Quant Trading

www.quant-trading.co

Análisis de componentes principales (ACP)

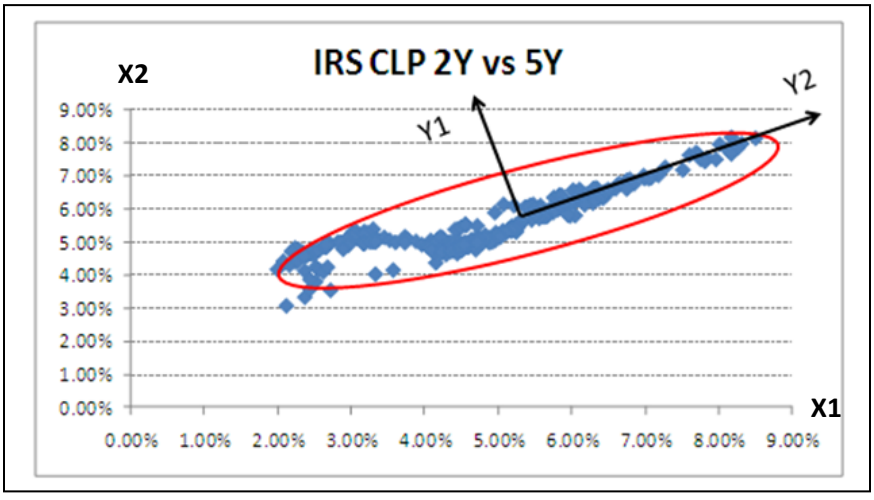
El Análisis de componentes principales (ACP) es una técnica de estadística multivariada utilizada generalmente para reducir la dimensionalidad de un conjunto de datos. Esta técnica también puede verse como un método para transformar variables correlacionadas en variables no correlacionadas.

El ACP hace uso de la matriz de varianzas y covarianzas del conjunto de datos. Si este conjunto está compuesto por p variables pues entonces necesitaríamos p componentes para explicar la varianza total del sistema. Sin embargo es posible utilizar k componentes (con k<p) para explicar una gran porción de la varianza total. Este último hecho es el que nos ayuda a reducir la dimensionalidad del sistema.

En los mercados financieros esta técnica se utiliza con mucha frecuencia en el análisis de la curva de tasas de interés. La curva en general está compuesta por diferentes nodos según su plazo al vencimiento. Estos nodos normalmente se mueven en conjunto y tienen una alta correlación. Si utilizáramos los nodos de cada mes hasta 1 año y luego los de cada año hasta 30 años tendríamos un total de 42 nodos. En claro que modelar un número tan alto de variables es difícil. Sin embargo vamos a ver como usando el ACP podemos reducir este número a tan sólo 2 o 3 variables.

Una manera sencilla de entender la metodología de cálculo del análisis de componentes principales es a través de un ejemplo en 2 dimensiones. Lógicamente no tiene mucho sentido aplicar este método en 2 dimensiones debido a que no vamos a obtener ningún beneficio para reducir la dimensionalidad del sistema. Sin embargo debido a su fácil representación gráfica el ejemplo en 2 dimensiones puede ser bastante ilustrativo. En la Figura 1 podemos observar un diagrama de dispersión entre las tasas de los Interest Rate Swaps (IRS) de Chile a 2 y 5 años. Estas variables las podemos llamar X1 y X2. En nuestro curso de **DERIVADOS DE TASAS DE INTERÉS INTERMEDIO** explicamos con detalle el funcionamiento de los IRS.

Figura 1



Fuente: Bloomberg

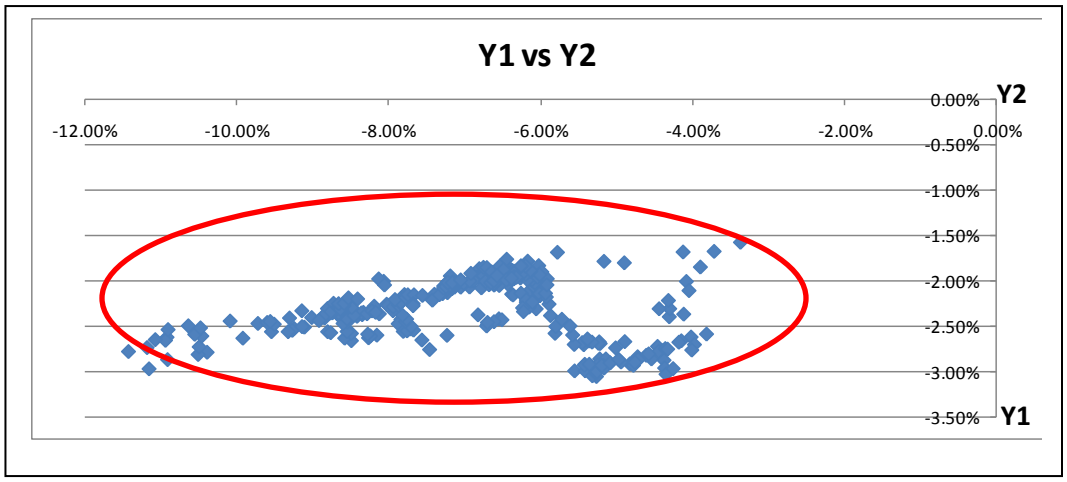
Como es de esperarse ambas tasas (X1 y X2) tienen una correlación alta. La elipse que está dibujada en rojo abarca casi toda la nube de puntos correspondiente al conjunto de datos. Y1 y Y2 son los vectores propios asociados a estos datos y se pueden interpretar como unos ejes de coordenadas rotados. Allí podemos observar que estos dos vectores son perpendiculares. Esto es equivalente a decir que los vectores son ortogonales y por lo tanto no están correlacionados. El tamaño de cada vector propio se conoce como valor propio.

