Modelo SSA: Análisis Espectro Singular

Adriana Abrego
Analítica Financiera

Como hemos visto, existen diferentes métodos para el análisis de series de tiempo. La mayoría de éstos, son paramétricos (ejemplo que requieran consideraciones de distribuciones particulares en su análisis, tales como el ARIMA o ARMA). Existen métodos alternativos, los no paramétricos que son neutrales en cuanto especificaciones de linealidad, estacionariedad y normalidad (dentro estos ejemplos la mayoría de los modelos de Machine Learning). Dentro de estos modelos, está el Análisis de Espectro Singular, o Singular Spectrum Analysis (SSA, por sus siglas en inglés), que es un método relativamente nuevo, particularmente no paramétrico que ha probado tener capacidad en muchas aplicaciones de series de tiempo, desde las ramas económicas hasta las ramas de la física. Así, en esta lectura aprenderemos sobre el método SSA, para lo cual conoceremos su uso incremental debido al crecimiento del Big Data. Seguidamente aprenderemos sobre la utilidad de este método y sus dos fases, como lo son el mapeo y descomposición y la reconstrucción. Finalmente, aprenderemos sobre el scatterplot, el cual nos permite identificar eigentriples que corresponden a los componentes harmónicos de la serie.

El uso del SSA se ha también incrementado

Crecimiento del Big Data que usualmente incrementa el ruido en las series de tiempo (Hassani *et al.*, 2018), que finalmente resulta en una distorsión de la señal y con ello demerita el proceso de pronóstico.

Las condiciones de volatilidad en los mercados que asegura que las series, no sean en su mayoría estacionarias. Esto, **especialmente en periodos de crecimiento seguidos por recesiones**.

La transformación de los datos para llevarlos a una condición de estacionariedad lleva a que se pierda información.

El SSA no está limitado por esta condición, lo que supera las restricciones usuales de modelos paramétricos en relación con la estructura de los datos. El SSA es muy útil pues **permite la**

descomposición de la serie extrayendo sus componentes como tendencia, componentes estacionales y componentes cíclicos, como comentaremos a continuación.

El proceso puede emplear series univariadas o multivariadas; donde el proceso general consta de dos fases:

1. Mapeo y descomposición. Se descompone la serie original en un pequeño número de componentes interpretables, como la variación lenta de la tendencia, componentes oscilatorios y ruido.

Esta descomposición es particularmente de tipo descomposición singular[1], que básicamente busca descubrir estructuras en los datos; en nuestro caso, una estructura de tendencia que pudiera ser muy compleja de modelar con métodos tradicionales; o componentes también de estacionalidad). Partiendo de una serie univariada, la serie se mapea en un espacio diferente, que llamaremos de trayectorias, se forma así una matriz con unas características que dependen de la longitud de la ventana L, de análisis. Esta longitud puede ir de $2 \leq L < N/2$, donde N es la longitud de nuestra serie. Posteriormente, se aplica una descomposición singular a la serie, que posee como objetivo caracterizar los componentes de las distintas partes independientes, esto es, con el fin de identificar la tendencia, la estacionalidad y el ruido, para de esta manera poder separar el ruido de toda la señal.

Reconstrucción. Aquí, se reconstruye la serie y se genera un pronóstico. Es decir, la serie filtrada se reconstituye y se emplea entonces para el pronóstico A continuación, se explicarán ambas fases del SSA.

[1] Para entender más qué es la descomposición singular, te recomiendo ver este documento: https://ichi.pro/es/aprendizaje-automatico-descomposicion-de-valores-singulares-svd-y-analisis-de-componentes-principales-pca-32186091037841

Fase 1:

Inicialmente la serie de tiempo unidimensional es tomada y transformada en una serie multidimensional de $x = [x_1, x_2, \ldots . x_k]$ vectores donde K = N-L+1. (T) es el tamaño de la serie de tiempo original y (L) es un parámetro suministrado, el cual, se conoce como ventana y es un valor entre 2 y T, en la literatura se establece que el valor de (L) no debe ser mayor a T/2. De esta manera, se obtiene una matriz x llamada matriz de trayectoria esta matriz tiene la forma de una matriz de Hankel. Esta matriz pasa por un proceso llamado *singular value descomposition*, donde adquiere la forma de $\sum_{l=1}^{d} X_l$ donde $x_i = \sqrt{\lambda_i} U_i V_i^{'}$ los valores de $(\sqrt{\lambda_i}, U_i, V_i^{'})$ se conocen como eigentriples.

