PESQUIMAT, Revista de la F.C.M. de la
Universidad Nacional Mayor de San Marcos
Vol. II - N°1 Pgs. 77-83 Lima – Perú Ag. 1999

ANÁLISIS DE COMPONENTES PRINCIPALES EN LA DETECCIÓN DE DATOS DISCORDANTES

Seier, E.; Cambillo E.; Cárdenas A.; Adriazola Y.
Laboratorio de Series de Tiempo

RESUMEN

A pesar de ser el análisis de componentes principales, es una de las técnicas multivariantes más antiguas del análisis multivariado, su uso hoy en día sigue en boga y es muy importante. Se presenta su utilización en la detección de datos discordantes y una aplicación a un conjunto de datos socioeconómicos del Perú.

PALABRAS CLAVES: Datos influyentes, datos potencialmente influyentes, primeras componentes principales, últimas componentes principales.

Introducción

El análisis de componentes principales (ACP), consiste en que si las p variables originales son altamente correlacionadas, es posible sustituirlas por q variables ($q<p$) tal que las nuevas variables contengan la mayor información posible contenida en los datos.

La idea básica del ACP [1], es construir combinaciones lineales del vector p variado de tal manera que la primera combinación lineal tenga la mayor variabilidad, más como dicha combinación se establece en base a los coeficientes, el problema no esta claramente definido debido que la varianza puede ser aumentada arbitrariamente por lo que es necesario considerar una combinación lineal normalizada. La segunda combinación lineal debe tener la mayor varianza después de la primera combinación lineal. El procedimiento continua hasta encontrar combinaciones lineales, las cuales deben ser no correlacionadas y de varianza decreciente, dichas combinaciones lineales son llamadas de componentes principales.

El desarrollo matemático del ACP consiste en calcular los autovalores y autovectores de la matriz de covarianza (o de correlaciones), los autovalores son las varianzas de las componentes principales y los autovectores son los coeficientes de las combinaciones lineales, llamadas componentes principales.

Respecto a los datos discordantes, estos pueden ser de diferentes tipos, lo cual muchas veces complica toda búsqueda de las direcciones en las cuales ellos pueden ocurrir, sin embargo existen razones para realizar una inspección visual de las direcciones definidas por

las primeras componentes principales y las últimas para detectar los datos discordantes multivariantes.

Generalmente las últimas componentes principales son más probables a proporcionar información adicional que no ha sido evaluada del gráfico de pares de variables originales, mientras que los gráficos de las primeras componentes principales nos permiten detectar datos discordantes que inflan la varianza y covarianza. No obstante si un dato discordante es la causa del incremento en una o más de las varianzas de las variables originales, entonces dicha observación sería extrema en dichas variables originales y por lo tanto serían fácilmente detectables mirando el gráfico de las variables originales. Muy por el contrario un gráfico de las últimas componentes puede detectar datos discordantes los cuales no ha podido identificarse a partir de las variables originales.

Justificación

Una estructura de correlación fuerte entre variables implica que existe una función lineal de las variables con varianza pequeña (es decir las últimas componentes principales). Supongamos que se tiene dos variables x_1 y x_2 y que existe una correlación fuerte y positiva entre ellas, luego es entonces posible escribir

$$x_2 = \beta x_1 + \varepsilon$$

donde x_1 y x_2 son por ejemplo las alturas y los pesos de los niños, β es una constante positiva, ε es el error, una variable aleatoria con varianza menor que x_1 y x_2, por lo tanto la función lineal de la forma,

$$x_2 - \beta x_1$$

tendrá una varianza pequeña y la segunda componente principal en un ACP de x_1 y x_2 tendrá una forma similar a $a_{22}x_2 - a_{12}x_1$ donde a_{22} y $a_{12} > 0$. La evaluación de las CP para cada una de las observaciones permitirá detectar observaciones las cuales serán datos discordantes con respecto a la estructura de la correlación existente en los datos, aunque no necesariamente con respecto a las variables originales independientemente.

Este argumento puede generalizarse cuando el número de variables es grande ($p>2$), mediante el examen de las últimas componentes principales, las cuales son capaces de detectar observaciones que no se enmarcan dentro de la estructura de correlación impuesta por el conjunto total de datos.

