Traballo Fin de Máster

Análise cluster de series de tempo baseada en modelos

Ana Mayán Carneiro

Máster en Técnicas Estatísticas

Curso 2019-2020

Proposta de Traballo Fin de Máster

Título en galego: Análise cluster de series de tempo baseada en modelos
Título en español: Análisis cluster de series de tiempo basado en modelos
English title: Cluster analysis of time series based on models
Modalidade: Modalidade A
Autora: Ana Mayán Carneiro, Universidade da Coruña
Directores: José Antonio Vilar Fernández, Universidade da Coruña; Borja Raúl Lafuente Rego, Universidade da Coruña
Breve resumo do traballo: Unha vía de interese para desenvolver análise cluster baseada en modelos mixtos é asumir que os datos seguen unha mixtura de distribución de xeito que cada compoñente nesta mixtura describe a natureza probabilística do grupo ou cluster. Cando os datos son series temporais, esta vía non é tan sinxela porque habitualmente as realización das series son longas e isto tradúcese nun problema de alta dimensión no procedemento de análise cluster. Neste proxecto proponse explorar esta vía de análise cluster asumindo series autorregresivas e aproximando o modelo mixto subxacente por máxima verosimilitude mediante algoritmos EM. Desenvolverase código en R para implementar as solucións propostas e realizarase unha análise comparativa das mesmas mediante datos simulados.
Recomendacións:
Outras observacións:

Don José Antonio Vilar Fernández, Catedrático da Universidade da Coruña, e Don Borja Raúl Lafuente Rego, Investigador Asociado da Universidade da Coruña, informan que o Traballo Fin de Máster titulado

Análise cluster de series de tempo baseada en modelos

foi realizado baixo a súa dirección por dona Ana Mayán Carneiro para o Máster en Técnicas Estatísticas. Estimando que o traballo está terminado, dan a súa conformidade para a súa presentación e defensa ante un tribunal.

En A Coruña, a 8 de setembro de 2020.

O director:	O director:	A autora:
Don José Antonio Vilar Fernández	Don Borja Raúl Lafuente Rego	Dona Ana Mayán Carneiro

Índice xeral

Resumo

Unha vía de interese para desenvolver análise cluster baseada en modelos mixtos é asumir que os datos seguen unha mixtura de distribución de xeito que cada compoñente nesta mixtura describe a natureza probabilística do grupo ou cluster. Cando os datos son series temporais, esta vía non é tan sinxela porque habitualmente as realización das series son longas e isto tradúcese nun problema de alta dimensión no procedemento de análise cluster. Neste proxecto proponse explorar esta vía de análise cluster asumindo series autorregresivas e aproximando o modelo mixto subxacente por máxima verosimilitude mediante algoritmos EM. Desenvolverase código en R para implementar as solucións propostas e realizarase unha análise comparativa das mesmas mediante datos simulados.

Resumen

Una vía de interés para desarrollar análisis cluster basado en modelos es asumir que los datos siguen una mixtura de distribuciones de modo que cada componente en esta mixtura describe la naturaleza probabilística del grupo o cluster. Cuando los datos son series temporales, esta vía no es sencilla porque habitualmente las realizaciones de las series son largas y ello se traduce en un problema de alta dimensión en el procedimiento de análisis cluster. En este proyecto se propone explorar esta vía de análisis cluster asumiendo series autorregresivas y aproximando el modelo mixto subyacente por máxima verosimilitud mediante algoritmos EM. Se desarrollará código en R para implementar las soluciones propuestas y se realizará un análisis comparativo de las mismas mediante datos simulados.

Abstract

One way of interest in developing model-based cluster analysis is to assume that the data follow a mixture of distributions so that each component in this mixture describes the probabilistic nature of the group or cluster. When the data are time series, this route is not easy because the series realizations are usually long and this translates into a high-dimensional problem in the cluster analysis procedure.

In this project it is proposed to explore this path of cluster analysis assuming autoregressive series and approximating the underlying mixed model by maximum likelihood using EM algorithms. R code will be developed to implement the proposed solutions and a comparative analysis of them will be carried out using simulated data.

Introdución

O presente Traballo Fin de Máster ten como obxectivo aplicar os coñecementos adquiridos ao longo do Máster en Técnicas Estatísticas organizado polas tres universidades galegas, pero máis concretamente no campo da análise cluster de series temporais.

A análise cluster está composta por procedementos estatísticos que teñen como fin o de agrupar un conxunto de datos en varios clusters de tal xeito que os elementos que se atopen no mesmo cluster presenten características comúns e sexan o máis homoxéneos posible, así como entre os distintos clusters teñan características diferentes. En particular, este traballo céntrase na análise cluster de series temporais, a cal ten como obxectivo dividir un conxunto de series de tempo en diferentes grupos ou clusters. As series temporais son datos dinámicos, o que supón unha complexidade adicional ao problema de desenvolver cluster de xeito que moitas das técnicas cluster que se empregan poderían ser erroneamente aplicadas se se traballa con series temporais xa que as métricas que se utilizan deben diferenciar a conduta no tempo e non simplemente a conduta estática nun instante de tempo. A agrupación para asignar as series a cada cluster basease na similitude que teñen as series entre si, así as series no mesmo grupo serán máis similares entre si que as series dos diferentes grupos. Estas dificultades adicionais que supón traballar con datos dinámicos complican o concepto de similaridade e increméntase a dimensionalidade dos datos.

A análise cluster é un tema clásico de análise multivariante, é unha técnica estatística moi antiga, pero o problema de cluster de series temporais xorde máis recentemente, sobre a última década do século XX. A partir desa data empézase a ter preocupación e interese sobre este tema e o número de publicacións aumenta exponencialmente. En particular recibe esta atención porque o número de aplicacións en distintas áreas de coñecemento como pode ser a economía, o medio ambiente, a medicina, a minería de datos, o recoñecemento de patróns, a intelixencia artificial, etc., é enorme.

No enfoque tradicional de cluster os elementos só poden ir a un cluster, pero esa división en determinados contextos pode ser excesiva porque pode darse o caso no que un punto é equidistante de dous centroides. Neste traballo realízase cluster *soft* de series temporais, o cal presenta a vantaxe fundamental de non facer unha partición no sentido estrito dos datos ao permitir solapamento entre

grupos, é dicir, que unha serie temporal pertenza a máis dun grupo, o que pode ser de particular interese en moitas aplicacións onde non está tan nítida a partición entre grupos.

Neste traballo faise unha revisión teórica sobre o cluster de series, incidindo nun novo método proposto por Lafuente (2017), así como unha parte de avaliación da conduta de diferentes criterios mediante probas de simulación. A estrutura xeral do traballo é a seguinte. No Capítulo 1 preséntanse uns resultados preliminares sobre a análise espectral de series de tempo. No Capítulo 2 exponse a problemática da análise cluster de series temporais así como algunhas medidas de disimilitude diferenciadas por categorías: procedementos libres do modelo e baseados no modelo; tamén se mostran as diferencias entre o cluster *hard* e o cluster *soft*. O capítulo 3 céntrase no cluster *soft* baseado en modelos mixtos coa exposición do algoritmo EM, o cal é a chave para o seguinte Capítulo 4, no cal se fai un estudo de simulación para comparar este algoritmo con outras funcións e ver o seu comportamento. E finalízase este TFM cunhas breves conclusións e futuras liñas de investigación no Capítulo 5.

Obxectivos principais

Os obxectivos que se perseguen neste traballo son os seguintes:

- Revisión, non exhaustiva pero completa, de diferentes técnicas para desenvolver análise cluster *soft* de series temporais baseadas en diferentes camiños: o dominio temporal e o dominio espectral.
- Establecer con claridade as diferenzas substanciais entre o enfoque *hard* e o enfoque *soft* en clustering, enfatizando as vantaxes do enfoque *soft*.
- Presentar un camiño novidoso para desenvolver cluster *soft* de series temporais baseado nunha técnica que se desenvolve no dominio espectral e que utiliza o algoritmo EM.
- Análise comparativa baseada en simulacións de Montecarlo entre diferentes procedementos con especial énfase no algoritmo EM.

Capítulo 1

Resultados preliminares

Neste capítulo expóñense algúns resultados teóricos, xa coñecidos, que resultan de interese para o desenvolvemento deste traballo. En concreto trátase da análise espectral de series temporais. A análise espectral é unha forma alternativa de tratar a análise de series temporais, en lugar de traballar no dominio do tempo trabállase no dominio das frecuencias. Esencialmente serve para descubrir as eventuais periodicidades ocultas que pode ter unha serie de datos. Por exemplo con datos de temperatura existe unha periodicidade anual e trimestral a causa das variacións climáticas, ou con datos económicos dun produto do Nadal que vai ter un ciclo cada doce meses.

1.1. Análise espectral de series de tempo

A representación espectral dun proceso estacionario $X = \{X(t), t \in \mathbb{Z}\}$ esencialmente descompón X nunha suma de compoñentes sinusoidais con coeficientes aleatorios e incorrelados con distintas frecuencias. Estas sumas van a ter máis peso de acordo ás frecuencias máis importantes, é dicir, ten en conta os ciclos. É importante poder chegar a coñecer para cada frecuencia a súa correspondente sinusoidal; haberá frecuencias nas que o ciclo sexa máis potente que noutras polo que é interesante saber cada período substancial na serie que se repite. Isto é conveniente para modelizar unha serie en termos de variacións periódicas regulares que subxacen.

A descomposición espectral é, no ámbito das series de tempo, un concepto análogo á representación de Fourier das funcións determinísticas. A análise de procesos estacionarios mediante a súa representación espectral denomínase habitualmente análise no dominio de frecuencias ou análise espectral. Resulta equivalente á análise no dominio de tempo baseado na función de autocovarianzas, pero proporciona unha forma diferente de analizar os procesos que pode resultar máis interesante e útil en algunhas aplicacións (Priestley, 1989; Shumway, 2006).

A continuación descríbense moi brevemente algúns aspectos esenciais da teoría espectral de procesos estacionarios, os cales serán de utilidade para o desenvolvemento deste traballo. A onda sinusoidal é da forma:

$$Acos(2\pi\omega t + \phi)$$

sendo A a amplitude, é dicir o alto das ondas, o impacto en termos de escala; ω a frecuencia, que é a inversa do período e vai indicar cales son os valores da frecuencia máis relevantes e ϕ indica a fase inicial da oscilación. Hai moitas sinusoidais diferentes pero o que interesa coñecer é a que ten maior peso, é dicir, a que domina.

Exemplo 1.1

Sendo $x_t = Acos(2\pi\omega t + \phi) + \omega_t$, amósase aquí un exemplo onde os parámetros A e ϕ son descoñecidos:

$$Acos(2\pi\omega t + \phi) = Acos(\phi)cos(2\pi\omega t) - Asin(\phi)sin(2\pi\omega t) = \beta_1 cos(2\pi\omega t) + \beta_2 sin(2\pi\omega t)$$

onde $\beta_1 = Acos(\phi)$ e $\beta_2 = -Asin(\phi)$. Se por exemplo se conta cunha frecuencia $\omega = 1/50$, o modelo pode ser escrito como unha regresión:

$$x_t = \beta_1 cos(2\pi t/50) + \beta_2 sin(2\pi t/50) + w_t$$

Unha vez que se consegue despexar β_1 e β_2 chégase a ter unha idea do que está pasando por detrás do ruído. Trataríase como un problema de regresión facendo un axuste para obter $\hat{\beta}_1$ e $\hat{\beta}_2$. A continuación na Figura 1.1 móstrase unha serie periódica simulada aplicando diferentes ruídos.

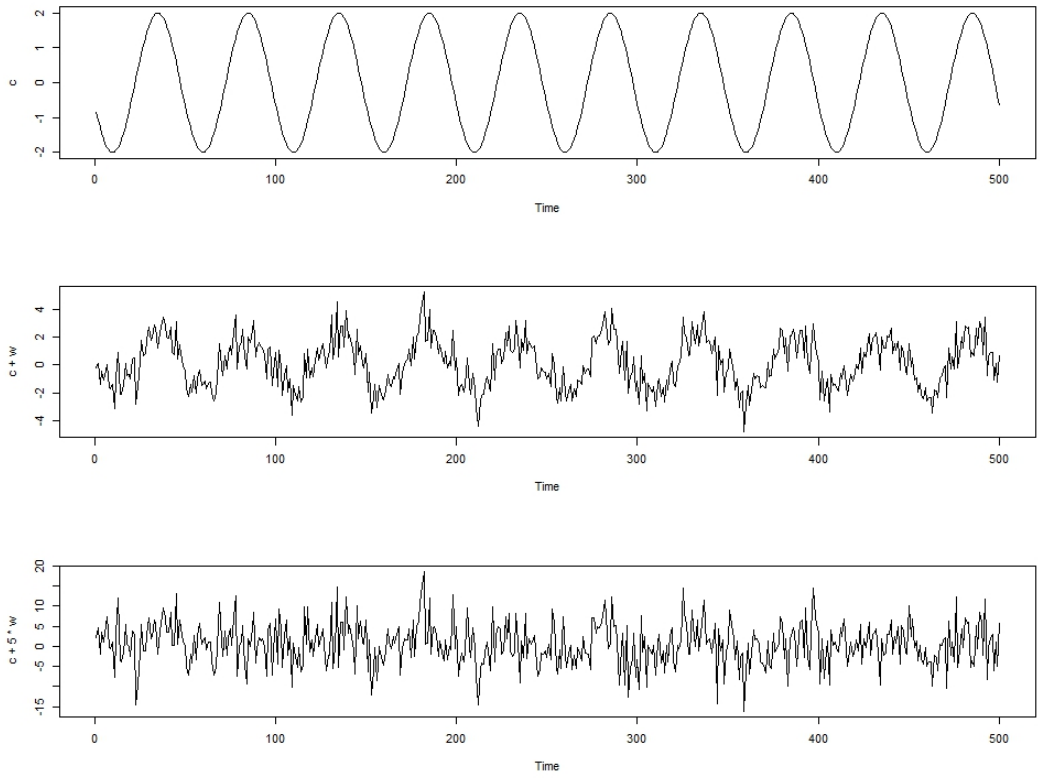

Figura 1.1: Onda coseno con 50 puntos de período (parte superior) en comparación coa onda coseno contaminada con ruído branco gaussiano, $\sigma_w = 1$ (parte central) e $\sigma_w = 5$ (parte inferior)

Como se amosa nos paneis inferiores da Figura 1.1, unha vez que se suman os ruídos escurécese a serie. O grao en que esta queda oculta depende da súa amplitude e do tamaño de σ_w; canto máis grande é esta relación da amplitude e σ_w, máis fácil é detectar a serie. Como se aprecia no panel central da figura, a serie é facilmente discernible, mentres que no panel inferior con un $\sigma_w = 5$ queda oculta. Polo xeral a serie queda oculta polo ruído (Shumway, 2006).

1.1.1. A densidade espectral

Sexa $X = \{X(t), t \in \mathbb{Z}\}$ un proceso de media cero e con función de autocovarianzas $\gamma(\cdot)$ absolutamente sumable, é dicir:

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty.$$

Defínese a densidade espectral do proceso $X = \{X(t), t \in \mathbb{Z}\}$ como a función $f(\cdot)$ dada por:

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h) e^{-i\lambda h}, -\infty < \lambda < \infty.$$

A sumabilidade de $|\gamma(\cdot)|$ garante que a serie anterior converxe absolutamente. Ademais, posto que as funcións $cos(\cdot)$ e $sen(\cdot)$ teñen ámbalas dúas período 2π, a función de densidade espectral é periódica de igual período, polo que é suficiente definila no intervalo $(-\pi, \pi]$. En particular, a función de densidade espectral verifica as seguintes propiedades:

- $f(\cdot)$ é par, é dicir, $f(\lambda) = f(-\lambda)$ para todo $\lambda \in (-\pi, \pi]$.
- $f(\lambda) \geq 0$ para todo $\lambda \in (-\pi, \pi]$.
- A función de autocovarianzas do proceso X_t pode expresarse como:

$$\gamma(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda) d\lambda = \int_{-\pi}^{\pi} cos(h\lambda) f(\lambda) d\lambda, \forall h \in \mathbb{Z}$$

Con todo, hai que ter en conta que non toda función de autocovarianzas ten asociada unha densidade espectral. En xeral, existirá unha función F en $(-\pi, \pi]$, continua á dereita, non decrecente e non acotada, con $F(-\pi) = 0$, tal que:

$$\gamma(h) = \int_{-\pi}^{\pi} e^{ih\lambda} dF(\lambda), \forall h \in \mathbb{Z}$$

Así definida, F é a función de distribución espectral de $\gamma(\cdot)$. Se $F(\lambda)$ pode expresarse como $F(\lambda) = \int_{-\pi}^{\lambda} f(x)dx$ dirase que a serie de tempo ten espectro continuo e que f é a súa densidade espectral. Se F é unha distribución discreta, dirase que a serie ten espectro discreto.

