La selección de variables a través de componentes principales: estudio de un caso

Ester Gutiérrez, Luis Onieva

Dpto. de Organización Industrial y Gestión de Empresas. Escuela Técnica Superior de Ingenieros. Universidad de Sevilla. Calle Enríquez de Ribera, 1. 41092. Sevilla. egm@esi.us.es; onieva@esi.us.es

Resumen

En muchas ocasiones la aplicación de análisis de componentes principales a un número elevado de variables puede dificultar la interpretación de los componentes principales como combinaciones lineales de las variables originales. Este trabajo describe varios métodos de selección de variables e investiga, a partir del estudio de un caso real, qué método es preferible para este objetivo según diversos indicadores de eficiencia. A pesar de que algunos métodos son mejores que otros, el mensaje principal de este trabajo es que la metodología más prudente es confiar en uno o dos métodos, considerando que tanto la relación existente entre las variables como la elección del indicador de información son aspectos importantes en la selección del mejor método.

Keywords: Componentes principales; Selección de variables; Medidas de eficiencia.

1. Introducción

El análisis de datos multivariantes es objeto de estudio en distintas disciplinas, tales como la economía, sociología, ingeniería, medicina y biología, entre otras. En muchas ocasiones, ante la ausencia de conocimientos previos y como medida de seguridad, se seleccionan o miden tantas variables como sea posible con el objetivo de no excluir del estudio ninguna variable que pudiera ser importante. Entonces, el investigador se enfrentará a una gran cantidad de datos agrupados en un número elevado de variables, que denominaremos p, las cuales pueden ser difíciles de analizar e interpretar. Por ello, sería deseable una reducción de la dimensión del espacio, que conserve gran parte de la información de espacio original.

Las técnicas multivariantes que tratan el problema de la reducción de la información suelen utilizarse sin tener en cuenta las posibles alteraciones que pudieran sufrir las características principales y la estructura subyacente del conjunto de datos. El análisis de componentes principales(ACP) es una de las técnicas de reducción que aborda el problema de encontrar un subespacio de dimensión menor que p, tal que al proyectar sobre él los puntos conserve su estructura con la menor distorsión posible. Para este objetivo, el ACP es una de las técnicas habitualmente empleadas con variables continuas, pudiendose encontrar un tratamiento más extenso del procedimiento en los trabajos de Jackson(1991), Jolliffe(2002), y Peña (2002).

ACP tiene como objetivo: dadas n observaciones de p variables, transformar las variables originales, en general correladas, $x_1, x_2, \ldots, x_p$ en un menor número de variables, incorreladas, denominadas *componentes principales* (CPs), $y_1, y_2, \ldots, y_p$ construidas como combinaciones lineales de las originales.

$$\mathbf{y}=\Delta'\mathbf{x} \qquad (1)$$

donde $\boldsymbol{x}$ e $\boldsymbol{y}$ son vectores aleatorios de dimensión p y Δ es una matriz ortogonal pxp cuya *j-ésima* columna , δ_j, es el *j-ésimo* vector propio del *j-ésimo* mayor valor propio de la matriz de varianzas covarianzas de $\boldsymbol{x}$. La transformación mediante CPs, según la expresión (1) conduce a

variables incorreladas, obtenidas mediante la maximización de la varianza de las proyecciones. La varianza del *j-ésimo* CP es el mayor valor propio de la matriz de varianzas-covarianzas, $\mathrm{Var}(y_j)=\lambda_j$. A partir de lo anterior, la proporción de variabilidad explicada por los primeros *k* componentes viene dada por la expresión:

$$\frac{\sum_{j=1}^{k}\lambda_j}{\sum_{j=1}^{p}\lambda_j} \qquad (2)$$

Cuando, para un número reducido de *k*, el ratio dado por la expresión (2) sea elevado, el ACP permite representar óptimamente en un espacio de dimensión *k* observaciones de un espacio general *p*-dimensional. Por otra parte, el ACP proporciona también la posibilidad de eliminar un número de variables originales reteniendo sólo un subconjunto de las *p* variables iniciales.

