

Capítulo 14

Análisis usando el factor de Bayes, para múltiples puntos de cambio temporales de un conjunto de datos simulados de un proceso Poisson

Bulmaro Juárez-Hernández, Lucila Muñoz-Merino y Víctor Hugo Vázquez-Guevara

Facultad de Ciencias Físico Matemáticas,
Benemérita Universidad Autónoma de Puebla,
bjuarez@fcfm.buap.mx, 216570304@alumnos.fcfm.buap.mx,
vvazquez@fcfm.buap.mx

Resumen. En este trabajo se realiza un análisis de puntos de cambio utilizando el factor de Bayes. Para llevar a cabo este análisis, se simularon datos de un proceso Poisson con un programa en el que se usaron instrucciones de las librerías de los paquetes INLA y Poisson de R, a los datos de este proceso se les aplicó el método de bisección y el factor de Bayes para detectar los puntos de cambio.

Abstract. In this work, an analysis of change points is carried out using the Bayes factor. To bring about this analysis, data from a Poisson process was simulated with a program in which instructions from the INLA and Poisson package libraries of R was used, the bisection method and the Bayes factor was applied to the data of this process for detect change points.

Palabras clave: Factor de Bayes, Poisson y método de bisección.

14.1. Introducción

Chen y Gupta [3] definen a un punto de cambio, en una sucesión de datos $\{x_{t_i}\}$, $i = 1, \dots, n$ observados y ordenados respecto al tiempo, como aquel, para el cual las observaciones siguen una distribución F_1 , antes de dicho punto, y en otro posterior la distribución es F_2 . Es decir, desde el punto de vista estadístico, la sucesión de observaciones muestra un comportamiento no homogéneo. El problema de punto de cambio es considerado como uno de los problemas centrales de inferencia estadística, pues relaciona a la teoría de control estadístico, a las pruebas de hipótesis (al detectar si existe algún cambio en la sucesión de variables aleatorias observadas), y a la teoría de estimación (al estimar el número de cambios y sus correspondientes localizaciones). Esto bajo los enfoques clásico y Bayesiano [3].

Los problemas de puntos de cambio originalmente surgieron en control de calidad y en general pueden ser encontrados en la modelación matemática de diversas disciplinas tales como Medio Ambiente, Epidemiología, Procesos de señal sísmica, Economía, Finanzas, Geología, Medicina, Biología, Física, etc.

En general el problema de puntos de cambio se visualiza de la forma siguiente [3]:

Sean X_1, \dots, X_n una colección de vectores (variables) aleatorios independientes con funciones de distribución de probabilidad F_1, \dots, F_n , respectivamente. Entonces el problema de puntos de cambio consiste en probar la hipótesis nula H_0 de la no existencia de cambio contra la alternativa H_a de que existe al menos un punto de cambio lo cual se expresa de la siguiente manera:

$$H_0 : F_1 = F_2 = \dots = F_n \text{ (en todos los puntos)}$$

vs

$$H_a : F_1 = F_2 = \dots = F_{k_1} \neq F_{k_1+1} = \dots = F_{k_2} \neq F_{k_2+1} = \dots = F_{k_q} \neq F_{k_q+1} = \dots = F_n.$$

Donde $1 < k_1 < k_2, \dots, k_q < n$, q es el número desconocido de puntos de cambio y k_1, k_2, \dots, k_q son las posiciones desconocidas respectivas que tienen que ser estimadas. Si las distribuciones F_1, F_2, \dots, F_n pertenecen a una familia paramétrica común $F(\theta)$, donde $\theta \in \mathbb{R}^p$, entonces el problema de puntos de cambio consiste en probar la hipótesis nula H_0 sobre la no existencia de cambio en los parámetros θ_i , $i = 1, \dots, n$ de la población contra la alternativa H_a de que existe al menos un punto de cambio; lo cual se expresa de la siguiente forma:

$$H_0 : \theta_1 = \theta_2 = \dots = \theta_n = \theta, \text{ desconocido}$$

vs

$$H_a : \theta_1 = \dots = \theta_{k_1} \neq \theta_{k_1+1} = \dots = \theta_{k_2} \neq \theta_{k_2+1} = \dots = \theta_{k_q} \neq \theta_{k_q+1} = \dots = \theta_n.$$

donde q y k_1, k_2, \dots, k_n tienen que ser estimados. Estas hipótesis revelan los aspectos de inferencia de puntos de cambio para determinar si existe algún punto de cambio en el proceso, estimar el número de ellos y sus respectivas posiciones.

