UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Tesis de Licenciatura

Algoritmos para detección de cambios en series de tiempo

Pablo Martín Herrera

Director: Juan Lucas Bali

2022

Agradecimientos

Primero a Lucas, no solo me ayudó un montón como matemático y director de tesis sino también como persona. Me llevo mucho esfuerzo terminar esta tesis y sin el apoyo de Lucas hubiese sido muchísimo mas difícil, siempre voy a decir lo mismo, tuve el mejor director que podía tener.

A mis padres por haberme brindado la mejor educación que pudieron y haberme apoyado en las decisiones que tome sobre mi carrera. A Sebastián Grynberg que me transmitió el amor que tenía por la matemática, de no haberlo conocido a él no estoy seguro de que hubiese llegado a hacer esta carrera.

A Diana por haberme ayudado y acompañado durate la carrera y estar siempre como mi amiga. A Euge por siempre darme una mano y enseñarme cosas, si no me hubiese dicho de hacer la tesis en este área no se como me hubiese recibido tampoco.

A todos mis ex compañeros de Despegar: René, Gise, Richard, Mauro, Lu que siempre estuvieron incentivandome para que me reciba y que me alegra mucho que me hayan quedado como amigues.

Índice general

Capítulo 1

Introducción

En este trabajo estudiaremos algunos métodos para resolver el problema de detección de *changepoints* en series de tiempo.

El análisis de series de tiempo es un tópico investigado intensamente. Los supuestos generales que subyacen en estas investigaciones son que las propiedades o parámetros que describen los datos son constantes o cambian lentamente con el tiempo. Es decir, se espera una cierta regularidad en la escala normalmente temporal. Por otro lado, surgen en la práctica muchos problemas en áreas como control de calidad, procesamiento de señales, detección de fallas y monitoreo en plantas industriales que pueden ser modelados con la ayuda de modelos perimétricos cuyos parámetros están sujetos a cambios abruptos en tiempos desconocidos. Por cambios abruptos nos refreímos a cambios en las características de la serie que ocurren rápidamente respecto al período de muestreo de las mediciones.

Por otro lado existen problemáticas en las cuales se quieren detectar cambios pero no es posible aplicar estos modelos. Al analizar series de tiempo, dos períodos cercanos pueden mostrar tendencias significativamente distintas. Los cambios de tendencia son comunes en series de tiempo climáticas y cruciales cunado se investiga cambio climático. Sin embargo existen relativamente pocos métodos para detectar cambios de tendencia, en particular cambios de tendencia online.

Una definición de changepoint en estas situaciones podría ser la siguiente, sea $\{x\}_t$ una serie de tiempo que venía distribuyéndose con la forma $x_t = \alpha + \beta t + \varepsilon_t$, donde β y α son parámetros desconocidos y $\varepsilon_t \sim \mathcal{N}(0, \epsilon)$ con ϵ positivo. Diremos que existe un changepoint si hay una variación por ejemplo en la pendiente β. Se analizará este caso posteriormente, precisando mejor esta definición.

Existen distintos métodos de detección cuando hablamos de changepoints. En algunos casos se busca detectar de forma *online*, en otros *offline*. La diferencia entre estos métodos es que el primero busca descubrir cambios "en tiempo real" y el segundo se aplica una vez que se tiene la serie de tiempo completa. Es decir, contamos con la ventaja de "saber el futuro", a diferencia de los esquemas online en donde se busca determinar el cambio en el menor tiempo posible. En estos casos la medida de performance debería incorporar esta demora, medida en cantidad de períodos, hasta poder detectar el changepoint. Por supuesto es también importante detectar un changepoint genuino (no tener falsos positivos) y no perder en la medida de lo posible ningún changepoint (falso negativo). En este estudio nos centraremos principalmente en el primer tipo de detección, el esquema online.

Estos problemas no son nada triviales, no solo por la dificultad que acarrea desarrollar un método para resolverlos sino porque tampoco resulta sencillo definirlos. Al plantearnos la pregunta ¿qué es un changepoint?, inmediatamente nos encontramos que podemos tener mas de una definición para esto. Como mencionamos anteriormente la noción de changepoint esta fuertemente relacionado con la problemática especifica a estudiar.

Observamos que cada método de detección cuenta con su propia definición de changepoint, que incluye la construcción del mecanismo generador de la serie de tiempo como así también lo que consistiría, para ese definición constructiva, un changepoint.

Esta tesis propone analizar dos métodos bastante utilizados para la detección de changepoints, radicalmente distintos en cuanto a su propuesta y definición de mecanismo generador, siendo el primero de ellos un esquema Bayesiano que puede resultar especialmente eficiente en el caso de contar con buena información *a priori* sobre las características generatrices de la serie de tiempo. El mismo se basa en el trabajo de [AM07]. El segundo método es conceptualmente más sencillo, orientado a la detección de cambios de tendencia en la serie, siendo una propuesta de raíz frecuentista con una propuesta orientada a un test de hipótesis, construido en función de la elaboración que se puede encontrar en [Zuo+19].

El presente trabajo de tesis se organiza como sigue: en el Capitulo 2 presentaremos una introducción teórica a la inferencia bayesiana, que servirá de base para la construcción del primer método que enfocaremos. En el capitulo 3 mencionaremos los elementos de la teoría de modelos gráficos que utilizamos para armar todo el modelo para el desarrollo teórico del algoritmo de detección bayesiana. En el Capitulo 4 desarrollaremos el modelo en profundidad y demostraremos de donde surgen los pasos que dan lugar al algoritmo, al finalizar realizaremos algunos ejemplos para testear su performance. En el Capitulo 5 introduciremos el marco teórico y desarrollaremos el algoritmo frecuentista para captar cambios de tendencia, finalizando con algunos ejemplos. En el Capitulo 6 estudiaremos dos casos de reales, por último en el Capitulo 7 daremos las conclusiones y los posibles pasos a seguir posteriores a este trabajo.

Nos gustaría terminar esta introducción remarcando donde consideramos que se encuentra el valor y el aporte de este trabajo. Gran parte de esta producción se basa en el trabajo de [AM07] en el que propone un algoritmo de detección de cambios con un enfoque bayesiano. Ese trabajo realiza una serie de deducciones que adolecen de una estructuración quizás no del todo formal o rigurosa, un aspecto que pretendemos acá completar a través de una formalización más precisa y rigurosa del esquema generador de datos, basado en modelos gráficos. Muchos de los pasos formales que se encuentran en ese trabajo son intuitivas, pero que no resultan tan sencillos de justificar de una forma directa y concluyente. Pretendemos con esta tesis suplir ese faltante con ese marco teórico requerido.

Es importante destacar esto pues consideramos que representa un aporte original en el área, a juzgar por la investigación bibliográfica realizada por el autor y el director de esta tesis.

Capítulo 2

Introducción a la inferencia Bayesiana

2.1. Motivación

El objetivo de este capitulo es dar una breve descripción de los principios y resultados de la inferencia Bayesiana que utilizaremos en esta tesis. Es decir, aplicado a la problemática de la detección de *changepoints* en una serie de tiempo. No es nuestro cometido meternos en los aspectos más intrínsecos, fundamentales o filosóficos de esta disciplina sino que nos abocaremos a repasar las definiciones y métodos más frecuentemente utilizados y que usaremos.

A modo de resumen, en el esquema frecuentista paramétrico usualmente se consideran $x_1, \ldots, x_n$ una muestra aleatoria, en donde $x_i \sim p_\theta$, siendo P una distribución que es perfectamente conocida salvo por el valor exacto de $\theta \in \Theta \subset \mathbb{R}^d$, el parámetro de la distribución.

En rigor de verdad podríamos considerar algo más general inclusive, teniendo $\mathbf{x} = (x_1, \ldots, x_n)$ un vector aleatorio tal que $\mathbf{x} \sim \mathbf{P}_\theta$. Es decir, se podría relajar las hipótesis de independencia e idéntica distribución.

La vuelta de tuerca que ofrece el esquema de inferencia Bayesiana es considerar que el parámetro es, a su vez, una variable aleatoria con una distribución que podría o no ser perfectamente conocida. Para simplificar la exposición supondremos que $\theta \sim p(\theta)$ tiene una distribución *a priori* completamente conocida. En algunos casos esto se puede relajar dando lugar por ejemplo a lo que se conoce como modelos jerárquicos. Omitiremos esas definiciones en lo que cabe en esta tesis.

2.2. Regla de Bayes

A continuación hablaremos sobre algunos conceptos específicos de la inferencia bayesiana. Más detalles sobre esta teoría se pueden encontrar en [Gel14].

Comenzaremos por hacer algunos comentarios sobre notación. Primero, $p(\cdot|\cdot)$ denotará una densidad condicional de probabilidad donde sus argumentos estarán determinados por el contexto. De la misma forma $p(\cdot)$ denotará una densidad marginal de probabilidad. Utilizaremos el termino 'densidad' y 'distribución' de forma intercambiable. Del mismo modo, la notación para distribuciones continuas y discretas será indistinta.

Para poder sacar conclusiones acera de θ, dado $\mathbf{x}$, debemos primero tener un modelo de probabilidad conjunta para ambos. La función de densidad conjunta puede ser escrita como el producto de dos funciones llamadas *distribución a priori*, que notaremos $p(\theta)$, y la *distribución de la muestra o verosimilitud*, denotada por $p(\mathbf{x}|\theta)$. Es decir,

$$p(\mathbf{x}, \theta) = p(\theta)p(\mathbf{x}|\theta).$$

Condicionando y despejando con respecto a $\mathbf{x}$ obtenemos la siguiente distribución conocida como *distribución posterior*:

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}, \theta)}{p(\mathbf{x})} = \frac{p(\theta)p(\mathbf{x}|\theta)}{p(\mathbf{x})}, \tag{2.1}$$

donde $p(\mathbf{x}) = \sum_\theta p(\theta)p(\mathbf{x}|\theta)$ es la suma (o integral en el caso continuo) sobre todos los valores de θ. Una forma equivalente a (2.1) omite el factor $p(\mathbf{x})$ ya que, fijado $\mathbf{x}$, este no depende de θ y puede ser considerado constante. Esto da lugar a la *distribución posterior no normalizada*:

$$p(\theta|\mathbf{x}) \propto p(\theta)p(\mathbf{x}|\theta). \tag{2.2}$$

El segundo termino en esta expresión $p(\mathbf{x}|\theta)$ se toma como una función de θ, no de $\mathbf{x}$. Estas expresiones forman parte del núcleo de la inferencia bayesiana, la tarea principal para cualquier aplicación es desarrollar un modelo para la distribución conjunta $p(\theta, \mathbf{x})$ y luego computar de forma apropiada $p(\theta|\mathbf{x})$.

2.3. Función de Likelihood/Verosimilitud

Una vez elegido el modelo de probabilidad la regla de Bayes nos dice que los datos $\mathbf{x}$ afectan a la distribución posterior (2.2) sólo a través de $p(\mathbf{x}|\theta)$. Cuando consideramos a esta última como función de θ fijando los valores de $\mathbf{x}$, la llamamos *función de verosimilitud*.

En lo que resta del capítulo utilizaremos el ejemplo detallado a continuación para hablar de las distintas distribuciones. Supongamos que tenemos una muestra aleatoria proveniente de una distribución normal $\mathcal{N}(\mu, \sigma^2)$ donde σ^2 es conocido y μ es el parámetro. La función de verosimilitud queda de la siguiente forma:

$$\begin{aligned} p(\mathbf{x}|\mu) &= \prod_{i=1}^{N} p(\mathbf{x}_i|\mu) \\ &= \prod_{i=1}^{N} \frac{1}{(2\pi\sigma^2)^{1/2}} \ \exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} \ \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right\}. \end{aligned}$$

2.4. Distribucion *a priori*

Como decíamos, al encarar un problema con el enfoque bayesiano el parámetro θ es considerado una variable aleatoria con una distribución conocida $p(\theta|\alpha)$, siendo α lo que llamaremos un *hiperparámetro*. Este se supone conocido e independiente de los datos observados.

En general, tanto la función de verosimilitud como la distribución a priori provienen del criterio utilizado para resolver el problema y responden a creencias previas del experimentador. Con frecuencia este criterio es revisado posteriormente para ver si es necesario modificar algún supuesto.

En el ejemplo anterior el parámetro del modelo θ es representado por μ. Este, a su vez, también sigue una distribución normal $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$ con función de densidad

$$p(\mu) = \frac{1}{(2\pi\sigma_0^2)^{1/2}} \ \exp\left\{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\}.$$

Los parámetros μ_0 y σ_0^2 son los hiperparámetros (conocidos) de la distribución a priori de μ. El valor μ_0 representa la "creencia" (subjetiva) sobre el posible valor de μ y σ_0 nos indica la incertidumbre que acarrea esa creencia. A valores altos de σ_0, la incertidumbre será mayor y por ende, a través de la verosimilitud tenderemos a "creer" menos del a priori y más de los datos.

2.5. Posterior: el problema de la estimación

En muchos casos el problema de la estimación consiste en determinar una distribución. Para tal fin, se caracteriza lo que hemos dado a conocer como la distribución posterior (2.1), esta representa la actualización de la distribución a priori en virtud de haber observado $\mathbf{x}$. Es decir, se modifican las "creencias" previas del investigador en función de los nuevos resultados que aportan los datos $\mathbf{x}$.

En nuestro ejemplo al multiplicar la distribución a priori por la verosimilitud para obtener la distribución posterior, el resultado es nuevamente una distribución normal sobre la variable μ

$$
\begin{aligned}
p(\mu|\mathbf{x}) &= p(\mathbf{x}|\mu,\sigma)p(\mu) \\
&= \frac{1}{(2\pi\sigma^2)^{N/2}(2\pi\sigma_0^2)^{1/2}} \ \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 + \frac{1}{\sigma_0^2}(\mu-\mu_0)^2\right)\right\}.
\end{aligned}
$$

Completando cuadrados en la expresión anterior obtenemos que la distribución posterior depende de $\mathbf{x}$ solo a través de la media muestral $\bar{x} = \frac{1}{n}\sum_i x_i$. Lo cual quiere decir que $\bar{x}$ es el estadístico suficiente para esta distribución, hablaremos de esta idea mas adelante de esto cuando mencionemos las distribuciones a priori conjugadas. Los detalles más finos de la cuenta se pueden ver en [Gel14]. De esta forma obtenemos

$$p(\mu|\mathbf{x}) \propto exp\left\{-\frac{1}{2\sigma_n}(\mu-\mu_n)\right\}$$

donde

$$\mu_n = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\sigma_0} + \frac{n}{\sigma^2}} \quad \text{y} \quad \frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}.$$

Cuando la distribución posterior y la distribución a priori tienen la misma forma funcional decimos son distribuciones *conjugadas*.

2.6. El problema de la predicción

A diferencia de la estimación, al enfrentarnos con el problema de la predicción no buscamos determinar el valor puntual o de la distribución de θ, sino de un posible valor futuro x_{n+1} habiendo observado una muestra $x_1, \ldots, x_n$.

Antes de que los datos $\mathbf{x}$ sean considerados, la distribución de un dato desconocido (pero observable) x es la siguiente

$$p(x) = \int p(x,\theta)d\theta = \int p(x|\theta)p(\theta)d\theta.$$

Notemos que aquí hemos usado que $p(x|\theta,\alpha) = p(x|\theta)$ ya que conociendo un valor de θ no es necesario saber nada en particular sobre el hiperparámetro α. En general, obviaremos escribir el hiperparámetro para simplificar la notación a menos que por contexto sea necesario que se explicite el mismo.

A esta distribución se la suele llamar *marginal de x*, pero también se le puede dar el nombre *distribución predictiva a priori*: a priori porque no esta condicionada sobre ningún dato anterior y predictiva porque es la distribución de un dato observable.

Supongamos que tenemos $\mathbf{x} = (x_1, \ldots, x_n)$ datos observados de una muestra aleatoria, con $x_i \sim p(x_i|\theta)$, con $\theta \sim p(\theta)$ la distribución a priori del parámetro.

La distribución *posterior predictiva* consistirá en la distribución de una nueva observación, x_{n+1}, habiendo observado $x_1, \ldots, x_n$, sin imponer ningún tipo de supuesto sobre el parámetro θ. Es decir, queremos obtener la distribución $p(x_{n+1}|\mathbf{x})$, de un dato futuro dados los datos ya observados.

Haremos un esbozo de la cuenta que habría que hacer para obtener esta distribución

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

Cabe destacar que en el último paso simplificar $\mathbf{x}$ en la condición proviene de suponer que dado el parámetro θ, x_{n+1} es independiente de los datos anteriores. Notemos que con esta expresión podemos establecer criterios de predicción de futuros valores de x_{n+1}, y a la vez con la distribución posterior podemos establecer una estimación sobre los posibles valores de θ.

De esta manera tenemos una distribución para caracterizar un próximo valor de la sucesión en función de los anteriores. La inferencia predictiva permite generalmente la construcción de intervalos de predicción para una observación futura.

En el caso del ejemplo nos quedaría de esta forma

$$
p(x_{n+1}|\mathbf{x}) = \int p(x_{n+1}|\mu, \sigma^2)p(\mu|\mu_N, \sigma_N)d\mu,
$$

como la posterior y la a priori son gaussianas podemos utilizar el siguiente resultado para distribuciones normales multivariadas.

Lema 1. *Sean $\boldsymbol{x}$ e $\boldsymbol{y}$ dos vectores aleatorias tales que $\boldsymbol{x} \sim \mathcal{N}(\mu, \Sigma_x)$ e $\boldsymbol{y}|_x \sim \mathcal{N}(A^\intercal \boldsymbol{x} + b, \Sigma_{y_{|x}})$ entonces $y \sim \mathcal{N}(A^\intercal \mu + b, \Sigma_{y_{|x}} + A^\intercal \Sigma_x A)$.*

Si reemplazamos por las variables de nuestro ejemplo

$$
x = \mu,\ \mu = \mu_N,\ \xi = \sigma_N^2,\ y = x_{n+1},\ A^\intercal = 1,\ b = 0,\ \Sigma_{y_{|x}} = \sigma
$$

obtenemos que

$$
p(x_{n+1}|\mathbf{x}) \propto exp\left\{ -\frac{1}{2(\sigma^2 + \sigma_N)}(x_{n+1} - \mu_N)^2 \right\}.
$$

2.7. Priors Conjugadas

Como ya mencionamos, la distribución a priori $p(\theta)$ es conjugada de la función de verosimilitud $p(\mathbf{x}|\theta)$ si la distribución posterior $p(\theta|\mathbf{x})$ tiene su misma forma funcional.

