Matriz de Varianzas Vs. Matriz de Correlaciones en el Análisis de Componentes Principales: Un Enfoque Analítico

Autor:

Jorge Mauricio Oviedo [1]

Resumen: es bien sabido que el Enfoque de las Componentes Principales presenta ciertas disparidades en sus resultados a la hora de decidir si utilizar la Matriz de Varianzas y Covarianzas o bien utilizar la Matriz de Correlaciones. Muchos argumentos se inclinan por preferir éste último en virtud de que elimina los efectos distorsivos generados por la presencia de variables representadas en distintas unidades de medida. En este artículo intentaremos ahondar la validez de tales conclusiones analizando en estricto rigor algebraico los efectos de utilizar uno y otro método. Así, se hallan fórmulas para los componentes principales en términos de valores paramétricos de las varianzas, covarianzas y coeficientes de correlación y se llevan a cabo estudios analíticos de sensibilidad ante cambio en los parámetros. Sus principales conclusiones son contrastadas en un ejemplo de aplicación de las mismas

Palabras clave: Componentes Principales, Matriz de Covarianzas, Matriz de Correlaciones, Análisis de Sensibilidad.

[1] joviedo@eco.unc.edu.ar

1.- Introducción

Ante el problema usual del análisis descriptivo de datos de reducir la dimensionalidad del problema conservando la mayor cantidad posible de información, el enfoque de las Componentes Principales ofrece una de las mejores respuestas y tal vez una de las mas sencillas y por ende también una de las mas utilizadas. El mismo consiste en hallar un conjunto menor de nuevas variables, combinaciones lineales de las originales, de modo tal que maximicen la variabilidad y por ende que conserven la mayor riqueza estadística de los datos originales.

Sin embargo, es bien sabido que el Enfoque de las Componentes Principales presenta ciertas disparidades en sus resultados a la hora de decidir si utilizar la Matriz de Varianzas y Covarianzas o bien utilizar la Matriz de Correlaciones. Muchos argumentos se inclinan por preferir éste último en virtud de que elimina los efectos distorsivos generados por la presencia de variables representadas en distintas unidades de medida. En este artículo intentaremos ahondar la validez de tales conclusiones analizando en estricto rigor los efectos de utilizar uno y otro método.

Para lleva a cabo dicha tarea se procederá a resolver los problemas de optimización que pretenden buscar una combinación de variables que minimiza la perdida de información o lo que es lo mismo que maximiza varianzas. Se resolverán de manera analítica para matrices de Varianzas y Covarianzas y Matrices de Correlación de tamaño 2 x 2. El trabajar con dicha generalidad nos permitirá hallar fórmulas para los componentes principales en términos de valores paramétricos de las varianzas, covarianzas y coeficientes de correlación. La importancia de lo anterior radica que al ser funciones de los parámetros antes mencionados podemos realizar estudios analíticos de sensibilidad y pudiendo así depurar los efectos de aumentos de variabilidad, de covarianzas y fundamentalmente el efecto unidad de medida.

El trabajo se estructura de la siguiente manera: en la sección siguiente se realizará un breve comentario del Método de las Componentes Principales para inmediatamente en la Sección Tercera proseguir con el tratamiento analítico del Caso de la Matriz de Varianzas y Covarianzas. El análisis para el caso de la Matriz de Correlaciones se trata en la sección Cuarta donde se extraen los principales resultados y se los compara con los de la sección anterior. En la Quinta Sección se proveerá de un caso de aplicación donde se volcarán los principales resultados encontrados. Conclusiones y Bibliografía se detallan al final del escrito.

2.- El Método de las Componentes Principales

Siguiendo a Peña [2002], el problema central en el análisis de datos multivariantes es la reducción de la dimensionalidad: si es posible describir con precisión los valores de p variables por un pequeño subconjunto r < p de ellas, se habrá reducido la dimensión del problema a costa de una pequeña pérdida de información.
El análisis de componentes principales tiene este objetivo: dadas *n* observaciones de *p* variables, se analiza si es posible representar adecuadamente esta información con un número menor de variables construidas como combinaciones lineales de las originales. Por ejemplo, con variables con alta dependencia es frecuente que un pequeño número de nuevas variables (menos del 20 por 100 de las originales) expliquen la mayor parte (más del 80 por 100 de la variabilidad original).
La técnica de componentes principales es debida a Hotelling (1933), aunque sus orígenes se encuentran en los ajustes ortogonales por mínimos cuadrados introducidos por K. Pearson (1901).
Su utilidad es doble:
1. Permite representar óptimamente en un espacio de dimensión pequeña observaciones de un espacio general p-dimensional. En este sentido, componentes principales es el primer paso para identificar las posibles variables latentes, o no observadas que generan los datos.
2. Permite transformar las variables originales, en general correladas, en nuevas variables incorreladas, facilitando la interpretación de los datos.
Se puede analizar desde tres puntos de vistas complementarios:
1) Enfoque descriptivo: Se desea encontrar un subespacio de dimensión menor que *p* tal que al proyectar sobre él los puntos conserven su estructura con la menor distorsión posible. Por ejemplo, si se trabaja en un espacio de dos dimensiones (*p*=2), se puede obtener una recta que pase por cerca de todos los puntos y las distancias entre ellos se mantengan aproximadamente en su proyección sobre la recta. Para conseguir esto, se debe exigir que las distancias entre los puntos originales y sus proyecciones sobre la recta sean lo mas pequeña posible.

2) Enfoque estadístico: representar puntos *p* dimensionales con la mínima pérdida de información en un espacio de dimensión uno es equivalente a sustituir las *p* variables originales por una nueva variable, *z* que resuma óptimamente la información. Esto supone que la nueva variable debe tener globalmente máxima correlación con las originales o, en otros términos debe permitir prever las variables originales con la máxima precisión. Desde este enfoque lo que se busca es una recta que minimiza las distancias ortogonales entre los puntos y la recta.
3) Enfoque geométrico: Si se consideran una nube de puntos, se tiene que estos puntos se sitúan alrededor de la recta formando una elipse o elipsoide (si se trabaja con una dimensión mayor a 2) y se pueden describirlos por su proyección en la dirección del eje mayor de la elipse.

Cálculo de los componentes

Supongamos que se dispone de los valores de *p*-variables en *n* elementos de una población dispuestos en una matriz **X** de dimensiones *n* × *p,* donde las columnas contienen las variables y las filas los elementos. Supondremos que previamente hemos restado a cada variable su media, de manera que las variables de la matriz **X** tienen media cero y su matriz de covarianzas vendrá dada por 1/*n* **X'X**.
El primer componente principal se define como la combinación lineal de las variables originales que tiene varianza máxima. Los valores de este primer componente de los n individuos se pueden representar por un vector **z** , tal que:

$$\mathbf{z} = \mathbf{Xc}$$

Si las variables originales tienen media cero, entonces **z** también tendrá madia nula. Su varianza será:

$$(1/n)\mathbf{z'z} = (1/n)\mathbf{c'X'Xc} = (1/n)\mathbf{c'Sc}$$

donde **S** es la matriz de varianzas y covarianzas de las observaciones. Dado que se puede incrementar sin límite la varianza con solo aumentar indefinidamente el módulo de **c**, se debe imponer una restricción al módulo de este último, y sin pérdida de generalidad se puede suponer que **c'c** =1.
De esta manera el problema de optimización lucirá de la siguiente manera:

$$\max_{\{\mathbf{c}\}} V = \mathbf{c'Sc}$$
$$st: \ \|\mathbf{c}\| = 1$$

Trabajando con los multiplicadores de Lagrange, derivando e igualando a cero se tiene:

$$\max_{\{\mathbf{c}\}} L = \mathbf{c'Sc} - \lambda(\mathbf{c'c} - 1)$$
$$\frac{\partial L}{\partial \mathbf{c}} = 2\mathbf{Sc} - 2\lambda\mathbf{c} = \mathbf{0}$$

Cuya solución es:

$$\boldsymbol{Sc} = \lambda \boldsymbol{c}$$

Que implica que c es un vector propio de **S** y λ su correspondiente valor propio. Para determinar que valor propio de **S** es el que maximiza la varianza, premultiplicamos ambos miembros de la condición de primer orden por **c**' con lo que se obtiene:

$$\mathbf{c'Sc} = \lambda\mathbf{c'c} = \lambda$$

De lo que se deduce que λ es la varianza de **c**. Cómo es ésa justamente la cantidad que se quiere maximizar, λ será el mayor valor propio de **S.** Su vector propio asociado, **c,** define los coeficientes de cada variable en la primer componente principal.

