EST-46114 Métodos Multivariados y Datos Categóricos

Sesion 05 - Analisis de Componentes Principales - Parte 1/3

Juan Carlos Martinez-Ovando

Maestria en Ciencia de Datos

ITAM

Objetivos

- Estudiaremos los fundamentos del analsis de componentes principales muestrales (SPCA) y su forma de calculo.
- Estudiaremos la intuicion geometrica / matricial destras del SPCA.
- Realizaremos una ilustracion.

Usos

El *SPCA* es una de las tecnicas de *ortogonalizacion* y *reduccion de dimensionalidad* mas empleada en la practica.
Se emplea en diferentes contextos, por ejemplo:

a. Ortogonalizar datos (matrices) para regresion (***)
b. Definir modelos de regresion donde $Y = M\beta + \varepsilon$, donde M es un conjunto de componentes de componentes principales de los datos originales X
c. Procesamiento y reconstruccion de se~nales e imagenes
d. Construccion de indices (sobre todo en las ciencias sociales)
e. Otras muchas aplicaciones...

Intuicion

- El SPCA es una tecnica que comunmente se emplea para realizar **analisis exploratorio de datos** y **reduccion de dimensionalidad**
- Usualmente, SPCA se asocia con el siguiente procedimiento

Tranformar un conjunto de observaciones multidimensionales y correlacionadas entre si, reemplazandolas por un numero de NUEVAS observaciones de dimension menor y no correlacionadas

- La interpretacion anterior no es enteramente correcta, pues SPCA es mas bien un procedimiento para la **ortogonalizacion de datos correlacionados** en la que

las nuevas variables son generadas mediante un procedimiento de rotacion de los datos originales, con cada componente dirigido hacia la dimension de los datos originales con la mayor variabilidad

SPCA y EDA

Supongamos que tenemos un conjunto de datos asociado con p variables y n observaciones,

$$\boldsymbol{X} = (x_{j1}, \ldots, x_{jp})_{j=1}^{n}.$$

- La interrelacion entre estas p variables podria visualizarse usando $\frac{p(p-1)}{2}$ diagramas de dispersion (esto solo para visualizar relaciones 1:1 entre las p variables)
- **Con SPCA puede obtenerse:**

(a) un simplificacion de la informacion contenida en las variables originales, pero en una menor dimension,

(b) un resumen de las dimensiones/variables originales que comparan la mayor cantidad de informacion, asi como la cuantificacion relativa a tal informacion para cada variable

Enfoques del SPCA

SPCA mapea la informacion de un espacio p-dimensional a otro de la misma dimension, i.e.

$$\text{PCA} : (X_1, \ldots, X_p) \to (Z_1, \ldots, Z_p)$$

–X_js correlacionadas, mientras que Z_js no correlacionadas–
Mapeo El mapeo a Z_j se obtiene como una **proyeccion lineal**,

$$Z_j =< \phi, X >= \phi' X = \sum_{i=1}^{p} \phi_{ji} X_i,$$

con **pesos normalizados**, i.e. $< \phi, \phi >= 1$, siendo $(\phi_{j1}, \ldots, \phi_{jp})$ el **vector de cargas** del j-esimo componente.
Formas de calculo del SPCA

1. Problema de optimizacion
2. Descomposicion singular
3. Descomposicion espectral

SPCA | Problema de optimizacion

Considerando un conjunto de datos en forma matricial $\boldsymbol{X} = (X_{ij})_{i=1,j=1}^{n,p}$ de dimension $n \times p$ (considerando $n >> p$)
La construccion del SPCA se obtiene como la solucion de un **problema de optimizacion**, donde:

- El **primer componente principal** se define como la solucion a

$$\max_{\phi_{11},\ldots,\phi_{1p}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{1j} x_{ij} \right)^2 \right\} \text{ s.a. } \sum_{j=1}^{p} \phi_{1j}^2 = 1.$$

- Los componentes subsecuentes, $\phi_2, \ldots, \phi_p$ se definen solucion la funcion objetivos sujeto a las restricicones anidadas, $< \phi_j, \phi_k >= 0$ para $k = 1, \ldots, (p-1)$.

