**Vitelleschi, María Susana**
**Quaglino, Marta Beatriz**
*Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística*

# COMPARACIÓN DE MÉTODOS ALTERNATIVOS PARA EL TRATAMIENTO DE DATOS FALTANTES EN LA CONSTRUCCIÓN DE MODELOS PCA

## I- INTRODUCCIÓN

Dada la complejidad de los fenómenos que se investigan en prácticamente todas las ciencias experimentales es usual que se midan varias variables sobre muchas unidades de observación, dando origen a grandes volúmenes de datos. Los métodos estadísticos multivariados son apropiados en estas situaciones ya que analizan simultáneamente toda la información.

Cuando se recogen grandes cantidades de datos es frecuente que presenten información faltante, es decir algunas observaciones no tienen el registro de alguna(s) variable(s). Las causas que lo originan pueden ser múltiples. En tales circunstancias se obtiene un conjunto de datos incompletos al que no es posible aplicarle técnicas estadísticas multivariadas clásicas como Análisis de Componente Principales (ACP), recurriéndose como alternativa estándar a la eliminación de 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.

En el presente trabajo se aplica ACP utilizando el algoritmo EM sobre la información proveniente de una colección de datos industriales. Si la matriz de datos originales presenta pérdidas en algunas variables, para algún individuo, igualmente el algoritmo considera la totalidad de la información existente sin acarrear pérdidas adicionales. Además, se ha demostrado que el algoritmo EM tiene mejores propiedades que otros, en situaciones, como la tratada en este trabajo.

A fin de evaluar el efecto de la presencia de 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 reducidas simulando al azar pérdida de información. Estas matrices incompletas se analizan con EM 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 observaciones procedentes de una planta de clasificación de mineral de hierro bruto de la empresa sueca LKAB. Para el tratamiento del presente trabajo se han seleccionado 175 observaciones y sólo 7 de las 12 variables del proceso, las que tienen información completa. Las variables consideradas fueron: carga total (ctot), carga en el triturador 30 (ctri30), carga en el triturador 40

(ctri40), efecto del triturador 30 (etri30), efecto del triturador 40 (etri40), carga de separador 3 (csep) y desecho de triturado (dtri).

## METODOLOGÍA

Análisis de Componentes Principales es una técnica que permite explicar la variabilidad existente en un conjunto de datos (N) multivariados con un gran número de variables (K) altamente correlacionadas, a través de un número inferior de nuevas variables no correlacionadas construidas como combinaciones lineales de las originales de modo tal que su variancia decrece de la primera a la última. Es decir, ACP es útil para representar, sintéticamente, observaciones de un espacio K-dimensional en otro de dimensión menor (A << K), trasformando las variables originales, en nuevas variables no correlacionadas llamadas Componentes Principales (CP) o variables latentes, las que facilitan la interpretación de los fenómenos estudiados. Las mismas se derivan de forma tal que la primera CP explique la mayor parte posible de la variación de los datos originales. Luego, se elige la segunda CP 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 CP, que coincide con el número de variables originales (K), aunque luego se utilicen unas pocas (A).

Los coeficientes de las CP se denominan cargas e informan cómo las variables originales son combinadas linealmente para formarlas, indicando la magnitud (pequeña o grande) y la manera (positiva o negativa) de su aporte en la combinación lineal.

Establecidas estas condiciones, se deduce matemáticamente que los coeficientes de las CP resultan ser las componentes de los autovectores normalizados asociados a los autovalores de la matriz de covariancias (o correlaciones) ordenados en forma decreciente.

Una particularidad importante a tener en cuenta es que si el cálculo de las CP se realiza con los datos originales, las diferentes escalas en que están medidas las variables originales pueden tener influencia decisiva en la dirección que se obtiene. Por tal motivo, previo al cálculo de las componentes principales, se lleva a cabo una transformación de los valores obtenidos para cada variable, a través del escalado a variancia igual a uno (estandarización), completada sin pérdida de generalidad, por el centrado con respecto al promedio.

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. Esto implica determinar el número A 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 observar qué sugieren la mayoría de ellos. La decisión sobre el número de componentes principales a utilizar depende, fundamentalmente, de cuánta información el investigador está dispuesto a perder, la cual es medida en términos de variancia no explicada. También se debe tener en cuenta el propósito del estudio y la interpretabilidad de las CP que son retenidas en el análisis. En ciertas situaciones, la decisión a priori del investigador será retener sólo una de ellas, la de mayor variabilidad, a fin de disponer de un indicador global que permita ordenar a las unidades de observación (productos, procesos, países, provincias. etc.) según el concepto complejo que la misma representa.