Puede demostrarse análogamente que el espacio de dimensión *r* que mejor representa a los puntos viene definido por los vectores propios asociados a los *r* mayores valores propios de **S**. Estas direcciones se denominan direcciones principales de los datos y a las nuevas variables por ellas definidas componentes principales. En general, la matriz **X** (y por tanto la **S**) tiene rango *p*, existiendo entonces tantas componentes principales como variables que se obtendrán calculando los valores propios o raíces características, $\lambda 1, \ldots, \lambda p$, de la matriz de varianzas y covarianzas de las variables, **S**, mediante:

$$\left|\mathbf{S} - \lambda \mathbf{I}\right| = 0$$

y sus vectores asociados son:

$$(\mathbf{S} - \lambda_i \mathbf{I})\mathbf{c_i} = \mathbf{0} \qquad i = 1,..p$$

Los términos $\lambda\imath$ son reales, al ser la matriz **S** simétrica, y positivos, ya que **S** es definida positiva. Por ser **S** simétrica si λi y λ_j son dos raíces distintas sus vectores asociados son ortogonales. Llamando **Z** a la matriz cuyas columnas son los valores de los *p* componentes en los *n* individuos, estas nuevas variables están relacionadas con las originales mediante:

$$\mathbf{Z = XA}$$

donde **A'A** = **I**. Calcular los componentes principales equivale a aplicar una transformación ortogonal **A** a las variables **X** (ejes originales) para obtener unas nuevas variables **Z** incorreladas entre sí. Esta operación puede interpretarse como elegir unos nuevos ejes coordenados, que coincidan con los "ejes naturales" de los datos.

Propiedades de los componentes

1.- Conservan la variabilidad inicial: la suma de la varianza de los componentes es igual a la suma de las varianzas de las variables originales, y la varianza generalizada de los componentes es igual a la original.
2.- La proporción de variabilidad explicada por un componente es el cociente entre su varianza, el valor propio asociado al vector propio que lo define, y la suma de los valores propios de la matriz.
3.- Las covarianzas entre cada componente principal y las variables **X** vienen dadas por el producto de las coordenadas del vector propio que define el componente por su valor propio:

$$Cov(z_i; x_1, ..., x_p) = \lambda_i c_i = (\lambda_i c_{i1}, ..., \lambda_i c_{ip})$$

donde c_i es el vector de coeficientes de la componente z_i.
4.- La correlación entre un componente principal y una variable **X** es proporcional es proporcional al coeficiente de esa variable en la definición del componente, y el componente de proporcionalidad es el cociente entre la desviación típica del componente y la desviación típica de la variable.
5.- Los *r* componentes principales ($r<p$) proporcionan la predicción lineal óptima con r variables del conjunto de variables *p*.
6.- Si estandarizamos los componentes principales, dividiendo cada uno por su desviación típica, se obtiene la estandarización multivariante de los datos originales.

Análisis normado o con correlaciones

Los componentes principales se obtienen maximizando la varianza de la proyección, pero cuando las escalas de medidas de las variables son muy distintas, la maximización de la varianza dependerá decisivamente de estas escalas de medida y las variables con valores más grandes tendrán más pesos en el análisis. Para solucionar este inconveniente, se debe estandarizar las variables entes de calcular los componentes, de manera que las magnitudes de los valores numéricos de las variables **X** sean similares. Así, en vez de utilizar la matriz de Covarianzas (**S**) se debe utilizar la Matriz de Correlación (**R**) en donde las varianzas son estandarizadas a uno y sus covarianzas sustituidas por los coeficientes de correlación. De esta manera el problema resolver será:

$$\max_{\{\mathbf{c}\}} V = \mathbf{c}'\mathbf{R}\,\mathbf{c}$$

$$st: \ \|\mathbf{c}\| = 1$$

Por lo tanto, la solución depende de las correlaciones y no de las varianzas.
Los componentes principales normados se obtienen calculando los vectores y valores propios de la matriz **R**, de coeficientes de correlación.
Siguiendo a Peña [2002], cuando las variables **X** originales están en distintas unidades conviene aplicar el análisis de la matriz de correlaciones. Cuando las variables tienen las mismas unidades, ambas alternativas son posibles. Si las diferencias entre las varianzas de las variables son informativas y queremos tenerlas en cuenta en el análisis, no debemos estandarizar las variables: por ejemplo, supongamos dos índices con la misma base pero uno fluctúa mucho y el otro es casi constante. Este hecho es informativo, y para tenerlo en cuenta no se deben estandarizar las variables, de manera que el índice de mayor variabilidad tenga más peso. Por el contrario, si las diferencias de variabilidad no son relevantes se eliminan con el análisis normado. En caso de duda, conviene realizar ambos análisis, y seleccionar aquél que conduzca a conclusiones más informativas.

Interpretación de los componentes

Un aspecto clave en el Análisis de las Componentes Principales es la interpretación de los factores, ya que ésta no viene dada a priori, sino que será deducida tras observar la relación de los factores con las variables iniciales (habrá, pues, que estudiar tanto el signo como la magnitud de las correlaciones). Esto no siempre es fácil, y será de vital importancia el conocimiento que el experto tenga sobre la materia de investigación.
Sin embargo, para que un factor sea fácilmente interpretable debe tener las siguientes características, que son difíciles de conseguir:
i) los coeficientes factoriales deben ser próximos a 1.
ii) Una variable debe tener coeficientes elevados sólo con un factor.
iii) No deben existir factores con coeficientes similares
Por otro lado, cuando existe una alta correlación positiva entre todas las variables, el primer componente principal tiene todas sus coordenadas del mismo signo y puede interpretarse como un promedio ponderado de todas las variables, o un facto global de tamaño. Las restantes componentes se interpretan como factores “de forma” y típicamente tienen coordenadas positivas y negativas, que implica que contraponen unos grupos de variables frente a otros. Estos factores de forma pueden frecuentemente escribirse como medias ponderadas de dos grupos de variables con distinto signo y contraponen las de un signo a las del otro.

Selección del número de componentes

Hay distintas reglas para seleccionar el número de componentes:
i) Realizar un gráfico de λ_i frente a *i*. Comenzar seleccionando componentes hasta que los restantes tengan aproximadamente el mismo valor de λ_i La idea es buscar un “codo” en el gráfico, es decir, un punto a partir del cual los valores propios son aproximadamente iguales. El criterio es quedarse con un número de componentes que excluya los asociados a valores pequeños y aproximadamente del mismo tamaño.
ii) Seleccionar componentes hasta cubrir una proporción determinada de varianza, como el 80 o el 90 por 100. Esta regla es arbitraria y debe aplicarse con cierto cuidado. Por ejemplo, es posible que un único componente de “tamaño” recoja el 90 por 100 de la variabilidad y, sin embargo, pueden existir otros componentes que sean muy adecuados para explicar la “forma” de las variables.

3.- Estudio Analítico del Método de las CP's utilizando la Matriz de Varianzas-Covarianzas

En esta sección llevaremos a cabo el estudio analítico del Método de las componentes Principales de la manera más general posible. Para ello hallaremos la estructura analítica de las mismas en función de las Varianzas de dos variables aleatorias y su respectiva covarianza. Así, suponiendo que existen sólo dos variables aleatorias se trata de encontrar una combinación lineal de ellas de modo tal que dicha combinación lineal conserve la mayor variabilidad posible de los datos originales. Resolviendo entonces, el siguiente problema en notación matricial:

$$\max_{\{c_1,c_2\}} V = [c_1, c_2] \begin{bmatrix} \sigma_1^{\,2} & Cov \\ Cov & \sigma_2^{\,2} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$

$$st: \quad \left\| [c_1, c_2] \right\| = 1$$

O bien en notación escalar

$$\max_{\{c_1,c_2\}} V = c_1^{\,2}\sigma_1^{\,2} + c_2^{\,2}\sigma_2^{\,2} + 2c_1c_2Cov$$

$$st: \quad c_1^{\,2} + c_2^{\,2} = 1$$

Donde c_1, c_2 representan las ponderaciones de las variables originales en las componentes principales, es decir las ponderaciones tal que maximicen la varianza.