Que dos distribuciones sean conjugadas tiene ventajas a la hora de computar la posterior. En el ejemplo vimos un caso particular de esta situación cuando la distribución a priori y la verosimilitud son normales. Sucede que las funciones pertenecientes a *familias exponenciales* tienen distribuciones conjugadas naturales. Para encontrar información más detallada acerca de a prioris conjugadas referimos al lector a [Rai00].

Por familias exponenciales nos referimos a la clase $\mathcal{F}$ de funciones de la forma

$$p(x_i|\theta) = f(x_i)g(\theta)\exp\left\{\phi(\theta)^T u(x_i)\right\}$$

donde x_i y θ son variables multidimensionales.

Los factores $\phi(\theta)$ y $u(x_i)$ son en general vectores de la misma dimensión que θ. El vector $\phi(\theta)$ se denomina *parámetro natural* de la familia $\mathcal{F}$. La función de verosimilitud correspondiente a una secuencia $\mathbf{x} = (x_1, x_2, ..., x_n)$ de observaciones independientes e idénticamente distribuidas es

$$p(\mathbf{x}|\theta) = \left\{\prod_{i=1}^{n} f(x_i)\right\} g(\theta)^n \exp\left\{\phi(\theta)^T \sum_{i=1}^{n} u(x_i)\right\}.$$

Para todo n y $\mathbf{x}$, esta expresión tiene una forma fija (como función de θ)

$$p(\mathbf{x}|\theta) \propto g(\theta)^n \exp\left\{\phi(\theta)^T t(\mathbf{x})\right\},$$

donde $t(\mathbf{x}) = \sum_{i=1}^{n} u(x_i)$ es conocido como *el estadístico suficiente* para θ. Suficiente, porque la función de verosimilitud para θ depende de $\mathbf{x}$ solo a través de $u(x_i)$. Los estadísticos suficientes son útiles a la hora de hacer manipulaciones algebraicas en la distribución posterior y la verosimilitud. Si la densidad a priori la podemos escribir como

$$p(\theta) \propto g(\theta)^\eta \exp\{\phi(\theta)^T \nu\},$$

entonces la distribución posterior es igual a

$$p(\theta|\mathbf{x}) \propto g(\theta)^{\eta+n} \exp\{\phi(\theta)^T (\nu + t(\mathbf{x})\}.$$

Lo que muestra que esta elección de distribución a priori es conjugada. Existen pocos ejemplos conocidos de otras distribuciones con conjugadas naturales, este hace que las familias exponenciales tengan una gran relevancia en la estadística bayesiana.

2.8. Ejemplos de modelos de muestras aleatorias normales

Parámetro generativo de la muestra x_t: μ (media)

- Función de verosimilitud: Normal con varianza σ^2 conocida.
- Distribución a priori conjugada: Normal con hiperparámetros μ_0, σ_0^2.
- Distribución posterior: Normal con parámetros

$$\mu' = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^n x_i}{\sigma^2} \right), \sigma^{2'} = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1}.$$

- Distribución posterior predictiva: $x_{n+1}|_{\mathbf{x}} \sim \mathcal{N}(\mu', \sigma^{2'} + \sigma^2)$.

Parámetro generativo de la muestra x_t: σ^2 (varianza)

- Función de verosimilitud: Normal con media μ conocida.
- Distribución a priori conjugada: Gamma Inversa con hiper parámetros α, β

$$f(\sigma^2; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\sigma^2} \right)^{\alpha+1} exp \left(\frac{-\beta}{\sigma^2} \right).$$

- Distribución posterior: Gamma Inversa con parámetros

$$\alpha' = \alpha + \frac{n}{2}, \quad \beta' = \beta + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2}.$$

- Distribución posterior predictiva: $x_{n+1}|_{\mathbf{x}} \sim t_{2\alpha'}(\mu, \sigma^2 = \frac{\beta'}{\alpha'})$.

Parámetros generativos de la muestra x_t: μ (media) y σ^2 (varianza)

- Función de verosimilitud: Normal.
- Distribución a priori conjugada: Normal-inversa-gamma con parámetros $\mu_0, \nu, \alpha, \beta$

$$f(\mu, \sigma; \mu_0, \nu, \alpha, \beta) = \frac{\sqrt{\nu}}{\sigma\sqrt{2\pi}} \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} exp\left(-\frac{2\beta + \nu(\mu - \mu_0)^2}{2\sigma^2}\right).$$

- Distribución posterior: Normal-inversa-gamma con parámetros

$$\mu_0' = \frac{\nu\mu_0 + n\bar{x}}{\nu + n}, \quad \nu' = \nu + n$$

$$\alpha' = \alpha + \frac{n}{2}, \quad \beta' = \beta + \sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{n\nu}{\nu + n}\frac{(\bar{x} - \mu_0)^2}{2}.$$

- Distribución posterior predictiva: $x_{n+1}|_{\mathbf{x}} \sim t_{2\alpha'}(\mu', \frac{\beta'(\nu'+1)}{\nu'\alpha'})$.

Capítulo 3

Introducción a Modelos Gráficos

3.1. Introducción

Un modelo gráfico es un modelo probabilístico para el cual la estructura de independencia condicional está codificada en un grafo. En un modelo gráfico, los vértices (o nodos) representan variables aleatorias, y las aristas codifican relaciones de independencia condicional entre los vértices asociados. El grafo caracteriza la forma en que la distribución conjunta se factoriza en el producto de muchas componentes menores, cada una de ellas contiene sólo un subconjunto de variables. En este capítulo vamos a introducir *modelos graficos dirigidos aciclicos* (que abreviaremos DAGs), en los cuales las aristas son dirigidas, es decir, son flechas. Un ejemplo de un modelo gráfico dirigido se muestra en la figura a continuación en la Figura 3.1.

$$X \longleftarrow Y \longrightarrow Z$$

Figura 3.1: un *grafo dirigido con vértices* $V = \{X,Y,Z\}$ *y aristas* $E = \{(Y,X), (Y,Z)\}$.

A continuación daremos algunas definiciones que nos servirán para nuestro problema en cuestión, se puede complementar esta teoría con lo que se puede encontrar en [Edw95] y en la tesis de licenciatura de Violeta Roizman [Roi17]. Definamos entonces independencia condicional:

Definición 1. *Sean X, Y y Z variables aleatorias. Entonces X e Y son* ***condicionalmente independientes*** *dado Z, y lo escribiremos como $X \perp Y|Z$, si*

$$p(x, y|z) = p(x|z)p(y|z)$$

para todo x,y y z.

Intuitivamente esto quiere decir que una vez conocido Z, Y no provee ninguna información extra acerca de X. Una definición equivalente es que $p(x|y, z) = p(x|z)$ para todo x,y y z.

Los grafos dirigidos son útiles para representar relaciones de independencia condicional entre variables. Algunos textos utilizan el término *redes Bayesianas* para referirse a grafos dirigidos dotados de una distribución de probabilidad. Sin embargo este nombre no es del todo acertado ya que se pueden usar métodos frecuentistas o Bayesianos para realizar inferencia estadística sobre grafos dirigidos.

3.1.1. Grafos Dirigidos Acíclicos (DAGs)

Formalmente, un **grafo dirigido** consiste en un conjunto de vértices V y un conjunto de aristas E conformado por pares ordenados de vértices. Es decir, $E \subset V \times V$. En nuestro caso, cada vértice corresponderá a una variable aleatoria. Si $(Y, X) \in E$ entonces hay una flecha que apunta de Y a X como muestra la Figura 3.1.

Si una flecha conecta dos variables X e Y (en cualquier sentido) decimos que X e Y son **adyacentes**. Si hay una flecha de X a Y entonces X es un **padre** de Y. El conjunto de todos los padres de X lo denotaremos como π_X. Un **camino dirigido** entre dos variables o vértices es una sucesión de aristas/flechas que apuntan en la misma dirección uniendo una variable con la otra como se muestra en la Figura 3.2.

Es decir, un camino dirigido es una sucesión $v_1, \ldots, v_n \subset V$ tal que $(v_k, v_{k+1}) \in E$ para todo $1 \leq k \leq n-1$.

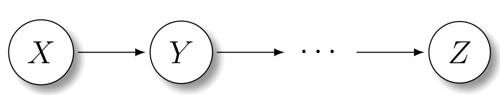


Figura 3.2: *un grafo con un camino dirigido.*

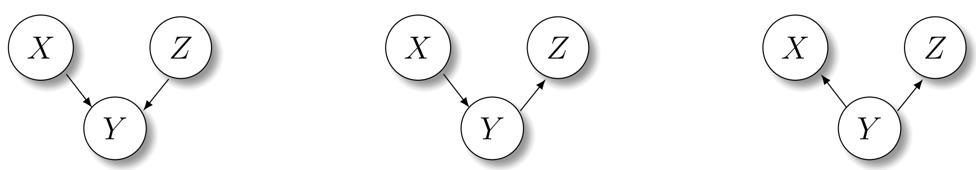
Una secuencia de vértices adyacentes que comienza en X y termina en Y pero ignorando la dirección de las flechas se llama **camino no dirigido**.

Un **circuito dirigido** o bien **ciclo dirigido** es un camino dirigido $v_1, \ldots, v_n$ tal que $v_1 = v_n$. Diremos que un grafo dirigido es **acíclico** si no contiene ciclos dirigidos de más de un vértice.

Desde ahora en adelante trabajaremos con grafos dirigidos acíclicos, un **DAG** por sus siglas en inglés, ya que es difícil proveer una semántica de probabilidad coherente para grafos dirigidos con ciclos, y tampoco necesitaremos hacer uso de ellos.

X es un **ancestro** de Y si existe un camino dirigido de X a Y. También decimos en este caso que Y es un *descendiente* de X. En el caso particular en que X es adyacente a Y, diremos que X es un **padre** de Y. Notaremos como π_Y al conjunto de padres de Y, los antecesores inmediatos.

Hay tres configuraciones básicas de conexiones de subgrafos con tres nodos, que permiten construir grafos mas grandes a partir de ellas. Una configuración de la forma de la Figura 3.3(a) se llama **colisionador** en Y (conexión cabeza-a-cabeza). Cuando una configuración no tiene esa forma se llama **no-colisionador** (conexión cabeza-a-cola o cola-a-cola), por ejemplo la Figura 3.3(b) y la Figura 3.3(c).



(a) (b) (c)

Figura 3.3: *(a) un colisionador en Y, (b), (c) no-colisionadores.*

La propiedad de colisionador es dependiente del camino. En la Figura 3.4, Y es un colisionador en el camino $\{X, Y, Z\}$ pero un no-colisionador en el camino $\{X, Y, W\}$.

Cuando las variables que apuntan a un colisionador no son adyacentes, decimos que el colisionador esta **desprotegido**.

Dos ejemplos muy importantes de DAGs son las cadenas de Markov y los Modelos de Markov Ocultos.

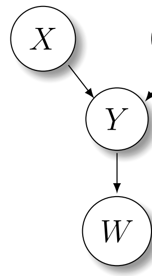


Figura 3.4: *un colisionador con un descendiente*

Cadenas de Markov

Sea G una cadena (grafo) como en la Figura 3.5. Entonces tenemos que $p(x) = p(x_0)p(x_1|x_0)p(x_2|x_1)...$. La distribución de cada variable depende solo del predecesor inmediato. Decimos que P es una cadena de Markov.

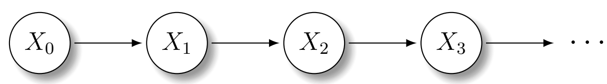


Figura 3.5: *Cadena de Markov.*

La idea de una cadena de Markov es representar una sucesión de eventos en donde el estado de la variable X_{t-1} (es decir, el estado a tiempo $t-1$) induce la distribución de X_t de forma independiente al resto de las variables anteriores.

Modelo de Markov oculto (HMM)

Un modelo oculto de Markov (HMM) involucra 2 conjuntos de variables $X_0, X_1, X_2, X_3, ...$ e $Y_0, Y_1, Y_2, Y_3, ...$ Las X_i's forman una cadena de Markov pero se suponen no observables. Las Y_i's son observables pero la distribucion de Y_i depende solo de X_i.

Un problema clásico de las cadenas ocultas de Markov es, habiendo observado una sucesión de observables $y_1, \ldots, y_n$, poder determinar el valor más probable de $x_1, \ldots, x_n$. Hay extensa bibliografía que trata estos temas, por ejemplo [Dym11].

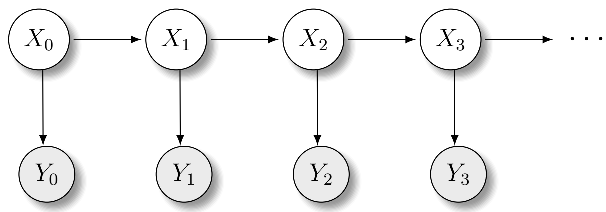


Figura 3.6: *Modelo oculto de Markov.*

3.1.2. Probabilidad y DAGs

Dotaremos de una semántica probabilística a los DAGs que hemos definido anteriormente.

Sea G un DAG con vértices $V = (X_1, ..., X_d)$. Para simplificar notación a veces representaremos $V = \{x_1, ..., x_d\}$. Si P es una distribución para V con función de probabilidad o de densidad $p(x)$, decimos que P es **Markov** a G, o que G **representa** a P si:

$$p(x) = \prod_{j=1}^{d} p(x_j|\pi_{x_j})$$

donde π_{x_j} es el conjunto de nodos padres de x_j. El conjunto de distribuciones que son representadas por G se nota $\mathcal{M}(G)$.

Lo interesante es que la probabilidad de un estado dado del grafo está determinado por la probabilidad del estado de cada vértice condicionado exclusivamente a sus antecesores inmediatos, sus padres. Pensemos que en una cadena de Markov es exactamente lo que pasa, puesto que la probabilidad de una sucesión $x_1, \ldots, x_n$ termina siendo $p(x_1)p(x_2|x_1)\ldots p(x_n|x_{n-1})$. Pero es ese un caso particular de markovianidad, en el sentido más general.

Por ejemplo, en la Figura 3.4 $P \in \mathcal{M}(G)$ si y sólo si su función de probabilidad $p(\cdot)$ se factoriza de la forma $p(x, y, z, w) = p(x)p(z)p(y|x, z)p(w|y)$.

El teorema siguiente dice que $P \in M(G)$ si y sólo si se cumple la **condición de Markov**. Dicho coloquialmente, la condición de Markov significa que cada variable W es independiente del "pasado" dado sus padres.

Teorema 2. *Para un grafo $G = (V, E)$, una distribución $P \in \mathcal{M}(G)$ si y solo si la siguiente condición de Markov se cumple para toda variable W:*

$$W \perp W'|\pi_W$$

donde W' denota todos las otras variables excepto los padres y los descendientes de W.

No es la intención de esta sección demostrar todos los resultados que vamos a enunciar, si dejaremos referencias en la bibliografía acerca de modelos gráficos. Vamos a enumerar todos los resultados que usaremos en el capitulo siguiente para demostrar como llegar al algoritmo de detección de changepoints.

Veamos un ejemplo, consideremos el siguiente DAG en la Figura 3.7 en este caso la función de probabilidad

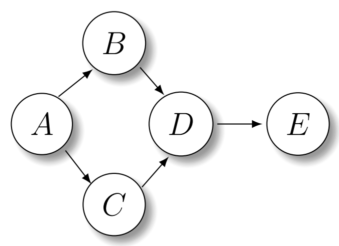


Figura 3.7: *D un colisionador , B y C no-colisionadores.*

se debe factorizar como $p(a, b, c, d, e) = p(a)p(b|a)p(c|a)p(d|b, c)p(e|d)$.

La condición de Markov nos dice que valen las siguientes independencias condicionales:

$$D \perp A \mid \{B, C\}, \quad E \perp \{A, B, C\} | \; D \;\; y \;\; B \perp C | \; A.$$

3.1.3. Más relaciones de independencia

La condición de Markov nos permite listar algunas relaciones de independencia condicional implicadas por un DAG. Estas relaciones pueden implicar otras relaciones de independencia. Por ejemplo en la siguiente figura:

resulta que (pero no es obvio) que las relaciones en este grafo implican que

$$\{X_4, X_5\} \perp X_2 \mid \{X_1, X_3\}$$

Ahora la pregunta es: ¿como hallamos estas relaciones? La respuesta: *d-separación*, que es una abreviación para **separación directa**. Vamos a

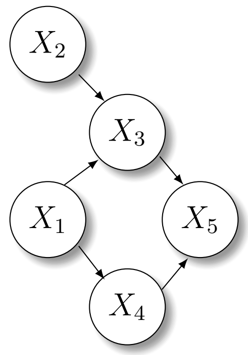


Figura 3.8: *otro DAG*

definir la noción de d-conexión primero, y diremos que dos vértices están **d-separados** si no hay ningún camino que los d-conecte.

Sea G un grafo dirigido, decimos que un camino π **d-conecta** a dos vértices X e Y condicionalmente a un conjunto C de vértices que no los contiene si ambos son extremos del camino y además:

1. todo vértice no-colisionador en π no pertenece a C y

2. todo vértice colisionador en π es un antecesor de C (es decir, antecesor de algún vértice de C) o pertenece a C.

Si no existe ningun camino que d-conecte a X y Y dado C, decimos que estos estan d-separados dado C. Sean dos conjuntos no vacíos de vértices A y B, estos estan d-separados dado C si para todo $x \in A$, $y \in B$, x e y estan d-separados dado C.

Consideremos la siguiente Figura 3.9.

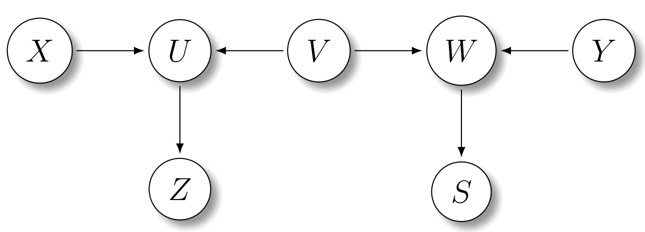


Figura 3.9: Ejemplo para d-separación

Podemos deducir de la definición de d-separación (o d-conexión) que:

- X e Y están d-separados dados el conjunto vacío, pues el único camino entre X e Y es: $U : X \to U \leftarrow V \to W \leftarrow Y$ que no cumple que todo vértice colisionador es antecesor o pertenece al conjunto vació y por lo tanto no d-conecta.

