## Divulgación

# Comparación de métodos para el tratamiento de información faltante en un análisis de componentes principales sobre datos biológicos.

RECIBIDO: 03/06/2008
ACEPTADO: 19/03/2009

**Quaglino, M. B. • Vitelleschi, M. S.**

Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística. Facultad de Ciencias Económicas y Estadística. Universidad Nacional de Rosario. Bvrd. Oroño 1261. (2000) Rosario. E-mail: mvitelle@fcecon.unr.edu.ar

**RESUMEN:** El Análisis de Componentes Principales (ACP) es una técnica para analizar datos multivariados, utilizada frecuentemente con el propósito de sintetizar y explorar realidades complejas. Clásicamente, su aplicación requiere información completa para todas las observaciones. Si se presentan matrices de datos incompletas, es usual descartar los individuos con información faltante, acarreando probablemente, pérdida de eficiencia.
El algoritmo Nonlinear Iterative Partial Least Squares (NIPALS), propuesto inicialmente en aplicaciones del área química, obtiene las componentes principales directamente a partir de las matrices de datos, aún existiendo pérdidas.
En este trabajo se analizan datos provenientes de niños con Leucemia Linfoblástica Aguda, sobre quienes se registraron mediciones fisiológicas y bioquímicas. Se comparan críticamente los resultados del ACP sobre la matriz original sin información faltante y sobre matrices obtenidas simulando pérdidas al azar, utilizando algoritmo NIPALS y el método de Casos Completos.
**PALABRAS CLAVE:** Análisis de Componentes Principales, algoritmo NIPALS, información faltante.

**SUMMARY:** *Comparison of methods for dealing with missing information in principal component analysis on biological data.*
The Principal Component Analysis (PCA) is a technique for analyzing multivariate data. It is often used to explore and synthesize complex realities.
Its implementation requires complete information for all observations. In the presence of incomplete data, the usual alternative is to exclude the individuals with missing information, deriving probably in a loss of efficiency. The Non-linear Iterative Partial Least Squares (NIPALS) algorithm, which was initially proposed for applications in the chemical area, obtains the principal components directly from data matrixes, even with missing information. This work analyzes information

on children with Acute Lymphoblastic Leukemia, where physiological and biochemical measurements were carried out. PCA results obtained by NIPALS algorithm and the Complete Case method are critically compared utilizing the original data matrix and matrixes constructed with missing information at random.
**KEYWORDS:** Principal Component Analysis, NIPALS algorithm, Missing information.

## Introducción

En distintas áreas del conocimiento es frecuente que para captar mejor la realidad de los fenómenos que se investigan, se midan varias variables sobre muchas unidades de observación, dando origen a una tabla de datos multivariados. Los métodos estadísticos multivariados son adecuados para el análisis de esta información en forma conjunta.

Uno de los problemas que presentan los datos multivariados para su análisis estadístico es el de la dimensionalidad. Si las variables son cuantitativas, el ACP es un método de gran utilidad en las primeras fases de análisis, con enfoque exploratorio, que permite representar esta información a través de un número inferior de nuevas variables no correlacionadas construidas como combinaciones lineales de las originales, llamadas componentes principales o variables latentes, las que facilitan la interpretación de los fenómenos estudiados.

Cuando se recoge información multivariada suele aparecer el problema de los datos faltantes: algunos individuos no tienen el registro de alguna(s) variable(s). Las causas que lo originan pueden ser múltiples. En las investigaciones clínicas, puede producirse cuando algunos pacientes se retiran por tener respuesta terapéutica insatisfactoria, por no responder a preguntas "sensibles", por fallas en equipos con los que se lleva a cabo el estudio, por intolerancia al tratamiento indicado, por errores de trascripción o ilegibilidad de una ficha, etc.. En tales circunstancias se obtiene un conjunto de datos incompletos. Para aplicar en esta situación técnicas estadísticas multivariadas clásicas como ACP, la alternativa estándar es eliminar los registros que presentaron información faltante. Sin embargo, actualmente, el investigador tiene la posibilidad de aplicar una variedad de procedimientos que permiten obtener las componentes principales haciendo uso de toda la información disponible (1-12).

