Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística
Programa Académico de Estadística
Estadística Aplicada III
Cali, septiembre de 2018

Laboratorio 2: ACP con datos faltantes

Kevin Steven García[a], Alejandro Vargas[b]

1. Introducción

En el presente informe veremos la aplicación del ACP con datos completos y datos faltantes a la matriz de datos correspondientes a las importaciones de los países sudamericanos provenientes de estados unidos entre 1991 y 2010. El ACP con datos faltantes se realizará a la matriz de datos mencionada con el 5,10,15, y 20 % de datos faltantes, por medio de dos algoritmos, el algoritmo NIPALS y el ACP-EM, se mostraran los resultados obtenidos y se comparará la eficiencia o la imputación obtenida de los dos algoritmos en cuanto a las estimaciones de los datos faltantes con respecto a la matriz de datos original. Además se analizará el comportamiento de los métodos en cuanto a valores y vectores propios obtenidos, porcentaje de Inercia explicado por los ejes, la estructura factorial en la nube de individuos y en la nube de variables, la ortogonalidad en las componentes y la ortonormalidad de los vectores propios obtenidos.

La base de datos sobre la cuál se va a trabajar es la siguiente:

Año	Colombia	Brasil	Chile	Argentina	Ecuador	Peru
1991	44.4	27.2	45.6	20.0	6.0	14.1
1992	75.5	11.8	58.9	22.6	17.8	14.4
1993	110.7	50.6	128.3	17.2	119.4	118.5
1994	80.3	70.6	102.2	15.2	154.9	146.1
1995	81.6	82.3	89.0	35.1	169.4	127.1
1996	76.4	97.4	185.0	51.0	75.5	129.0
1997	32.0	89.5	195.3	31.1	33.4	110.2
1998	55.5	63.1	66.3	24.4	9.7	66.7
1999	74.3	72.6	76.3	28.1	11.2	110.7
2000	84.5	76.2	80.1	29.5	11.8	110.2
2001	87.1	97.4	89.3	51.5	63.1	89.3
2002	89.3	89.5	72.4	40.3	66.3	70.2
2003	70.2	63.1	80.1	60.5	76.3	90.1
2004	90.1	66.3	70.5	39.1	20.0	64.5
2005	60.5	76.3	107.2	31.1	63.4	92.7
2006	140.3	20.0	63.4	50.2	101.2	120.8
2007	120.4	22.6	101.2	51.0	103.1	107.2
2008	130.2	17.2	103.1	42.5	66.7	70.8
2009	110.1	31.1	75.6	25.7	110.7	101.2
2010	120.2	24.4	68.9	60.3	110.2	110.8

[a]Código: 1533173. E-mail: kevin.chica@correounivalle.edu.co
[b]Código: 1525953. E-mail: jose.alejandro.vargas@correounivalle.edu.co

2. ACP datos completos

En el laboratorio 1 se vio paso por paso la aplicación del ACP, por lo tanto se mostraran solo los resultados obtenidos,es decir, los valores y vectores propios u, y las componentes principales Z· u.

Los valores y vectores propios obtenidos para la matriz de correlaciones asociada a la matriz de datos completa son:

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
2.1934092	1.9561781	0.9038789	0.5119470	0.2854407	0.1491461

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
-0.3266572	0.5708984	0.0068014	0.1170875	0.5332445	0.5189071
-0.1588587	-0.6033277	0.2009878	-0.5413708	0.1852719	0.4929051
-0.3381710	-0.4635960	0.0514265	0.7890499	-0.0779079	0.1985066
-0.2898593	0.2066273	0.8863988	-0.0448699	-0.2455986	-0.1589085
-0.5459391	0.1756044	-0.3785943	-0.2216216	-0.6626128	0.1990175
-0.6096157	-0.1470297	-0.1669623	-0.1395665	0.4193810	-0.6192861

