Análisis de conglomerados sobre series de tiempo. Utilizando una modificiación de la divergencia de Kullback-Liebler sobre la densidad espectral

Facultad de Ciencias Actuariales
Alumno: Campos Martínez Joaquín Nicolás
Asesor: Dr. Cuevas Covarrubias Carlos

Índice general

Introducción

Nuestro objetivo es crear un algoritmo, o una regla que se pueda utilizar como apoyo para decidir si una serie de tiempo es autorregresiva o de promedios móviles.

Para crear esta regla nos basamos en análisis discriminante, utilizando la medida de discrepancia de Kullback-Liebler. Está requiere de funciones de densidad, por eso tomamos la función de densidad espectral de la serie de tiempo.

La función de densidad espectral está en relación biyectiva con la función de autocorrelación parcial, por lo mismo hacer una comparacón entre funciones de densidad espectral es comparar la estructura de varianza de las series.

La metodología utilizada para obtener la función de densidad espectral se obtuvo del Priestley

1. Marco Teórico

1.1. Objetivo

Nuestro objetvio es elaborar un algoritmo que sirva como herramienta de apoyo para distinguir entre series de tiempo con componente autorregresiva y con componente de promedios móviles utilizando la divergencia de Kullback-Liebler sobre las funciones de densidad espectral como la medida de clasificación

1.2. Relevancia

Las series de tiempo pueden ser observadas en diferentes ámbitos, información de ventas, precios de acciones, tasas cambiarias, información sobre el clima, datos biomédicos, etc. (Saeed Aghabozorgi, 2015)

Las aplicaciones de análisis de conglomerados en estos ámbitos son variadas, por ejemplo, en medicina es importante reconocer la diferencia entre señales normales en un ECG, EEG, EMG de aquellas producidas por enfermedades. En sismología, para discriminar las ondas sísmicas, de los moviemientos naturales de la tierra (Corduas Marcella, 2015) o para monitorear violaciones al CTBT (Tratado de prohibición completa de los ensayos nucleares, en español) (Shumway, 2003).

Otras aplicaciones son, en astronomía las curvas de luz muestran el brillo de una estrella en un periodo de tiempo, en medicina la actividad cerebral, en el medio ambiente y urbanización los niveles de la marea, niveles de contaminantes en el aire ($PM_{2,5}, PM_{10}$) (Saeed Aghabozorgi, 2015).

En la práctica, uno no sabe *a priori* si una serie de tiempo tiene una componenente de promedios móviles o una autorregresiva. En la metodología Box-Jenkins no hay una forma clara de identificar esto, y nos basamos en las funciones de autocorrelación y autocorrelación parcial.

Nosotros queremos proponer un algoritmo que sirva como herramienta para comprobar si una seríe de tiempo contiene componentes AR o MA.

2. Algoritmo

Para desarrollar el algoritmo se realizaron varias tareas, primero simular series de tiempo MA o AR, luego estimar la función de autocorrelación parcial, después obtener el periodograma (estimador de la función de densidad espectral). Luego se obtienen promedios de las funciones de densidad espectral (Shumway, 2003), para poder utilizar la divergencia de Kullback-Liebler como una regla de decisión para clasificar una serie de tiempo en el grupo correspondiente, está es la metodología propuesta por (Yoshihide Kakizawa, 1998)

2.1. Simulador de series de tiempo

R tiene integrado simuladores, pero por objevitos didactivos el simulador se realizo utilizando únicamente la función rnorm y la serie se contruye con el supuesto de que $\mu 0$ es decir nuestros modelos son de las siguientes formas

- AR $Y_t = \sum_{i=1}^{n} \varphi_i Y_{t-i} + \varepsilon_t$ donde $Y_i = 0$ para $i = 1, \ldots, n$ y ε_t $norm(0, \sigma^2)$
- MA $Y_t = \sum_{i=1}^{n} \theta_i \varepsilon_{t-i} + \varepsilon_t$ donde $\varepsilon_i = 0$ para $i = 1 - n, \ldots, 0$ y ε_t $norm(0, \sigma^2)$ para $t = 1, \ldots, n$
- $Y_t = \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \theta_{12} Y_{t-1} Y_{t-2} + \varepsilon_t$ donde $Y_i = 0$ para $i = 1, 2$ y ε_t $norm(0, \sigma^2)$