Fase 2:

En esta fase se genera la agrupación de las matrices elementales X en subgrupos siendo $I = [i_1, i_2, \ldots . i_p]$ un grupo de índices correspondientes al grupo de la matriz X_i. De esta forma, se puede reescribir $X = \sum_{i=1}^{m} X_{Ii}$. Asimismo, la contribución de cada componente de X_{Ii} se puede medir por medio del correspondiente eigenvalue $\frac{\sum_{i \in I} \lambda_i}{\sum_{i=1}^{d} \lambda_i}$.

Además, en esta fase se genera el *diagonal averaging* donde cada matriz I es transferida a una serie de tiempo. Por lo que, se obtiene una matriz de Hankel (HZ). Al aplicar el diagonal averaging sobre todos los componentes de la matriz se obtiene la expansión:

$$y_t = \sum_{k=1}^{m} X_{Ii}$$

Cada componente contiene una parte de la serie original, los cuales pueden ser clasificados en 3 categorías: tendencia, componentes armónicos y ruido.

Debido a que el método busca poder identificar y separar los componentes de tendencia, ruido y estacionalidad. El reto principal consiste en poder separar los componentes de ruido y estacionalidad de aquellos que componen a la tendencia. Debido a esta dificultad, surgen herramientas graficas que permiten visualizar los componentes harmónicos de la serie, y así, poder discriminarlos y separarlos.

Scatterplot:

El scatterplot ayuda a identificar eigentriples que corresponden a los componentes harmónicos de la serie. En situaciones ideales los eigenvectores y los componentes principales forman secuencias de seno y coseno. En donde, si poseen igual secuencia, amplitud y fase se pueden generan puntos de un polígono regular. A continuación, en la figura 6, se pueden observar pares de componentes, los cuales, generan polígonos regulares de estas funciones con estas características.

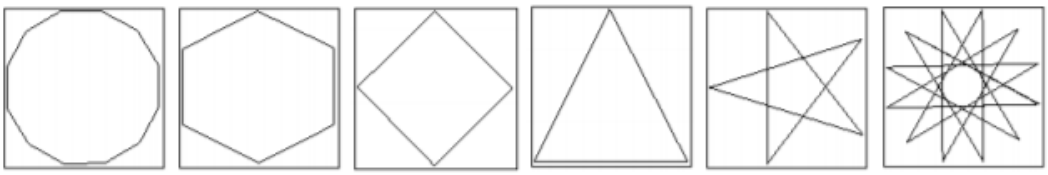

Figura 1. Formas de seno y coseno idealizadas.

Tomado de: Singular Spectrum Analysis: Methodology and Comparison, Hossein Hassani (2007, p. 247).

Obsérvese que en la práctica se espera que los scatterplot de los pares de eigenvectores tengan este tipo de formas o sean lo más parecidas posibles. En un escenario normal los Scatterplot pueden generar figuras del siguiente estilo:

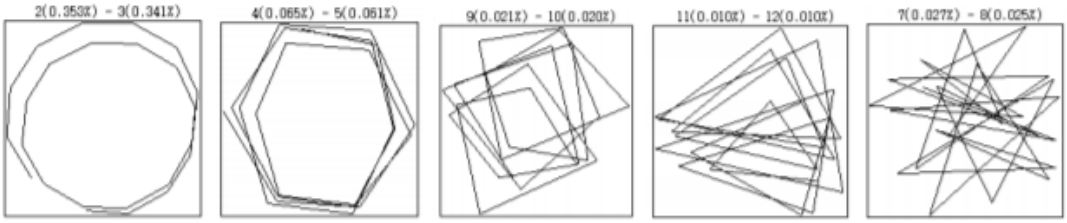

Figura 2. Ejemplos de pares de Eigenvectors.

Tomado de: Singular Spectrum Analysis: Methodology and Comparison, Hossein Hassani (2007, p. 247)

Teniendo en cuenta figura 7, los pares de eigenvectores a escoger son los que más se parezcan a figuras geométricas (diagrama 11). Para el diagrama 12 se deben seleccionar los pares 2-3 y 4-5.

