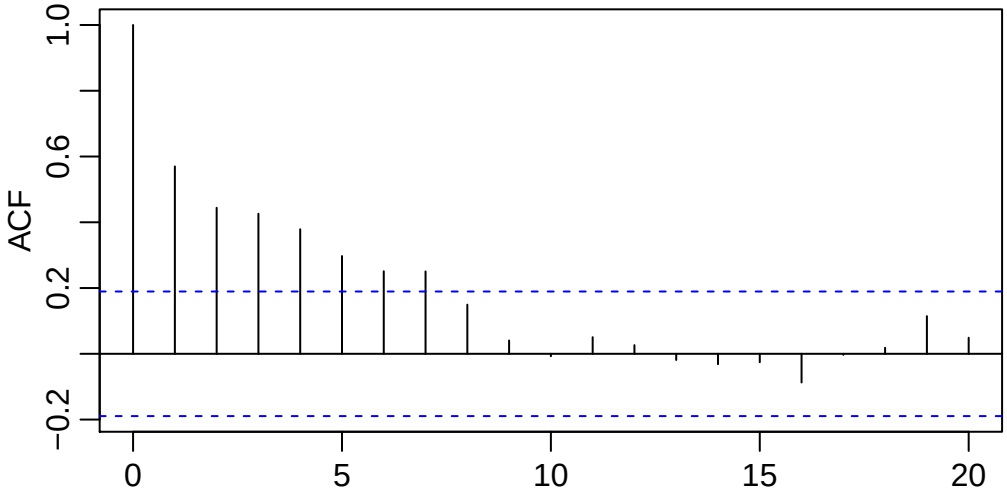


Figure 2: Serie de terremotos: función de autocorrelación de la muestra.

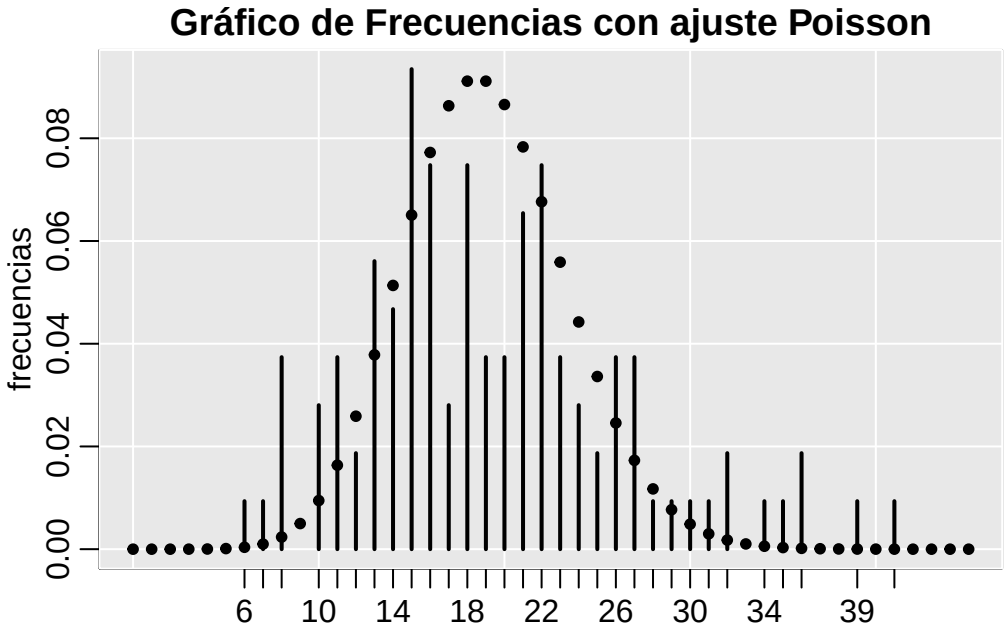


Figure 3: Diagrama de barras de las frecuencias relativas de los conteos

7. Inferencia Clásica

En esta sección se discute algunas metodologías usuales para el cálculo de los parámetros de los modelos ocultos de Markov, es decir las probabilidades de transición y las probabilidades de los estados dependientes. Usualmente en los modelos ocultos de Markov, se presentan dos problemas la estimación de los parámetros dada una secuencia de observaciones, y una vez estimados estos parámetros, obtener la correspondiente sucesión de estados ocultos.

El primer problema puede ser resuelto mediante la estimación máxima verisimilitud que proporciona los valores de los parámetros que maximizan la verosimilitud, sin embargo por la dificultad matemática de hallar esta estimación, se plantean métodos computacionales como el algoritmo EM (**?**), sin embargo a menudo se utiliza el algoritmo Baum - Welch (Baum et al., 1970), que es una versión particular del algoritmo EM pero más eficiente computacionalmente hablando. El segundo problema de obtener la sucesión de estados ocultos puede ser resuelto mediante al algoritmo Viterbi (Viterbi, 1967), que es un caso especial de algoritmos de inferencia aplicables a modelos gráficos.

7.1. Estimación de los parámetros

El PHMM descrito en la sección 3 depende del siguiente conjunto de parámetros: la distribución estacionaria inicial $\delta = (\delta_1, ..., \delta_m)'$, las probabilidades de transición $\gamma_{ij}(i, j \in S_X)$ y las probabilidades de los estados - dependientes $p(y_i)(y \in \mathbb{N}; \in S_X)$. Ahora buscamos algunos estimadores de estos parámetros. En particular, buscamos los estimadores de máxima verosimilitud de las $m^2 - m$ probabilidades de transición γ_{ij} con $i \neq j$, i.e. los elementos fuera de la diagonal de la matriz γ (los elementos diagonales se obtienen por diferencia, ya que cada fila de γ suma a uno: $\gamma_{ij} = 1 - \sum_{j \in S_X} \gamma_{ij}$, para algún $i \neq j \in S_X$ y el estimador de máxima verosimilitud para los m parámetros λ_i de la distribución Poisson, entrando las probabilidades de los estados-dependientes $p(y_i)$. Usando la matriz estimada γ, entonces obtenemos el estimador de la distribución inicial a partir de δ desde la igualdad $\delta'\Gamma$ (siendo δ la distribución estacionaria). Sea θ denota el vector de los parámetros desconocidos a estimar con el método de máxima verosimilitud,

$$\theta = (\gamma_{1,2}, \gamma_{1,3}, ...\gamma_{m,m-1}, \lambda_1, ..., \lambda_m)'$$

y sea Θ el espacio de parámetros. Sea $y = (y_1, ..., y_T)'$ un vector de los datos observados i.e. la secuencia de las T realizaciones del proceso estocástico $\{Y_t\}_{t \in \mathbb{N}}$. Sea $x = (i_1, ..., i_T)'$ el vector de estados no observados de la cadena $\{X_t\}_{t \in \mathbb{N}}$; por lo tanto $(i_1, y_1, ..., i_T, y_T)'$ es el vector de los datos completos. La función de verosimilitud de los datos $L_T(\theta)$ esta definida como la probabilidad conjunta de las T observaciones y los T estados no observados. Aplicando las propiedades de Markov de dependencia, independencia condicional y condiciones de dependencia contemporáneas, obtenemos fácilmente:

$$L_T(\theta) = P(Y_1, ..., Y_T = y_T) = \delta_{i1} p(y_{i_1}) \prod_{t=2}^{T} \gamma_{i_{t-1}, it} p(y_{it})$$