Las ventajas de seleccionar un subconjunto de variables que representen la variación total del conjunto de variables original es doble: por una parte, proporciona la posibilidad de no tener que evaluar o medir determinadas variables cuyo acceso puede ser difícil y costoso, y por otra, permite una mejor interpretación de los CPs.

El trabajo presentado se estructura como sigue: después de esta breve introducción, se realiza una revisión de los principales métodos de selección de variables, para en la tercera sección describir dos medidas que permiten evaluar los resultados proporcionados por los distintos métodos de selección. En la cuarta sección se efectúa un análisis, a partir de una base de datos real, de la metodología propuesta. Por último, se finaliza el trabajo con las conclusiones más relevantes derivadas del mismo.

2. Métodos de Selección de Variables

En gran parte de las investigaciones es práctica habitual, antes de proceder al análisis detallado de los datos, realizar una reducción previa del número de variables a estudiar. Esta decisión es de gran utilidad para el investigador ya que le permite descartar aquellas variables que introducen complicaciones en el conjunto de datos sin proporcionar suficiente información adicional.

Los intentos de afrontar el problema de reducción de variables se remontan a los trabajos de Beale *et al.* (1967) y Jolliffe(1972,1973), estos autores presentan diversos métodos de selección de conjuntos de *k* variables cuya finalidad es conservar la mayor cantidad de información de los datos originales. Entre estos métodos se encontraban métodos basados en el denominado *análisis de interdependencia* (denotados como A1 y A2), en CPs (B1, B2, B3 y B4) y en análisis de conglomerados (C1 y C2). Algunos de estos métodos fueron comparados a través de datos reales y simulados, siendo aquéllos basados en CPs, los que proporcionaron mejores resultados.

McCabe(1984) introdujo una aproximación diferente al problema mediante el concepto de *variables principales*, asociado al subconjunto de variables que optimiza alguno de los cuatro criterios de optimalidad definidos en su trabajo. De los cuatro métodos propuestos, McCabe(1984) sólo aconseja utilizar el primero cuando se pretenda realizar una búsqueda exhaustiva entre todas las posibles combinaciones de variables.

Posteriormente se presentaron varios procedimientos de selección de variables, como el

Krzanowski (1987) que combinó ACP y análisis Procrustes, o el de Al-Kandari *et al*. (2001) que propusieron cuatro métodos basados en el análisis de la covarianza.

Las propuestas realizadas para resolver el problema de selección de variables han sido varias, este apartado recoge los principales métodos de selección de subconjuntos de variables, muchos de los cuales se basan en el ACP. En cada uno ellos la descripción del metodo aconseja retener k de las p variables originales.

Considérese el componente principal $y_j=\delta_{j1}x_1+\delta_{j2}x_2+\ldots+\delta_{jp}x_p$, donde los coeficientes δ_{j1}, δ_{j2}, …, δ_{jp} se denominan *coeficientes* o *pesos principales*, siendo estos utilizados en la interpretación de los componentes principales. Así, para un determinado componente se considerará que una variable es importante si el valor del coeficiente es elevado, siendo ignoradas o despreciadas las variables con un coeficiente pequeño.

Los métodos de selección de variables analizados en este trabajo se describen brevemente a continuación:

2.1. Método B1

Este método de eliminación de variables fue propuesto por Beale *et al.*(1967, p.359). El método, al cual de ahora en adelante nos referiremos como B1B, asocia una variable a cada uno de los últimos m_1^* (=p-m_1) componentes, según la variable que tenga el coeficiente más elevado en el componente (autovector) bajo consideración y no haya sido anteriormente asociada a un componente previo. Este proceso se continua, en un segundo ACP sobre las m_1 variables restantes, eliminándose m_2^* (=p-m_1-m_2). El proceso finaliza cuando no son necesarias eliminar más variables, es decir, las variables que quedan son las deseadas. También, Jolliffe(1972) se refirió a la versión *forward* (B1F) de este método, sin justificarlo en ningún conjunto de datos reales ni simulados. La descripción del método B1F es similar al método B1B, pero en vez de eliminar variables a partir del último componente, las variables seleccionadas quedan determinadas a partir del primer componente.