En el presente trabajo se aplica ACP utilizando el algoritmo NIPALS (13) sobre la información proveniente de las historias clínicas de niños con Leucemia Linfoblástica Aguda (LLA) atendidos en un hospital de la ciudad de Rosario durante el año 2003. Con este algoritmo no se requiere, como en los clásicos, el cálculo de la matriz de variancias o de correlaciones previo a la obtención de las Componentes Principales, sino que ellas se derivan directamente a partir de la matriz de datos originales. Si ésta presenta pérdidas en alguna variable, para algún individuo, igualmente se considera la totalidad de la información existente en dicha matriz sin acarrear pérdidas adicionales. Además, este algoritmo se ha mostrado con mejores propiedades que otros, en situaciones, como la tratada en este trabajo, donde el tamaño muestral es pequeño en relación a la cantidad de variables (14).

A fin de evaluar el efecto de posibles pérdidas en la recolección de datos, se comparan los resultados del ACP obtenidos de la

matriz de datos originales (sin información faltante) y de matrices de datos obtenidas simulando al azar un pequeño porcentaje de pérdidas. Estas matrices incompletas se analizan con NIPALS y con el clásico método de Casos Completos que descarta a los individuos con información perdida, produciendo una reducción del tamaño de muestra original. La eficiencia de los métodos se compara a través de los cambios producidos en la estructura de las componentes principales y en la magnitud de la variabilidad representada por ellas, características importantes para su interpretación.

**Material**

La información considerada en este trabajo corresponde a una muestra de 31 niños con Leucemia Linfoblástica Aguda (LLA) atendidos en un hospital de la ciudad de Rosario durante el año 2003. Las variables recogidas fueron: edad (EDAD), recuento de leucocitos (GLOBA), recuento de células indiferenciadas dentro de la sangre periférica (BLAS), cantidad de hemoglobina (HEMO), recuento de plaquetas (PLAQ), incremento del tamaño normal del hígado (HEPA) e incremento del tamaño normal del bazo (SPLE).

**Metodología**

En la literatura clásica de Análisis Multivariado (15, 16, 17) se presenta el Análisis de Componentes Principales como un método de proyección que fue diseñado para resumir y visualizar la variación sistemática de un conjunto de K variables correlacionadas, transformándolo en uno nuevo, de variables no correlacionadas. Estas nuevas variables son combinaciones lineales de las originales, su variancia decrece de la primera a la última y se derivan de forma tal que la primera componente principal explique gran parte de la variación de los datos originales. Luego, se elige la segunda componente principal de modo que sea ortogonal con la primera y explique la máxima variabilidad restante posible, una vez descontada la explicada por la primera componente principal y así sucesivamente. Se procede de esta manera hasta obtener el conjunto total de componentes principales, que coincide con el número de variables originales.

ACP es una técnica matemática que no requiere de modelos estadísticos para explicar la estructura de error. En particular, no se realiza ningún supuesto sobre la distribución de probabilidad de las variables originales, aunque pueden tener más utilidad las componentes principales en el caso que las observaciones provengan de una distribución conjunta normal multivariada.

Sea $\mathbf{X}$ la matriz de datos de N filas y K columnas. Dicha matriz puede considerarse como una colección de K vectores columnas $\mathbf{X}_i^*$ de orden Nx1 que representan a las mediciones de las K variables a través del conjunto de los individuos seleccionados o de N vectores filas $\mathbf{X}_j^T$ de orden 1xK que representan a los individuos u objetos medidos, es decir:

$$\mathbf{X} = \left[ \mathbf{x}_1^* \quad . \quad . \quad . \quad \mathbf{x}_K^* \right] = \begin{bmatrix} \mathbf{x}_1^T \\ . \\ . \\ . \\ \mathbf{x}_N^T \end{bmatrix}$$

La información que se necesita para aplicar ACP, está contenida en la matriz de covariancias ($\mathbf{\Sigma}$). Si algunas de las variables $x_i$, i = 1, ..., K, presentan mucha variabilidad, esto significa que la matriz de covariancias $\mathbf{\Sigma}$ tendrá valores dominantes en su diagonal principal y afectarán los resultados del análisis. En estas situaciones se sugiere, previo al cálculo de las componentes principales,

estandarizar las variables $x_i$, i = 1, ..., K, de modo que la variancia de cada una de ellas sea igual a uno y en este caso las componentes principales se obtendrán a través de la matriz de correlaciones (**R**).