En xeral, pódese demostrar que calquera proceso estacionario é o resultado da superposición dunha cantidade infinita de compoñentes sinusoidais:

$$X(t) = \int_{(-\pi,\pi]} e^{it\lambda} dZ(\lambda),$$

onde $\{Z(\lambda), -\pi < \lambda \leq \pi\}$ é un proceso de valores complexos con incrementos incorrelados. A representación anterior dun proceso estacionario de media cero $X = \{X(t), t \in \mathbb{Z}\}$ coñécese como a representación espectral do proceso e é comparable á representación espectral da función de autocovarianzas $\gamma(\cdot)$. Como consecuencia desta expresión, pode deducirse que un salto na función de distribución espectral (ou, equivalentemente, un pico na densidade espectral) nunha frecuencia $\pm\omega$ indica a presencia na serie de tempo dun compoñente sinusoidal de frecuencia ω (e período $2\pi/\omega$) (Shumway, 2006).

1.1.2. O periodograma

Resulta útil obter aproximacións tanto da función de autocovarianzas como do espectro dunha serie. Dada $\mathbf{X}_n = (X_1, ..., X_n)^t$ unha realización parcial dun proceso estacionario de media cero, $X = \{X(t), t \in \mathbb{Z}\}$, a función de autocovarianzas da mostra pode utilizarse como unha estimación de $\gamma(\cdot)$, mentras que o periodograma $I_n(\cdot)$ resulta ser o análogo mostral da densidade espectral $f(\cdot)$. Sexa $\mathbf{X}_n = (X_1, X_2, ..., X_n)^t \in C^n$. E sexa $\lambda_k = \frac{2\pi k}{n}$, onde k recorre os enteiros entre $-N$ e N, con $N = [\frac{n-1}{2}]$. É dicir:

$$\lambda_k = \frac{2\pi k}{n}, k = -\left[\frac{n-1}{2}\right], ..., \left[\frac{n-1}{2}\right].$$

Os valores λ_k reciben o nome de frecuencias de Fourier asociadas ao tamaño mostral n. Desta forma, os vectores:

$$\mathbf{e}_k = \frac{1}{\sqrt{n}}\left(e^{i\lambda_k}, e^{2i\lambda_k}, ..., e^{ni\lambda_k}\right), k = -\left[\frac{n-1}{2}\right], ..., \left[\frac{n-1}{2}\right]$$

forman unha base en C^n, de xeito que o vector $\mathbf{x} \in C^n$ pode expresarse como suma de n compoñentes:

$$x = \sum_{k=-[\frac{n-1}{2}]}^{[\frac{n-1}{2}]} a_k e_k,$$

onde

$$a_k = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} X_t e^{-it\lambda_k}.$$

A secuencia a_k coñécese como a transformada finita de Fourier do proceso X. As súas adecuadas propiedades teóricas (proporciona variables practicamente incorreladas) e a rapidez coa que pode calcularse mediante calquera dos algoritmos da transformada de Fourier, fan que desempeñe un papel fundamental na análise de series de tempo. Defínese o periodograma de $\mathbf{X}_n = (X_1, ..., X_n)^t$ como:

$$I_n(\lambda) = \frac{1}{2\pi n}\left|\sum_{t=1}^{n} X_t exp(-i\lambda t)\right|^2, \lambda \in [-\pi, \pi].$$

Mediante cálculos pódese demostrar o seguinte resultado:

Proposición 1.1.1 (Priestley (1989)) *Dada $\boldsymbol{X}_n = (X_1, ..., X_n)^t$ unha realización dun proceso estacionario de media cero $X = \{X(t), t \in \mathbb{Z}\}$, e λ_k unha das frecuencias de Fourier, $\lambda_k = \frac{2\pi k}{n}$, en $(-\pi, \pi], \lambda_k \neq 0$, entón:*

$$I_n(\lambda_k) = \frac{1}{2\pi} \sum_{|h|<n} \hat{\gamma}(h) e^{-ih\lambda_k}$$

onde $\hat{\gamma}(h)$ é a función de autocovarianzas mostral asociada a $\boldsymbol{X}_n$.

A comparación desta expresión coa definición da densidade espectral suxire utilizar o periodograma $I_n(\lambda)$ como un estimador natural de $f(\lambda)$. Con todo, estudos detallados das súas propiedades revelan que non se trata dun estimador consistente. A continuación expóñense algúns resultados neste sentido.

Teorema 1.1.1 (Priestley (1989)) *Sexa $X = \{X(t), t \in \mathbb{Z}\}$ unha serie de variables aleatorias independentes con cumulante de cuarta orde finita k_4, entón:*

$$cov(I_n(\lambda_1), I_n(\lambda_2)) = \frac{k_4}{4\pi^2 n} + \frac{\pi^4 x}{2\pi n}\{F_n(\lambda_1+\lambda_2) + F_n(\lambda_1+\lambda_2)\}$$

onde $F_n(\lambda)$ é o núcleo de Fejer dado por:

$$F_n(\lambda) = \frac{1}{2\pi n}\frac{sen^2(\frac{N\lambda}{2})}{sen^2(\frac{\lambda}{2})}.$$

Como consecuencia do teorema anterior dedúcese que, tomando $\lambda_1 = \lambda_2 = \lambda$:

$$Var(I_n(\lambda)) = \begin{cases} \frac{1}{4\pi^2}\left(\sigma_x^4 + \frac{k_4}{n}\right) + O(n^{-2}) & \lambda \neq 0, \pm\pi \\ \frac{1}{4\pi^2}\left(2\sigma_x^4 + \frac{k_4}{n}\right) & \lambda = 0, \pm\pi \end{cases}$$

Deste xeito, si X é un proceso normal (co cal $k_4 = 0$ e λ_k é unha das frecuencias de Fourier), obtense:

$$Var(I_n(\lambda_k)) = \begin{cases} \frac{1}{4\pi^2}\sigma_x^4 & \lambda \neq 0, \pm\pi \\ \frac{1}{4\pi^2}2\sigma_x^4 & \lambda = 0, \pm\pi \end{cases}$$

Baixo estas mesmas hipóteses, para $\lambda_1 \neq \pm\lambda_2$:

$$Cov(I_n(\lambda_1), I_n(\lambda_2)) = \begin{cases} 0 & \text{se } X \text{ é normal e } \lambda_1, \lambda_2 \text{ son múltiplos de } \frac{2\pi}{n} \\ O(n^{-2}) & \text{se } X \text{ é normal e } |\lambda_1 \pm \lambda_2| >> \frac{2\pi}{n} \\ O(n^{-1}) & \text{se } X \text{ é non normal e } |\lambda_1 \pm \lambda_2| >> \frac{2\pi}{n}, \text{ ou } \lambda_1, \lambda_2 \text{ son múltiplos de } \frac{2\pi}{n} \end{cases}$$

É dicir, incluso para procesos non normais, as ordenadas do periodograma son asintoticamente incorreladas se λ_1, λ_2 son múltiplos de $\frac{2\pi}{n}$ ou están suficientemente espaciadas.

O periodograma, sendo o estimador da densidade espectral, non conta cunhas propiedades moi boas xa que é asintoticamente insesgado pero inconsistente. Ademáis para distintas frecuencias, os valores do periodograma son asintoticamente independentes, o que explica a apariencia ruidosa deste.

Exemplo 1.2

Tendo presente o modelo do exemplo anterior:

$$Acos(2\pi\omega t + \phi) = Acos(\phi)cos(2\pi\omega t) - Asin(\phi)sin(2\pi\omega t) = \beta_1 cos(2\pi\omega t) + \beta_2 sin(2\pi\omega t)$$

onde $\beta_1 = Acos(\phi)$ e $\beta_2 = -Asin(\phi)$. As frecuencias de Fourier $\omega_j = \frac{j}{n}$ onde j indica o número de ciclo para todos os datos e n o período, axudan para a obtención do periodograma. Para cada frecuencia elévanse ao cadrado os coeficientes β_1 e β_2 e súmanse, así obtense unha medida apropiada de cal é o peso de unha determinada frecuencia na serie de tempo, é unha medida de correlación da serie de tempo e esa sinusoidal para unha frecuencia (Shumway, 2006).

Na Figura 1.2 amósase un exemplo con distintas frecuencias e a suma delas.

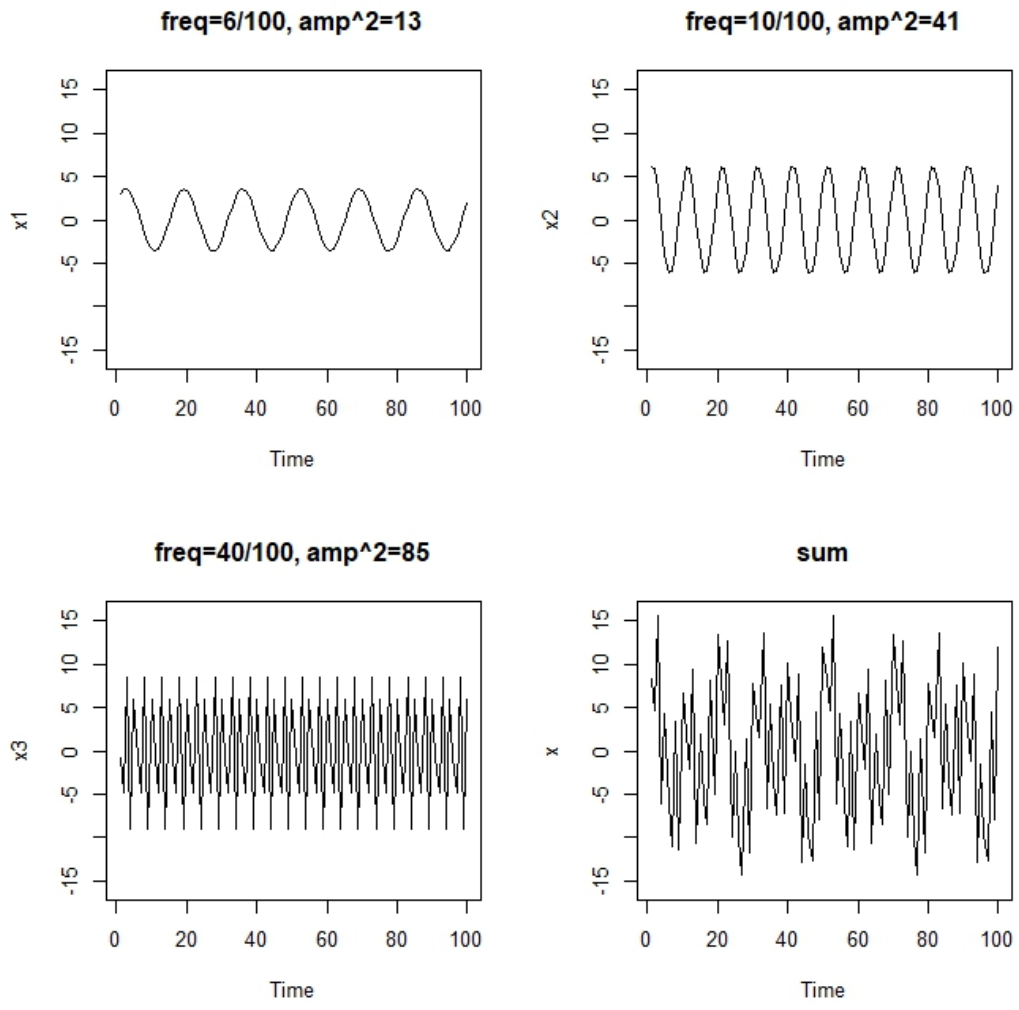

Figura 1.2: Compoñentes periódicas e a súa suma

A continuación realízase o gráfico do periodograma para todas as frecuencias, Figura 1.3, no que se pode ver onde alcanza os valores máis potentes. No momento no que hai unha forte correlación entre esa sinusoidal e a serie temporal indica a importancia desa frecuencia para a serie.

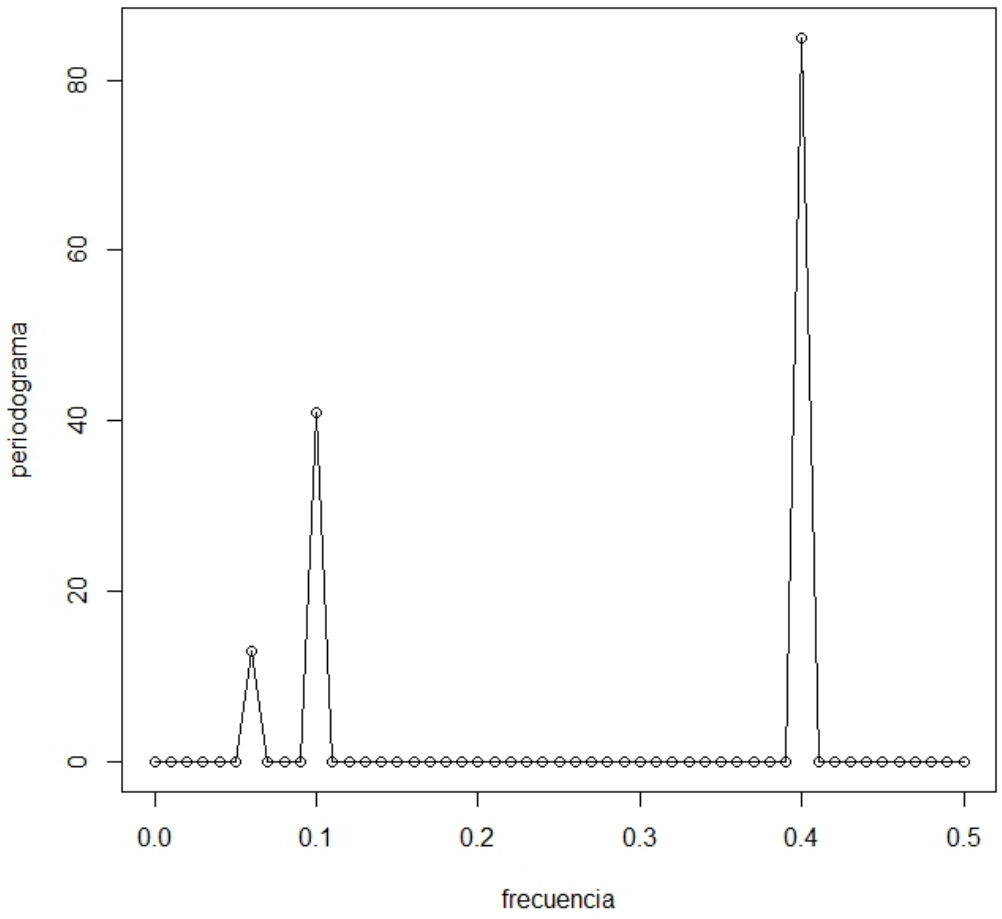

Figura 1.3: Periodograma

Na Figura 1.3 pódese observar como a frecuencia que domina é a do punto 0.4 pois é a máis potente do periodograma, seguida pola frecuencia en 0.1 e pola frecuencia en 0.07.

A partir da serie de tempo trátase de descubrir como un problema de regresión quen son os coeficientes β_1 e β_2 que fan que esa sinusoidal con frecuencia $1/n, 2/n...$ sexa a mellor.

O problema do periodograma é que é altamente variable polo que hai que suavizalo, e isto da lugar á densidade espectral, que vai a estar relacionada coas autocovarianzas e vai a indicar onde hai maior peso. Nos puntos onde máis alto sexa o periodograma nunha frecuencia indican que hai máis impacto na varianza da serie.

