USO DE TÉCNICAS DE ANÁLISIS DE DATOS EN LA DETERMINACIÓN DE VARIABLES SIGNIFICATIVAS EN LA EVOLUCIÓN DE LA EPIDEMIA COVID 19

Ignacio Garach Vélez

Universidad de Granada

5 de julio de 2023

Tabla de contenidos

Tabla de contenidos

Objetivos

- Iniciales:
  - Descripción del método de Análisis de Componentes Principales.
  - Aplicación a los datos de la epidemia en España.
  - Desarrollo de una aplicación web interactiva.

Objetivos

- Iniciales:
  - Descripción del método de Análisis de Componentes Principales.
  - Aplicación a los datos de la epidemia en España.
  - Desarrollo de una aplicación web interactiva.
- Adicionales:
  - Comparación de la variable *resumen* con las variables de partida.
  - Despliegue de la aplicación en la nube.

Tabla de contenidos

Variables aleatorias

Consideramos un conjunto de variables aleatorias $X_1, ..., X_n$.

Variable *Resumen*

Nueva variable aleatoria X_r combinación lineal de las anteriores de modo que recoja la máxima información y variabilidad de todas las demás.

$$X_r = a_1 X_1^N + a_2 X_2^N + ... + a_n X_n^N$$

Consideramos un conjunto de variables aleatorias $X_1, ..., X_n$.

Variable *Resumen*

Nueva variable aleatoria X_r combinación lineal de las anteriores de modo que recoja la máxima información y variabilidad de todas las demás.

$$X_r = a_1 X_1^N + a_2 X_2^N + ... + a_n X_n^N$$

- Estandarización: Obtenemos variables con escalas comparables y con la misma media y varianza.

$$X_i^N = \frac{X_i - \bar{X}_i}{\sqrt{Var[X_i]}}$$

- $E[X_i^N] = 0$ y $Var[X_i^N] = 1$

Matriz de covarianzas

Dadas 2 variables aleatorias estandarizadas, se tiene que

$$Cov[X,Y] = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

Matriz de covarianzas

Dadas 2 variables aleatorias estandarizadas, se tiene que

$$Cov[X,Y] = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

Por tanto, dado el nuevo conjunto de variables, $X_1^N, X_2^N ..., X_n^N$:

Matriz de covarianza

Codifica las covarianzas de cada par de variables:

$$\Sigma = \begin{bmatrix} Var[X_1^N] & Cov[X_1^N, X_2^N] & \cdots & Cov[X_1^N, X_n^N] \\ Cov[X_1^N, X_2^N] & Var[X_2^N] & \cdots & Cov[X_2^N, X_n^N] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[X_1^N, X_n^N] & Cov[X_2^N, X_n^N] & \cdots & Var[X_n^N] \end{bmatrix}$$

Matriz de covarianzas

Dadas 2 variables aleatorias estandarizadas, se tiene que

$$Cov[X,Y] = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

Por tanto, dado el nuevo conjunto de variables, $X_1^N, X_2^N ..., X_n^N$:

Matriz de covarianza

Codifica las covarianzas de cada par de variables:

$$\Sigma = \begin{bmatrix} Var[X_1^N] & E[X_1^N X_2^N] & \cdots & E[X_1^N X_n^N] \\ E[X_1^N X_2^N] & Var[X_2^N] & \cdots & E[X_2^N X_n^N] \\ \vdots & \vdots & \ddots & \vdots \\ E[X_1^N X_n^N] & E[X_2^N X_n^N] & \cdots & Var[X_n^N] \end{bmatrix}$$

Propiedades de la variable resumen

Linealidad de X_r

$$X_r = a_1 X_1^N + a_2 X_2^N + ... + a_n X_n^N$$

- Esperanza matemática:

$$\begin{aligned} E[X_r] &= E[a_1 X_1^N + a_2 X_2^N + \cdots + a_n X_n^N] \\ &= a_1 E[X_1^N] + a_2 E[X_2^N] + \cdots + a_n E[X_n^N] = 0 \end{aligned}$$