En este trabajo se desarrolla un programa para aplicar algunos procedimientos en la detección de puntos de cambio temporales, en particular usando el factor de Bayes. Para desarrollar este método se tomó como base los conceptos presentados en Altieri [1]. También se utiliza el proceso Poisson homogéneo, con el propósito de simular valores donde se encontrarán los puntos de cambio. Se analizan y comparan los resultados obtenidos. El mencionado programa fue creado utilizando algunas instrucciones del paquete INLA de R, el cual, se presenta en los libros de Gómez [4] y de Blangiardo y Cameletti [2] para lo cual se usaron las librerías INLA y Poisson.

14.2. Factor de Bayes

Si se presenta un problema de selección de modelos, en el que, se debe elegir entre dos posibles modelos, en base a un conjunto de datos observados D , la plausibilidad de la diferencia de dos modelos M_1 y M_2 , parametrizados por vectores de parámetros θ_1 y θ_2 se puede medir mediante el factor Bayes.

El factor de bayes se define como:

$$B = \frac{P(D|M_1)}{P(D|M_2)} = \frac{\int_{\theta_1} P(D|\theta_1, M_1)\Pi(\theta_1|M_1)d\theta_1}{\int_{\theta_2} P(D|\theta_2, M_2)\Pi(\theta_2|M_2)d\theta_2},$$

donde $P(D|M_1)$ se denomina verosimilitud marginal o verosimilitud integrada. Esto es similar a lo que se hace en las pruebas de la razón de verosimilitudes pero ahora, en lugar de maximizar la verosimilitud, el factor Bayes realiza un promedio ponderado mediante la

distribución de los parámetros.

Un valor de $B > 1$ significa que M_1 es apoyado por los datos más que M_2 .

En el caso del factor de Bayes, Jeffreys [5] estableció una escala de interpretación de B , la cual se muestra en la Tabla 14.1.

Tabla 14.1: Escala de interpretación de B , según Jeffreys.

B	Fuerza de la evidencia a favor de M_1
$B \leq 1$	Negativa apoya M_2
$1 < B \leq 3$	Muy escasa
$3 < B \leq 10$	Sustancial
$10 < B \leq 30$	Fuerte
$30 < B \leq 100$	Muy fuerte
> 100	Decisiva

Otra forma de considerar el factor de bayes es la siguiente: Supóngase dos hipótesis H_0 y H_1 , tales que, las densidades a priori son: $f_0 = P(H_0)$ y $f_1 = P(H_1)$. Después de observar una muestra aleatoria, las probabilidades a posteriori de ambas hipótesis son $\alpha_0 = P(H_0|x)$ y $\alpha_1 = P(H_1|x)$. Se define el factor de Bayes a favor de H_0 como

$$B = \frac{\frac{\alpha_0}{\alpha_1}}{\frac{f_0}{f_1}} = \frac{\alpha_0 f_1}{\alpha_1 f_0}.$$

Así, el factor de Bayes representa la plausibilidad a posteriori dividida entre la plausibilidad a priori. Nos informa de los cambios en nuestras creencias introducidas por los datos. Tiene la propiedad de que es casi objetivo y elimina parcialmente la influencia de la distribución a priori.

Como ejemplo, supóngase el contraste simple:

$$H_0 : \theta = \theta_0 \quad vs \quad H_1 : \theta = \theta_1.$$

Se tiene que las distribuciones a posteriori son:

$$\alpha_0 = P(H_0|x) = \frac{f_0 L(\theta_0|x)}{f_0 L(\theta_0|x) + f_1 L(\theta_1|x)},$$

$$\alpha_1 = P(H_1|x) = \frac{f_1 L(\theta_1|x)}{f_0 L(\theta_0|x) + f_1 L(\theta_1|x)}.$$

Entonces el factor de Bayes es:

$$B = \frac{\alpha_0 f_1}{\alpha_1 f_0} = \frac{f_0 L(\theta_0|x) f_1}{f_1 L(\theta_1|x) f_0} = \frac{L(\theta_0|x)}{L(\theta_1|x)}.$$

que coincide con la razón de verosimilitudes, de modo que, la distribución a priori no influiría, en este caso, en el factor de Bayes.