Exemplo 1.3

Aquí amósanse uns exemplos de cálculo de periodograma, para unhas series simuladas $AR(1), p = 0{,}9$; $AR(1), p = -0{,}9$; $AR(1), p = 0{,}2$; $AR(1), p = -0{,}2$ e $AR(1), p = 0$.

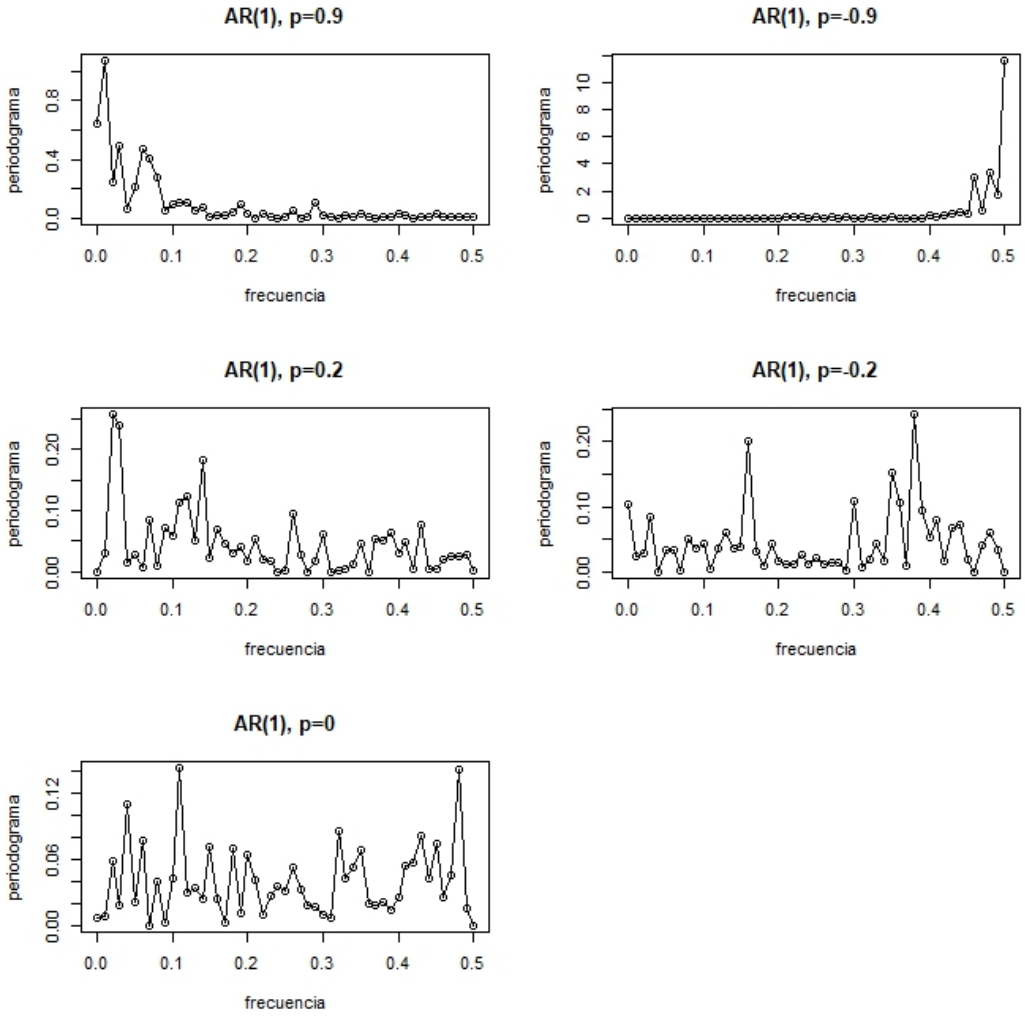

Figura 1.4: Exemplos de periodograma

As frecuencias que dominan no periodograma da serie $AR(1), p = 0{,}9$ sitúanse á esquerda, destacando a frecuencia 0.02, seguida polas frecuencias 0.04, 0.07 e 0.08; pola contra, na serie $AR(1), p = -0{,}9$ as frecuencias que dominan están á dereita, destacando na frecuencia 0.5 como a máis potente. Na serie $AR(1), p = 0{,}2$ as frecuencias que dominan son 0.03 e 0.04, e na serie $AR(1), p = -0{,}2$ dominan as frecuencias 0.38, 0.16. Na última das series simuladas $AR(1), p = 0$ as frecuencias que dominan no periodograma serían a 0.12 e 0.48.

1.1.3. Teoría espectral de procesos lineais

A análise espectral resulta útil no estudo de procesos lineais. A continuación abórdase a transmisión de procesos estocásticos a través de filtros lineais, co fin de mostrar a forma que toma a densidade espectral dun proceso lineal calquera.

Un proceso $X = \{X(t), t \in \mathbb{Z}\}$ é a saída dun filtro lineal invariante $\psi = \{\psi_j, j = 0, \pm 1, ...\}$ aplicado a un proceso de entrada $Z = \{Z(t), t \in \mathbb{Z}\}$ se:

$$X(t) = \sum_{j=-\infty}^{\infty} \psi_j Z(t-j), t = 0, \pm 1, ...$$

Dise entón que o proceso $X = \{X(t), t \in \mathbb{Z}\}$ é un proceso lineal.

Proposición 1.1.2 (Priestley (1989)) *Sexa $Z = \{Z(t), t \in \mathbb{Z}\}$ un proceso estacionario de media cero e densidade espectral $f_Z(\lambda)$. Sexa $\psi = \{\psi_j, j = 0, \pm 1, ...\}$ un filtro lineal invariante con $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$. Entón o proceso lineal*

$$X(t) = \sum_{j=-\infty}^{\infty} \psi_j Z(t-j)$$

é un proceso estacionario de media cero e densidade espectral

$$f(\lambda) = |\psi(e^{-i\lambda})|^2 f_Z(\lambda) = \psi(e^{-i\lambda})\psi(e^{i\lambda}) f_Z(\lambda),$$

onde $\psi(e^{-i\lambda}) = \sum_{j=-\infty}^{\infty} \psi_j e^{-ij\lambda}$. A función $\psi(e^{-i\cdot})$ denomínase función de transferencia do filtro, e $|\psi(e^{-i\cdot})|^2$ potencia da función de transferencia.

Do resultado anterior dedúcese que se $X = \{X(t), t \in \mathbb{Z}\}$ é un proceso lineal Gaussiano dado por

$$X(t) = \sum_{j=-\infty}^{\infty} \psi_j Z(t-j),$$

con $\{Z(j), j \in \mathbb{Z}\}$ unha secuencia de variables aleatorias independentes e identicamente distribuídas con distribución $N(0, \sigma_Z^2)$, a súa densidade espectral é necesariamente da forma

$$f(\lambda) = |\psi(\lambda)|^2 \frac{\sigma_Z^2}{2\pi},$$

con

$$\psi(\lambda) = \sum_{j=-\infty}^{\infty} \psi_j e^{-ij\lambda}.$$

Do mesmo xeito, o seguinte teorema establece a expresión asintótica do periodograma dun proceso

lineal.

Teorema 1.1.2 (Priestley (1989)) *Sexa* $X = \{X(t), t \in \mathbb{Z}\}$ *un proceso lineal xeral dado por*

$$X(t) = \sum_{j=-\infty}^{\infty} \psi_j Z(t-j),$$

sendo $Z = \{Z(t), t \in \mathbb{Z}\}$ *un proceso estacionario de variables aleatorias independentes con* $E(Z(t)) = 0, E(Z^2(t)) = \sigma_Z^2, E(Z^4(t)) < \infty$ *e* $\sum_{j=-\infty}^{\infty} |\psi_j||j|^\alpha < \infty, \alpha > 0$. *Entón tense que:*

$$I_n(\lambda) = f(\lambda)\frac{2\pi}{\sigma_Z^2} I_{n,Z}(\lambda) + R_n(\lambda),$$

sendo $I_{n,Z}(\lambda)$ *o periodograma asociado a* $(Z_1, ..., Z_n)^t$, *e onde* $E\left\{|R_n(\lambda)|^2\right\} = O(n^{-2\alpha})$ *uniformemente en* λ.

A expresión anterior permite obter unha expresión asintótica de $I_n(\lambda)$ directamente a partir dos resultados coñecidos para $I_{n,Z}(\lambda)$. En particular, para un proceso lineal Gaussiano, con $Z(t)$ i.i.d $N(0, \sigma_Z^2)$, tense para $\lambda_k = \frac{2\pi k}{n}$ que:

$$I_{n,Z}(\lambda_k) = \begin{cases} \frac{1}{4\pi}\sigma_Z^2 X_2^2 & k \neq 0, \frac{n}{2}, n \text{ par} \\ \\ \frac{1}{4\pi} 2\sigma_Z^2 X_1^2 & k = 0, \frac{n}{2} \end{cases}$$

A partir das expresións anteriores obtense que:

$$I_{n,Z}(\lambda_k) = \begin{cases} \frac{1}{2} f(\lambda_k) X_2^2 + R_{n,k} & k \neq 0, \frac{n}{2}, n \text{ par} \\ \\ f(\lambda_k) X_1^2 + R_{n,k} & k = 0, \frac{n}{2} \end{cases}$$

De modo que, se se ignora $R_{n,k}$, sobre as frecuencias de Fourier tense que:

$$E(I_n(\lambda_k)) = f(\lambda_k), \forall k = -N, ..., N$$

$$Var(I_n(\lambda_k)) = \begin{cases} f^2(\lambda_k) & k \neq 0, \frac{n}{2} \\ \\ 2f^2(\lambda_k) & k = 0, \frac{n}{2} \end{cases}$$

Posto que $Var(I_n)$ non converxe a cero, $I_n(\cdot)$ non é un estimador consistente de $f(\cdot)$.

De igual forma, pode establecerse a covarianza asintótica para as ordenadas do periodograma sobre as frecuencias de Fourier:

Teorema 1.1.3 (Priestley (1989)) *Sexa $X = \{X(t), t \in \mathbb{Z}\}$ un proceso lineal xeral coma no Teorema 1.1.2. Entón:*

$$Cov(I_n(\lambda_1), I_n(\lambda_2)) = \left\{ \frac{e}{n} + \frac{2\pi}{n}[F_n(\lambda_1 - \lambda_2)] \right\} f(\lambda_1) f(\lambda_2) + O(n^{-\alpha})$$

onde $e = \frac{k_4}{\sigma_Z^4} = E(Z_t^4) - 3, Fn$ é o núcleo de Fejer e o termo restante é de $O(n^{-\alpha})$ uniformemente en λ_1, λ_2.

En particular, se X é un proceso lineal normal as ordenadas do periodograma sobre as frecuencias de Fourier satisfan

$$cov(I_n(\lambda_t), I_n(\lambda_s)) = f(\lambda_t) f(\lambda_s) 1_{\{\lambda_t = \lambda_s\}} + O(n^{-2}).$$

Capítulo 2

Cluster de series temporais

2.1. Introdución

A análise cluster é unha ferramenta da análise de datos que agrupa elementos en conxuntos homoxéneos en función das similitudes entre eles. En particular, a análise cluster de series temporais ten como obxectivo particionar un conxunto de series de tempo en diferentes grupos ou clusters. Hai que ter en conta que a clasificación para asignar as series a cada grupo basease na similitude que teñen as series entre si, así as series no mesmo grupo serán máis similares entre si que as series dos diferentes grupos.

O agrupamento de series temporais é un problema central en moitos campos de aplicación, hoxe en día é unha área de investigación activa nunha ampla gama de campos como a economía, a medicina, a enxeñería ou a física entre outros. A análise cluster de series temporais realízase en moitas aplicacións reais como, por exemplo, determinar produtos con similares patróns de venda, identificar países con similar crecemento da poboación ou con similar temperatura, etc. Como expón Lafuente (2017) este tipo de problemas xorden dun xeito natural, polo que o crecente interese por este tema deu lugar a un gran número de contribucións nas últimas décadas, como poden ser: clasificación das series de produción industrial (Piccolo, 1990), comparación de datos sismolóxicos como no caso clásico de distinguir entre o sismo e as formas de onda de explosión nuclear (Kakizawa et al., 1998), cluster de dinámica ecolóxica (Li et al., 2001), comparación das series temporais hidrolóxicas diarias (Grimaldi, 2004), agrupamento de países industrializados segundo datos históricos de emisións de CO2 (Alonso et al., 2006), detección de comportamentos de resposta inmune semellante á progresión do número de células CD4 en pacientes con virus inmunodeficiente (VIH) (Douzal-Chouakria e Nagabhushan, 2007), identificación de xenes activos durante o proceso de división celular (Douzal-Chouakria et al., 2009), clasificación de datos quimiométricos (D'Urso e Giovanni, 2014), cluster baseado nas emisións diarias

de monóxido de nitróxeno (D'Urso et al., 2015), análise dos patróns de navegación dos usuarios que visitan sitios web de novas (García-Magariños e Vilar, 2015), entre outros.

Un tema moi importante no cluster de series temporais é determinar unha medida adecuada para avaliar a disimilitude entre dúas series temporais. Ao contrario da agrupación convencional en obxectos de datos estáticos as series temporais son intrínsecamente dinámicas, con estruturas de autocorrelación subxacentes e, polo tanto, a busca de semellanza debe rexerse polo comportamento da serie durante os seus períodos de observación.

A selección dunha métrica adecuada ten un papel fundamental pero tamén hai que afrontar outras dificultades na agrupación de series temporais. Por exemplo, moitas aplicacións de clustering na vida real implican un gran número de series moi longas, é dicir, enfróntase ao problema da alta dimensionalidade, de feito, as series temporais observadas conteñen frecuentemente miles de datos, que na análise de cluster tradúcense en miles de variables de clasificación. Os enfoques baseados en características están dirixidos a representar a estrutura dinámica de cada serie por un vector de característica de dimensión inferior, permitindo así unha redución da dimensionalidade e un aforro significativo no tempo de cálculo.

Cómpre mencionar que as medidas de disimilitude son a miúdo adaptadas ao obxectivo da análise cluster, o cal pode ser distinto en diferentes contextos. Así, o obxectivo pode ser discriminar entre os procesos estocásticos que xeran as series, outras veces o obxectivo pode ser de corte xeométrico para tratar de discriminar as formas de series, tamén pode ser discriminar entre as predicións das series. Dependendo cal sexa o obxectivo de clustering determinarase a métrica que interesa, pois a homoxeneidade entre grupos pode ser que as series foran xeradas co mesmo patrón ou que as series dan predicións iguais aínda que sexan de distinto patrón. Por exemplo, hai situacións prácticas nas que o interese real da agrupación baséase nas propiedades das previsións, como no caso de calquera problema de desenvolvemento sostible ou en situacións nas que a preocupación é alcanzar os valores obxectivos nun tempo previo especificado. Os traballos de Alonso et al. (2006) e Vilar et al. (2010) centráronse nesta idea e consideraron unha noción de disimilitude rexida polo desempeño de previsións futuras. En particular, dúas series temporais son semellantes se as súas previsións para un tempo futuro específico están próximas.

Debido a isto existe unha ampla gama de medidas para comparar series temporais e a elección da medida de disimilitude adecuada depende en gran parte da natureza da agrupación, é dicir, na determinación do obxectivo da agrupación. Unha vez que se determina a medida de disimilitude, pódese obter unha matriz inicial de diferencias en pares e entón usar un algoritmo de agrupación convencional para formar grupos de obxectos. De feito, a maioría dos enfoques de agrupamento de series de tempo revisadas por Liao (2005) son variacións dos procedementos xerais (por exemplo, un k-medias ou un clustering xerárquico) que usan unha serie de disimilitudes deseñadas especificamente

para tratar con series temporais.

Na análise de cluster, atendendo á asignación de cluster, considéranse dous paradigmas diferentes dependendo de se se constrúe unha partición *hard* ou unha partición *soft*. Os métodos de clustering tradicionais atribúen cada obxecto de datos exactamente a un cluster, producindo así unha partición *hard* dos datos en subconxuntos, pero esa división pode ser demasiado ríxida en determinados contextos nos que hai obxectos de datos equidistantes de dous ou máis grupos ou en presenza de grupos superpostos. As técnicas de cluster *fuzzy* proporcionan un enfoque máis versátil permitindo a pertenza gradual de obxectos de datos aos clusters. Na partición *soft* resultante os obxectos poden pertencer a varios clusters con niveis de adhesión específicos que indican a cantidade de conformidade na asignación de cada dato aos clusters. Así, o que diferencia o cluster *hard* do cluster *soft* é que no primeiro cada elemento se asigna a un único cluster, xa que non se permiten os solapamentos, mentres que o segundo permite dar un grao de credibilidade ou de asignación a todos os clusters, aínda que en algúns casos pode ter un grao de confianza cero.