- Varianza:

$$Var[X_r] = E[X_r^2] = E[(a_1 X_1^N + a_2 X_2^N + \cdots + a_n X_n^N)^2]$$

Varianza de X_r

Proposición

Sea X_r una variable aleatoria combinación lineal de $X_1^N, X_2^N ..., X_n^N$ y sea $v = (a_1, a_2, ..., a_n) \in \mathbb{R}^N$ el vector de coeficientes de dicha combinación lineal.
Consideremos también la matriz de covarianza Σ_{X_r} del conjunto de variables $X_1^N, X_2^N ..., X_n^N$. Entonces:

$$Var[X_r] = v^T \Sigma_{X_r} v \tag{1}$$

Varianza de X_r

Corolario

La varianza de las distintas combinaciones lineales $X_r = a_1 X_1^N + a_2 X_2^N + ... + a_n X_n^N$ *define una forma cuadrática en* $\mathbb{R}^n$

$$\begin{aligned} \mathcal{V} \colon \mathbb{R}^n &\longrightarrow \mathbb{R} \\ v &\longmapsto v^T \Sigma_{X_r} v = \langle \Sigma_{X_r} v, v \rangle \end{aligned}$$

Varianza de X_r

Corolario

La varianza de las distintas combinaciones lineales $X_r = a_1 X_1^N + a_2 X_2^N + ... + a_n X_n^N$ *define una forma cuadrática en* $\mathbb{R}^n$

$$\begin{aligned} \mathcal{V}\colon \mathbb{R}^n &\longrightarrow \mathbb{R} \\ v &\longmapsto v^T \Sigma_{X_r} v = \langle \Sigma_{X_r} v, v \rangle \end{aligned}$$

- $\mathcal{V}$ es semidefinida positiva.
- $\mathcal{V}$ no está acotada superiormente.

Optimización de la forma cuadrática $\mathcal{V}$

Teorema

Sea $\mathcal{V}: \mathbb{S}^{n-1} \longrightarrow \mathbb{R}$ *una función real continua definida en el compacto* $\mathbb{S}^{n-1}$*. Entonces* $\mathcal{V}$ *alcanza un valor máximo.*
Es decir, existe un $a \in \mathbb{S}^{n-1}$ *tal que* $\mathcal{V}(a) \geq \mathcal{V}(x) \qquad \forall x \in \mathbb{S}^{n-1}$

Teorema

Dada una matriz simétrica $\Sigma \in \mathbb{R}^{n \times n}$ *existe un vector* $u_1 \in \mathbb{S}^{n-1}$ *tal que maximiza la forma cuadrática* $\mathcal{V}(x) = x^T \Sigma x$ *en la esfera unitaria.*
Además u_1 *es un vector propio, asociado al valor propio real* $\lambda_1 = u_1^T \Sigma u_1$ *, es decir:*

$$u_1 = \arg \max_{\|x\|_2 = 1} x^T \Sigma x \tag{2}$$

$$\lambda_1 = \max_{\|x\|_2 = 1} x^T \Sigma x \tag{3}$$

Caso particular. Teorema de Perron.

Teorema

Si la matriz de covarianzas

$$\Sigma = \{Cov[X_i^N, X_j^N]\}_{i,j \in 1,\ldots,N} >> 0 \tag{4}$$

entonces todos los pares de variables están correladas y λ_1 *es un valor propio de Perron. Además, su vector propio asociado es positivo y por tanto, todos los pesos que generan la variable resumen* $X_r = a_1 X_1^N + \ldots + a_n X_n^N$ *son positivos. En particular, la variable resumen* X_r *correla positivamente con* X_i *para todo* $i = 1, 2, \ldots, n$

Casos particulares. Valor propios mínimo y máximo.

Proposición

Si $\lambda_1 = 1$ *entonces la matriz de covarianzas es la identidad* $\Sigma = I_n$ *y todas las variables de partida son incorreladas. Además la forma cuadrática* $\mathcal{V}$ *es constantemente 1.*

Proposición

Sea $\Sigma = \{Cov[X_i^N, X_j^N]\}_{i,j \in 1,\ldots,N} >> 0$ *y supongamos que* $\lambda_1 = N$ *es un valor propio de* Σ*. Entonces la matriz* Σ *tiene unos en todas sus entradas y por tanto, las variables* $X_i^N = X_j^N$ *en el espacio de probabilidad.*

Varianza explicada

Definición

El valor propio λ_1 que es el máximo valor que toma la forma cuadrática $\mathcal{V}$ en la esfera unitaria se denomina varianza explicada. Representa el número de variables del conjunto de partida que X_r consigue representar.