Es a partir de allí que el ACP es una herramienta entre otras, para detectar datos discordantes, dado que si se reduce la dimensión, un análisis gráfico de las primeras componentes, las cuales resumen la mayor variabilidad contenida en los datos, permite detectar datos discordantes. En algunas situaciones las últimas componentes principales permiten detectar datos discordantes que no han sido detectados en un simple gráfico de pares

de variables originales, puesto que las primeras componentes principales son siempre absorbidas por la mayor variabilidad de los datos originales.

Aplicación a datos socioeconomicos

En la presentación de la aplicación del ACP, se utilizara un conjunto de datos socioeconómicos del país, que fue utilizado por [3] identificándose a los departamentos de Lima y Madre de Dios como datos influyentes y potencialmente influyentes respectivamente. Las variables, observadas en los 24 departamentos son:

Porcentaje de viviendas sin construir (VSC)

Ingreso familiar promedio(IFP)

Porcentaje de la población agropecuaria (PEAG)

Tasa de analfabetismo (ANAL)

Tasa de migración (MIG)

Una pregunta que fluye es ¿Los dos departamentos son datos discordantes en todas las variables o en un sub conjunto de ellas? Para responder a dicha interrogante utilizaremos el ACP, en los 24 de departamentos; luego retirando uno a uno los departamentos de Lima y Madre de Dios y los dos departamentos al mismo tiempo.

En el cuadro N° 01, se presenta las matrices de correlación considerando la totalidad de los departamentos(a), retirando el departamento de Lima (b), retirando el departamento de Madre de Dios(c) y retirando los dos departamentos (d).

En la matriz de correlación con todos los departamentos, observamos correlaciones altas y negativas entre el porcentaje de la población agropecuaria y el porcentaje de viviendas sin construir(-0.81), ingreso familiar promedio(-0.82), indicando que en los departamentos en los cuales el porcentaje de población agropecuaria es grande, el porcentaje de viviendas sin construir y el ingreso familiar promedio disminuye; así mismo observamos una tasa de analfabetismo alta (0.85) a medida que el porcentaje de población agropecuaria en el departamento es mayor. También existe correlación alta e inversa entre el ingreso familiar promedio y la tasa de analfabetismo (-0.83), lo cual indica que a menor tasa de analfabetismo es mayor el ingreso familiar promedio. Existen además correlaciones relativamente bajas entre la tasa de migración y las siguientes variables: con la población agropecuaria (-0.55), viviendas sin construir (0.54), ingreso familiar promedio (0.54) y con tasa de analfabetismo(-0.49). Los departamentos con mayor tasa de migración tiene menor población agropecuaria y menor tasa de analfabetismo. Asimismo, departamentos con mayor tasa de migración tienen mayor ingreso familiar promedio y el número de viviendas sin construir crece.

Al retirar el departamento de Lima, disminuye la fuerza de la relación entre la tasa de migración y el porcentaje de la población agropecuaria(-0.35) y el porcentaje de viviendas sin construir(0.14). Esto pues Lima es un departamento con una alta tasa de migración. Al retirar

el departamento de Madre de Dios la correlación es más fuerte entre el ingreso familiar promedio y la tasa de analfabetismo(-0.91). Esto debido a que en Madre de Dios, los ingresos dependen en buen parte de la explotación del oro. Cuando se retiran los dos departamentos, se observa los cambios que cada uno de los departamentos ocasiono cuando fueron retirados independientemente.

En el Cuadro N° 02 se presenta el porcentaje de variación explicada por cada una de las componentes principales obtenidas de las matrices de correlación considerando los cuatro casos. En el cual podemos observar que cuando se retira el departamento de Lima, el porcentaje de variación explicada de la segunda componente principal se incrementa en casi un 20%, incremento que se mantiene cuando se retira ambos departamentos más no cuando se retira solamente el departamento de Madre de Dios, indicando que el departamento de Lima, es un dato discordante que posiblemente infla la varianza y puede ser detectado en el gráfico de las primeras componentes principales.