Resolviendo el problema anterior en términos algebraicos por medio de Maple 8, hallamos los valores de las componentes principales en función de las varianzas y covarianzas paramétricas del problema general. Resolviendo, simplificando y agrupando términos se arriban a los siguientes expresiones que indican como están conformadas las CP

```
T:=eigenvectors(A):
A:=factor(simplify(T[1,3][1])):
[c[1],c[2]]=factor(simplify(T[1,3][1]));
factor(simplify(T[2,3][1]));
```

Coeficientes de la Primer Componente Principal[2]

$$[c_1, c_2] = \left[\frac{1}{2} \frac{\sigma_1^{\,2} - \sigma_2^{\,2} + \sqrt{\sigma_1^{\,4} - 2\,\sigma_1^{\,2}\,\sigma_2^{\,2} + \sigma_2^{\,4} + 4\,Cov^2}}{Cov}, 1 \right]$$

Como puede apreciarse los coeficientes de primer componente principal depende de ambas varianzas y de su covarianza. Se puede dilucidar que mientras mayor sea la varianza de la primer variable mayor, mayor será la participación de ésta en la primer componente principal.

Coeficientes de la Segunda Componente Principal

$$\left[-\frac{1}{2} \frac{-\sigma_1^{\,2} + \sigma_2^{\,2} + \sqrt{\sigma_1^{\,4} - 2\,\sigma_1^{\,2}\,\sigma_2^{\,2} + \sigma_2^{\,4} + 4\,Cov^2}}{Cov}, 1 \right]$$

[2] Aquí se ha supuesto que la varianza de la primer variable es mayor que la de la segunda y el segundo coeficiente se ha normalizado a uno en lugar de tomar módulo unitario para el vector **c**.

Asimismo se computan también las expresiones de los valores propios, que, como se mostró en la sección anterior indican el valor de la varianza de cada una de las componentes principales.

Valores Propios

```
> with(linalg):
A:= [[(sigma[1])^2,Cov],[Cov,(sigma[2])^2]]:
lambda[1] = eigenvalues(A)[1];
lambda[1] = eigenvalues(A)[2];
```

$$\lambda_1 = \frac{1}{2}\sigma_1{}^2 + \frac{1}{2}\sigma_2{}^2 + \frac{1}{2}\sqrt{\sigma_1{}^4 - 2\,\sigma_1{}^2\,\sigma_2{}^2 + \sigma_2{}^4 + 4\,Cov^2}$$

$$\lambda_2 = \frac{1}{2}\sigma_1{}^2 + \frac{1}{2}\sigma_2{}^2 - \frac{1}{2}\sqrt{\sigma_1{}^4 - 2\,\sigma_1{}^2\,\sigma_2{}^2 + \sigma_2{}^4 + 4\,Cov^2}$$

Expresiones que a su vez pueden expresarse como:

$$\lambda_1 = \frac{\sigma_1{}^2 + \sigma_2{}^2}{2} + \frac{\sqrt{(\sigma_1{}^2 - \sigma_2{}^2)^2 + 4Cov^2}}{2}$$

$$\lambda_2 = \frac{\sigma_1{}^2 + \sigma_2{}^2}{2} - \frac{\sqrt{(\sigma_1{}^2 - \sigma_2{}^2)^2 + 4Cov^2}}{2}$$

De las mismas se desprende que la varianza de la primer componente será mayor en la medida que mayor sean la diferencia entre las varianzas de ambas variables y mayor su covarianza.

Proporción de la Varianza Explicada por la Primer Componente:

Otro elemento útil a analizar es la Proporción (*P*) de la varianza explicada por la Primer Componente, elemento que computamos a continuación en función de todos los parámetros del problema.

```
P:=lambda[1]/(lambda[1]+lambda[2])=simplify(T[1,1]/(T[1,1]+T[2,1]));
```

$$P := \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{1}{2}\,\frac{\sigma_1{}^2 + \sigma_2{}^2 + \sqrt{\sigma_1{}^4 - 2\,\sigma_1{}^2\,\sigma_2{}^2 + \sigma_2{}^4 + 4\,Cov^2}}{\sigma_1{}^2 + \sigma_2{}^2}$$

Análisis de Estática Comparativa:

Habiendo previamente calculado las expresiones de los valores y vectores propios como así también de las proporciones de la varianza total explicada por la primer componente principal, estamos pues en condiciones de efectuar un preciso análisis de sensibilidad. Para ello, responderemos a la pregunta de que suceden con las expresiones anteriores si aumenta la varianza de la primer variable o si aumenta la covarianza entre ambas.

Efectos sobre las coordenadas de la primer componente principal ante cambios en la varianza de la primer variable

```
sigma[1]->c[1](sigma[1]);
diff(c[1](sigma[1]),sigma[1])=factor(simplify(diff(A[1],sigma[1])));
Cov->c[1](Cov);
diff(c[1](Cov),Cov)=factor(simplify(diff(A[1],Cov)));
```

$$\frac{d}{d\sigma_1} c_1(\sigma_1) = \frac{\sigma_1 \left(\sigma_1^{\,2} - \sigma_2^{\,2} + \sqrt{\sigma_1^{\,4} - 2\,\sigma_1^{\,2}\,\sigma_2^{\,2} + \sigma_2^{\,4} + 4\,Cov^2}\right)}{\sqrt{\sigma_1^{\,4} - 2\,\sigma_1^{\,2}\,\sigma_2^{\,2} + \sigma_2^{\,4} + 4\,Cov^2}\;Cov}$$

Como se puede observar, si la varianza de la primer variable es mayor que la de la segunda, un aumento de la primer varianza ocasiona un aumento en la ponderación que recibe la primer componente principal en la primer variable.
Aquí merece la pena efectuar una importante observación. Muchas veces un simple cambio en las unidades en que se han realizados las mediciones de una variable puede contribuir a aumentar la varianza de la misma. En otras palabras, un simple cambio de unidad de medida introduce, vía los cambios en la varianza, un distorsión positiva en las componentes principales pudiendo desvirtuar la utilidad del Método Estadístico. En este sentido es altamente aconsejable en los casos en que las variables estén expresadas en diferentes unidades de medida trabajar con la Matriz de Correlaciones, cuyo tratamiento analizamos mas adelante, para aislar así los efectos nocivos de las distintas unidades de medida.
Por otro lado, en los casos en que las variables se encuentren en la misma unidad de medida, debe asegurarse que la información que proveen las diferencias en las varianzas al análisis estadístico sea un elemento útil al propósito del análisis, para usar el enfoque de la Matriz de Varianzas y Covarianzas. Esto es así, ya que al usar la Matriz de Correlación no solo se anulan los efectos de unidad de medida si no toda otra información referente a la diferencia de varianza que, en algunos casos, podría aportar riqueza al estudio bajo análisis y su no inclusión podría distorsionar el significado y en análisis de las componentes.