Nota: Se supone que las X_{ij}s son estandarizadas, de manera que $\frac{1}{n} \sum_{i=1}^{n} X_{ij} = 0$.

SPCA | Descomposicion singular

Partiendo de la matris de datos, $\boldsymbol{X}$, podemos considerar la descomposicion en valores singulares dado por,

$$\boldsymbol{X}_{(n\times p)} = \boldsymbol{U}_{(n\times n)}\boldsymbol{S}_{(n\times p)}\boldsymbol{V}_{(p\times p)},$$

donde

- $\boldsymbol{S}_{(n\times p)}$ es una matriz rectangular con valores singulares en la diagonal correspondientes a $(\boldsymbol{X}'\boldsymbol{X})$ y $(\boldsymbol{X}\boldsymbol{X}')$
- $\boldsymbol{V}_{(p\times p)}$ es una matrixz ortogonal con vectores columnas correspondientes a los `eigenvectores derechos` de $\boldsymbol{S}_{(n\times p)}$ (a.k.a. los vectores ϕ_js)

SPCA | Descomposicion espectral

Podemos considerar la matriz muestral asociada con $\boldsymbol{X}$, i.e.

$$\boldsymbol{\Sigma} = \frac{1}{n}(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Junto con ella, la descomposicion espectral,

$$\boldsymbol{\Sigma} = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^{-1},$$

en `eigenvalores`, $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_p\}$, y `eigenvectores`, $\boldsymbol{\Phi} = (\phi_1, \ldots, \phi_p)$.

SPCA | Supuestos

- **Linealidad:** La nueva base es una combinación lineal de la base original
- **Suficiencia:** La media y la varianza son estadísticas suficientes. PCA supone que estas estadísticas describen totalmente la distribución de los datos a lo largo del eje (es decir, la distribución normal).
- **Variablidad:** Las grandes variaciones tienen una dinámica importante: alta varianza significa señal, baja varianza significa ruido. Esto significa que el PCA implica que la dinámica tiene una alta relación señal / ruido.
- **Ortonormalidad:** Los componentes que se contruyen seran ortonormales, a.k.a. estocastimanete independientes y de norma unitaria

Suponemos que la matriz de covarianza muestral de X está dada por $\frac{1}{n}(X'X)$.

SPCA | Intuicion 1/13

Consideremos el siguiente conjunto de datos

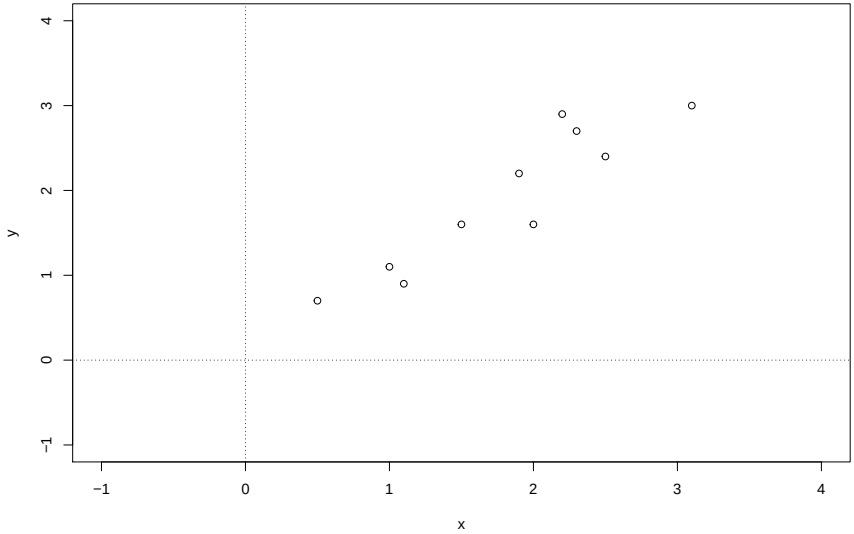
```
x <- c(2.5,.5,2.2,1.9,3.1,2.3,2,1,1.5,1.1)
y <- c(2.4,0.7,2.9,2.2,3.0,2.7,1.6,1.1,1.6,.9)
print("Covarianza")
```

```
## [1] "Covarianza"
```