En el cuadro N° 03 se presentan los coeficientes de la primera componente principal , en los cuatro casos considerados. No se observan diferencias significativas. Este componente puede interpretarse como un indicador que contrasta la tasa de migración, porcentaje de viviendas sin construir, ingreso familiar promedio frente a la población agropecuaria y la tasa de analfabetismo. Valores altos indican que el departamento en cuestión tiene una tasa alta de migración, un alto porcentaje de viviendas sin construir y ingresos familiar mas alto que el promedio.

Los coeficientes de la segunda componente principal se presentan en el cuadro N° 04, observándose cambios en los casos considerados. Cuando se consideran todos los departamentos, esta componente principal puede interpretarse como un indicador de la tasa de migración , tasa de analfabetismo frente el ingreso familiar promedio. Retirando el departamento de Lima puede interpretarse como indicador de la tasa de migración frente a porcentaje de viviendas sin construir. Mientras que al retirar el departamento de Madre de Dios, como un indicador de la tasa de migración y la tasa de analfabetismo para finalmente interpretarse cuando son retirados los dos departamentos como un indicador de la tasa de migración frente el porcentaje de viviendas sin construir. Observándose, que el retiro de cada uno de los departamentos afectan diferente a la segunda componente.

En el cuadro N° 05, se presentan los coeficientes de tercera componente principal, en el caso de considerar todos los departamentos, esta componente puede interpretarse como un indicador de la tasa de migración, el ingreso familiar promedio frente el porcentaje de viviendas sin construir. Al retirar el departamento de Lima, es un indicador entre el ingreso familiar promedio frente a la tasa de migración y el porcentaje de viviendas sin construir. Retirando el departamento de Madre de Dios, puede interpretarse como un indicador del porcentaje de viviendas sin construir, la tasa de analfabetismo. Mientras que cuando se retiran los dos departamentos, esta componente principal puede interpretarse como un indicador de la tasa de migración, porcentaje de viviendas sin construir y la tasa de analfabetismo.

Los coeficientes de la cuarta componente principal, se presenta en el cuadro N° 06, en él se observa que esta componente puede interpretarse como un indicador del ingreso familiar promedio y la tasa de analfabetismo cuando se consideran todos los departamentos. Al retirar Lima, es similar su interpretación, a diferencia que cuando se retira Madre de Dios, se observa como un indicador del ingreso familiar promedio , la tasa de analfabetismo y la población agropecuaria. Cuando son retirados los dos departamentos, la cuarta componente se interpreta similar al caso cuando se retira Madre de Dios.

En el cuadro N° 07, se presentan los coeficientes de la quinta componente principal en los cuatro casos considerados. La quinta componente puede interpretarse como un indicador de la población agropecuaria, el porcentaje de viviendas sin construir y el ingreso familiar promedio cuando se consideran todos lo departamentos. Al retirar Lima, es similar su interpretación; a diferencia que cuando es retirado Madre de Dios, se convierte en un indicador de la tasa de analfabetismo frente la población agropecuaria y el porcentaje de viviendas sin construir. Similar interpretación se da cuando son retirados los dos departamentos.

Lima es un dato discordante que puede identificarse del análisis de las primeras componentes principales, este hecho puede también observarse del gráfico de las primeras componentes principales, tal como se presenta en el gráfico N° 01, en el cual se presenta el gráfico de las dos primeras componentes principales y el departamento que se comporta diferente a la mayoría de los departamentos es Lima; él cual es un dato discordante que infla la varianza y es un dato influyente[3]. El departamento de Madre de Dios, es un dato discordante que puede identificarse del análisis de las últimas componente principales, este hecho puede observarse del gráfico de las últimas componentes principales, tal como se presenta en el gráfico N° 02, en el cual se presenta el gráfico de la tercera y cuarta componente principal y el departamento de Madre de Dios, se comporta en forma diferente a la mayoría de los departamentos; éste es un departamento que no fue posible detectarlos de los gráficos de pares de variables y este es un dato potencialmente influyente [3]

Conclusión

Es posible detectar datos discordantes influyentes y potencialmente influyente utilizando los resultados del ACP, realizando gráficos de las primeras componentes principales para identificar puntos influyentes y gráficos de las últimas componentes principales para identificar puntos potencialmente influyentes.

