Análisis de Componentes Principales en presencia de datos faltantes: el principio de datos disponibles

Principal Components Analysis in the presence of missing data: the principle of available data

V. M. Gonzalez-Rojas ; G. Conde-Arango , A. F. Ochoa-Muñoz
DOI: https://doi.org/10.22517/23447214.20591
Artículo de investigación científica y tecnológica

***Resumen*— En este trabajo proponemos utilizar el principio de datos disponibles derivado del algoritmo NIPALS (Nonlinear estimation by Iterative Partial Least Square) para trabajar el Análisis de Componentes Principales (ACP) en presencia de datos faltantes. Esta propuesta es importante puesto que no realiza imputación de datos, ni se descartan individuos de la base datos; el método propuesto trabaja con los elementos pares disponibles para conformar las matrices de cuasicorrelación en R^p y en R^n; la descomposición espectral de estas matrices permite a través de las relaciones de transición realizar un ACP convencional. Del estudio de simulación realizado se encontró que a medida que aumenta el porcentaje de datos faltantes disminuye la inercia explicada en el primer plano factorial. Se desarrolló el algoritmo de solución bajo el entorno de programación R y se anexa el código para uso libre.**

***Palabras claves*—ACP, datos faltantes, datos disponibles, NIPALS, relaciones de transición.**

***Abstract*— In this paper we propose to use the principle of available data derived from the NIPALS (Nonlinear estimation by Iterative Partial Least Square) algorithm to work on the Principal Components Analysis (PCA) in the presence of missing data. This proposal is important since it does not perform data imputation, nor are individuals discarded from the database; the proposed method works with the available pairs to form the quasicorrelation matrices in R^p and in R^n; the spectral decomposition of these matrices allows through the transition relations to realize a conventional PCA. From the simulation study carried out, it was found that as the percentage of missing data increases, the inertia explained in the first factorial plane decreases. The solution algorithm was developed under the R programming environment and the code is appended for free use.**

***Index Terms*— PCA, missing data, available data, NIPALS, transition relations.**

I. INTRODUCCIÓN

EL Análisis de Componentes Principales (ACP), es por excelencia el método más utilizado en análisis multivariado de datos. De hecho, casi todos los demás métodos multivariados, descansan en el ACP como es el caso de los Análisis de Correspondencias, Canónico, Interbaterías, Análisis Factorial Múltiple, STATIS, entre otros [1] Sin embargo, en presencia de datos faltantes (NA) estos métodos no funcionan, por lo que se requiere que se imputen los valores omitidos, o se eliminen los registros comprometidos con los NA. La imputación de datos faltantes puede conducir a resultados inestables [2].

Recientemente, los métodos PLS (Partial Least Squares) y en especial el algoritmo NIPALS [3] presenta el principio de "datos disponibles" para tratar de remediar esta situación, sin embargo, se evidencia que las componentes así obtenidas pierden la propiedad de ortogonalidad y por consiguiente comprometen las relaciones de transición.

En este artículo se presenta una solución al problema del ACP de una matriz de datos en presencia de NA, sin hacer imputación, usando el principio de datos disponibles y las propiedades derivadas de las relaciones de transición.

Siendo $Z_{n,p}$ la matriz de datos completos estandarizados y $N_{n,n}$ la matriz diagonal con los pesos de los individuos generalmente $1/n$, nos valemos del hecho de que la descomposición espectral (d.e) de la matriz de correlaciones $Z'NZ$ nos entrega los valores λ_α y vectores propios u_α en R^p(donde $\alpha = 1,2,\dots,p$). Análogamente, en R^n la d.e de la matriz $N^{1/2}ZZ'N^{1/2}$ nos permite obtener los mismos valores propios λ_α y los vectores propios $v_\alpha \epsilon R^n$.

Usaremos las propiedades de transición y el principio de datos disponibles para conformar estas dos matrices simétricas, ya que mediante descomposición espectral (d.e) es inmediato el cálculo de los valores propios, las componentes principales, las coordenadas de las variables y las contribuciones de individuos y variables como elementos básicos para el ACP.

El análisis de los resultados se debe realizar de la manera acostumbrada sin perder de vista que estos fueron obtenidos de la información disponible en una matriz de datos faltantes.

Observe que, con este procedimiento, se tiene una respuesta metodológica rápida y eficaz al problema de realizar un ACP

Este manuscrito fue enviado el 22 de enero de 2020 y aceptado 23 de junio 2021.

en matrices de datos incompletos. Se presenta el algoritmo de solución bajo el entorno de programación R [4], el cual a su vez funcionará adecuadamente y de manera general en matrices de datos con y sin datos faltantes.

El capítulo dos presenta las metodologías donde se describen el ACP y sus principales propiedades, el algoritmo NIPALS con el principio de los datos disponibles. El capítulo tres se refiere a los resultados y se muestra la consistencia en cuanto a la conservación de las propiedades derivadas del ACP con datos faltantes. Finalmente, el capítulo cuatro recoge las conclusiones y recomendaciones.

II. METODOLOGÍAS

Para garantizar la ortogonalidad de los factores derivados de un ACP con datos faltantes, se propone a partir de las relaciones de transición conformar las matrices con datos disponibles asociadas a los sistemas de valores y vectores propios en R^p y R^n respectivamente:

$$Z'NZu = \lambda u$$
$$N^{1/2}ZZ'N^{1/2}v = \lambda v$$

Este procedimiento garantiza los mismos valores propios λ_α en cada eje α del mismo rango en ambos espacios, conllevando la ortogonalidad de los vectores propios.

A. *Analisis De Componentes Principales (Acp)*

Con el ACP se obtienen representaciones sintéticas de un conjunto de datos cuantitativos expresados en una matriz de orden $n.p$ describiendo las *similitudes (distancias)* entre individuos, las *correlaciones* entre variables y las relaciones entre individuos y variables, en espacios de menor (reducción) dimensión generalmente en el primer plano factorial [5,1].

La matriz inicial de datos brutos $B_{n,p}$ con *n* filas (nube de individuos) y p columnas (nube de variables métricas) generalmente es transformada (centrada: $X_{n,p}$ o estandarizada: $Z_{n,p}$) y luego sometida al análisis de datos [6]. En el primer caso su término general es x_{ij} ($i_{\text{ésima}}$ observación de la $j_{\text{ésima}}$ variable); note que un vector fila x_i pertenece al espacio R^p y un vector columna $X_j \in R^n$, con $i = 1,\dots,n$; $j = 1,\dots,p$.