Efectos sobre las coordenadas de la primer componente principal ante cambios en la Covarianza

$$\frac{d}{dCov} c_1(Cov) = -\frac{1}{2} \frac{(\sigma_1 - \sigma_2)(\sigma_1 + \sigma_2)\left(\sigma_1^{\,2} - \sigma_2^{\,2} + \sqrt{\sigma_1^{\,4} - 2\,\sigma_1^{\,2}\,\sigma_2^{\,2} + \sigma_2^{\,4} + 4\,Cov^2}\right)}{\sqrt{\sigma_1^{\,4} - 2\,\sigma_1^{\,2}\,\sigma_2^{\,2} + \sigma_2^{\,4} + 4\,Cov^2}\;Cov^2}$$

Como se observa, si $\sigma_1 > \sigma_2$ un aumento en la correlación entre las variables reduce la participación de la 1er variable en la 1er componente. Esto a primera vista parece contradictorio, pero si analizamos la función objetivo de donde surgen las componentes principales como solución del problema de optimización, tenemos

$$\max_{\{c_1\}} V = c_1^{\,2}\sigma_1^{\,2} + (1 - c_1^{\,2})\sigma_2^{\,2} + 2c_1\sqrt{1 - c_1^{\,2}}\,Cov$$

Así, observando la función objetivo y suponiendo que $\sigma_1 > \sigma_2$ la función tiene dos factores por lo cual puede crecer: por un lado el efecto σ_1, el cual se maximiza concediendo valores a c_1 cercanos a uno, y por otro lado el efecto Covarianza el que exigen elegir valores de c_1 cercanos a 0.5. De

esta manera, el proceso de optimización asignará a c_1 valores entre 0.5 y 1 dependiendo de cuan fuerte sea el efecto covarianza.
Visto de este modo, un incremento de la Covarianza, hace que el segundo efecto mencionado anteriormente sea mayor conduciendo al proceso de optimización a seleccionar valores mas cercanos a 0.5, y por lo tanto menores que antes. Esto explicaría porque un aumento en la correlación entre las variables reduce la participación de la 1er variable en la 1er componente

Efectos sobre la proporción de la Varianza Total explicada por la Primer Componente Principal ante un cambio en la varianza de la Primer variable.

```
VV:=simplify(T[1,1]/(T[1,1]+T[2,1])):
QQ:=eval(VV,sigma[1]^2=a):
AA:=eval(QQ,sigma[1]^4=a^2):
sigma->P(sigma);
diff(P(sigma),sigma)=eval(factor(diff(AA,a)),a=sigma[1]^2);
Cov->P(Cov);
diff(P(Cov),Cov)=factor(simplify(diff(simplify(VV),Cov)));
```

$$\frac{d}{d\sigma}\mathrm{P}(\sigma) = -\frac{\sigma_2{}^4 - \sigma_1{}^2\,\sigma_2{}^2 + 2\,Cov^2}{\sqrt{\sigma_1{}^4 - 2\,\sigma_1{}^2\,\sigma_2{}^2 + \sigma_2{}^4 + 4\,Cov^2}\,(\sigma_1{}^2 + \sigma_2{}^2)^2}$$

Recordando que $Cov^2 = \rho^2\sigma_1{}^2\sigma_2{}^2$

$$\frac{d}{d\sigma_1}P(\sigma_1) = \frac{\sigma_1{}^2\sigma_2{}^2(1-2\rho^2) - \sigma_2{}^4}{\sqrt{(\sigma_1{}^2 - \sigma_2{}^2)^2 + 4Cov^2}\,(\sigma_1{}^2 - \sigma_2{}^2)^2}$$

Expresión que indica que si estamos en presencia de una fuerte correlación entre las variables un incremento de la Varianza de la Primer Variable tiende a reducir la proporción explicada por la Primer Componente Principal.
Este hecho tiene importantes implicancias en materia práctica:
En primer lugar, si las varianzas de las variables son mayores que uno, utilizar la Matriz de Correlaciones, que indirectamente estandariza las varianzas de todas las variables a uno, tiende a incrementar la proporción de la varianza total explicada por la Primer componente
En segundo lugar, si las varianzas son menores que uno (hecho que puede suceder cuando las variables están expresadas en porcentajes tal y como sucede en el ejemplo que ilustraremos al final de trabajo), utilizar la Matriz de Covarianzas reduce la proporción explicada por la primer componente dificultando así la validez del método, el proceso de selección del número de componentes y fundamentalmente su interpretación.

Efectos sobre la proporción de la Varianza Total explicada por la Primer componente ante un cambio en la covarianza

```
Cov->P(Cov);
diff(P(Cov),Cov)=factor(simplify(diff(simplify(VV),Cov)));
```

$$\frac{d}{dCov}\mathrm{P}(Cov) = \frac{2\,Cov}{\sqrt{\sigma_1^{\,4} - 2\,\sigma_1^{\,2}\,\sigma_2^{\,2} + \sigma_2^{\,4} + 4\,Cov^2}\,(\sigma_1^{\,2} + \sigma_2^{\,2})}$$

Expresión que simplemente nos dice que si la covarianza aumenta, aumenta el porcentaje explicado por la 1er Componente Principal, resultado que es altamente intuitivo.

4.- Estudio Analítico del Método de las CP's utilizando la Matriz de Correlaciones

En el caso de que se trabaje con la Matriz de Correlaciones, el problema de optimización del cual surgen las componentes principales luce como sigue:

$$\max_{\{c_1,c_2\}} V = \begin{bmatrix} c_1, c_2 \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$

$$st: \quad \left\| \begin{bmatrix} c_1, c_2 \end{bmatrix} \right\| = 1$$

O bien en notación escalar

$$\max_{\{c_1,c_2\}} V = c_1^{\,2} + c_2^{\,2} + 2c_1c_2\rho$$

$$st: \quad c_1^{\,2} + c_2^{\,2} = 1$$

Resolviendo el problema anterior en términos algebraicos por medio de Maple 8, hallamos los valores de las componentes principales en función del coeficiente de correlación paramétrico del problema general. Resolviendo, simplificando y agrupando términos se arriban a los siguientes expresiones que indican como están conformadas las CP

```
> B:=[[1,rho],[rho,1]];
 BB:=eigenvalues(B):
 lambda[1]=BB[1];
 lambda[2]=BB[2];
 BBB:=eigenvectors(B):
 [c[1],c[2]]=BBB[2,3][1];
 [c[1],c[2]]=BBB[1,3][1];
 lambda[1]/(lambda[1]+lambda[2])=factor(simplify(BB[1]/(BB[1]+BB[2])));
```

Primer Componente Principal

$$[c_1, c_2] = [1, 1]$$

Segunda Componente Principal

$$[c_1, c_2] = [-1, 1]$$

Valores Propios

$$\lambda_1 = 1 + \rho$$

$$\lambda_2 = 1 - \rho$$

Proporción de la Varianza Explicada por la Primer Componente:

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{1}{2} + \frac{\rho}{2}$$

De esta manera puede observarse que en el caso de utilizar la Matriz de Correlaciones los efectos de las varianzas son eliminados pues las mismas han sido estandarizadas a 1. Esto tiene la ventaja que se eliminan los problemas originados en el uso de diferentes unidades de medida pero por otro lado elimina información referente a la diferencia de varianzas entre variables que en algunos casos puede resultar ser una información útil de considerar.
En cuando a los efectos de modificar el coeficiente de correlación, las participaciones de cada variable en la conformación de las componentes principales son fijas e independientemente de las correlaciones, conformando una recta de 45º con los ejes originales.
Sin embargo, un incremento en la correlación de las variables aumenta el porcentaje de la varianza total explicada por la Primer Componente Principal4.

Conclusiones Generales:

De lo analizado anteriormente resulta aconsejable en los casos en que las variables estén expresadas en diferentes unidades de medida trabajar con la Matriz de Correlaciones, para aislar así los efectos nocivos de las distintas unidades de medida.
Por otro lado, en los casos en que las variables se encuentren en la misma unidad de medida, debe asegurarse que la información que proveen las diferencias en las varianzas al análisis estadístico sea un elemento útil al propósito del análisis, para usar el enfoque de la Matriz de Varianzas y Covarianzas. Esto es así, ya que al usar la Matriz de Correlación no solo se anulan los efectos de unidad de medida si no toda otra información referente a la diferencia de varianza que, en algunos casos, podría aportar riqueza al estudio bajo análisis y su no inclusión podría distorsionar el significado y en análisis de las componentes.
Adicionalmente merece la pena destacar dos conclusiones importantes:
En primer lugar, si las varianzas de las variables son mayores que uno, utilizar la Matriz de Correlaciones, que indirectamente estandariza las varianzas de todas las variables a uno, tiende a incrementar la proporción de la varianza total explicada por la Primer componente
En segundo lugar, si las varianzas son menores que uno (hecho que puede suceder cuando las variables están expresadas en porcentajes tal y como sucede en el ejemplo que ilustraremos al final de trabajo), utilizar la Matriz de Covarianzas reduce la proporción explicada por la primer componente dificultando así la validez del método, el proceso de selección del número de componentes y fundamentalmente su interpretación.