CUADRO N° 01.- MATRIZ DE CORRELACIÓN

a) Todos los departamentos

	MIG	PEAG	VSC	IFP	ANAL
MIG	1.00				
PEAG	-0.55	1.00			
VSC	***0.54***	**-0.81**	1.00		
IFP	0.54	**-0.82**	0.51	1.00	
ANAL	-0.49	**0.85**	**-0.62**	**-0.83**	1.00

b) Retirando Lima

	MIG	PEAG	VSC	IFP	ANAL
MIG	1.00				
PEAG	-0.35	1.00			
VSC	***0.14***	**-0.76**	1.00		
IFP	0.53	**-0.80**	0.40	1.00	
ANAL	-0.52	**0.85**	**-0.58**	**-0.82**	1.00

c)Retirando Madre de Dios

	MIG	PEAG	VSC	ING	ANAL
MIG	1.00				
PEAG	-0.55	1.00			
VSC	0.56	**-0.89**	1.00		
ING	0.64	**-0.87**	0.75	1.00	
ANAL	-0.49	**0.85**	**-0.67**	**-0.91**	1.00

d) Retirando Lima y Madre de Dios

	MIG	PEAG	VSC	ING	ANAL
MIG	1.00				
PEAG	-0.32	1.00			
VSC	**0.17**	**-0.86**	1.00		
ING	0.56	**-0.83**	0.67	1.00	
ANAL	-0.51	**0.84**	**-0.64**	**-0.91**	1.00

CUADRO N° 02.- PORCENTAJE DE VARIACION EXPLICADA POR LAS COMPONENTES PRINCIPALES

Componente	Todos los departamentos		Retirando Lima		Retirando Madre de Dios		Retirando ambos departamentos	
1	73.18		67.64		78.11		72.38	
2	85.52	12.34	86.79	**19.15**	89.84	11.73	90.87	18.49
3	95.76	10.24	95.40	8.61	97.31	7.47	96.61	5.74
4	98.79	3.04	98.49	3.09	98.81	1.50	98.51	1.90
5	100.00	1.21	100.00	1.51	100.00	1.19	100.00	1.49

CUADRO N° 03. - COEFICIENTES DE LA PRIMERA COMPONENTE PRINCIPAL

Componente	Todos los departamentos	Retirando Lima	Retirando Madre de Dios	Retirando ambos departamentos
MIG	*0.369909*	0.314638	0.360318	0.289668
PEAG	-0.499408	-0.511044	-0.481930	-0.492009
VSC	*0.425768*	0.386630	0.446491	0.427878
ING	*0.458969*	0.480112	0.481444	0.499088
ANAL	-0.470987	-0.509750	-0.454719	-0.491783

CUADRO N° 04.- COEFFICIENTES DE LA SEGUNDA COMPONENTE PRINCIPAL

Componente	Todos los departamentos	Retirando Lima	Retirando Madre dαe Dios	Retirando ambos departamentos
MIG	**0.791246**	**0.73355**	**0.912283**	**0.82740**
PEAG	0.157353	0.24637	0.230209	0.26799
VSC	0.291023	*-0.59609*	-0.090651	*-0.47397*
ING	*-0.346461*	0.20821	-0.066440	0.11288
ANAL	**0.380052**	-0.05023	**0.319548**	-0.07859

CUADRO N° 05 . - COEFICIENTES DE LA TERCERA COMPONENTE PRINCIPAL

Componente	Todos los departamentos	Retirando Lima	Retirando Madre de Dios	Retirando ambos departamentos
MIG	**0.465118**	**-0.585168**	0.081020	*-0.464709*
PEAG	0.223833	-0.123292	0.228770	0.029728
VSC	*-0.728577*	*-0.568128*	*-0.733553*	*-0.631232*
ING	**0.438168**	**0.543115**	**0.345884**	**0.311938**
ANAL	-0.103679	-0.156954	**-0.532318**	*-0.536090*