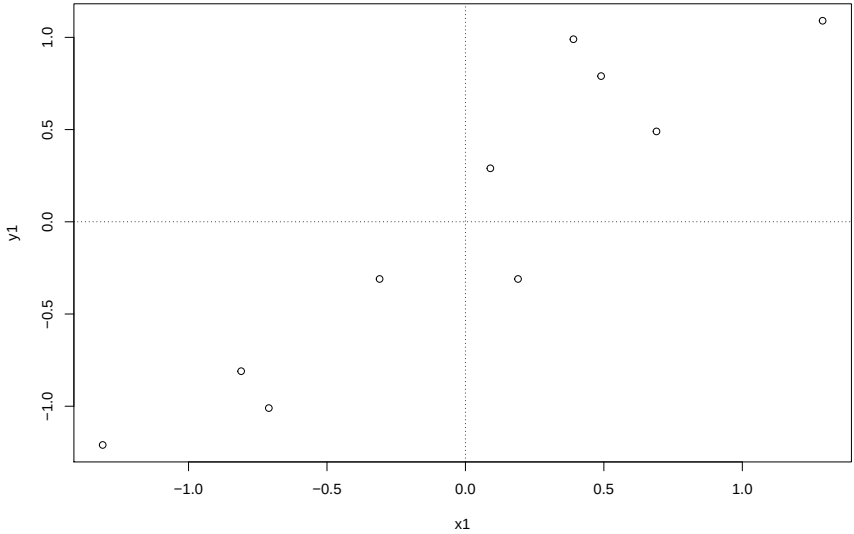
```
cov(x,y)
```

```
## [1] 0.6154444
```

SPCA | Intuicion 2/13



SPCA | Intuicion 3/13



```
## [1] "Sample covariance matrix"

##           [,1]      [,2]
## [1,] 0.6165556 0.6154444
## [2,] 0.6154444 0.7165556

## [1] "Determinant"

## [1] 0.06302444
```

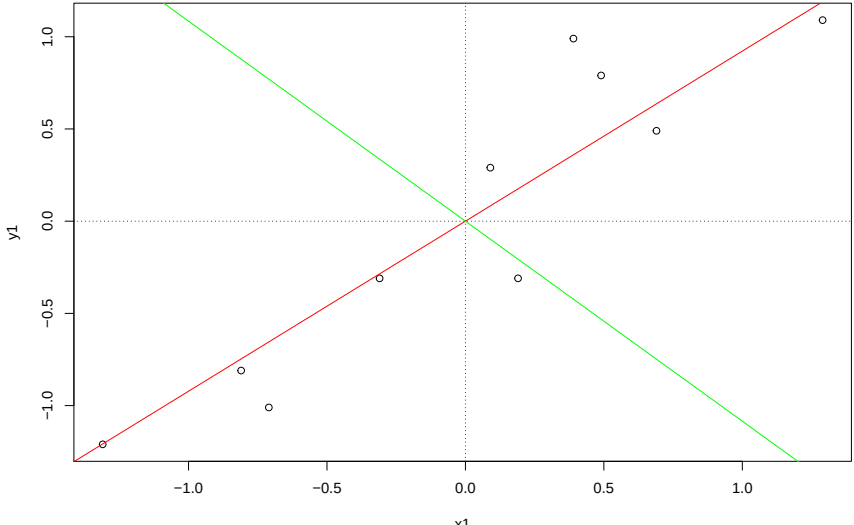
```
## eigen() decomposition
## $values
## [1] 1.2840277 0.0490834
##
## $vectors
##           [,1]       [,2]
## [1,] 0.6778734 -0.7351787
## [2,] 0.7351787  0.6778734

## [1] "Producto interior"

##      [,1]
## [1,]    0
```

SPCA | Intuicion 6/13

Visualizamos la descomposicion espectral. La linea roja representa el primer componente principal, la linea verde representa el segundo componente principal.