El conjunto de nuevas variables $t_1$, $t_2$, ..., $t_k$, (CP), no correlacionadas entre sí y con variancia decreciente ($\lambda_j$ con j= 1, ..., K) se expresan:

$$t_j = p_{1j}\, x_1 + p_{2j}\, x_2 + \ldots + p_{Kj}\, x_K = \mathbf{p}_j^T\, \mathbf{x}$$

donde $p_j^T = \left[ p_{1j}, p_{2j}, \ldots, p_{Kj} \right]$ es un vector de constantes, al que se le impone que $p_j^T\, p_j = 1$. Este procedimiento de normalización asegura que la transformación global sea ortogonal.

El vector de constantes $\mathbf{p}_j$ resulta ser el vector propio normalizado de la matriz de covariancias $\Sigma$ o correlaciones **R**, asociados al j-ésimo valor propio ($\lambda_j$). Se denota con **P** la matriz ortogonal de orden KxK cuyas columnas son los vectores $\mathbf{p}_h$, variando h=1,2,...,K, habiéndolos previamente ordenado en forma decreciente según los valores propios asociados a ellos. Dicha matriz es $P = [p_1,\ p_2, \ldots,\ p_k]$. A cada $\mathbf{p}_h$ se lo denomina vector de cargas.

Las cargas informan cómo las variables originales son combinadas linealmente para formar las componentes principales, indicando la magnitud (pequeña o grande) y la manera (positiva o negativa) de su aporte en la combinación lineal.

Generalmente, en las aplicaciones de componentes principales es necesario identificar a los individuos en el nuevo espacio de coordenadas. Se denomina vector de "scores" a aquel que contiene las coordenadas de las N observaciones sobre la h-ésima componente principal simbolizándolo con $\mathbf{t}_h$ se tiene:

$$\mathbf{t}_h = \mathbf{X}\, \mathbf{p}_h \qquad h=1,\ldots,K \qquad (1)$$

En la Figura 1 se presentan, esquemáticamente, las matrices que intervienen en la ecuación (1), relacionando las coordenadas de los individuos en ambos espacios.

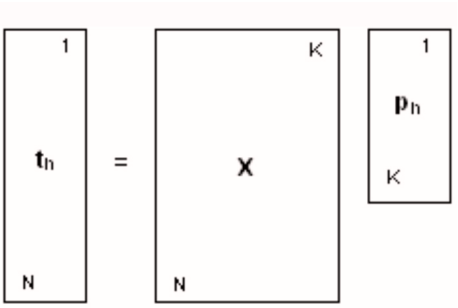
Considerando los K vectores $t_h$ simultáneamente, con h=1,....,K, puede escribirse:

$$\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_K] = [\mathbf{X}\mathbf{p}_1, \mathbf{X}\mathbf{p}_2, \ldots, \mathbf{X}\mathbf{p}_K] = \mathbf{XP}$$

De esta manera se forma la matriz de "scores" **T** de orden NxK, cuyas columnas son los vectores $\mathbf{t}_h$.

Una vez hallado el nuevo espacio de las componentes principales, el próximo paso consiste en elegir un subconjunto de ellas que sean capaces de retener gran parte de la información de la nube de puntos del espacio original. Eso implica determinar el número de componentes principales que serán analizadas. Para tal fin existen diferentes criterios, algunos basados en gráficos, otros a través de tests paramétricos basados en supuestos distribucionales o a través de tests no paramétricos. No existe un criterio que sea mejor en todas las situaciones. Diversos autores proponen aplicar varios criterios simultáneamente y ob-

**Figura 1:** Representación esquemática de la ecuación (1)



servar qué sugieren la mayoría de ellos. La decisión sobre el número de componentes principales a utilizar depende fundamentalmente de cuánta información, medida en términos de variancia no explicada, el investigador está dispuesto a perder. También se debe tener en cuenta el propósito del estudio y la interpretabilidad de las componentes principales que son retenidas en el análisis.