Donde $p(y_{it})$ es la probabilidad de los estados dependientes y_t condicionado sobre el estad $i_t (t = 1, ..., T;$

$$p(y_{it}) = e^{-\lambda_{it}} \frac{\lambda_{it}^{y_t}}{y_t!} \tag{1}$$

Para encontrar el estimador de máxima verosimilitud de θ deberíamos resolver el sistema de verosimilitud, pero es muy difícil encontrar analíticamente la solución, entonces debemos utilizar un algoritmo numérico. Para realizar el algoritmo EM (MacDonald and Zucchini, 2009), se basa en un procedimiento iterativo con dos pasos en cada iteración: el primer paso, paso **E**, proporciona el cálculo de una Expectativa; El segundo, paso **M**, proporciona una Maximización.

Sea $Q(\theta;\theta)'$ la función definida en el paso **E**:

$$Q(\theta;\theta)' = E_{\theta'}(logL_T(\theta)|y)$$

para algún vector θ perteneciente al espacio de parámetros Θ. En Dempster, Laird y Rubin (**?**) se demuestra que una condición suficiente para maximizar L_T es máximizar $Q(\theta;\theta)'$ con respecto a *theta*. Sin entrar en detalles, el esquema iterativo del algoritmo EM es el siguiente. Sea $\theta^{(k)}$ el vector estimado obtenido en el k^{th} iteración.

$$\theta^{(k)} = (\gamma_{1,2}^{(k)}, \gamma_{1,3}^{(k)}, ..., \gamma_{m,m-1}^{(k)}, ..., \lambda_m^{(k)})'$$

en la $(k+1)^{th}$ iteración, el E y M pasos son definidos como sigue:

- Paso E - dar $\theta^{(k)}$, calcular

$$Q(\theta;\theta^{(k)}) = E_{\theta^{(k)}}(logL_T(\theta)|y)$$

- Paso M - buscar $\theta^{(k+1)}$, para que maximice $Q(\theta;\theta^k)$ i.e. tal que

$$Q(\theta^{(k+1)};\theta^{(k)} \geq Q(\theta;\theta^{(k)})$$

para algún $\theta \in \Theta$.

Los pasos E y M deben repetirse de una manera alterna hasta que la secuencia de valores de la log-verosimilitud $\left\{logL_T(\theta^{(k)})\right\}$ converja, es decir, hasta que la diferencia:

$$logL_T(\theta^{(k+1)}) - logL_T(\theta^{(k)})$$

Sea menor o igual que un valor arbitrario ϵ. Cuando se cumplen ciertas condiciones de regularidad en el espacio de parámetros Θ y en las funciones $L_T(\theta)$ y $Q(\theta;\theta)'$ son satisfechas (ver Wu, 1983, pag. 94-96), nosotros podemos decir que, si el algoritmo converge en la $(k+1)^{th}$ iteración entonces $\left(\theta^{k+1}; logL_T(\theta^{k+1})\right)$ es un punto estacionario y $\theta^{(k+1)} = (\gamma_{1,2}^{(k+1)}, \gamma_{1,3}^{(k+1)}, ..., \gamma_{m,m-1}^{(k+1)}, ..., \lambda_m^{(k+1)})'$ es el estimador de máxima verosimilitud de los parámetros desconocidos θ. En el PHMMs, una condición suficiente para que las condiciones de Wu se mantengan es que los parámetros de Poisson $\lambda_i (i = 1, 2, ..., m)$ sean estrictamente positivos. Para HMMs, la superficie de la log-verosimilitud es irregular y se caracteriza por muchos máximos locales o puntos estacionarios; entonces, el punto stadionario al que converge el algoritmo EM no puede ser el máximo global. Por lo tanto, para identificar el máximo global, la elección del punto de referencia es de vital importancia.

Implementando el algoritmo, la búsqueda de los estimadores de los parámetros desconocidos con el algoritmo EM puede simplificarse usando las probabilidades *forward* y *backward*, introducidas por (Baum et al., 1970). La probabilidad *forward*, denotada por α_i, es la probabilidad conjunta del pasado y las presentes observaciones y el estado actual de la cadena:

$$\alpha_t(i) = P(Y_1 = y_1, Y_2 = y_2 ..., X_t = i)$$

mientras que las probabilidades *backward*, denotadas po $\beta_t(i)$, es la probabilidad de las observaciones futuras condicionado sobre estado actual de la cadena:

$$\beta_t(i) = P(Y_{t+1} = y_{t+1}, ..., Y_T = y_T | X_t = i).$$

Las probabilidades $\alpha_t(i)$ y $\beta_t(i)$ se pueden obtener recursivamente como sigue:

$$\begin{aligned} \alpha_1(i) &= \delta p(y_1); i = 1, ..., m, \\ \alpha_t(j) &= \left(\sum_{i \in S_X} \alpha_{t-1}(i) \gamma_{i,j} p(y_{t,j}) \right); j = 1, ..., m \end{aligned} \tag{2}$$

para las probabilidades *forward* y

$$\begin{aligned} \beta_T(i) &= 1; i = 1, ..., m, \\ \beta : i(i) &= \sum_{j \in S_X} p(y_{t+1,j}) \beta_{t+1}(j) \gamma_{i,j}; t = T-1, ..., 1; i = 1, ..., m, \end{aligned} \tag{3}$$

para las probabilidades *backward* (ver MacDonald and Zucchini, 2009, pag. 66-67). Entonces, obtenenmos la siguiente expresión para la función $Q(\theta; \theta^{(k)})$ en el paso E de la $(k+1)^{th}$, iteración del algoritmo EM

$$\begin{aligned} Q\left(\theta; \theta^{(k)}\right) &= E_{\theta^{(k)}}(log L_T(\theta) | y) \\ &= \sum_{i \in S_X} \frac{\alpha_1^{(k)}(1) \beta_1^{(k)}(i)}{\sum_{i \in S_X} \alpha_t^{(k)}(l) \beta_t^{(k)}(l)} log \delta_i + \sum_{i \in S_X} \sum_{j \in S_X} \frac{\sum_{t=1}^{T-1} \alpha_t^{(k)}(i) \gamma_{i,j}^{(k)} p(y_{t+1,j}) \beta_{t+1}^{(k)}(j)}{\sum_{t \in S_X} \alpha_t^{(k)}(l) \beta_t^{(k)}(l)} log \gamma_{i,j} \\ &+ \sum_{i \in S_X} \frac{\sum_{i=1}^{T} \alpha_1^{(k)}(1) \beta_1^{(k)}(i)}{\sum_{i \in S_X} \alpha_t^{(k)}(l) \beta_t^{(k)}(l)} log p(y_{t,i}) \end{aligned} \tag{4}$$