Definición

El porcentaje de varianza explicada por la variable resumen viene dado por el cociente entre el mayor valor propio λ_1 y el número de variables de partida. Representa el porcentaje de variables del conjunto original que es capaz de resumir.

$$\%VarianzaExp = \frac{Var[X_r]}{\sum_{i=1}^{n} Var[X_i^N]} \cdot 100 = \frac{\lambda_1}{n} \cdot 100$$

Algoritmo

Dado un conjunto de datos de tamaño m con n variables $X_1, \cdots, X_n$:

1. *Estandarizamos cada variable para que tengan media 0 y varianza 1.*
2. *Calculamos la matriz de covarianzas $\Sigma_{X_1,\cdots,X_n}$.*
3. *Diagonalizamos la matriz para calcular los coeficientes que definen cada componente.*
4. *Calculamos las componentes como es producto de los coeficientes por las variables.*

Tabla de contenidos

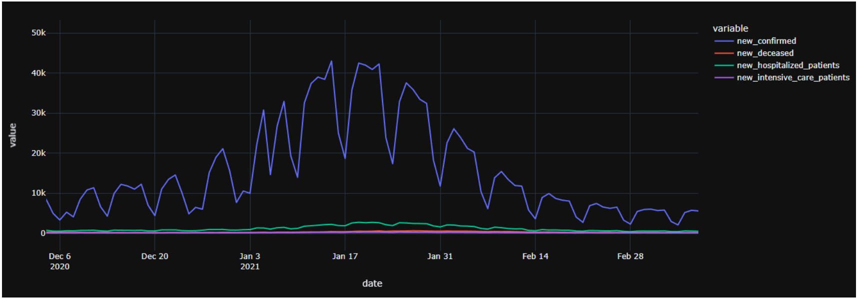
Variables consideradas

- Salud
- Sociales y sobre medidas gubernamentales
- Movilidad

Fuentes:

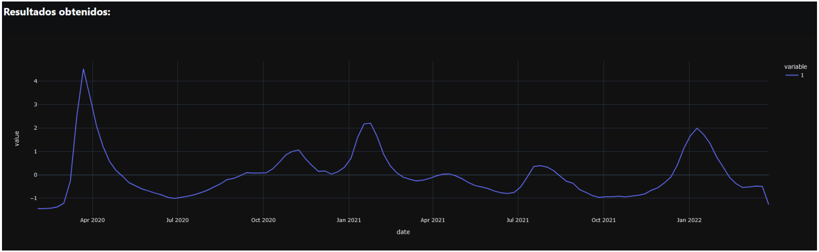
- COVID-19 Open-Data Google
- OWID COVID-19 Repository

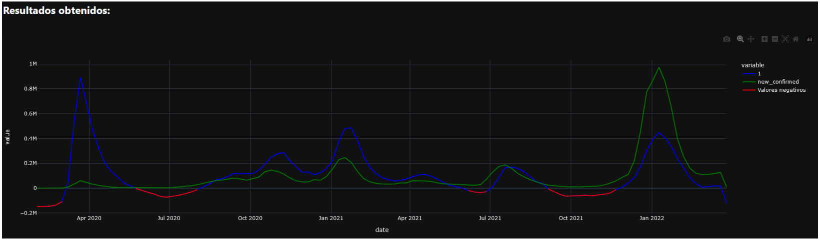
Limpieza de datos



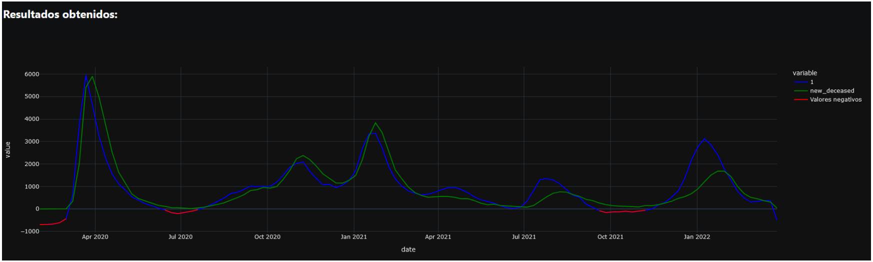
- Agrupación semanal

Análisis global





Análisis global



- Aumenta el desfase entre la componente y los nuevos fallecidos.
- Nuevas cepas $\Rightarrow$ Mayor transmisión y menor letalidad.