- X e Y están d-conectados dados $\{Z, S\}$, ya que los colisionadores en U son Z y W que son antecesores de $\{Z, S\}$ y V no pertenece a $\{Z, S\}$.

- X e Y es tan d-separados dado $\{Z, S, V\}$, pues el único no-colisionador en U es V y este se encuentra en $\{Z, S, V\}$.

Para concluir esta sección enunciaremos un teorema que vamos a utilizar en el capitulo siguiente para poder deducir una independencia condicional, podemos encontrar mas al respecto en [Pea00].

Teorema 3. *Sean A,B,C conjuntos disjuntos de vértices en un DAG. Si A y B están d-separados por C, entonces A es independiente de B condicional a C en toda distribución compatible con el DAG. Es decir,*

$$A \perp B \mid C. \tag{3.1}$$

En nuestro ejemplo anterior esto nos dice que se cumple la siguiente relación de independencia condicional

$$X \perp Y \mid \{Z, S\}$$

En el capitulo siguiente este resultado sera clave para demostrar uno de los pasos del algoritmo propuesto por [AM07].

Capítulo 4

Propuesta Bayesiana

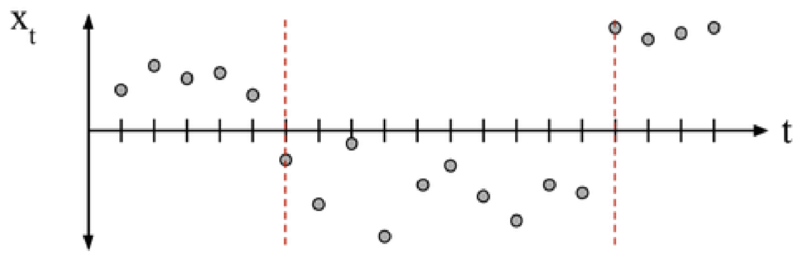
4.1. Introducción

Presentaremos un modelo de series de tiempo basado en la construcción de [AM07], en pos de completar los aspectos formales omitidos en dicho trabajo, con un enfoque basado en de modelos gráficos introducida en en el capítulo anterior. Nos centraremos en identificar cambios en los parámetros generativos de una secuencia de datos. Cuando se produzca este fenómeno diremos que ha ocurrido un changepoint. Esta construcción presupone que los datos de la serie de tiempo son independientes y que van siendo generados a partir de una distribución fija con un parámetro también fijo. En el momento en que ocurre un changepoint, este parámetro podría verse alterado en función de una distribución *prior* que se supondrá conocida y que reflejará parte del conocimiento que se tiene a priori del fenómeno por parte del investigador. De ahí la naturaleza bayesiana si se quiere de este enfoque.

Pasemos a formalizar las ideas.

Asumiremos que una secuencia de observaciones $x_1, x_2, x_3, ..., x_T$ está dividida en un producto de intervalos no solapados I_k con $k \in \{1, 2, ..., n\}$. Los changepoints se producen en los bordes de estos intervalos. Para cada intervalo I_k los datos $(x_t)_{t \in I_k}$ son variables aleatorias independientes idénticamente distribuidas para alguna familia de distribución $p(x_t|\theta_{I_k})$. Bajo este modelo tendremos un mecanismo generador de parámetros θ_{I_k} con distribución $p_\alpha(\theta_{I_k})$ siendo α un hiper-parámetro fijo.

A modo de ejemplo supongamos que tenemos una secuencia de datos x_t tal que $x_t|_{\theta_{I_k}} \sim N(\theta_{I_k}, \sigma_0^2)$ con σ_0^2 un valor fijo y conocido. Diremos que el



parámetro generativo de esta serie de tiempo es la media de una distribución normal, con varianza conocida.

A su vez θ_{I_k} tendrá una distribución normal que dependerá de un hiperparámetro conocido en este caso, $\alpha = (\mu_1, \sigma_1^2)$ o sea que $\theta_{I_k} \sim N(\mu_1, \sigma_1^2)$.

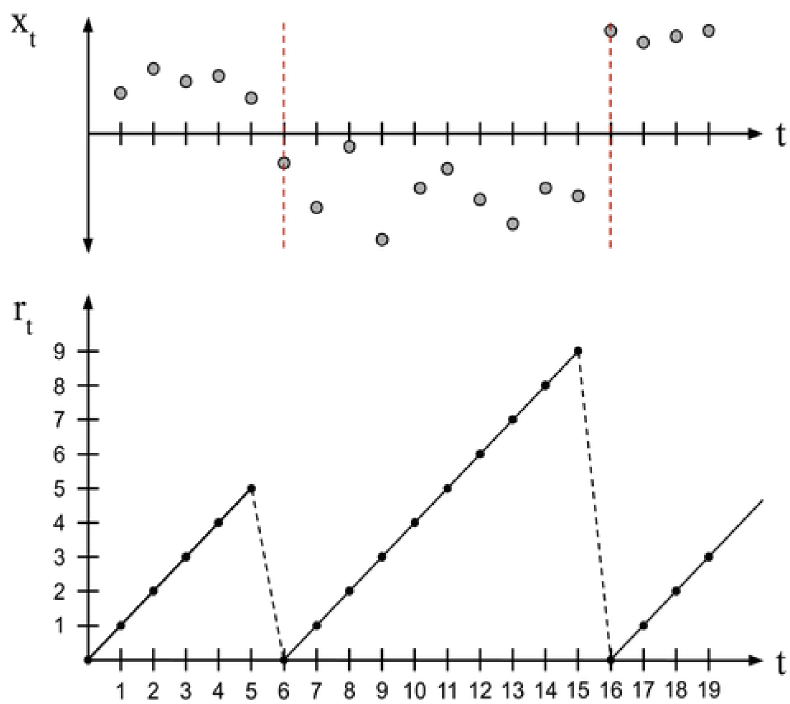
Cada intervalo I_k contará con un parámetro generativo θ_{I_k} que pensaremos de la siguiente forma: para cada tiempo t tendremos un parámetro generativo θ_t y será tal que $\theta_{I_k} = \theta_t = \theta_l$ si $t, l \in I_k$. De ahora en mas hablaremos de θ_t que será el parámetro generativo de x_t a tiempo t.

Con cierta probabilidad, que describiremos más adelante, ocurrirán changepoints que obligarán a "rebarajar" la distribución a priori para generar nuevos parámetros θ_t para las observaciones que vienen a continuación. Diremos que entre el tiempo t y $t+1$ se produjo un *changepoint* si existe una variación entre los parámetros generativos θ_t y θ_{t+1}. Es decir, entre ambos tiempos el parámetro θ_{t+1} volvió a ser generado por la distribución a priori $p_\alpha(\theta)$. El nuevo parámetro puede ser igual al anterior (lo será con probabilidad $p_\alpha(\theta_{t+1} = \theta_t)$) o no, en ambos casos igual diremos que se producirá un changepoint. Solo que de ser iguales no habria forma de detectarlo, pero ese caso no es en general relevante, por ejemplo si la distribución generadora $p_\alpha(\theta)$ es continua en cuyo caso la probabilidad de que en un rebarajamiento se obtenga el mismo parámetro será 0.

Así pues, observando los valores x_t buscamos determinar los tiempos t que son candidatos de estar sujetos a un cambio de distribución.

Notemos que no necesariamente vamos a suponer que x_t sea una variable continua o discreta, o incluso por fuera de estas dos clases, aunque a fin de simplificar la exposición supondremos algunos de estos dos casos aunque es posible generalizar las ideas usando la adecuada definición abstracta de medida apropiada según el contexto. Lo mismo pasará con la distribución de θ_t según el a priori.

Vamos a introducir ahora una de las variables mas importantes en nuestro



modelo. Denominaremos **runlength** (r_t) al tiempo transcurrido desde el último changepoint al finalizar el tiempo t (lo cual significa después de que haya sido relevado el dato x_t). Puede tomar dos valores:

$$r_t = \begin{cases} 0 & \text{si existe un changepoint al finalizar el tiempo t.} \\ r_{t-1}+1 & \text{caso contrario.} \end{cases}$$

Así pues, r_t es una variable en general creciente que cae bruscamente a cero en los changepoints.

Es importante aclarar que $r_t = 0$ quiere decir que el punto x_t es el último elemento de la sección actual de la partición y el punto x_{t+1} (todavía no relevado) es el primer punto de la nueva sección. Los gráficos que agregamos son iguales a los del paper (para mantener la fidelidad), en realidad en nuestro modelo la caída a 0 es un tiempo antes.

Prior de Runlength

Vamos a mencionar algunas cuestiones acerca de la distribución de r_t ya que su carácter recursivo es lo que le da potencia al algoritmo que veremos en este capítulo.

Supondremos que la sucesión de observaciones empieza con un changepoint, es decir $r_0 = 0$, al no haber observaciones anteriores esto no implica una pérdida de generalidad en nuestro modelo.

El primer supuesto será que los r_t forman una cadena de Markov. Eso significa que la distribución condicional $p(r_t|r_0, \ldots, r_{t-1}) = p(r_t|r_{t-1})$. Tenemos dos transiciones posibles de r_{t-1} a r_t, o bien seguimos en la misma partición y $r_t = r_{t-1} + 1$ u ocurre un changepoint, con lo cual r_t pasará a valer cero.

$$P(r_t = \ell|r_{t-1}) = \begin{cases} H(t, r_{t-1} + 1) & \text{si } \ell = 0 \\ 1 - H(t, r_{t-1} + 1) & \text{si } \ell = r_{t-1} + 1 \\ 0 & \text{en otro caso.} \end{cases}$$

La función $H(t, k)$ caracteriza la probabilidad de que ocurra un changepoint a tiempo t, habiendo alcanzado longitud k en el runlength actual.

Introducimos como hipótesis adicional la **homogeneidad** de la cadena de Markov. Esto nos permite independizar H con respecto al tiempo t. Es decir $H(t, k) = H(k)$ esta función se conoce como *Hazard* (proveniente de teoría de análisis de supervivencia, para más información consultar [KK12]) pues representa probabilidad de que un proceso ocurra a tiempo t dado que no ocurrió todavía a tiempo $t - 1$.

Más aún vamos a suponer que H es constante. No depende del valor del runlength actual y que es un valor fijo $\frac{1}{\lambda}$. Esta es una hipótesis importante a la hora de probar independencias acerca de r_t con otras variables.

4.2. Arquitectura del Modelo

Emplearemos el modelo gráfico representado de la Figura 4.1. Tiene algunas similitudes con el hidden Markov model de la Figura. 3.6.

Los únicos observables en este modelo (sombreados con gris) serán los x_t, cuya distribución solo dependerá de θ_t. Bajo este criterio podemos calcular la distribución x_1 a modo de ejemplo valiéndonos de probabilidad total

$$p(x_1) = \int_{\Theta} p(x_1, \theta_1) d\theta_1 = \int_{\Theta} p(x_1|\theta_1) p(\theta_1) d\theta_1.$$

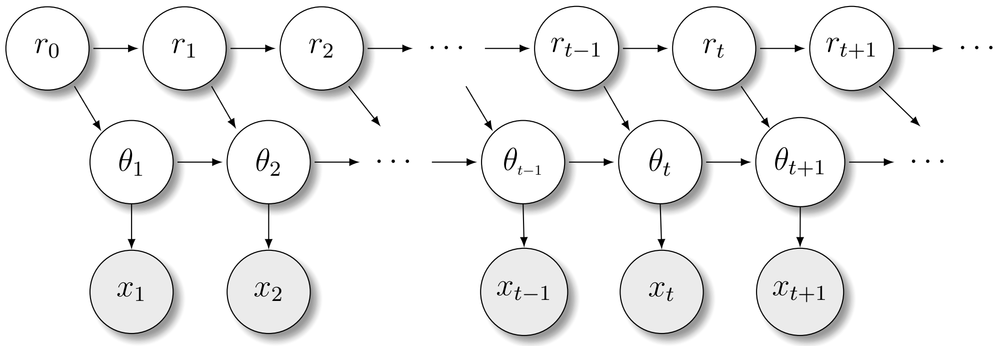


Figura 4.1: Modelo gráfico del esquema bayesiano de changepoints.

Tengamos presente la condición de Markov que utilizaremos a lo largo de todo el capítulo.

Para un grafo $G = (V, E)$, recordemos que una distribución $P \in \mathcal{M}(G)$ si y sólo si la siguiente condición se cumple para toda variable W

$$W \perp W' | \pi_W$$

donde W' denota cualquier variable excepto los padres y los descendientes de W.

Esta propiedad implica para nuestro modelo las siguientes independencias condicionales:

1. $x_t \perp W'_{x_t} | \pi_{x_t} \quad \text{donde} \quad \pi_{x_t} = \{\theta_t\}.$

2. $r_t \perp W'_{r_t} | \pi_{r_t} \quad \text{donde} \quad \pi_{r_t} = \{r_{t-1}\}.$

3. $\theta_t \perp W'_{\theta_t} | \pi_{\theta_t} \quad \text{donde} \quad \pi_{\theta_t} = \{\theta_{t-1}, r_{t-1}\}.$

En cada caso W'_* es distinto de los padres y los descendientes de la variable correspondiente.

Agregaremos una hipótesis extra que resulta propia del modelo que queremos proponer:

$$\theta_t \perp \theta_{t-1} \mid \{r_{t-1} = 0\}.$$

Asumimos esto ya que en nuestro modelo una vez ocurrido un changepoint el parámetro θ se vuelve a generar a partir de la distribución a priori, independientemente del valor que tenia el parámetro anteriormente. Una formulación equivalente que terminaremos usando es la siguiente

$$\theta_t \perp W'_{\theta_t} \cup \theta_{t-1} \mid \{r_{t-1} = 0\}. \tag{4.1}$$

Demostración.

$$\begin{aligned} p(\theta_t | W'_{\theta_t}, \theta_{t-1}, r_{t-1} = 0) &= p(\theta_t | \theta_{t-1}, r_{t-1} = 0) \\ &= p(\theta_t | r_{t-1} = 0) \end{aligned}$$

Para la primera igualdad aplicamos Markov ya que $\{r_{t-1}, \theta_{t-1}\}$ son los padres de θ_t. Luego simplemente utilizamos la hipótesis 4.1. □

Esta es la versión de la hipótesis que utilizaremos para demostrar los pasos del algoritmo en la sección siguiente.

4.3. Algoritmo

4.3.1. Descripción informal

Queremos entonces llegar a un algoritmo *online* que nos permita detectar cuando se produce un changepoint. Comenzaremos con una lista que tendrá en su primera posición el valor 0, en la cual iremos guardando los candidatos a checkpoints que detecte el algoritmo. Recordemos que sin perdida de generalidad comenzamos suponiendo que arrancamos en un changepoint a tiempo 0.

En cada iteración el algoritmo relevará un nuevo dato x_t, luego calculará para cada posible valor de r_t (que puede tomar valores entre 0 y t) las probabilidades $p(r_t|x_{1:t})$. El valor de r_t para el cual esta probabilidad se maximice será el que supondremos como el valor del actual runlength.

Supongamos que estamos a tiempo $t = 20$, se releva x_{20} y al calcular $p(r_t|x_{1:t})$ el valor que maximiza la probabilidad es $r_t = 3$, esto querrá decir que asumiremos que en el tiempo $t = 17$ se produjo el último changepoint. En ese momento el algoritmo chequea si $t = 17$ se encuentra en la lista de valores de posibles checkpoints, de no ser así lo agrega y comienzo la próxima iteración para $t = 21$.

Tengamos presente que en la practica la probabilidad de que $r_t = 0$ no suele maximizarse, ya que esto predice que el dato que viene va a ser un dato de una nueva partición. Lo que suele suceder es que se maximiza $p(r_t = 1|x_1, ..., x_t)$ o para algún valor mayor a 1.

4.3.2. Formalización de los pasos del algoritmo

El algoritmo de deteccion de changepoint que estudiamos en este capitulo se compone de una serie de pasos, que se desprenden a partir de caracterizar la distribución condicional (dado que se observaron $x_1, \ldots, x_t$) del valor del runlength r_t a tiempo t. Es decir, buscamos calcular la distribución de $p(r_t|x_{1:t})$. Podemos por definición escribir esta probabilidad condicional del siguiente modo

$$p(r_t|x_{1:t}) = \frac{p(r_t, x_{1:t})}{p(x_{1:t})}.$$

Nos enfocaremos sobre el numerador ya que buscamos detectar para que valor de r_t se maximiza esta probabilidad. Reescribámoslo utilizando probabilidad total

$$\begin{aligned}
p(r_t, x_{1:t}) &= p(r_t, x_t, x_{1:t-1}) \\
&= \sum_{r_{t-1}=0}^{t-1} p(r_t, r_{t-1}, x_t, x_{1:t-1}) \\
&= \sum_{r_{t-1}=0}^{t-1} p(r_t, x_t \mid r_{t-1}, x_{1:t-1})\ p(r_{t-1}, x_{1:t-1}) \\
&= \sum_{r_{t-1}=0}^{t-1} p(x_t \mid r_t, r_{t-1}, x_{1:t-1})\ p(r_t \mid r_{t-1}, x_{1:t-1})\ p(r_{t-1}, x_{1:t-1}) \\
&= \sum_{r_{t-1}=0}^{t-1} p(x_t \mid \cancel{r_t}, r_{t-1}, x_{1:t-1})\ p(r_t \mid r_{t-1}, \cancel{x_{1:t-1}})\ p(r_{t-1}, x_{1:t-1}) \\
&= \sum_{r_{t-1}=0}^{t-1} p(x_t \mid r_{t-1}, x_{t-r_{t-1}:t-1})\ p(r_t \mid r_{t-1})\ p(r_{t-1}, x_{1:t-1}).
\end{aligned}$$

Observaciones:

- Escribimos la sumatoria para todos los posibles valores de r_{t-1} pero en realidad existen solo dos casos:

  1. Si $r_t = 0$ entonces r_{t-1} puede tomar cualquier valor entre $\{0, ..., t-1\}$ obteniendo una sumatoria con más de un valor no nulo.