donde $p(y_{t,i}^{(k)}), \alpha_t^{(k)}$ y $\beta_t^{(k)}(i)$ son calculados de acuerdo a las formulas (1),(2) y (3), respectivamente, usando los valores del parámetro obtenido en la k^{th} iteración; mientras $\delta^{(k)}$ es calculado con $\delta'^{(k)} = \delta'^{(k)} \Gamma^{(k)}$. Debe notarse que δ, por el supuesto de estacionariedad, contiene información sobre la matriz de probabilidad de transición Γ ya que $\sum_{i \in S_X}$, para cualquier $j \in S_X$. Sin embargo, para T grande, el efecto, de δ es insignificante (ver Basawa, 2014). Además, en el paso M de la $(k+1)^{th}$ iteración, obtenemos $\theta^{(k+1)}$, podemos ignorar el primer apéndice en (4) al maximizar $Q\left(\theta; \theta^{k}\right)$ con respecto a los $m^2 - m$ paramétros $\gamma_{i,j}$s.

La expresión para el estimador de máxima verosimilitud de $\gamma_{i,j}$ obtenidos en la $(k+1)^{th}$ iteración de el algoritmo EM esta dado por:

$$\gamma_{i,j}^{(k+1)} = \frac{\sum_{t=1}^{T-1} \alpha_t^{(k)}(i) \gamma_{i,j}^{(k)} p(y_{t+1,j}) \beta_{t+1}^{(k)}(j)}{\sum_{t=1}^{T-1} \alpha_t^{(k)}(i) \beta_t^{(k)}(i)}, \tag{5}$$

para cualquier estado i y cualquier estado $j, j \neq i$, de la cadena de Markov $\{X_t\}$. El estimador de maxima verosimilitud de λ_i obtenido de la $(k+1)^{th}$ iteración de el algoritmo EM,esta dada

por:[2]

$$\lambda_i^{(k+1)} = \frac{\sum_{t=1}^{T} \alpha_t^{(k)}(i)\beta_t^{(k)}(i)y_t}{\sum_{t=1}^{T} \alpha_t^{(k)}(i)\beta_t^{(k)}(i)}, \tag{6}$$

para algún estado i de la cadena de Markov $\{X_t\}$. Leroux (1992) y Bickel, Ritov, Rydén (1998) provaron que los estimadores en (5) y en (6) son consistentes y asintóticamente normales.

7.2. Aplicación del algoritmo EM al PHMM

Ahora ajustamos el modelo de tres estados por el algoritmo EM, como se ha descrito anteriormente, a los datos de los terremotos. Para el modelo de tres estados, los valores iniciales de las probabilidades de transición fuera de la diagonal se toman como 0.1, y el resto, es el complemento dado que la suma de las filas debe ser uno. El valor inicial de δ, es (0.5,0.5,0.5) dando igual peso a todos componentes. Dado que la media muestral fue 19.36, 10,20 y 30 son valores iniciales plausibles. para los estados dependientes de las medias λ_1 y λ_2. A continuación se muestra los resultados obtenidos y con el algoritmo Viterbi, se muestra la secuencia más probable de los estados dado el conjunto de datos observados.

```
library(HiddenMarkov)
# 4.1 Crea un objeto de modelo de Markov ocultos de tiempo discreto con la función "dthmm".
Pi <- rbind(c(.8,.1,.1),c(.1,.8,.1),c(.1,.1,.8))
mod.HM <- dthmm(Terremotos,Pi=Pi,delta=c(1/3,1/3,1/3),"pois", list(lambda=c(10,20,30)))
```

```
ajuste.mod.HM <- BaumWelch(mod.HM)
```

Sean las estimaciones del vector de probalidades $\delta = (1, 0, 0)$, para el vector de medias las estimaciones son $\lambda = (13.134, 19.714, 29.711)$ y matriz de transición.

$$\mathbf{\Gamma} = \begin{pmatrix} 0.9392 & 0.0321 & 0.0285 \\ 0.4042 & 0.9063 & 0.0532 \\ 0.1881 & 0.1904 & 0.8095 \end{pmatrix}$$

```
Vit <- Viterbi(ajuste.mod.HM)
stargazer::stargazer(rbind(Vit[1:20],Vit[21:40],Vit[41:60],Vit[61:80],Vit[81:100],
                     Vit[101:107],rep("",13)),header = FALSE,column.sep.width="1pt",
                 title = "Calculo de los estados más probables. Algoritmo Viterbi.")
```

8. Selección y Verificación del Modelo

En el HMM básico con m estados, el aumento de m siempre mejora el ajuste del modelo (juzgado por la verosimilitud). Pero junto con la mejora viene un aumento cuadrático en el número de parámetros, y la mejora en el ajuste tiene que ser negociado contra este aumento. Por lo tanto, se necesita un criterio para la selección del modelo. En algunos casos, es razonable reducir el número de parámetros haciendo suposiciones sobre las distribuciones dependientes del estado o

[2]la formula de $\lambda_i^{(k+1)}$ es fácilmente obtenida derivando $Q(\theta;\theta^{(k)})$ en (4) con respecto a λ_i y poniendo la derivada igual a 0.

Table 3: Calculo de los estados más probables. Algoritmo Viterbi.

1	1	1	1	1	3	3	3	3	3	3	2	2	2	2	2	2	2	1
1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	3	3	3	3	3	3	3	3	3	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	3	3	3	2	2	2	2	2	2	2	2
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

de la cadena de Markov (MacDonald and Zucchini, 2009, pag.97). En esta sección tambien se dara un breve resumen de la selección de modelos en HMMs, y luego se describira el uso de pseudo-residuos para verificar las deficiencias en el modelo seleccionado.

8.1. Selección del modelo por AIC y BIC

Un problema que surge naturalmente cuando se usan modelos de Markov ocultos (u otros) es el de seleccionar un modelo apropiado (por ejemplo, de elegir el número apropiado de estados m, a veces descrito como el **orden** del HMM. La estimación de orden para un HMM no es ni trivial ni establecida (ver cap. 15, MacDonald and Zucchini, 2009), por lo tanto se necesita algún criterio para la comparación de modelos. El material descrito a continuación se basa en (Zucchini, 2000), que da una descripción introductoria de la selección del modelo. Celeux y Durand [-celeux2008] presentan y discuten varias técnicas de selección de modelos para seleccionar el número de estados en un HMM.

Describimos los dos enfoques más populares para la selección de modelos. En el enfoque frecuencial se selecciona la familia estimada más cercana a la modelo operativo.