1) *Análisis de la nube de individuos*

En el cálculo de la distancia $d_M^2(b_i, b_{i'}) = \sum_j m_j(b_{ij} - b_{i'j})^2$ entre los puntos fila b_i *e* $b_{i'}$, es posible considerar la importancia de las columnas mediante las ponderaciones m_j que se encuentran en la matriz diagonal *M* simétrica definida positiva[2] denominada *métrica*.

Haciendo[3] $l = i - i'$ entonces

$$\|l\|_M = < l, l >_M = l'Ml = d_M^2(i, i') .$$

Se generaliza el concepto de norma o distancia euclidea a partir de M.

Definición. La inercia (~varianza) como medida de dispersión alrededor del centro de gravedad $g = (\bar{b}_1,..,\bar{b}_p)$ para la nube de individuos en el hiperespacio se define como

$$I_n = \sum_{i=1}^{n} p_i d_M^2(i, g)$$

Cada individuo posee su propio peso p_i, pero en muchos estudios estos se asumen uniformes, $p_i = 1/n$, los cuales se pueden fijar en la matriz diagonal N.

Generalmente, se analiza la matriz de datos estandarizados $Z_{n,p}$ para evitar la afectación de las variables debido a las unidades de medida [7,8]. Así, con datos de partida $z_{ij} = \frac{b_{ij} - \bar{b}_j}{s_j}$ normalizados[4], se puede considerar implícitas las ponderaciones $1/s_j^2$ en el cálculo de la distancia, conllevando como métrica $M = I$, con lo cual $u'u = 1$.

En este caso normado, la *inercia total* de la nube de individuos respecto a $g = (0, \dots, 0)$ en el hiperespacio está dada por

$$I_n = \sum_{i=1}^{n} p_i d^2(z_i, O) = \sum_i \frac{1}{n} \sum_j z_{ij}^2 = \sum_j v(Z_j) = traza\,(Z'NZ) = p.$$

Esta inercia total se descompone en forma gradual y maximal en p nuevos ejes u_α, a través de la varianza generada por las *n proyecciones ortogonales* contenidas en el vector $\psi_\alpha = ZMu_\alpha = Zu_\alpha$. Geométricamente (Fig 1):

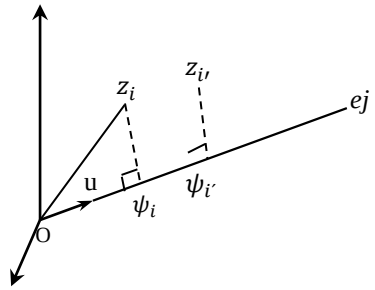


Fig 1. Proyección ortogonal del individuo *i* sobre *u* ($z_i'.u =$

[2] $x'Mx \geq 0$; $es\ 0\ (definida)\ sii\ x = 0, simetrica\ si\ x'Ml = l'Mx$.

[3] Para facilitar notación algunas veces se denota un vector fila con subíndice *i* ; análogamente el subíndice *j* para columnas.

[4] Observe que la estructura de correlaciones de los datos originales no cambia al ser estandarizados.

La coordenada de la proyección, ψ_i, es el *escalar* que contrae o dilata u hasta obtener la ortogonalidad pues,

$$(z_i - \psi_i u)'(\psi_i u) = 0 \Rightarrow \psi_i z_i' u - \psi_i^2 u'u = 0 \Rightarrow \psi_i = z_i' u \, .$$

El teorema de Pitágoras aplicado a cada uno de los *n* triángulos rectángulos del tipo $Oz_i\psi_i$ en la Fig 2, conduce a la relación

$$\sum_{i=1}^{n} {z_i\psi_i}^2 = \sum_i {Oz_i}^2 - \sum_i {O\psi_i}^2$$

Premultiplicando por *1/n*, y ya que $\frac{1}{n}\sum_i {Oz_i}^2$ es la cantidad fija (observada), minimizar $\frac{1}{n}\sum_i {z_i\psi_i}^2$ (proyección óptima) es equivalente a maximizar la cantidad $\frac{1}{n}\sum_i {O\psi_i}^2 = \psi' N\psi = u'Z'NZu$ que es la *inercia (varianza*[5]*) de la nube* N_I *proyectada* sobre el eje de dirección u.

Buscar el máximo de $\frac{1}{n}\sum_i {O\psi_i}^2$ equivale a encontrar *u* tal que maximice $u'Z'NZu$ bajo la restricción $u'u = 1$. Para el lagrangiano $L(u) = u'Z'NZu - \lambda(u'u - 1)$, la derivada $\partial L/\partial u = 0$, conduce a resolver el sistema de valores y vectores propios

$$Z'NZu = \lambda u \quad (2.1)$$

Así, $u'Z'NZu = \lambda$ es la inercia para maximizar y debe corresponder al valor propio más grande de la matriz de correlaciones $Z'NZ = R$.

Sea u_1 en R^p el vector propio correspondiente al mayor valor propio λ_1; el subespacio en R^p de dos dimensiones que mejor se ajusta a la nube, contiene ortogonalmente a $u_2 (u_1'u_2 = 0)$ con $\lambda_2 \leq \lambda_1$ bajo $u_2'u_2 = 1$.

Se busca de manera análoga el mejor subespacio de dimensión $q \leq p$ que recoja gradualmente y de forma decreciente las proporciones de inercia proyectadas. La orientación (*signo*) de los ejes es *arbitraria*, pues no afecta la forma de la nube y respeta las distancias.