2. Si $r_t = l$ entonces $r_{t-1} = l - 1$. Por lo tanto $p(r_t | r_{t-1}) = 0$ para todo valor de $r_{t-1} \neq l - 1$.

- En el último paso hay 3 igualdades, que nos dedicaremos a demostrar en esta sección:

  (a) $p(r_t \mid r_{t-1}, \cancel{x_{1:t-1}}) = p(r_t \mid r_{t-1})$

  (b) $p(x_t \mid \cancel{r_t}, r_{t-1}, x_{1:t-1}) = p(x_t \mid r_{t-1}, x_{1:t-1})$

  (c) $p(x_t \mid r_{t-1}, x_{1:t-1}) = p(x_t \mid r_{t-1}, x_{t-r_{t-1}:t-1})$.

Comenzaremos demostrando

$$(a)\ p(r_t \mid r_{t-1}, \cancel{x_{1:t-1}}) = p(r_t \mid r_{t-1})$$

Demostración. El planteo del modelo gráfico es el que permite responder esto. Como r_{t-1} es el único padre de r_t, condicionado a este tenemos que r_t es independiente de todo el resto de las variables que no sean sus descendientes. Como podemos observar en la Figura 4.1, ninguno de los $x_{1:t}$ son descendientes de r_t por lo tanto vale la independencia. □

Para probar la igualdad (b) queremos ver que dados $\{r_{t-1}, x_{1:t-1}\}$ la variable x_t no depende del valor que tome r_t. Además de la demostración queremos brindar una idea intuitiva de por qué esto sucede. Para eso es importante tener en mente como es el DAG que representa nuestro modelo en la Figura 4.2 y recordar que la variable r_t que se evalúa al finalizar el tiempo t, luego que x_t haya sido relevado. El valor que toma r_t nos dice lo siguiente:

1. Si $r_t = r_{t-1} + 1$ seguimos en la misma partición lo cual quiere decir que el parámetro θ_{t+1} es igual a θ_t y que el nuevo dato x_{t+1} va a ser generado con el mismo valor de θ con el cual fue generado x_t.

2. Si $r_t = 0$ lo que sucede es que estamos frente a un changepoint (entre los tiempos t y $t + 1$) y que el próximo dato x_{t+1} es el primer dato de una nueva partición que va a ser generado con un parámetro θ_{t+1} nuevo.

En ambos casos el valor de r_t nos brinda información acerca “futuro” y no sobre el valor de x_t.

Existen varias formas de demostrar esto, un enfoque interesante lo podemos encontrar basándonos en las ideas del siguiente artículo [GVP13], que describe una forma algorítmica para deducir independencias con d-separación. Vamos a utilizar la misma idea de d-separación pero con otra técnica.

$$(b)\ p(x_t \mid \cancel{r_t}, r_{t-1}, x_{1:t-1}) = p(x_t \mid r_{t-1}, x_{1:t-1})$$

Demostración. Probaremos que x_t y r_t están d-separados dados $\{r_{t-1}, x_{1:t-1}\}$ y se desprenderá del teorema (3.1) que vale la independencia. Utilizaremos la definición de d-separación vista en el capitulo anterior, dos vértices están d-separados dado un conjunto si no están d-conectados.

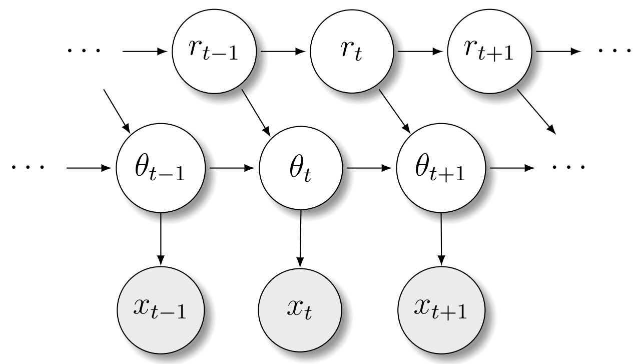


Figura 4.2: Zoom sobre θ_t

Veamos que ningún camino no dirigido d-conecta a x_t y r_t dados $\{r_{t-1}, x_{1:t-1}\}$. Recordemos que un camino no dirigido es un conjunto de vértices distintos y adyacentes.

Si nos enfocamos alrededor de la variable θ_t en el grafo de nuestro modelo obtenemos la Figura 4.2. A partir de este gráfico notamos que se puede caracterizar cualquier camino U entre x_t y r_t, primero tenemos que pasar por θ_t y luego tenemos tres opciones posibles:

- Ir hacia arriba a la izquierda a r_{t-1}.
- Ir hacia la izquierda a θ_{t-1}.
- Ir hacia la derecha a θ_{t+1}.

Veamos entonces que ninguna de estas opciones puede generar un camino que d-conecte a x_t y r_t. Por definición para que x_t y r_t estén d-conectados dado $\{r_{t-1}, x_{1:t-1}\}$, todo no-colisionador en U no debe pertenecer a $\{r_{t-1}, x_{1:t-1}\}$ y todo colisionador en U debe ser un antecesor de $\{r_{t-1}, x_{1:t-1}\}$ o pertenecer a $\{r_{t-1}, x_{1:t-1}\}$.

En los primeros dos casos sucede que U contiene a r_{t-1} como no colisionador y por ende no d-conecta

- $U : x_t \leftarrow \theta_t \leftarrow r_{t-1} \rightarrow r_t$.
- $U : x_t \leftarrow \theta_t \leftarrow \theta_{t-1} \leftarrow ... \leftarrow \theta_k \leftarrow r_{k-1} \rightarrow ... \rightarrow r_{t-1} \rightarrow r_t$, con $0 < k < t-1$.

En el tercer caso tenemos que U contiene un colisionador θ_j con $j > t$ por lo tanto tampoco d-conecta

- $U : x_t \leftarrow \theta_t \rightarrow \theta_{t+1} \leftarrow r_t$
- $U : x_t \leftarrow \theta_t \rightarrow \theta_{t+1} \rightarrow ... \rightarrow \theta_j \leftarrow r_{j-1} \leftarrow ... \leftarrow r_t$.

Finalmente concluimos que x_t y r_t son condicionalmente independientes dado $\{r_{t-1}, x_{1:t-1}\}$ y por eso podemos quitar a r_t de la condición.

□

Para finalizar queremos ver que se cumple (c), es decir que solo aportan información a x_t los valores x_i que fueron generados con su mismo parámetro

$$p(x_t \mid r_{t-1}, x_{1:t-1}) = p(x_t \mid r_t, x_{t-r_{t-1}:t-1}).$$

Demostración. Para simplificar notación escribiremos r_{t-1} en vez de $r_{t-1} = k$ cuando sea necesario. Comenzaremos probando la igualdad cuando $k = 0$ para fijar ideas, posteriormente lo haremos para otros valores de k. Utilizando probabilidad total tenemos que podemos escribir $p(x_t|r_{t-1}, x_{1:t-1})$ de la siguiente forma

$$\begin{aligned} p(x_t|r_{t-1}, x_{1:t-1}) &= \int_\Theta p(x_t, \theta_t|r_{t-1}, x_{1:t-1}) d\theta_t \\ &= \int_\Theta p(x_t|\theta_t, r_{t-1}, x_{1:t-1}) p(\theta_t|r_{t-1}, x_{1:t-1}) d\theta_t. \end{aligned} \tag{4.2}$$

Si logramos quitar $x_{1:t-k-1}$ en ambas probabilidades dentro de la integral (4.2) tendríamos exactamente escrito $p(x_t|r_{t-1}, x_{t-k:t-1})$ y por lo tanto quedaría probada la igualdad. Recordemos que cuando $k = 0$, $x_{t:t-1}$ es el conjunto vacío.

Miremos $p(x_t|\theta_t, r_{t-1}, x_{1:t-1})$, por Markov tenemos que x_t es condicionalmente independiente a $\{r_{t-1}, x_{1:t-1}\}$ dado θ_t, por lo tanto

$$p(x_t|\theta_t, r_{t-1}, x_{1:t-1}) = p(x_t|\theta_t, r_{t-1}). \tag{4.3}$$

Cabe destacar que podríamos quitar r_{t-1} del condicional pero para obtener la igualdad que queremos, lo dejamos. Ahora nos falta ver que podemos quitar $x_{1:t-1}$ del condicional en $p(\theta_t|r_{t-1}, x_{1:t-1})$.

Para eso usaremos la hipótesis del modelo que mencionamos en (4.1), que nos dice que dado que ocurrió un changepoint ($r_{t-1} = 0$) el parámetro θ_t es independiente de θ_{t-1} y de cualquier conjunto que no contenga a descendientes de θ_t (ni padres pero sabemos que tiene solamente dos). Como estamos en el caso que $r_{t-1} = 0$, y $x_{1:t-1}$ no son descendientes de θ_t por hipótesis podemos quitarlos

$$p(\theta_t|r_{t-1}, x_{1:t-1}) = p(\theta_t|r_{t-1}).$$

Recapitulando, demostramos lo siguiente

$$\begin{aligned}
p(x_t|r_{t-1}, x_{1:t-1}) &= \int_\Theta p(x_t, \theta_t|r_{t-1}, x_{1:t-1}) d\theta_t \\
&= \int_\Theta p(x_t|\theta_t, r_{t-1}, x_{1:t-1}) p(\theta_t|r_{t-1}, x_{1:t-1}) d\theta_t. \\
&= \int_\Theta p(x_t|\theta_t, r_{t-1}) p(\theta_t|r_{t-1},) d\theta_t. \\
&= p(x_t|r_{t-1}) \\
&= p(x_t|r_{t-1}, x_{t:t-1})
\end{aligned}$$

Agregamos la ultima igualdad aunque sea redundante para escribir exactamente lo que estaba en el enunciado, recordando que $x_{t:t-1}$ es el conjunto vacío.

Resta ver que se cumple la hipótesis para todo valor $0 < k < t$. Supongamos $r_{t-1} = k$, empecemos de la misma forma que antes escribiendo la igualdad (4.2), aplicando la propiedad de Markov realizamos la siguiente modificación a la igualdad (4.3)

$$p(x_t|\theta_t, r_{t-1}, x_{1:t-1}) = p(x_t|\theta_t, r_{t-1}, x_{t-k:t-1}),$$

dado que tenemos ambos padres de x_t podemos desechar cualquiera de los $x_{1:t-1}$ anteriores, pero solo necesitamos tirar los primeros valores hasta $t - k$.

Si logramos hacer lo mismo para $p(\theta_t|r_{t-1}, x_{1:t-1})$ terminamos la demostración. No podemos utilizar el mismo argumento que antes ya que $r_{t-1} = k$, desarrollemos entonces esta probabilidad

$$p(\theta_t|r_{t-1}, x_{1:t-1}) = \int_\Theta p(\theta_t|\theta_{t-1}, r_{t-1}, x_{1:t-1}) p(\theta_{t-1}|r_{t-1}, x_{1:t-1}) d\theta_{t-1}.$$

Podemos aplicar Markov para $p(\theta_t|\theta_{t-1}, r_{t-1}, x_{1:t-1})$, pues el conjunto de padres de θ_t se encuentra en la condición

$$p(\theta_t|\theta_{t-1}, r_{t-1}, x_{1:t-1}) = p(\theta_t|\theta_{t-1}, r_{t-1}, x_{t-k:t-1}).$$

Solo nos queda $p(\theta_{t-1}|r_{t-1}, x_{1:t-1})$, notemos que estamos en la misma situacion que teniamos recien con θ_t. No podemos utilizar Markov, pero podemos volver a desarrollar esta probabilidad

$$p(\theta_{t-1}|r_{t-1}, x_{1:t-1}) = \int_\Theta p(\theta_{t-1}|\theta_{t-2}, r_{t-1}, x_{1:t-1})p(\theta_{t-2}|r_{t-1}, x_{1:t-1})d\theta_{t-2}.$$

Hay que tener presente que si en algun momento logramos cancelar dentro de la integral en ambas probabilidades los $x_{1:t-k-1}$, terminamos. Para $p(\theta_{t-1}|\theta_{t-2}, r_{t-1}, x_{1:t-1})$ logramos hacer aparecer uno de los dos padres de θ_{t-1} veamos que también podemos hacer aparecer el otro. Para eso vamos a utilizar la siguiente igualdad

$$p(r_{t-1} = k) = \sum_{r_{t-2}} p(r_{t-1} = k|r_{t-2})p(r_{t-2}) = p(r_{t-1} = k, r_{t-2} = k-1).$$

Vale pues $p(r_{t-1} = k|r_{t-2} = j) = 0$ para todo $j \neq k-1$ (y además estamos considerando k no nulo). Por ende podemos escribir lo siguiente

$$p(\theta_{t-1}|\theta_{t-2}, r_{t-1} = k, x_{1:t-1}) = p(\theta_{t-1}|\theta_{t-2}, r_{t-1} = k, r_{t-2} = k-1, x_{1:t-1})$$

gracias a que $\{\theta_{t-2}, r_{t-1} = k, x_{1:t-1}\}$ y $\{\theta_{t-2}, r_{t-1} = k, r_{t-2} = k-1, x_{1:t-1}\}$ son el mismo evento. Si el runlength a tiempo $t-1$ tiene longitud k entonces el valor del runlength a tiempo $t-2$ era $k-1$ (no vale la recíproca).

Para simplificar la notación de ahora en más abreviaremos el conjunto $\{r_{t-1} = k, r_{t-2} = k-1, ..., r_{t-k-1} = 0\}$ como $\{r_{t-1}, r_{t-2}, ..., r_{t-k-1}\}$ de ser necesario.

Ahora tenemos a ambos padres de θ_{t-1} en el condicional y podemos eliminar los x_i que deseemos

$$p(\theta_{t-1}|\theta_{t-2}, r_{t-1}, x_{1:t-1}) = p(\theta_{t-1}|\theta_{t-2}, r_{t-1}, x_{t-k:t-1}).$$

Podemos seguir utilizando probabilidad total de forma que $p(\theta_t|r_{t-1}, x_{1:t})$ nos queda escrito como

$$\int_{\Theta^k} \left(\prod_{i=t-k+1}^{t-1} p(\theta_i|\theta_{i-1}, r_{t-1}, x_{1:t-1}) \right) p(\theta_{t-k}|r_{t-1}, x_{1:t-1}) d\theta_{t-k:t-1}.$$

Para todos los elementos dentro de la productoria se cumple

$$p(\theta_i|\theta_{i-1}, r_{t-1}, x_{1:t-1}) = p(\theta_i|\theta_{i-1}, r_{t-1}, x_{t-k:t-1}).$$

ya que $\{r_{t-1} = k\} = \{r_{t-1} = k, ..., r_{i-1} = k - (t - i)\}$ son el mismo conjunto siempre tenemos a los padres de θ_i en la condición, lo que nos permite quitar $x_{1:t-k-1}$. Para concluir solo necesitamos ver que sucede con $p(\theta_{t-k}|r_{t-1}, x_{1:t-1})$ como mencionamos anteriormente vale

$$p(\theta_{t-k}|r_{t-1} = k, x_{1:t-1}) = p(\theta_{t-k}|r_{t-1} = k, ..., r_{t-k-1} = 0, x_{1:t-1})$$

ahora si estamos en una situación análoga a la primera demostración que hicimos cuando $r_{t-1} = 0$. Como $r_{t-k-1} = 0$ esta en la condición nos volvemos a valer de la hipótesis (4.1) para decir que θ_{t-k} es independiente a cualquier conjunto que no contenga a sus descendientes ni a sus padres. Por hipótesis entonces θ_{t-k} es independiente de $\{r_{t-1} = k, ..., r_{t-k} = 1, x_{1:t-1}\}$, lo cual nos permite retirar $x_{1:t-k-1}$ de la condición terminando la demostración.

Realicemos un raconto final de que fue lo que probamos:

- Escribimos $p(x_t|r_{t-1}, x_{1:t-1})$ como su integral (4.2) utilizando probabilidad total.

- Quitamos $x_{1:t-k-1}$ de $p(x_t|\theta_t, r_{t-1}, x_{1:t-1})$ por Markov.

- Reescribimos a $p(\theta_t|r_{t-1}, x_{1:t-1})$ utilizando probabilidad total k veces.

- Quitamos $x_{1:t-k-1}$ de cada uno de los elementos $p(\theta_i|\theta_{i-1} r_{t-1}, x_{1:t-1})$ dentro de la productoria utilizando Markov pues $\{r_{t-1} = k\} = \{r_{t-1} = k, r_{t-2} = k - 1, ..., r_{t-k-1} = 0\}$.

- Quitamos $x_{1:t-k-1}$ de $p(\theta_{t-k}|r_{t-1}, x_{1:t-1})$ por hipótesis de modelo (4.1).

$$\begin{aligned}
p(x_t|r_{t-1}, x_{1:t-1}) &= \int_{\Theta} p(x_t, \theta_t|r_{t-1}, x_{1:t-1}) d\theta_t \\
&= \int_{\Theta} p(x_t|\theta_t, r_{t-1}, x_{1:t-1}) p(\theta_t|r_{t-1}, x_{1:t-1}) d\theta_t. \\
&= \int_{\Theta} p(x_t|\theta_t, r_{t-1}, x_{t-k:t-1}) p(\theta_t|r_{t-1}, x_{t-k:t-1}) d\theta_t. \\
&= p(x_t|r_{t-1}, x_{t-k:t-1})
\end{aligned}$$

Demostrando así todos los pasos que dan forma al algoritmo bayesiano de detección de checkpoints.

$\square$

4.3.3. Pasos

Input: constante de Hazard H, parámetros de la distribución a priori α.

Output: a medida que se ejecuta, paso a paso, va actualizando una lista de changepoints detectados.