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

- **Criterio de la proporción de la variancia acumulada**: Este criterio determina el

número de componentes principales que serán retenidas en el análisis, estableciendo un porcentaje mínimo, m, de la variación total de los datos originales que se desea explicar con las componentes principales y se selecciona el menor número de ellas que explica al menos ese porcentaje fijado. Dados los valores propios de la matriz de covariancias, ordenados en forma decreciente $\lambda_1 \geq ... \geq \lambda_K$, el porcentaje acumulado por los primeros A valores propios es:

$$z = \frac{\sum_{j=1}^{A} \lambda_j}{\sum_{j=1}^{K} \lambda_j} \ 100$$

Por lo tanto, se elige el mínimo A, tal que $z \geq m$. Este criterio también puede ser aplicado si las componentes principales son calculadas a partir de la matriz de correlaciones, sólo que en este caso $\sum_{j=1}^{K} \lambda_j = K$. Diferentes autores establecen el valor de m entre el 75% y el 85%.

• **Regla del valor propio mayor que uno**: Esta regla es aplicable a datos estandarizados y sugiere retener en el análisis sólo aquellas componentes principales cuyos valores propios sean mayores que uno. Es decir, establece un valor mínimo de la variancia de cada componente principal, el cual es igual al de la variancia de cada variable original.

• **Criterio basado en el gráfico "Scree"**: Consiste en construir 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, de tal forma que a la izquierda del mismo la pendiente de la línea es empinada y a la derecha es suave, convirtiéndose horizontal al eje de las abscisas. Pueden presentarse situaciones donde no es posible identificar el punto de quiebre, dado que la línea quebrada que une a los valores propios se asemeja a una curva suave.

• **Criterio de Horn**: Si las componentes principales fueron obtenidas a través de la matriz de correlaciones y el punto de quiebre del gráfico "Scree" no puede ser determinado, Horn propuso un procedimiento llamado análisis paralelo. El mismo consiste en generar J muestras aleatorias de tamaño N provenientes de distribuciones normales K-variadas con matrices de correlaciones iguales a la identidad. A partir de cada una de estas muestras se realiza un análisis de componentes principales y es de esperar que cada uno de los K valores propios sea igual a uno. Sin embargo, debido a los errores de muestreo algunos valores propios serán mayores que uno y otros menores. Se representa el promedio de los valores propios correspondientes a cada una de las K componentes principales obtenidas a través de las J muestras en un gráfico que a su vez contiene el gráfico "Scree" de los valores propios de la matriz de correlaciones obtenida del conjunto de datos en estudio. Horn establece el punto de corte donde ambos gráficos se interceptan.

• **Prueba de esfericidad de Anderson**: Si las variables originales $x_i$, i = 1, ..., K siguen una distribución conjunta normal, Anderson plantó un test de hipótesis para evaluar si los valores propios de la matriz de covariancias[1], a partir del A+1-ésimo son iguales, es decir que la variabilidad es constante en las últimas (K-A) dimensiones. La igualdad de estos

[1] Se recalca que la distribución derivada por Anderson es sólo válida para los valores propios de la matriz de covariancias.

valores propios señala una nube de puntos esférica en dicho subespacio, en la que no se pueden reconocer direcciones principales de variabilidad. La hipótesis nula evaluada es:

$$H_0 : \lambda_{A+1} = ... = \lambda_K .$$

Si la hipótesis nula es cierta, la estadística:

$$\chi^2 = -(N-1)\sum_{i=A+1}^{K} \log(\hat{\lambda}_i) + (K-A)(N-1)\log\left\{\left(\sum_{i=A+1}^{K} \hat{\lambda}_i\right) \Big/ (K-A)\right\},$$

sigue una distribución asintótica $\chi^2$ con [½ (K-A) (K-A+1) –1] grados de libertad, siendo $\hat{\lambda}_i$ con i = 1, ..., K, los valores propios de la matriz de covariancias estimada a partir de una muestra aleatoria proveniente de una distribución normal.

Las CP constituyen la base de un espacio de menor dimensión sobre el que puede proyectarse la nube de puntos original. Dicha proyección sobre el espacio de menor dimensión permitirá apreciar la presencia de agrupamientos de unidades, observaciones anómalas ("outliers"), etc. Estas últimas se clasifican en fuertes y moderadas. El gráfico de los individuos valorizados en las CP (gráfico de los "scores") se utiliza para detectar los "outliers" fuertes trazando el elipsoide de confianza definido por la estadística $T^2$ de Hotelling asociada a un determinado nivel de confianza. Los "outliers" fuertes aparecerán a una distancia del centro del elipsoide mayor que cualquier punto de dicho elipsoide, es decir son las observaciones que se encuentran fuera del área definida por el mismo. En cambio, los "outliers" moderados se detectan a través de los residuos de cada observación según el modelo de ACP. Para tal fin se establece un valor para el límite inferior y otro para el superior, generalmente –3 y 3 respectivamente. Cada observación será un "outlier" moderado si el valor de su residuo no está comprendido en el intervalo [-3; 3].