Entonces, las componentes serán:

$$C = \begin{pmatrix} 3.57940293 & 0.2200186 & -0.45449358 & 0.07133223 & -0.720973151 & -0.20047529 \\ 2.98416620 & 1.1042398 & -0.46390509 & 0.71180698 & -0.455553845 & 0.19923050 \\ -1.18011107 & -0.0116174 & -1.74610834 & 0.70549306 & 0.310200546 & 0.48646527 \\ -1.55340630 & -0.7593625 & -2.19242499 & -0.62366337 & -0.204020462 & -0.20732998 \\ -1.75941910 & -0.3826535 & -0.86317511 & -1.17272788 & -0.865691407 & 0.12542483 \\ -1.95192515 & -2.1096857 & 1.13135115 & 0.91256447 & -0.033529896 & 0.19059358 \\ -0.24144158 & -3.3571269 & 0.21738043 & 1.42840546 & -0.256638938 & -0.32533218 \\ 1.98337243 & -0.7162408 & -0.17289755 & -0.37637158 & 0.202418839 & -0.25420678 \\ 0.73442430 & -0.7828704 & -0.07627906 & -0.46668505 & 1.061208739 & -0.52216426 \\ 0.53228046 & -0.6705084 & 0.04502957 & -0.41734440 & 1.233682980 & -0.25313938 \\ -0.37260533 & -0.5712985 & 1.32752250 & -0.83881402 & 0.043950908 & 0.54991653 \\ 0.34063969 & -0.2236844 & 0.59420850 & -0.93477773 & -0.009209529 & 0.85547004 \\ -0.25306113 & 0.1018299 & 1.54180034 & -0.54482210 & -0.818951578 & -0.47394187 \\ 1.13138487 & 0.1457486 & 0.74310059 & -0.28906437 & 0.448604684 & 0.38436376 \\ 0.26399863 & -1.2144046 & -0.14020651 & -0.12147196 & -0.240048044 & -0.04377519 \\ -1.30421117 & 2.4709380 & 0.21944271 & 0.01303094 & 0.504568106 & -0.33221632 \\ -1.21980212 & 1.6156676 & 0.38826932 & 0.71976837 & -0.148517945 & -0.21359085 \\ -0.07690509 & 1.8051120 & 0.27156821 & 1.24900452 & 0.206211966 & 0.49705097 \\ -0.35915931 & 1.2125308 & -1.25387195 & 0.04770938 & 0.035716022 & 0.03874664 \\ -1.27762217 & 2.1233680 & 0.88368884 & -0.07337295 & -0.293427995 & -0.50109002 \end{pmatrix}$$

3. ACP datos faltantes

3.1. NIPALS

El algoritmo NIPALS realiza una descomposición singular de una matriz de datos, mediante secuencias iterativas convergentes de proyecciones ortogonales, el cuál utiliza la descomposición derivada del ACP que permite realizar la reconstitución de la matriz con los datos faltantes estimados. Gonzales (2014).

La siguiente figura es el esquema básico del algoritmo NIPALS

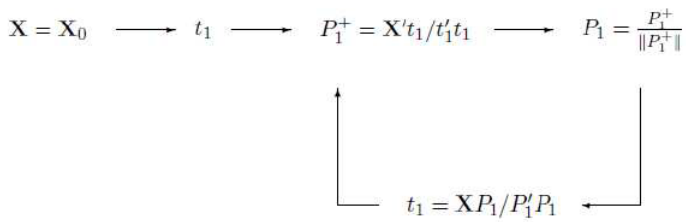

FIGURA 1: Esquema del algoritmo NIPALS

Para desarrollar nuestro código NIPALS en R Core Team (2017), nos basamos en el siguiente pseudocódigo propuesto por Gonzales (2014)

Etapa 1. $X_0 = X_h$

Etapa 2. $h = 1, 2, ..., a$:

Etapa 2.1. $t_h = 1^{\underline{a}}$ columna de X_{h-1}

Etapa 2.2. :Repetir hasta la convergencia de P_h

Etapa 2.2.1 Para $j = 1, 2, ..., p$:

$$P_{hj} = \frac{\sum_{\{i:x_{ji} \ e \ t_{hi} \ existen\}} x_{h-1,ji} t_{hi}}{\{i:x_{ji} \ e \ t_{hi} \ existen\} t_{hi}^2} \qquad [cov(t_h, x_{h-1,j})/s_{th}^2]$$

Etapa 2.2.2 normar P_h a 1.

Etapa 2.2.3 Para $i = 1, 2, ..., n$: $\quad t_{hi} = \frac{\sum_{\{j:x_{ji} \ existe\}} x_{h-1,ji} P_{hj}}{\sum_{\{j:x_{ji} \ existe\} P_{hj}^2}}$

Etapa 2.3. $X_h = X_{h-1} - t_h P_h'$

Fin

Para realizar este algoritmo NIPALS en matrices con datos faltantes, generamos NA's de forma aleatoria (se generaron pares de números correspondientes a las filas y columnas de la matriz y se reemplazaron por NA) en la matriz de datos correspondiente a la base de datos dada en la introducción , con 5,10,15, y 20 % de NA's o datos faltantes. Esas matrices son las siguientes:

- Matrices con el 5 % y 10 % de datos faltantes respectivamente:

Año	Colombia	Brasil	Chile	Argentina	Ecuador	Peru
1991	44.4	27.2	45.6	20.0	6.0	14.1
1992	75.5	11.8	58.9	22.6	NA	14.4
1993	110.7	50.6	128.3	17.2	119.4	118.5
1994	80.3	70.6	102.2	15.2	154.9	146.1
1995	81.6	82.3	89.0	35.1	169.4	127.1
1996	76.4	97.4	185.0	51.0	75.5	129.0
1997	32.0	89.5	195.3	31.1	33.4	NA
1998	55.5	63.1	66.3	24.4	9.7	66.7
1999	74.3	72.6	76.3	28.1	11.2	110.7
2000	84.5	76.2	80.1	29.5	11.8	110.2
2001	87.1	97.4	89.3	51.5	63.1	NA
2002	89.3	89.5	72.4	40.3	66.3	70.2
2003	70.2	63.1	80.1	60.5	76.3	90.1
2004	NA	66.3	70.5	39.1	20.0	64.5
2005	60.5	NA	107.2	31.1	63.4	92.7
2006	140.3	20.0	63.4	50.2	101.2	120.8
2007	120.4	22.6	101.2	51.0	103.1	107.2
2008	130.2	NA	103.1	42.5	66.7	70.8
2009	110.1	31.1	75.6	25.7	110.7	101.2
2010	120.2	24.4	68.9	60.3	110.2	110.8

Año	Colombia	Brasil	Chile	Argentina	Ecuador	Peru
1991	44.4	27.2	45.6	NA	6.0	14.1
1992	75.5	NA	58.9	22.6	17.8	14.4
1993	NA	50.6	NA	17.2	119.4	118.5
1994	80.3	70.6	102.2	15.2	154.9	146.1
1995	81.6	82.3	89.0	35.1	169.4	127.1
1996	76.4	97.4	185.0	51.0	75.5	129.0
1997	32.0	89.5	195.3	31.1	33.4	110.2
1998	55.5	63.1	66.3	24.4	9.7	66.7
1999	NA	72.6	NA	28.1	11.2	110.7
2000	84.5	76.2	80.1	29.5	11.8	110.2
2001	87.1	97.4	NA	51.5	63.1	89.3
2002	89.3	89.5	72.4	40.3	66.3	70.2
2003	70.2	63.1	80.1	60.5	76.3	90.1
2004	90.1	66.3	NA	39.1	20.0	64.5
2005	60.5	76.3	107.2	31.1	63.4	92.7
2006	140.3	20.0	63.4	50.2	101.2	120.8
2007	120.4	22.6	101.2	51.0	103.1	107.2
2008	NA	17.2	103.1	42.5	66.7	70.8
2009	NA	31.1	75.6	25.7	110.7	101.2
2010	120.2	24.4	68.9	NA	110.2	110.8

- Matrices con el 15 % y 20 % de datos faltantes respectivamente:

Año	Colombia	Brasil	Chile	Argentina	Ecuador	Peru
1991	NA	27.2	45.6	NA	6.0	14.1
1992	75.5	NA	58.9	22.6	17.8	14.4
1993	110.7	50.6	128.3	NA	119.4	118.5
1994	80.3	70.6	NA	15.2	154.9	146.1
1995	81.6	82.3	89.0	35.1	169.4	NA
1996	76.4	97.4	185.0	51.0	75.5	NA
1997	32.0	89.5	195.3	31.1	33.4	NA
1998	55.5	63.1	NA	NA	9.7	66.7
1999	74.3	72.6	76.3	28.1	11.2	NA
2000	84.5	76.2	80.1	29.5	NA	110.2
2001	87.1	97.4	89.3	51.5	63.1	89.3
2002	89.3	NA	72.4	40.3	66.3	70.2
2003	70.2	63.1	80.1	60.5	76.3	90.1
2004	90.1	66.3	NA	39.1	20.0	64.5
2005	60.5	76.3	NA	31.1	63.4	92.7
2006	140.3	20.0	63.4	50.2	101.2	120.8
2007	120.4	22.6	101.2	51.0	103.1	107.2
2008	NA	17.2	NA	42.5	66.7	70.8
2009	110.1	31.1	NA	25.7	110.7	101.2
2010	120.2	24.4	68.9	60.3	110.2	110.8