5.- Caso de Aplicación: Relación entre crecimiento y otras variables

En esta sección aplicaremos todos conocimientos y conclusiones expuestos en secciones anteriores. Para ello utilizaremos datos correspondientes a distintas variables macroeconómicas de diversos países para el año 2000, obtenidos en la página web del Banco Mundial, en el enlace de datos estadísticos. Las variables fueron codificadas de la siguiente manera:

Código	*Variable*
expbser	Exportaciones de bs y servicios (proporción del PBI)
Tcrec	Tasa de crecimiento del PBI (proporción anual)
expaltec	Exportación de alta tecnología (proporción de exportaciones de manufacturas)
impbser	Importaciones de bs y servicios (proporción PBI)
partVAind	Participación del valor agregado industrial (proporción de PBI)
Crecpob	Crecimiento poblacional (tasa anual)
Inv	Inversión (proporción de PBI)

paises	Expbser	TasaPBI	Exaltec	Impbser	Partind	Crecpob	invers
Argentina	0,1089	-0,0079	0,0904	0,1152	0,2806	0,0087	0,1619
Australia	0,2294	0,0175	0,1522	0,2281	0,2583	0,0113	0,2145
Belgica	0,855	0,0372	0,0985	0,8229	0,2784	0,0025	0,2155
Brasil	0,1066	0,044	0,1861	0,1218	0,2788	0,0123	0,2154
Canada	0,46	0,0456	0,186	0,4078	0,2896	0,0088	0,2077
Chile	0,2975	0,044	0,0344	0,2876	0,3466	0,0128	0,2249
China	0,2587	0,08	0,1858	0,232	0,5022	0,0071	0,3633
Estonia	0,9365	0,073	0,2984	0,9774	0,2901	-0,0044	0,2776
Etiopia	0,1508	0,0595	0,1655	0,3004	0,0944	0,0239	0,1585
Finlandia	0,4296	0,0553	0,2733	0,3372	0,3449	0,0014	0,2059
Francia	0,2855	0,0379	0,2425	0,2731	0,2548	0,0046	0,2087
Alemania	0,3376	0,0286	0,1771	0,3338	0,3041	0,0015	0,2187
Guatemala	0,202	0,0361	0,079	0,2895	0,1979	0,0264	0,178
Guinea	0,236	0,019	0,001	0,2872	0,3589	0,0224	0,2196
Honduras	0,4186	0,0524	0,0185	0,559	0,3208	0,0262	0,3099
India	0,1389	0,0394	0,0501	0,1465	0,266	0,0168	0,2267
Italia	0,2829	0,0314	0,0917	0,2731	0,2904	0,0008	0,202
Japón	0,1076	0,028	0,2835	0,0934	0,321	0,0017	0,2615
México	0,31	0,0657	0,224	0,3294	0,2801	0,0142	0,2373
Perú	0,1611	0,0282	0,0361	0,1802	0,2757	0,0146	0,2026
España	0,3014	0,0418	0,0764	0,3241	0,3023	0,0074	0,2568
Suecia	0,4583	0,0436	0,2213	0,402	0,2951	0,0014	0,1846
ReinoUnido	0,279	0,0308	0,32	0,2984	0,2849	0,0025	0,1731
EEUU	0,1128	0,0378	0,3352	0,1502	0,2445	0,0127	0,2072
Turquia	0,2405	0,0736	0,0486	0,3153	0,2529	0,0169	0,2451
Suiza	0,4623	0,0316	0,1931	0,4142	0,2671	0,0056	0,2087

En virtud de las características de los datos, especialmente al hecho de que todos están expresados en porcentajes, las diferencias que observemos en las varianzas de cada una de las variables estarán limpias del efecto distorsivo que causan las diferentes unidades de medida en que podrían estar expresados los datos.

En segundo lugar, al estar las variables contenidas en el intervalo [0, 1], su matriz de varianzas y covarianzas también lo estará, por lo que, acorde a lo analizado en secciones anteriores, utilizar las matriz de correlaciones en vez de la de Varianzas y Covarianzas, incrementará las varianzas originales al ser estandarizas a uno. Como se vio, esto tendería a disminuir la proporción explicada por las primeras componentes.
A continuación llevaremos a cabo la aplicación de las Componentes Principales por ambos vías: Matriz de Varianzas-Covarianzas y Matriz de Correlación

Caso Uno: Análisis de Componentes Principales partiendo de la Matriz de Covarianzas

En esta subsección expondremos los principales resultados[3] y análisis gráficos como consecuencia de utilizar la Matriz de Covarianzas.

Variable	Varianza
Expbser	4,22E-04
TasaPBI	3,68E-02
Exaltec	9,97E-03
Impbser	4,00E-04
Partind	4,49E-03
Crecpob	7,17E-01
invers	1,99E-03

Tal como se anticipó líneas más arriba, al estar las variables originales expresadas en proporciones o tasas, su varianzas son todas menores que uno como lo indica la tabla anterior. Llevando a cabo el cómputo de los coeficientes que conforman las Componentes principales, se obtienen los siguientes resultados:

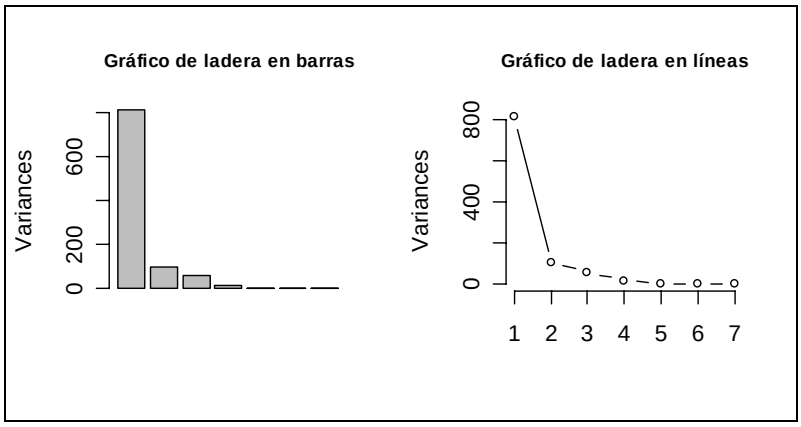
U	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Expbser	0.71597792	0.06930815	0.14140377	-0.52094059	0.42078294	0.07744158	0.0902352
TasaPBI	0.02454577	0.02229512	0.05099265	0.27001307	0.44967753	-0.84247385	-0.10686934
Exaltec	0.04575544	0.98528963	0.00861694	0.13133386	-0.0910377	0.01701609	0.03479294
Impbser	0.6949282	-0.13438165	-0.19382075	0.49175826	-0.4567485	-0.07148659	-0.07677355
Partind	0.0181776	-0.02746129	0.86464094	-0.12706781	-0.43272983	-0.21880207	-0.00598462
Crecpob	-0.01156496	-0.04738944	-0.03345618	0.08763351	-0.04160386	-0.1227999	0.9859061
invers	0.03591908	-0.05334184	0.43708206	0.61058433	0.46350705	0.46463595	0.0358492

Por otro lado, calculando las proporciones y las proporciones acumuladas de la varianza total explicada por cada una de las Componentes Principales, se construye la siguiente tabla donde se resumen medidas importantes de los valores propios:

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**
Standard Desviation	0.285	0.100	0.0764	0.0413	0.01803	0.01242	0.00447
Proportion of Variance	0.817	0.101	0.0589	0.0172	0.00328	0.00156	0.00020
Cumulative Proportion	0.817	0.919	0.9777	0.9949	0.99816	0.99972	0.99992

Según lo que se observa en la tabla anterior, se puede concluir que entre las dos primeras componentes se acumula casi el 92 % de la variación. Los gráficos de ladera, que a continuación se muestran, permiten corroborar la afirmación anterior.