Separability:

En este se contempla que se pueden separar los componentes de la señal. Para medirlo se utiliza un valor llamado el w-correlation, el cual, permite calcular la dependencia entre dos valores Y_T^1, Y_T^2. Se expresa por medio de la siguiente fórmula:

$$p(w) = \frac{Y_T^1, Y_T^2}{\left|\left|Y_T^1\right|\right|_w \left|\left|Y_T^2\right|\right|_w}$$

Donde,

$$\left|\left|Y_T^i\right|\right|_w = \sqrt{Y_T^i, Y_T^i, (Y_T^i, Y_T^i)}$$

Si el valor absoluto del w-correlation es pequeño, entonces estas series son ortogonales. Sin embargo, si el valor de w-correlation es grande las series no son ortogonales. Con esto en mente, si dos componentes poseen w-correlation de cero o cercano, quiere decir que son separables. Además, si poseen un w-correlation alto esto puede indicar que los componentes deben estar concentrados en un grupo y por lo tanto podrían tener ruido. Para medir el w-correlation en la selección de componentes del SSA, se puede utilizar una matriz de w-correlaciones como la que se muestra a continuación:

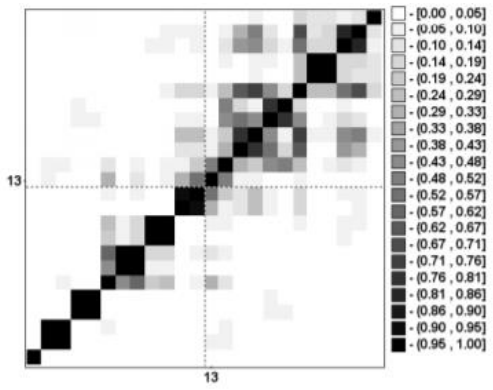

Figura 3. Ejemplos de pares de Eigenvectors.
Tomado de: Singular Spectrum Analysis: Methodology and Comparison, Hossein Hassani (2007, p. 249)

Esta matriz nos permite observar que los valores cuya diagonal sea negra y alrededor de esta los valores sean blancos, son componentes que se pueden incluir pues se pueden separar. Sin embargo, si en la gráfica se observan recuadros grises esto quiere decir que estos componentes no se pueden separar (existe ruido), porque el w-correlation es alto y por lo tanto no se pueden incluir en el SSA.

De esta manera, una vez seleccionados los componentes en la fase de reconstrucción y realizado el promedio diagonal, para poder obtener datos en la forma de serie de tiempo univariada (es decir, pasamos de unos datos tipo matriz de trayectorias a serie de tiempo univariada reconstruida o filtrada), generamos los pronósticos (Aldana Cortes & Guerrero Arango, 2021).

En R, el paquete de library(Rssa) de Golyandina (et al., 2018) se puede emplear para aplicar las fases de mapeo, descomposición y recontrucciónd de la serie. El pronóstico se puede realizar mediante la función de forecast() o predict(), muy similar a la manera que se emplea para los modelos ARIMA o holt Winters. Se proporciona en el primer argumento, el objeto resultado de la descomposición, y los demás parámetros que indican la selección de los componentes a emplear en el pronóstico, que son resultado de la fase de reconstrucción.

Referencias bibliográficas

Aldana Cortes, N. A., & Guerrero Arango, C. A. (2021). Proyecto de Grado: Seguro Indexado a la Precipitación para Cultivos de Banano en Antioquia y Magdalena y la Influencia que Tiene en Cambio Climático sobre estos Integrantes. Bogotá: Uniandes.

Golyandina, N., Korobeynikov, A. & Zhigljavsky, A. 2018. Singular Spectrum Analysis with R. Springer

Hassani, H., & Mahmooudvand, R. 2018. Singular Spectrum Analysis Using R. Palgrave, McMillan.

Hassani, H. (2007). Singular Spectrum Analysis: Methodology and Comparison. Journal of Data Science, vol. 5, núm. 2, pp. 239-257. Recuperado de http://www.jds-online.com/files/JDS-396.pdf