Para ello se define una discrepancia (una medida de la *falta de ajuste*) entre los modelos operativos y los modelos $\Delta(f, \hat{g_1})$ y $\Delta(f, \hat{g_2})$. Estas discrepancias dependen del modelo operativo f, que es desconocido, por lo que no es posible determinar cuál de las dos discrepancias es menor, es decir, qué modelo debe ser seleccionado. En su lugar, se basa la selección en estimadores de las discrepancias esperadas, $\hat{E_f}(\Delta(f, \hat{g_1}))$ y $\hat{E_f}(\Delta(f, \hat{g_2}))$, que se denominan criterios de selección de modelos. Al elegir la discrepancia de Kullback-Leibler, y bajo las condiciones listadas en el Apéndice A de Linhart y Zucchini (1986), el criterio de selección del modelo simplifica el criterio de información Akaike (AIC):

$$AIC = -2logL + 2p$$

Donde log L es la log-verosimilitud del modelo ajustado y p denota el número de parámetros del modelo. El primer término es una medida de ajuste y disminuye con el número creciente de estados m. El segundo término es un término de pena, y aumenta con el aumento m.

El enfoque bayesiano de la selección de modelos, selecciona la familia cuya estimación sea la más probablemente verdadera. En una primera etapa, antes de considerar las observaciones, se especifican los priores, que son las probabilidades $Pr(f \in G_1)$ y $Pr(f \in G_2)$ tal que f proviene de la familia aproximada. En un segundo paso se calcula y compara los posteriores, que son las probabilidades de que f pertenezca a la familia aproximada, dadas las observaciones, $Pr(f \in G_1|x^{(T)})$ y $Pr(f \in G_2|x^{(T)})$. Bajo ciertas condiciones (ver, por ejemplo, Wasserman, 2000), este enfoque da como resultado el criterio de información bayesiano (BIC) que difiere de AIC en la penalización de términos:

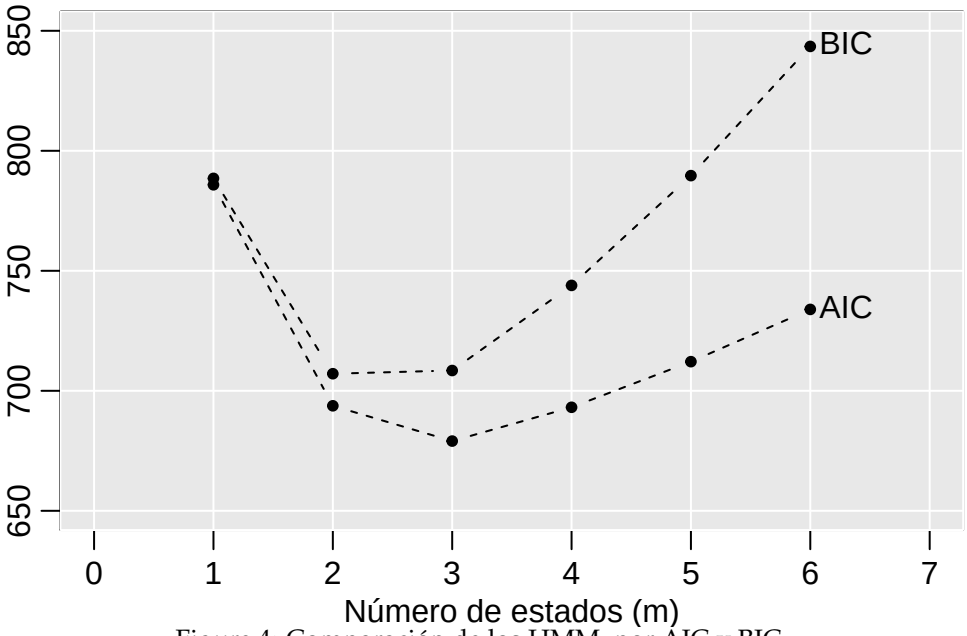


Figure 4: Comparación de los HMM, por AIC y BIC.

$$BIC = -2logL + Tlogp$$

Donde $logL$ y p son como en el AIC, y T es el número de observaciones. En comparación con el AIC, la penalización de términos en el BIC tiene más peso para $T > \epsilon^2$, que se mantiene en la mayoría de las aplicaciones. Por lo tanto, el BIC a menudo favorece modelos con menos parámetros que el AIC. A continuación se muestra en la figura 4 y en la tabla 3 donde se ajustaron seis modelos Ocultos de Markov con diferentes número de estados m. Los resultados indica que el modelo de tres estados con k parámetros debiera ser el modelo seleccionado.

```
par(mar=c(2,2,1,.5)+.5, mgp=c(1.6,.6,0))
plot(c(AIC(mod1),AIC(mod2),AIC(mod3),AIC(mod4),AIC(mod5),AIC(mod6)),
type = "b", xlim = c(0,7),ylim=c(650,850),pch=20, lty=2,ylab="",xlab="Número de estados (m)")
lines(c(BIC(mod1),BIC(mod2),BIC(mod3),BIC(mod4),BIC(mod5),BIC(mod6)), type = "b", pch=20,lty=2)
rect(par("usr")[1], par("usr")[3],par("usr")[2],par("usr")[4],col=gray(.9,.9),border='white')
grid(lty=1, col='white')
text(6.3,845,labels = "BIC");text(6.3,735,labels = "AIC")
lines(c(AIC(mod1),AIC(mod2),AIC(mod3),AIC(mod4),AIC(mod5),AIC(mod6)), type = "b", pch=20,lty=2)
lines(c(BIC(mod1),BIC(mod2),BIC(mod3),BIC(mod4),BIC(mod5),BIC(mod6)), type = "b", pch=20,lty=2)
```

```
stargazer::stargazer(cbind(Modelo = c("HM-1 Estado","HM-2 Estado","HM-3 Estado",
"HM-4 Estado","HM-5 Estado","HM-6 Estado"),k=c(1,4,9,16,25,36),
AIC=round(c(AIC(mod1),AIC(mod2),AIC(mod3),AIC(mod4),AIC(mod5),AIC(mod6)),2),
BIC=round(c(BIC(mod1),BIC(mod2),BIC(mod3),BIC(mod4),BIC(mod5),BIC(mod6)),2),
"-logL"=-1*round(c(logLik(mod1),logLik(mod2),logLik(mod3),logLik(mod4),logLik(mod5),
logLik(mod6)),4)),title = "Comparación de los HMM, por AIC y BIC",
header = FALSE)
```

Table 4: Comparación de los HMM, por AIC y BIC

Modelo	k	AIC	BIC	-logL
HM-1 Estado	1	785.84	788.51	391.9189
HM-2 Estado	4	693.76	707.12	341.8787
HM-3 Estado	9	679.05	708.46	328.5275
HM-4 Estado	16	693.13	743.91	327.5652
HM-5 Estado	25	712.14	789.65	327.0708
HM-6 Estado	36	733.91	843.49	325.954

En primer lugar aunque quizás sea obvio a priori que ni siquiera se debe intentar modelar con 25 o 36 parámetros para 107 observaciones, y observaciones dependientes en eso, es interesante explorar las funciones de verosimilitud en el caso de HMM con Cinco y seis estados. La probabilidad parece ser altamente multimodal en estos casos, y es fácil encontrar varios máximos locales usando valores de partida diferentes. Una estrategia que parece tener éxito en estos casos es iniciar todas las probabilidades de transición o diagonal a valores pequeños (como 0.01) y espaciar los medios dependientes del estado en un rango algo menor que el rango de las observaciones. Según AIC y BIC, el modelo con tres estados es el más apropiado. Pero, más generalmente, el modelo seleccionado puede depender del criterio de selección adoptado.