[3] Para las salidas numéricas y gráficos se utilizaron sencillas rutinas escritas en R.



Si se analizan los coeficientes de los vectores propios asociados a las componentes principales que resultaron más relevantes se puede detectar la contribución de cada variable a la correspondiente componente. Para facilitar este análisis se construyen los siguientes gráficos referidos a la contribución de cada variable a cada componente y la correlación entre las variables originales y las componentes. Este último nos permitirá también interpretar el significado de los nuevos ejes de las componentes.

Contribución de cada Variable a cada Componente Principal

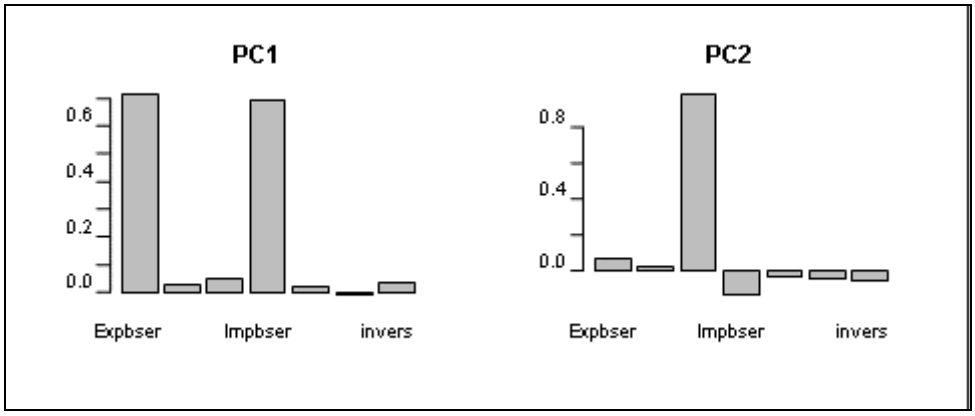
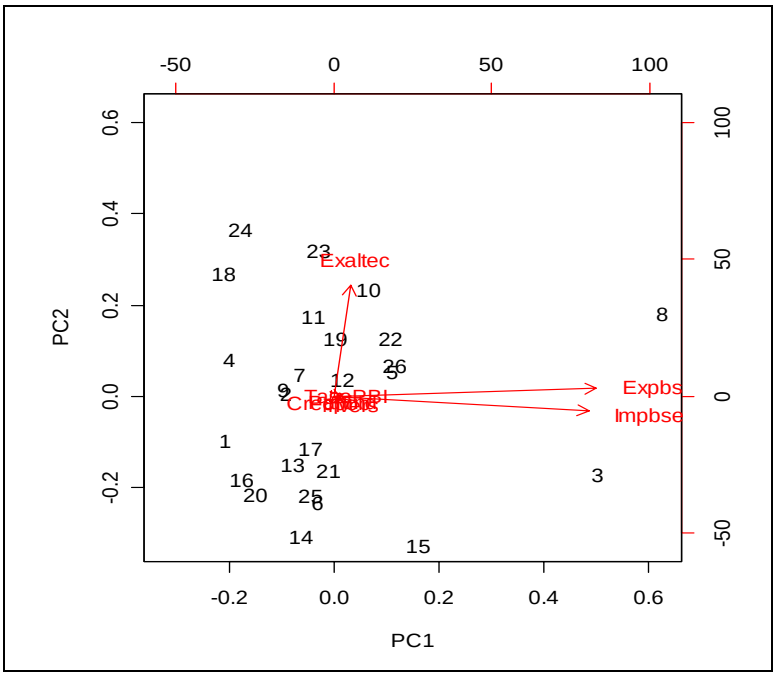


Tabla de los cuadrados de las correlaciones

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	Totales
Expbser	0.98374301	0.00114401	2.76E-03	0.01097499	0.00134869	2.18E-05	3.13E-06	1
TasaPBI	0.13270426	0.01358725	4.13E-02	0.33841262	0.17678643	2.97E-01	5.04E-04	1
Exaltec	0.01702439	0.97970229	4.35E-05	0.00295587	0.00026751	4.47E-06	1.97E-06	1
Impbser	0.97796183	0.00453839	5.48E-03	0.01032027	0.00167692	1.96E-05	2.39E-06	1
Partind	0.00595823	0.00168759	9.71E-01	0.00613566	0.01340273	1.64E-03	1.29E-07	1
Crecpob	0.15108199	0.31482379	9.11E-02	0.18281434	0.0077608	3.23E-02	2.20E-01	1
Invers	0.05260547	0.01439782	5.61E-01	0.32034448	0.03477028	1.67E-02	1.05E-05	1

Correlaciones entre las componentes principales y las variables originales



Es importante tener en cuenta para la interpretación del último gráfico, que la representación de las variables como vectores de dos coordenadas es tal que el ángulo que se forma entre ambos vectores equivale, aproximadamente, a la correlación entre las variables que los mismos representan. Por ejemplo, si los vectores de ambas variables forman un ángulo agudo, las variables estarán correlacionadas positivamente, y mientras menor sea ese ángulo mayor será la correlación. En tanto, si los vectores que representan ambas variables forman un ángulo obtuso cercano al ángulo de un giro, dichas variables estarán correlacionadas de manera negativa y por último, cuando forman un ángulo recto, se puede decir que las variables representadas por dichos vectores no están correlacionadas, es decir, son incorreladas.
De la información que se desprende de los gráficos anteriores se puede decir que las variables Exportaciones e Importaciones de Bienes y Servicios, fuerte y positivamente correlacionadas, están claramente representadas en la primera componente principal. En otras palabras, esto último nos dice que la Primer Componente Principal está referida a al Grado de Apertura de un país.
Por otro lado, Exportación de Alta Tecnología es la variable que más contribuye a la segunda componente principal y por lo tanto es la interpretación directa de esta última.
El resto de las Componentes es difícil de interpretar a la vez que agregan muy poco valor de varianza total explicada, donde las 2 primeras explican casi el 92% de la Variabilidad Total.

Se puede concluir entonces, que de las 7 variables originales, la dimensión se puede reducir a 2. La primera dimensión está marcada por el grado de Apertura de las economías y la segunda por las Exportaciones de Alta Tecnología. El resto de las variables no añaden demasiada información a la ya proporcionada por las anteriormente mencionadas.

Por último re-expresamos las coordenadas de cada país en los nuevos ejes

Reexpresión de las coordenadas de cada país en los nuevos ejes (Vectores Propios)