2.2. Método B4

Este método descrito en Jolliffe (1972, p.164; 1973, p.22) selecciona las variables asociando una variable, la que tenga el valor de su coeficiente más elevado, a cada uno de los k primeros CPs. Las p-k variables restantes son eliminadas. A diferencia del método B1F sólo se se practica un único ACP.

Al-Kandari *et al*. (2001) propusieron, entre otros, los métodos para seleccionar variables, descritos a continuación:

- Método basado en las combinaciones de las puntuaciones (P1): en este método, las puntuaciones obtenidas de todas las p variables originales son ordenadas en orden decreciente, seleccionando aquellas k variables, con las puntuaciones más elevadas.

- Método basado en el promedio de las puntuaciones (P2): este método selecciona aquellas k variables que tengan, por término medio, las mayores puntuaciones en valor absoluto considerando los q CPs.

Además de estos procedimientos, también se han evaluado otros que se expondrán en el apartado que se inicia a continuación.

3. Medidas de eficiencia

Después de aplicar cada uno de los métodos propuestos de selección de variables, se obtendrá un número extenso de subconjuntos potenciales de variables a retener. Así pues, la pregunta que surge es la de cómo elegir el subconjunto más representativo del conjunto de variables originales. Con el objetivo de facilitar una respuesta, se precisa alguna medida o indicador de eficiencia que evalúe la proximidad entre el conjunto de variables originales y el subconjunto de variables propuestas por cada uno de los métodos. A partir de estas medidas, se considerará que el subconjunto con mayor índice de eficiencia será el mejor subconjunto de tamaño *k* de variables a retener.

Las ventajas asociadas al empleo de estos indicadores son dos, la primera, viene dada de la interpretación que se puede realizar del subespacio definido por los *k* CPs seleccionados, más adecuada que la dada por los CPs considerados individualmente. La segunda ventaja evita la necesidad de preocuparse de los diferentes subconjuntos de variables que es mejor para los diferentes CPs, ya que existe la posibilidad de que la unión de estos subconjuntos pueda conducir a seleccionar un subconjunto de variables de mayor tamaño que el estrictamente necesario. En este trabajo se proporcionará la definición de estos indicadores, una descripción detallada de la elaboración de los mismos se detalla en Cadima *et al*.(2001).

Supóngase que deseamos representar el subespacio definido por los *q* CPs mediante un subconjunto de *k* variables. La matriz de las proyecciones ortogonales en ese subespacio viene expresada por:

$$\boldsymbol{P}_q = \frac{1}{n-1}\boldsymbol{X}\ {}_q^{-1}\boldsymbol{X}' \tag{3}$$

donde $\boldsymbol{S}_{\cdot} = \sum_{i=1}^{q} l_{\cdot}\boldsymbol{a}_{\cdot i}\boldsymbol{a}_{i}'$ es la suma de los primeros *q* términos de la descomposición espectral de $\boldsymbol{S}$, y $\boldsymbol{S}_q = \sum_{j=1}^{q} l_j^{-1}\boldsymbol{a}_j\boldsymbol{a}_j'$ es la inversa generalizada de $\boldsymbol{S}_q$. La matriz de las proyecciones ortogonales en el subespacio definido por el subconjunto de *k* variables es:

$$\boldsymbol{P}_k = \frac{1}{n-1}\boldsymbol{X}\ \boldsymbol{I}_k\boldsymbol{S}_k^{-1}\boldsymbol{I}'_k\ \boldsymbol{X}' \tag{4}$$

donde $\boldsymbol{I}_k$ matriz identidad de orden *k* y $\boldsymbol{S}_k^{-1}$ es la inversa de la submatriz de tamaño (*k*x*k*) de *S* correspondiente a las *k* variables seleccionadas.