2.2. Densidad espectral

El estimador que utilizamos de la función de densidad espectral es el periodograma, que requiere el estimador de la función de autocorrelación. El periodograma es un estimador incosistente dado que para una serie de tamaño N, los valores en n de la función de autocorrelación, cuando n se acerca a N están basados en pocas observaciones y tienden a tener mayor varianza. Aun así es el estimador más sencillo, y es la forma muestral de

modelo teórico.

$$I_N^*(w) = \frac{1}{2\pi} \sum_{s=-(N-1)}^{N-1} \hat{R}(s) e^{isw}$$

Donde

$$\hat{R}(s) = \frac{1}{N} \sum_{t=1}^{N-|s|} (y_t - \bar{y})(y_{t+|s|} - \bar{y})$$

$I_N^*(w)$ es la densidad no normaliza, de modo

$$\hat{f}(w) = \frac{I_N^*(w)}{\sigma_y^2}$$

Nótese que $\hat{R}(0) = \sigma_y^2$, es decir, la varianza de la serie de tiempo, de modo que el periodograma'

2.3. Series de tiempo

Una simplificación de la definición de una serie de tiempo es: un proceso que varía en el tiempo, observaciones tomadas secuencialmente en el tiempo. En la vida real podemos observar varios fenómenos de este tipo, como los mencionados arriba.

Definición. Una serie de tiempo es una colección de de observaciones hechas en una secuencia de tiempo. Se dice que una serie de tiempo es continua si las observaciones son en tiempos continuos. Se dice discreta si cuando las observaciones se toman en tiempos específicos, usualmente separados a intervalos iguales

2.4. Análisis de conglomerados

«Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups.» (Pradeep Rai, 2010)

Análisis de conglomerados no es una técnica asociada a un único tipo de problemas. Pero es asociado naturalmente a las ideas de grupos homogéneos, clases de equivalencia, datos multimodales.

Punj & Stewart identifican algunas de las principales aplicaciones en el campo de investigación de mercado como segmentación de mercado, identificación de grupos homogéneos de consumidores, desarrollo de nuevos productos potenciales, selección de un mercado de prueba, como técnica general de reducción.

Las técnicas aplicadas son varias, y algunas de las más tratadas en la literatura son conglomerados jerárquicos, y algunos algoritmos de k-centroides.

Tenemos que reconocer que nuestro objeto de estudio son observaciones multivariadas, medidas en p variables, que pueden ser categóricas (respuestas a una encuesta, sí, no) o numéricas (valores continuos, como estatura y peso)

2.4.1. Conglomerados Jerárquicos

Se separa en 2 grandes categorías, aglomerativos y divisivos. El primer tipo considera cada observación, elemento de la muestra, como un conglomerado y comienza a agruparlos basado en alguna regla. El segundo comienza con toda la muestra como un gran conglomerado y comienza a separarlo en distintos conglomerados, en su mayoría conforme a una función objetivo.

Un análisis de conglomerados jerárquico es sensible, o depende principalmente de 2 factores.

- La medida de **similitud o distancia** elegida
- El algoritmo para agrupar elementos, la medida de **distancia o similitud** entre conglomerados

Conglomerados jerárquicos aglomerativos suelen ser los típicos ejemplos de libro de texto por su facilidad para explicarse, sin embargo esto no significa que su utilidad sea menor.

Primero debemos abordar el tema de distancia o simulitud.

Definición. Dado un conjunto X una medida de distancia es una función $d : X \times X \to \mathbb{R}$ tal que cumple las siguientes propiedades

- $d(x, y) = 0$ si y sólo si $x = y$
- $d(x, y) = d(y, x)$ para todas $x, y \in X$
- $d(x, y) + d(y, z) \geq d(x, z)$ para todas $x, y, z \in X$

La propiedad de no negatividad $d(x, y) \geq 0$ para todas $x, y \in X$ queda definida por las propiedades anteriores.

La idea de distancia suele ser intuitiva, y se asocia comúnmente con la distancia euclidiana, sin emabargo existen muchas otras formas de distancia, como la suma de valores absoulutos de las componentes. En cambio, la simulitud, aunque la idea es intuitiva, no suele ser tan clara cuando uno maneja valores numéricos o cateogóricos.