Algunos de los criterios más utilizados que orientan a la selección del número de componentes principales son:

• la proporción de la variancia acumulada, la cual consiste en establecer un porcentaje mínimo de la variación total de los datos originales que se desea explicar con las componentes principales y seleccionar el menor número de ellas que satisface ese porcentaje fijado.

• el gráfico "Scree", basado en un gráfico en el cual se representan en el eje de las abscisas el número de orden del valor propio y en el de las ordenadas los valores propios de la matriz de covariancias (o de correlaciones) ordenados de mayor a menor. La cantidad de componentes principales que se retendrán en el análisis está dada por el número de orden del valor propio donde la línea que los une forma un codo. Dicho punto es denominado punto de quiebre.

• los valores propios mayores que uno, esto se aplica a datos estandarizados y sugiere retener en el análisis sólo aquellas componentes principales cuyos valores propios sean mayores que uno.

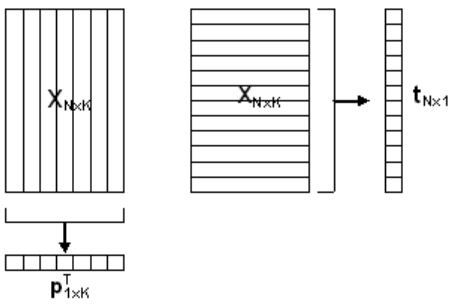
### Algoritmo NIPALS

El algoritmo "Nonlinear Iterative Partial Least Squares" (NIPALS) se diferencia de los algoritmos clásicos en que para obtener cada combinación lineal que origina a cada componente principal, parte de la matriz de datos y no de la matriz de covariancias o correlaciones. Consiste en un método secuencial mediante el cual, en cada ciclo, se calcula una componente principal. Cada iteración de este algoritmo consiste en una regresión lineal de las columnas de la matriz de datos **X** sobre un vector de "scores" **t** para obtener un vector de cargas **p**, seguida de una regresión lineal de las filas de la matriz de datos **X** sobre el vector de cargas para reestimar **t**. Así se continúa hasta que se alcanza la convergencia. Los "scores" y las cargas son proyecciones de la matriz **X** en vectores, es decir cada columna de **X** es proyectada en un elemento del vector p y cada fila de **X** es proyectada en un elemento del vector **t**, siendo esquematizado en la Figura 2 (13).

A continuación se detallan los pasos que se realizan en cada ciclo del algoritmo, h=1,2,...,A; con A≤K (13):

**1.** Se selecciona una cualquiera de las K columnas de la matriz **X** y se la iguala a un vector $\mathbf{t}_h$.

**Figura 2:** Representación esquemática de los "scores" y las cargas obtenidos como proyecciones de la matriz **X** en vectores.



**2.** Se utiliza el vector $\mathbf{t}_h$ para predecir la matriz **X** con el siguiente modelo de regresión: $\mathbf{X} = \mathbf{t}_h \mathbf{b}_h^T + \mathbf{U}$. El estimador mínimo cuadrático de $\mathbf{b}_h^T$ es

$$\hat{\mathbf{b}}_h^T = \left(\mathbf{t}_h^T \mathbf{t}_h\right)^{-1} \mathbf{t}_h^T \mathbf{X}$$

que constituye la proyección de las columnas de **X** sobre la dirección de $\mathbf{t}_h$, definida en el espacio de las N observaciones.