Y	PC1	PC2	PC3	PC4	PC5	PC6	PC7
1	-30.0594072	-4.8496295	-2.36659405	-5.5394109	-3.0742373	1.3121827	-0.27218121
2	-13.0956275	0.3824237	-2.30593719	-1.2494984	0.8112922	2.30353173	0.35019992
3	72.8832353	-8.5405916	-3.12291513	-5.0345627	0.5559844	0.85994017	0.15541097
4	-29.0154257	4.2931752	-0.00921328	0.8842451	0.5328009	-0.48142661	-0.01749259
5	16.1616723	2.9209256	0.0617638	-4.0576752	1.6034344	-0.47521952	0.57911437
6	-4.3628185	-11.7978555	5.62173838	-3.1765764	-0.1219577	-1.49466237	-0.05284075
7	-9.436956	2.5394014	25.9868678	5.4942996	0.7308301	-1.07695523	0.00700893
8	90.7099262	9.3931456	-0.86120446	5.4920394	-0.8806055	0.42483918	-0.44705843
9	-14.0485857	0.9592514	-21.2543444	6.4830232	0.4135937	-1.50828929	-0.17712397
10	9.6047754	12.1749767	5.85250677	-5.4148109	0.7446864	-2.07733378	0.27778143
11	-5.5080162	9.181473	-2.73673575	-0.5906712	1.1222839	0.74066575	-0.07214977
12	2.2474946	2.088308	1.4298413	-1.4728609	-0.9376678	0.806438	-0.49548307
13	-11.338183	-7.5144291	-10.7199722	1.4093184	-0.7842271	-0.59892959	0.57212033
14	-9.0158485	-15.6163305	5.40333279	-1.5019248	-4.3294869	-0.55198923	0.46635735
15	23.3588097	-16.5995917	3.54256838	9.514373	-1.8994478	1.11763856	0.4524551
16	-25.6077923	-9.2715279	-0.79975631	-0.6020544	2.8539575	0.49810233	0.15213994
17	-6.3550761	-5.7532217	-0.13859927	-3.5060033	0.2581799	-0.03204112	-0.97079198
18	-30.256484	13.9312389	6.25691536	2.4684045	0.6206015	2.59182922	-0.18560873
19	0.2797807	6.5663754	0.04978921	2.9182099	1.1911152	-1.18822727	0.38908517
20	-21.8343506	-10.862564	-1.41549702	-2.2016426	0.3450201	0.04386636	-0.14438223
21	-1.3200263	-8.1509235	2.57641162	1.3707261	1.3117749	1.04921939	-0.51974933
22	15.7290879	6.603882	-0.33898899	-5.3899151	0.107601	-1.32113907	-0.06317771
23	-3.9487851	16.5337925	-2.23485528	-0.7563406	-4.2754771	-1.04741148	-0.33219431
24	-26.0231935	18.7673847	-3.70059469	3.6873038	-1.0371282	0.5041865	0.4360053
25	-6.4838226	-10.9699121	-2.79046915	4.6006236	2.3900633	-1.66526498	-0.51405857
26	16.7356166	3.590823	-1.98605821	-3.8286192	1.7470162	1.26644965	0.42661382

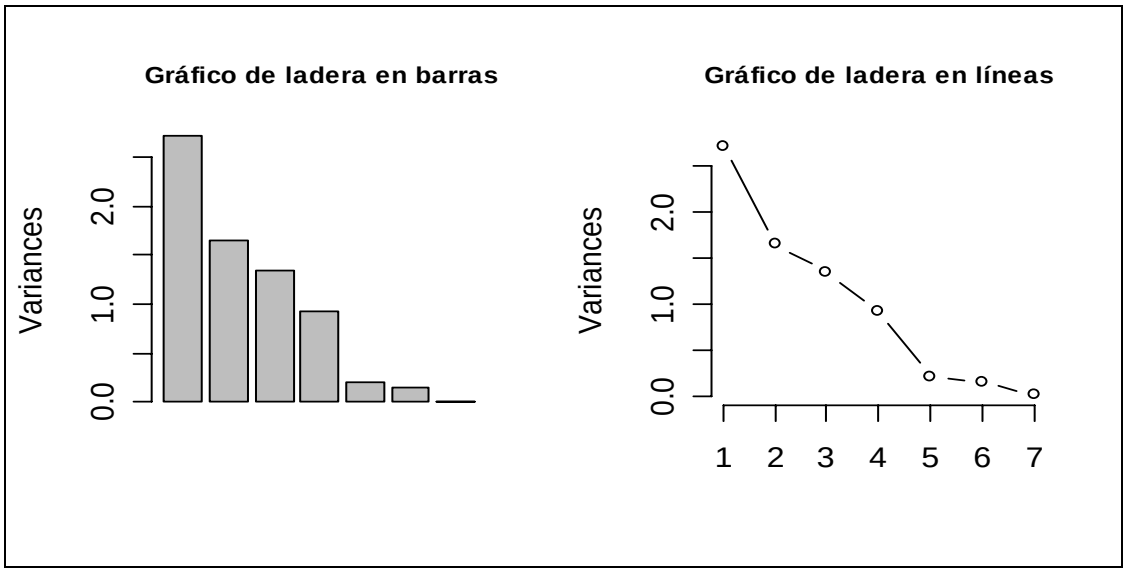
Caso Dos: Análisis de componentes principales partiendo de la matriz de correlaciones

En esta subsección expondremos los principales resultados y análisis gráficos como consecuencia de utilizar la Matriz de Covarianzas.
Llevando a cabo el cómputo de las Componentes principales, se obtienen los siguientes resultados:

U	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Expbser	0.5174919	0.2421862	-0.2887648	-0.24244788	0.03443993	0.16569417	-0.70904245
TasaPBI	0.3637413	-0.2638644	-0.1860476	0.67076054	-0.54331727	0.13282811	0.02640947
Exaltec	0.206716	0.399972	0.474827	0.53045599	0.49043554	0.22314078	-0.01130541
Impbser	0.4826025	0.19773	-0.4440337	-0.15379344	0.23073449	0.03593031	0.67279261
Partind	0.2774793	-0.4571682	0.4455175	-0.36874038	-0.04689921	0.6005723	0.12907566
Crecpob	-0.343503	-0.366206	-0.4997419	0.21545457	0.48840905	0.44642482	-0.11788841
Invers	0.3595395	-0.5740532	0.1055777	0.06223508	0.41123215	-0.58637628	-0.11500189

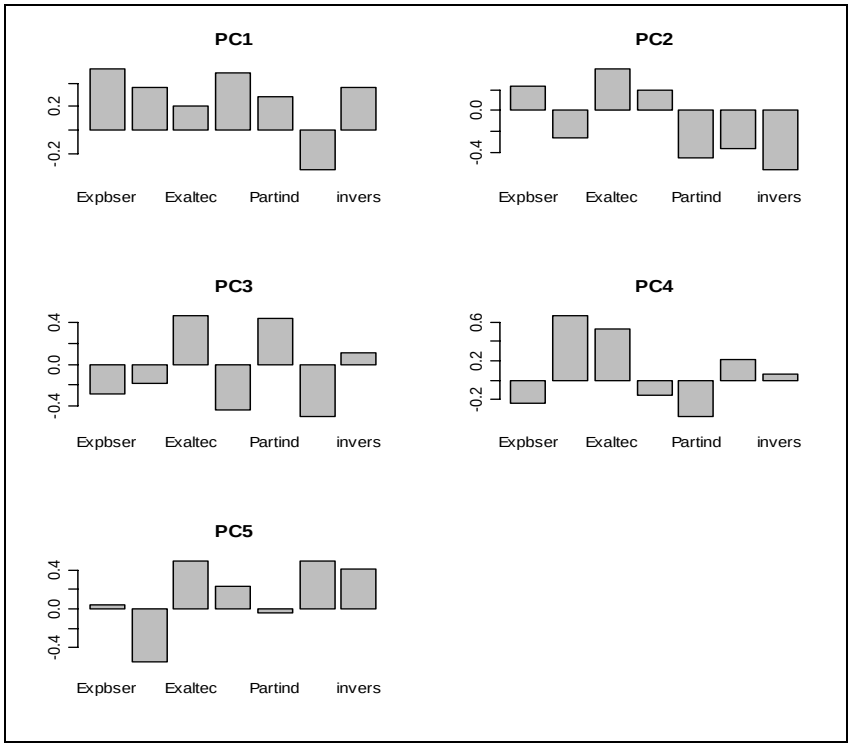
Por otro lado, calculando las proporciones y las proporciones acumuladas de la varianza total explicada por cada una de las Componentes Principales, se construye la siguiente tabla donde se resumen medidas importantes de los valores propios:

Componente	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.649	1.284	1.159	0.96	0.458	0.3849	0.09763
Proportion of Variance	0.388	0.236	0.192	0.132	0.03	0.0212	0.00136
Cumulative Proportion	0.388	0.624	0.816	0.947	0.977	0.9986	1