1. **Inizialización**

$$P(r_0 = 0) \leftarrow 1, H \leftarrow cte, checkpoints = \{0\}, t \leftarrow 0$$

2. **Nueva Observación** x_t

3. **Calculo de Probabilidades Predictivas** $\pi_t \leftarrow p(x_t|r_{t-1}, x_{r_{t-1}:t-1})$

4. **Calculo de Probabilidades de Crecimiento** $p(r_t = r_{t-1}+1, x_{1:t}) \leftarrow p(r_{t-1}, x_{1:t-1})\ \pi_t\ (1 - H)$

5. **Calculo de Probabilidades de Changepoint** $p(r_t = 0, x_{1:t}) \leftarrow \sum_{r_{t-1}} p(r_{t-1}, x_{1:t-1})\ \pi_t\ H$

6. **Calculo de Probabilidad Marginal de los Datos x** $p(x_{1:t}) \leftarrow \sum_{r_t} p(r_t, x_{1:t})$

7. **Determinación de la distribución del runlength** $p(r_t|x_{1:t}) \leftarrow p(r_t, x_{1:t})/p(x_{1:t})$

8. **Determinar tiempo c_t candidato a último changepoint** $c_t \leftarrow argmax_t\ p(r_t|x_{1:t})$

9. **Chequear si c_t esta en checkpointArray, de lo contrario agregarlo**

   if $c_t \notin checkpoints$ **then**

   $checkpoints \leftarrow checkpoints \cup \{c_t\}$

   end if

   $t \leftarrow t + 1$

10. **Volver a paso 2.**

4.4. Experimentos sintéticos de la propuesta bayesiana

En esta sección nos enfocaremos en algunos experimentos realizados para testear la performance del algoritmo. Haremos algunas observaciones al final del capítulo y ahondaremos un poco más en las conclusiones de la tesis.

Tanto las series, como los tiempos en los cuales se producen los changepoints son generados de forma aleatoria. La distribución de changepoints esta dada por la función de Hazard (de parámetro constante) explicada en el capitulo anterior, y la distribución de cada serie de tiempo se especifica en cada caso.

Marcaremos los changepoints "reales" con una linea vertical roja y con una linea punteada púrpura cuando el algoritmo detecte un changepoint.

Métricas

En términos de evaluar algoritmos de detección de changpoints, existen un numero distintos de enfoques. En papers como Buntain, Natoli, and Zivkovic [CZ14] se basan primordialmente en métricas de clasificación binaria, otro ejemplo es el paper de Kawahara and Sugiyama [KS09] en el cual utilizan variaciones de este tema enfocándose en curvas (ROC). En este caso nosotros seguiremos los lineamientos utilizados en [Cha17].

Indice de Rand

En este caso utilizaremos el índice de Rand para medir los experimentos. Esta medida esta diseñada para calcular la similitud entre dos clusters de puntos. Es usada para calcular la precisión dado un modelo de clustering, cuando se lo compara al cluster original. Esta métrica fue propuesta primero por Rand en "Objective Criteria for the Evaluation of Clustering Methods" [Ran71]. También fue usada por Mattenson y James en "A nonparametric approach for multiple change point analysis of multivariate data" [MJ12] con el propósito de evaluar distintos enfoques de detección de cambios. El índice de Rand esta definido como:

$$R = \frac{a+b}{a+b+c+d}$$

Dado un set de datos S, particionado (clusterizado) mediante dos métodos diferentes, a los cuales llamaremos X e Y, se puede definir lo siguiente:

- a = numero total de pares que fueron particionados en el mismo subconjunto por X e Y.
- b = numero total de pares que fueron particionados en distintos subconjuntos por X e Y.
- c = numero total de pares que fueron particionados en el mismo subconjunto en X pero en un distinto subconjunto en Y.
- d = numero total de pares que fueron particionados en el mismo subconjunto en Y pero en diferentes subconjuntos en X.

Intuitivamente se puede decir que a + b es el numero total de acuerdos entre los métodos X e Y, mientras c + d es el numero total de desacuerdos. Esta cuenta retorna 0 para clusters completamente diferentes y 1 para clusters idénticos.

4.4.1. Cambios de media

Descripción

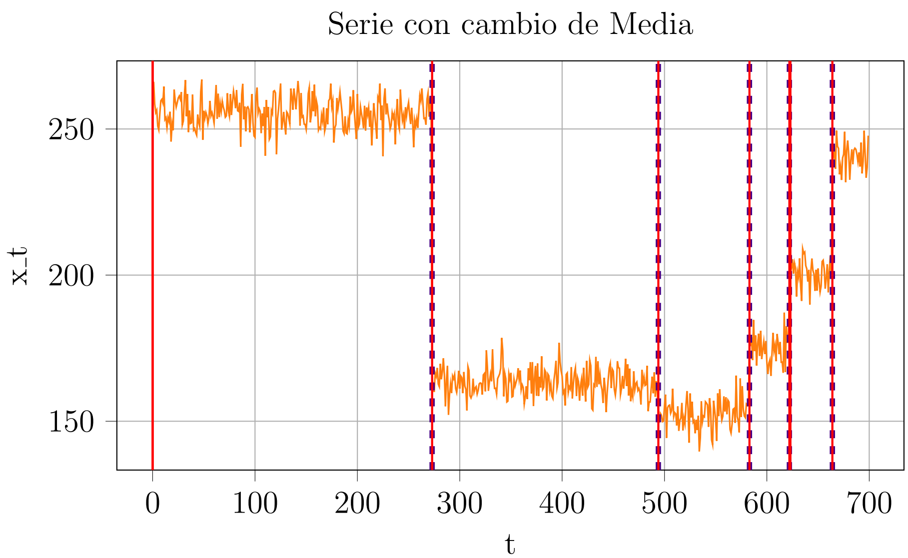
En esta serie de tiempo se producirán changepoints resultado de cambios en uno de los parámetros generativos de la serie: la media. A partir de la distribución a priori de $\mu \sim \mathcal{N}(\mu_0, \sigma_0)$ se genera una nueva media cuando se produce un changepoint.

$$X : \{x_t \sim \mathcal{N}(\mu_t, \sigma) | 1 \leq t \leq T\} \tag{4.4}$$

Parámetros

- T = 700, $H = 0.005$, $\sigma = 5$.
- changepoints t = {0, 273, 494, 583, 622, 623, 664}.
- $\mu \sim \mathcal{N}(\mu_0 = 200, \sigma_0 = 60)$.
- μ = {256.17, 163.50, 153.07, 173.49, 198.54, 200.87, 240.09}.

- Índice de Rand promedio = 0.978



4.4.2. Serie de tiempo con cambios de varianza

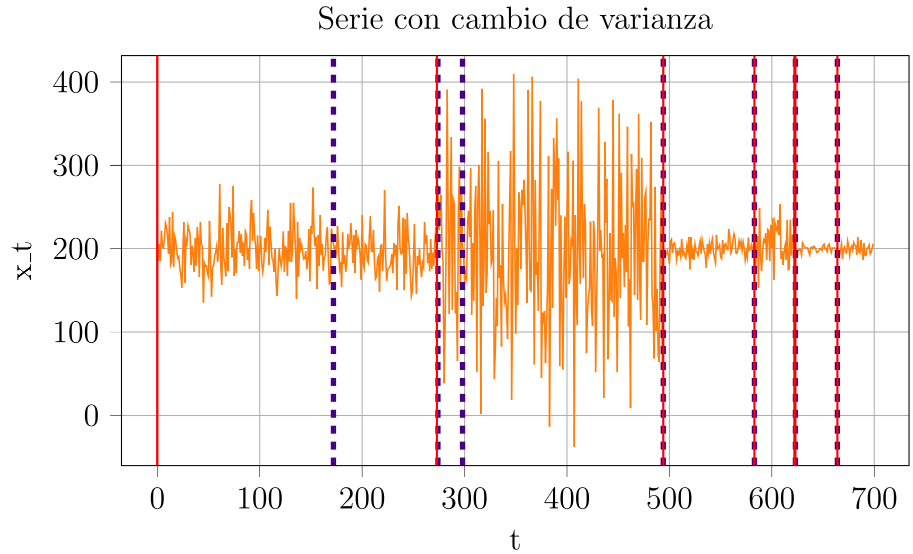
Descripción

Serie de tiempo con cambios en la varianza (μ es constante y conocido). A partir de la distribución a priori de $\sigma \sim \Gamma^{-1}(\alpha, \beta)$ se genera una nueva varianza cuando se produce un changepoint.

$$X : \{x_t \sim \mathcal{N}(\mu, \sigma_t) | 1 \leq t \leq T\} \tag{4.5}$$

Parámetros

- T = 700 $H = 0.005$, $\mu = 200$.
- changepoints t = {0, 273, 494, 583, 622, 623, 664}.
- $\sigma \sim \Gamma^{-1}(\alpha = 2, \beta = 22)$.
- σ = {25.85, 94.22, 9.02, 27.33, 37.69, 3.06, 9.00}.
- Indice de Rand promedio = 0.837



4.4.3. Serie de tiempo con cambios de media y varianza

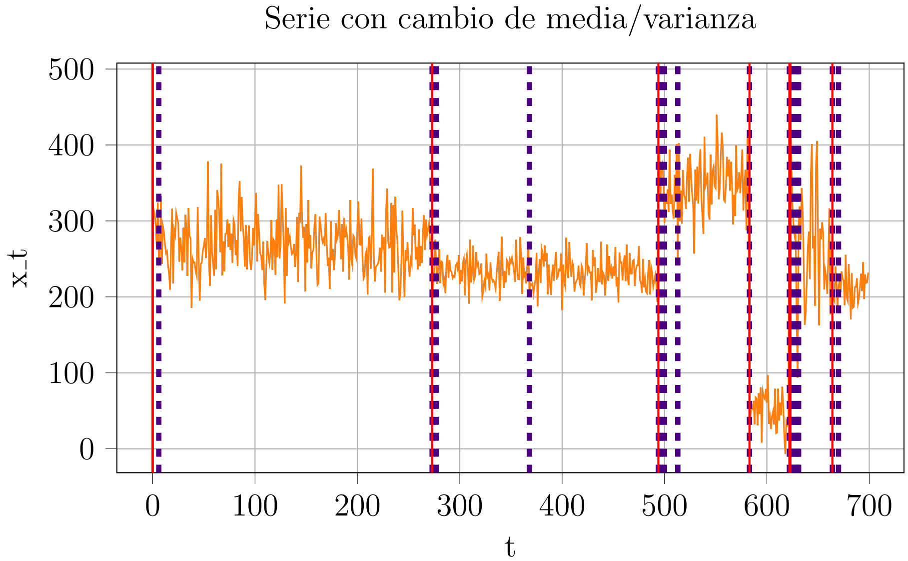
Descripción

En esta serie cambiarán ambos parámetros, la media y la varianza. A partir de la distribución a priori de $(\mu, \sigma) \sim N - \Gamma^{-1}(\mu_0, \nu, \alpha, \beta)$ se generarán una nueva media y varianza cuando se produzca un changepoint.

$$X : \{x_t \sim \mathcal{N}(\mu_t, \sigma_t) | 1 \leq t \leq T\} \tag{4.6}$$

Parámetros

- T = 700, $H = 0.005$.
- changepoints t = {0, 273, 494, 583, 622, 623, 664}.
- $(\mu, \sigma) \sim N - \Gamma^{-1}(\mu_0 = 200, \nu = 0.1, \alpha = 10, \beta = 400)$.
- μ ={273.08, 233.95, 347.60, 45.06, 593.53, 273.49, 211.26}, σ = {35.14, 20.39, 35.09, 26.80, 46.98, 73.28, 28.02}.
- Indice de Rand promedio = 0.973



4.4.4. Serie de tiempo sin changepoints

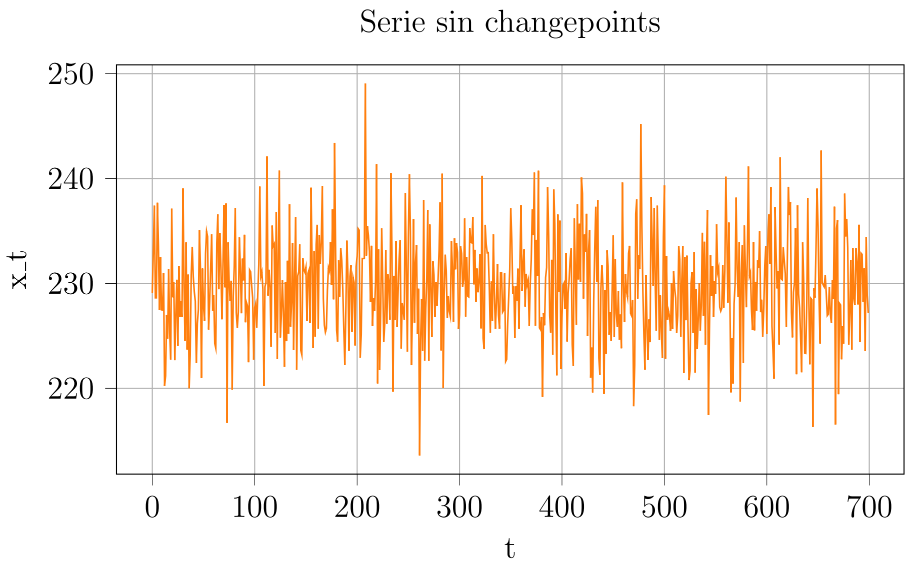
Descripción

Serie de tiempo que sin changepoints. Buscamos observar si se detectan falsos positivos y con que frecuencia. La serie tendrá la siguiente forma

$$X : \{x_t \sim \mathcal{N}(\mu, \sigma) | 1 \leq t \leq T\} \tag{4.7}$$

Parámetros

- T = 700.
- changepoints t = {}.
- $\mu \sim \mathcal{N}(\mu_0 = 200, \sigma_0 = 60)$.
- $x_t \sim \mathcal{N}(\mu = 229, 80, \sigma = 5)$.



4.4.5. Serie de tiempo con outliers

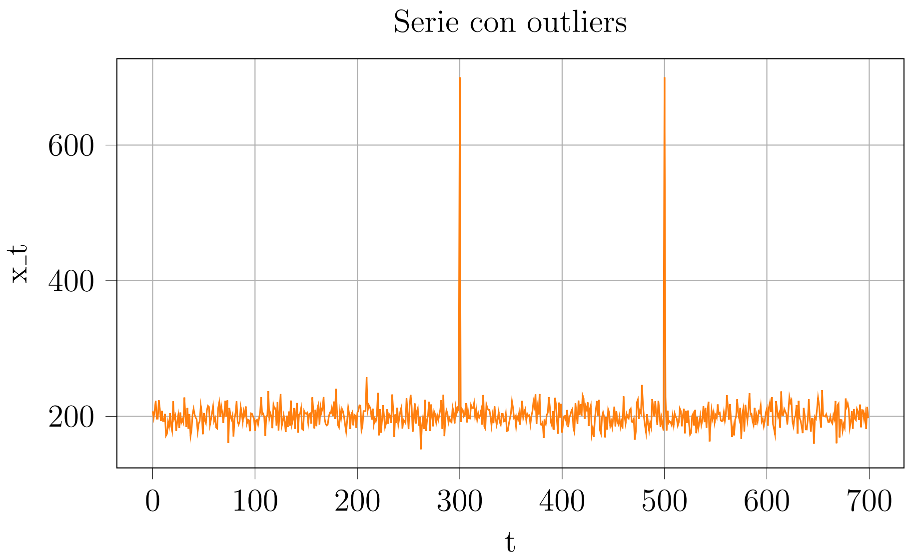
Descripción

Serie de tiempo que sin changepoints con outliers. Buscamos observar si el modelo detecta changepoints cuando se producen outliers. La serie tendrá la siguiente forma salvo en dos puntos agregados a mano.

$$X : \{x_t \sim \mathcal{N}(\mu, \sigma) | 1 \leq t \leq T\} \tag{4.8}$$

Parámetros

- T = 700,
- changepoints t = {}.
- outliers t = {300,500}.
- $x_t \sim \mathcal{N}(\mu = 200, \sigma = 15)$.


Serie con outliers
x_t
600
400
200
0
100
200
300
400
500
600
700
t

4.4.6. Observaciones

En general notamos es que el algoritmo tiene muy buena performance en los casos de prueba. Si miramos la métrica de Rand, podemos notar que el algoritmo posee mucha mejor precisión en los casos que se encuentran involucrados cambios de medias en la distribución.

Algo que se desprende de trabajar con muchos ejemplos, que dependiendo la situación en la que se este utilizando el algoritmo, los falsos positivos podrían llegar a representar un problema, ese es un tema interesante para agregar a trabajos futuros.

Recordemos que el algoritmo puede detectar los changepoints con un retraso temporal, que en este trabajo no estamos tomando en cuenta pero se puede medir. Sin embargo un consideramos una propiedad a destacar, que el algoritmo cuando detecta un cambio también detecta a que tiempo fue y lo señala.

Agregamos dos ejemplos que no fueron medidos con el indice de Rand por un tema de costos. Uno es evaluar una serie sin cambios, en estos casos el algoritmo casi no presenta falsos positivos. En el otro ejemplo introdujimos outliers manualmente, para ver cuan robusto es el modelo a estos casos excepcionales. El algoritmo se porta bien con outliers aislados con cambios de hasta 3 veces la media. Recién al probar con una media de 1000 es cuando detectó un changepoint.

Existen una gran cantidad de experimentos que se pueden realizar. No es la intención de esta tesis testear todos los distintos casos que puedan surgir, pero si nos parece interesante mencionar algunos de ellos como por ejemplo:

- Setear hiper parámetros para las funciones a priori pero generar los parámetros con otros hiper parámetros distintos.
- Generar datos de las series de tiempo con distintas distribuciones.
- Generar datos de la serie de tiempo con tendencias.

Hablaremos mas de esto en la sección de conclusiones cuando mencionemos posibles formas de continuar este trabajo.

Capítulo 5

Propuesta Frecuentista

5.1. Introducción

En el capitulo anterior vimos un tipo particular de changepoint que surgía a partir de un cambio en los parámetros generativos de una secuencia de datos. Esto implica suponer que los datos pertenecen a un tipo de distribución paramétrica en particular, siendo un cambio de dichos parámetros el indicativo del changepoint.

Un punto en donde podría presentar problemas el algoritmo anterior es cuando exista una "violación" del principio de idéntica distribución, que es propenso a ocurrir ante la presencia de tendencias marcadas en la serie de tiempo. Para tal fin, la propuesta de esta sección apunta a detectar una naturaleza distinta de changepoints resultado de variaciones significativas en la pendiente de la serie.

El algoritmo propuesto en esta ocasión es una modificación del algoritmo offline publicado en [Zuo+19]. Al convertirlo en online se pierden algunos beneficios, en aras de poder ganar la facilidad de detectar lo más pronto posible el punto de cambio y no una vez que se observó toda la serie. Desde lo práctico tiene varias diferencias con la propuesta bayesiana, entre ellas podríamos destacar su implementación más sencilla y que no requiere conocer una distribución a priori ya que que es una construcción puramente frecuentista. Una posible desventaja es que su performance estará en cierto punto "garantizada" en la medida que se preserven las hipótesis generativas del proceso, por ejemplo la normalidad de la distribución de los errores que caen por fuera de la curva de tendencia.