La primera medida de proximidad de los dos espacios considera la matriz de correlación entre $\boldsymbol{P}_q$ y $\boldsymbol{P}_k$ definida como:

$$\boldsymbol{P}_k = \frac{1}{n-1}\boldsymbol{X}\ \boldsymbol{I}_k\boldsymbol{S}_k^{-1}\boldsymbol{I}'_k\ \boldsymbol{X}' \tag{5}$$

Esta medida de eficiencia es conocida también como el Coeficiente de Determinación Generalizado de Yanai *C*(*DG*). El indicador *CDG* es la matriz de correlaciones de las proyecciones ortogonales definida por los dos subespacios y puede interpretarse como la media del cuadrado de las correlaciones canónicas entre dos conjuntos de variables que representan a esos subespacios. Como resultado de (5), el *CDG* puede escribirse como:

$$corr\,(\boldsymbol{P}_q, \boldsymbol{P}_k) = \frac{tr\,(\boldsymbol{P}_q \boldsymbol{P}_k)}{\sqrt{tr\,(\boldsymbol{P}'_q \boldsymbol{P}_q)\,tr\,(\boldsymbol{P}_k{}' \boldsymbol{P}_k)}} \tag{6}$$

donde $(r_m)_i$ representa el coeficiente de correlación múltiple entre el *j-ésimo* componente principal y el subconjunto de *k* variables, y el sumatorio se extiende sobre los *q* CPs (*j*=1,...,*q*) seleccionados. Para los distintos subespacios el indicador *CDG* puede alcanzar valores entre cero (si los subespacios son mutuamente ortogonales) y uno (si *q*=*k* y todos los subespacios coinciden). Este indicador se calcula a partir de la correlación múltiple de los CPs con las *k* variables, de forma que valores elevados de *CDG* serán indicativos de un buen ajuste.

El segundo indicador es el coeficiente r_m, y se basa en la semejanza de la descomposición espectral de la matriz de datos original *p*-dimensional y la matriz obtenida por regresión de variables originales sobre un subconjunto *k*-variables. El coeficiente r_m se define según la expresión:

$$r_m = \sqrt{\frac{\sum_{i=1}^{p} \lambda_i \left(r_m\right)_i^2}{\sum_{i=1}^{p} \lambda_i}} \tag{7}$$

donde λ_i representa el *i-ésimo* mayor valor propio de la matriz varianzas-covarianzas y *r* representa el coeficiente de correlación múltiple entre el *i-ésimo* CP y el subconjunto de *k* variables.

El cuadrado de la expresión (7), r_m^2, puede interpretarse como la proporción de variabilidad total explicada por el subconjunto de *k* variables. Además, el indicador r_m es equivalente al criterio basado en minimizar la traza de la matriz de varianzas- covarianzas de las variables descartadas, propuesto por McCabe(1984). Ambos indicadores, (6) y (7), se definen a través de los promedios ponderados del cuadrado de las correlaciones múltiples entre cada CP y el conjunto de variables seleccionadas. En el segundo indicador, las ponderaciones son los autovalores de $\boldsymbol{S}$ y por tanto las varianzas de los CPs. Sin embargo, para el primer indicador las ponderaciones son positivas e iguales para los primeros *q* CPs. Por lo tanto cuando los CPs alcancen valores propios pequeños no será aconsejable utilizar como medida de ajuste el indicador r_m siendo más apropiado, desde el punto de vista de la capacidad explicativa, el indicador CDG.

A partir de un conjunto de *p* variables, el número de veces que habría que calcular estos indicadores depende precisamente de *p*, de forma que cuanto mayor sea *p* más difícil será calcular los indicadores para todos los posibles subconjuntos de *k*-variables. Por ejemplo, para 13 variables el número de todos los posibles subconjuntos viene dado por

2^{13}-1=8191, mientras que para 29 variables supera los 50 millones. Resulta, pues, evidente que examinar todos los subconjuntos posibles es demasiado costoso, en términos de tiempo y computación. Con el objetivo de resolver estas posibles limitaciones son necesarias otras aproximaciones que aunque no garanticen necesariamente encontrar el subconjunto óptimo de variables, sí que proporcionen resultados razonables. Para salvar esta limitación, este trabajo propone diversos procedimientos basados en algoritmos de pasos sucesivos, selección hacia delante y recocido simulado. A continuación se comentarán cada uno de ellos.