**3.** Se define el vector $\mathbf{p}_h = \hat{\mathbf{b}}_h$.

**4.** Se normaliza el vector $\mathbf{p}_h$ a longitud uno.

**5.** Se utiliza el vector $\mathbf{p}_h$ para predecir la matriz $\mathbf{X}^T$ a partir de un modelo de regresión diferente: $\mathbf{X}^T = \mathbf{p}_h \mathbf{b}_h^T + \mathbf{F}$. Ahora, el estimador mínimo cuadrático de $\mathbf{b}_h^T$ es

$$\hat{\mathbf{b}}_h^T = \left(\mathbf{p}_h^T \mathbf{p}_h\right)^{-1} \mathbf{p}_h^T \mathbf{X}^T \quad ,$$

y su transpuesto es

$$\hat{\mathbf{b}}_h = \mathbf{X}\, \mathbf{p}_h \left(\mathbf{p}_h^T \mathbf{p}_h\right)^{-1}$$

De esta manera se obtuvo la proyección de las filas de la matriz de datos **X** sobre la dirección del vector $\mathbf{p}_h$, definida en el espacio de las K variables.

**6.** Se define $\mathbf{t}_h = \hat{\mathbf{b}}_h$.

**7.** Se calcula la norma cuadrática de la diferencia entre el vector $\mathbf{t}_h$ usado en el paso 1 con el obtenido en el 6.

**8.** Se compara la norma cuadrática obtenida en el paso 7 con algún valor de tolerancia prefijado.

Si la diferencia en el paso 8 es mayor que el nivel de tolerancia se regresa al paso 2; caso contrario, se ha obtenido la h-ésima componente principal.

Se asigna a X el resultado de $\mathbf{X} - \mathbf{t}_h \mathbf{p}_h$ y se vuelve al paso 1. Este proceso se repite A veces.

Una vez completado los A ciclos los vectores $\mathbf{p}_h$ y $\mathbf{t}_h$ son las columnas h-ésimas de las matrices **P** y **T**, respectivamente, y la matriz **E** es el resultado de extraer de **X** la parte explicada por cada una de las A componentes principales, es decir:

$$\mathbf{E} = \mathbf{X} - \mathbf{t}_1 \mathbf{p}_1^T - \mathbf{t}_2 \mathbf{p}_2^T - \ldots . - \mathbf{t}_A \mathbf{p}_A^T \qquad (2)$$

En la mayoría de las situaciones, este algoritmo, converge. Si no hay convergencia es porque existen dos o más valores propios muy similares, en cuyo caso la dirección de las componentes principales no está definida.

Cuando hay datos faltantes en alguna fila o columna de la matriz de datos **X**, la regresión iterativa se desarrolla empleando sólo los datos presentes, es decir ignorando los faltantes. Este procedimiento puede ser interpretado de diferentes formas (2). Una de ellas consiste en que dicho procedimiento es equivalente a asignar en cada iteración el valor nulo a los residuos correspondientes de los elementos perdidos en la función objetivo mínimo-cuadrática [ecuación (2)] o, alternativamente, a reemplazar cada dato faltante por su proyección perpendicular sobre la estimación actual del vector de cargas o "scores" en cada iteración.

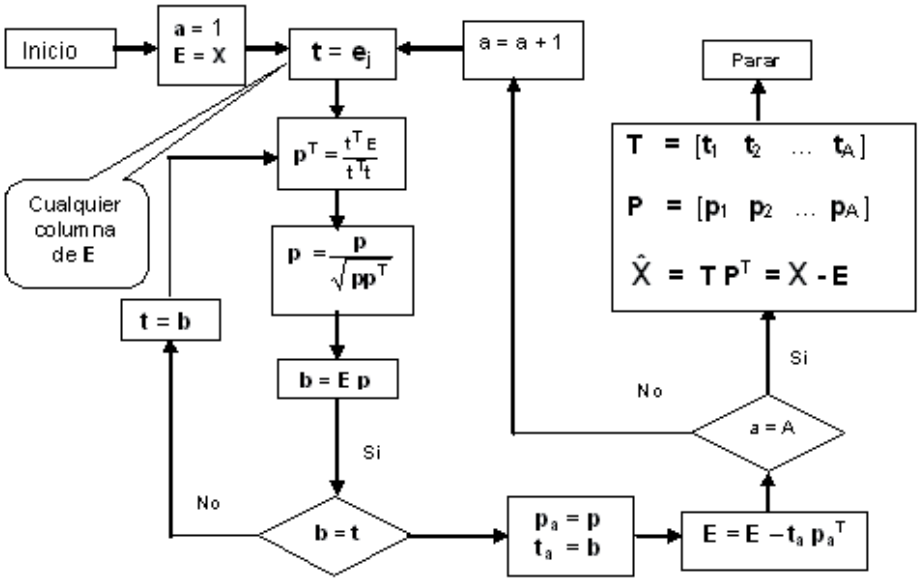
Este algoritmo asume que, en cada componente, los datos faltantes se hallan representados por el apropiado producto entre un vector de cargas y un vector de "scores" sin tener en cuenta las componentes aún no calculadas.