Así, el factor de Bayes para el punto de cambio cuando se divide en dos segmentos está dado por la razón de verosimilitudes:

$$\frac{L_0}{L_1} = \frac{Q_1 Q_2}{L_1},$$

donde Q_1 es la verosimilitud del segmento 1 y Q_2 es la verosimilitud del segmento 2 bajo la hipótesis alternativa y L_1 es la verosimilitud bajo la hipótesis nula.

Así, aplicando logaritmos se tiene:

$$\ln(B) = \ln(Q_1) + \ln(Q_2) - \ln(L_1).$$

14.3. Proceso de Poisson homogéneo

Para aplicar el factor de Bayes que determina cuales son los puntos de cambio, se simula un proceso Poisson homogéneo y se trabaja con los datos obtenidos. Ahora, un proceso de Poisson está definido de la siguiente forma:

Definición: Una colección de variables aleatorias $\{N(t) : t \geq 0\}$ (definidas en un espacio de probabilidad (Ω, F, P)) se llama proceso de Poisson (homogéneo) con intensidad $\lambda > 0$ si satisfacen las siguientes propiedades:

- i) $P(N(0) = 0) = 1$.
- ii) Para todo $0 < s < t$, $N(t) - N(s)$ tiene distribución de Poisson de parámetro $\lambda(t-s)$.
- iii) Para todo $0 \leq t_1 < \dots < t_n$, $n \geq 1$ (es decir, para todo conjunto finito de tiempos), las variables aleatorias $N(t_n) - N(t_{n-1}), \dots, N(t_2) - N(t_1), N(t_1) - N(0), N(0)$, son independientes. Esta propiedad se conoce como propiedad de incrementos independientes.

14.4. Resultados y discusión

14.4.1. Análisis de múltiples puntos de cambio temporales con el factor de Bayes

Con el objetivo de detectar, analizar y comparar resultados de puntos de cambio, se hicieron programas para detectar múltiples puntos de cambio, entre ellos, uno en el que se detectan 5 puntos de cambio, los datos utilizados fueron simulaciones de un proceso Poisson obteniéndose un conjunto de 60 datos con 6 valores diferentes para el parámetro λ , otro en el que se detectan 6 puntos de cambio, aquí los datos utilizados fueron simulaciones de un proceso Poisson obteniéndose un conjunto de 60 datos con 7 valores diferentes para el parámetro λ . Se pretende, en este problema, detectar los cambios que se generan con los distintos valores para el parámetro de intensidad λ , utilizando el método de bisección y el factor de Bayes, además, se hizo un comparativo de los resultados al utilizar las a priori: uniforme, log-gamma y Gaussiana.

Análisis del caso con 5 puntos de cambio

Se corrió un programa en R, el programa para detectar múltiples puntos de cambio temporales, en total de 5 puntos de cambio, se utilizaron 60 datos y el segmento más pequeño fue de 4 puntos. Se simuló un proceso Poisson uniforme con diferentes valores para

el parámetro λ y se aproximó a la distribución posteriori con el paquete de R, INLA, el cual aproxima con series de Taylor. Para detectar los puntos de cambio se utilizó el método de bisección y el factor de Bayes.

Se utilizaron tres distribuciones a priori una uniforme, una loggamma y una gaussiana. Se muestra una gráfica con los 5 puntos de cambio en la Figura 14.1.