2.2. Disimilitude entre series temporais

A determinación dunha medida de disimilitude adecuada entre obxectos é un problema chave na análise de clusters así como un problema particularmente sensible ao tratar datos de series temporais. As diferenzas comunmente utilizadas no cluster convencional ignoran a evolución temporal da serie e poden producir resultados insatisfactorios nun contexto da serie temporal.

Algunhas medidas de disimilitude están implementadas no paquete TSclust (Montero e Vilar, 2014), dispoñible en *http://CRAN.R-project.org/ package=TSclust*, o cal contén medidas de disimilitude ou diferenza usadas máis frecuentemente, incluindo medidas libres do modelo, medidas baseadas en modelos, medidas baseadas na complexidade e medidas baseadas na predición introducida por Vilar et al. (2010). Algunhas destas medidas funcionan no dominio do tempo e outras desenvólvense no dominio da frecuencia. Tamén algunhas funcionan baixo certas condicións de regularidade, mentres que outras son aplicables en contextos máis xerais polo que os usuarios do paquete TSclust deben analizar detidamente que medidas específicas son máis adecuadas para captar semellanza no seu problema de agrupación.

A continuación descríbense algunhas destas medidas segundo dúas categorías distintas: procedementos libres do modelo e procedementos baseados no modelo. Aínda que estas categorías non son exhaustivas, xa que existen outros criterios para definir métricas, son as máis utilizadas.

2.2.1. Procedementos libres do modelo

As distancias do modelo libre inclúen principalmente distancias entre as observacións en bruto e distancias baseadas na comparación de características extraídas da serie temporal orixinal. Un enfoque natural para medir a disimilitude entre X_t e Y_t é substituir os valores observados por un vector de características de dimensión inferior e logo avaliar unha distancia convencional entre os vectores extraídos. Esta aproximación intuitiva presenta algunhas vantaxes, incluindo que non se requiren suposicións sobre os procesos xeradores, a aplicabilidade a series non balanceadas e unha baixa complexidade computacional. As características extraídas pódense obter tanto no dominio do tempo como no dominio da frecuencia.

Unha aproximación para medir a proximidade entre X_t e Y_t é considerar métricas convencionais baseadas na proximidade dos seus valores en determinados puntos do tempo.

Distancia de Minkowski

A distancia de Minkowski de orde q, sendo q un enteiro positivo, tamén chamado distancia $L_q - norm$, é definida por

$$d_{L_q}(X_t, Y_t) = \left(\sum_{t=1}^{T} (X_t - Y_t)^q \right)^{1/q}.$$

A distancia de Minkowski é normalmente usada con $q = 2$ (distancia euclídea) ou $q = 1$ (distancia de Manhattan). Esta métrica é moi sensible ás transformacións como o cambio ou o escalado do tempo (estiramento ou encollemento do eixe tempo). Por outra banda, a noción de proximidade depende da proximidade dos valores observados nos puntos correspondentes do tempo para que as observacións sexan tratadas coma se fosen independentes. En particular, d_{L_q} é invariante ás permutacións ao longo do tempo.

Distancias baseadas na correlación

Outra métrica serían as distancias baseadas na correlación. Un primeiro criterio de disimilitude é considerar o factor de correlación de Pearson entre X_t e Y_t dado por

$$COR(X_t, Y_t) = \frac{\sum_{t=1}^{T} (X_t - \bar{X}_T)(Y_t - \bar{Y}_T)}{\sqrt{\sum_{t=1}^{T} (X_t - \bar{X}_T)^2} \sqrt{\sum_{t=1}^{T} (Y_t - \bar{Y}_T)^2}},$$

sendo $\bar{X}_T$ e $\bar{Y}_T$ os valores medios das realizacións na serie X_t e Y_t respectivamente. Golay et al. (2005) constrúen un algoritmo de k-medias *fuzzy* usando as seguintes distancias baseadas na correlación cruzada:

$$d_{COR,1}(X_T, Y_T) = \sqrt{2(1 - COR(X_T, Y_T))},$$

e

$$d_{COR,2}(X_T, Y_T) = \sqrt{\left(\frac{1 - COR(X_T, Y_T)}{1 + COR(X_T, Y_T)}\right)^{\beta}}, \text{con } \beta \geq 0.$$

Distancias baseadas na autocorrelación

Varios autores como Galeano e Peña (2000), Caiado et al. (2006), D'Urso e Maharaj (2009) consideraron as medidas baseadas nas funcións de autocorrelación estimadas.

Sexan $\hat{p}_{X_T} = (\hat{p}_1, X_T, ..., \hat{p}_L, X_T)^T$ e $\hat{p}_{Y_T} = (\hat{p}_1, Y_T, ..., \hat{p}_L, Y_T)^T$ os vectores estimados de autocorrelación de X_t e Y_t respectivamente, para algún L tal que $\hat{p}_t, X_T \approx 0$ e $\hat{p}_t, Y_T \approx 0$ para $i > L$. Galeano e Peña (2000) definen esta distancia entre X_t e Y_t como segue.

$$d_{ACF}(X_T, Y_T) = \sqrt{(\hat{p}_{X_T} - \hat{p}_{Y_T})^T \Omega (\hat{p}_{X_T} - \hat{p}_{Y_T})},$$

onde Ω é unha matriz de pesos. Algunhas opcións que ten Ω son:

(i) Considerar os pesos uniformes $\Omega = I$. Neste caso d_{ACF} convértese na distancia euclídea entre as funcións de autocorrelación estimadas:

$$d_{ACFU}(X_T, Y_T) = \sqrt{\sum_{t=1}^{L} (\hat{p}_t, X_T - \hat{p}_t, Y_T)^2}.$$

(ii) Considerar pesos xeométricos que se desintegran co retraso de autocorrelación, de xeito que d_{ACF} tome a forma:

$$d_{ACFG}(X_T, Y_T) = \sqrt{\sum_{t=1}^{L} (1-p)^t (\hat{p}_t, X_T - \hat{p}_t, Y_T)^2}, \text{con } 0 < p < 1.$$

As distancias análogas pódense construír considerando as funcións de autocorrelación parcial (PACFs) en vez das ACF. Así a notación d_{ACFU} e d_{ACFG} servirá para denotar a distancia euclídea entre os coeficientes de autocorrelación parcial estimados con pesos uniformes e con pesos xeométricos que se desintegran, respectivamente.

Distancias baseadas na autocovarianza cuantil

Sexa $X_1, ..., X_T$ un tramo observado dun proceso estritamente estacionario $X_t; t \in \mathbb{Z}$. Denótase por F a distribución marxinal de X_t e por $q_\tau = F^{-1}(\tau), \tau \in [0, 1]$, a correspondente función cuantil. Fíxase $l \in \mathbb{Z}$ e un par arbitrario de cuantiles $(\tau, \tau') \in [0, 1]^2$, considerando a covarianza cruzada das funcións

do indicador $I(X_t \leq q_\tau)$ e $I(X_{t+l} \leq q_{\tau'})$ dada por

$$\gamma_l(\tau, \tau') = cov\left\{I(X_t \leq q_\tau), I(X_{t+l} \leq q_{\tau'})\right\} = \mathbb{P}(X_t \leq q_\tau, X_{t+l} \leq q_{\tau'} \leq q_{\tau'}) - \tau\tau'.$$

A función $\gamma_l(\tau, \tau')$, con $(\tau, \tau') \in [0,1]^2$ é a función de autocovarianzas cuantil (QAF) con retardo l que pode verse como unha xeneralización da función de autocovarianza clásica. A función de autocovarianzas cuantil captura a estrutura de dependencia secuencial dunha serie de tempo, pois representa as características de serie relacionadas á distribución conxunta de X_t, X_{t+l} que as autocovarianzas simples non poden detectar.

As autocovarianzas cuantiles proporcionan unha visión máis ampla da dependencia das series que outras características extraídas. Estas abarcan moitas propiedades interesantes, incluíndo robustez frente á inexistencia de momentos, traballar de maneira correcta con distribucións marxinais con colas pesadas, detección de características non lineais e cambios en formas condicionais, entre outros.

Un estimador de $\gamma_l(\tau, \tau')$ pode construírse substituíndo os cuantiles teóricos polos correspondentes cuantiles empíricos $\widehat{q}_\tau$ e $\widehat{q}_{\tau'}$ obtendo a realización observada $X_1, ..., X_T$. Desta forma, o estimador QAF vén dado por

$$\widehat{\gamma}_l(\tau, \tau') = \frac{1}{T-l}\sum_{t=1}^{T-l} I(X_t \leq \widehat{q}_\tau) I(X_{t+l} \leq \widehat{q}_{\tau'}) - \tau\tau',$$

onde os cuantiles empíricos $\widehat{q}_\alpha$, para $0 \leq \alpha \leq 1$ poden verse formalmente como a solución dun problema de minimización dado por

$$\widehat{q}_\alpha = \arg\, min_{q \in \mathbb{R}} \sum_{t=1}^{T} \rho_\alpha(X_t - q),$$

con $\rho_\alpha(x) = x(\alpha - I(x \leq 0))$.

Distancias baseadas en periodogramas

Ata agora, todas as métricas traballan no dominio temporal, pero o enfoque de dominio de frecuencia tamén ofrece unha alternativa interesante para medir a diferenza entre series temporais. A idea chave é avaliar a disimilitude entre as correspondentes representacións espectrais da serie.

Sexan $I_{X_T}(\lambda_k) = T^{-1}|\sum_{t=1}^{T} X_t e^{-i\lambda_k t}|^2$ e $I_{Y_T}(\lambda_k) = T^{-1}|\sum_{t=1}^{T} Y_t e^{-i\lambda_k t}|^2$ os periodogramas de X_T e Y_T, respectivamente, en frecuencias $\lambda_k = 2\pi k/T, k = 1, ..., n$, con $n = [(T-1)/2]$.

Foron analizadas tres medidas de disimilitude baseadas en periodogramas por Caiado et al. (2006).

(i) A distancia euclídea entre as ordenadas de periodogramas:

$$d_P(X_T, Y_T) = \frac{1}{n}\sqrt{\sum_{k=1}^{n}(I_{X_T}(\lambda_k) - I_{Y_T}(\lambda_k))^2}.$$

(ii) Se non interesa a escala de proceso na súa estrutura de correlación, pódense obter mellores resultados empregando a distancia euclídea entre as ordenadas do periodograma normalizadas:

$$d_{NP}(X_T, Y_T) = \frac{1}{n}\sqrt{\sum_{k=1}^{n}(NI_{X_T}(\lambda_k) - NI_{Y_T}(\lambda_k))^2},$$

onde $NI_{X_T}(\lambda_k) = I_{X_T}(\lambda_k)/\hat{\gamma}0, X_T$ e $NI_{Y_T}(\lambda_k) = I_{Y_T}(\lambda_k)/\hat{\gamma}0, Y_T$ sendo $\hat{\gamma}0, X_T$ e $\hat{\gamma}0, Y_T$ as variacións da mostra de X_T e Y_T, respectivamente.

(iii) Ao ser a varianza das ordenadas do periodograma proporcional ao valor do espectro nas frecuencias correspondentes ten sentido usar o logaritmo do periodograma normalizado:

$$d_{LNP}(X_T, Y_T) = \frac{1}{n}\sqrt{\sum_{k=1}^{n}(logNI_{X_T}(\lambda_k) - logNI_{Y_T}(\lambda_k))^2}.$$

Casado de Lucas (2010) considera unha medida de distancia baseada nas versións acumulativas dos periodogramas, é dicir, os periodogramas integrados. Casado de Lucas argumenta que os enfoques baseados no periodograma integrado presentan varias vantaxes respecto das baseadas nos periodogramas. En particular,

- O periodograma é un estimador asimptótico imparcial pero inconsistente da densidade espectral mentres que o periodograma integrado é un estimador consistente da distribución espectral.

- Desde un punto de vista teórico, a distribución espectral sempre existe, pero a densidade espectral só existe baixo distribucións absolutamente continuas.

- O periodograma integrado determina completamente o proceso estocástico.

En Casado de Lucas (2010) propóñense as seguintes distancias baseadas no periodograma integrado, unha normalizada e outra non normalizada.

$$d_{IP}(X_T, Y_T) = \int_{-\pi}^{\pi} |F_{X_T}(\lambda) - F_{Y_T}(\lambda)| d\lambda,$$

onde $F_{X_T}(\lambda_j) = C_{X_T}^{-1}\sum_{i=1}^{j} I_{X_T}(\lambda_i)$ e $F_{Y_T}(\lambda_j) = C_{Y_T}^{-1}\sum_{i=1}^{j} I_{Y_T}(\lambda_i)$, sendo $C_{X_T} = \sum_i I_{X_T}(\lambda_i)$ e $C_{Y_T} = \sum_i I_{Y_T}(\lambda_i)$ para a versión normalizada, e $C_{X_T} = C_{Y_T} = 1$ para a versión non normalizada.

A versión normalizada dá máis peso á forma das curvas mentres que a non normalizada considera a escala. Casado de Lucas suxire usar a versión normalizada cando os gráficos das funcións tenden a cruzarse e os non normalizados cando non.

2.2.2. Procedementos baseados no modelo

As medidas de disimilitude baseadas en modelos asumen que os modelos subxacentes son xerados a partir de estruturas paramétricas específicas. O enfoque principal na literatura é asumir que os procesos xeradores de X_T e Y_T seguen modelos ARIMA invertibles. Nese caso, a idea é poñer un modelo ARIMA en cada serie e medir a diferenza entre os modelos introducidos. O primeiro paso require a estimación da estrutura e dos parámetros dos modelos ARIMA. Suponse que a estrutura é dada ou estimada automaticamente usando, por exemplo, o criterio de información de Akaike (AIC) ou o criterio de información bayesiano de Schawartz (BIC). Os valores dos parámetros comunícanse usando estimadores de mínimos cadrados xeneralizados. A continuación móstranse algunhas das medidas de disimilitude máis relevantes baixo o suposto de modelos ARIMA subxacentes.

Distancia de Piccolo

A medida de disimilitude de Piccolo (1990) está na clase de procesos ARIMA invertibles como a distancia euclídea entre os operadores $AR(\infty)$ que aproximan as correspondentes estruturas ARIMA. Piccolo argumenta que as expansións autorregresivas transmiten toda a información útil sobre a estrutura estocástica deste tipo de procesos (agás os valores iniciais). Se a serie non é estacionaria, realízase a desintegración para facela estacionaria e se a serie ten estacionalidade, entón debe ser eliminada antes dunha análise posterior. Un criterio definido como AIC ou BIC úsase para modelos truncados de $AR(\infty)$ de ordes k_1 e k_2 que aproximan os procesos de xeración de X_T e Y_T, respectivamente. Esta visión permite superar o problema de obter aproximacións ARMA *ad hoc* para cada unha das series sometida a clustering.

Se $\hat{\Pi}_{X_T} = (\hat{\pi}_1, X_T, ..., \hat{\pi}_{k_1}, X_T)^T$ e $\hat{\Pi}_{Y_T} = (\hat{\pi}_1, Y_T, ..., \hat{\pi}_{k_2}, Y_T)^T$ denotan os vectores de $AR(k_1)$ e $AR(K_2)$ para X_T e Y_T , respectivamente, entón a distancia de Piccolo toma a forma

$$d_{PIC}(X_T, Y_T) = \sqrt{\sum_{j=1}^{k} \left(\hat{\pi}'_{j,X_T} - \hat{\pi}'_{j,Y_T}\right)^2},$$

onde $k = max(k_1, k_2), \hat{\pi}'_{j,X_T} = \hat{\pi}'_{j,X_T}$ si $j \leq k_1$ e $\hat{\pi}'_{j,X_T} = 0$ en outro caso, e analogamente $\hat{\pi}'_{j,Y_T} = \hat{\pi}'_{j,Y_T}$ si $j \leq k_2$, e $\hat{\pi}'_{j,Y_T} = 0$ en outro caso.