Selección de pasos sucesivos (*stepwise*)

La selección paso a paso es quizá el método de selección automática de variables utilizado con mayor profusión, particularmente en el análisis de regresión. Este método no requiere el cálculo de todas las posibles combinaciones de variables. En este trabajo el algoritmo de selección construido para el problema propuesto busca los mejores subconjuntos de variables según los indicadores r_m y *CDG*, los cuales se nombrarán ahora en adelante como Sr y Sg, respectivamente. De esta forma, se determina el mejor subconjunto de k variables para cada uno de los criterios.

3.1. Selección hacia delante *(forward)*

El método de selección hacia delante es una versión simplificada de la selección *stepwise*, ya que si una variable entra a formar parte del subconjunto de variables seleccionadas no podrá ser retirada. Al igual, que en procedimiento *stepwise*, también se ha construido un algoritmo que selecciona las variables según los dos indicadores propuestos, denótense éstos por Fr y Fg, respectivamente. El enfoque de los algoritmos de selección de pasos sucesivos y hacia delante es el adoptado en el trabajo de Neter *et al*.(1990).

3.2. Selección basada en el recocido simulado (*simulated annealing*)

El proceso se inicia con la selección aleatoria de un subconjunto inicial de k variables a partir del conjunto original de p variables. Esta técnica de optimización, a diferencia de otras, permite alcanzar óptimos locales con menor probabilidad, sin garantizar el óptimo global (Aarts *et al*. (1985)).

Los métodos de selección introducidos en este trabajo se pueden clasificar en dos grupos. El primero considera cinco métodos B1B, B1F, B4, P1 y P2, que utilizan ACP y se basan en la magnitud de las coeficientes principales. El segundo considera los algoritmos Sr, Sg, Fr, Fg, Ar, Ag y se encargan de buscar el "mejor" subconjunto o grupo de variables según los dos indicadores de eficiencia anteriormente definidos.

4. Un análisis empírico

Esta sección ilustra la metodología propuesta a partir de una base de datos real, que integra 13 indicadores de energía evaluados por la OCDE, correspondientes al año 2003. Estos indicadores se refieren a todos los países que componen dicha organización, a excepción de Bélgica, Eslovaquia, Hungría, Islandia, Luxemburgo, Méjico y Polonia, ya que para estos no se disponía de una información completa. Las 13 variables, de x_1 a x_{13}, son respectivamente:

x_1-Energía total consumida

x_2-Energía total consumida en el sector industrial

x_3-Energía total consumida en el sector transporte

x_4- Energía total consumida en el resto de sectores

x_5-Energía total generada

x_6-Emisiones de CO_2 procedentes del uso de la energía

x_7-Precio del petróleo crudo importado

x_8-Producción de crudo importado

x_9-Producción total de energía

x_{10}-Contribución al suministro de energía por parte de las energías renovables

x_{11}-Suministro de energía total primaria per capita

x_{12}-Energía total primaria suministrada

x_{13}-Energía total primaria suministrada por unidad de producto interior bruto

Los resultados de los vectores propios a partir de la matriz de correlaciones, los autovalores y la variabilidad explicada por los primeros siete CPs se indican en la tabla 1. Se observa que el primer CP explica el 70.49% de la variabilidad total de los datos, los dos primeros el 89.83% siendo del 99.03% para los cinco primeros componentes. Esta matriz constituye una buena aproximación de la matriz completa de datos de origen. Asimismo, después del sexto vector propio la variabilidad explicada disminuye, lo que muestra la escasa capacidad explicativa de los CPs restantes. El primer componente tiene valores de los coeficientes elevados para diversas variables, destacándose las variables x_5(energía total generada) y x_6(emisiones de CO_2 procedentes del uso de energía). El segundo componente distingue los países respecto a la producción total de energía (x_9) y, en menor medida, respecto a la producción de crudo importado (x_8).

