EST-46114 Métodos Multivariados y Datos Categóricos

Sesion 06 - Analisis de Componentes Principales - Parte 2/3

Juan Carlos Martinez-Ovando

Maestria en Ciencia de Datos

ITAM

Objetivos

- Estudiaremos el procedimiento inferencial asociado con PCA (extension del SPCA).
- Realizaremos una ilustracion del procedimiento inferencial.

Preambulo

PCA es usado en muchas ocaciones para producir estadisticas e indicadores oficiales, e.g.:

- Estadiscas demograficas (Mexico)
- Incidencia criminal (Nigeria)
- Aprovechamiento escolar (Colombia, Mexico)
- Habilidades sociales (Argentina)
- Analisis de rendimientos financieros (Internacional)
- Entre muchos otros...

Motivavion

- Reconociendo que muchos indicadores oficiales (incluyendo los construidos con PCA), **Manski (2015)** abogo por la inclusion de `medidas de error` (a.k.a. dispersion) en la publicacion de indicadores estadisticos.
- Esto reconociendo el aspecto inferencial asociado con su construccion.

Revisen la rectura complementaria de esta sesion.

PCA | Recapitulacion

Recodremos, a partir de un conjunto de datos, matriz $\boldsymbol{X}$ de dimension $(n \times p)$, PCA se contruye a partir de la asociacion de la `estructura espectral` de $\boldsymbol{\Sigma}$, la matriz de varianzas y covarianzas subyacente a $\boldsymbol{X}$.
Es decir, los componentes principales de $\boldsymbol{X}$ corresponden a la transformacion

$$c_{j_{(n\times 1)}} = \boldsymbol{X}\boldsymbol{v}_j,$$

para $j = 1, \ldots, p$, donde

$$(e_j, \boldsymbol{v}_j)_{j=1}^{p},$$

con los eigenvalores asociados con $\boldsymbol{X}$ y $\boldsymbol{\Sigma}$.

PCA | Observaciones

- Cuando la matriz $\boldsymbol{X}$ de dimension $(n \times p)$ es una `muestra de datos`, la informacion poblacional acerca de las interacciones de las p variables es incompleta.
- Informacion incompleta poblacional involucra que PCA constriuido a partir de $\boldsymbol{X}$ o $\boldsymbol{\Sigma}$ sea en realidad construido a partir de $\hat{\boldsymbol{\Sigma}}$ (un estimador puntual de Σ).
- Aun cuando $\boldsymbol{X}$ refiera a un censo poblacional, la estructura de codependencia entre las p-dimensiones de estudio es incierta (i.e. la codependencia puede no ser lineal).
- En los escenarios anteriores, se descarta el aspecto temporal/dinamico del PCA implementado.

PCA | Ejemplo SPCA Swiss 1/2

En el caso de los datos Swiss, realizamos SPCA como:

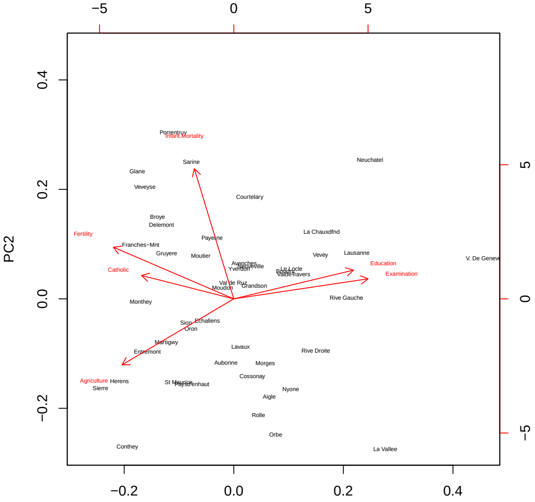
```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     1.7888 1.0901 0.9207 0.66252 0.45225 0.34765
## Proportion of Variance 0.5333 0.1981 0.1413 0.07315 0.03409 0.02014
## Cumulative Proportion  0.5333 0.7313 0.8726 0.94577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

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

```
##                    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
```

PCA | Ejemplo SPCA Swiss 2/2

