EST-46114 Métodos Multivariados

Sesión - Examen Parcial 1 - Soluciones

Juan Carlos Martínez-Ovando

Maestría en Ciencia de Datos

Pregunta 1.a.

El análisis de componentes principales poblacionales (PPCA) es equivalente al análisis de componentes principales muestrales (SPCA).

R: No. Mientas que PPCA describe la covarialidad poblacional de N individuos a traves de la descomposicion espectral de la covariabilidad de p-mediciones, $\boldsymbol{\Lambda} = diag\{\lambda_1, \ldots, \lambda_p\}$ *y* $\boldsymbol{V} = (\boldsymbol{v}'_1, \ldots, \boldsymbol{v}'_p)$*, considerando que* ***no existe incertidumbre en*** $\boldsymbol{\Lambda}$ *y* $\boldsymbol{V}$*,*

en SPCA el desconocimiento acerca de la covarialidad poblacional es ***descrito muestralmente*** *con* $\boldsymbol{\Lambda}_n$ *y* $\boldsymbol{V}_n$*, calculadas con n-observaciones, las cuales pueden:*

- transferir sesgos en los datos
- trasnferir incertidumbre derivada de informacion incompleta
- transferir errores dervidados del procedimiento de estimacion (*metodológico* y *computacional*)

Pregunta 1.b.

Los resultados de aplicar PCA sobre estas dos matrices de covarianzas muestrales,

$$\Sigma_1 = \begin{bmatrix} 1 & 0.1 & 0.3 \\ 0.1 & 2 & 0.2 \\ 0.3 & 0.2 & 3 \end{bmatrix}, \text{ y } \Sigma_2 = \begin{bmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 3 & 0.1 \\ 0.3 & 0.1 & 2 \end{bmatrix},$$

respectivamente, son equivalentes.

No. Σ_1 y Σ_2 son ambas muy parecidas. Ambas son matrices positivo definidas, en principio.

Pregunta 1.b.

Bajo lo anterior, tendran asociada la misma descomposicion espectral si son derivadas de rotacion de las columnas originales. Ese no es el caso, ya que si

$$\Sigma_1 = \begin{bmatrix} \sigma_{xx} & & \\ \sigma_{yx} & \sigma_{yy} & \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{bmatrix},$$

observamos que Σ_2 ***no esta asociada con la rotacion inducida por intercambiar*** *las variables* y *y* z*, i.e.*

$$\Sigma_2 \neq \begin{bmatrix} \sigma_{xx} & & \\ \sigma_{zx} & \sigma_{zz} & \\ \sigma_{yx} & \sigma_{yz} & \sigma_{yy} \end{bmatrix}.$$

Pregunta 1.b.

Numericamente tenemos:

```
Sigma1 <- matrix(c(1,0.1,0.3,
                   0.1,2,0.2,
                   0.3,0.2,3),ncol=3,nrow=3)
Sigma2 <- matrix(c(1,0.2,0.3,
                   0.2,3,0.1,
                   0.3,0.1,2),ncol=3,nrow=3)
Sigma1_eigen <- eigen(Sigma1)
Sigma2_eigen <- eigen(Sigma2)
```

Pregunta 1.b.

Eigenvalores

```
Sigma1_eigen$values
```

```
## [1] 3.0856729 1.9631139 0.9512132
```

```
Sigma2_eigen$values
```

```
## [1] 3.0365310 2.0598371 0.9036319
```

Pregunta 1.b.

Eigenvectores

```
Sigma1_eigen$vectors
```

```
##            [,1]        [,2]        [,3]
## [1,] -0.1487448 -0.03928735  0.98809489
## [2,] -0.1923881 -0.97896799 -0.06788596
## [3,] -0.9699803  0.20019537 -0.13805793
```

```
Sigma2_eigen$vectors
```

```
##            [,1]       [,2]        [,3]
## [1,] -0.1156563 -0.2422244  0.96330212
## [2,] -0.9849427  0.1534267 -0.07967505
## [3,] -0.1284970 -0.9580123 -0.25632188
```

Pregunta 1.c.

Los resultados del modelo de factores (FA) son invariantes a la forma en cómo se modela la media de todas dimensiones originales en un conjunto de datos.