Tabla 1. Vectores propios, autovalores y porcentaje de varianza total explicada por los primeros siete componentes.

Variable	CP1	CP2	CP3	CP4	CP5	CP6	CP7
x_1	0.2019	-0.0532	-0.0577	-0.0520	0.4789	0.0229	-0.0047
x_2	0.0595	-0.0606	-0.0764	-0.1945	-0.1058	0.4098	-0.2738
x_3	0.0809	0.0332	0.0733	0.1632	0.1656	-0.6427	-0.3707
x_4	0.0615	-0.0257	-0.0546	-0.0207	0.4191	0.2558	0.6399
x_5	0.5229	-0.0953	-0.7699	0.0149	-0.2970	-0.0370	0.0040
x_6	0.7365	-0.1473	0.6098	-0.1006	-0.2050	0.0490	0.0596
x_7	-0.0002	0.0038	-0.0029	-0.0007	-0.0275	-0.0022	0.0204
x_8	0.0415	0.3807	-0.0545	-0.8686	0.1531	-0.0932	-0.1392
x_9	0.2040	0.8974	0.0180	0.3529	-0.0303	0.1215	0.0387
x_{10}	-0.0016	0.0463	-0.0281	-0.1555	-0.2014	-0.5603	0.5790
x_{11}	0.0005	0.0067	-0.0063	0.0087	-0.0109	0.0220	0.0049
x_{12}	0.2935	-0.0929	-0.1172	0.1370	0.6004	-0.1067	-0.1340
x_{13}	0.0593	0.0001	-0.0001	0.0005	-0.0004	0.0012	-0.0005
Autovalores	2.4198	0.6640	0.1583	0.0964	0.0611	0.0309	0.0025
% Varianza	70.49	19.34	4.61	2.81	1.78	0.90	0.07

La tabla 2 presenta los resultados de las medidas de eficiencia para cada método de selección de variables. Si examinamos todos los subconjuntos posibles de dos variables, encontramos que la combinación de las variables x_9(producción total de energía) y x_{12}(energía total primaria suministrada) es la elección óptima para los indicadores (4) y (5). El indicador r_m para estas variables coincide con los métodos Ar y Ag y considera un porcentaje elevado, el 99.76%, de la variabilidad total. El indicador *CDG* obtenido a partir de los subespacios definidos por los dos primeros CPs y las variables x_9 y x_{12} es 0.9939. El significado práctico de estos resultados se puede ilustrar mediante la comparación de los 24 países, tal como se representa en la figura 1. El gráfico de la izquierda de la figura 1 representa el conjunto de las observaciones sobre un espacio de bidimensional definido a través de las dos primeras componentes. En el gráfico situado a la derecha se representa el conjunto de países sobre el diagrama de los únicos dos componentes principales de los datos proyectados, $\boldsymbol{P_k X}$, el cual se obtiene mediante la regresión de las 13 variables sobre las variables x_9 y x_{12}. Es decir, el gráfico de la derecha ilustra las estimaciones de la matriz de datos original n x p a partir de la regresión de las variables x_9 y x_{12}. La comparación de ambos diagramas evidencian la proximidad existente, a partir del indicador r_m=0.9986, entre la matriz de datos original y la matriz de datos estimada a partir de las variables x_9 y x_{12}. De esta forma, a través del uso de indicadores de eficiencia se muestra que es posible realizar una simplificación en el análisis de los datos.

Tabla 2. Valores de los indicadores r_m y CDG para los subconjuntos de k variables (k=2, 3, 4, 5).