- El **primer eigenvector** (linea roja) parece un ajuste lineal, que nos muestra como se relaciona con los datos, pero el otro vector propio no parece estar relacionado con los datos . . . Por que?
- Si nos fijamos en los **eigenvectores**, el primero es mucho mayor que el segundo: el **eigenvalor** mas grande identifica la componente principal del conjunto de datos.

Una vez encontrados los **eigenvectores**, debemos ordenarlos coin base en sus **eigenvalores**. Esto nos da los componentes por orden de importancia! Podemos decidir ignorar los componentes con menor `significado/variabilidad capturada`, sin embargo, perderemos informacion pero no tanto si sus valores son pequenos.

En nuestro ejemplo en 2D, solo tenemos dos opciones para ordenar los `eigenvalores`. Iluestremos este procedimiento:

```
##            [,1]
## [1,] 0.6778734
## [2,] 0.7351787
```

```
##            [,1]       [,2]
## [1,] 0.6778734 -0.7351787
## [2,] 0.7351787  0.6778734
```

Derivamos el nuevo conjunto de datos transformados. Si X es el conjunto de `datos` originales y P es el vector de atributos, entonces la transpuesta de los datos transformados es XP' define a las nuevas variables:

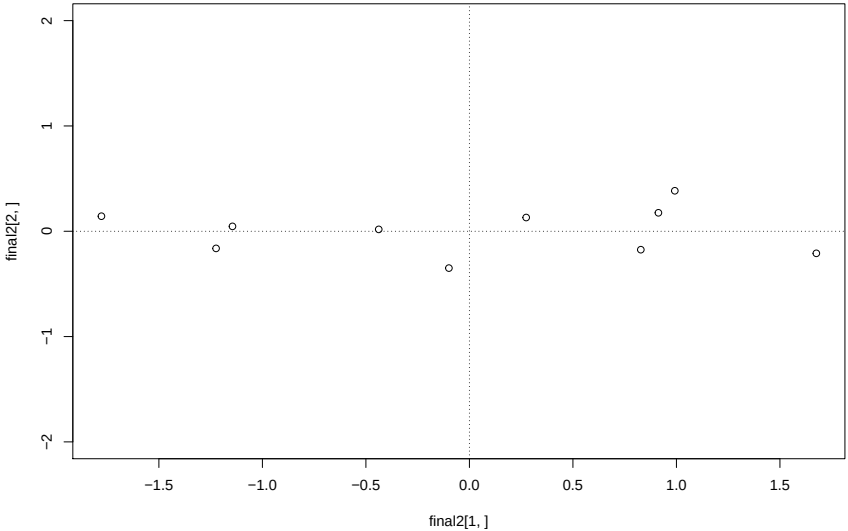
```
##             [,1]       [,2]       [,3]       [,4]       [,5]       [,6]
## [1,]  0.8279702 -1.7775803 0.9921975 0.2742104  1.6758014 0.9129491
## [2,] -0.1751153  0.1428572 0.3843750 0.1304172 -0.2094985 0.1752824
##              [,7]        [,8]        [,9]       [,10]
## [1,] -0.09910944 -1.14457216 -0.43804614 -1.2238206
## [2,] -0.34982470  0.04641726  0.01776463 -0.1626753
```

```
##               [,1]          [,2]
## [1,] 1.284028e+00 4.636018e-17
## [2,] 4.636018e-17 4.908340e-02
```

Estos conjuntos de datos finales son los datos originales en términos de los vectores que elegimos, es decir, ya no estan sobre los ejes x y y originales, sino que utilizan los vectores propios elegidos como su nuevo eje.

```
##            [,1]      [,2]       [,3]       [,4]      [,5]       [,6]
## [1,] 0.8279702 -1.77758 0.9921975 0.2742104 1.675801 0.9129491 -0.09
##            [,8]       [,9]      [,10]
## [1,] -1.144572 -0.4380461 -1.223821
```