Exiten algunos intentos sobre la definición de una medida de similitud, entre ellos mencionaremos como ejemplo la definicón presentada por Chen, Ma y Zhang (Chen Shihyen, 2009)

Definición. Dado un conjunto X una medida de similitud es una función $s : X \times X \to \mathbb{R}$ que cumple las siguientes propiedades

- $s(x, y) = s(y, x)$ para todo $x, y \in X$
- $s(x, x) \geq 0$ para todo $x \in X$
- $s(x, x) \geq s(x, y)$ para todo $x, y \in X$
- $s(x, y) + s(y, z) \leq s(x, z) + s(y, y)$ para todo $x, y, z \in X$
- $s(x, x) = s(y, y) = s(x, y)$ si y sólo si $x = y$

La propiedad de simetría es intuitiva, se espera que un objeto sea tan similar a otro, como el otro al primero. La segunda propiedad no es necesaria, pero su importancia radica en el concepto que tenemos de similitud, un objeto no puede tener similitud negativa con respecto de si mismo. La tercer propiedad parte de la idea de que un ningún obteto es tan similar a otro como a sí mismo.

La cuarta propiedad sería el equivalente a la desigualdad del triángulo, y aunque a primera vista parece poco comprensible, es más fácil analogarla con la idea de intersección de conjuntos, podemos decir que s es una medida de todo lo que tienen en común 2 objetos x, y. La última propiedad nos dice que sólo cuando 2 objetos son iguales se alcanzan las cuotas superiores dela similitud.

No es importante recordar esta definición de similitud, pues nosotros trabajaremos algo más parecido a una distancia que a una similitud.

Lo siguiente es construir una matriz de distancias o similitudes entre los objetos. Si denotamos $s_{ij} = s(x_i, x_j)$ y $d_{ij} = d(x_i, x_j)$, y tenemos un total de n observaciones, nuestra matriz de similitud es

$$\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{bmatrix}$$

y nuesta matriz de distancias es

$$\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}$$

En un algoritmo aglomerativo el siguiente paso es juntar las 2 observaciones, distintas, más similares (menos distantes) para formar un conglomerado, luego se recalcula la distancia del resto de los conglomerados (observaciones) al conglomerado recién formado.

¿Cómo se calcula la distancia entre 2 conglomerados?. Existen distintas reglas para calcular la distancia, algunas de las más conocidas son

- Vecino más cercano (single linkage), es decir, la distancia entre 2 conglomerados es la distancia más corta de sus elementos
- Vecino más lejano (complete linkage), es decir, la distancia entre 2 conglomerados es la distancia más grande entre sus elementos
- Promedio (average linkage), media de las distancias entr los elementos de ls conglomerados.
- Centroide, se obtiene el centro de masa de los conglomerados y se calcula la distancia entre los mismo

Este proceso se repite hasta tener toda la muestra en un conglomerado, y cuando uno decida hacer el corte, cuando hay r conglomerados, donde el usuario escoge r.

2.4.2. K-centroides

Los algoritomos de k-centroides están basados en la minimización de una función de suma de cuadrados, es decir la suma de cuadrados de la distancia de los elementos de un conglomerado a su centroide.

En comparación con un análsis jerárquico, k-centroides depende de 2 configuraciones

- La primera es el número de conglomerados, k
- El tipo de centroide, centro de gravedad, medoide.

En el caso de **k-means**, el centroide es el centro de gravedad y

$$E = \sum_{h=1}^{k} \sum_{i=1}^{n} u_{i,h} d(x_i, \mu_h)^2$$

Es la función que deseamos optimizar (Hana, 2014). Donde los u_{ih} clumplen que $\sum_{h=1}^{k} u_{ih} = 1$ para $i = 1, \ldots, n$ y $\sum_{i=0}^{n} u_{ih} = 1$ para $h = 1, \ldots, k$, y μ_h es la la media del h-ésimo conglomerado. En el caso de medoides sustituimos μ_h por m_h, donde m_h es el medoide del h-ésimo.

Hasta ahora sólo hemos presentado metodologías en las que una observación no puede ser parte de más de un conglomerado sin embargo existen otros algoritmos que asignan «grados de pertenencia» en lugar de un conglomerado fijo.

2.5. Análisis espectral

Inferencias basadas en la función de densidad espectral se dice que es un análisis in el Dominio de las frecuencias.