8.2. Introducción a los pseudo-residuales

Para hablar de los pseudo-residuales primero se necestia definir los siguientes resultados. Sea X una variable aleatoria con función de distribución continua F. Entonces $U \sim F(X)$ se distribuye uniformemente en el intervalo unitario $[0, 1]$. Los **pseudo-residuales uniformes** de una observación x_t de una variable aleatoria continua X_t se define como la probabilidad, bajo el modelo ajustado, de obtener una observación menor o igual a x_t:

$$u_t = Pr(X_t \leq x_t) = F_{X_t}(x_t)$$

Es decir, u_t es la observación x_t transformada por su función de distribución bajo el modelo.Aunque el pseudo-residuo uniforme es útil de esta manera, tiene un inconveniente si se usa para la identificación de valores aleatorios. Por ejemplo, si se consideran los valores cercanos a 0 o 1 en una gráfica, es difícil ver si un valor es muy improbable o no.

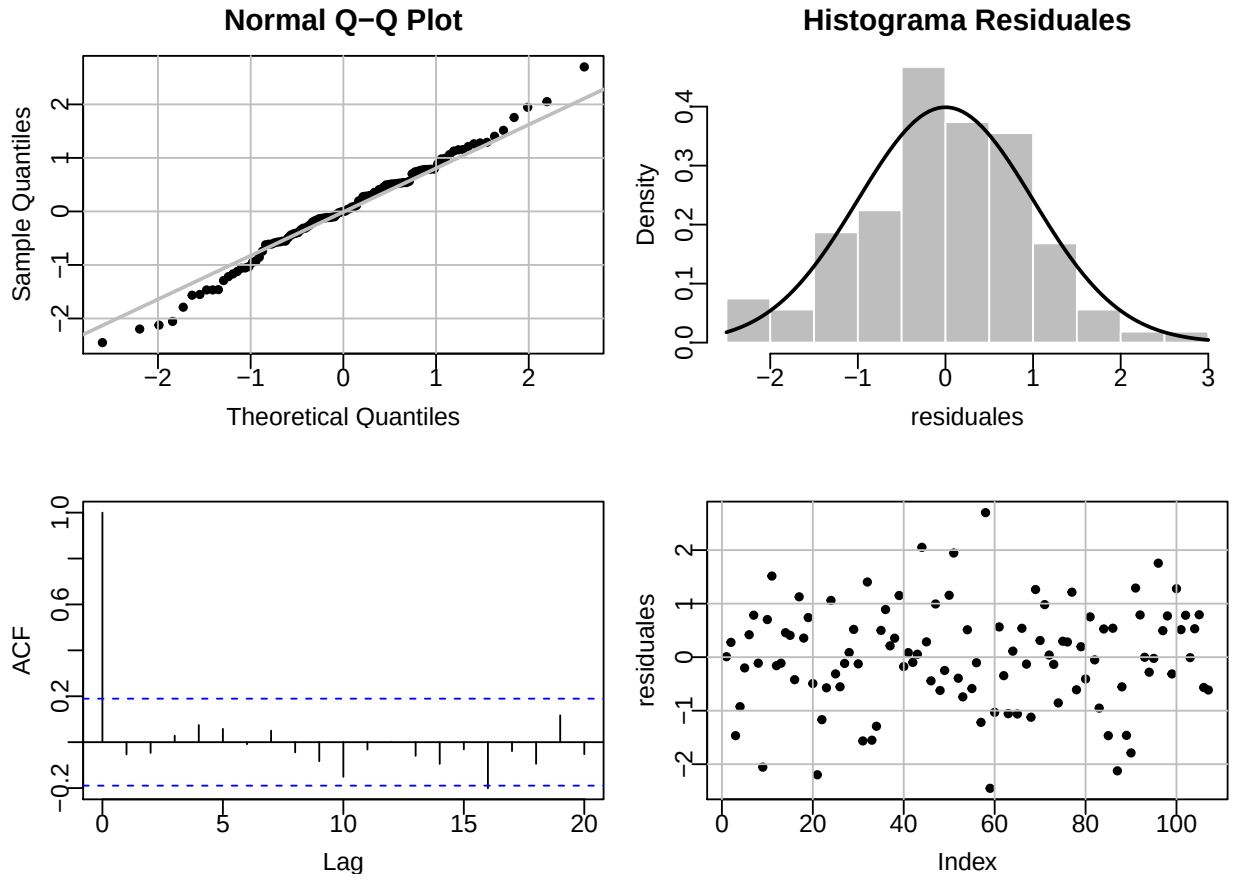
Esta deficiencia de pseudo-residuos uniformes puede, sin embargo, ser fácilmente remediada usando el siguiente resultado (MacDonald and Zucchini, 2009). Sea Φ la función de distribución de una distribución normal estándar y X una variable aleatoria con función de distribución F. Entonces $Z = \Phi^{-1}(F(X))$ se distribuye normal estándar. Ahora definimos los pseudo-residuales normales como:

$$Z = \Phi^{-1}(u_t) = \Phi^{-1}(F_{X_t}(x_t))$$

Si las observaciones $x_1, ..., x_T$ fueron de hecho generados por el modelo $X_t \sim F_t$, los pseudo-residuales normales z_t seguirían una distribución normal estándar. Por lo tanto, se puede comprobar el modelo mediante el análisis visual del histograma o qq-plot de los pseudoresiduales normales, o realizando pruebas de normalidad. Esta versión normal de pseudo-residuos tiene la ventaja frente a los pseudo-residuos uniformes de que el valor absoluto del residuo aumenta con la desviación creciente de la mediana y que las observaciones extremas pueden identificarse más

fácilmente en una escala normal. En la figura 5, se muestran los pseudo-residuales normales de un HMM de tres estados, ajustado con el paquete HiddenMarkov como vemos los residuales se distribuyen normalmente.

```
residuales=residuals(ajuste.mod.HM)
par(mar=c(2,2,2,.5)+.5, mgp=c(1.6,.3,0),mfrow=c(2,2))
# Gráfico 1
qqnorm(residuales,pch=20);qqline(residuales,col="gray",lwd=2)
grid(lty=1, col='gray')
# Gráfico 2
hist(residuales,col="gray",border="white",freq = F,main="Histograma Residuales")
curve(dnorm(x),lwd=2,add=T)
# Gráfico 3
acf(residuales,main="")
# Gráfico 4
plot(residuales,pch=20)
grid(lty=1, col='gray')
```