El análisis efectúa una *traslación* y *rotación* alrededor del origen y obtiene un sistema de vectores ortonormados u_1, u_2, ... , u_p que pasan lo más cerca de la nube; es decir, se diagonaliza (*descomposición espectral*) la matriz de correlaciones $Z'NZ = UDU'$; U es ortogonal $(U'U = UU' = I)$ conteniendo los vectores propios $u_\propto$ y D es diagonal con los valores propios λ_α de R (*semidefinida positiva*). Observe entonces que

$$traza(Z'NZ) = traza(UDU') = traza(D) = \sum_\alpha \lambda_\alpha = p = I_n$$

Las coordenadas de los *n* puntos individuos proyectados sobre el eje 'canónico' u_α son los *n* componentes[6] del vector (factor) $\psi_\alpha = Zu_\alpha$ el cual es una combinación lineal (subespacio generado) de las variables iniciales; Si $\alpha = 1$,

$$\psi_1 = u_{11}Z_1 + u_{12}Z_2 + \cdots + u_{1p}Z_p$$

es la primera *Componente Principal*. Los pesos $u_{\alpha j}$ asociados a las variables, las ubicarán al lado del eje con mayor correlación y los individuos con valores máximos en estas variables seguirán sus direcciones ubicándose en los extremos correspondientes, y en sentido contrario aquellos individuos con valores mínimos.

2) Análisis de la nube de puntos variables (Estandarizadas)

En R^n el análisis de los puntos variables con métrica *N* se hace respecto al origen *O*, con lo cual están a una distancia 1 del origen, esto es $d^2(j,O) = \sum_{i=1}^{n} \frac{1}{n}(z_{ij})^2 = j'Nj = \|j\|_N = 1$; es decir, todas las variables están sobre una hiperesfera de radio 1, ver Fig 2.

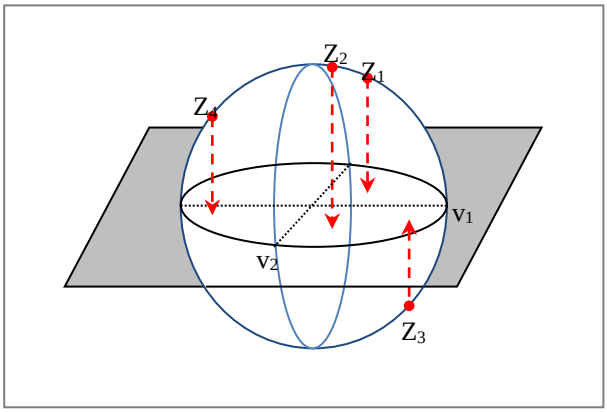


Fig 2. Proyección de variables sobre el plano director

En el *círculo de correlaciones* (Fig 3), el primer eje está asociado con las variables Z_3 *y* Z_4 relacionadas inversamente; y el segundo eje está ligado básicamente a Z_2. Note que Z_1 afecta poco a los dos ejes.

[5]Con variables centradas $E(\psi_\alpha) = 0$, $v(\psi_\alpha) = \psi_\alpha' N\psi_\alpha = \lambda_\alpha$; cada *componente* explica una proporción λ_p/p de la variabilidad total, pues $\lambda_1 + \cdots + \lambda_p = p$.

[6] Las *componentes principales* también son denominadas *variables latentes*.

Procediendo en forma análoga a la nube de individuos, la distancia $h = j - k$ entre dos variables, ahora con métrica N está dada por:

$$d_N^2(j,k) = h'Nh = \sum_{i=1}^{n} \frac{1}{n}(z_{ij} - z_{ik})^2 = 2(1 - r_{jk}),$$

lo cual implica $0 \leq d^2(j,k) \leq 4$.

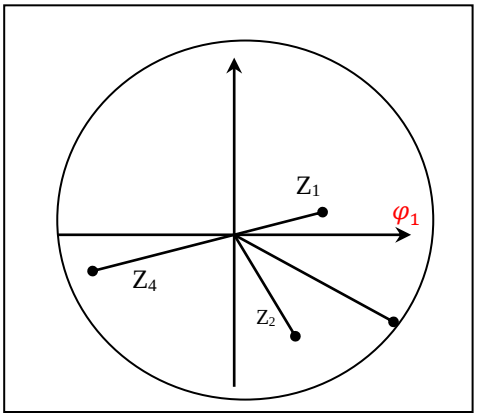


Fig 3. Círculo de correlación en el plano $_{1,2}$

Dos puntos variables próximos indican que toman valores muy relacionados en el conjunto de individuos; así, el espacio de las variables es una representación de la matriz de correlaciones.

En el hiperespacio de las variables, ahora con métrica N y pesos[7] $M = I$; se tiene la inercia $I_p = \sum_j d_N^2(j,O) = \sum_j j'Nj = \sum_j v(j) = p$, la cual se proyecta sobre el vector $\dot{v}$ tal que

$$\varphi = Z'N\dot{v} \qquad (2.2)$$

con varianza $\varphi'M\varphi = \varphi'\varphi = \dot{v}'N'ZZ'N\dot{v}$ a maximizar bajo la restricción $\dot{v}'N\dot{v} = 1$. Esto implica diagonalizar la matriz *no simétrica* asociada al sistema $ZZ'N\dot{v} = \lambda\dot{v}$. Sin embargo, bajo la transformación $v = N^{1/2}\dot{v}$ se resuelve el sistema simétrico

$$N^{1/2}ZZ'N^{1/2}v = \lambda v, \quad \text{bajo} \quad v'v = 1 \quad (2.3)$$

3) *Relaciones de Transición*

Realizando el cambio $Y = N^{1/2}Z$ en (2.1) y en (2.3) se tiene respectivamente:

$$Y'Y\underbrace{u} = \lambda u \quad ; \quad YY'v = \lambda v \Rightarrow Y'Y\underbrace{Y'v} = \lambda Y'v$$

Así, $u = kY'v$, y por tanto $u'u = k^2v'YY'v = k^2\lambda = 1$ esto es $k = 1/\sqrt{\lambda}$; entonces

$$u = \frac{1}{\sqrt{\lambda}}Y'v = \frac{1}{\sqrt{\lambda}}Z'N^{1/2}v\,.$$

Análogamente, $u\sqrt{\lambda} = Y'v$, $Yu\sqrt{\lambda} = YY'v$ se tiene $v = Yu/\sqrt{\lambda}$ tal que en el eje α

$$v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}N^{1/2}Zu_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}N^{1/2}\psi_\alpha \quad , \quad v \in R^n$$