5.2. Marco teórico

Abordaremos el marco teórico que da lugar al algoritmo de detección de changepoints en el trabajo de [Zuo+19]. El objetivo es entender qué representa el estadístico utilizado para el algoritmo, marcar las hipótesis del mismo y deducir su distribución.

Sean $Y_1 : \{y_i^1 = \beta_1 i + \alpha_1 + \epsilon_i^1 \mid 1 \leq i \leq n\}$ de longitud n e $Y_2 : \{y_j^2 = \beta_2 j + \alpha_2 + \epsilon_j^2 \mid 1 \leq j \leq m\}$ de longitud m, dos series de tiempo reducidas a su variante casi más sencilla como ventanas de regresión simple a las cuales queremos realizarle un test estadístico para medir la diferencia de sus pendientes. Sean β_1 y β_2 las pendientes de las series Y_1 e Y_2, α_1 y α_2 los intercepts (ordenada a la origen) y ϵ_i^1 y ϵ_j^2 los errores, que asumiremos variables aleatorias independientes normalmente distribuidas con media 0 y varianza σ^2. Esta última hipótesis de *homoscedasticidad* es fundamental, la varianza de los errores es la misma para cada observación y para cada serie.

Consideraremos la concatenación de ambas series $Y_1 \cup Y_2$ como una porción de la serie de tiempo. Diremos que se produce un changepoint cuando ocurre un cambio de pendiente ($\beta_1 \neq \beta_2$), el cual buscamos detectar con un test de hipótesis sobre los estimadores de cuadrados mínimos de las mismas. Así pues, nuestra hipótesis nula será $\beta_1 = \beta_2$ siendo la alternativa que ambas pendientes son distintas.

Una heurística razonable para la construcción de un estadístico buscando testear $H_0 : \theta = \theta_0$, consiste en usar la estandarización natural derivada del test de Wald

$$T = \frac{\widehat{\theta} - \theta_0}{\widehat{SD}(\widehat{\theta})}$$

donde $\widehat{\theta}$ es un estimador de θ y $\widehat{SD}(\widehat{\theta})$ es un estimador del desvío estándar del estimador $\widehat{\theta}$. Bajo criterios de regularidad adecuados, se puede ver que esto se aproximará por una distribución normal estándar. En nuestro caso, recurriremos en un ajuste más fino a través de una distribución t de Student.

El test de hipótesis propuesto por [Zuo+19] para la diferencia de pendientes viene dado por el siguiente estadístico:

$$t_{slope} = \frac{C^{1/2}(\hat{\beta}_1 - \hat{\beta}_2)}{S_{\beta_1,\beta_2}},$$

donde

$$C = \frac{NM}{N+M} \quad con \quad N = \sum_{i=1}^{n} \left(i - \frac{(n+1)}{2} \right)^2 \; , \; M = \sum_{j=1}^{m} \left(j - \frac{(m+1)}{2} \right)^2$$

y

$$S^2_{\beta_1,\beta_2} = \frac{1}{n+m-4} \left(\sum_{i=1}^{n} (y_i^1 - \hat{y}_i^1)^2 + \sum_{j=1}^{m} (y_j^2 - \hat{y}_j^2)^2 \right).$$

Recordemos que la distribución t de Student es el el cociente dado por

$$t = \frac{Z}{\sqrt{U/n}}$$

siendo Z y U independientes, Z con distribución normal estándar y U con distribución χ^2_n.

Vamos a dividir en cuatro partes el resto de la sección para darle un cierre a estas consideraciones:

- Regresión (obtención de los estimadores).
- Caracterización distribucional del Numerador.
- Caracterización distribucional del Denominador.
- Independedncia entre ambos.

Con todos estos ingredientes tendremos que bajo H_0 el estadístico tendrá distribución t de Student.

Regresion

Utilizaremos la siguiente notación matricial, clásica en la literatura de modelo lineal

$$Y = X\beta + \varepsilon \tag{5.1}$$

$$\begin{bmatrix} y_1^1 \\ \vdots \\ y_n^1 \\ y_1^2 \\ \vdots \\ y_m^2 \end{bmatrix} = \left[\begin{array}{cc|cc} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & n & 0 & 0 \\ \hline 0 & 0 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & m \end{array}\right] \cdot \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1^1 \\ \vdots \\ \epsilon^1 n \\ \epsilon_1^2 \\ \vdots \\ \epsilon_m^2 \end{bmatrix}.$$

En este caso $Y = (y_1^1, ..., y_n^1, y_1^2, ..., y_m^2)^\intercal$ representa a nuestra serie de tiempo, en cada posición tenemos una variable aleatoria escalar observable. X es una matriz de constantes conocidas de dimensión $(n+m) \times 4$ que tiene rango máximo (es fácil verificar que las cuatro columnas son linealmente independientes), lo cual implica que $X^\intercal X$ es no singular. $\beta = (\alpha_1, \beta_1, \alpha_2, \beta_2)^\intercal$ es un vector de parámetros desconocidos.

Por último ε al cual nos referiremos como *error*, un vector de variables aleatorias independientes idénticamente distribuidas no observables. Como hipótesis tendrán distribución normal multivariada con media igual a 0 y varianza igual a $\sigma^2 I_{n+m}$. Es decir, cada ϵ_i tendrá distribución normal de media 0 e idéntica varianza σ^2.

Según la teoría clásica de regresión lineal, sobre la cual podemos encontrar en más el respecto en [Ame85], el estimador de cuadrados mínimos de β está dado por

$$\hat{\beta} = (X^\intercal X)^{-1} X^\intercal Y. \tag{5.2}$$

Reemplazando Y por su expresión de (5.1) podemos escribir $\hat{\beta}$ en función de ε y β

$$\hat{\beta} = \beta + (X^\intercal X)^{-1} X^\intercal \varepsilon. \tag{5.3}$$

De la igualdad anterior resulta inmediato que si ε es un vector con distribución normal multivariada, entonces también lo es el estimador de cuadrados mínimos $\hat{\beta}$. Calculemos su varianza usando la expresión (5.3)

$$\begin{aligned} V[\hat{\beta}] &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^\intercal] \\ &= E[(X^\intercal X)^{-1} X^\intercal \varepsilon \varepsilon^\intercal X (X^\intercal X)^{-1}] \\ &= \sigma^2 (X^\intercal X)^{-1}. \end{aligned}$$

Todo lo anterior nos dice que $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^\intercal X)^{-1})$. Teniendo la expresión de $\hat{\beta}$ podemos definir los *residuos* de la siguiente forma

$$\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\beta}. \tag{5.4}$$

Al igual que antes podemos reemplazar los valores de Y y $\hat{\beta}$ de (5.1) y (5.3) para escribir a $\hat{\varepsilon}$ en función de ε:

$$\begin{aligned}
\hat{\varepsilon} &= Y - X\hat{\beta} \\
&= X\beta + \varepsilon - X\hat{\beta} \\
&= X\beta + \varepsilon - X(\beta + (X^\intercal X)^{-1}X^\intercal \varepsilon) \\
&= \varepsilon - X(X^\intercal X)^{-1}X^\intercal \varepsilon \\
&= (I - X(X^\intercal X)^{-1}X^\intercal)\varepsilon.
\end{aligned}$$

Llamaremos $P = X(X^\intercal X)^{-1}X^\intercal$ y $M = I - P$. Con esto podemos escribir a Y en dos componentes *ortogonales* usando (5.4)

$$Y = X\hat{\beta} + \hat{\varepsilon} = PY + MY.$$

Como $\hat{\epsilon}$ es ortogonal a X (eso es porque $\hat{\epsilon}^\intercal X = 0$) el estimador de cuadrados mínimos puede ser pensado como la proyección ortogonal de Y sobre el espacio generado por las columnas de X. Estas matrices P y M tienen ciertas propiedades cuyas demostraciones podemos encontrar en [Bel70], mencionaremos las más importantes.

Lema 4. *Sea X una matriz de $k \times r$ de rango completo y definimos $P = X(X^\intercal X)^{-1}X^\intercal$:*

1. *$P = P^\intercal = P^2$ (P es un proyector).*

2. *P es de rango completo.*

3. *Si $x = Xc$ para algun vector c, entonces $Px = x$ (proyecta como la identidad en el espacio columna de X).*

4. *$M = I - X(X^\intercal X)^{-1}X^\intercal$ también es simétrica idempotente de rango $k - r$.*

Hasta acá enumeramos los resultados que usaremos de la teoría clásica de regresión lineal. En la próxima sección veremos cual es la distribución del numerador de t_{slope}.

Numerador

El estadistico de la diferencia entre las pendientes de cuadrados mínimos puede ser pensado de la siguiente forma

$$\hat{\beta}_1 - \hat{\beta}_2 = (0, 1, 0, -1)\hat{\beta}.$$

Como vimos previamente $\hat{\beta}$ tiene distribución normal multivariada por ende la diferencia de pendientes es un funcional linear de una normal multivariada ergo normal. Veamos cuales son su media y su varianza

$$E[\hat{\beta}_1 - \hat{\beta}_2] = \beta_1 - \beta_2.$$

La esperanza coincide con la diferencia de las pendientes, calculemos su varianza

$$V[\hat{\beta}_1 - \hat{\beta}_2] = V[(0, 1, 0, -1)\hat{\beta}]. \tag{5.5}$$

Reemplazando la expresión de $\hat{\beta}$ de (5.2) y usando la siguiente propiedad de la varianza: $V[x + a] = V[x], a = cte$, obtenemos

$$(5.5) = V[(0, 1, 0, -1)(X^\intercal X)^{-1}X^\intercal \varepsilon].$$

Llamemos $a^\intercal = (0, 1, 0, -1)(X^\intercal X)^{-1}X^\intercal$ y usando el siguiente lema podemos encontrar en [Was03]

Lema 5. *Sea x un vector aleatorio de dimension $k \times 1$ y a un vector de constantes de igual dimensión, entonces:*

$$V[a^\intercal x] = a^\intercal Cov[x]a.$$

Donde $Cov[x]$ es la *matriz de varianzas-covarianzas de x* definida como la esperanza de la matriz $Cov[x] = E[(x - E[x])(x - E[x])^\intercal]$. En particular como $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_{n+m})$ tenemos que $Cov[\varepsilon] = \sigma^2 I_{n+m}$ pues estamos suponiendo que los ε_i son todos independientes. Se desprende

$$\begin{aligned} V[\hat{\beta}_1 - \hat{\beta}_2] &= \sigma^2(0, 1, 0, -1)(X^\intercal X)^{-1}X^\intercal I_{n+m} X(X^\intercal X)^{-1}(0, 1, 0, -1)^\intercal \\ &= \sigma^2(0, 1, 0, -1)(X^\intercal X)^{-1}(0, 1, 0, -1)^\intercal. \end{aligned} \tag{5.6}$$

Para terminar de explicitar la varianza resta calcular $(X^\intercal X)^{-1}$. Como X es una matriz con 4 bloques entonces $X^\intercal X$ también lo es, los bloques superior derecho e inferior izquierdo son matrices nulas de 2×2 y los bloques superior izquierdo e inferior izquierdo son dos matrices de 2×2 invertibles, ya que $(X^\intercal X)$ es no singular.

$$X^\intercal X = \left[\begin{array}{c|c} (X^\intercal X)_1 & \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix} \\ \hline \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix} & (X^\intercal X)_2 \end{array}\right].$$

Utilizamos $(X^\intercal X)_1$ como notación para el bloque superior izquierdo

$$(X^\intercal X)_1 = \begin{bmatrix} 1 \cdots 1 \\ 1 \cdots n \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n i \\ \sum_{i=1}^n i & \sum_{i=1}^n i^2 \end{bmatrix},$$

de manera análoga el segundo bloque nos queda

$$(X^\intercal X)_2 = \begin{bmatrix} 1 \cdots 1 \\ 1 \cdots m \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & m \end{bmatrix} = \begin{bmatrix} m & \sum_{j=1}^m j \\ \sum_{j=1}^m j & \sum_{j=1}^m j^2 \end{bmatrix}.$$

Entonces invertir la matriz $X^\intercal X$ es simplemente invertir cada bloque, como cada bloque es de 2×2 sus inversas tienen la siguiente forma

$$(X^\intercal X)_1^{-1} = \frac{1}{det((X^\intercal X)_1)} \begin{bmatrix} \sum_{i=1}^n i^2 & -\sum_{i=1}^n i \\ -\sum_{i=1}^n i & n \end{bmatrix}$$

$$(X^\intercal X)_2^{-1} = \frac{1}{det((X^\intercal X)_2)} \begin{bmatrix} \sum_{j=1}^m j^2 & -\sum_{j=1}^m j \\ -\sum_{j=1}^m j & m \end{bmatrix}$$

siendo sus determinantes

$$det((X^\intercal X)_1) = n \sum_{i=1}^n i^2 - \left(\sum_{i=1}^n i \right)^2$$

$$det((X^\intercal X)_2) = m \sum_{j=1}^m j^2 - \left(\sum_{j=1}^m j \right)^2.$$

La matriz finalmente queda

$$(X^\intercal X)^{-1} = \left[\begin{array}{c|c} (X^\intercal X)_1^{-1} & \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix} \\ \hline \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix} & (X^\intercal X)_2^{-1} \end{array}\right].$$

La igualdad (5.6) luego de multiplicar $(1, 0, -1, 0)^\intercal (X^\intercal X)^{-1}(1, 0, -1, 0)$ queda de la siguiente forma

$$V[\hat{\beta}_1 - \hat{\beta}_2] = \sigma^2 \left(\frac{n}{n\sum_{i=1}^n i^2 - (\sum_{i=1}^n i)^2} + \frac{m}{m\sum_{j=1}^m j^2 - (\sum_{j=1}^m j)^2} \right).$$

Podemos reescribir los denominadores para tener la misma notación que en [Zuo+19]

$$\begin{aligned} n\sum_{i=1}^n i^2 - \left(\sum_{i=1}^n i\right)^2 &= n\sum_{i=1}^n i^2 - \left(\frac{n(n+1)}{2}\right)^2 \\ &= n\left(\sum_{i=1}^n i^2 - n\left(\frac{(n+1)}{2}\right)^2\right) \\ &= n\left(\sum_{i=1}^n i^2 - \left(\frac{(n+1)}{2}\right)^2\right) \\ &= n\sum_{i=1}^n \left(i - \frac{(n+1)}{2}\right)^2. \end{aligned}$$

Así la varianza resulta

$$V[\hat{\beta}_1 - \hat{\beta}_2] = \sigma^2 \left(\frac{n}{nN} + \frac{m}{mM}\right) = \sigma^2 \frac{1}{C}.$$

Ahora bien, el numerador de t_{slope} esta dado por $C^{1/2}(\hat{\beta}_1 - \hat{\beta}_2)$. Recordemos que estamos bajo la hipótesis nula y por lo tanto $\beta_1 - \beta_2 = 0$. Solo resta para decir que el numerador posee distribución $\mathcal{N}(0, 1)$, dividir por σ. Supongamos entonces que el numerador lo escribimos como $\sigma^{-1}C^{1/2}(\hat{\beta}_1 - \hat{\beta}_2)$ y a continuación veremos que podemos cancelar σ cuando dividamos por el denominador.

Denominador

Para ver cual es la distribución de S_{β_1,β_2} veamos de donde sale. Según la teoría de regresión lineal el estimador de σ^2 se define

$$\hat{\sigma}^2 = (n+m-4)^{-1}\hat{\varepsilon}^\intercal\hat{\varepsilon}.$$

Veamos que distribución tiene $\frac{\hat{\varepsilon}^\intercal\hat{\varepsilon}}{\sigma^2}$ para esto conviene escribirlo de la siguiente forma

$$\frac{\hat{\varepsilon}^\intercal\hat{\varepsilon}}{\sigma^2} = \frac{\epsilon^\intercal M^\intercal M\epsilon}{\sigma^2} = \frac{\epsilon^\intercal M\epsilon}{\sigma^2}.$$

Utilizamos que M es simétrica e idempotente. Es importante destacar que dividimos dos veces por σ para estandarizar ε lo que nos permite aplicar el siguiente lema, que podemos encontrar en [Rao47]

Lema 6. *Sea $z \sim \mathcal{N}(0, I_n)$ y A una matriz simétrica idempotente de rango n. Entonces $z^\intercal Az \sim \chi^2_n$.*

El lema anterior implica que $\frac{\epsilon^\intercal M\epsilon}{\sigma^2} \sim \chi^2_{(n+m-4)}$. Como $\hat{\epsilon} = Y - \hat{Y}$ tenemos la siguiente igualdad

$$\frac{\hat{\varepsilon}^\intercal\hat{\varepsilon}}{\sigma^2} = \frac{(Y-\hat{Y})^\intercal(Y-\hat{Y})}{\sigma^2} = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}\left(y_i^1 - \hat{y}_i^1\right) + \sum_{j=1}^{m}\left(y_j^2 - \hat{y}_j^2\right)\right).$$

Tomando raíz cuadrada y dividiendo por sus grados de libertad la expresión anterior es igual a $\frac{S_{\beta_1,\beta_2}}{\sigma}$. Ubicándolo en el denominador podemos ver que el estadístico t_{slope} propuesto tiene la siguente forma

$$t_{slope} = \frac{\sigma^{-1}C^{1/2}(\hat{\beta}_1 - \hat{\beta}_2)}{\sigma^{-1}S_{\beta_1,\beta_2}} = \frac{C^{1/2}(\hat{\beta}_1 - \hat{\beta}_2)}{S_{\beta_1,\beta_2}}.$$

Deducimos entonces que su distribución es t de Student si combinamos estos resultados con el siguiente lema que podemos encontrar en [Rao47]

Lema 7. *Sea $z \sim \mathcal{N}(0,1)$ y $w \sim \chi^2_n$ variables aleatorias independientes, entonces $n^{1/2}zw^{1/2}$ tiene distribución t de Student con n grados de libertad.*

Para concluir esta sección resta que probemos la independencia entre el numerador y el denominador.