9. Inferencia Bayesiana

Existen varios enfoques para la inferencia bayesiana en modelos ocultos de Markov; Véase, por ejemplo, (Chib, 1996), más recientemente a seguimos a (Scott, 2002). Nuestro propósito es demostrar una aplicación de la inferencia bayesiana en los Poisson-HMMs. Hay obstáculos que

deben superarse, como el cambio de símbolos y la dificultad para estimar m, el número de estados, y algunos de estos para modelos específicos. Consideramos un Poisson - HMM $\{Y_t\}$ sobre m estados, que subyacen a la cadena de Markov $\{X_t\}$. Denotamos a las medias de los estados dependientes por $\lambda = (\lambda_1, .., \lambda_m)$, y matriz de transición de probabilidad de la cadena de Markov por Γ. Dada una secuencia de observaciones $Y_1, Y_2, ..., Y_T$, un m fijo, y una distribución prior sobre los parámetros λ y Γ el objetivo de esta sección es estimar la distribución posterior de estos parámetros mediante el muestreador de Gibbs. La distribución prior que nosotros asumimos para los parámetros son de la siguiente forma. La r-ésimo fila de Γ, su rango es entre (0,1), por lo tanto Γ se asume que tiene distribución Dirichlet con vector de parámetros v_r, y los incrementos $\tau_j = \lambda_j - \lambda_{j-1}$ con $\lambda_0 = 0$ tienen distribución gamma con parámetro de forma a_j y parámetro de escala b_j. Además, las filas de Γ y los cuantiles de τ_j son asumidos mutuamente independientes en sus distribuciones prior. La notación utilizada es como sigue. Se dice que las variables aleatorias $D_1, ..., D_m$ tienen una distribución de Dirichlet con el vector de parámetro $(v_1, ..., v_m)$ si la densidad conjunta es proporcional a

$$y_1^{v_{1-1}} y_2^{v_{1-1}} \cdots \left(1 - \sum_{i=1}^{m-1} y_i\right)$$

Más precisamente sobre la dimensión $m-1$, esto es sobre el subespacio $\mathbb{R}^{m-1}$, definida por $\sum_{i=1}^{m-1} \leq 1, y_i \geq 0$. Una variable aleatoria X, se dice que tiene distribución gamma xon parámetro de forma α y parámetro de escala b, si su densidad es positiva (para todo x).

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

Con esta parámetrización, X tiene media a/b, variana a/b^2 y coeficiente de variación $(c.v)1/sart(a)$ Si fuera posible observar la cadena de Markov, la actualización de las probabilidades de transición de la matriz Γ ería directa. Aquí, sin embargo, tenemos que generar rutas de la muestra de la cadena de Markov con el fin de actualizar el orden de la cadena en Γ. ??? Una parte importante de la estructura del modelo de Scott, que copiamos, es ésta. Se considera que cada conteo observado y_t es la suma $\sum_j x_{jt}$ de las contribuciones de hasta m regímenes, siendo la contribución del régimen j a $y_t, ..., y_{jt}$. Obsérvese que, si la cadena de Markov está en el estado i en un momento dado, los regímenes $i+1, ..., m$ para estar activos. Este es un uso inusual de la palabra "régimen", pero conveniente aquí.

Instantaneamante la parámetrización del modelo, en términos de los m estados-dependientes de la media λ_i, parametrizamos en términos de incrementos no negativos $\tau = (\tau_1, ..., \tau_m)$, donde $\tau_j = \lambda_j - \lambda_{j-1}$ (con $\lambda_0 = 0$). Equivalente a

$$\lambda_i = \sum_{j=1}^{i} \tau_j$$

Esto tiene el efecto de colocar el λ_j en orden creciente, que es útil para evitar el problema técnico conocido como labelSwitching. Para una explicación de este problema, véase, por ejemplo, (MacDonald and Zucchini, 2009). La variable aleatoria τ_j puede describirse como la contribución media del régimen j, si está activa, al recuento observado en un momento dado. En el esquema, procedemos como sigue.

- Dado los recuentos observados $X^{(T)}$ y los valores actuales de los parámetros Γ y λ, se genera una trayectoria muestral de la cadena de Markov (MC).

- Utilizamos este camino de muestra para descomponer los recuentos observados en contribuciones de régimen (simuladas).

Con la trayectoria muestral del MC disponible, y las contribuciones de régimen, ahora podemos actualizar Γ y τ, por lo tanto λ.

Las etapas anteriores se repiten un gran número de veces y, después de un "periodo de burning", las muestras de valores de Γ y λ resultantes proporcionan las estimaciones requeridas de sus distribuciones posteriores. Ahora denotamos a θ para representar tanto Γ como λ.

9.1. Paquete R2OpenBugs

```
library(R2OpenBUGS)
model.file <- file.path("C:/Users/USER/Desktop/Proyecto", "HMM.txt")
```

9.1.1. Modelo sintaxis BUGS

Para estimar los parámetros con el muestreador de Gibbs, es necesario, incluir adecuadamente los parámetros como se muestra a continuación.

$$\delta_i \sim U(0,m) V_i \sim U(0,1) S_i \sim Categorico(\Gamma_{i-1}) X_i \sim Pois(\lambda_{S_i}) \tau \sim gamma(1, 0.08) \Gamma \sim Dirichlet(V_i)$$

```
HMM<- function(){
  for(i in 1:m){
    delta[i] <- 1/m
    v[i] <- 1}
  s[1] ~ dcat(delta[])
  for (i in 2:100){
    s[i] ~ dcat(Gamma[s[i-1],])}
  states[1] ~ dcat(Gamma[s[100],])
  x[1]~dpois(lambda[states[1]])
  for(i in 2:n){
    states[i]~dcat(Gamma[states[i-1],])
    x[i]~dpois(lambda[states[i]])}
  for(i in 1:m){
    tau[i]~dgamma(1,0.08)
    Gamma[i,1:m]~ddirch(v[])}
  lambda[1]<-tau[1]
  for(i in 2:m){
    lambda[i]<-lambda[i-1]+tau[i]}}
# Escribir el modelo en formato .txt
write.model(HMM, model.file)
model.HMM = paste(getwd(),"HMM.txt", sep="/")
```

```
# Nombre correspondiente a las variables nombradas en el modelo
x <- Terremotos
n <- length(x)
```

```
m <- 3
data <- list("x", "n", "m" )
```

```
# Correr el muestreador de Gibbs
dir <- "C:/Program Files (x86)/OpenBUGS/OpenBUGS323/OpenBugs.exe"
eq.sims <- bugs(data,inits=NULL,model.file=model.HMM,summary.only = TRUE,
                n.burnin = 50,n.thin = 5,OpenBUGS.pgm = dir,
                parameters=c("tau","lambda","Gamma"),n.iter=1000,n.chains=1)
stargazer::stargazer(round(cbind(eq.sims$stats[,-c(4,6)],Clasico=
c(ajuste.mod.HM$Pi[1,1],ajuste.mod.HM$Pi[1,2],ajuste.mod.HM$Pi[1,3],
  ajuste.mod.HM$Pi[2,1],ajuste.mod.HM$Pi[2,2],ajuste.mod.HM$Pi[2,3],
  ajuste.mod.HM$Pi[3,1],ajuste.mod.HM$Pi[3,2],ajuste.mod.HM$Pi[3,3],
  BIC(mod3),
  ajuste.mod.HM$pm$lambda[1],ajuste.mod.HM$pm$lambda[2],ajuste.mod.HM$pm$lambda[3],
  ajuste.mod.HM$delta[1],ajuste.mod.HM$delta[2],ajuste.mod.HM$delta[3])),4),header = FALSE,
  title = "Ajuste Bayesiano vs Clásico")
```

