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

MÉTODOS DE PROYECCIÓN MULTIVARIADOS FRENTE A INFORMACIÓN FALTANTE

I- INTRODUCCIÓN

Para captar mejor la realidad de diversos fenómenos complejos que se investigan en distintas áreas, es frecuente que se midan varias variables sobre muchas unidades de observación, dando origen a una tabla de datos multivariados. Si las variables medidas son cuantitativas, esta información puede ser representada gráficamente por un conjunto de puntos (filas de la matriz de datos), en un espacio de dimensión igual al número de variables (columnas de la misma matriz). Este gráfico, de ser posible su visualización, contendría información acerca de los parecidos entre individuos graficados, asociaciones entre las variables medidas, existencia de grupos de individuos, valores extremos o aberrantes, etc. El Análisis de Componentes Principales (ACP) representa la búsqueda de un sub-espacio de proyección que muestre con mínima deformación al conjunto de puntos del espacio original, resumiendo la información brindada por ellos, con mínima pérdida.

El ACP es una de las técnicas de análisis multivariado más utilizadas, ya sea como fin en sí mismo, a fin de explorar realidades complejas, o como paso previo de filtrado para aplicar con mayor eficiencia otras técnicas multivariadas como análisis de conglomerados, modelos lineales, ecuaciones estructurales, etc. La posibilidad de aplicar esta técnica independientemente de supuestos distribucionales, su fácil interpretación a través de simples conceptos geométricos y su robustez frente a distintas escalas de medida, son algunas de las características que han favorecido su amplia difusión. Sin embargo, hasta hace unos años un requisito para su aplicación era la disponibilidad de conjuntos de datos sin información perdida y que el número de elementos observados superara al número de variables. Los trabajos publicados en la década del 80 por Wold, Geladi y Kowalski, en el área de la industria química, superaron los requisitos anteriormente expuestos y plantearon a las componentes principales con un enfoque diferente, no sólo como un método de proyección diseñado para resumir y visualizar la variación sistemática de un conjunto de variables correlacionadas, sino para construir un modelo explicativo de las variables originales en el que las componentes principales intervienen como variables independientes y puede ser utiliza-

do para predecir valores futuros. La construcción de tal modelo se obtiene a partir de un algoritmo iterativo que puede aplicarse aun cuando la matriz de datos tenga información faltante.

La aparición del problema de información faltante es frecuente cuando se trabaja con conjuntos de datos con gran número de variables registradas en muchos objetos. Por ejemplo, en el área de las ciencias sociales, al recoger información por medio de encuestas, es probable que algunos encuestados dejen de responder ciertos items; en el área industrial, pueden ocurrir fallos de comunicación en la transmisión de los datos recogidos por sensores; en el área de ciencias agrarias, algunas parcelas de cultivo bajo experimento podrían ser afectadas por un periodo de sequía; en el área médica, algunos pacientes que participan de investigaciones clínicas podrían retirarse por tener respuesta terapéutica insatisfactoria; etc.. Esta problemática ha recibido mucha atención e importancia en los últimos años, dado que los esfuerzos del analista deben tender a no descartar las unidades con información incompleta porque esto produciría una reducción en los tamaños muestrales con la correspondiente pérdida de precisión en los análisis estadísticos realizados.

En el presente trabajo se aplica la técnica de Componentes Principales sobre la información proveniente de una colección de datos industriales completos, y se evalúa el efecto de posibles pérdidas en la recolección de datos, comparando estos resultados con los derivados de matrices de datos reducidas por pérdidas generadas al azar. Sobre las matrices de datos incompletas se aplican dos algoritmos, NIPALS y un algoritmo clásico, que descarta a los individuos con información perdida produciendo una importante reducción del tamaño de muestra original.

II- METODOLOGÍA

II.1- Análisis de Componentes Principales

En la literatura clásica de Análisis Multivariado se presenta el Análisis de Componentes Principales como un método 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 varianza 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 componen-

te 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.

Sea **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. La información que se necesita para aplicar ACP está contenida en la matriz de covariancias de dichas variables, la cual se expresa, previo centrado de los valores originales en sus promedios, como $S = \frac{XX'}{n}$. Si algunas de las variables x_i, i = 1, ..., K, presentan mucha variabilidad, la matriz de covariancias tendrá valores dominantes en su diagonal principal y este hecho afecta los resultados del análisis. En estas situaciones se sugiere, previo al cálculo de las componentes principales, estandarizar las variables $\mathbf{x}_i$, 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.