El método de bisección consiste en dividir al conjunto de los datos en dos subconjuntos con la misma (o aproximada) cantidad de datos y busca puntos de cambio, enseguida se va a la izquierda y también se divide al subconjunto a la mitad y busca puntos de cambio, posteriormente se divide el lado derecho y así se sigue sucesivamente yendo a la izquierda y a la derecha, dividiendo los respectivos subconjuntos. Para aplicar el método se usan 60 datos simulados y se llega a la división más pequeña que fue de cuatro datos, es decir se tienen dos pasos para hacer las divisiones y en cada paso, de las divisiones respectivas, se obtienen los posibles puntos de cambio en los subconjuntos, que en este caso resultaron ser 15, finalmente usando el factor de Bayes se determinan cuales son los puntos de cambio reales para este proceso, el resultado se muestra en la Tabla 14.2.

Los puntos de cambio para cuando los datos son generados con valores del parámetro $\lambda = 2, 1, 4, 7, 6, 1$, cuyas divisiones del conjunto de datos o segmentos cuando se utiliza una a priori uniforme quedan con los siguientes números de datos 8, 7, 15, 15, 8, 7, respectivamente. El resultado para la detección de los puntos de cambio fueron, como se muestra en la Tabla 14.2, esto es, los puntos de cambio, en este caso, son los primeros cuatro y el séptimo.

Con los mismos puntos de cambio y con la distribución a priori loggamma con parámetro 0.01, el resultado se muestra en la Tabla 14.2. Se puede observar que en comparación con la uniforme las log-verosimilitudes de la loggamma son más pequeñas, sin embargo, también detectan los mismos puntos de cambio.

Con la distribución a priori gaussiana con media cero y parámetro de precisión 0.001, para el mismo número de puntos de cambio, se detectaron los cinco puntos de cambio, se puede observar en los datos de la Tabla 14.2, son los primeros cuatro y el siete, aunque el valor de la log-verosimilitud del quinto punto de cambio que se encuentra en la séptima posición es menor en comparación con las siguientes log-verosimilitudes que se encuentran en las posiciones 8, 10, 12 y 14, como se puede observar en la Tabla 14.2.

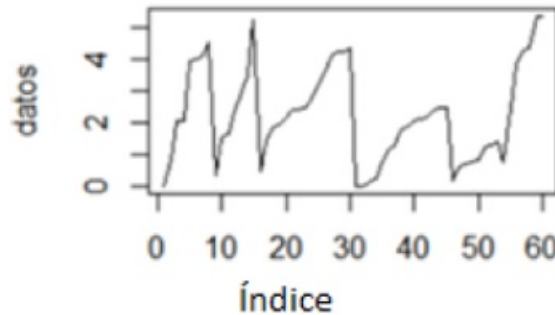


Figura 14.1: Gráfica de 5 puntos de cambio

Tabla 14.2: Detención de cinco puntos de cambio.

Num	uniforme	loggamma	Gaussiana
1	49.565011	43.644575	42.972975
2	24.392282	25.829232	26.592785
3	13.940830	26.041694	28.329433
4	16.000633	9.990938	11.641769
5	12.478724	2.387813	4.477951
6	12.478724	2.387813	4.477951
7	12.572475	4.498905	5.523642
8	5.701123	1.835601	5.801758
9	4.732525	1.384310	4.912561
10	5.701123	1.835601	5.801758
11	4.732525	1.384310	4.912561
12	5.701123	1.835601	5.801758
13	4.732525	1.384310	4.912561
14	5.701123	1.835601	5.801758
15	4.732525	1.384310	4.912561

Los puntos de cambio se muestran en la Figura 14.1 y se encuentran en 8, 15, 30, 45 y 53.

Análisis del caso con 6 puntos de cambio

También para este caso, se corrió un programa en R, el programa para detectar múltiples puntos de cambio, aquí el número total de puntos de cambio es 6, se generaron y utilizaron 60 datos, y el segmento más pequeño fue de 3 puntos. Esto es, se simuló un proceso Poisson uniforme con diferentes valores para el parámetro λ y se aproximó a la distribución a posteriori con el paquete de R, INLA. Para detectar los puntos de cambio, se utilizó el método de bisección y el factor de Bayes.

Nuevamente se utilizaron las distribuciones a priori: uniforme, loggamma y gaussiana.