No. Recordemos que en modelo FA, la incertidumbre de $\boldsymbol{X} = (X_1, \ldots, X_p)$ *se describe con una distribución gaussiana p-dimensional, donde la* ***matriz de varianzas y covarianzas*** *es de la forma*

$$\begin{aligned}\boldsymbol{\Omega} &= \boldsymbol{\Lambda}'\boldsymbol{\Lambda} + \boldsymbol{\Sigma}, \\ &= \mathbb{E}_{N_p}\left\{(\boldsymbol{X} - \boldsymbol{\mu})^2\right\},\end{aligned}$$

siendo

$$\boldsymbol{\mu} = \mathbb{E}_{N_p}(\boldsymbol{X}).$$

Pregunta 1.c.

En la práctica, en el caso del modelo FA aplicado a las serie de datos de tipos de cambio, vimos que la descomposición en factores está determinada fundamentalmente por la forma de calcular/modelar los niveles medios obsertvacion a observacion.

Pregunta 1.d.

Realizar predicciones de $\boldsymbol{X} = (X_1, \ldots, X_p)$, no observado, con base en datos $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, donde $\boldsymbol{x}_n = (x_{1n}, \ldots, x_{pn})$ para $n = 1, \ldots, N$, empleando PCA reduce incertidubre respecto a realizar la predicción empleando un modelo gaussiano p-dimensional.

No necsariamente. Tenemos dos casos:

Caso 1) *Si la estructura de variabilidad en las p-dimensiones es conocida (ya sea conocida o conocida nula), la prediccion de las p- variables podria realizarse por separado y calcular la prediccion conjunta a través de la transformacion afín,*

$$\boldsymbol{\Lambda}'\boldsymbol{X}.$$

Caso 2) *Si la estructura de correlacion es desconocida, el esfuerzo predictivo de un metodo y otro es equivalente. Aunque, el esfuerzo computacional en la prediccion via PCA es más elevado.*

Pregunta 1.e.

El resultado en el que después de aplicar PCA a un conjunto de datos se obtiene que la estructura de los datos originales y la de los componentes principales resultante es la misma, es evidencia exclusiva que los datos originales son ortogonales.

No necesariamente.

- En el caso **ortogonal**, desde luego, esta aseveracion se cumple.
- En el caso funcional, donde las relaciones de las p-variables originales tienen una estructura compleja (*no euclidiada*, como textos o imágenes), aplicar PCA involucra solo la rotacion de las variables originales.

Pregunta 1.e.

El resultado en el que después de aplicar PCA a un conjunto de datos se obtiene que la estructura de los datos originales y la de los componentes principales resultante es la misma, es evidencia exclusiva que los datos originales son ortogonales.

No necesariamente.

- En el caso **ortogonal**, desde luego, esta aseveracion se cumple.
- En el caso funcional, donde las relaciones de las p-variables originales tienen una estructura compleja (*no euclidiada*, como textos o imágenes), aplicar PCA involucra solo la rotacion de las variables originales.

Pregunta 1.f.

La matriz de cargas $\mathbf{\Lambda}^{(k)}$ del modelo FA con k factores, aplicado a un vector p-dimensional de mediciones, incluye toda la información acerca de la estructura de correlación de las p mediciones.

*No. La matriz $\mathbf{\Lambda}$ incluye informacion relevante respecto al comovimiento de las p-variables, pero no respecto a la **escala** marginal y agregada.*

*La **matriz de covarianzas** es,*

$$\mathbf{\Omega} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Sigma},$$

mientras que la matriz de correlaciones es

$$\boldsymbol{\rho} = (\rho_{kj})_{k,j=1}^{p}$$

con

$$\rho_{kj} = \frac{\omega_{kj}}{(\omega_{jj}\omega_{kk})^{1/2}}.$$

Pregunta 1.g.

Si el número de componentes retenidos aumenta, por ejemplo, de m a $(m+1)$, las primeras m componentes del SPCA no habran de cambiar. Este es también el caso en el modelo de análisis de factores (FA).

No. En SPCA la reducción de dimensionalidad es arbitraria y no tiene forma de incorporar caso-a-caso la pérdida de información inducida por la reducción de dimensionalidad.

En el modelo FA, se incorpora impícitamente a través de $\mathbf{\Lambda}^{(m)}$ y $\mathbf{\Sigma}^{(m)}$ la cantidad de información asociada con la reducción de p a m dimensiones $(m << p)$.

Por lo que no existe asociación directa entre los componentes FA_ms anidados.

Pregunta 4.a.

En SPCA, ¿cómo podemos recuperar la matriz de covarianzas muestral a partir de los *eigenvectores* y *eigenvectores* muestrales?