Tabla de contenidos

Metodología de desarrollo incremental

Disposición de prototipos continua a los que añadir funcionalidad en cada secuencia de desarrollo lineal.

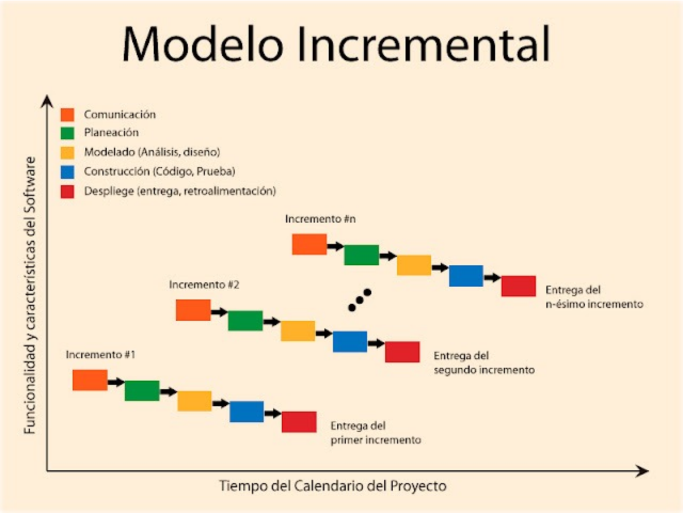
Metodología de desarrollo incremental

Disposición de prototipos continua a los que añadir funcionalidad en cada secuencia de desarrollo lineal.

Ventajas:

- Software operativo desde la primera etapa.
- Flexibilidad a cambios de requisitos.
- Iteraciones reducidas $\Rightarrow$ Facilidad en depuración y test
- Fases solapadas

Metodología de desarrollo incremental



Análisis de requisitos

Requisitos funcionales:

- **RF1**: Se deben poder visualizar las distintas series temporales de datos, ya sea de forma diaria o semanal, con o sin normalizar.
- **RF2**: El usuario debe poder escoger los rangos temporales en los que explorar los datos.
- **RF3**: El sistema debe permitir activar o desactivar las variables que se quieren visualizar.
- **RF4**: El sistema debe permitir seleccionar grupos de variables comparables.
- **RF5**: El sistema debe permitir descargar los datos de la pandemia que se muestran.
- **RF6**: El sistema debe permitir elegir a que rango aplicar el método de análisis.
- **RF7**: Los resultados se representarán en una nueva gráfica independiente. Además, se mostrará la varianza explicada y un breve informe.
- **RF8**: El sistema debe permitir comparar el resultado con las variables de partida.
- **RF9**: El sistema debe permitir descargar la primera componente principal resultante.

Análisis de requisitos

Requisitos no funcionales:

- **RNF1**: La gráficas deben ser interactivas, permitir zooms y translaciones.
- **RNF2**: Las gráficas deben actualizarse en tiempo real.
- **RNF3**: Se mostrarán unas indicaciones del funcionamiento de la aplicación.
- **RNF4**: Se debe poder utilizar en un navegador web.

Requisitos de información:

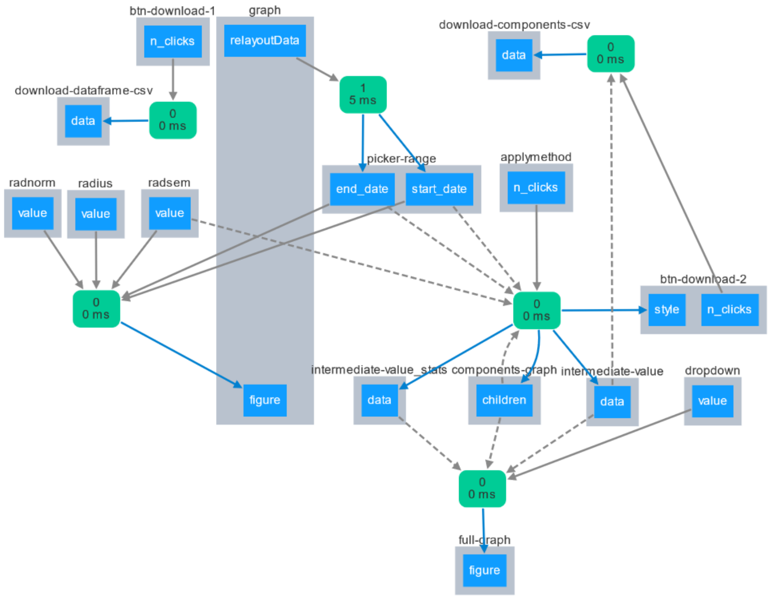
- **RI1**: Los ficheros generados por la aplicación tanto con los datos originales como los resultantes deben estar en formato .CSV.
- **RI2**: Los gráficas deben poder exportarse en cada estado en formato .PNG.

Recursos software utilizados

python™

plotly | Dash

Sistema de callbacks



Despliegue en GCP

- Modelo de computación *cloud*
- GCP App Engine

```
dash==2.8.1
pandas==1.3.4
numpy==1.20.3
scikit-learn==0.24.2
plotly==5.9.0
gunicorn
```

Fichero de dependencias requirements.txt

```
service: default
runtime: python37

basic_scaling:
    max_instances: 2
    idle_timeout: 10m

resources:
    cpu: 1
    memory_gb: 1
    disk_size_gb: 2

entrypoint: gunicorn -b 0.0.0.0:8080 pcatfg:server
```

Fichero de infraestructura app.yaml

https://pca-tfg-igarachv.oa.r.appspot.com

Conclusiones

Objetivos cumplidos:

- Método para extraer una variable significativa del estado de la pandemia.
- Aplicación a los datos de España.
- Desarrollo y despliegue de la aplicación web.

Conclusiones

Objetivos cumplidos:

- Método para extraer una variable significativa del estado de la pandemia.
- Aplicación a los datos de España.
- Desarrollo y despliegue de la aplicación web.

Posible desarrollo futuro:

- Relacionar con el modelo epidemiológico SIR.
- Libertad de datos en la aplicación.

Referencias

[1] Alfaro-Martínez, J. J., García del Pozo, J. S., et al. (2023). Estudio de la incidencia de COVID-19 en España y su relación geográfica provincial. Journal of Healthcare Quality Research.

[2] Fernández-Granda, C. (2019). Principal Component Analysis. En Mathematical Tools for Data Science. NYU's Center for Data Science. Spring 2019.

[3] Hodge, M., UK ONS Data Science Campues. (2022). Guía de despliegue de Dash en GCP. Disponible en:https://datasciencecampus.github.io/deploy-dash-with-gcp/. Accedido el 2023-06-11.

[4] Jolliffe, I. T. (2002). Principal Component Analysis. Springer Series in Statistics. Springer, New York, NY, 1, 1-6.

[5] Luque Martínez, T. (2000). Técnicas de análisis de datos en investigación de Mercados. Editorial Pirámide, 2, 40-58.

[6] Mathieu, E., Ritchie, H., et al. (2020). Coronavirus Pandemic Data (COVID-19). Repositorio Covid-19 OWID. Disponible en: https://ourworldindata.org/coronavirus. Accedido el 2022-12-16.

[7] Ortiz, M., Modelos de Desarrollo de Software. . Disponible en: http://isw-udistrital.blogspot.com/2012/09/ingenieria-de-software-i.html. Accedido el 2023-04-06.

[8] Pressman, R. (2010). Software Engineering: A Practitioner's Approach. Boston: McGraw Hill. 48–49.

[9] Wahltinez, O., et al. (2020). COVID-19 Google Open-Data: curating a fine-grained, global-scale data repository for SARS-CoV-2. Repositorio Covid-19 Open Data. Disponible en: https://github.com/GoogleCloudPlatform/covid-19-open-data/tree/main . Accedido el 2022-12-15.

Gracias por su atención.