Las CP o nuevas variables t_1, t_2, ..., t_k , no correlacionadas entre sí y con variancia decreciente (λ_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 $\mathbf{p}_j^T = \left[\mathbf{p}_{1j}, \mathbf{p}_{2j}, \ldots, \mathbf{p}_{Kj} \right]$ es un vector de constantes, al que se le impone que $\mathbf{p}_j^T\, \mathbf{p}_j = 1$. 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 (o la de correlaciones), asociados al j-ésimo valor propio (λ_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:

$$\mathbf{P} = \left[\mathbf{p_1},\ \mathbf{p_2}, \ldots,\ \mathbf{p_k} \right].$$

A cada $\mathbf{p}_h$ se lo denomina vector de cargas (aún cuando en otras aplicaciones multivariadas, esta palabra se reserva para denotar correlaciones). 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:

$$\mathbf{t}_h = \mathbf{X}\,\mathbf{p}_h \text{ , donde } h = 1, ..., K$$

Considerando los K vectores $\mathbf{t}_h$ simultáneamente, como columnas de una matriz **T**, puede escribirse:

$$\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_K] = [\mathbf{Xp}_1, \mathbf{Xp}_2, \ldots, \mathbf{Xp}_K] = \mathbf{X}\,\mathbf{P}.$$

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 del conjunto 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 como Timm (1975), Morrison (2005), Sharma (1996), entre otros, 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, 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. 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 (regiones, países, empresas, etc.) según el concepto complejo que la misma representa. Si el investigador aplica ACP previo a la utilización de otra técnica estadística, el criterio de selección del número adecuado de componentes principales puede no coincidir con el aplicado cuando ellas son el fin último del análisis. Por ejemplo si se utilizan las variables latentes para aplicar Análisis de Regresión, se debe examinar principalmente, la interpretabilidad de las componentes principales a fin de poder interpretar el modelo. En cambio, si el objetivo principal es reducir el número de las variables originales, para utilizarlas en un Análisis Cluster, el investigador puede retener un número mayor, aunque no sea clara su interpretación, a fin de no perder una cantidad sustancial de información, porque a posteriori, los clusters o grupos pueden interpretarse en término de las variables originales, obviando a las componentes principales utilizadas.

En cualquier caso, se recurre al concepto de retener la información sustancial contenida en la matriz de datos, sin embargo la noción de sustancial es arbitraria, dependiendo del propósito de la investigación. La cantidad no explicada de la variación total de los datos

originales es usada como una medida de pérdida de información. 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.

• **Prueba de esfericidad de Anderson**: Si las variables originales x_i, i = 1, ..., K siguen una distribución conjunta normal, Anderson (1963) plantea 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 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:

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

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

sigue una distribución asintótica χ^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.

• **Prueba ε de Ibanez:** Esta prueba consiste en realizar dos análisis de componentes principales, uno al conjunto formado por las K variables originales x_i, i = 1, ..., K y otro análisis al conjunto constituido por las K variables originales más una variable arbitraria ε que contiene datos generados al azar de una distribución uniforme. Si a partir de la componente principal A+1 la variable arbitraria ε participa con una carga alta, el número significativo de componentes principales no puede ser superior a A, pues las demás componentes principales explicarían una variabilidad inferior a la que es debida a la variable arbitraria ε. Ibanez dio únicamente una justificación empírica de esta prueba, la cual sólo proporciona una cota superior para la dimensión A.

• **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 (1965) 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 pro-

pios de la matriz de correlaciones obtenida del conjunto de datos en estudio. Horn establece el punto de corte donde ambos gráficos se interceptan.

II-2. 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 re-estimar 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. A continuación se detallan los pasos que se realizan en cada ciclo del algoritmo, h=1, 2, ..., A; con A≤K :

1. Se selecciona una cualquiera de las K columnas de la matriz **X** y se la iguala a un vector $\mathbf{t}_h$.
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 $\hat{\mathbf{b}}_h^T = \left(\mathbf{p}_h^T \mathbf{p}_h\right)^{-1} \mathbf{p}_h^T \mathbf{X}^T$ 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 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:

$$\mathbf{E} = \mathbf{X} - \mathbf{t}_1 \mathbf{p}_1^T - \mathbf{t}_2 \mathbf{p}_2^T - .... - \mathbf{t}_A \mathbf{p}_A^T$$

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. 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 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.

Cuando el algoritmo NIPALS aplicado a una matriz cualquiera **X** converge, se demuestra que la solución lograda es igual a la obtenida por medio del cálculo de valores propios y vectores propios de la matriz simétrica $\mathbf{X}\mathbf{X}^T$.

III- RESULTADOS

La información considerada en este trabajo corresponde a una colección de datos industriales reales extraído de la aplicación informática SIMCAP-8.0, que consta de 230 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 sólo 8 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), desecho de triturado (dtri), y desecho total (dtot). Las dos primeras Componentes Principales calculadas a partir de esta matriz de datos originales acumulan el 91.2% del total de la variancia de las 8 variables del proceso y sus variancias son respectivamente λ_1 = 6.644 y λ_2 = 0.712. La matriz de cargas para estas dos CP se muestra 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) y moderadamente del desecho total (dtot).