Para este caso y siguiendo el proceso descrito en la sección anterior, se obtuvo que con el modelo uniforme se detectaron 6 puntos de cambio para $\lambda = 2, 1, 4, 7, 6, 1, 2$ y en las divisiones de datos o segmentos de 8, 7, 15, 15, 8, 4, 3. Los puntos de cambio que se detectaron son los primeros cuatro, el séptimo y el último. Los resultados se muestran en la segunda columna de la Tabla 14.3.

Ahora con la distribución a priori loggamma con parámetro 0.01, el resultado se muestra en la Tabla 14.3. Se puede observar que en comparación con la uniforme las log-verosimilitudes de la loggamma son más pequeñas, sin embargo también detectan los mismos puntos de cambio.

Con la distribución a priori gaussiana con media cero y parámetro de precisión 0.001, se detectaron los seis puntos de cambio, son los primeros cuatro, el siete y el 15, como se puede observar, en la última columna de la Tabla 14.3. La diferencia entre esta y la uniforme es que, el último punto de cambio, tiene el log-verosimilitud menor a los valores de los log-verosimilitud anteriores, en donde no hay punto de cambio.

Tabla 14.3: Detención de seis puntos de cambio.

Num	uniforme	loggamma	Gaussiana
1	37.382628	44.717137	38.258157
2	23.107912	28.913661	26.506732
3	22.987741	27.851420	26.529322
4	16.466681	5.045601	11.798824
5	12.478724	2.387813	4.477951
6	12.478724	2.387813	4.477951
7	16.555598	3.062915	12.572375
8	5.701123	1.835601	5.801758
9	4.732525	1.384310	4.912561
10	5.701123	1.835601	5.801758
11	4.732525	1.384310	4.912561
12	5.701123	1.835601	5.801758
13	4.732525	1.384310	4.912561
14	5.701123	1.835601	5.801758
15	4.733385	1.720949	4.280097

Los puntos de cambio se muestran en la Figura 14.2, y se encuentran en los puntos: 8, 15, 30, 45, 53 y 57.

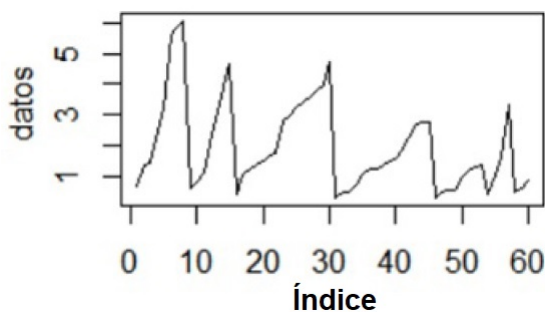


Figura 14.2: Gráfica con 6 puntos de cambio

El programa se hizo en R, es etiquetado como Programa 1 y se presenta al final, en el Apéndice A.

14.5. Conclusiones

Como el propósito de este trabajo era verificar que el método del factor de Bayes es una herramienta que funciona adecuadamente para detectar puntos de cambio, se procedió a simular procesos de Poisson, obteniéndose datos para valores diferentes del parámetro, después se elaboró un programa usando el método de bisección y el factor de Bayes para detectar los puntos de cambio considerando los datos simulados y así, numéricamente detectar los puntos de cambio y comparar los resultados de las distribuciones a priori utilizadas. De los resultados obtenidos se concluye que el método del factor de Bayes

detecta bien los puntos de cambio, pero su debilidad es que detecta cambios hasta una división que contiene al menos cuatro datos, para menos datos en la división ya no se detectan los puntos de cambio.

Bibliografía

- [1] Altieri Linda, *A Bayesian changepoint analysis on spatio-temporal point processes*, Tesis , 2015.
- [2] Blangiardo Marta y Cameletti Michela *Spatial and Spatio-temporal Bayesian Models with R-INLA*, John Wiley and Sons. 2015.
- [3] Chen, J., y Gupta, A. *Parametric Statistical Changepoint Analysis*, Bogota: Birkhauser. 2012.
- [4] Gómez Rubio Virgilio, *Bayesian inference with INLA*, Chapman and Hall, 2020.
- [5] Jeffreys H., *Theory of probability*, Oxford: Oxford University Press; 1961.