Igual que en el espacio fila, se busca el eje más relacionado con el conjunto de variables originales que maximice la inercia proyectada, obteniendo las coordenadas de las variables proyectadas ortogonalmente sobre el α-ésimo de dirección v_α ; mediante (2.2),

$$\varphi_\alpha = Z'N^{1/2}v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}Z'NZu_\alpha = \sqrt{\lambda_\alpha}u_\alpha \qquad (2.4)$$

Retomando la parte central de (2.4), la coordenada de la j-ésima variable es

$$\varphi_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}}Z_j'N\psi_\alpha = \sum_{i=1}^{n}\frac{b_{ij} - \bar{b}_j}{s_j}\frac{1}{n}\frac{\psi_{\alpha i}}{\sqrt{\lambda_\alpha}} = cor(j,\psi_\alpha),$$

Note que $|\varphi| \leq 1$, y que la coordenada de una variable es el *coeficiente de correlación* de la variable con el factor ψ_α. Esta correlación se lee directamente en el círculo a través de la coordenada de la variable j en el eje α [5, 9]. Además, $v(\varphi_\alpha) = \varphi_\alpha'\varphi_\alpha = \sum_j cor^2(\psi_\infty, Z_j) = \lambda_\alpha$ es máxima y corresponde a la varianza explicada o redundancia $Rd(Z,\psi_\infty)$.

Otras relaciones de interés son:

$$\psi_\alpha = Zu_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}ZZ'N^{1/2}v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}Z\varphi_\alpha = \sqrt{\lambda_\alpha}N^{-1/2}v_\alpha$$

Observe que los factores φ_α e ψ_α son colineales a los vectores propios u_α e v_α . De la relación anterior y de (2.1) se tiene que (ver NIPALS):

$$u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}Z'N^{1/2}v_\alpha = \frac{1}{\lambda_\alpha}Z'N\psi_\alpha = Z'\psi_\alpha/[\psi_\alpha'\psi_\alpha]$$

B. Algoritmo NIPALS

El algoritmo NIPALS (Nonlinear estimation by Iterative Partial Least Square) es la base de la regresión PLS [10]. Fundamentalmente realiza la descomposición singular de una matriz de datos, mediante secuencias iterativas convergentes de proyecciones ortogonales [concepto geométrico de regresión simple], se tiene equivalencia con los resultados del ACP.

[7] Ya que la estandarización introduce los pesos de las variables $1/S_j^2$, implica que luego se tome $M = I$.

Sea $X_{n,p}$ la matriz de datos de rango $a \leq p$ cuyas columnas $X_1, \dots, X_p$ se suponen centradas o estandarizadas (bajo S_n). La reconstitución derivada del ACP conlleva a $X = \sum_h^a t_h P_h'$ donde t es la componente principal (scores) y P_h' el vector propio (loadings) en el eje h.

$$[X_1 \dots X_p] = t_1 P_1' + \cdots + t_a P_a' \quad . \qquad (2.5)$$

Así, la columna $X_j = \sum_h^a P_{hj} t_h \; j = 1, \dots p$ y la i-ésima fila $x_i = \sum_h^a t_{hi} P_h \; i = 1, n$.

Observe entonces que si $h = 1$, la columna j se expresa como $X_j = P_{1j} t_1$ es decir $P_{hj} = X_j' t_h$ es el coeficiente (pendiente[8]) en la regresión de X_j sobre t_h . En el espacio de las filas, t_{hi} es el coeficiente de la regresión sin constante del individuo x_i sobre P_h .

Para $h > 1$, P_{hj} es el coeficiente de regresión de t_h en la regresión simple del vector deflactado $X_j - \sum_l^{h-1} P_{lj} t_l$ sobre t_h y t_{hi} el de P_h en la regresión de $x_i - \sum_l^{h-1} t_{li} P_l$ sobre P_h .

El objetivo de cualquier algoritmo PLS es el procedimiento iterativo para calcular los parámetros P_h del modelo. Para cada componente las cargas son computadas, una como función de la otra, a través del procedimiento iterativo [11].

1) Pseudocódigo NIPALS con datos completos

Se presenta a continuación el pseudocódigo asociado al algoritmo NIPALS, donde se destaca la etapa 2.2, la cual es la parte esencial del procedimiento.

Etapa 1. $X_0 = X_h$

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

Etapa 2.1. t_h = 1ª columna de X_{h-1} [o $\bar{X}$]

Etapa 2.2. Repetir hasta la convergencia de P_h

Etapa 2.2.1 $P_h = \frac{X_{h-1}' t_h}{t_h' t_h} \left[u = \frac{X'Xu}{\lambda} \; , \; \lambda = \frac{1}{n} t't \right]$

Etapa 2.2.2 normar P_h a 1

Etapa 2.2.3 $t_h = X_{h-1} P_h$ [t = Xu]

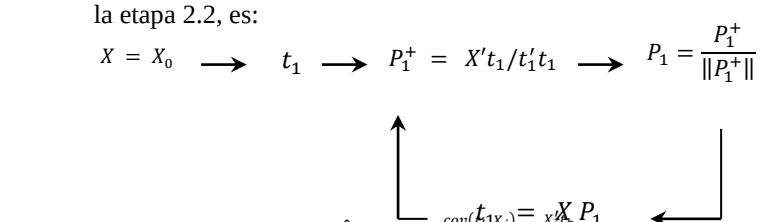
Etapa 2.3. $X_h = X_{h-1} - t_h P_h'$ [garantiza ortogonalidad]

Siguiente h.

Fin

2) Descripción pseudocódigo NIPALS

El flujograma asociado al procedimiento de convergencia en la etapa 2.2, es:



[8]De la regresión simple $\hat{\beta}_1 = \hat{p}_{hj} = \frac{cov(t_h, x_j)}{S_{t_h}^2} = \frac{x_j' t_h}{\|t\|^2} = X_j' t_h = r$ si *x* e *t* estandarizadas.

Luego se construirán una serie de tablas deflactadas notadas X_h cuyas columnas son $X_{h1}, \dots, X_{hp}$; la i-ésima fila se notará $x_{hi}' = (x_{h1i}, \dots, x_{hpi})$. El algoritmo inicia tomando X_{01} como la 1ª componente principal t_1 para calcular P_h con el cual recalcula t_1 iterando hasta la convergencia.