```
biplot(swiss.pca, cex = 0.4)
```



Planteamiento Pensemos que cada renglon de la matriz $\boldsymbol{X}$ es una observacion de

$$\boldsymbol{X}_{j\cdot} \sim N_p(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\Lambda}),$$

para $j = 1, \ldots, n$.

- Adoptamos el supuesto de `homogeneidad` e `independencia estocastica` en la enunciacion anterior.

Complementariamente
El desconocmiento acerca de $(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ es expresado como:

$$(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \sim \text{N-Wi}(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{m}_0, s_0, a_0, \boldsymbol{B}_0),$$

para ciertos valores de $\boldsymbol{m}_0, s_0, a_0, \boldsymbol{B}_0$.

PCA | Swiss Inferencial 2/

Aprendizaje
Asi, al consolidar las `fuentes de informacion` tenemos:

$$(\boldsymbol{\mu}, \boldsymbol{\Lambda} | \boldsymbol{X}) \sim \text{N-Wi}(\boldsymbol{\mu}, \boldsymbol{\Lambda} | \boldsymbol{m}_n, s_n, a_n, \boldsymbol{B}_n),$$

con $\boldsymbol{m}_n, s_n, a_n, \boldsymbol{B}_n$ expresados analiticamente en la presentacion de la `Sesion 04`.

PCA inferencial
Con base en el aprendizaje realizado, se define para un componente principal $c_{j\cdot(n\times 1)}$ como

$$\begin{aligned} c_{j\cdot} &= \boldsymbol{X} v_j \\ &= \boldsymbol{X} v_j(\mu, \Lambda), \end{aligned}$$

donde $v_j(\mu, \Lambda)$ **hace explicito** que el `eigenvector` asociado desta en funcion del *parametro desconocido* Λ^{-1}.

- Por consiguiente, al ser definidos $(e_j, \boldsymbol{v}_j)_{j=1}^{p}$ en funcion de $\boldsymbol{\Lambda}$ (*aleatoria*), los define como **desconocidos y aleatorios** tambien.
- La distribucion de $(\boldsymbol{e}, \boldsymbol{V}) = (e_j, \boldsymbol{v}_j)_{j=1}^{p}$ esta definida por $\text{N-Wi}(\boldsymbol{\mu}, \boldsymbol{\Lambda} | \boldsymbol{m}_n, s_n, a_n, \boldsymbol{B}_n)$.

Siguiendo el ultimo punto, el calculo de

$$\mathbb{P}(\boldsymbol{e}, \boldsymbol{V} | \boldsymbol{m}_n, s_n, a_n, \boldsymbol{B}_n),$$

es *dificil de obtener analiticamente*.

- Press (2005) sugiere emplear `distribuciones asintoticas` basadas en el CLT (vea Teoremas 9.3.2 y 9.3.3). Eso solo para la distribucion de los `eigenvalores` $\boldsymbol{e} = (e_1, \ldots, e_p)$, mas no para los `eigenvectores`.
- En la practica, esta distribucion puede aproximarse via muestras de Monte Carlo de N-Wi$(\boldsymbol{\mu}, \boldsymbol{\Lambda} | \boldsymbol{m}_n, s_n, a_n, \boldsymbol{B}_n)$ de acuerdo al siguiente algoritmo.

Algotimo

0. Actuarlizar la distrinucion de los parametros (μ, Λ) con base en $\boldsymbol{X}$. Fijar el tamanio de la muestra simulada M.

Para $m = 1, \ldots, M$,

1. Generar los `pseudodatos aleatorios` $(\boldsymbol{\mu}^{(m)}, \boldsymbol{\Lambda}^{(m)})$ como
$$\begin{aligned} \boldsymbol{\Lambda}^{(m)} &\sim \mathrm{Wi}(\boldsymbol{\Lambda}|a_n, \boldsymbol{B}_n), \\ \boldsymbol{\mu}^{(m)}|\boldsymbol{\Lambda}^{(m)} &\sim \mathrm{N}(\boldsymbol{\mu}|\boldsymbol{m}_n, s_n\boldsymbol{\Lambda}^{(m)}). \end{aligned}$$
2. Generar la `descomposicion espectral`
$$(e_j^{(m)}, \boldsymbol{v}_j^{(m)})_{j=1}^p,$$
de $\boldsymbol{\Sigma}^{(m)} = \left(\boldsymbol{\Lambda}^{(m)}\right)^{-1}$.
3. Generar los componentes principales $\boldsymbol{C}^{(m)} = (\boldsymbol{c}_{1\cdot}^{(m)}, \ldots, \boldsymbol{c}_{p\cdot}^{(m)})$ como
$$\boldsymbol{c}_{j\cdot}^{(m)} = \boldsymbol{X}\boldsymbol{v}_j^{(m)}.$$

Comentarios

- Las muestras

$$\left(e_j^{(m)}, \boldsymbol{v}_j^{(m)}\right)_{j=1,m=1}^{p,M}$$

  son iid de

$$\mathbb{P}(\boldsymbol{e}, \boldsymbol{V} | \boldsymbol{m}_n, s_n, a_n, \boldsymbol{B}_n).$$

- Las muestras

$$\left(\boldsymbol{C}^{(m)}\right)_{m=1}^{M} = (\boldsymbol{c}_{1\cdot}^{(m)}, \ldots, \boldsymbol{c}_{p\cdot}^{(m)})_{m=1}^{M}$$

  son iid de

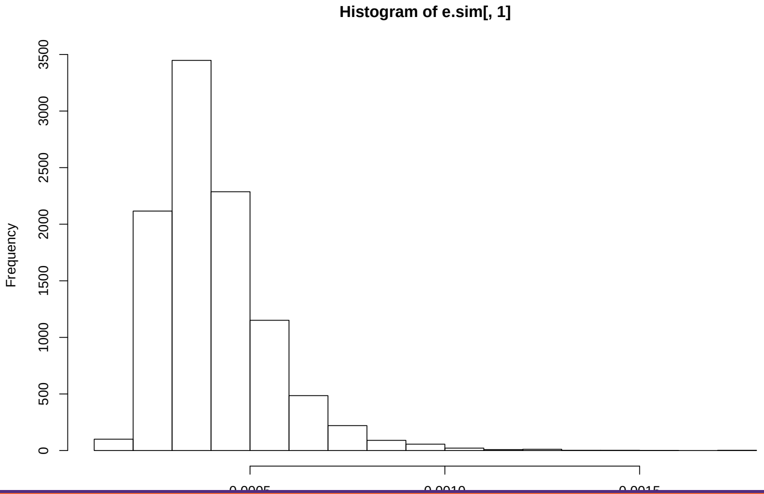
$$\mathbb{P}(\boldsymbol{C} | \boldsymbol{m}_n, s_n, a_n, \boldsymbol{B}_n).$$

PCA | Swiss Inferencial 7/