Independencia

Nos vamos a valer del siguiente resultado que podemos encontrar también en [Rao47]

Lema 8. *Sea* $z \sim \mathcal{N}(0, I)$ *entonces* $c^\intercal z$ *y* $z^\intercal A z$ *son independientes si* $Ac = 0$.

Veamos que podemos aplicar este lema para eso recordemos podemos escribir en función de ε la diferencia entre las pendientes y $S^2_{\beta_1,\beta_2}$

$$\hat{\beta}_1 - \hat{\beta}_2 = (0, 1, 0, -1)\hat{\beta} = (0, 1, 0, -1)(\beta + (X^\intercal X)^{-1} X^\intercal \varepsilon)$$

$$S^2_{\beta_1,\beta_2} = \hat{\varepsilon}^\intercal \hat{\varepsilon} = \varepsilon^\intercal M \varepsilon.$$

Obviaremos β porque es constante y no afecta a la independencia, el resultado se reduce a probar que $Ac = 0$ donde $A = M$ y $c = X(X^\intercal X)^{-1}(0, 1, 0, -1)^\intercal$.

Lo cual es inmediato por la propiedad 3 del Lema (4) de matrices pues $M = I - P$ y debido a que P proyecta sobre las columnas de X vale que si $c = Xa$ entonces $Pc = c$. En nuestro caso en particular tenemos que

$$(I - P)c = c - Pc = c - c = 0.$$

Queda entonces demostrado que el estadístico propuesto tiene t_{slope} efectivamente posee distribución t de Student con $n + m - 4$ grados de libertad.

5.3. Propuesta de algoritmo de detección online

Presentamos una variante online a [Zuo+19]. Previo a declarar los pasos vamos a definir la ventana temporal τ, cuya longitud depende de los conocimientos previos acerca del problema. Dada la naturaleza del algoritmo este no podrá detectar changepoints que sucedan de forma consecutiva en ventanas menores a 2τ. El otro parámetro que es importante destacar es ν, el algoritmo realiza un test de hipótesis y declara un changepoint al rechazar la hipótesis nula, con un test de nivel ν. Este parámetro suele ser llamado α le cambiamos la notación para que no exista confusión con los intercepts. En los gráficos de cada experimento indicaremos con una linea vertical punteada cuando el algoritmo rechace la hipótesis nula. Recordemos que a diferencia del algoritmo bayesiano, este no indica en que punto se produce el changepoint, solo detecta que hubo un cambio de pendiente en algún momento anterior dentro de la ventana temporal.

5.3.1. Pasos

Input: ventana de tiempo τ, nivel del test ν.

Output: a medida que se ejecuta, paso a paso, va actualizando una lista de changepoints detectados.

1. **Inizialización** $\tau \leftarrow cte, \nu \leftarrow cte, checkpoints = \{\}, t \leftarrow 0$
2. **Nueva Observación** y_t
3. **Calculo de estadístico** $t_{slope} \leftarrow \frac{C^{1/2}(\hat{\beta}_1 - \hat{\beta}_2)}{S_{\beta_1,\beta_2}}$
4. **Test de Hipótesis**
   if Se rechaza H_0 **then**
      $checkpoints \leftarrow checkpoints \cup \{t\}$
      $t \leftarrow t + 2\tau$
   else
   $t \leftarrow t + 1$
   end if
5. **Volver a paso 2.**

5.4. Experimentos sintéticos de la propuesta

De la misma forma que hicimos en el capítulo anterior presentaremos algunos casos que nos parecieron relevantes para testear la performance del algoritmo, mencionando algunas conclusiones en un apartado al final del mismo.

5.4.1. Serie de tiempo sin changepoints

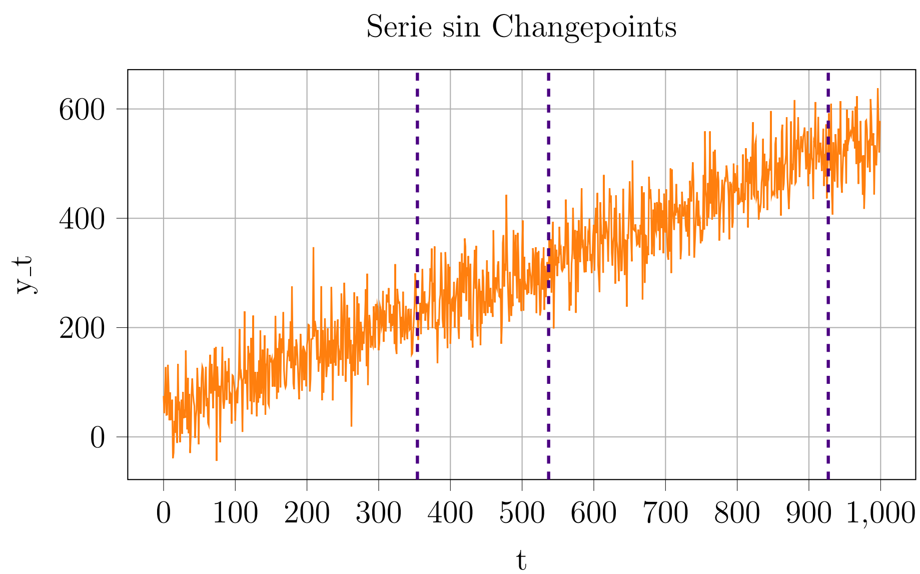
Descripción

Testeamos el algoritmo sobre una serie que no posee changepoints. Buscamos observar si se detectan falsos positivos y con que frecuencia. La serie tendrá la siguiente forma:

$$Y : \{y_i = \beta i + \alpha + \epsilon_i \mid 1 \leq i \leq n\} \tag{5.7}$$

Parámetros

- $n = 1000,\ \tau = 50,\ \nu = 0{,}05,\ \beta = 0{,}5,\ \alpha = 50$
- $\epsilon_i \sim \mathcal{N}(0, 50)$ (random seed = 42)



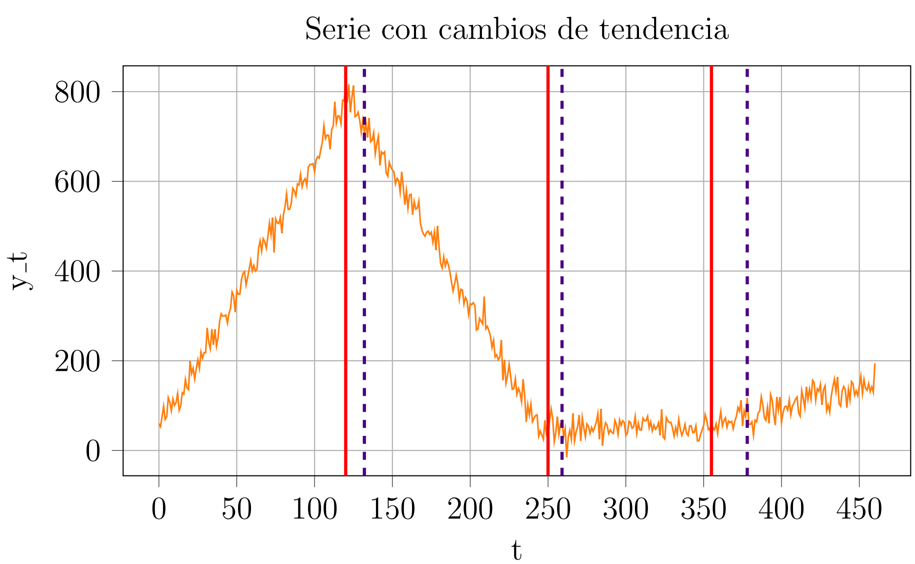
5.4.2. Trend changepoints

Descripción

En el siguiente ejemplo tenemos 4 subseries concatenadas Y_i con $i \in \{1, 2, 3, 4\}$. Cada subserie tiene la misma forma que en (5.7) pero con sus respectivos parámetros. Las ubicaciones de los changepoints fueron asignadas manualmente con una una distancia mayor a 2 τ entre si.

Parámetros

- $n = 461,\ \tau = 50,\ \nu = 0{,}01$
- changepoints: t = $\{120,\ 250,\ 355\}$
- $\beta_1 = 6,\ \beta_2 = -6,\ \beta_3 = 0,\ \beta_4 = 1$
- $\alpha_1 = 50,\ \alpha_2 = 800,\ \alpha_3 = 50,\ \alpha_4 = 50$
- $\epsilon_i \sim \mathcal{N}(0, 50)$ (random seed = 42)



Parámetros

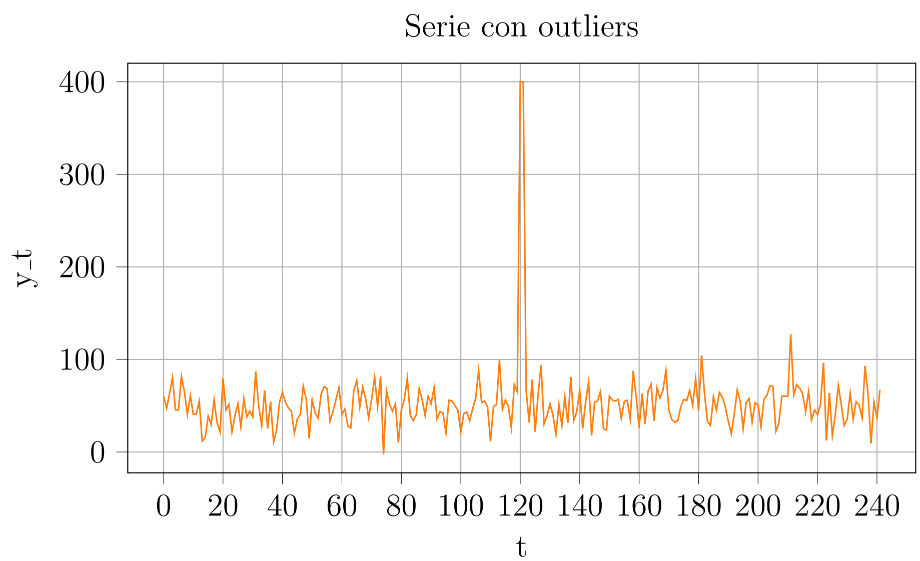
5.4.3. Outliers

Descripción

Serie de tiempo que no posee changepoints. Contiene dos outliers consecutivos.

Parámetros

- $n = 242,\ \tau = 50,\ \nu = 0{,}01,\ \beta = 0,\ \alpha = 50$
- outliers: posición = {121,122} valor = 400
- $\epsilon_i \sim \mathcal{N}(0, 20)$ (random seed = 42)



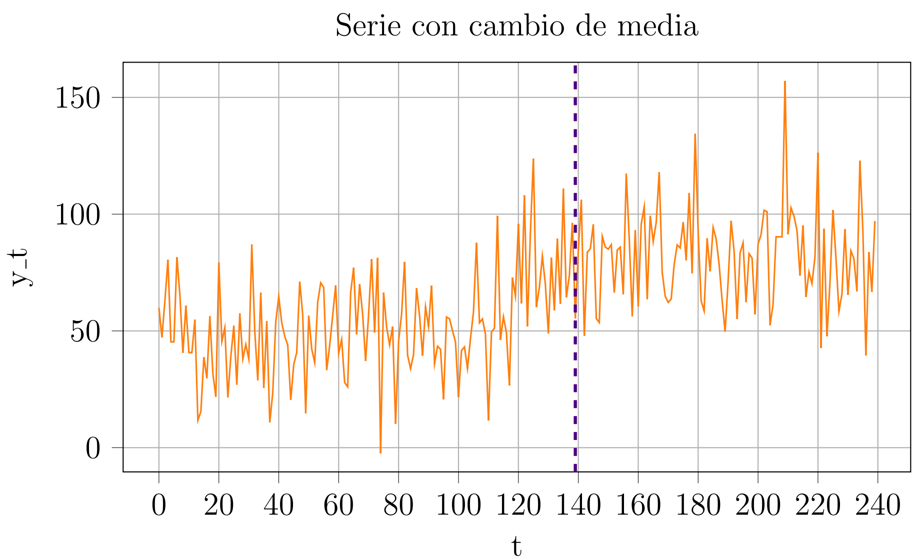
5.4.4. Cambio de Media

Descripción

Dos subseries concatenadas de igual pendiente. Esta serie posee un changepoint en el cual cambia el valor de α. Podemos pensar la serie de tiempo de la siguiente forma $y_t \sim \mathcal{N}(\alpha_j, \sigma)$ donde j = {1,2}, σ se encuentra fijo.

Parámetros

- $n = 240,\ \tau = 50,\ \nu = 0{,}01$
- changepoints: t = $\{120\}$
- $\beta_1,\ \beta_2 = 0$
- $\alpha_1 = 50,\ \alpha_2 = 80$
- $\epsilon_i \sim \mathcal{N}(0, 20)$ (random seed = 42)



5.4.5. Cambio de Varianza

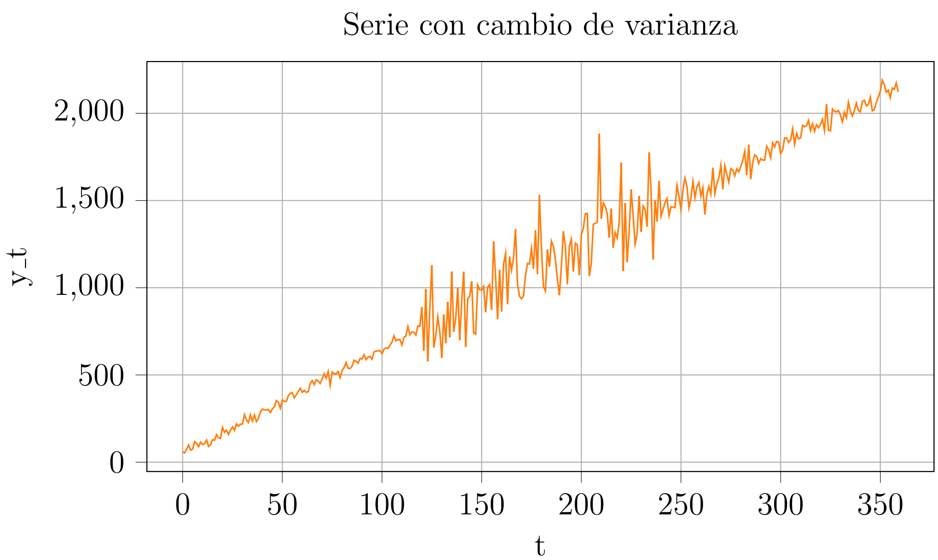
Descripción

Serie de tiempo sin changepoints, pero con un cambio en la varianza de la distribución de ϵ_i.

Parámetros

- $n = 360,\ \tau = 50,\ \nu = 0{,}01$
- cambios σ: t = $\{120,\ 240\}$

- $\beta_1, \beta_2, \beta_3 = 6$
- $\alpha_1, \alpha_2, \alpha_3 = 80$
- $\epsilon_i^1 \sim \mathcal{N}(0, 20),\ \epsilon_i^2 \sim \mathcal{N}(0, 200),\ \epsilon_i^2 \sim \mathcal{N}(0, 50)$ (random seed = 42)



5.4.6. Observaciones

El primer experimento no presenta changepoints sin embargo el algoritmo detecta en tres ocasiones cambios (falsos positivos). La causa de los falsos positivos tiene la siguiente explicación: un changepoint se declara cuando se rechaza la hipótesis nula para el test estadístico de t_{slope} Como el test tiene un cierto nivel de potencia, en este caso 0,05, es dable generar falsos positivos por el ruido propio de la serie. En condiciones ideales el test debería haber detectado $1000 * 0{,}05 = 50$ changepoints. Sin embargo esto no sucede ya que ciertas hipótesis no se cumplen como por ejemplo las ventanas no son independientes entre sí, lo cual genera que los test realizados cada vez que se releva un nuevo dato tengan muchos puntos solapados, esto no obstante es algo que termina beneficiando al test al reducir la cantidad de falsos positivos.

En el siguiente caso se generaron 3 changepoints, es importante destacar que estos se encuentren a una distancia mayor o igual a 2τ por el siguiente motivo: el algoritmo selecciona dos ventanas de tiempo de tamaño τ para ser comparadas, al rechazar H_0 levanta una alarma, luego salta al tiempo $t + 2\tau$. Lo cual quiere decir que apagamos el algoritmo, ya que si lo dejamos correr seguiría rechazando la hipótesis nula al relevar los datos siguientes. Esto a su vez trae una desventaja a la hora de volver a prender el test. Si el changepoint ocurrió hace mucho tiempo (durante la primer ventana) se podría detectar un cambio pero con mucho retraso.

En este ejemplo donde ocurren 3 changepoints podemos notar un par de cosas más: El tiempo que tarda el algoritmo en detectar cambios crece cuando la diferencia entre las pendientes (a misma varianza) es menor. El algoritmo detecta cambios de pendientes mas rápido a medida que la diferencia entre las pendientes crece, lo cual es esperable.

Cuando el algoritmo es puesto a prueba con outliers, demostró ser relativamente robusto. En otros test que no fueron presentados en esta tesis por cuestiones de alcance pudimos constatar que deben suceder varios outliers consecutivos para que salte una alarma, se probo hasta con 3 outliers y no se detectaron cambios. Encontramos que la explicación de este fenómeno se debe a que la diferencia entre las pendientes de cuadrados mínimos frente a un outlier puede aumentar, pero a su vez también crece el estimador de la varianza S_{β_1,β_2} y se compensan. Consideramos esta robustez una ventaja de este algoritmo, ya que buscamos detectar cambios de tendencias y un outlier no entra dentro de esa categoría.

El siguiente experimento que realizamos fue en base a la siguiente pregunta:

¿Cómo se comportaría el algoritmo si en vez de una diferencia entre pendientes existiese una diferencia entre los intercepts?. Este caso también contempla uno de los casos vistos en el capitulo del algoritmo bayesiano, ya que cambiar los intercepts manteniendo las pendientes nulas es lo mismo que tener una serie de tiempo generada a partir de una distribución normal cuyos parámetros generativos se modifican.

Para este experimento el algoritmo bayesiano corre con la ventaja de que da una estimación de cuando ocurrió el changepoint. Sin embargo la propuesta frecuentista probó detectar cambios de medias, que no implican necesariamente cambios en tendencias.