Año	Colombia	Brasil	Chile	Argentina	Ecuador	Peru
1991	44.4	27.2	45.6	NA	6.0	14.1
1992	75.5	11.8	58.9	22.6	17.8	14.4
1993	NA	50.6	128.3	NA	119.4	118.5
1994	80.3	70.6	102.2	15.2	154.9	146.1
1995	81.6	82.3	89.0	35.1	169.4	NA
1996	76.4	97.4	NA	51.0	NA	129.0
1997	32.0	89.5	NA	NA	33.4	110.2
1998	55.5	63.1	66.3	24.4	9.7	66.7
1999	74.3	72.6	NA	28.1	NA	110.7
2000	NA	76.2	80.1	29.5	11.8	110.2
2001	87.1	NA	89.3	51.5	63.1	NA
2002	89.3	89.5	72.4	40.3	NA	NA
2003	70.2	63.1	80.1	60.5	76.3	90.1
2004	90.1	66.3	70.5	39.1	20.0	NA
2005	60.5	NA	107.2	31.1	63.4	92.7
2006	140.3	20.0	63.4	50.2	101.2	120.8
2007	120.4	NA	NA	51.0	103.1	107.2
2008	130.2	17.2	103.1	42.5	66.7	70.8
2009	110.1	31.1	75.6	25.7	NA	NA
2010	120.2	24.4	68.9	60.3	NA	110.8

3.2. Resultados

Los valores propios obtenidos mediante el algoritmo NIPALS para las matrices con 5,10,15 y 20 % son:

% NA	λ1	λ2	λ3	λ4	λ5	λ6
5%	2,3547663	1,8468149	0,8981045	0,5496384	0,2795083	0,1209178
10%	2,0492213	1,8400072	1,0695745	0,4218713	0,2685668	0,1102202
15%	2,6192857	1,8931495	0,8460214	0,3352409	0,1818002	0,1172560
20%	2,8817975	1,7503809	0,7013755	0,4482426	0,2283079	0,0390586

En cuanto a los vectores propios obtenidos, por cuestiones de espacio, compararemos solo los tres primeros vectores propios de cada matriz:

5%			10%			15%			20%		
λ1	λ2	λ3	λ1	λ2	λ3	λ1	λ2	λ3	λ1	λ2	λ3
0,429852	0,482038	0,007478	0,484462	0,332624	0,228735	0,541642	0,212081	-0,003868	0,457017	0,453990	0,274918
0,072541	-0,627325	0,174591	-0,215357	-0,618792	0,071501	-0,377450	-0,445339	-0,160451	0,026206	-0,680218	-0,332918
0,190994	-0,550795	0,063925	0,059908	-0,578469	0,212304	-0,176297	-0,586110	-0,307341	0,396460	-0,275012	0,090973
0,309390	0,135118	0,900199	0,230694	-0,008894	0,865033	0,253567	0,158517	-0,909668	0,594395	0,236216	-0,696302
0,614496	0,040961	-0,362995	0,607419	-0,058545	-0,334772	0,454672	-0,409706	0,226803	0,389608	-0,200052	0,551248
0,547868	-0,225371	-0,152487	0,541435	-0,410275	-0,192719	0,511911	-0,469237	0,029089	0,358005	-0,399683	0,128834

Las componentes principales, también mostraremos las primeras tres para cada matriz:

5%			10%			15%			20%		
C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3
-3,383383	1,005828	-0,503124	-2,687992	1,921901	-0,955318	-1,660809	2,650744	0,702454	-4,057761	1,225500	0,472312
-3,170772	1,682689	-0,253596	-2,425476	1,727894	-0,122038	-2,332173	1,540168	0,927777	-2,508760	2,055053	0,207893
1,100503	-0,257444	-1,712558	1,009942	0,004336	-1,916570	1,500936	-0,894782	-0,818451	3,504455	0,277890	-0,152691
1,397917	-1,003520	-2,165248	1,449271	-1,069060	-2,385866	0,915464	-2,988067	1,925428	0,450791	-2,049283	2,316199
1,719708	-0,722982	-0,837304	1,610912	-0,934097	-1,066339	0,588331	-1,767889	0,800821	0,825102	-1,668399	1,195674
1,350247	-2,491395	1,164828	0,662112	-2,659141	1,152210	-1,178519	-1,986640	-1,599009	1,129349	-1,752291	-0,751256
-1,199498	-3,203311	0,292846	-1,251143	-2,939085	-0,013715	-3,447847	-1,636462	-0,551307	-1,895599	-2,001891	-0,234656
-2,064231	-0,213941	-0,217689	-1,920584	0,330140	-0,414289	-1,801518	0,323464	-0,243204	-2,110640	-0,135306	-0,463583
-0,923939	-0,509152	-0,096537	-0,918395	-1,006763	-0,262824	-1,617601	0,878347	0,397730	-0,693327	-0,917205	0,391566
-0,710296	-0,458470	0,024442	-0,675107	-0,301287	-0,097015	-0,101776	-0,632084	0,815349	-0,868038	-0,963995	-0,570836
0,451344	-0,759175	1,261783	-0,067886	-1,127103	1,428805	-0,325463	-0,368079	-1,088631	0,827088	0,065064	-0,698842
-0,311046	-0,160874	0,560416	-0,461420	-0,019737	0,514442	-0,149984	0,742351	0,031755	-0,012032	-0,763083	-0,510395
0,260481	0,066707	1,568230	0,231771	-0,006920	1,471616	0,146168	0,170142	-1,402135	0,736469	-0,018195	-1,275903
-1,390731	0,060332	0,709943	-0,933495	0,558371	0,856769	-0,899340	0,839420	-0,315576	-0,540867	0,294127	-0,739107
-0,551296	-1,123437	-0,149252	-0,662338	-0,837402	-0,333049	-1,002490	-1,108777	0,220980	-0,285517	-1,565333	-0,087737
1,683820	2,171089	0,293683	2,412082	1,747589	0,726690	2,803132	0,969454	-0,138441	1,519369	1,760108	0,703389
1,416131	1,317941	0,465472	1,918166	1,053951	0,881821	2,020439	0,413484	-0,465955	1,788242	0,552835	0,340316
0,452212	1,257342	0,455087	0,102658	1,263747	0,721762	0,520454	1,800461	0,257750	1,043003	1,677925	0,664486
0,585679	1,135097	-1,218811	0,874131	1,056392	-1,094310	1,241362	0,143055	1,561128	-0,446901	1,198518	1,233460
1,611154	1,822445	0,967061	1,820901	1,421643	0,206124	2,426357	0,828832	-0,874358	1,661081	1,605027	-0,354433

3.3. EM-PCA

El método Análisis de Componentes Principales vía EM (ACP-EM), es un método propuesto por Josse and Husson (2012), él cual realiza un Análisis de Componentes principales en presencia de datos faltantes. Particularmente, los datos faltantes son estimados inicialmente por valores promedio y luego se desea minimizar la distancia entre los valores originales (Z) y los valores estimados ($\Psi u'$). Ochoa (2018)

Para realizar nuestro código de este algoritmo en R Core Team (2017), seguimos el pseudocódigo propuesto por Ochoa (2018)

1. Iniciación L=0

   Z^0 los datos faltantes son reemplazados por valores iniciales, como por ejemplo la media.

2. Paso L

2.1 Realice un ACP para calcular Ψ^L,u^L, tomando q dimensiones (para garantizar un proceso iterativo).

2.2 Los valores faltantes son imputados via reconstitución de la matriz $Z^L = \Psi^L u'^L$. Los valores observados son los mismos y los faltantes se reemplazan por la imputación.

3. Se repiten el paso 2.1 y 2.2 hasta la convergencia.

Referencias

Gonzales, V. M. (2014), ANÁLISIS CONJUNTO DE MÚLTIPLES TABLAS DE DATOS MIXTOS MEDIANTE PLS, PhD thesis, UNIVERSITAT POLITÉCNICA DE CATALUNYA, Barcelona.

Kassambara, A. & Mundt, F. (2017), *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.5.
*https://CRAN.R-project.org/package=factoextra

Lê, S., Josse, J. & Husson, F. (2008), 'FactoMineR: A package for multivariate analysis', *Journal of Statistical Software* **25**(1), 1–18.

Ludovic Lebart, Alain Morineau, M. P. (1995), *Statistique exploratoire multidimensionnelle*, Dunod, Paris.

Ochoa, A. F. (2018), Análisis de correspondencias múltiples en presencia de datos faltantes: el principio de datos disponibles del algoritmo nipals (acmpdd), Master's thesis, Universidad del Valle.

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*https://www.R-project.org/

Wickham, H. (2009), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
*http://ggplot2.org

Wickham, H. & Bryan, J. (2018), *readxl: Read Excel Files*. R package version 1.1.0.
*https://CRAN.R-project.org/package=readxl

Zelaya, J. T. (n.d.), *ANÁLISIS MULTIVARIADO DE DATOS*, Universidad de Costa Rica.