SPCA | Intuicion 11/13



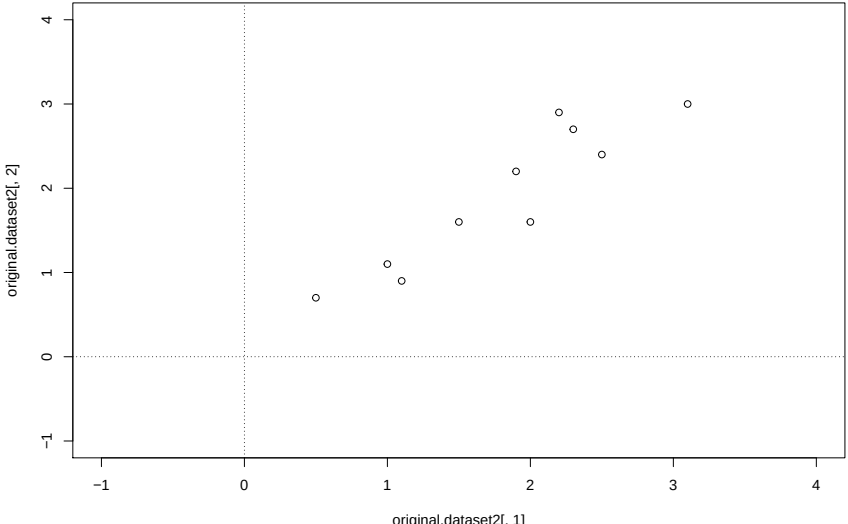
SPCA | Intuicion 12/13

Recuperando los `vectores originales`...

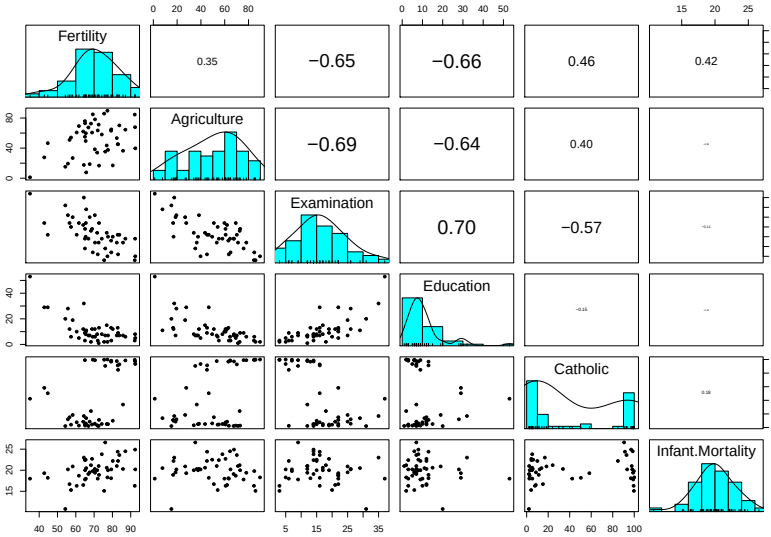
```
##       [,1] [,2]
##  [1,]  2.5  2.4
##  [2,]  0.5  0.7
##  [3,]  2.2  2.9
##  [4,]  1.9  2.2
##  [5,]  3.1  3.0
##  [6,]  2.3  2.7
##  [7,]  2.0  1.6
##  [8,]  1.0  1.1
##  [9,]  1.5  1.6
## [10,]  1.1  0.9
```

Operacion PCA

Recuperando los `vectores originales`...



Ilustracion swiss 1/



Ilustracion swiss 2/

Usando la funcion prcomp de r-base...

```
## Importance of components:
##                             PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     1.7888 1.0901 0.9207 0.66252 0.45225 0.34765
## Proportion of Variance 0.5333 0.1981 0.1413 0.07315 0.03409 0.02014
## Cumulative Proportion  0.5333 0.7313 0.8726 0.94577 0.97986 1.00000
```