Otro gráfico de utilidad es el de los pesos de cada variable en cada componente (gráfico de las cargas), que evidencia las relaciones entre las variables originales. Cuando dichas variables están posicionadas cerca y sobre el mismo cuadrante, ellas están positivamente correlacionadas, mientras que si están posicionadas sobre cuadrantes diagonalmente opuestos, las variables están negativamente correlacionadas. La información que da origen a dicho gráfico, se refiere a la contribución de cada variable original en cada CP. Si la proyección de una variable sobre cada eje de coordenada (el cual representa a una CP) se ubica cerca del origen de coordenadas, esto indica que esa variable contribuye muy poco en la formación de esa CP, mientras que si la proyección se ubica lejos del origen, esta variable tiene una alta contribución en la formación de esa componente principal.

## ALGORITMO EM

El algoritmo EM (Expectation Maximization) es un método iterativo general para obtener la estimación máximo verosímil en una gran diversidad de problemas incluyendo la estimación máximo verosímil en situaciones en las cuales existen datos faltantes y el mecanismo de pérdida es MAR.

Cada iteración del algoritmo EM consiste en un paso E (Esperanza) y un paso M (Maximización). Estos pasos son conceptualmente fáciles de construir y cada uno tiene una interpretación estadística directa. En el paso M se desarrolla la estimación máximo verosímil de los parámetros como si no hubiera información faltante. En el paso E se calcula la esperanza condicional de los estadísticos suficientes, dados los datos observados y la estimación obtenida de los parámetros en el paso anterior.

La forma general del funcionamiento del algoritmo EM consiste en los siguientes pasos:

1. Asignar un valor inicial a los datos faltantes.
2. Estimar los parámetros de interés.
3. Re-estimar los valores perdidos mediante la esperanza condicional, dado el valor estimado de los parámetros y los datos observados.
4. Comparar los resultados de los dos últimos pasos.
5. Regresar al paso 2 si no se alcanza la convergencia.

Cuando el algoritmo EM es aplicado a la construcción de un modelo PCA con datos faltantes adopta la siguiente forma:

1. Asignar un valor inicial a los datos faltantes y designar con $\tilde{\mathbf{X}}_{\mathbf{d}} = \{\hat{x}_{ij,\,paso\,d}\}$ la matriz "completa". Se establece en d=0 al paso inicial.
2. Centrar y escalar la matriz $\tilde{\mathbf{X}}_{\mathbf{d}}$ .
3. Construir el modelo PCA con A componentes principales sobre la matriz $\tilde{\mathbf{X}}_{d}$ y reestimar los datos faltantes a través de:

$$\hat{x}_{ij} = \sum_{a=1}^{A} t_{ia}\, p_{aj}^{T} \quad .$$

La aproximación se denota con: $\tilde{\mathbf{X}}_{\mathbf{d+1}} = \{\hat{x}_{ij,\,paso\,d+1}\}$ .

4. Calcular la diferencia en valor absoluto entre $\hat{x}_{ij,paso\,d}$ y $\hat{x}_{ij,paso\,(d+1)}$ para cada dato faltante. Se realiza la suma de dichas diferencias y se la compara con un valor preestablecido. Si dicha suma es menor que ese valor entonces se ha logrado la convergencia; caso contrario se regresa al paso 2.

Una desventaja de este algoritmo que puede converger en forma muy lenta si existe una gran proporción de datos faltantes.

## IV- RESULTADOS

Las componentes principales se calcularon a partir de la matriz de datos originales (M1) previa estandarización, a través del algoritmo EM. M1 contiene información de 7 variables sobre 175 observaciones. Las variancias de las dos primeras componentes principales fueron $\lambda_1$ = 5,999 y $\lambda_2$ = 0,606, representando, aproximadamente, el 94% de la variabilidad total de los datos, 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. Se observa que todas las variables influyen similarmente en la primera dirección principal, mientras que la segunda componente es fuertemente dependiente de la carga del separador 3 (csep).