Table 5: Ajuste Bayesiano vs Clásico

	mean	sd	val2.5pc	val97.5pc	Clasico
Gamma[1,1]	0.855	0.080	0.684	0.970	0.939
Gamma[1,2]	0.095	0.072	0.006	0.287	0.032
Gamma[1,3]	0.050	0.041	-0.025	0.156	0.029
Gamma[2,1]	0.096	0.067	0.007	0.246	0.040
Gamma[2,2]	0.813	0.083	0.633	0.934	0.906
Gamma[2,3]	0.092	0.059	0.002	0.238	0.053
Gamma[3,1]	0.080	0.080	0.003	0.237	0
Gamma[3,2]	0.230	0.123	0.045	0.499	0.190
Gamma[3,3]	0.691	0.140	0.403	0.913	0.810
deviance	611.400	10.710	591.100	636.700	708.456
lambda[1]	13.240	0.853	11.580	15.010	13.134
lambda[2]	19.940	1.288	17.680	22.610	19.713
lambda[3]	29.700	2.053	26.200	33.750	29.712
tau[1]	13.240	0.853	11.580	15.010	1
tau[2]	6.696	1.226	4.321	9.195	0
tau[3]	9.762	1.904	5.987	13.500	0

```
eq.sims2 <- bugs(data,inits=NULL,model.file=model.HMM,codaPkg = TRUE,
                 n.burnin = 50,n.thin = 5,OpenBUGS.pgm = dir,
                 parameters=c("tau","lambda","Gamma"),n.iter=1000,n.chains=1)
```

```
library(coda)
cadena <- as.mcmc(as.matrix(line.coda))
heidel.diag(cadena)
```

```
##
##             Stationarity start     p-value
##             test         iteration
## Gamma[1,1] passed        1         0.767
## Gamma[1,2] passed        1         0.420
## Gamma[1,3] passed        1         0.975
## Gamma[2,1] passed        1         0.104
## Gamma[2,2] passed       96         0.116
## Gamma[2,3] passed        1         0.210
## Gamma[3,1] passed        1         0.165
## Gamma[3,2] passed        1         0.676
## Gamma[3,3] passed        1         0.990
## deviance   passed        1         0.776
## lambda[1]  passed        1         0.988
## lambda[2]  passed        1         0.842
## lambda[3]  passed        1         0.996
## tau[1]     passed        1         0.988
## tau[2]     passed        1         0.674
## tau[3]     passed        1         0.292
##
##             Halfwidth Mean     Halfwidth
##             test
## Gamma[1,1] passed      0.8551 0.01848
## Gamma[1,2] failed      0.0948 0.01319
## Gamma[1,3] passed      0.0501 0.00439
## Gamma[2,1] failed      0.0958 0.00986
## Gamma[2,2] passed      0.8080 0.01228
## Gamma[2,3] passed      0.0916 0.00834
## Gamma[3,1] failed      0.0795 0.01058
## Gamma[3,2] passed      0.2298 0.01516
## Gamma[3,3] passed      0.6907 0.02325
## deviance   passed    611.4135 1.22630
## lambda[1]  passed     13.2421 0.16967
## lambda[2]  passed     19.9384 0.25346
## lambda[3]  passed     29.7008 0.29180
## tau[1]     passed     13.2421 0.16967
## tau[2]     passed      6.6964 0.16857
## tau[3]     passed      9.7621 0.21110
```

10. Conclusiones

Sugerimos PHMMs como un enfoque más general que la distribución de Poisson y el proceso de Poisson para modelar el número de terremotos mayores en el mundo. PHMMs permite modelar la sobredispersión en los datos de conteo y explicar la variabilidad, cambiando el parámetro de Poisson de acuerdo a una cadena de Markov no observada.

En esta aplicación, la dimensión m del estado-espacio de la cadena de Markov ha sido estimada por el Criterio de Información de Alcaike (AIC) y el Criterio de Información de Bayes (BIC).

Como lo señalan ZUcchini y MacDonald (2009), los investigadores y los profesionales tienden

a utilizar EM o maximización numérica directa, pero no ambos, para realizar la estimación de máxima verosimilitud en HMMs, y cada enfoque tiene sus méritos.

Específicamente en el contexto de los HMM el algoritmo EM es fácil de programar, ya que no está involucrada ninguna evaluación directa de la verosimilitud o sus derivados, además calcula las probabilidades *forward* y *backward*. Por otro lado el enfoque bayesiano es exigente Computacionalmente, por ejemplo MCMC y en ciertos otros aspectos. El modelo necesita ser parametrizado de una manera especifica

10.1. Comentarios Finales

Me pareció muy interesante realizar este trabajo y abordar especialmente este tema, ya que fue raro encontrar artículos que relacionaran los procesos estocásticos con la estadística bayesiana, el cual esta área es de mi interés. Pienso que hay una gran parte aun por explorar, pues los avances bayesianos igual que en los modelos ocultos de Markov, han sido bastante recientes debido a los avances computacionales. El libro de Zucchini el cual me base gran parte de la teoría es del 2009,hace apenas ocho años, y el paquete R2OpenBugs utilizado para realizar las estimaciones bayesianas por medio de OpenBugs, tuvo su última actualización el 22 de febrero del 2017, por otra parte en mi revisión bibliográfica son muy escaso los artículos relacionados con los HMM, por lo que todavía hay bastante, por explorar pues aunque, los HMM datan de la década de los 70s, en reconocimiento de patrones del habla, su diversas aplicaciones en otras áreas hasta ahora empiezan a difundirse.

10.2. Recomendaciones

Como vimos para las estimaciones de los parámetros del modelo Poisson oculto de Markov, se utilizaron dos metodologías, la inferencia clásica (los algoritmos EM y una derivación de este el algoritmo Baum-Welch, pero más eficiente computacionalmente) y la inferencia bayesiana (con el muestreador de Gibbs). Aunque las estimaciones fueron muy parecidas en ambos caso, se recomienda para las personas interesadas en enfocarse en la parte bayesiana ver el libro de Zucchini (2009), pues no se alcanzaron abordar algunos temas, como la estimación bayesiana del número de estados a traves del factor de Bayes, la maximización numérica directa, el pronóstico de las, etc. dsitribuciones.