Por último se realizó un experimento cambiando la varianza de los ε_i (violación del principio de homoscedasticidad). Este experimento así como el de la media y los outliers buscan registrar como se comporta el algoritmo fuera de sus hipótesis. Para este escenario no se detectaron cambios incluso con modificaciones de hasta cambios del orden de 10 veces la varianza anterior.

Unos comentarios finales de estos experimentos se van a encontrar en las conclusiones de esta tesis.

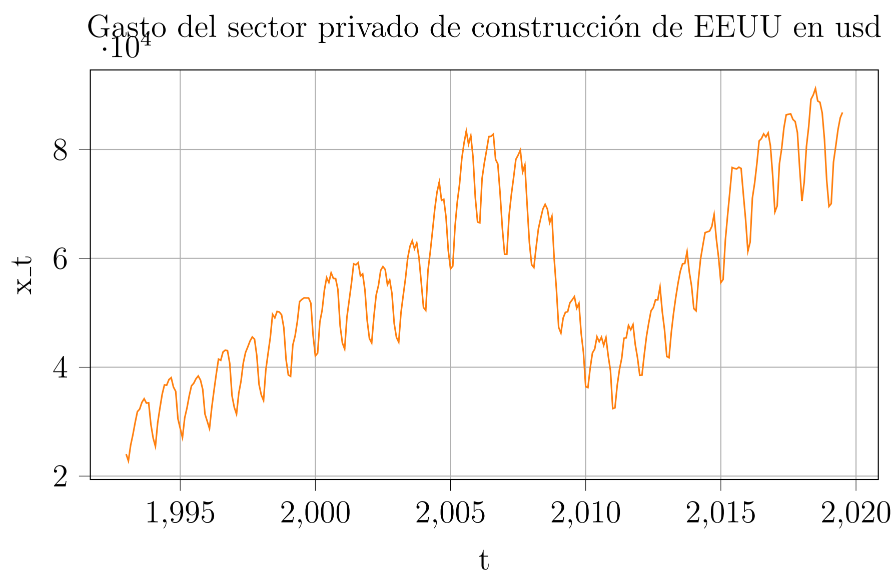
Capítulo 6

Casos Reales

En este último capitulo presentaremos dos casos con datos reales a los cuales les aplicamos las propuestas ofrecidas en esta tesis. Son ejemplos bastante canónicos en la literatura de benchmarks de changepoints.

6.0.1. Gasto en construcción de los EEUU

El siguiente set representa el gasto total en construcción privada en los Estados Unidos, los datos fueron obtenidos a partir del U.S. Census Bureau y se pueden leer en [BW20]. Este censo provee estimaciones mensuales del valor total en 50 estados y el distrito de Columbia, acerca de trabajos hechos cada mes en nuevas estructuras o mejoras en estructuras existentes para el sector publico y privado.



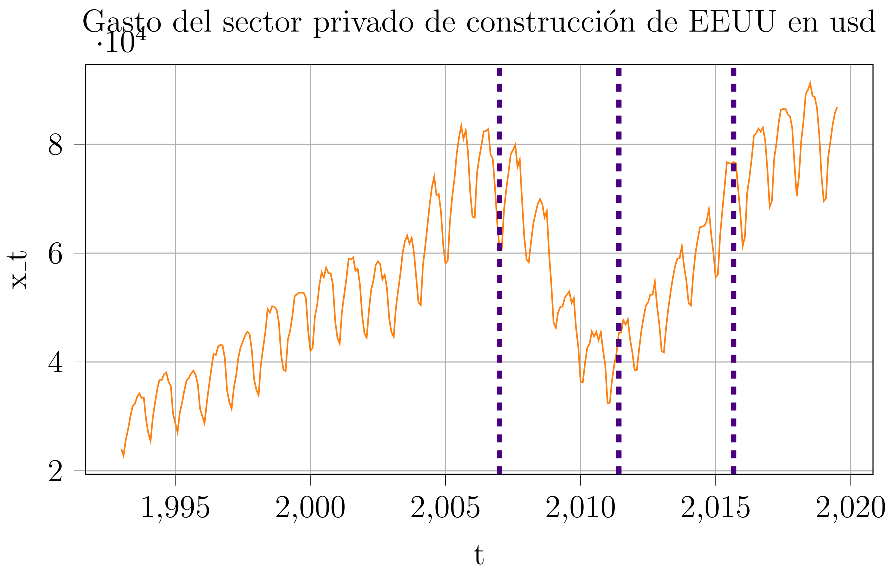
Los datos incluyen el costo de la mano de obra y los materiales, trabajos de arquitectura e ingeniería, intereses e impuestos pagados durante la construcción y las ganancias de los contratistas.

La recolección de datos y estimación de las actividades comienza el primer día de cada mes después del mes de referencia y continua por tres semanas. Los datos reportados y estimados son para las actividades realizadas el mes anterior, esta encuesta se realiza desde 1960.

El Bureau de Análisis Económico utiliza esta data directamente para producir las estadísticas de PBI. Otros agencias gubernamentales y relacionadas con la construcción utilizan estos datos pronósticos económicos, análisis de mercado y la toma de decisiones financieras.

Aplicación de algoritmo de detección de changepoint basado en paper de Zuo

Al aplicar el algoritmo frecuentista podemos observar los siguientes resultados



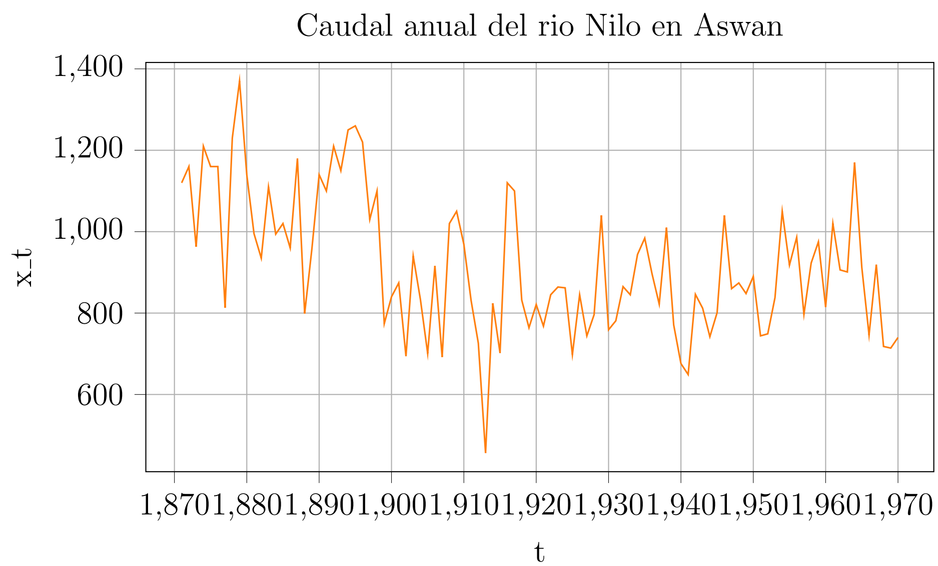
Se detectan 3 changepoints en la serie al usar una ventana de tiempo $\tau = 30$. Los primeros dos puntos detectados por el algoritmo, corresponden a épocas de cambio en la serie de tiempo, que podrían correlacionarse con fenómenos propios de la economía norteamericana en este período, por ejemplo la crisis subprime del 2008. El ultimo de los changepoints podríamos considerarlo un falso positivo, sin embargo observando con detenimiento el final de la serie no descartamos la posibilidad de que la pendiente de la recta de cuadrados mínimos se estuviese aplanando. Hacen falta mas datos para chequear esto aun así consideramos que existe cierta probabilidad no nula de que efectivamente sea otro punto de cambio. Otro indicio de esto encontramos al mirar los últimos 3 mínimos locales de la serie, tienen valores muy similares a comparación de los mínimos locales anteriores.

Es importante mencionar que en este ejemplo no estamos bajo las hipótesis del algoritmo ya que esta serie no tiene la forma $y_t = \beta t + \alpha + \varepsilon_t$ por la presencia de un cierto fenómeno estacionario, consideramos que esto una fortaleza del algoritmo.

Podemos concluir que en este ejemplo el algoritmo demuestra una buena performance, no solo puede detectar cambios estando fuera de sus hipótesis sino que también lo logra rápidamente.

6.0.2. Caudal anual del río Nilo

El siguiente set de datos representa el volumen anual del rió Nilo (las unidades se encuentran en $10^8 m^3$) recaudados en Aswan, Egipto entre los años 1871 y 1970.



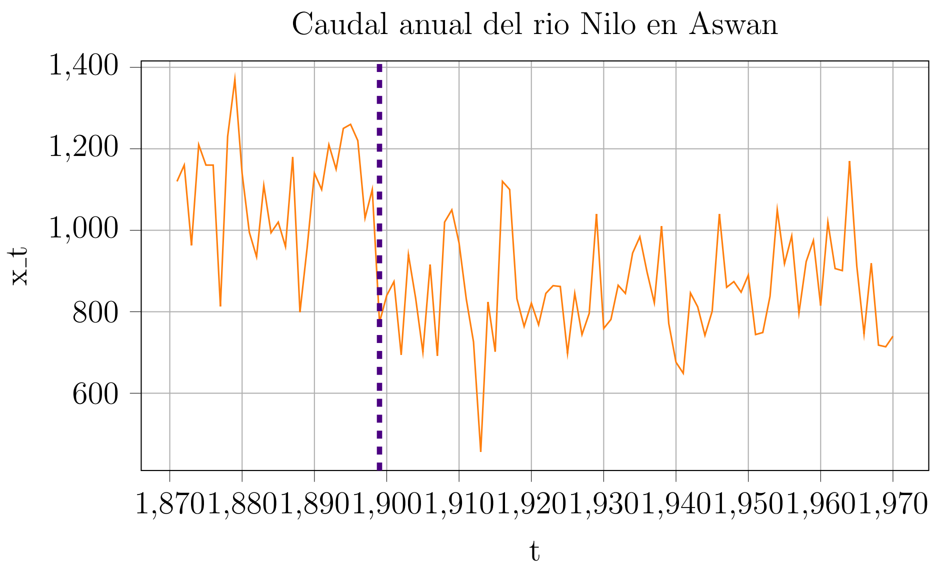
Los datos fueron extraídos del libro [Dur01]. El cual aplica métodos de análisis de espacio para series de tiempo. Anteriormente el problema de changepoint del río Nilo también fue analizado en [Cob78] y [Bal93] con un enfoque especifico a changepoints.

Alrededor del año 1898 el flujo anual cae abruptamente de 1100 a 800 debido a la construcción de una represa. Testearemos si este cambio es detectado por el método bayesiano basado en el paper de Adams.

Aplicación de algoritmo bayesiano

Existen varias consideraciones antes de aplicar este algoritmo empezando por que tipo de cambio queremos detectar ya que para cada uno de estos se utilizan funciones a priori y conjugadas diferentes. En este caso podemos ver toda la data de antemano, esto es inevitable de todos modos por lo tanto decidimos detectar cambios de media. Al igual que en el caso anterior tampoco estamos exactamente bajo las hipótesis ya que esta serie no esta dada por una distribución normal, pero podemos ver que se comporta bien.

Para estimar los parámetros de la distribución a priori tomamos la media de los datos antes del punto de cambio y elegimos una varianza teniendo en cuenta cuanto era el salto efectivamente de las medias. La varianza de la serie de tiempo fue estimada promediando las varianzas muestrales de la serie antes y después del cambio. Por ultimo definimos una constante de Hazard igual a 0.001. Consideramos que estas son desventajas del algoritmo bayesiano, sin embargo en la practica estas decisiones siempre deberán tomarse.



Bajo estos supuestos el algoritmo detecta exactamente el salto que se produce en la serie en el año que se construye la represa. Tampoco detecta falsos positivos en este caso, eso es otro punto a favor.

Como vimos en los casos de test y en este caso real el algoritmo se porta muy bien, no obstante remarcamos que el estimar los parámetros de la distribución a priori, la varianza y la constante de hazard son trabajos no triviales que se deben tener en cuenta a la hora de aplicar esta solución.

Capítulo 7

Conclusiones

A lo largo de esta tesis estudiamos diferentes métodos para detectar changepoints en series de tiempo. Lo que comenzó siendo un trabajo buscando aplicaciones prácticas termino adquiriendo un enfoque mixto. Al momento de entender por que estos algoritmos funcionan encontramos dos situaciones muy distintas.

Por un lado el algoritmo propuesto en [Zuo+19] se desprende directamente de la teoría de regresión lineal clásica, sin embargo no encontramos forma de hacer esto con el algoritmo del paper de Adams. Es en la construcción de un marco teórico utilizando las herramientas proporcionadas por los modelos gráficos y en la demostración de las diferentes igualdades mencionadas en el algoritmo de Adams donde este trabajo encuentra su valor principal.

A su vez pudimos explorar distintas aplicaciones de los distintos algoritmos, tanto en casos de prueba como en dos casos reales. Acerca de esto tenemos varias cosas a destacar que podemos dividirlas en pros y contras.

Propuesta bayesiana

Pros

- Posee muy buena performance bajo sus hipótesis en los casos de prueba.
- Puede detectar distintos tipos de cambios, media, varianza y ambos a la vez, o el parámetro que caracterice a la distribución en cuestión.
- Detecta los cambios rápidamente e indica el tiempo en el que se producen.

- Detectó correctamente el changepoint cuando se lo aplicó a un caso real no muy lejano pero por fuera de sus hipótesis.

Contras

- Depende fuertemente de elegir correctamente la función a priori y la constante de Hazard.
- Para cada tipo de changepoint se requiere programar un algoritmo distinto.
- Depende de la existencia de funciones a priori conjugadas para el calculo de las probabilidades. Si bien esto es relajable, el cálculo del posterior predictive se tornaría bastante más complejo, requieriendo técnicas del tipo Markov Chain Monte Carlo para estimar distribuciones.

Algoritmo frecuentista

Pros

- Posee buena performance bajo sus hipótesis en los casos de prueba.
- Fácil de programar y corre velozmente.
- Detectó correctamente el changepoint cuando se lo aplicó a un caso real fuera de sus hipótesis.
- Puede ser útil para detectar otros tipos de changepoints que no sean cambios de tendencias, por ejemplo cambios de medias.
- Es parcialmente robusto a outliers, dependiendo del ancho de la ventana.

Contras

- Encuentra los changepoints con retraso, muy dependiente del ancho del a ventana.
- No detecta checkpoints cercanos (en ventanas mas chicas que 2τ.
- Al volver a prender el algoritmo si el changepoint se encuentra en la primer ventana de tiempo detecta un cambio de pendientes pero no reconoce esto.

De cara a trabajos futuros existen muchos otros tipos de test que se pueden realizar, mencionaremos algunos a continuación. Se pueden hacer experimentos probando cuando los parámetros de la a priori se encuentran lejos de los "reales", probar distintos tipos de distribuciones fue de las hipótesis, probar método Monte Carlo cuando no existen a prioris conjugadas. Para el algoritmo frecuentista se puede modificar la implementación online para no apagar el algoritmo una vez que detecta un changepoint, por ejemplo achicando la ventana temporal durante cierto intervalo. También sería posible considerar casos más generales, quizás con genuinas series de tiempo como podrían ser procesos ARIMA.

Bibliografía

[AM07] Ryan Prescott Adams y David J.C. MacKay. "Bayesian Online Changepoint Detection". En: arXiv (2007). URL: https://arxiv.org/abs/0710.3742v1.

[Ame85] Takeshi Amemiya. Advanced Econometrics. Harvard University Press, 1985. ISBN: 0674005600.

[Bal93] N. S. Balke. Detecting level shifts in time series. Business y Economic Statist, 1993.

[Bel70] Richard Bellman. Introduction to Matrix Analisys. McGraw-Hill Book Company, 1970.

[BW20] Gerrit J.J. van den Burg y Christopher K.I. Williams. "An Evaluation of Change Point Detection Algorithms". En: arXiv (2020). URL: https://arxiv.org/abs/2003.06222.

[Cha17] Matt Chapman. A Meta-Analysis of Metrics for Change Point Detection Algorithms. 2017.

[Cob78] G. W. Cobb. The problem of the Nile: conditional solution to a change point problem. Biometrika, 1978.

[CZ14] Christopher Natoli Cody Buntain y Miroslav Zivkovic. A Brief Comparison of Algorithms for 2014. URL: https://github.com/%20cbuntain/ChangePointDetection.

[Dur01] J. Durbin. Time Series Analysis by State Space Methods. Oxford University Press, 2001.

[Dym11] P. Dymarski. Hidden Markov Models: Theory and Applications. IntechOpen, 2011. ISBN: 9789533072081. URL: https://books.google.com.ar/books?id=GS2RDwAAQBAJ.

[Edw95] David Edwards. Introduction to Graphical Models. Springer, 1995.

[GVP13] Dan Geiger, Tom S. Verma y Judea Pearl. "d-Separation: From Theorems to Algorithms". En: CoRR abs/1304.1505 (2013). arXiv: 1304.1505. URL: http://arxiv.org/abs/1304.1505.

[Gel14] Andrew Gelman. Bayesian Data Analysis. Taylor y Francis Group, 2014.

[KS09] Yoshinobu Kawahara y Masashi Sugiyama. Change-point detection in time-series da
2009.

[KK12] David Kleinbaum y Mitchel Klein. Survival Analysis. Springer, 2012.

[MJ12] David S. Matteson y Nicholas A. James. A nonparametric approach for multiple cha
arXiv: 1306.4933, 2012.

[Pea00] Judea Pearl. Causality. Cambridge University Press, 2000.

[Rai00] Howard Raiffa. Applied Statistical Decision Theory. Wiley Classics Library Edition, 2000.

[Ran71] William M. Rand. Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association, 1971. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1971.10482356.

[Rao47] C.R. Rao. Large sample tests of statistica hypotheses concerning several parameter
Proceeding of the Cambridge Philosophical Society, 1947.

[Roi17] Violeta Roizman. Seleccion de modelos graficos no dirigidos en el contexto de alta
Universidad de Buenos Aires, 2017. URL: http://cms.dm.uba.ar/academico/carreras/licenciatura/tesis/2017/Roizman.pdf.

[Was03] Larry Wasserman. All of Statistics - A Concise Course on Statistical Inference. Springer, 2003.

[Zuo+19] Bin Zuo y col. "A new statistical method for detecting trend turning". En: A new statistical method for detecting trend turning (2019). URL: https://link.springer.com/article/10.1007/s00704-019-02817-9.