Ademais de satisfacer as propiedades dunha distancia (non negatividade, simetría e triangularidade), d_{PIC} sempre existe para calquera proceso ARIMA invertible onde $\sum \pi_j, \sum ||\pi_j||$ e $\sum \pi_j^2$ son cantidades ben definidas.

Distancia de Maharaj

Para a clase de procesos ARMA invertibles e estacionarios, Maharaj (1996, 2000) introduciu dúas

medidas de discrepancia baseadas en test de hipóteses para determinar se dúas series temporais teñen ou non procesos de xeración significativamente diferentes. A primeira destas métricas está dada polo test estatístico

$$d_{MAH}(X_T, Y_T) = \sqrt{T}\left(\hat{\Pi}'_{X_T} - \hat{\Pi}'_{Y_T}\right)^T \hat{V}^{-1}(\hat{\Pi}'_{X_T} - \hat{\Pi}'_{Y_T}),$$

onde $\hat{\Pi}'_{X_T}$ e $\hat{\Pi}'_{Y_T}$ son as estimacións dos parámetros X_T e Y_T, respectivamente, sendo o k dado como na distancia de Piccolo, e $\hat{V}$ é un estimador de $V = \sigma^2_{X_T} R^{-1}_{X_T}(k) + \sigma^2_{Y_T} R^{-1}_{Y_T}(k)$, con $\sigma^2_{X_T}$ e $\sigma^2_{Y_T}$ que denotan as variacións dos procesos de ruído branco asociados con X_T e Y_T, e R_{XT} as matrices de covarianza da mostra de ambas series.

Maharaj demostrou que o d_{MAH} está asintóticamente distribuído baixo a hipótese nula de igualdade de procesos xeradores, asumindo $\Pi_{X_T} = \Pi_{Y_T}$. Polo tanto, a disimilitude entre $\hat{\Pi}'_{X_T}$ e $\hat{\Pi}'_{Y_T}$ tamén se pode medir a través do p-valor asociado, considerando

$$d_{MAH,p}(X_T, Y_T) = P(\chi^2_k > d_{MAH}(X_T, Y_T)).$$

Tanto o test estatístico d_{MAH} coma o p-valor asociado $d_{MAH,p}$ satisfán as propiedades de non negatividade e simetría para que calquera delas poida ser usada como medida de disimilitude entre X_T e Y_T. Aínda que d_{MAH} e d_{PIC} avalían a disimilitude entre dúas series comparando as súas aproximacións autorregresivas, hai unha diferenza substancial entre elas: a distancia de Piccolo non ten en conta a varianza dos procesos de ruído branco asociados á serie observada, mentres que o test estatístico de Maharaj implica estas variacións na súa definición. É importante ser consciente deste feito cando se usan estas medidas de disimilitude para levar a cabo a agrupación porque d_{MAH} será detectada pola unidade de escala.

Tamén hai que destacar que se se desenvolve un algoritmo xerárquico a partir da matriz de pares dos p-valores de $d_{MAH,p}$, entón proporciona un criterio de homoxeneidade de agrupamento previamente especificando un nivel de significación α (por exemplo, 5 % ou 1 %). As series con p-valores asociados maiores que α agruparanse xuntas, o que implica que só as series cuxas estruturas dinámicas non sexan significativamente diferentes do nivel α situaranse no mesmo grupo.

As medidas d_{MAH} e $d_{MAH,p}$ proceden dun test de hipóteses deseñado para comparar dúas series temporais independentes. Para superar esta limitación, Maharaj (2000) introduciu un novo procedemento no test que se pode aplicar ás series temporais que non son necesariamente independentes. Neste caso, considérase un modelo agrupado que inclúe colectivamente os modelos introducidos en X_T e Y_T e estímase o vector combinado $2k$ parámetros AR $\Pi = (\Pi_{X_T}, \Pi_{Y_T})$ empregando mínimos cadrados xeneralizados. Supoñendo que os dous modelos están correlacionados no mesmo tempo pero non correlacionados entre as observacións, a proposta no test estatístico ($d_{MAH}ext$) distribúese asintotica-

mente como χ^2 con k graos de liberdade. Coma antes, pódese construír unha medida de disimilitude ($d_{MAH}ext, p$) baseada nos p-valores asociados a este novo test.

2.3. Cluster *hard* versus cluster *soft*

2.3.1. Concepto

Como xa se comentou na introdución deste capítulo, atendendo á asignación de cluster considéranse dous paradigmas diferentes: *hard* e *soft*. O enfoque *hard* é un método máis tradicional de clustering no que os datos se asignan a un único cluster e o enfoque *soft* un método máis versátil no que os datos poden pertencer a varios clusters. Neste segundo enfoque a cada dato asígnaselle un vector de valores de credibilidade de pertenza a cada cluster. Si hai k clusters, pois cada dato ten asignado un vector k dimensional, onde a compoñente j-ésima nos di o grao de credibilidade de que ese dato pertenza ao cluster j-ésimo. Para abordar este problema e aproximarse a unha solución *soft* utilízase o chamado cluster *fuzzy* no que os valores de credibilidade de pertenza aos clusters reciben o nome de *membership*. Eses *membership* existen en calquera procedemento *soft*, no cluster *fuzzy* trátase de detectar eses *membership* optimizando unha función obxectivo que incorpora un parámetro $m \geq 1$ indicando o grao de solapamento que se está disposto a asumir. Valores elevados de m conducen a membership máis baixos incrementando o grao de confusión (*fuzziness*) en tanto que valores baixos de m conducen no límite ($m = 1$) a solucións *hard*.

Algúns autores motivan a adopción da lóxica *fuzzy* na agrupación de series temporais. D'Urso e Maharaj (2009) argumentan que a dinámica dunha serie temporal pode cambiar ao longo do tempo de xeito que poida pertencer a clusters distintos durante diferentes períodos de tempo, é dicir, dun xeito difuso. Aielli e Caporin (2013) motivan un clustering *soft* baseado en modelos mixtos argumentando que se a semellanza está baseada en parámetros dinámicos estimados, entón a estimación de erro xera variabilidade causando grupos superpostos. Aínda que os métodos *hard* recibiron unha maior atención na literatura de clasificación de series temporais, varias contribucións recentes adoptaron o enfoque *fuzzy* combinado con distintos criterios de disimilitude entre series, incluídas as distancias baseadas en funcións de autocorrelación (D'Urso e Maharaj, 2009), características extraídas no dominio da frecuencia como o periodograma normalizado e os coeficientes cepstral (Maharaj e D'Urso, 2011), aproximacións autorregresivas (D'Urso et al., 2013), e estimadores de coeficientes de GARCH (D'Urso et al., 2016).

Neste traballo realízase clasificación *soft* empregando modelos mixtos, os cales utilizan o algoritmo *Expectation Maximization* (EM). O algoritmo EM, que se expón en detalle máis adiante, consiste en executar iterativamente dúas etapas ata que deixa de mellorarse a función obxectivo ou se satisface

unha regla de parada previamente establecida. Rematada a iteración s, na iteracción *(s+1)* da etapa *Expectation* calcula o valor esperado das variables latentes z, que indican a probabilidade que teñen as series de tempo de pertencer a un determinado grupo. Na seguinte etapa de *Maximization* calcúlanse os centros dos clusters e as probabilidades a priori maximizando a log-verosimilitude da etapa anterior; e o algoritmo itera ata lograr a converxencia.

2.3.2. Cluster *hard*: algoritmos k-means e PAM

Na análise cluster existen distintos enfoques ou xeitos de proceder para desenvolver cluster. Os dous máis importantes son os métodos xerárquicos e os métodos de partición ou partitivos.

Nos métodos xerárquicos o obxectivo é estruturar os elementos dun conxunto de forma xerárquica pola súa similitude. As observacións ordénanse en niveis, de forma que os niveis superiores conteñen aos inferiores. Esta estrutura xerárquica adóitase representar en forma de árbore (dendrograma). A estrutura de asociación entre os elementos vai a permitir separar os elementos en grupos homoxéneos. Os algoritmos xerárquicos son de dous tipos:

- Aglomerativos: Parten das observacións individuais e van agrupando casos ata chegar á formación de grupos homoxéneos.
- Divisivos: Parten dun cluster inicial con todas as observacións e van dividindo ata chegar a grupos con unha soa observación.

Nos métodos de partición dispóñense de datos heteroxéneos que se queren agrupar nun número de grupos homoxéneos prefixado consonte a algún criterio, de maneira que: cada elemento pertenza a un e só un dos grupos, todo elemento quede clasificado, e cada grupo sexa internamente homoxéneo. Entre os métodos de partición destacan o algoritmo k-means e o algoritmo PAM.
O algoritmo de k-means realiza catro etapas:

- Seleccionar k puntos como centros dos grupos iniciais: escollendo k observacións ao azar, tomando como centros as k observacións máis afastadas entres si, utilizando unha selección a priori, etc.
- Calcular as distancias euclídeas de cada observación ao centro dos k grupos, e asignar cada elemento ao grupo máis próximo. A asignación realízase secuencialmente e ao introducir un novo elemento nun grupo recalcúlase a nova media do grupo.
- Definir un criterio de homoxeneidade e comprobar si reasignando un a un cada elemento dun grupo a outro mellora o criterio.
- Se non é posible mellorar o criterio de homoxeneidade, termínase o proceso.

Por outra banda, o algoritmo PAM, un procedemento de partición en torno a medoides, é máis rápido e traballa con k-medoides. O algoritmo k-means é sensible á presenza de outliers, pero o algoritmo PAM utiliza medoides en lugar de centroides, isto é, tomar como referencia un obxecto xa existente no cluster (idealmente, o obxecto máis central do cluster). Este algoritmo divide os datos conformados por n obxectos en k grupos, sendo k coñecido de antemán, e as súas etapas son as que seguen:

- Seleccionar arbitrariamente k dos n puntos como o medoide.
- Asociar cada punto restante ao medoide máis próximo.
- Seleccionar aleatoriamente un obxecto non-medoide e calcular o custo total do intercambio
- Intercambiar o medoide seleccionado ao inicio polo seleccionado no paso anterior se mellora a calidade. Noutro caso, desfacer ese intercambio.

2.3.3. Cluster *soft*: versión *fuzzy* dos algoritmos k-means e PAM

Adóptase un enfoque *fuzzy* para ter en conta a incerteza intrínseca (non estocástica) derivada do agrupamento de datos tan complexos como series temporais e para capturar a natureza de conmutación ou deriva dalgunhas series temporais no proceso de agrupación, isto quere dicir que os procesos subxacentes poden mudar co tempo de xeito que as series poden estar nun cluster durante un período de tempo pero axustarse mellor a outros clusters noutros períodos de tempo. A agrupación *fuzzy* permite asignar unha serie temporal a dous ou máis clusters, cun grao de adhesión que representa a incerteza relacionada coa asignación da serie temporal a cada cluster, máis formalmente, constrúe unha matriz en función da adhesión cuxo elemento (i, j) representa o grao de pertenza da i-ésima observación ao j-ésimo cluster.

Debido á dificultade de identificar un límite claro entre os clusters en problemas do mundo real, a agrupación *fuzzy* parece máis atractiva que a clasificación determinista de métodos de agrupación non superposicionais. A aproximación *fuzzy* é preferible á aproximación probabilística, por exemplo, o enfoque de mestura finita fai suposicións de distribución rigorosas en datos dentro de clusters descoñecidos e, pola contra, no método de agrupación *fuzzy* non se debe asumir a priori ningunha forma específica de distribución de datos observados (dentro de cada cluster) para o método proposto.

Ademais, conta con maior sensibilidade na captura dos detalles que caracterizan as series temporais. En moitos casos, xa que a dinámica das series temporais está a deriva ou cambia, os enfoques de agrupación estándar probablemente perdan esta estrutura subxacente. Este método tamén conta con máis adaptación na definición do prototipo da serie temporal. Isto pódese apreciar mellor cando os patróns de tempo observados non difiren demasiado uns dos outros. Neste caso, a definición *fuzzy* dos

clusters permite distinguir as estruturas subxacentes, se é probable que existan no conxunto dado de series temporais.

No ano 1969, Ruspini realizou unha aplicación pioneira do concepto de conxuntos *fuzzy* para a análise de clusters. A teoría deste tipo de clustering desenvolveuse rapidamente e o potencial do clustering *fuzzy* suxeriuse para unha ampla gama de aplicacións (Ruspini, 2019).

Clustering k-means *fuzzy*

O clustering k-means *fuzzy* é unha técnica de optimización dunha función debidamente elixida e é unha xeneralización directa do clustering de k-means. Sexa $X = (x_1, x_2, ..., x_n)$ unha matriz de datos $n \times p$ onde x_i é o vector p dimensional que representa as coordenadas da i-ésima observación, e sexa $U = (u_{ij})$ unha matriz de membros $n \times k$, sendo k o número de clusters e cuxos elementos satisfán as seguintes condicións:

$$u_{ij} \in [0, 1], \quad \forall i, j,$$

$$\sum_{j=1}^{k} u_{ij} = 1 \in [0, 1], \quad \forall i,$$

Un k-tupla $(u_{i1}, ..., u_{ik})$ representa a pertenza á i-ésima observación onde U pode ser interpretado coma o grao de pertenenza da i-ésima observación ao j-ésimo cluster. O clustering *fuzzy* é un percorrido dende o conxunto de matrices de datos ao conxunto de matrices de adhesión, e se a restrición

$$u_{ij} = 1 \text{ ou } 0$$

se engade ás anteriores, a solución resultante redúcese ao habitual particionamento de n observacións para k clusters (clustering *hard*). A técnica máis fácil de implementar do clustering *fuzzy* é a agrupación funcional obxectiva que minimiza a funcionalidade escollida axeitadamente.

O clustering k-means *fuzzy* adopta como criterio funcional o criterio de erro de mínimos cadrados

$$J(U, X, V) = \sum_{i=1}^{n} \sum_{j=1}^{k} (u_{ij})^m d(x_i, v_j)^2$$

onde $V = (v_1, v_2, ..., v_k)$ é un conxunto dos vectores valor típico dos clusters, $d(x_i, v_j) = ||x_i - v_j||$ é unha norma de produto interior arbitraria (normalmente distancia euclídea) e m é a constante de axuste que determina a imprecisión ou (*fuzziness*) da solución.

Clustering PAM *fuzzy*

Neste algoritmo hai dúas cuestións fundamentais, en primeiro lugar, débese determinar o número de clusters dende o principio para realizar as agrupacións, pero nos conxuntos de datos reais este número descoñécese e, en segundo lugar, a aleatoriedade dos valores inicias como puntos centrais produce diferentes clusters en cada paso polo que este algoritmo é sensible aos puntos iniciais, para resolver isto pódese realizar un ensaio e seleccionar os que teñen mellor saída.

A tarefa de agrupación é útil para resumir adecuadamente a información nun conxunto de series temporais. En vez de considerar todas as series temporais pódese analizar series temporais prototípicas, é dicir, series temporais que manteñen as características principais de series de tempo similares clasificadas no mesmo grupo. Para iso, adóptase o enfoque de Particionamento en torno a medoides (PAM) nun marco *fuzzy*, este é o enfoque *fuzzy* C-medoides (FCMD). Con este enfoque os prototipos de cada serie de tempo medoide, son series de tempo realmente observadas e non unha serie temporal virtual, como os centroides cun enfoque *fuzzy* k-means. A posibilidade de obter series temporais representativas non ficticias nos clusters é moi atractiva e útil nunha ampla gama de aplicacións. Isto é moi importante para a interpretación dos grupos seleccionados.

De feito, en moitos problemas de agrupación interesa particularmente unha caracterización dos grupos mediante obxectos típicos ou representativos. Estes son obxectos que representan os distintos aspectos estruturais do conxunto de obxectos que están sendo investigados. Pode haber distintas razóns para buscar obxectos representativos, estes non só proporcionan unha caracterización dos clusters pois utilízanse especialmente cando é máis económico ou cómodo usar un pequeno conxunto de k obxectos en vez do gran volume co que se comeza unha investigación.