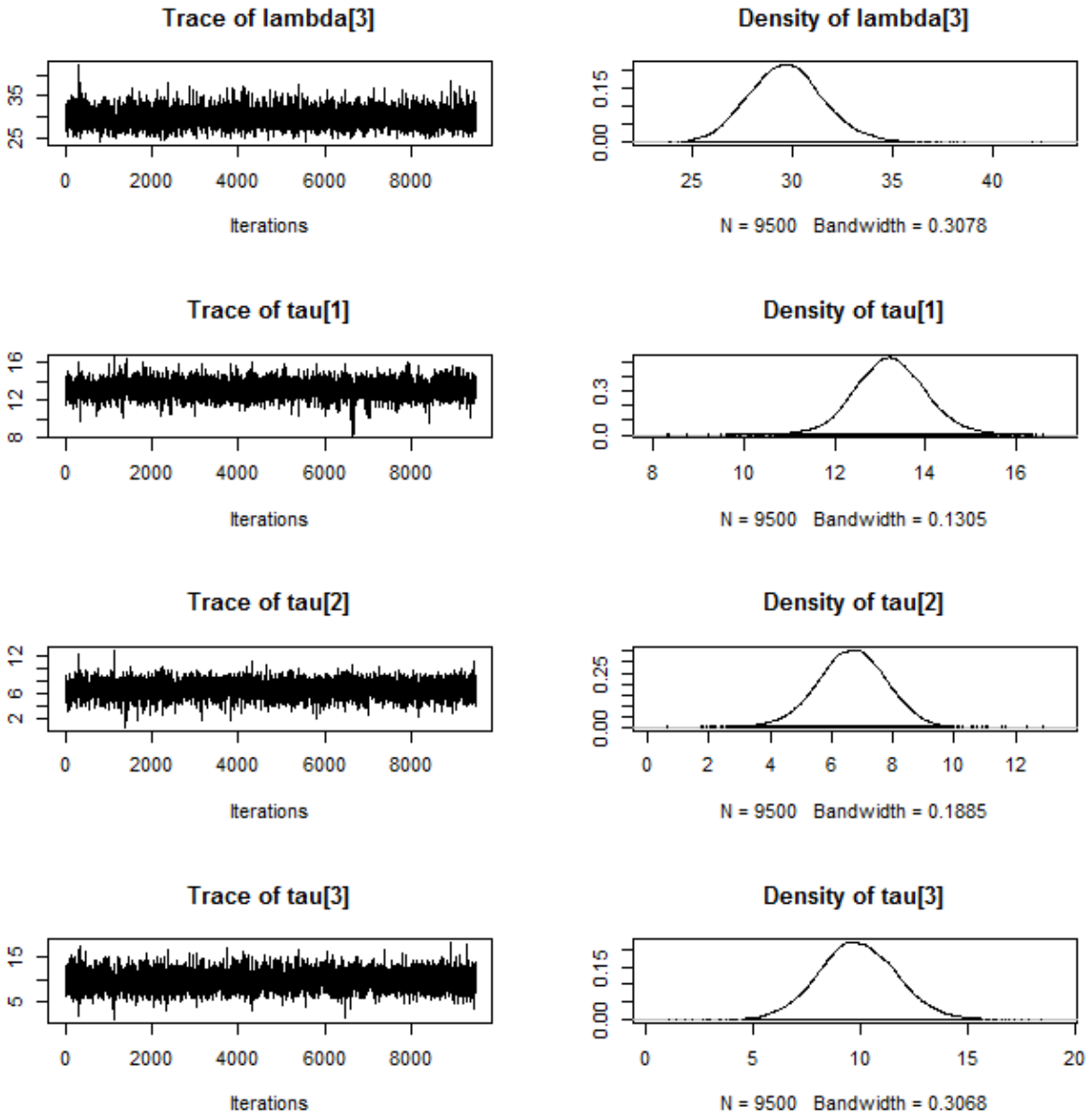


Figure 5: Traza de la Cadena

References

Albert, James H and Siddhartha Chib. 1993. "Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts." *Journal of Business & Economic Statistics* 11(1):1–15.

Basawa, Ishwar V. 2014. *Statistical Inferences for Stochasic Processes: Theory and Methods*. Elsevier.

Baum, Leonard E., Ted Petrie, George Soules and Norman Weiss. 1970. "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains." *Ann. Math. Statist.* 41(1):164–171.

Beal, Matthew J., Zoubin Ghahramani and Carl E. Rasmussen. 2002. The Infinite Hidden Markov Model. In *Machine Learning*. MIT Press pp. 29–245.

Bickel, Peter J, Ya'acov Ritov, Tobias Ryden et al. 1998. "Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models." *The Annals of Statistics* 26(4):1614–1635.

Bilmes, Jeff. 1998. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical report.

Chib, Siddhartha. 1996. "Calculating posterior distributions and modal estimates in Markov mixture models." *Journal of Econometrics* 75(1):79–97.

Churchill, Gary A. 1989. "Stochastic models for heterogeneous DNA sequences." *Bulletin of Mathematical Biology* 51(1):79–94.

Ephraim, Y. and N. Merhav. 2006. "Hidden Markov Processes." *IEEE Trans. Inf. Theor.* 48(6):1518–1569.

Grimmett, Geoffrey and David Stirzaker. 2001. *Probability and random processes*. Oxford university press.

Guttorp, Peter and Vladimir N Minin. 1995. *Stochastic modeling of scientific data*. CRC Press.

Hamilton, James D. 1989. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica* 57(2):357–384.

Leroux, Brian G. and Martin L. Puterman. 1992. "Maximum-Penalized-Likelihood Estimation for Independent and Markov- Dependent Mixture Models." *Biometrics* 48(2):545–558.

Linhart, H and W Zucchini. 1986. *Model Selection*. New York, NY, USA: John Wiley & Sons, Inc.

MacDonald, Iain L and Walter Zucchini. 2009. *Hidden Markov Models for Time Series: An Introduction Using R (Monographs on statistics and applied probability; 110)*. CRC press.

Rabiner, Lawrence R. 1990. Readings in Speech Recognition. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. chapter A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, pp. 267–296.

Romberg, J. K., Hyeokho Choi and R. G. Baraniuk. 2001. "Bayesian Tree-structured Image Modeling Using Wavelet-domain Hidden Markov Models." *Trans. Img. Proc.* 10(7):1056–1068.

Scott, Steven L. 2002. "Bayesian methods for hidden Markov models: Recursive computing in the 21st century." *Journal of the American Statistical Association* 97(457):337–351.

Viterbi, Andrew. 1967. "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm." *IEEE transactions on Information Theory* 13(2):260–269.

Wasserman, Larry. 2000. "Bayesian model selection and model averaging." *Journal of mathematical psychology* 44(1):92–107.

Wu, CF Jeff. 1983. "On the convergence properties of the EM algorithm." *The Annals of statistics* pp. 95–103.

Zucchini, Walter. 2000. "An introduction to model selection." *Journal of mathematical psychology* 44(1):41–61.