> *Si* $\boldsymbol{\Lambda} = diag\{\lambda_1, \ldots, \lambda_p\}$ *y* $\boldsymbol{V} = (\boldsymbol{v}_1', \ldots, \boldsymbol{v}_p')$ *son las matrices de* $eigenvalores$ *y* $eigenvectores$ *asociados con SPCA, entonces*
>
> $$\boldsymbol{\Omega} = \boldsymbol{V}'\boldsymbol{\Lambda}\boldsymbol{V}.$$

Pregunta 4.b.

Si $\boldsymbol{X} = (X_1, \ldots, X_p)$ denota un vector p-dimensional en escala original sobre el cual se aplica el modelo FA con k factores fijos ($k << p$). Describe el efecto que tendrá sobre los factores el reescalar cada columna de $\boldsymbol{X}$ por $c_1, \ldots, c_p$ valores fijos positivos.

Si definimos la matriz

$$\boldsymbol{C} = diag\{c_1, \ldots, c_p\},$$

el vector reescalado queda expresado como

$$\boldsymbol{Y} = \boldsymbol{C}\boldsymbol{X}.$$

Si $\boldsymbol{\Lambda}^{(m)}$ y $\boldsymbol{\Sigma}^{(m)}$ son las matrices de cargas y dispersiones del modelo FA m-dimensional para $\boldsymbol{X}$, entonces, al aplicar la transformación afín en $\boldsymbol{C}$ sobre

$$\boldsymbol{X}|\boldsymbol{f}^{(m)} \sim \mathrm{N}_p(\boldsymbol{x}|\boldsymbol{\Lambda}^{(m)}\boldsymbol{f}^{(m)}, \boldsymbol{\Sigma}^{(m)}),$$

por lo que

$$\boldsymbol{\Omega}_{\boldsymbol{Y}} = \boldsymbol{C}\boldsymbol{\Lambda}^{(m)}\boldsymbol{\Lambda}^{(m)'}\boldsymbol{C} + \boldsymbol{C}\boldsymbol{\Sigma}^{(m)}\boldsymbol{C}.$$

Pregunta 4.b.

Es decir, las matrices de cargas y dispersiones para $\boldsymbol{Y}$ *son ahora,*

$$\begin{aligned} \boldsymbol{\Lambda}_{\boldsymbol{Y}}^{(m)} &= \boldsymbol{C}\boldsymbol{\Lambda}^{(m)} \\ \boldsymbol{\Sigma}_{\boldsymbol{Y}}^{(m)} &= \boldsymbol{C}\boldsymbol{\Sigma}^{(m)}\boldsymbol{C}. \end{aligned}$$

Asi, FA es prácticamente no afectado por cambio de escala en las variables originales.

Es decir, el modelo FA sobre matrices de correlación y sobre matrices de covarianzas arroja resultados análogos. No así PCA, por cierto.

Pregunta 4.c.

Desarrolla la expresión para la matriz de covarianzas en el modelo FA si se considera que los factores puedan estar correlacionados.

Este es un problema importante. Lo vimos parcialmente en el caso de volatilidad estocástica; ***pero lo examinaremos con detalle el martes entrante****.*

Especificacion, para $\boldsymbol{X}_t$*, con* $t = 1, \ldots, T$*, y* m *fijo,*

$$\begin{aligned} \boldsymbol{X}_t | \boldsymbol{f}_t^{(m)} &\sim N_p(\boldsymbol{x} | \boldsymbol{\Lambda}^{(m)} \boldsymbol{f}_t^{(m)}, \boldsymbol{\Sigma}^{(m)}) \\ \boldsymbol{f}_t &\sim N_m(\boldsymbol{f} | \boldsymbol{0}, \boldsymbol{G}^{(m)}), \end{aligned}$$

con $\boldsymbol{G}^{(m)}$ $(m \times m)$ *no diagonal, pero simética positivo definida.*

Pregunta 4.c.

La forma de la relación es

$$\boldsymbol{\Omega} = \boldsymbol{\Lambda}^{(m)} \boldsymbol{G}^{(m)} \boldsymbol{\Lambda}^{(m)\prime} + \boldsymbol{\Sigma}^{(m)}.$$

Vean:

- `est46114_2019a_parcial_1_soluciones_auxiliar2.pdf` (artículo)
  Referencia: Tucker (1940) "The role of correlated factors in factor analysis", *Psychometrika*.
- `est46114_2019a_parcial_1_soluciones_auxiliar3.pdf` (desarrollo)

Preguntas 2 y 3

Revisen laboratorio `markdown`...

`est46114_2019a_parcial_1_soluciones_auxiliar1.pdf`

Table of Contents