En la Figura 3 se muestra, esquemáticamente, la construcción de un modelo PCA con A componentes a través del algoritmo NIPALS (2).

**Resultados**

Las componentes principales se calcularon a partir de la matriz de datos originales (M1) previa estandarización, a través del

**Figura 3:** Representación esquemática del algoritmo NIPALS para la construcción de un modelo PCA con A componentes.




algoritmo NIPALS con el software SIMCA P-8.0. M1 contiene información de 7 variables sobre 31 individuos. Las variancias de las dos primeras componentes principales fueron $\lambda_1 = 2.551$ y $\lambda_2 = 1.650$, representando, aproximadamente, el 60% de la variabilidad total entre los niños, criterio aceptable para representar con sólo dos nuevas variables latentes, la información contenida en M1. Los coeficientes de ambas componentes se muestran en la Tabla 1.

La primera componente principal identifica el grado de avance de la enfermedad, estando asociada en forma positiva a un mayor número de leucocitos, a un mayor número de blastos y al aumento del tamaño del hígado y plaquetas. La segunda componente principal separa a los niños de acuerdo a la edad, la cantidad de hemoglobina y plaquetas.

Con el objeto de comparar en esta aplicación el efecto que tendría la existencia de faltantes en la información original, se gene-

**Tabla 1:** Coeficientes de las dos primeras componentes principales en la matriz original (M1).

| Componente | Variables | | | | | | |
|---|---|---|---|---|---|---|---|
| | EDAD | GLOBLA | BLAS | HEMO | PLAQ | HEPA | SPLE |
| 1 | -0,105 | 0,474 | 0,502 | -0,289 | -0,265 | 0,354 | 0,483 |
| 2 | 0,464 | 0,347 | 0,367 | 0,572 | 0,448 | -0,007 | -0,027 |

raron pérdidas al azar en un 20% de cada una de las variables: "BLAS" y "SPLE", lo cual significó una pérdida global del 5,5% sobre el total de datos (12 / 217). Las variantes en la reducción del tamaño de muestra en las situaciones simuladas (100 repeticiones) abarcan desde el 20% al 40% del tamaño de muestra original (n=31).

A partir de cada matriz simulada, se calcularon las CP mediante el algoritmo NIPALS y Casos Completos, contrastando los resultados con los derivados a partir del conjunto original. Para evaluar el impacto de la pérdida de información se observaron los cambios producidos en las variancias de las dos primeras CP (valor absoluto de las diferencias) y la distancia euclídea entre sus vectores de cargas. Estas medidas enfocan los aspectos más importantes en la interpretación de un ACP: variabilidad explicada por las componentes y la estructura de la combinación lineal, que define su interpretación. La Tabla 2 resume los resultados obtenidos, a través de los valores máximos y mínimos de discrepancia en las 100 repeticiones.