Teorema 1 (Wiener-Khintchine). Para cualquier proceso estocástico estacionario con función de autocovarianza $\gamma(k)$ existe una función monotamente creciente $F(W)$ tal que

$$\gamma(k) = \int_0^{\pi} \cos(wk) dF(w)$$

Llamada la representación espectral de la función de autocovarianza, que involucra un tipo de integral llamada de Stieltjes

$F(w)$ tiene una interpretación física directa, es la contribución a la varianza que se puede atribuir a las frecuencias en el rango $(0, w)$ $F(w)$ es monótona en el intervalo (o, π), así se puede descomponer en 2 funciones $F_1(w)$ y $F_2(w)$ talque

$$F(w) = F_1(w) + F_2(w)$$

Donde $F_1(w)$ es una función continua no decreciente y $F_2(w)$ es una función escalonada no decreciente (WOLD descomposition). Además $F_1(w)$ se relaciona con la componente puramente no determinista del proceso y $F_2(w)$ se relaciona con la componente determinista. Nuestro estudio es sobre los procesos puramente indeterminados, donde $F_2(w) = 0$, de tal forma que $F(w)$ es continua en $(0, \pi)$. La potencia está directamente relacionado al cuadrado de la amplitud de oscilación. y la varianza es la potencia total

Definición. La forma normalizada de $F(w)$ está dada por

$$F^* = \frac{F(w)}{\sigma_x^2}$$

De modo que $F^*(w)$ es la proporción de varianza atribuida por las frecuencias en el rango $(0, w)$. Como $F^*(\pi) = 1$ y es además monóna creciente tenemos que $F^*(w)$ tiene propiedades similares a una función de distribución acumulada. (Chatfield, 1989)

2.6. Kullback-Liebler

Definición. Dadas dos funciones de densidad de probabilidad f, g definidas sobre un mismo espacio medible, y que son absolutamente continuas entre sí, la divergencia de Kullback-Liebler se define como

$$d_{KL}(f, g) = \int f(x) \ln \left(\frac{f(x)}{g(x)} \right) dx$$

La divergencia de Kullback-Liebler como una medida de divergencia dirigida. Una manera de entender lo anterio es primero considerar un modelo verdadero f y un modelo de aproximación g, entonces la divergencia de Kullback-Liebler nos habla de cuanta información se pierde utilizando g como aproximación (Kenneth P. Burnham, 2007). La divergencia de Kullback-Liebler no cumple con las propiedades de distancia mencionadas anteriormente.

La propiedad más cercana a una distancia, es la positividad y la nulidad cuando f y g son iguales exceptuando en puntos de probabilidad 0.

Ahora podemos definir la divergencia de Jeffreys

Definición. Dadas dos funciones de densidad de probabilidad f, g definidas sobre un mismo espacio medible, y que son absolutamente continuas entre sí, la divergencia de Jeffreys se define como

$$d_J(f, g) = d_{KL}(f, g) + d_{KL}(g, f) = \int (f(x) - g(x)) \ln \left(\frac{f(x)}{g(x)} \right) dx$$

En este caso, tenemos dos modelos, y la divergencia de Jeffreys nos indica la dificultad para distinguir entre ambos (Kullback, 1978). Esta nueva medida de divergencia soluciona el problema de la simétria'

2.7. Antecedentes

Hay múltiples artículos sobre análisis de conglomerados, y algunos de ellos se enfocan en series de tiempo. *Time-series clustering - A decade review* (Saeed Aghabozorgi, 2015) tiene como objetivo comparar los enfoques más populares.
«This review will expose four main components of time-series clustering and is aimed to

represent an updated investigation on the trend of improvements in efficiency, quality and complexity of clustering time-series approaches during the last decade and enlighten new paths for future works» (Saeed Aghabozorgi, 2015)

Clustering of time series data - a survey busca resumir investigación realizada en el tema y sus aplicaciones en varios campos. Incluye aspectos basicos de los algoritmos más generales que se utilizan en estudios de conglomerados, medidas de similitud y disimilitud, evaluación de conglomerados. (Liao, 2005)

Ambos artículos distinguen 3 principales categorizaciones de los acercamientos a conglomeración de series de tiempo, trabajar con los datos brutos, extraer «características» de los datos, basarse en un modelo al que se ajusten los datos.

«Computational Models of Music Similarity and their Application in Music Information Retrieval» es una tesis que busca mostrar el desarrollo de las técnicas de descubrimiento de música basado en medidas de similitud y disimilitud. Uno de los enfoques consiste en utilizar de función de densidad espectral. Sin embargo también utiliza técnicas aplicables a series de tiempo. (Pampalk, 2006).

3. Alcance del proyecto

3.1. Metodología

Ya mencionamos los conceptos principales detrás del proyecto, *series de tiempo*, *análisis espectral*, *divergencia de Kullback-Liebler* y *análisis de conglomerados*. La meta del proyecto es utilizar los conceptos anteriores sobre distintas series de tiempo, precios de acciones, clima, *música*, etc y analizar los resultados obtenidos.