Capítulo 3

Cluster *soft* baseado en modelos mixtos

3.1. Introdución

Ademais de enfoques de agrupación *fuzzy*, outros algoritmos de agrupación pertencentes ao dominio da computación *soft* foron propostos e aplicados con éxito nas últimas décadas. Unha alternativa de clasificación *soft* é o enfoque baseado en modelos mixtos que utiliza o algoritmo *Expectation-Maximization* (EM). Unha posible vía realizando un clustering baseado en modelos é considerar que a distribución subxacente ten a forma dunha mestura adecuada de distribucións paramétricas, onde cada compoñente da mestura describe a natureza probabilística dun grupo específico no conxunto de datos. No caso de series temporais este enfoque non é sinxelo debido á alta dimensionalidade dos datos.

Como expón Bouveyron (2014) a análise cluster baseada en modelos mixtos con datos estáticos convértese nunha técnica de referencia, destacando os traballos de McLachlan e Basford (1988), McLachlan e Peel (2000), Banfield e Raftery (1993), Fraley (1998) e Fraley e Raftery (2002). En espazos de alta dimensión os métodos de agrupamento baseados en modelos mostran algunha deficiencia e están sobre-parametrizados polo que se podería reducir a dimensión coa menor perda posible de información. Un dos métodos utilizados para reducir a dimensión é a análise de compoñentes principais (PCA), que se leva a cabo antes de proceder coa tarefa de agrupación.

A continuación, para acadar o terceiro dos obxectivos deste traballo, describirase un novo procedemento de cluster de series temporais baseado en modelos mixtos, enfatizando as súas principais características.

3.2. Un modelo mixto baseado no dominio da frecuencia

Wong e Li (2000) consideran un modelo mixto gaussiano autoregresivo de primeira orde para datos de series temporais, e máis tarde Chen e Maitra (2011) estenden este modelo para incluír información de variables explicativas e consideran series de tempo de autoregresión máis xerais. Os dous procedementos traballan no dominio do tempo e aproveitan a forma razoablemente sinxela (baseada en $p+1$ parámetros) da matriz de covarianza dun AR(p). A pesar diso, a aproximación tradicional de estimar os parámetros por máxima verosimilitude usando o algoritmo *Expectation-Maximization* (EM) esixe un alto custo computacional porque a matriz de varianzas covarianzas que hai que estimar conta con moitos parámetros e ten unha gran dimensión.

Resulta interesante desenvolver métodos alternativos para realizar a agrupación de series de tempo baseadas en modelos mixtos. Nesta liña Lafuente (2017) propón analizar o dominio da frecuencia e considerar a representación asintótica do log-periodograma mediante un modelo de regresión non paramétrico con erros de distribución log-exponencial, supoñendo que as series temporais dentro dun mesmo cluster se caracterizan por unha densidade espectral específica. A estimación do modelo mixto implica aproximacións non paramétricas dos log-periodogramas de cada grupo e estimadores das probabilidades de pertencer aos grupos. Para obter estas estimacións emprégase o algoritmo EM.

A continuación amósase como chegar ao modelo mixto no contexto espectral. Sexa S un conxunto de n realizacións de series temporais estacionarias univariantes con media cero denotadas por $X_t^{(i)} = \left\{X_1^{(i)}, ..., X_{T_i}^{(i)}\right\}$, onde $i = 1, ..., n$. Suponse por simplicidade $T_i = T$, para todo i. Considerar as representacións espectrais correspondentes a través dos log-periodogramas $I_k^{(i)}$, $i = 1, ..., n$, avaliado nas frecuencias de Fourier $\lambda_k, k = 1, ..., M$, con $M = [(T-1)/2]$. Para cada serie temporal a secuencia de log-periodogramas $Y_k^i = log(I_k^i) - C_0$, con $C_0 = -0{,}57721$ sendo a constante de Euler, admite aproximadamente o modelo de regresión non paramétrico dado por

$$Y_k^i = m^i(\lambda_k) + \epsilon_k^i$$

onde $m^i(\cdot) = log(f^i(\cdot))$ denota o logaritmo da densidade espectral para a serie i-ésima, e os erros ϵ_k^i son asintóticamente independentes e identicamente distribuídos con función de densidade de probabilidade $\varphi(\lambda) = exp(\lambda - exp(\lambda))$.

Asumindo a existencia de C grupos homoxéneos para as n series, é dicir, a existencia de C densidades espectrais diferentes, $\mathbf{f} = \left\{f^1(\cdot), ..., f^C(\cdot)\right\}$, entón calquera serie observada de S satisfai

$$Y_k^i = m^c(\lambda_k) + \epsilon_k^i,$$

para $i = 1, ..., n$, $k = 1, ..., M$ e algún $c = 1, ..., C$.

Sexa $\boldsymbol{\pi} = (\pi_1, ..., \pi_C)^t$ o vector das probabilidades a priori de pertenza a cada cluster, é dicir, $\pi_c = \mathrm{P}(\mathbf{X}_t^{(i)} \in \text{grupo } c)$, para todo $i = 1, ..., n$ e $c = 1, ..., C$.

Denótese por $\Theta = \left\{\pi_1, ..., \pi_{C-1}, m^1(\cdot), ..., m^C(\cdot)\right\}$ o conxunto de parámetros e funcións descoñecidos que determinan a estrutura probabilística das n series temporais observadas. Da ecuación anterior conclúese que a función de densidade de probabilidade dos erros, digamos $g(\cdot)$, pode ser escrita como

$$g(\epsilon_k^i/\Theta) = \sum_{c=1}^{C} \pi_c \varphi(Y_k^i - m^c(\lambda_k)),$$

para $i = 1, ..., n$ e $k = 1, ..., M$.

Esta ecuación establece que a densidade dos erros dos modelos de regresión non paramétrica ten a forma dunha mestura finita de distribucións cuxo c-ésimo coeficiente representa a probabilidade de que a serie temporal corresponda ao c-ésimo cluster. Segundo esta ecuación, a verosimilitude do conxunto de parámetros descoñecidos e os log-espectros, Θ, dados os datos, $Y \equiv \left\{(\lambda_k, Y_k^i), k = 1, ..., M, i = 1, ..., n\right\}$ está dada por

$$L(\Theta/Y) = \prod_{i=1}^{n} \prod_{k=1}^{M} \sum_{c=1}^{C} \pi_c \varphi(Y_k^i - m^c(\lambda_k))$$

e a correspondente log-verosimilitude por

$$\mathcal{L}(\Theta/Y) = \log L(\Theta/Y) = \sum_{i=1}^{n} \sum_{k=1}^{M} log \left(\sum_{c=1}^{C} \pi_c \varphi(Y_k^i - m^c(\lambda_k)) \right)$$

Non obstante, os elementos $m^c \in \Theta$ son realmente funcións, o que suxire abordar o problema como un problema de optimización local asumindo que os log-espectros son funcións suaves. Así, as aproximacións non paramétricas de tipo núcleo para $m^c(\cdot)$ poden obterse maximizando a función da log-verosimilitude local en lugar da función da log-verosimilitude. Usando a desigualdade de Jensen para funcións cóncavas, a función de log-verosimilitude local toma a forma

$$\begin{aligned} \ell(\Theta/Y)(\lambda) &= \sum_{i=1}^{n} \sum_{k=1}^{M} log \left(\sum_{c=1}^{C} \pi_c \varphi(Y_k^i - m^c(\lambda)) \right) K_h(\lambda_k - \lambda) \qquad (3.1) \\ &\geq \sum_{i=1}^{n} \sum_{k=1}^{M} \sum_{c=1}^{C} log(\pi_c \varphi(Y_k^i - m^c(\lambda))) K_h(\lambda_k - \lambda), \end{aligned}$$

onde $K_h(\cdot) = \frac{1}{h} K(\frac{\cdot}{h})$ é a función kernel $K(\cdot)$ reescalada polo ancho de banda h.

3.3. Algoritmo EM

A maximización da función de verosimilitude local $\ell(\Theta/Y)(\cdot)$ dada en (3.1) realízase usando o algoritmo *Expectation-Maximization* (EM). Aquí realizaranse axustes locais constantes (media ponderada) en lugar de usar polinomios de orde superior para producir solucións pechadas no paso M do algoritmo EM.

No marco EM, o problema de modelo mixto está formulado como un problema de datos incompletos. Os datos observados considéranse incompletos xa que a cada dato se lle asocia un valor non observado ou unha variable latente, especificando a compoñente de mestura ao que pertence. Para formular o problema en termos de datos completos asígnanse etiquetas $(z_{i1}, ..., z_{iC}), c = 1, ..., C$, á serie i-ésima, para todo $i = 1, ..., n$, onde $z_{ic} = 1$ se a serie temporal pertence ao clúster c e 0 noutro caso. No que segue Z denotará a matriz $n \times C$ cuxa i-ésima fila é o vector $\mathbf{Z}^{(i)} = (z_{i1}, ..., z_{iC})^t$ con $z_{ic} = 1_{\left\{\mathbf{X}_t^{(i)} \in \text{ grupo } c\right\}}$. Así, os datos completos son $\left\{\mathbf{X}_t^{(i)}, \mathbf{Z}^{(i)}\right\}$, e a log-verosimilitude local cos datos completos é da forma

$$\ell(\Theta/Y, Z)(\lambda) = \sum_{i=1}^{n}\sum_{c=1}^{C} z_{ic} \sum_{k=1}^{M} log\left\{\pi_c \varphi(Y_k^i - m^c(\lambda))\right\} K_h(\lambda_k - \lambda).$$

Os valores esperados das etiquetas $\{z_{ic}\}$ condicionadas aos estimadores máis recentes de Θ (estimacións para π e m^c obtidas no paso M) son calculados e actualizados iterativamente no paso de *Expectation* (paso E). A iteración (s + 1)-ésima do procedemento EM detállase a continuación.

Ao final da s-ésima iteración, as estimacións $\Theta_s = \left\{\pi_1^{(s)}, ..., \pi_{C-1}^{(s)}, m^{1(s)}(\cdot), ..., m^{C(s)}(\cdot)\right\}$ están dispoñibles. Os pasos E e M proceden do seguinte xeito.

Paso E. Segundo as estimacións da iteración s, tense

$$z_{ic}^{(s+1)} = \mathbb{E}(z_{ic}/\Theta_s, Y) = \mathbb{P}(X_t^{(i)} \in \text{ grupo } c/\Theta_s, Y),$$

para cada $c = 1, ..., C$ e $i = 1, ..., n$. O enfoque estándar para estimar esta expectativa é usar a regra de Bayes,

$$z_{ic}^{(s+1)} = \frac{\pi_c^{(s)} \prod_{k=1}^{M} \varphi(Y_k^i - m^{c(s)}(\lambda_k))}{\sum_{c'=1}^{C} \pi_{c'}^{(s)} \prod_{k=1}^{M} \varphi(Y_k^i - m^{c'(s)}(\lambda_k))}$$

$$= \frac{\pi_c^{(s)} \prod_{k=1}^{M} exp(Y_k^i - m^{c(s)}(\lambda_k) - exp(Y_k^i - m^{c(s)}(\lambda_k)))}{\sum_{c'=1}^{C} \pi_{c'}^{(s)} \prod_{k=1}^{M} exp(Y_k^i - m^{c'(s)}(\lambda_k) - exp(Y_k^i - m^{c'(s)}(\lambda_k)))}$$

para $i = 1, ..., n$ e $c = 1, ..., C$.

Aínda que esta expresión proporciona unha solución pechada para a estimación de z_{ic}, Lafuente (2017) atopou algúns problemas de corte computacional cando realizou probas con datos simulados.

Estes problemas están intrinsecamente relacionados coas colas pesadas do produto das distribucións exponenciais, o que resulta en valores próximos a cero do numerador de $z_{ic}^{(s+1)}$ para todos os c diferentes do clúster verdadeiro. Deste xeito, se unha serie temporal é equidistante de todos os clusters, entón sempre hai un cluster (o cluster máis próximo) que recibe un valor de pertenza igual a 1. Ademais dunha asignación de *membership* inestable, este comportamento non é desexable no clustering *soft*, onde se podería esperar que os graos de adhesión estean distribuídos uniformemente sobre os grupos.

Proponse unha nova aproximación para estimar $\mathbb{P}(\Theta_s, Y|X_t^{(i)} \in \text{ grupo } c)$. Para cada serie $X_t^{(i)}, i = 1, ..., n$ calcular as estimacións da densidade do núcleo $\tilde{\varphi}_c^i$ baseada nos erros $Y_k^i - m^c(\lambda_k)$, para $c = 1, ..., C$. Entón, defínese

$$\mathbb{P}(\Theta_s, Y|X_t^{(i)} \in \text{ grupo } c) = P_{ic} = \frac{1/KLD(\varphi, \tilde{\varphi}_c^i)}{\sum_{c'=1}^{C} 1/KLD(\varphi, \tilde{\varphi}_{c'}^i)},$$

onde $KLD(\cdot)$ denota a diverxencia de Kullback-Leibler entre dúas distribucións de probabilidade (Kullback e Leibler, 1951). En realidade, KLD non é unha métrica. É sempre non negativo e é igual a cero se e só se as dúas distribucións son idénticas, pero non é simétrica e tampouco satisfai a desigualdade triangular. Non obstante, este feito non é importante porque a principal preocupación é medir a información perdida cando as densidades estimadas $\tilde{\varphi}_c^i$ se utilizan para aproximar a densidade de referencia φ. Noutras palabras, os roles que xogan $\tilde{\varphi}_c^i$ e φ son diferentes. De todas formas, podería usarse calquera outra distancia entre distribucións. Por último, ter en conta que a diverxencia de Kullback-Leibler toma valores entre 0 e ∞ adoptando o criterio de fixación $P_{ic} = 1$ se $KLD(\varphi, \tilde{\varphi}_c^i) = 0$ e $P_{ic} = 0$ cando $KLD(\varphi, \tilde{\varphi}_c^i) = \infty$. Unha vez calculado o P_{ic}, as probabilidades a posteriori son definidas por

$$z_{ic}^{(s+1)} = \frac{\pi_c P_{ic}}{\sum_{c'=1}^{C} \pi_{c'} P_{ic'}}.$$

Paso M. Este paso proporciona estimacións de parámetros actualizados $\Theta_{(s+1)}$ maximizando a función de log-verosimilitude local completa esperada cos valores das variables latentes $z_{ic}^{(s+1)}$ obtidas no paso E. Selecciónase unha cuadrícula de frecuencias espaciada regularmente para λ, $\lambda \in \{\gamma_1, \gamma_2, ..., \gamma_r\}$, logo a función obxectivo ten a seguinte forma

$$\ell(\Theta/Y, Z)(\lambda) = \sum_{i=1}^{n}\sum_{c=1}^{C} z_{ic}^{(s+1)} \sum_{k=1}^{M} log\left\{\pi_c \varphi(Y_k^i - m^c(\lambda))\right\} K_h(\lambda_k - \lambda)$$

$$= \sum_{i=1}^{n}\sum_{c=1}^{C} z_{ic}^{(s+1)} \left\{ log\pi_c + \sum_{k=1}^{M} log\left\{\pi_c \varphi(Y_k^i - m^c(\lambda))\right\} K_h(\lambda_k - \lambda)\right\}$$

$$= \underbrace{\sum_{i=1}^{n}\sum_{c=1}^{C} z_{ic}^{(s+1)} log\pi_c}_{(A)}$$

$$+ \underbrace{\sum_{i=1}^{n}\sum_{c=1}^{C} z_{ic}^{(s+1)} \sum_{k=1}^{M} exp\left\{Y_k^i - m^c(\lambda) - exp\left\{Y_k^i - m^c(\lambda)\right\}\right\} K_h(\lambda_k - \lambda)}_{(B)},$$

para $\lambda = \gamma_j$, $j = 1, ..., r$

A optimización realízase maximizando os termos A e B por separado. En canto ao termo A, a optimización faise mediante o procedemento multiplicador de Lagrange. O problema de optimización restrinxido é dado por

$$\max_{\pi} \sum_{i=1}^{n}\sum_{c=1}^{C} z_{ic}^{(s+1)} \log \pi_c$$

suxeito a $\sum_{c=1}^{C} \pi_c = 1, \pi_c \geq 0$ para $c = 1, ..., C$, de xeito que a función Lagrangiana toma a forma

$$R(\pi, \beta) = \sum_{i=1}^{n}\sum_{c=1}^{C} z_{ic} log\pi_c + \beta\left(\sum_{c=1}^{C} \pi_c - 1\right),$$

onde β denota o multiplicador de Lagrange descoñecido. Para obter os puntos críticos de $R(\pi, \beta)$, o sistema de ecuacións simultáneas que inclúen as derivadas parciais respecto de π_c e β é igual a cero, debe resolverse

$$\frac{\partial R}{\partial \pi_c} = \frac{1}{\pi_c}\sum_{i=1}^{n} z_{ic}^{(s+1)} + \beta = 0,$$

$$\frac{\partial R}{\partial \beta} = \sum_{c=1}^{C} \pi_c - 1 = 0$$

As solucións danse por $\pi_c^{(s+1)} = -\frac{1}{\beta}\sum_{i=1}^{n} z_{ic}^{(s+1)}$ e $\hat{\beta} = -\frac{1}{\sum_{i=1}^{n}\sum_{c=1}^{C} z_{ic}^{(s+1)}}$, e por conseguinte

$$\pi_c^{(s+1)} = \frac{\sum_{i=1}^{n} z_{ic}^{(s+1)}}{\sum_{i=1}^{n}\sum_{c=1}^{C} z_{ic}^{(s+1)}}$$

Por outra banda, a maximización do termo B calcúlase directamente establecendo en cero a primeira derivada con respeto a $m_c(\lambda)$ e atopando, como resultado, os estimadores

$$m_c^{(s+1)}(\lambda) = log\left[\frac{\sum_{i=1}^{n} z_{ic}^{(s+1)} \sum_{k=1}^{M} exp(Y_k^i) K_h(\lambda_k - \lambda)}{\sum_{i=1}^{n} z_{ic}^{(s+1)} \sum_{k=1}^{M} K_h(\lambda_k - \lambda)}\right]$$

$$= log\left(\sum_{i=1}^{n} w_{ic}^{(s+1)\hat{f}i,(s+1)}(\lambda)\right),$$

para $c = 1, ..., C$ e λ na cuadrícula seleccionada, onde $w_{ic}^{(s+1)} = z_{ic}^{(s+1)} / \sum_{i=1}^{n} z_{ic}^{(s+1)}$ e $\hat{f}i, (s+1)(\lambda)$ é a estimación de Nadaraya-Watson do espectro con parámetro de suavización h e núcleo K. Hai que destacar que a maximización da log-verosimilitude local completa no paso M leva a expresións de forma pechada para actualizar os centroides e as probabilidades a priori, o que resulta nunha menor complexidade computacional.