La Tabla 2 muestra las situaciones extremas en relación a los cambios ocurridos a partir de un Análisis de Componentes Principales con sólo el 5% de información faltante en la matriz correspondiente al problema tratado. Las discrepancias (diferencias de variancias y distancias entre vectores de cargas) entre los ACP a partir de datos con y sin pérdidas son siempre menores cuando se utiliza el algoritmo NIPALS. Con Casos Completos la diferencia de la variancia de la CP1 con respecto del valor 2.551 obtenido en la muestra sin pérdidas osciló, en valores absolutos, entre 0.044 y 0.510. Con NIPALS esa diferencia varió entre 0.006 y 0.324, produciendo una mejora relativa que va del 13.64% al 63.53%. Con respecto de la discrepancia entre los vectores de carga, la ganancia relativa del algoritmo NIPALS vs Casos Com-

**Tabla 2:** Comparación de los ACP obtenidos a partir de la matriz original (M1) y de matrices con pérdidas analizadas según algoritmos NIPALS y Casos Completos.

* Porcentaje de la "diferencia" o la "distancia" de NIPALS respecto de Casos Completos

| **Medidas de comparación** (datos sin pérdida vs datos con pérdida | | **Métodos de obtención de ACP** | | **Indicador de mejora relativa*** |
|---|---|---|---|---|
| | | **NIPALS** | **Casos Completos** | |
| Diferencia | Variancia de CP1 | 0.324 | 0.510 | 63.53 |
| Máxima en la | Variancia de CP2 | 0.113 | 0.154 | 73.38 |
| Distancia | Coeficientes de CP1 | 0.345 | 1.960 | 17.60 |
| Máxima entre los | Coeficientes de CP2 | 0.353 | 0.814 | 43.37 |
| Diferencia | Variancia de CP1 | 0.006 | 0.044 | 13.64 |
| Mínima en la | Variancia de CP2 | 0.004 | 0.031 | 12.90 |
| Distancia | Coeficientes de CP1 | 0.031 | 0.162 | 19.13 |
| Mínima entre los | Coeficientes de CP2 | 0.128 | 0.294 | 43.54 |

pletos fue de alrededor del 18% para la primera CP y de 43% para la segunda.

## Discusión

La existencia de información faltante frente a la aplicación de técnicas multivariadas como ACP, constituye un desafío para el investigador, que no debe ser obviado utilizando procedimientos poco eficientes como la eliminación de las unidades incompletas. En los últimos años este problema ha sido abordando desde distintas perspectivas, surgiendo diferentes propuestas metodológicas (métodos basados en máxima verosimilitud, métodos robustos, bayesianos, NIPALS, EM, entre otros) para el tratamiento de la información faltante (1-12).

El algoritmo NIPALS es un sencillo procedimiento iterativo, fácil de implementar con simples rutinas de estimación mínimo cuadrática de modelos lineales, que puede ser aplicado tanto sobre matrices de datos completas, como con faltantes. Además, las autoras han demostrado que es más eficiente que el EM en aplicaciones de ACP sobre matrices con información faltante de dimensión reducida en cuanto al número de individuos (14, 18 y 19).

El empleo del algoritmo NIPALS permitió analizar con una técnica clásica de análisis multivariado, la información correspondiente a un pequeño conjunto de datos biológicos tanto en el caso de manejar información completa, como cuando en la matriz original se provocaron pérdidas al azar. En este último caso, se pudo considerar la información total disponible sin necesidad de descartar individuos por no tener las mediciones correspondientes a algunas variables.

Se llevó a cabo un estudio por simulación produciendo un porcentaje de pérdidas de sólo un 5% sobre el total de información de la matriz original, seleccionando al azar 6 datos en cada una de dos variables. Esto ocasionó la pérdida de 6 a 12 individuos, lo cual representa entre el 20 y el 40% del tamaño de muestra original. Se evaluó el efecto que produjo esta pérdida de información sobre el ACP obtenido por dos métodos diferentes NIPALS y Casos Completos.

Los resultados obtenidos muestran que en la situación estudiada, si se hubiera reducido el tamaño muestral, las conclusiones hubieran podido variar sustancialmente, tanto en el porcentaje de variancia explicada por las Componentes Principales, como en su interpretación. Las discrepancias entre variabilidades y entre los vectores de coeficientes, son más acentuadas en la primera Componente Principal, siendo ésta la componente más importante.