Tal como se estudió inicialmente, en la etapa 2.2.1 P_{hj} representa, antes de la normalización, el coeficiente (pendiente) de la regresión de $X_{h-1,j}$ sobre la componente t_h.

Análogamente, en la etapa 2.2.3, t_{hi} representa el coeficiente de regresión (sin constante) de $x_{h-1,i}$ sobre P_h ; y ya que $P_h' P_h = 1$, t_{hi} también es el largo de la proyección ortogonal de $X_{h-1,i}$ sobre P_h .

Para $h = 1$ se obtiene el primer eje factorial P_1 y la primera componente principal t_1 de $X'X$. Ya que la matriz $X_1 = X - t_1 P_1'$ representa el residuo de la regresión de X sobre la primera componente principal, de (3.1), el vector propio P_2 de la matriz $X_1' X_1 \,/\, n$ asociado al valor propio más grande, corresponde al vector propio de $X'X/n$ asociado al segundo valor propio más grande λ_2 .

Las relaciones cíclicas de la etapa 2.2 muestran que en el límite se verifican las ecuaciones:

$\frac{1}{n} X_{h-1}' X_h P_h = \lambda_h P_h$; $\frac{1}{n} X_h X_{h-1}' t_h = \lambda_h t_h \Rightarrow \lambda_h = \frac{1}{n} t_h' t_h$

Una vez se consigue la convergencia, en la etapa 2.3 se deflacta la matriz precedente para garantizar la ortogonalidad de las siguientes componentes. Con datos completos el divisor $t_h' t_h$ en la etapa 2.2.1 no es necesario.

La convergencia en NIPALS se puede conseguir con P_h tal como se ha expuesto, pero también con los t_h que representarán los vectores propios en R^n y los resultados serán equivalentes [11, 3].

Así, el problema del ACP bajo NIPALS es resolver una serie de regresiones simples locales hasta alcanzar la convergencia de los coeficientes de regresión P_{hj} y t_{hi} que es el nuevo valor proporcionado de la regresión sin constante de $x_{h-1,i}$ sobre la 'nueva' variable P_h después de su normalización.

La principal característica del NIPALS es que trabaja respecto a una serie de productos escalares como suma de productos de los elementos emparejados. Esto permite manejar datos faltantes, agregando en cada operación los pares disponibles. Geométricamente el procedimiento 'toma' los elementos omitidos como si ellos cayeran sobre la recta de regresión; no son puntos de apalancamiento [3].

Así, con datos faltantes se obtiene sin embargo las componentes t_h y los vectores P_h que permiten luego 'reconstituir' la matriz de datos mediante $\hat{X}$ y de ésta, estimar los datos faltantes utilizando la fórmula de reconstitución (2.5) derivada del ACP : $\hat{x}_{ji} = \sum_l^h t_{li} p_{lj}$.

3) Pseudocódigo NIPALS con datos faltantes

Etapa 1. $X_0 = X_h$

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

Etapa 2.1. th = 1ª columna de X_{h-1}

Etapa 2.2. Repetir hasta la convergencia de P_h

Etapa 2.2.1 para j=1,2,…,p:

$$P_{hj} = \frac{\sum_{\{i:x_{ji}et_{hi}existen\}} x_{h-1,ji}t_{hi}}{\{i:x_{ji}et_{hi}existen\}t_{hi}^2}$$

Etapa 2.2.2 normar P_h a 1.

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

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

Fin

En las etapas 2.2.1 y 2.2.3 se calculan las pendientes de las rectas de mínimos cuadrados pasando por el origen de la nube de puntos sobre los datos disponibles. Los P_{hj} y los t_{hi} deben conservar en sus posiciones j e i, la característica de dato faltante dada por x_{ij} , la cual se puede expresar con 0.

4) Propuesta Metodológica de Solución: el principio de datos disponibles en ACP

El método propuesto a través de la función fACPna(), permite el procesamiento de matrices con datos completos y datos faltantes (ver Anexo 1).

El método tiene el mismo principio expuesto en el Análisis de Correspondencias Múltiples usando el principio de datos disponibles [2] y en el Análisis Interbaterías vía PLS [12]. Recientemente Patel y colaboradores [13] acuden al uso de NIPALS sin usar técnicas de imputación en modelos de espacio estado. De esta forma, se destaca el principio de datos de disponibles de NIPALS como una solución novedosa que sigue siendo utilizada para hacer frente al problema de NA.

Se conforma la matriz de datos disponibles estandarizados $\mathbb{Z}_{n,p}$ y por consiguiente la matriz simétrica $\mathbb{Z}'N\mathbb{Z}$ que ahora es una "matriz de cuasi-correlaciones", cuya d.e permite obtener los valores y vectores propios $\lambda_\propto$ y $\mathbb{u}_\alpha \epsilon R^p$. Análogamente la d.e de la matriz $N^{1/2}\mathbb{Z}\mathbb{Z}'N^{1/2}$ entrega los mismos valores propios $\lambda_\propto$ y los vectores propios $\mathbb{v}_\alpha \epsilon R^n$.

De las relaciones de transición es inmediato el cálculo de las componentes $\psi_\alpha^* = \sqrt{\lambda_\propto}\, N^{-1/2}\mathbb{v}_\alpha$ y de las coordenadas de las variables $\varphi_\alpha^* = \sqrt{\lambda_\propto}\, \mathbb{u}_\alpha$, con lo cual se puede realizar los análisis gráficos acostumbrados como el "circulo de correlaciones" y la representación simultánea. Igualmente se tendrá las contribuciones absolutas de individuos y variables para identificar aquellos responsables de la conformación de los principales ejes de representación.

III. RESULTADOS

En esta sección se realizó un ACP con datos completos y datos faltantes. La base de datos que se utilizo es la que se denomina *carscomplete* y se encuentra en el paquete *plsdepot* del software R [14]. Esta base de datos tiene 27 automóviles y 6 variables cuantitativas, las cuales son Cilindraje (CILIN), Potencia (PUISS), Velocidad (VITES), Peso (POIDS), Longitud (LONG) y Altura (Large).