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	0.379	-0.072
ctri30	0.370	-0.171
ctri40	0.376	-0.196
etri30	0.361	-0.178
etri40	0.361	-0.212
csep	0.265	0.784
dtri	0.371	-0.142
dtot	0.325	0.464

Con el objeto de comparar en esta aplicación el efecto de la existencia de mediciones faltantes en la información original, se generaron pérdidas completamente al azar, aproximadamente, en un 40% 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 simulaciones realizadas, van desde el 6.1% (113/1840) al 6.9% (128/1840) y las reducciones del tamaño de muestra abarcan desde el 47% (n=122) al 58% (n=97) del tamaño de muestra original (n=230). A partir de cada matriz simulada se calcularon las componentes principales a través de NIPALS y un algoritmo clásico que obliga a descartar los individuos con información faltante. Los resultados se compararon cada vez, con los obtenidos a partir del conjunto de datos originales, enfocando dos aspectos diferentes la variabilidad explicada por las componentes y la estructura de la combinación lineal, dos de los aspectos más importantes en la interpretación de un ACP. Dado que en el con-

junto original de datos sin pérdida las dos primeras componentes principales explican el 91% de la variabilidad total, la comparación se limita a ellas.

La Tabla 2 muestra las medidas de discrepancia definidas para la comparación: diferencia máxima y mínima de primer y segundo autovalor (entre el resultado original y el de las cien matrices con pérdidas aleatorias) y diferencias máxima y mínima entre los coeficientes del vector de cargas (idem).

Tabla 2. Comparación de resultados de ACP obtenidos a partir de la matriz original y de matrices con pérdidas analizadas según algoritmos NIPALS y clásico.

Medida de Comparación de ACP sobre matrices completas versus:		**ACP sobre matrices con pérdidas**		**Clásico vs NIPALS (%)**
		Algoritmo NIPALS	**Algoritmo clásico**	
Valores máximos	Diferencia con λ_1	-0.055	0.168	305
	Diferencia con λ_2	-0.019	-0.152	800
	Diferencia entre coeficientes de $\mathbf{p}_1$	0.012	0.030	250
	Diferencia entre coeficientes de $\mathbf{p}_2$	0.073	0.071	973
Valores mínimos	Diferencia con λ_1	-0.005	-0.006	120
	Diferencia con λ_2	-0.002	-0.004	200
	Diferencia entre coeficientes de $\mathbf{p}_1$	0.005	0.008	160
	Diferencia entre coeficientes de $\mathbf{p}_2$	0.023	0.040	174

IV- DISCUSIÓN

La utilización del algoritmo NIPALS para la derivación 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. En los resultados derivados a través de un procedimiento de simulación de pérdidas al azar, los cambios que se hubieran obtenido en un resultado fundamental, como es el del valor de las variancias explicadas por cada componente, son muy importantes. Lo mismo sucede con las componentes del vector de cargas, con las que se estimarían los scores de los individuos en el nuevo espacio de las CP.

En un procedimiento clásico de obtención de las componentes principales, a menos que se realice un paso previo de imputar los datos faltantes, las unidades con información

incompleta hubieran sido descartadas. El trabajo realizado muestra que en la situación estudiada si se hubiera reducido el tamaño muestral, las conclusiones obtenidas hubieran podido variar sustancialmente, tanto en el porcentaje de variancia explicada por las componentes principales como en su interpretación. Si bien estos son resultados limitados por los escenarios elegidos para la comparación, es de prever que puedan extenderse a otras situaciones, indicando al analista, las ventajas de utilizar métodos de estimación que, frente a información incompleta, hagan posible el utilizar toda la información disponible.

V- REFERENCIAS BIBLIOGRÁFICAS

Anderson, T. W.. (1963). "Asymptotic theory for principal component analysis". Annals of Mathematical Statistics, vol. 34 pp. 122-148.

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

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

Sharma, S.. (1996). *"Applied multivariate techniques".* Wiley. New York.

Tatsuoka, M. M.. (1971). *"Multivariate analysis: techiques for educational and psychological research"*. Wiley. New York.

Timm, N. H.. (1975). *"Multivariate analysis with applications in education and psychology"*. Brooks-Cole, Monterey.

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

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

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.

Arteaga, F.. (2003). "Control estadístico multivariado de procesos con datos faltantes mediante análisis de componentes principales". Tesis Doctoral. Dto.de Estadística e Investigación Operativa Aplicadas y Calidad, Univ. Politécnica de Valencia, España.