CUADRO N° 06 . - COEFICIENTES DE LA CUARTA COMPONENTE PRINCIPAL

Componente	Todos los departamentos	Retirando Lima	Retirando Madre de Dios	Retirando ambos departamentos
MIG	-0.111421	0.109027	-0.174606	-0.124535
PEAG	-0.248653	-0.166722	**0.430295**	**0.550827**
VSC	0.074227	0.168419	0.185945	0.217483
ING	**0.554024**	**0.495267**	**0.767539**	**0.733578**
ANAL	**0.783138**	**0.828651**	**0.400832**	**0.369264**

CUADRO N° 07 . - COEFICIENTES DE LA QUINTA COMPONENTE PRINCIPAL

Componente	Todos los departamentos	Retirando Lima	Retirando Madre de Dios	Retirando ambos departamentos
MIG	-0.091361	-0.092728	-0.029280	-0.005161
PEAG	**0.783520**	**0.796956**	***-0.690830***	***-0.617900***
VSC	**0.444628**	**0.379543**	***-0.468770***	***-0.382760***
ING	**0.412769**	**0.431135**	0.234610	0.320500
ANAL	-0.098337	-0.162274	**0.497082**	**0.607390**

GRAFICO N° 01 .- GRAFICO DE LAS DOS PRIMERAS COMPONENTES PRINCIPALES

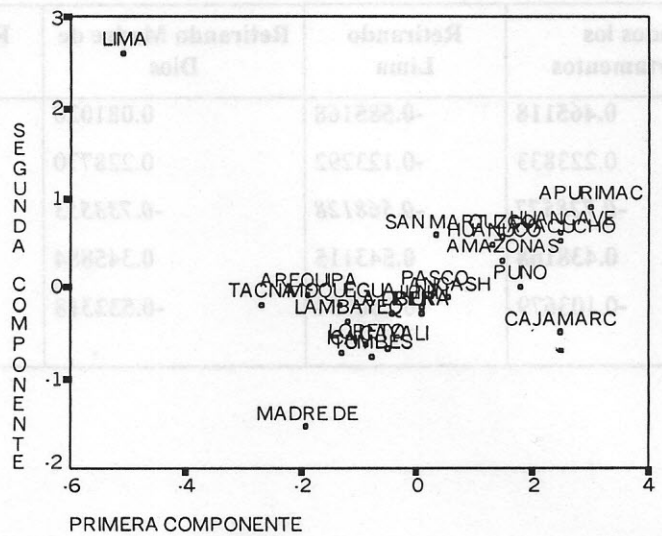
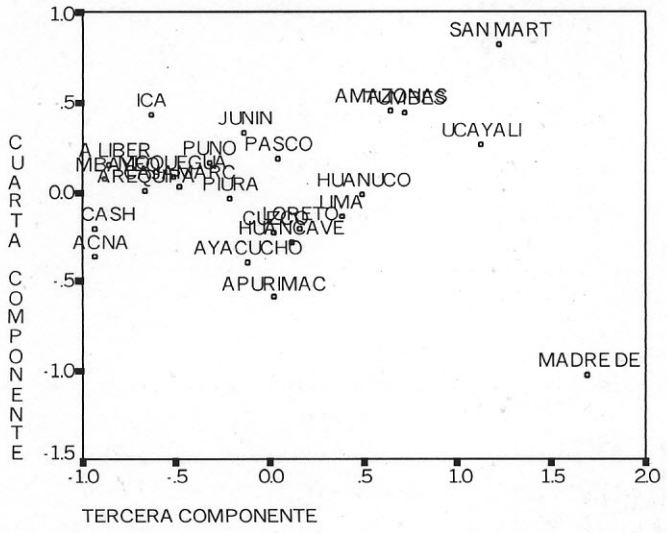


GRAFICO N° 02 .- GRAFICO DE LA TERCERA Y CUARTA COMPONENTE PRINCIPAL



BIBLIOGRAFÍA

1. Jhonson, R.; Wicher, D. (1980) *Applied Multivariate Statistical Analysis.*
2. Jollife, I.T. (1986) *Principal Componente Analysis*. Springer Verlay.
3. Seier,E.; Cárdenas, A.; Cambillo, E.; Adriazola, Y.; Medina, F. (1994). *Análisis de Influencia*. Laboratorio de Series de Tiempo.