K	%Varianza *k* PCs	Subconjunto	Método	r_m	%Varianze(r_m^2)	*CDG*
2	89.83%	10,13	B1B	0.2929	8.58%	0.1711
		1,5	B1F	0.9968	99.35%	0.5041
		1,10	B4	0.9966	99.33%	0.6039
		1,12	P1	0.9963	99.25%	0.5167
		3,2	P2	0.9968	99.37%	0.6241
		1,9	Fr-Fg-Sr-Sg	0.9986	99.71%	0.9919
		9,12	Ar-Ag	**0.9986**	**99.72%**	**0.9939**
3	94.44%	8,10,13	B1B	0.8811	77.63%	0.4250
		1,5,12	B1F	0.9970	99.39%	0.6075
		1,10,13	B4	0.9968	99.36%	0.4380
		1,4,12	P1	0.9965	99.31%	0.4061
		3,2,8	P2	0.9984	99.69%	0.7210
		1,5,9	Fr-Fg	0.9995	99.90%	0.9149
		5,6,9	Sr-Sg-Ar-Ag	**0.9998**	**99.96%**	**0.9974**
4	97.24%	7,8,10,13	B1B	0.8867	78.62%	0.3968
		1,5,9,12	B1F	0.9994	99.87%	0.7677
		1,7,10,13	B4	0.9970	99.41%	0.3918
		1,4,12,13	P1	0.9967	99.34%	0.3978
		3,2,8,9	P2	0.9989	99.77%	0.8027
		1,5,6,9	Fr	0.9998	99.96%	0.7515
		1,5,8,9	Fg	0.9994	99.88%	0.9237
		5,6,8,9	Sr-Sg-Ar-Ag	**0.9999**	**99.98%**	**0.9936**
5	99.03%	7,8,9,10,13	B1B	0.9791	95.87%	0.4994
		1,4,5,9,12	B1F	0.9995	99.90%	0.7330
		1,7,8,10,13	B4	0.9984	99.67%	0.4976
		1,4,7,12,13	P1	0.9972	99.44%	0.4404
		3,2,5,8,9	P2	0.9994	99.88%	0.7654
		1,5,6,8,9	Fr-Fg	0.9999	99.98%	0.9615
		5,6,8,9,12	Sr-Sg-Ar-Ag	**0.9999**	**99.99%**	**0.9790**

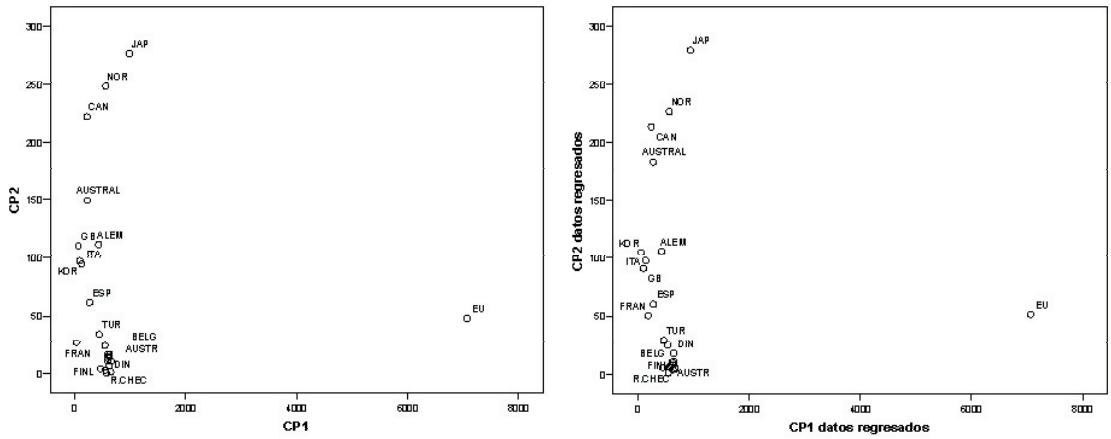

Figura 1. Proyección de las observaciones de energía en los planos definidos por las dos primeras componentes a partir de las 13 variables (diagrama de dispersión bidimensional situado a la izquierda) y regresadas sobre las variables producción total de energía y energía total primaria suministrada (diagrama de dispersión bidimensional situado a la derecha).