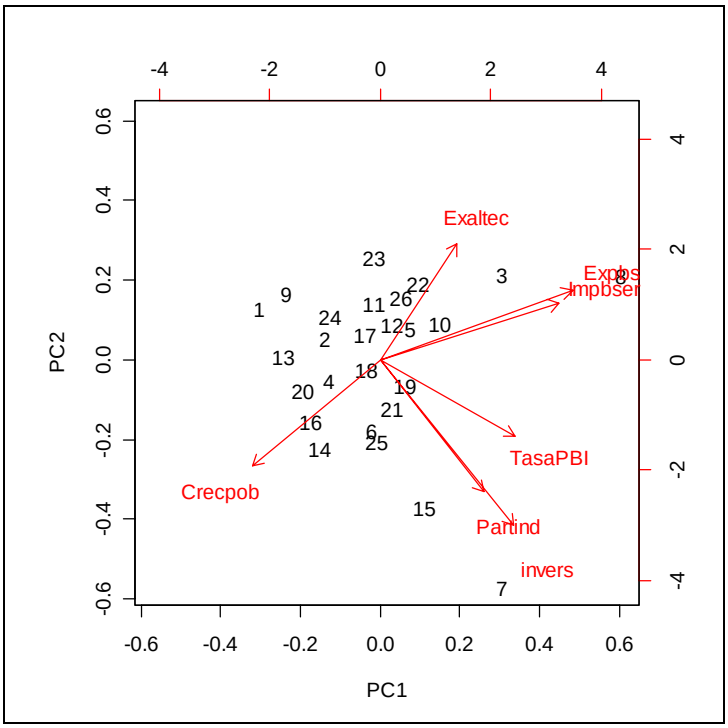


Según lo que se observa en la tabla anterior, la primer componente principal solo representa el 38% de la variabilidad total en comparación al 81% que representaba en el caso de análisis de Matriz de Covarianzas. Por otro lado, las dos primeras componentes representan solo el 62% contra el 92% analizado en el caso anterior. Así, al estar todas las variables en la misma unidad de medida, el haber utilizado la Matriz de Correlaciones nos condujo a perder información importante contenida en las diferencias de varianzas de las variables originales que ahora fueron estandarizadas a uno. A su vez, al ser las varianzas menores que uno, se redujo la proporción explicada por las primeras componentes tal y como lo anticiparon los análisis algebraicos de secciones anteriores.

Continuando con el análisis, si se analizan los coeficientes de los vectores propios asociados a las componentes principales que resultaron más relevantes se puede detectar la contribución de cada variable a la correspondiente componente. Para facilitar este análisis se construyen los siguientes gráficos referidos a la contribución de cada variable a cada componente y la correlación entre las variables originales y las componentes. Este último nos permitirá también interpretar el significado de los nuevos ejes de las componentes.



Contribución de cada variable a cada componente considerada.



De la información que se desprende de los gráficos anteriores podemos apreciar que la escasa proporción que representan de la varianza total, no solo se hace difícil interpretar los significados de cada componente principal si no que pone en duda la utilidad del método cuando se lo analiza por medio la matriz de Correlaciones.
En los gráficos se observa la escasa correlación entre las componentes principales y las variables originales corroborando lo dicho anteriormente.
De esta manera, verificamos las recomendaciones que se desprendieron de los análisis algebraicos de las secciones anteriores que, ante casos como los anteriores, sugieren la utilización de la Matriz de Varianzas y Covarianzas.

Por último re-expresamos las coordenadas de cada país en los nuevos ejes

Y	PC1	PC2	PC3	PC4	PC5	PC6	PC7
1	-2.56314696	0.84653967	0.80743339	-1.74587581	0.16120626	-0.02699409	0.09667466
2	-1.15122529	0.36070555	0.25804976	-0.49600266	0.54649595	-0.35010906	-0.12622666
3	2.56967936	1.43029795	-1.73266282	-1.62084219	0.00329532	-0.01936173	-0.09253189
4	-1.09263534	-0.31243399	0.65013998	0.75136371	-0.12914693	0.01558892	-0.00415251
5	0.64197378	0.54117046	-0.23724408	0.01077258	-0.06625654	0.37632352	-0.17005046
6	-0.18906638	-1.13768013	-0.26411803	-0.75437266	-0.58215987	0.36900359	0.011049
7	2.56665841	-3.72377352	1.98291363	0.58988769	-0.08960044	0.18862023	0.01226082
8	5.06955689	1.40641807	-0.95462007	0.32515906	0.32994882	-0.30681552	0.13090512
9	-1.99176604	1.1090923	-2.10137471	2.22439952	-0.18294461	-0.14758206	0.09796919
10	1.27430234	0.61790379	1.08319824	0.40908266	-0.48065159	0.73049985	-0.08502019
11	-0.11709155	0.94070292	0.66928726	0.43714335	0.03619429	-0.26188465	-0.04911118
12	0.25919241	0.59706717	0.77469382	-0.67979375	-0.06393796	-0.27281839	0.08214745
13	-2.02771245	0.06898235	-1.74708541	0.41596818	0.30254316	0.33885853	-0.0235909
14	-1.26931369	-1.43116757	-0.59041469	-1.56407622	0.44719409	0.7574512	0.0950027
15	0.93228856	-2.40387427	-1.9546014	-0.29521344	1.00993577	-0.10483857	0.06390827
16	-1.46490525	-0.98435024	-0.37614895	0.01139009	-0.26041561	-0.31589903	-0.13988727
17	-0.31173704	0.43064934	0.46363921	-0.89023112	-0.82681014	-0.43930239	0.10670611
18	-0.28401114	-0.15207236	2.3452859	0.29266067	0.55446084	-0.66877562	-0.02631086
19	0.51534213	-0.38718818	-0.17829558	1.38376543	0.0265923	0.27560692	-0.06260451
20	-1.63406803	-0.48760464	-0.30291178	-0.64972099	-0.32531172	-0.11286823	-0.00551479
21	0.24696465	-0.77364607	-0.02975851	-0.48995201	-0.25688854	-0.64371393	0.02014009
22	0.79544896	1.28226559	0.38377923	-0.11586451	-0.48679775	0.40204123	-0.01726847
23	-0.12393449	1.71016463	1.29938702	0.32021157	0.17560402	0.48870068	0.21887103
24	-1.04165971	0.72782455	1.07674539	1.48507286	0.78419047	0.13768651	-0.00590689
25	-0.06626725	-1.32615354	-1.28184185	1.0500183	-0.84012785	-0.32428294	0.05044589
26	0.45713312	1.05016016	-0.04347495	-0.40495033	0.21338826	-0.08513497	-0.17790375

6.- Conclusiones

A lo largo de éste artículo hemos analizado en detalle las propiedades algebraicas que permiten comparar el Método de las Componentes Principales cuando se trabaja con la matriz de Covarianzas de cuando se lo hace con la Matriz de Correlaciones. A través de desarrollos matemáticos rigurosos y su previa corroboración por medio de ejemplos y aplicaciones numéricas para variables referidas a datos macroeconómicos de diversos países del mundo se arribaron a una serie de recomendaciones que resumimos a continuación.

En los casos en que las variables estén expresadas en diferentes unidades de medida trabajar con la Matriz de Correlaciones, para aislar así los efectos nocivos de las distintas unidades de medida. Por otro lado, en los casos en que las variables se encuentren en la misma unidad de medida, debe asegurarse que la información que proveen las diferencias en las varianzas al análisis estadístico sea un elemento útil al propósito del análisis, para usar el enfoque de la Matriz de Varianzas y Covarianzas. Esto es así, ya que al usar la Matriz de Correlación no solo se anulan los efectos de unidad de medida si no toda otra información referente a la diferencia de varianza que, en algunos casos, podría aportar riqueza al estudio bajo análisis y su no inclusión podría distorsionar el significado y en análisis de las componentes.

Adicionalmente merece la pena destacar dos conclusiones importantes:

En primer lugar, si las varianzas de las variables son mayores que uno, utilizar la Matriz de Correlaciones, que indirectamente estandariza las varianzas de todas las variables a uno, tiende a incrementar la proporción de la varianza total explicada por la Primer componente

En segundo lugar, si las varianzas son menores que uno (hecho que puede suceder cuando las variables están expresadas en porcentajes tal y como sucede en el ejemplo que ilustraremos al final de trabajo), utilizar la Matriz de Covarianzas reduce la proporción explicada por la primer componente dificultando así la validez del método, el proceso de selección del número de componentes y fundamentalmente su interpretación.

BIBLIOGRAFÍA

Peña, Daniel. (2002)*: Análisis de datos multivariantes.* McGraw-Hill, Madrid.