Estes dous pasos do algoritmo EM aplícanse iterativamente ata que se cumpre un criterio de detención. Pódese seleccionar varias opcións para determinar este criterio. Neste caso, a regra de detención foi que a probabilidade de rexistro de datos non aumenta significativamente, é dicir

$$\frac{logL(\Theta_{s+1}, Y) - logL(\Theta_s, Y)}{|logL(\Theta_s, Y)|} < \epsilon$$

para algúns valores prefixados e suficientemente pequenos $\epsilon > 0$, ou alternativamente alcanzar un número máximo de iteracións. Unha vez que o algoritmo EM converxe, os valores z_{ic}, para $c = 1, ..., C$ proporcionan a secuencia dos graos de *membership* para a i-ésima serie temporal, $i = 1, ..., n$. De feito, o procedemento EM require valores iniciais para as probabilidades a priori π_c e os centroides $m_c(\cdot)$, $c = 1, ..., C$.

Capítulo 4

Estudo de simulación

4.1. Introdución

Nesta parte do traballo faise uso das técnicas de simulación estatística, coa axuda do software estatístico R, para avaliar e comparar empiricamente a conduta dalgúns dos procedementos de cluster *soft* descritos nos capítulos previos, con especial énfase en observar se a técnica baseada en modelos mixtos descrita no Capítulo 3 resulta ou non competitiva. O primeiro estudo considera diferentes escenarios de análise cluster de series temporais, onde cada cluster vén caracterizado por un patrón xerador específico. Neste contexto simúlanse series dos diferentes grupos e procédese a desenvolver análise cluster con diferentes algoritmos, incluindo obviamente aquel baseado en modelos mixtos no dominio da frecuencia. Considéranse series de diferentes lonxitudes. O segundo estudo de simulación realízase co fin de ver o comportamento do algoritmo EM cando o escenario proposto conta con unha serie equidistante.

A totalidade de simulacións realízanse con series temporais que seguen un modelo autoregresivo (AR). Este modelo é unha representación dun proceso aleatorio onde a variable de interés depende das súas observacións pasadas, é dicir, depende linealmente dos seus valores anteriores. A definición formal para un proceso autoregresivo de orden 1 ou AR(1) é:

$$X_t = c + \phi_1 X_{t-1} + a_t,$$

onde c e ϕ_1 son constantes e as innovacións a_t conforman un proceso de ruído branco con media cero e varianza finita σ_a^2. Con isto verifícase que o proceso AR(1) explica o valor actual X_t como unha función lineal de un valor pasado X_{t-1}.

4.2. Comparativa entre distintos algoritmos

Neste primeiro estudo de simulación preténdese facer unha comparativa entre distintas métricas e algoritmos para chequear a calidade da clasificación cluster que se obtén para cada unha delas. Especificamente, examinar o comportamento do procedemento baseado en modelos mixtos empregando o algoritmo EM mediante a súa comparación con outras técnicas *fuzzy* propostas na literatura. Máis concretamente, e coa intención de considerar métricas representativas de diferentes enfoques, considéranse: unha métrica baseada na función de autocorrelación (ACF) e unha métrica baseada na función de autocovarianzas cuantil (QAF), as cales traballan no dominio temporal, unha métrica baseada no logaritmo do periodograma normalizado (LPN) que traballa no dominio das frecuencias, e unha métrica con coeficientes autoregresivos estimados (AR), proposta por Maharaj (2000).

Para isto, propuxéronse 3 escenarios diferentes. En todos eles, os clusters caracterízanse por modelos AR(1):

$$X_t = \phi X_{t-1} + a_t$$

de xeito que en cada clúster c_i o parámetro autoregresivo $\phi \in U(a_i, b_i)$, con diferentes rangos para a distribución uniforme. Especificamente:

Escenario	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1	$\phi \in U(0, 0.2)$	$\phi \in U(0.4, 0.6)$	$\phi \in U(0.8, 1)$		
2	$\phi \in U$(-1, -0.7)	$\phi \in U$(-0.4, 0.4)	$\phi \in U(0.7, 1)$		
3	$\phi \in U$(-0.9, -0.7)	$\phi \in U$(-0.5, -0.3)	$\phi \in U$(-0.1, 0.1)	$\phi \in U(0.3, 0.5)$	$\phi \in U(0.7, 0.9)$

En cada escenario, 10 series son simuladas dende cada cluster consonte ao modelo especificado na táboa anterior de tal xeito que ϕ se obtén de forma independente para cada serie a partir dunha distribución uniforme.

Os diferentes soportes para as uniformes que determinan os valores de ϕ teñen por obxecto introducir algunha variabilidade na selección do parámetro e crear clusters máis ou menos separados. En definitiva, introducir diferentes niveis de dificultade para caracterizar de xeito único os clusters de pertenza. Nótese que os dous primeiros escenarios contan con 3 clusters e o último con 5 clusters, a intención disto é observar como as diferentes técnicas se ven afectadas polo número de clusters.

Comezouse simulando en R as series temporais destes escenarios, xerándonse 100 repeticións de cada conxunto de series considerando diferentes lonxitudes: $L = 100$, $L = 250$ e $L = 500$. A selección da ventá na estimación non paramétrica xoga un papel fundamental e coa idea de que os resultados

non fosen excesivamente afectados por este parámetro, realizáronse probas preliminares para designar o parámetro razoablemente axeitado para cada lonxitude seleccionada. En concreto, fixéronse probas sobre unha grella de valores equiespaciados entre 0.8 e 1.6 no caso de series de lonxitude 100, entre 0.4 a 0.8 no caso de series de lonxitude 250 e entre 0.2 a 0.6 no caso de series de lonxitude 500.

Agás para o procedemento baseado en modelos mixtos, onde se usaron as tres lonxitudes, para o resto de procedementos as simulacións limitáronse a dúas lonxitudes de series, $L = 100$ e $L = 500$, por entender que para estes procedementos a influenza da lonxitude das series está suficientemente tratada na literatura. Para determinar o parámetro m indicativo do nivel de *fuzziness* requirido para o resto de procedementos (nótese que no caso de modelos mixtos non é necesario), realizouse un estudo previo con distintos valores de m. Concretamente cos valores 1.5, 1.8, 2.0 e 2.2. Preséntanse aquí os resultados para os valores 1.5 e 2 por amosar a mellor conduta.

De cada simulación obtense unha matriz (u_{ij}) de dimensión $C \times S$, sendo C o número de clusters ($C = 3$ nos escenarios 1-2 e $C = 5$ no 3) e S o total de series sometidas a clustering ($S = 30$ ou 50 dependendo do escenario). Así a j-ésima columna proporciona os membership da j-ésima serie.

Cada simulación replicouse 100 veces, de xeito que en cada caso dispoñerase de 100 matrices de membership.

O obxectivo é coñecer o éxito da asignación aos clusters correctos dos distintos algoritmos. Nos escenarios 1 e 2 considéranse 3 umbrais diferentes de clasificación que se detallan a continuación. En concreto, a j-ésima serie é asignada ao i-ésimo cluster se o membership u_{ij} satisfai:

$$\text{(i) } u_{ij} > 0.5$$

$$\text{(ii) } u_{ij} > 0.4$$

$$\text{(iii) } u_{ij} = \max_{1 \leq r \leq C} u_{rj}$$

No escenario 3, o cal conta con cinco clusters, considéranse 3 umbrais diferentes de clasificación, tendo en conta que para este caso os umbrais son máis baixos que nos escenarios anteriores porque ao ter máis clusters as probabilidades repártense máis. Polo tanto, a j-ésima serie é asignada ao i-ésimo cluster se o membership u_{ij} satisfai:

$$\text{(i) } u_{ij} > 0.4$$

$$\text{(ii) } u_{ij} > 0.3$$

$$\text{(iii) } u_{ij} = \max_{1 \leq r \leq C} u_{rj}$$

Para chegar a ese obxectivo primeiro analízase o éxito da clasificación como a proporción do total de

series agrupadas correctamente, logo selecciónase, para cada cluster teórico, o cluster solución no que hai un maior número de series dese cluster teórico e considéranse esas series como ben clasificadas. Repítese o proceso para cada un dos clusters teóricos. Con isto tense o número total de series ben clasificadas e, finalmente, divídese ese número entre o total de series para obter a proporción de series correctamente clasificadas.

As táboas, que a continuación se expoñen, mostran a media e a correspondente desviación típica, entre parénteses. Na primeira columna aparecen as 5 métricas a comparar e nas columnas seguintes aparecen os resultados da clasificación en cada cluster para os umbrais fixados. Na fila do algoritmo EM pódese ver os resultados de cada unha das tres lonxitudes expostas anteriormente para cada umbral proposto. Para as seguintes métricas aparecen dúas filas dentro de cada unha, estas indican o valor do parámetro m co que foron calculadas, polo tanto hai unha fila para $m = 1.5$ e outra para $m = 2$, e os resultados de cada umbral están divididos para as lonxitudes $L = 100$ e $L = 500$.

Escenario 1		**>0.5**			**>0.4**			**Máx.**		
		L=100	L=250	L=500	L=100	L=250	L=500	L=100	L=250	L=500
EM		0.501 (.142)	0.644 (.116)	0.723 (.117)	0.694 (.126)	0.813 (.108)	0.836 (.132)	0.742 (.131)	0.833 (.111)	0.843 (.132)
		L=100		L=500	L=100		L=500	L=100		L=500
ACF	m=1.5	0.566 (.099)		0.544 (.048)	0.566 (.101)		0.544 (.048)	0.566 (.101)		0.544 (.048)
	m=2	0.565 (.105)		0.544 (.057)	0.571 (.107)		0.547 (.059)	0.571 (.107)		0.545 (.059)
AR	m=1.5	0.872 (.123)		0.997 (.009)	0.879 (.123)		0.997 (.009)	0.879 (.123)		0.998 (.009)
	m=2	0.838 (.148)		0.995 (.013)	0.879 (.126)		0.998 (.008)	0.882 (.127)		0.998 (.008)
LPN	m=1.5	0.460 (.121)		0.473 (.130)	0.573 (.131)		0.599 (.136)	0.582 (.137)		0.608 (.141)
	m=2	0.160 (.109)		0.119 (.087)	0.490 (.132)		0.509 (.136)	0.579 (.130)		0.620 (.121)
QAF	m=1.5	0.874 (.118)		0.996 (.013)	0.877 (.118)		0.996 (.013)	0.877 (.118)		0.996 (.013)
	m=2	0.857 (.113)		0.994 (.015)	0.877 (.115)		0.995 (.014)	0.877 (.115)		0.995 (.014)

Cadro 4.1: Taxas promedio de boa clasificación con diferentes algoritmos *soft* sobre 100 réplicas do estudo de simulación proposto no Escenario 1. Considéranse diferentes umbrais e lonxitudes de series. En parénteses a desviación estándar das taxas.

Do Cadro 4.1 dedúcese, en primeiro lugar, que as taxas de asignación correcta aos clusters melloran a medida que se incrementa a lonxitude das series. En tal caso aumenta a precisión dos estimadores dos parámetros empregados por cada métrica e de aí a mellora da clasificación. Tan só no caso do cluster *fuzzy* coa métrica baseada nas autocorrelacións (ACF) non se observa este feito. A mellora tamén se fai patente cando reducimos o umbral para os membership. Umbrais elevados supoñen asignacións claras, pouco difusas, e polo tanto maior dificultade para que se obteñan en escenarios con clusters menos separables. Cómpre subliñar que os procedementos empregados son de cluster *soft* e que os umbrais se introducen a título ilustrativo, co obxecto de dispor de taxas de éxito como se tratase de cluster *hard* e obter así unha idea da conduta dos diferentes procedementos. Por exemplo, para o umbral Máx., o

criterio *hard* asociado é asignar a serie ao cluster co maior membership. Trátase polo tanto do criterio menos restritivo e obviamente co maior risco de erro en escenarios con clusters pouco distanciados.

Agás da métrica ACF, todos os demáis procedementos amosan taxas máis altas a medida que se baixa o umbral, concluíndo así que todos eles tenden a identificar correctamente o cluster de pertenza. As mellores taxas acádanse cos procedementos *fuzzy* baseados nas métricas AR e QAF. A métrica AR basease en asumir un modelo AR que é xusto o caso aquí, logo era de prever que esta métrica traballaría ben. A métrica QAF tamén amosa unha moi boa conduta practicamente equiparable á métrica AR pero non asume estrutura AR para as series, o que ten maior valor se cabe. O procedemento baseado en modelos mixtos empregando EM ocupa unha posición intermedia en termos dos resultados acadados: taxas de éxito claramente superiores aos procedementos *fuzzy* baseados nas métricas LPN e ACF pero inferiores aos baseados en QAF e AR. Cómpre lembrar que esta vía non precisa seleccionar o parámetro m, se ben os resultados neste escenario das métricas QAF e AR non se ven afectados significativamente pola elección deste parámetro.

Escenario 2		**>0.5**			**>0.4**			**Máx.**		
		L=100	L=250	L=500	L=100	L=250	L=500	L=100	L=250	L=500
EM		0.502 (.163)	0.701 (.104)	0.739 (.112)	0.722 (.136)	0.812 (.113)	0.835 (.123)	0.776 (.133)	0.828 (0.118)	0.862 (.109)
		L=100		L=500	L=100		L=500	L=100		L=500
ACF	m=1.5	0.856 (.115)		0.875 (.122)	0.857 (.115)		0.876 (.122)	0.857 (.115)		0.876 (.122)
	m=2	0.842 (.115)		0.857 (.125)	0.854 (.118)		0.872 (.125)	0.854 (.118)		0.872 (.124)
AR	m=1.5	0.960 (.056)		0.988 (.020)	0.961 (.056)		0.988 (.020)	0.961 (.056)		0.988 (.020)
	m=2	0.954 (.057)		0.984 (.022)	0.962 (.056)		0.988 (.019)	0.962 (.056)		0.988 (.019)
LPN	m=1.5	0.791 (.161)		0.864 (.161)	0.839 (.160)		0.899 (.145)	0.842 (.159)		0.899 (.145)
	m=2	0.568 (.161)		0.556 (.159)	0.819 (.166)		0.886 (.161)	0.845 (.161)		0.903 (.154)
QAF	m=1.5	0.967 (.046)		0.995 (.014)	0.967 (.046)		0.995 (.014)	0.967 (.046)		0.995 (.014)
	m=2	0.965 (.034)		0.992 (.016)	0.969 (.033)		0.995 (.014)	0.969 (.033)		0.995 (.014)

Cadro 4.2: Taxas promedio de boa clasificación con diferentes algoritmos *soft* sobre 100 réplicas do estudo de simulación proposto no Escenario 2. Considéranse diferentes umbrais e lonxitudes de series. En parénteses a desviación estándar das taxas.