14.6. Apéndice

Programa.

```
rm(list = ls())
library(poisson)
library(INLA)
vector <- -hpp.event.times(2, 15, num.sims = 1, t0 = 0)
plot(vector)
vector1 <- -hpp.event.times(1, 15, num.sims = 1, t0 = 0)
vector2 <- -hpp.event.times(4, 15, num.sims = 1, t0 = 0)
vector4 <- -hpp.event.times(6, 15, num.sims = 1, t0 = 0)
datos <- -c(vector, vector1, vector2, vector4)
plot(datos, type = "l")
datosA = datos[1 : 30]
datosB = datos[31 : 60]
datos
datosA
datosB
vec_res = rep(0, 20)
factor_b <- -function(datosA, datosB, datos){
p1 <- -data.frame("num" = seq(1, length(datosA), 1), "datosA" = datosA)
p2 <- -data.frame("num" = seq(1, length(datosB), 1), "datosB" = datosB)
mp1 <- -inla(num f(datosA, model = "ar1"), data = p1, family = "poisson")
mp2 <- -inla(num f(datosB, model = "ar1"), data = p2, family = "poisson")
```

```
midf <- data.frame("num" = seq(1, length(datos), 1), "datos" = datos)
mp <- inla(num ~ f(datos, model = "ar1"), data = midf, family = "poisson")
respuesta <- as.numeric(mp1$mlik[1, 1]) + as.numeric(mp2$mlik[1, 1]) - as.numeric(mp$mlik[1, 1])
return(respuesta)
}
```

```
vec_res[1] = factor_b(datosA, datosB, datos)
vec_res
```

```
datosC = datosA[1 : 15]
datosD = datosA[16 : 30]
vec_res[2] = factor_b(datosC, datosD, datosA)
vec_res
datosE = datosB[1 : 15]
datosF = datosB[16 : 30]
vec_res[3] = factor_b(datosE, datosF, datosB)
vec_res
datosG = datosC[1 : 8]
datosH = datosC[9 : 15]
vec_res[4] = factor_b(datosG, datosH, datosC)
vec_res
datosI = datosD[1 : 8]
datosJ = datosD[9 : 15]
vec_res[5] = factor_b(datosI, datosJ, datosD)
vec_res
datosK = datosE[1 : 8]
datosL = datosE[9 : 15]
vec_res[6] = factor_b(datosK, datosL, datosE)
vec_res
datosM = datosF[1 : 8]
datosN = datosF[9 : 15]
vec_res[7] = factor_b(datosM, datosN, datosF)
vec_res
datosO = datosG[1 : 4]
datosP = datosG[5 : 8]
vec_res[8] = factor_b(datosO, datosP, datosG)
vec_res
datosQ = datosH[1 : 4]
datosR = datosH[5 : 7]
vec_res[9] = factor_b(datosQ, datosR, datosH)
vec_res
datosS = datosI[1 : 4]
datosT = datosI[5 : 8]
vec_res[10] = factor_b(datosS, datosT, datosI)
vec_res
datosU = datosJ[1 : 4]
datosV = datosJ[5 : 7]
vec_res[11] = factor_b(datosU, datosV, datosJ)
vec_res
datosX = datosK[1 : 4]
datosY = datosK[5 : 8]
vec_res[12] = factor_b(datosX, datosY, datosK)
```

```

vec_res
datosW = datosL[1 : 4]
datosZ = datosL[5 : 7]
vec_res[13] = factor_b(datosW, datosZ, datosL)
vec_res
datosAA = datosM[1 : 4]
datosBB = datosM[5 : 8]
vec_res[14] = factor_b(datosAA, datosBB, datosM)
vec_res
datosCC = datosN[1 : 4]
datosDD = datosN[5 : 7]
vec_res[15] = factor_b(datosCC, datosDD, datosN)
vec_res

```

Para utilizar otra función apriori se cambia en el programa por la siguiente instrucción, donde se incluye la distribución apriori a utilizar y sus parámetros, ya el paquete INLA tiene las distribuciones apriori a utilizar.

```

prec.prior <- list(prec = list(prior = "logtgaussian", param = c(0, 0, 0.001))) mp1 <-
  inla(num f(datosA, model = "ar1", hyper = prec.prior), data = p1, family = "poisson")

```