```
M <- 10000
mu.sim <- matrix(NA,nrow=M, ncol=ncol(data))
Lambda.sim <- array(NA,dim=c(M,ncol(data),ncol(data)))
e.sim <- matrix(NA,nrow=M, ncol=ncol(data))
V.sim <- array(NA,dim=c(M,ncol(data),ncol(data)))
C.sim <- array(NA,dim=c(M,nrow(data),ncol(data)))
m <- 1; X <- as.matrix(data)
for(m in 1:M){
  # Simulacion (mu,Lambda)
  Lambda.sim[m,,] <- rWishart(1, output[[3]], output[[4]])
  mu.sim[m,] <- mvrnorm(1, mu=output[[1]], Sigma=solve(output[[2]]*Lamb
  # Simulacion (e,V) + C
  eigen_aux <- eigen(solve(Lambda.sim[m,,]))
  e.sim[m,] <- eigen_aux$values
  V.sim[m,,] <- eigen_aux$vectors
  C.sim[m,,] <- X %*% V.sim[m,,]
}
```

Inferencia sobre el eigenvenctor e_1

```
hist(e.sim[,1])
```



PCA | Swiss Inferencial 9/

Inferencia sobre el `eigenvenctor` e_1

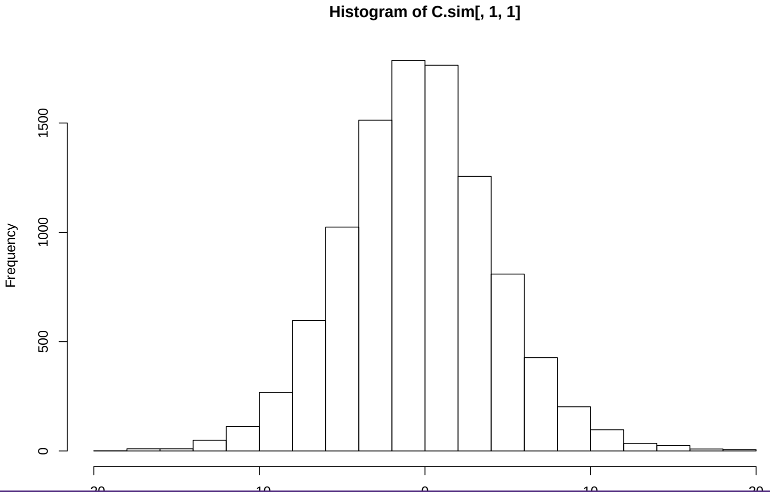
```
summary(e.sim[,2])
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## 5.912e-05 1.277e-04 1.545e-04 1.630e-04 1.895e-04 4.855e-04
```

PCA | Swiss Inferencial 10/

Inferencia sobre el componente principal uno $c_{i,1}$ de la observacion $i = 1$

```
hist(C.sim[,1,1])
```



PCA | Swiss Inferencial 11/

Inferencia sobre el componente principal uno $c_{i,1}$ de la observacion $i = 1$

```
summary(C.sim[,1,1])
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -19.1027  -3.3611   -0.3913   -0.3644    2.5175   19.4049
```

Ejercicio

Consideren el archivo `est46114_s06_data.xls` con los tipos de cambio reales (respecto a USD) de varias economias (periodo 1970-2010).

1. Implementen el *procedimiento inferencial PCA* considerando distribuciones iniciales no informativas para $\boldsymbol{\mu}, \boldsymbol{\Lambda}$.
2. Reporten que economia tiene el mayor peso esperado en la descomposicion PCA.
3. Reporten que economia tiene la mayor consistencia en estimacion de los c_js correspondientes.

Lecturas complementarias

- Maski (2015) “Communicating Uncertainty in Official Economic Statistics”. `est46114_s06_suplemento.pdf`
- Press (2005) *Applied Multivariate analysis*, Caps. 3 y 5.

Table of Contents