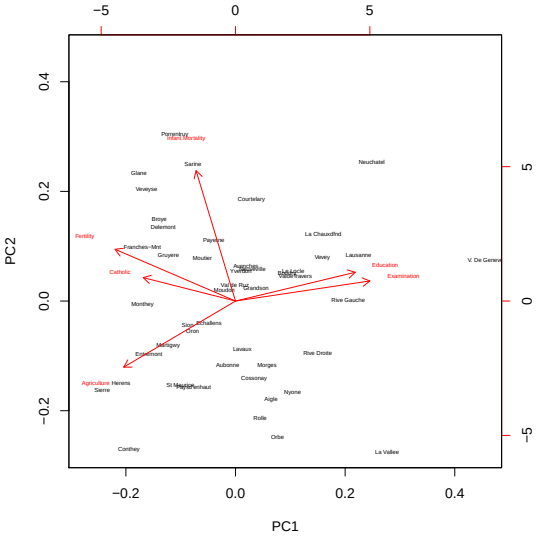
```
## [1] "center"   "rotation" "scale"    "sdev"     "x"
```

Ilustracion `swiss` 3/

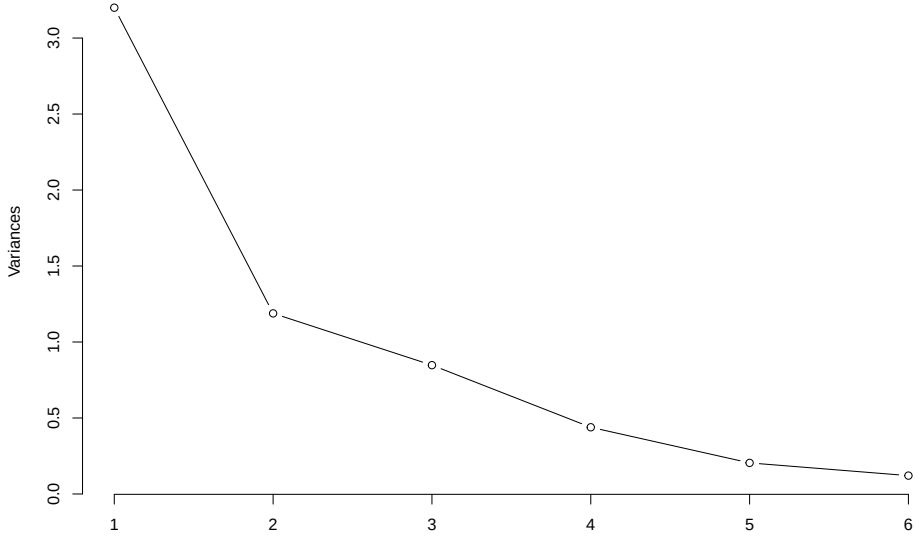
Primeros tres componentes...

```
##                    PC1   PC2   PC3
## Fertility        -0.46  0.32 -0.17
## Agriculture      -0.42 -0.41  0.04
## Examination       0.51  0.13 -0.09
## Education         0.45  0.18  0.53
## Catholic         -0.35  0.15  0.81
## Infant.Mortality -0.15  0.81 -0.16
```

Ilustracion `swiss` 4/



Ilustracion `swiss` 5/



Advertencias

- SPCA **no es invariante ante cambios de escala.**
- SPCA es muestral, por lo que esta sesgado hacia la informacion en $\boldsymbol{X}$ y, en particular, respecto al error de estimacion de la matriz de varianzas y covarianzas, $\boldsymbol{S}$.
- Aun cuando SPCA sea aplicado a un censo de poblacion, subyace la incertidumbre acerca de la estimacion de $\boldsymbol{S}$.
- Las relaciones de asociacion en PCA y SPCA se suponen **lineales solamente**.

Siguiente sesion

- Construir PCA incorporando **incertidumbre epistemica** (a.k.a. error de estimacion de la matriz de covarianzas)
- Extender PCA para **objetos no matriciales** (e.g. `functional PCA for imagenes or documents`)
- Definir PCA para datos `ralos` (i.e. `sparce data`, caso $n << p$)
- Vincular PCA con `analisis de factores` y `analisis de correlacion canonica`

Lectura complementaria

- Everitt & Hothorn (2006) “An Introduction to Multivariate Analysis with R” - Cap. 6
- Izenman (2008) “Modern Multivariate Statistical Techniques” - Sec. 7.2-7.5

Lectura extendida

- Johnstone & Lu (2004) *Sparse Principal Components Analysis*. `est46114_s05_suplemento.pdf`

Table of Contents