A. ACP con datos completos

Al realizar un ACP en la base de datos completa se obtienen 6 valores propios, los cuales son los siguientes: 4.6560, 0.9152, 0.2404, 0.1027, 0.0646 y 0.02096. Observe que la inercia recogida en el primer plano factorial representa el 92.8% de la inercia total. En la Tabla I, se muestra los tres primeros vectores propios u_1, u_3 y u_3. Estos vectores son de mucha ayuda en el ACP puesto que con ellos se pueden calcular las coordenadas factoriales $\varphi_\alpha = \sqrt{\lambda_\alpha}u_\alpha$

TABLA I.
TRES PRIMEROS VECTORES PROPIOS DE R^P CON DATOS COMPLETOS

Variable	Eje 1	Eje 2	Eje 3
CILIN	-0.4442	0.0339	0.4014
PUISS	-0.4144	0.4212	0.0395
VITES	-0.3435	0.6634	-0.3699
POIDS	-0.4303	-0.2551	0.4844
LONG	-0.4302	-0.2955	-0.0439
LARGE	-0.3776	-0.4783	-0.6810

B. ACP con datos faltantes (fACPna)

Se contaminó la base original con el 20% de datos faltantes distribuidos aleatoriamente (ver Anexo 2) en la matriz de datos y se obtienen los valores propios asociados al ACP bajo el principio de datos disponibles, los valores propios son: 4.1416, 0.9699, 0.4827, 0.1872, 0.1244 y 0.0940.

Con datos faltantes, la inercia recogida en el primer plano factorial ahora representa el 85%, con lo cual se insinúa pérdida del poder descriptivo en R^2 como consecuencia de la falta de información.

TABLA II.
TRES PRIMEROS VECTORES PROPIOS DE R^P CON DATOS FALTANTES

Variable	Eje 1	Eje 2	Eje 3
CILIN	-0.4493	0.0264	0.4380
PUISS	-0.4432	0.3357	0.1637
VITES	-0.3070	0.7543	-0.2291
POIDS	-0.4095	-0.3988	0.4230
LONG	-0.4355	-0.2512	-0.2274
LARGE	-0.3871	-0.3088	-0.7054

En la Tabla II, se tienen los tres primeros vectores propios u_1, u_3 y u_3 en presencia de datos faltantes. Estos vectores son ortogonales como consecuencia del cumplimiento de las relaciones de transición en los espacios definidos por la nube de individuos y de variables.

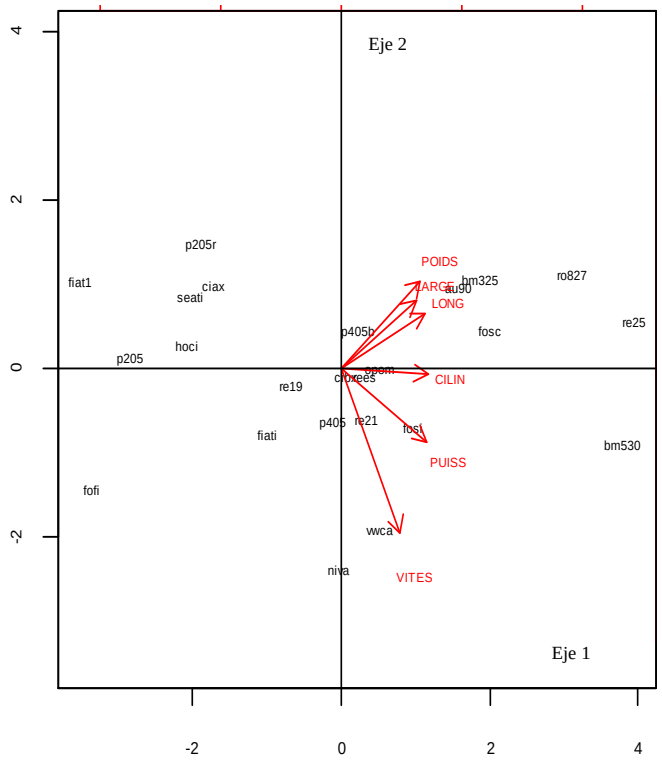


Fig 4. Representación Simultánea, datos faltantes

Se observa en la Fig 4 que los autos hoci, bm325, au90, r0827, re25, fosc presentan mayor cilindraje y potencia. También se observa el factor tamaño, con lo cual el índice subyacente derivado de la primera componente principal se refiere al desempeño "performance" de los autos. Así, los autos antes citados son los de mejor performance mientras que los autos fiat1, fofi, p205 son los de peor desempeño.

C. *Estudio de Simulación de la inercia principal y la correlación entre coordenadas factoriales según el porcentaje de datos faltantes*

En esta sección se realizó un estudio de simulación en donde fue de interés analizar qué pasa con el porcentaje de inercia explicado, con los ejes factoriales 1 y 2; teniendo en cuenta porcentajes del 5%, 10%, …, 30% con información faltante generada aleatoriamente (ver Anexo 2). De esta manera se simularon 1000 matrices con cada uno de los porcentajes correspondientes, es decir se tendrán 6000 matrices simuladas. En la Fig 5, se observa que a medida que aumenta el porcentaje de datos faltantes se disminuye el porcentaje de inercia explicado en el primer plano factorial, lo cual hace referencia a que a medida que hay más datos faltantes es más difícil explicar las relaciones entre individuos y las variables. Es importante mencionar que la línea de referencia corresponde al porcentaje de inercia con datos completos, el cual es 0.9285.

Ahora bien en la Fig 6, se observa que la correlación entre los coordenadas de las variables con datos completos y datos faltantes en el eje 1 $(cor(\varphi_1, \varphi_1^*))$, son correlaciones que oscilan desde 0.75 a 1, se observa que a medida que aumenta el porcentaje de datos faltantes se disminuye la correlación entre las variables y aumenta el número de casos donde se disminuye dicha correlación, esto se puede relacionar con individuos que tienen todos sus valores faltantes, los cuales quedan situados en el origen del plano y provocan cambios en las coordenadas de las variables. Este mismo comportamiento se observa en la Fig 7, donde se compara la correlación ahora en el eje 2, para este eje se observa más perdida de correlación la cual está influenciada también por la perdida en el porcentaje de varianza explicado. Sin embargo, se observa que la mediana de las correlaciones está ubicada en un valor deseado para la correlación (0.8 a 1).