Nuestra propuesta es que utilizar la divergencia de Kullback-Liebler como medida de distancia, y comparar con respecto de otros acercamientos a la conglomeración de series de tiempo.

Nos centraremos en conglomerados jerárquicos, sin embargo con la idea detras de la divergencia de Kullback-Liebler, la aplicación de k-medoides suena como una idea intuitiva, donde queremos minimizar la divergencia de los elementos de un conglomerado con respecto de un «centroide» (medoide).

La información a utilizar será extraída de Yahoo finance o un servicio similar, con fines puramente didácticos. Eventualmente deseamos crear un algoritmo y aplicación que permita observar los resultados obtenidos de la aplicación de un algoritmo basado en las ideas de este proyecto.

La misma idea desea ser aplicada, sobre series de tiempo basadas en música, pero no se considerá dentro del alcance de este trabajo.

Para el análisis de datos se utilizará el software R, y las librerías tuneR y seewave, aunque no se descartan otras aplicaciones que puedan ser más útiles.

Por cuestiones de optimización, no se utilizarán las series completas, sino que se reduciran de manera que sean comparables, longitudes iguales y a tiempos iguales.

3.2. Fuentes de datos

La cantidad de información disponible actualmente es impresionante, y se puede obtener de distintas formas, mediante bases de datos online, minando información, o simplemente checando el registro de nuestras acciones.

Nuestra principal fuente son los precios de acciones, adquiridos mediante Yahoo Finance ©. Y en la medida de lo posible se buscará utilizar otras bases de datos disponibles. Y también archivos de audio, música, obtenidos mediante compra en distintos servicios iTunes Store©, Bandcamp©, o adquiridos gratuitamente.

Algunos de los ejemplos que realizaremos serán estilo «libro» utilizando bases de datos populares disponibles en el sitio web de la *International Association for Statistical Computing*

3.3. Pasos siguientes

- 13-Sep Presentación de los avances, Actualización del escrito
- 20-Sep Anexos de los temas complejos
- 27-Sep Implementar en código el clasificador
- 4-Oct Experimentos y resultados con series AR y MA simuladas
- 11-Oct Experimentos y resultados con series reales
- 18-Oct Correcciones y actualización de escrito
- 25-Oct Verificación de escrito, sobre detalles finales
- 1-Nov Entregar borrador de presentación y escrito a sinodales
- 8-Nov Dar diseño a cartel (Imágenes, gráficas y diseño en general)
- 15-Nov Ajustes Finales para presentación de cartel
- 22-Nov Detallar por Completo escrito Final
- 29-Nov Detallar por Completo escrito Final

Referencias

Chatfield, C. (1989). *The analysis of time series : an introduction.*

Chen Shihyen, Z. K., Ma Bin. (2009). *On the similarity mectric and the distance metric.* https://www.journals.elsevier.com/theoretical-computer-science. ([Online; Obtenido 20-November-2017])

Corduas Marcella, P. D. (2015). *Time series clustering and classification by the autoregressive metric.* http://www.sciencedirect.com/science/article/pii/S0306437915000733?via%3Dihub. ([Online; Obtenido 01-Septiembre-2017])

Hana, R. (2014). *Cluster analysis of economic data.* (University of Economics, Prague, Czech Republic)

Kenneth P. Burnham, D. R. A. (2007). *Model selection and multimodel inference: A practical information-theoretic approach.* ([Online; Obtenido 28-November-2017])

Kullback, S. (1978). *Information theory and statistics.* ([Online; Obtenido 30-November-2017])

Liao, T. W. (2005). *Clustering of time series data - a survey.* ([Online; Obtenido 09-Octubre-2017])

Pampalk, E. (2006). *Computational models of music similarity and their application in music information retrieval.* ([Online; Obtenido 09-Octubre-2017])

Pradeep Rai, S. S. (2010). *A survey of clustering techniques.* http://www.ijcaonline.org/volume7/number12/pxc3871808.pdf. ([Online; Obtenido 01-Septiembre-2017])

Saeed Aghabozorgi, T. Y. W., Ali Seyed Shirkhorshidi. (2015). *Time-series clustering – a decade review.* http://www.sciencedirect.com/science/article/pii/S0306437915000733?via%3Dihub. ([Online; Obtenido 01-Septiembre-2017])

Shumway, R. H. (2003). *Time-frequency clustering and discrimination analysis.*

Yoshihide Kakizawa, M. T., Robert H. Shumway. (1998). *Discrimiantion and clustering for multivariate time series.*

Anexos A: Teoría

4. Anexo B: Código