En el caso en el que la selección de variables estuviera basada en las variables con mayor

puntuación se elegirían los subconjuntos de variables x_1, x_{12} y x_3, x_2, alcanzándose peores resultados para ambos indicadores peores resultados.

El porcentaje de varianza explicada para tres componentes principales, es de 94.44% y más del 99% es el porcentaje de varianza explicada por la proyección de los datos en el subespacio determinado únicamente por estas tres variables, estas son, energía total generada (variable x_5), emisiones de CO_2 procedentes del uso de la energía (variable x_6) y producción total de energía (variable x_9).

Obsérvese que el conjunto de tres variables *óptimo* no considera la variable x_{12} (energía total primaria suministrada), que sí pertenecía al subconjunto *óptimo* de dos variables. Además, el subconjunto de tres variables obtenido mediante el algoritmo *annealing* proporciona el valor óptimo del indicador GCD, 0.9974. Cabe señalar, que no siempre los resultados conducirán, para cada *k*, a un único subconjunto de variables, tal como ha sucedido en el análisis practicado.

5. Conclusiones

La motivación de este trabajo ha sido doble, por una parte, presentar los principales métodos de selección de variables propuestos hasta ahora y por otra, exponer la posibilidad de medir, a través de indicadores la información suministrada por los distintos subconjuntos de variables propuestos cada uno de estos métodos. Entre las ventajas asociadas a la practica de esta metodología, se podrían citar, la posibilidad de que los datos sean más fácilmente interpretables, la reducción del esfuerzo dedicado en futuros estudios a la recolección de datos, así como la de facilitar las relaciones subyacentes entre las variables.

Los resultados que se muestran en este trabajo permiten concluir que los métodos de selección basados en los algoritmos *stepwise*, *forward* y *annealing* proporcionan mejores subconjuntos de variables desde el punto de vista de la capacidad explicativa del conjunto de datos al que representan.

Finalmente, habría que señalar que el interés central del trabajo se ha basado en explicar, de una forma parsimoniosa, la variabilidad del conjunto de datos original, sin tener pretensiones de realizar un proceso de inferencia estadística. En el caso de ser este el objetivo sería aconsejable incorporar técnicas de validación cruzada al proceso de selección del subconjunto óptimo.

Referencias

Aarts, E.; Korst, J. (1989). Simulated Annealing and Boltsman Machines-A Stochastic Approach to Combinatorial Optimzation and Neural Computing, Chichester_Wiley Interscience Series in Discrete Mathematics and Optimzation.

Al-Kandari, N.M.; Jolliffe, I.T. (2001). Variable selection and interpretation in covariance principal components. Communications in Statistics. 30, 2, 339-354.

Beale et al. (1967). The discarding of variables in multivariate analysis. Biometrika, 54, 3/4, 357-366.

Jolliffe, I.T. (1972). Discarding variables in a Principal Component Analysis I: Artificial Data. Applied Statistics, 21, 2, 160-173.

Jolliffe, I.T. (1973). Discarding variables in a Principal Component Analysis II: Artificial Data. Applied Statistics, 22, 1, 21-31.

Jolliffe, I.T. (2002). Principal Component Analysis, (2nd edition) Springer-Verlag, New York.

Jackson, E.J. (1991).A User's guide to Principal Components, John Wiley and Sons, Inc. New York.

McCabe, G.P. (1984). Principal Variables. Technometrics, 26, 2, 137-143.

Neter, J., Wasserman, W.; Kutner, M.H. (1990). Applied Linear Statistical Models. Chicago: Irwin.

Peña, D. (2002). Análisis de Datos Multivariantes, Mc-Graw Hill, Madrid.

Krzanowski, W.J. (1987). Selection of variables to preserve multivariate data structure, using principal components. Applied Statistics, 36, 22-33.