Para este ejemplo de simulación, cuando se tiene la más alta proporción de datos faltantes (30%), la probabilidad de tener individuos con todos sus elementos faltantes es del orden de 10^{-37}, es decir, es muy poco frecuente encontrar individuos con todos sus elementos faltantes. Esta situación es importante analizarla porque en caso de presentarse se deberá excluir estos individuos del análisis. El cálculo de la probabilidad se puede pensar como "de cuantas formas se pueden repartir datos faltantes en una matriz, de tal forma que un individuo o más queden con todos sus elementos como NA". Las combinatorias facilitan este cálculo.

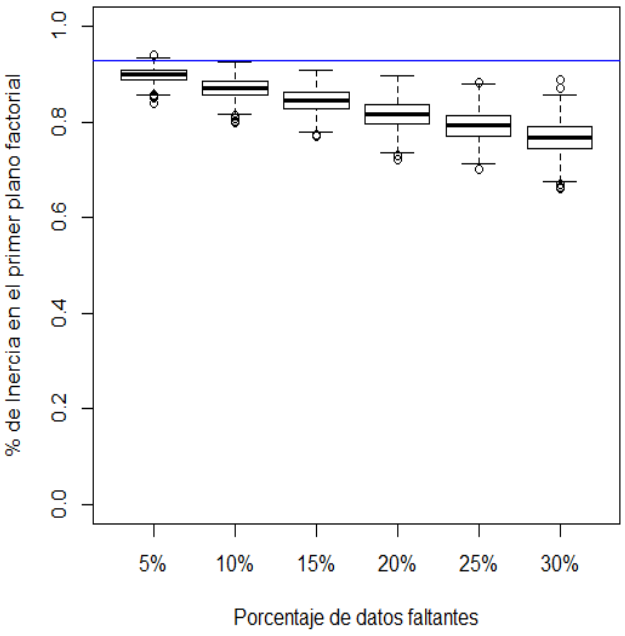


Fig 5. Porcentaje de varianza explicado según el porcentaje de datos faltantes

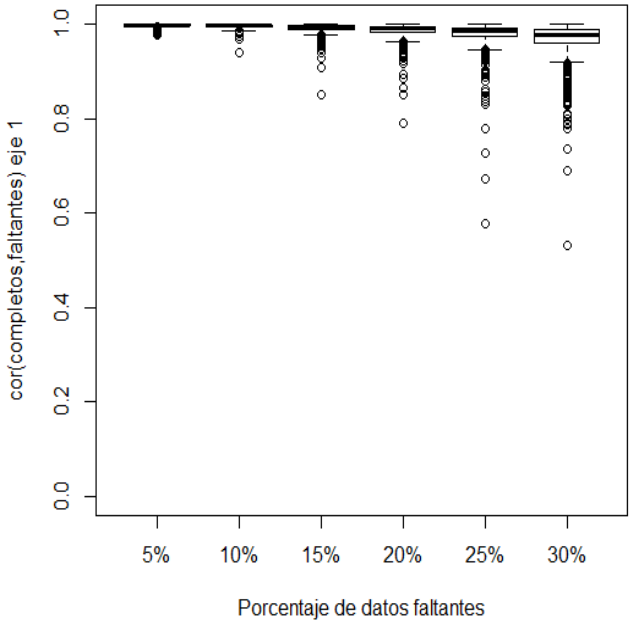


Fig 6. Correlación entre las componentes de las variables con datos completos y faltantes eje 1

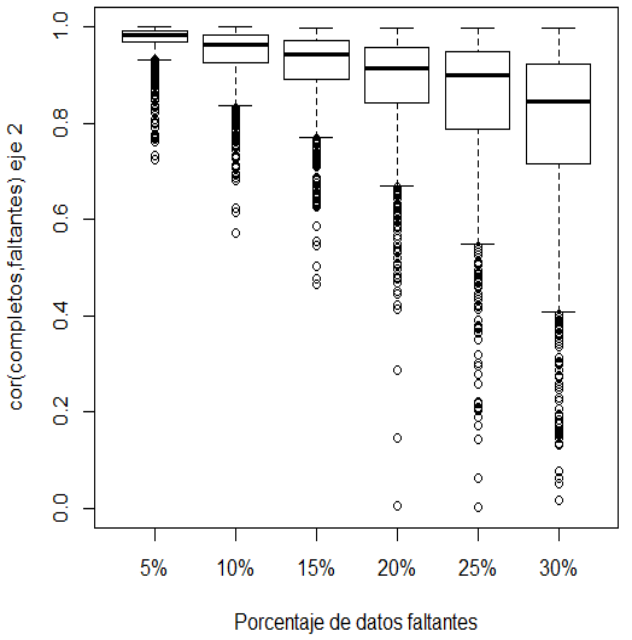


Fig 7. Correlación entre las componentes de las variables con datos completos y faltantes eje 2

IV. CONCLUSIONES

La función fACPna() que solo tiene como parámetro de entrada la matriz de datos X con datos completos o faltantes, permitió desarrollar un ACP conservando todas las propiedades asociadas con las relaciones de transición. Se usó el principio de datos disponibles de NIPALS sin hacer imputación de datos, lo cual es una alternativa diferente al manejo de NA y es una solución que se está implementando en modelos avanzados [13].

Del proceso de simulación se concluye que a mayor porcentaje de datos faltantes conlleva más perdida de inercia y de poder descriptivo en el primer plano principal, resultados que también fueron encontrados en otros estudios [2, 12]

También se encontró que a mayor porcentaje de datos faltantes se disminuye la correlación entre las primeras componentes con datos completos y faltantes; esto se observa mucho más marcado en el eje 2, en donde hay correlaciones muy pequeñas. No obstante, se destaca que las medianas de las correlaciones del proceso de simulación son aceptables, puesto que son superiores a un 0.80 lo cual indica que la estimación con nuestra propuesta fACPna() no se aleja mucho del valor real.

Se destaca la ortonormalidad de los vectores propios conseguidos con el método propuesto en este trabajo de investigación; a diferencia de los vectores propios que se obtienen usando la función nipals() de librerías de R (ade4, plsdepot, entre otros).

Se recomienda contrastar los resultados con los obtenidos con los derivados del paquete FactoMineR que imputa los datos faltantes con la media disponible de las variables.