Este trabajo destaca la importancia, frente a la presencia de información faltante, de utilizar métodos de análisis adecuados, que permitan la inclusión de todos los datos disponibles, sin descartar individuos por no disponer de su información completa, especialmente cuando las muestras son pequeñas.

## Bibliografía

**1.** Nelson, P.; Taylor, P. and MacGregor, J. (1996). "Missing data methods in PCA and PLS: score calculations with incomplete observations". Chemometrics and Intellegent Laboratory Systems, vol. 37, pp. 45-65.

**2.** Arteaga, F. (2003). "Control estadístico multivariado de procesos con datos faltantes mediante análisis de componentes principales". Tesis Doctoral. Departamento de Estadística e Investigación Operativo Aplicadas y Calidad, Universidad Politécnica de Valencia, Valencia, España.

**3.** Kiers, H. A. (1997). "Weighted least squares fitting using ordinary least squares algorithms". Psychometrika, vol. 62, Nº 2, pp. 251-266.

**4.** Walczak, B. and Massart, D. (2001). "Dealing with missing data: Part I". Chemometrics and Intellegent Laboratory Systems, vol. 58, pp. 15-27.

**5.** Stacklies, W.; Redestig, H.; Scholz, M.; Walter, D. and Selbig, J. (2007). "pcaMethods—a bioconductor package providing PCA methods for incomplete data". Bioinformatics, vol 23, Nº 9, pp. 1164-1167.

**6.** Raiko, T.; Ilin, A. and Karhunen, J. (2007). "Principal component analysis for large scale problems with lots of missing values". Lecture Notes in Computer Science, vol. 4701, Springer-Verlag, pp. 691-698.

**8.** Gabriel, K. R. and Zamir, S. (1979). "Lower rank approximation of matrices by least squares with any choice of weights". Technometric, vol. 21, Nº 4, pp. 489- 498.

**9.** Nelson, P.; Taylor, P. and MacGregor, J. (1996). "Missing data methods in PCA and PLS: score calculations with incomplete observations". Chemometrics and Intellegent Laboratory Systems, vol. 37, pp. 45-65.

**10.** Rännar, S.; Geladi, P.; Lindgren, F. and Wold, S. (1995). "A PLS Kernel algorithm for data sets with many variables and fewer objects. Part II: cross-validation, missing data and examples". Journal of Chemometrics, vol. 9, pp. 459-470.

**11.** Rubin. D. (1991). "EM and beyond". Psychometrika, vol. 56, Nº 2, pp. 241-254.

**12.** Stanimirova, I.; Daszykowski, B. and Walczak, B. (2007). "Dealing with missing values and outliers in principal component analysis". Talanta, vol. 72, pp. 172-178.

**13.** Geladi, P. and Kowalski, B. R. (1986). "Partial least squares regression: a tutorial". Analytica Chimica Acta, vol. 185, pp.1-17.

**14.** Vitelleschi, M. (2008). "Modelos PCA a partir de conjuntos de datos con información faltante. ¿Se afectan sus propiedades?". Tesis de Maestría dirigida por la Dra. Marta Quaglino. Facultad de Ciencias Económicas y Estadística, Universidad Nacional de Rosario, Rosario, Argentina.

**15.** Morrison, D. (2004). "Multivariate statistical methods". (4º edición). Duxbury Press.

**16.** Peña, D. (2002). "Análisis de datos multivariantes". McGraw-Hill, New York.

**17.** Wold, S.; Eriksson, L.; Johansson, E. and Kettaneh-Wold, N. (1999). "Introduction to multi- and megavariate data analysis using projection methods (PCA and PLS)". Umetrics, Sweden.

**18.** Quaglino, M. y Vitelleschi, M. (2007). "Multivariate analysis with incomplete information. Characterization of children with leukemia". Biocell, vol. 32, pp. A13.

**19.** Quaglino, M. y Vitelleschi, M. (2008). "Efecto de la pérdida de información sobre las componentes principales obtenidas a través del algoritmo NIPALS". Octavo Congreso Latinoamericano de Sociedades de Estadística. Montevideo, Uruguay. vol. 1, pp.261.