Neste Escenario 2 a tendencia é a mesma que no anterior, xa que se obteñen mellores resultados a medida que se aumenta a lonxitude das series, e con umbrais menos restritivos. Pero obsérvase que os resultados amosan unha maior media para todas as métricas a comparar e para todos os umbrais, isto débese a que os clusters están claramente máis separados que no Escenario 1, o que explica esta mellora das taxas de éxito observadas para todas as métricas. As funcións AR e QAF son as que amosan mellores resultados, o cluster *fuzzy* baseado en QAF chega incluso a mellorar á propia métrica AR (aínda que os resultados seguen a ser moi parellos). Estas dúas métricas melloran ostensiblemente respecto ao Escenario 1 en tanto que o procedemnto de modelos mixtos non presenta unha mellora

significativa. As función ACF e LPN melloraron moito para este escenario, subindo entre 0.3 e 0.4 a media con respecto ao escenario anterior, superando incluso ao algoritmo EM.

Escenario 3		**>0.4**			**>0.3**			**Máx.**		
		L=100	L=250	L=500	L=100	L=250	L=500	L=100	L=250	L=500
EM		0.393 (.099)	0.530 (.083)	0.619 (.100)	0.595 (.096)	0.709 (.096)	0.776 (.099)	0.686 (.088)	0.757 (.091)	0.792 (.096)
		L=100		L=500	L=100		L=500	L=100		L=500
ACF	m=1.5	0.711 (.106)		0.784 (.108)	0.711 (.106)		0.784 (.108)	0.711 (.106)		0.784 (.108)
	m=2	0.698 (.101)		0.781 (.106)	0.713 (.102)		0.781 (.106)	0.713 (.102)		0.781 (.106)
AR	m=1.5	0.825 (.109)		0.996 (.014)	0.828 (.108)		0.996 (.014)	0.828 (.108)		0.996 (.014)
	m=2	0.811 (.113)		0.996 (.018)	0.834 (.112)		0.996 (.016)	0.834 (.112)		0.996 (.016)
LPN	m=1.5	0.412 (.083)		0.361 (.068)	0.611 (.096)		0.689 (.119)	0.645 (.095)		0.791 (.108)
	m=2	0.102 (.039)		0.071 (.025)	0.328 (.100)		0.237 (.126)	0.609 (.105)		0.699 (.119)
QAF	m=1.5	0.841 (.097)		0.991 (.014)	0.842 (.096)		0.991 (.014)	0.842 (.096)		0.991 (.014)
	m=2	0.839 (.094)		0.991 (.014)	0.841 (.093)		0.991 (.014)	0.841 (.093)		0.991 (.014)

Cadro 4.3: Taxas promedio de boa clasificación con diferentes algoritmos *soft* sobre 100 réplicas do estudo de simulación proposto no Escenario 3. Considéranse diferentes umbrais e lonxitudes de series. En parénteses a desviación estándar das taxas.

Para o Escenario 3 os umbrais fixados son distintos aos demais, xa que como se explicou arriba este escenario conta con 5 clusters e, polo tanto, considerouse oportuno reducir a 0.4 e 0.3 o valor para o éxito de clasificación en cada membership. Con isto, a tendencia continúa a ser a mesma, xa que se obteñen mellores resultados a medida que se aumenta a lonxitude das series, e con umbrais menos restritivos, pero xa non se obteñen tan bos resultados como nos escenarios anteriores. Para este escenario é normal que os resultados empeoren xa que é moito máis complexo que os anteriores e, a pesar diso, os resultados continúan sendo moi bos.

4.3. Algoritmo EM con serie equidistante

A segunda parte do estudo de simulación está deseñada para examinar se os procedementos *soft* son capaces de detectar adecuadamente a equidistancia dunha serie a dous clusters, propiedade que non permiten os procedementos de cluster *hard* que ubicarían a serie equidistante nun dos cluster distorsionando ademais o patrón representativo do mesmo. Especificamente, créase un novo escenario (Escenario 4) con dous clusters con estrutura autoregresiva, como nos escenarios previos, e unha serie con coeficiente autoregresivo equidistante destes que definen os dous clusters.

Escenario	Cluster 1	Cluster 2	Serie equidistante
4	$\phi \in U(0.15, 0.2)$	$\phi \in U(0.8, 0.95)$	$\phi = (0.5)$

Coma nos anteriores escenarios simúlanse 10 series para cada un dos dous clusters e 1 serie equidistante. De igual xeito que nas simulacións previas, con distintas lonxitudes de 100, 250 e 500 e con un número de repeticións igual a 100. O obxectivo é comprobar como funciona o algoritmo EM cando a clasificación se complica con unha serie equidistante.

Neste caso os umbrais para os membership en orde a determinar o cluster de pertenza de cada serie son: unha serie pertence ao cluster c_i se o i-ésimo membership supera o 0.7 mentres que noutro caso se considera equidistante aos dous clústers, toda vez que ambos membership se moven entre 0.3 e 0.7.

Na seguinte táboa amósase a taxa de éxito medida como o promedio da proporción de series ben clasificadas en cada cluster:

	Éxito cluster 1 (>0.7)	**Éxito cluster 2 (>0.7)**	**Éxito serie equidistante (0.3 – 0.7)**
L=100	68.3%	78.1%	84%
L=250	93.8%	90.7%	87%
L=500	99.3%	90.1%	96%

Cadro 4.4: Éxito de clasificación en porcentaxe (Umbral 1)

Como se pode comprobar a porcentaxe de éxito tende a mellorar segundo se aumenta a lonxitude das series simuladas, tanto para o cluster 1, como para o cluster 2, así como para a serie equidistante. O algoritmo EM funciona moi ben á hora de detectar esa serie equidistante con porcentaxes superiores ao 80 %.

Na seguinte táboa amósase a porcentaxe de éxitos que se obtivo cando se considera un umbral de 0.6 en lugar de 0.7, e acotando o rango entre 0.4 e 0.6 para considerar a equidistante:

	Éxito cluster 1 (>0.6)	**Éxito cluster 2 (>0.6)**	**Éxito serie equidistante (0.4 – 0.6)**
L=100	87.2%	85.1%	71%
L=250	98.5%	91.8%	80%
L=500	99.6%	91.1%	86%

Cadro 4.5: Éxito de clasificación en porcentaxe (Umbral 2)

Obsérvase de novo que hai unha tendencia de máis éxito canto máis grande sexa a lonxitude das series simuladas. Pero pódese apreciar que o éxito da serie equidistante para este umbral fixado é menor que co umbral anterior, ata un 10 % menos de éxitos, xa que se acota cara o rango de 0.5. O algoritmo EM segue a funcionar moi ben á hora de detectar esa serie equidistante con porcentaxes superiores ao 70 %.

Capítulo 5

Conclusións

Neste traballo explórase o comportamento dun procedemento de analise cluster *soft* de series de tempo baseado en modelos mixtos dentro do domino de frecuencias.

Atendendo aos obxectivos principais do traballo, en primeiro lugar realízase unha revisión dalgúns dos principais resultados da análise espectral, con especial énfase no concepto do periodograma, como análogo mostral da densidade espectral, o cal non é consistente e é altamente variable polo que se reemplaza por unha versión suavizada do mesmo.

Preséntase o problema xeral de análise cluster de series temporais, un tema con numerosas aplicacións en diferentes ramas de coñecemento e con características específicas que fan do seu tratamento unha tarefa complexa. En particular, a necesidade de determinar unha medida axeitada para avaliar a disimilitude entre series temporais e o problema da alta dimensionalidade inherente á observación de series temporais. Establécense as diferencias entre os enfoques *hard* e *soft* en clustering. A diferencia da vía *hard* ou estándar, o cluster *soft* permite asignar obxectos a varios clusters simultaneamente, resultando así un enfoque máis flexible e de particular utilidade en algúns problemas onde resulta natural atopar clusters con certo grao de solapamento. Tras describir os procedementos *fuzzy* como vía máis usualmente empregada en clustering *soft*, preséntase unha técnica máis novidosa proposta en Lafuente (2017).

A vía proposta por Lafuente (2017) toma a vantaxe da modelización do log-periodograma para series estacionarias. No presente traballo descríbese en detalle o procedemento de clustering proposto, que fai uso do algoritmo EM para acadar os coeficientes que conducen ao modelo maximizando a verosimilitude dos rexistros, e equivalentemente as probabilidades de pertenza a cada cluster.

Na parte práctica deste traballo, que atende ao cuarto obxectivo, realízase un pequeno estudo de simulación con modelos autoregresivos de orde 1 para examinar a conduta do procedemento baseado en modelos mixtos. Os resultados amosan un algoritmo razoablemente competitivo, con taxas de boa

clasificación elevadas, que melloran coa lonxitude das series a separabilidade dos clusters. Comparado con outras vías propostas na literatura ocupa un lugar intermedio, mellorando a procedementos *fuzzy* baseados en métricas que avalían distancias en termos de autocorrelacións (D'Urso e Maharaj, 2009) ou log-periodogramas (Caiado et al., 2006) estimados, pero conducindo a resultados peores que procedementos *fuzzy* baseados en comparar estimacións de coeficientes autoregresivos (Maharaj, 2000) ou autocovarianzas cuantil (Vilar e Lafuente, 2017). Naturalmente, a técnica de Maharaj (2000) era xa esperada para arroxar bos resultados por estar especificamente deseñada para modelos autoregresivos. Ao mesmo tempo cómpre subliñar que, por construción, a técnica considerando modelos mixtos non require pre-establecer un parámetro de *fuzziness*, como si ocorre co resto de procedementos *fuzzy*.

Para finalizar, pódese plantexar para un traballo futuro a ampliación deste estudo de simulación. Así, poderían simularse series para uns escenarios que sigan un modelo non lineal e realizar a comparativa coas mesmas métricas, xa que posiblemente houbera cambios nos resultados destas. Tamén sería interesante aplicar o algoritmo EM a datos reais para comprobar a clasificación cluster que fai destes.

Bibliografía

[1] Aielli, G. P. e Caporin, M. (2013). Fast clustering of GARCH processes via gaussian mixture models. *Math. Comput. Simul., 94,* 205-222.

[2] Alonso, A. M., Berrendero, J. R., Hernández, A., e Justel, A. (2006). Time series clustering based on forecast densities. *Comput. Stat. Data Anal., 51(2),* 762-776.

[3] Bouveyron, C., Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics and Data Analysis, 71,* 52-78.

[4] Caiado, J., Crato, N., e Peña, D. (2006). A periodogram-based metric for time series classification. *Comput. Stat. Data Anal., 50(10),* 2668-2684.

[5] Casado de Lucas, D. (2010). Classification techniques for time series and functional data. (Tese doutoral).

[6] Chen, W. e Maitra, R. (2011). Model-based clustering of regression time series data via apecm - an aecm algorithm sung to an even faster beat. *Stat. Anal. Data Min., 4(6),* 567-578.

[7] Douzal-Chouakria, A., Diallo, A., e Giroud, F. (2009). Adaptive clustering for time series: Application for identifying cell cycle expressed genes. *Comput. Statist. Data Anal., 53(4),* 1414-1426.

[8] Douzal-Chouakria, A. e Nagabhushan, P. N. (2007). Adaptive dissimilarity index for measuring time series proximity. *Adv. Data Anal. Classif., 1(1),* 5-21.

[9] D'Urso, P. e De Giovanni, L. (2014). Robust clustering of imprecise data. *Chemometrics Intell. Lab. Syst., 136,* 58-80.

[10] D'Urso, P., De Giovanni, L., e Massari, R. (2015). Time series clustering by a robust autoregressive metric with application to air pollution. *Chemometrics Intell. Lab. Syst., 141,* 107-124.

[11] D'Urso, P., De Giovanni, L. e Massari, R. (2016). GARCH-based robust clustering of time series. *ScienceDirect, 305,* 1-28.

[12] D'Urso, P. e Maharaj, E. A. (2009). Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets Syst., 160(24),* 3565-3589.

[13] Galeano, P. e Peña, D. (2000). Multivariate analysis in vector time series. *Resenhas, 4(4),* 383-403.

[14] García-Magariños, M. e Vilar, J. A. (2015). A framework for dissimilarity-based partitioning clustering of categorical time series. *Data Min. Knowl. Discov., 29(2),* 466-502.

[15] Golay, X., Kollias, S., Stoll, G., Meier, D., Valavanis, A., e Boesiger, P. (2005). A new correlation-based fuzzy logic clustering algorithm for fmri. *Magn. Reson. Med., 40(2),* 249-260.

[16] Grimaldi, S. (2004). Linear parametric models applied to daily hydrological series. *J. Hydrol. Eng., 9(5),* 383-391.

[17] Kakizawa, Y., Shumway, R. H., e Taniguchi, M. (1998). Discrimination and clustering for multivariate time series. *J. Amer. Statist. Assoc., 93(441),* 328-340.

[18] Kullback, S. e Leibler, R. A. (1951). On information and suficiency. *Ann. Math. Statist., 22(1),* 79-86.

[19] Lafuente Rego, B. (2017). New methodological contributions in time series clustering (Tese doutoral). Universidade da Coruña.

[20] Li, C., Biswas, G., Dale, M., e Dale, P. (2001). Building models of ecological dynamics using hmm based temporal data clustering - A preliminary study. *F. Hoffmann et al. (Eds.): Advances in Intelligent Data Analysis, IDA 2001,* 53-62.

[21] Liao, T. W. (2005). Clustering of time series data: A survey. *Pattern Recognit., 38(11),* 1857-1874.

[22] Maharaj, E. A. (1996). A significance test for classifying ARMA models. *J. Statist. Comput. Simulation, 54(4),* 305-331.

[23] Maharaj, E. A. (2000). Cluster of time series. *J. Classif., 17(2),* 297-314.

[24] Maharaj, E. A. e D'Urso, P. (2011). Fuzzy clustering of time series in the frequency domain. *Inf. Sci., 181(7),* 1187-1211.

[25] Montero, P. e Vilar, J. (2014). TSclust: An R Package for Time Series Clustering. *Journal Of Statistical Software, 62(1).*

[26] Ohashi, Y. (1985). Fuzzy Clustering and Robust Estimation. *Proceedings of the first Fuzzy Sistems Symposium.*

[27] Piccolo, D. (1990). A distance measure for classifying arima models. *J. Time Series Anal., 11(2),* 153-164.

[28] Priestley, M. B. (1989). Spectral Analysis of Time Series. Probability and Newline Mathematical Statistics. A Series of Monograph an Text books. *Academic Press.*

[29] Ruppert, D., Sheather, S. J., e Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc., 90(432),* 1257-1270.

[30] Ruspini, E. H., Bezdek, J. C. e Keller, J. M. (2019). Fuzzy Clustering: A Historical Perspective. *IEEE Computational Intelligence Magazine, 14(1),* 45-55.

[31] Shumway, R. e Stoffer, D. (2006). Time series analysis and its applications. With R Examples. USA: Springer.

[32] Vilar, J. A., Alonso, A. M., e Vilar, J. M. (2010). Non-linear time series clustering based on non-parametric forecast densities. *Comput. Statist. Data Anal., 54(11),* 2850-2865.

[33] Vilar, J. A., Lafuente Rego, B. e D'Urso, P. (2017). Quantile autocovariances: A powerful tool for hard and soft partitional clustering of time series. *Fuzzy Sets Syst.*

[34] Wong, C. S. e Li, W. K. (2000). On a mixture autoregressive model. *J. R. Stat. Soc. Series B Stat. Methodol., 62(1),* 95-115.