Para trabajos futuros puede ser útil extender el método propuesto a los modelos de regresión, métodos de clasificación supervisada y no supervisada. Además, es importante considerar como trabajar el método propuesto en el contexto de grandes volúmenes de datos (Big Data).

REFERENCIAS

[1]. Lebart L, Morineau A. y Piron, M. Statistique Exploratoire Multidimensionnelle,Dunod, Paris. 2006
[2]. [Ochoa-Muñoz, A. F., González-Rojas, V. M., & Pardo, C. E. Missing data in multiple correspondence analysis under the available data principle of the NIPALS algorithm. *Dyna*, *86*(211), 249-257. 2019 https://doi.org/10.15446/dyna.v86n211.80261
[3]. Tenenhaus, M. Le Regression PLS: Théorie et Pratique. EditionsTechnip. Paris. 1998
[4]. Team, R. C. R: A language and environment for statistical computing. 2013.
[5]. Aluja, T. Aprender de los Datos: El Análisis de Componentes Principales. EUB. Barcelona. 1999.
[6]. Diaz, L. G. Estadística Multivariada Inferencia y Métodos. Universidad Nacional de Colombia, Facultad de Ciencias. 2002.
[7]. Trejos-Zelaya, J., Castillo-Elizondo, W., & Gonzáles-Varela, J. Análisis multivariado de datos: métodos y aplicaciones. Editorial UCR. 2014.
[8]. Pardo, C.E., Ortiz J. Análisis Multivariado de datos en R. Universidad Nacional de Colombia, Bogotá. 2007
[9]. Hair, J.F.;Black, B.; Babin,B.;Anderson, R.E.; Tatham, R. Multivariate Data Analysis. 6th Edition. 2005.

[10]. Wold, H. Nonlinear iterative partial least squares (NIPALS) modelling: some current developments. In *Multivariate Analysis–III* (pp. 383-407). 1973
[11]. González Rojas, V. M. Análisis conjunto de múltiples tablas de datos mixtos mediante PLS. Tesis Doctoral, Universidad Politécnica de Cataluña. 2014
[12]. González Rojas, V. M. Inter-Battery Factor Analysis via PLS: The Missing Data Case. Revista Colombiana de Estadística, *39*(2), 247-266. 2016 https://doi.org/10.15446/rce.v39n2.52724
[13]. [13] Patel, N., Mhaskar, P., & Corbett, B. Subspace based model identification for missing data. AIChE Journal, 66(10), e16538. 2020. https://doi.org/10.1002/aic.16538
[14]. Sanchez, G., & Sanchez, M. G. Package 'plsdepot'. *Partial Least Squares (PLS) Data Analysis Methods*, *v. 0.1*, *17*. 2016.
[15]. https://cran.r-project.org/web/packages/plsdepot/plsdepot.pdf

Víctor Manuel González-Rojas, es profesional en Estadística desde 1990, Magíster en Estadística desde 2003 y Doctor en Estadística. Obtuvo su título de Doctor en 2014 en la Universidad Politécnica de Cataluña, España y ha trabajado como profesor (actualmente titular) de la Escuela de Estadística de la Universidad del Valle, Colombia durante casi 20 años. Sus principales temas de interés son el análisis de datos multivariados y las series temporales.
ORCID: https://orcid.org/0000-0002-6526-7879

Gabriel Conde-Arango. es Matemático desde 1976, profesional en Estadística desde 1990, Magíster en Ingeniería de Sistemas desde 1998. Ha trabajado como profesor asistente de la Escuela de Estadística de la Universidad del Valle, Colombia durante casi 20 años. Sus principales temas de interés son los procesos estocásticos, la simulación y la modelación estadística.

ORCID: 0000-0002-4299-0212

Andrés Felipe Ochoa-Muñoz, es profesional en Estadística desde 2014, y Magíster en Estadística desde 2018, de la Universidad del Valle, Colombia. Actualmente realiza sus estudios de Doctorado en Estadística de la Universidad de Valparaíso, Chile. Y sus temas de interés son la modelación estadística, métodos de clasificación y el análisis multivariado.
ORCID: https://orcid.org/0000-0002-0003-1347

ANEXOS

Anexo 1. Código en R para ACP con datos faltantes.

```
fACPna   <- function(X)
{
   Xi <- X; ni <- 0;  p <- ncol(Xi);  n <- nrow(Xi)
   Mj <- colMeans(Xi,na.rm=TRUE)
   Sj <- colSums(Xi,na.rm=TRUE)
   nj <- sj/Mj    # vector con n° d.d en c|colj
   X. <- t(t(scale(Xi))/sqrt(nj-1))
   P <- matrix(0,p,p);  p1o <- matrix(0,1,p)
   T <- matrix(0,n,n);  t1o <- matrix(0,n,1)
   for(h in 1:p)    #     X.'X.   cuasi. correlacion   disponible
   {
        t1 <- X.[,h]
      for(j in 1:p)
      {
            j1 <- na.omit(cbind(X.[,j],t1))
      p1o[j] <- sum(j1[,1]*j1[,2])
        }
        P[h,] <- p1o
   }
   for(i in 1:n)     #    X.X.'
   {
        p1 <- X.[i,]
         for(m in 1:n)
        {
            i1 <- na.omit(cbind(X.[m,],p1))
            t1o[m] <- sum(i1[,1]*i1[,2])
        }
            T[i,] <- t1o
    }
   deP<- eigen(P); Lu <- deP$values; U <- deP$vectors
   deT<- eigen(T);  Lv <- deT$values[1:p]     # Lv=Lu
           V <- deT$vectors[,1:p]
   RACPna <- list(Lu,Lv,U,V);  return(rACPna)
} #    end    fACPna()
```

Anexo 2. Código en R para generar datos faltantes aleatoriamente

```
   Xo <- read.table("troue.txt",header=T,row.names=1)

   fmd <- function(Xo,a)    #    a: % NAs , md: miss data
   {
      X. <- as.matrix(Xo)
      n <- nrow(X.); p <- ncol(X.); N <- n*p
      m <- sample(N,round(a*N,0)); d <- length(m)
     for(j in 1:d)   {
       X.[m[j]] <- NA
      }
      return(X.)
   }

   X <- fmd(Xo,0.1)    # X con 10%  NAs
```