**Tabla 1:** Matriz de cargas de las dos primeras componentes principales, calculadas a partir de información completa

| Variables | $\mathbf{p}_{01}$ | $\mathbf{p}_{02}$ |
|---|---|---|
| ctot | 0.397 | 0.013 |
| ctri30 | 0.396 | -0.025 |
| ctri40 | 0.395 | -0.211 |
| etri30 | 0.383 | -0.064 |
| etri40 | 0.382 | -0.280 |
| csep | 0.291 | 0.925 |
| dtri | 0.394 | -0.127 |

Con el objeto de comparar en esta aplicación el efecto que tendría la existencia de faltantes en la información original, se generaron pérdidas completamente al azar en, aproximadamente, un 14% en las diferentes variables de la base de datos disponible (con la restricción de que cada unidad de observación tuviera al menos dos mediciones. Las variantes de pérdida global en las 100 matrices con pérdidas generadas, van desde el 7,6% (93/1225) al 11,3% (138/1225) y las reducciones del tamaño de muestra abarcan desde el 35% (n=114) al 56% (n=77) del tamaño de muestra original (n=175).

A partir de cada matriz en las que se generaron las pédidas, se calcularon las componentes principales a través de EM y un algoritmo clásico que obliga a descartar los individuos con información faltante. 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 matrices que surgieron.

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

| **Medidas de comparación (datos sin pérdida vs datos con pérdida)** | | **Métodos de obtención de ACP** | | **Indicador de mejora relativa*** |
|---|---|---|---|---|
| | | **EM** | **Casos Completos** | |
| Diferencia Máxima en la | Variancia de $CP_1$ | 0,073 | 0,101 | 72,2 |
| | Variancia de $CP_2$ | 0,054 | 0,087 | 62,1 |
| Distancia Máxima entre los | Coeficientes de $CP_1$ | 0,012 | 0,060 | 20,0 |
| | Coeficientes de $CP_2$ | 0,094 | 0,171 | 55,0 |
| Diferencia Mínima en la | Variancia de $CP_1$ | 0,018 | 0,051 | 35,3 |
| | Variancia de $CP_2$ | 0,002 | 0,018 | 11,1 |
| Distancia Mínima entre los | Coeficientes de $CP_1$ | 0,002 | 0,009 | 22,2 |
| | Coeficientes de $CP_2$ | 0,028 | 0,043 | 65,1 |

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

La Tabla 2 muestra las situaciones extremas en relación a los cambios ocurridos a partir de un Análisis de Componentes Principales con pérdida de información 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 EM. Con Casos Completos la diferencia de la variancia de la $CP_1$ con respecto del valor 5,999 obtenido en la muestra sin pérdidas osciló, en valores absolutos, entre 0,051 y 0,101. Con EM esa diferencia varió entre 0,018 y 0,073, produciendo una mejora relativa que va del 35,3% al 72,2%. Con respecto de la discrepancia entre los vectores de carga, la ganancia relativa del algoritmo EM vs Casos Completos varió del 20% al 22,2 para la primera CP y del 55,0% al 65,1% para la segunda.

## V- 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.

La utilización del algoritmo EM para la obtención de las componentes principales frente a información faltante, posibilita considerar el conjunto total de información disponible sin tener que desechar unidades por no tener información completa sobre todas las variables en estudio.

Se llevó a cabo un estudio comparativo entre el algoritmo EM y el método de Casos Completos, generando repetidamente pérdida de información mediante mecanismos aleatorios. En las 100 repeticiones realizadas se produjo pérdidas entre 61 y 98 observaciones, lo cual representa del 35% al 56% del tamaño de muestra original. Se evaluó el efecto que

produjo esta pérdida de información sobre el ACP obtenido por los dos métodos diferentes.

Los resultados obtenidos muestran que en la situación estudiada, si se hubiera reducido el tamaño muestral al usar Casos Completos, las conclusiones hubieran podido variar sustancialmente, tanto en el porcentaje de variancia explicada por las Componentes Principales, como en su interpretación.

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 observaciones por no disponer de su información completa.

## VI- REFERENCIAS BIBLIOGRÁFICAS


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.

Dempster, A.; Laird, N. and Rubin, D.. (1977). "Maximum likelihood estimation from incomplete data via the EM algorithm". Journal of the Royal Statistical Society, Serie B, vol. 39, pp. 1-38.

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.

Little, R. and Rubin, D.. (2002). "Statistical analysis with missing data". Second Edition. Wiley. New York.

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

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

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.



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.



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

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.

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.

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

Schafer, J.. (1997). "Analysis of Incomplete multivariate data". Chapman & Hall. London.

Stacklies, W.; Redestig, H.; Schol,z, 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.

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

Vitelleschi, M. (2008). "Modelos PCA a partir de conjuntos de datos con información faltante. ¿Se afectan sus propiedades?". Facultad de Ciencias Económicas y Estadística, Universidad Nacional de Rosario, Rosario, Argentina.

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

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.