Y1 y Y2 son en este caso las componentes principales del conjunto de datos. Normalmente las componentes principales se ordenan de mayor a menor según las magnitudes de sus valores propios, es decir de su variabilidad. En este caso Y2 sería la primera componente principal ya que posee la mayor variabilidad. El análisis se puede extender a p dimensiones. Por ejemplo en el caso de p=3 no estaríamos hablando de una elipse sino de un elipsoide en 3 dimensiones (como un balón de fútbol americano).

En la Figura 2 podemos observar las variables Y1 y Y2 rotadas. Esto se logra básicamente a partir de la transformación lineal del tipo.

$$Y = A'X$$

Figura 2



Básicamente lo que se hace allí es premultiplicar X por una matriz A transpuesta para rotar el conjunto de datos. Los elementos de la matriz A son de gran importancia y pueden ser utilizados en diversas aplicaciones como lo ilustraremos a continuación. En el Anexo 1 se explica con detalle el método de cálculo tanto para p=2 como para un p general.

Vamos a ilustrar 2 de las aplicaciones del ACP usando una base de datos con información de la evolución de las tasas de los IRS chilenos. Allí tenemos un total de 12 nodos (los dos primeros son las tasas de 3 y 6 meses y luego tenemos tasas de algunos años siendo 20 años el vencimiento más largo). Contamos con un total de 333 observaciones lo cual nos da aproximadamente 7.5 años de observaciones semanales.

Al aplicar la función prcomp en R a esta base de datos obtenemos la matriz A que se ilustra en la Figura 3. En la Figura 4 podemos observar las desviaciones estándar y varianzas de las 12 componentes principales.

Figura 3

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12
CHSWPC.Curncy	(0.50)	(0.38)	0.61	(0.21)	(0.22)	0.33	0.17	0.03	(0.07)	0.03	(0.03)	0.02
CHSWPF.Curncy	(0.50)	(0.27)	0.04	0.22	0.16	(0.55)	(0.50)	(0.12)	0.15	(0.06)	0.05	(0.07)
CHSWP1.Curncy	(0.45)	(0.11)	(0.42)	0.36	0.35	0.09	0.57	0.07	(0.12)	0.03	(0.00)	0.07
CHSWP2.Curncy	(0.33)	0.10	(0.48)	(0.07)	(0.33)	0.45	(0.42)	0.28	0.26	0.07	(0.06)	0.10
CHSWP3.Curncy	(0.25)	0.21	(0.25)	(0.30)	(0.24)	(0.04)	0.01	(0.38)	(0.47)	(0.18)	0.05	(0.53)
CHSWP4.Curncy	(0.20)	0.26	(0.06)	(0.35)	(0.13)	(0.27)	0.14	(0.25)	(0.01)	(0.15)	0.04	0.75
CHSWP5.Curncy	(0.17)	0.29	0.04	(0.32)	0.00	(0.30)	0.29	0.12	0.46	0.54	(0.07)	(0.29)
CHSWP7.Curncy	(0.13)	0.32	0.16	(0.16)	0.26	(0.04)	0.04	0.48	0.11	(0.67)	(0.19)	(0.14)
CHSWP8.Curncy	(0.12)	0.33	0.17	(0.01)	0.24	0.06	(0.18)	0.30	(0.34)	0.25	0.69	0.07
CHSWP10.Curncy	(0.11)	0.33	0.17	0.10	0.28	0.11	(0.25)	(0.09)	(0.35)	0.32	(0.66)	0.12
CHSWP15.Curncy	(0.09)	0.34	0.18	0.21	0.22	0.39	(0.03)	(0.58)	0.44	(0.14)	0.19	(0.07)
CHSWP20.Curncy	(0.08)	0.36	0.20	0.62	(0.61)	(0.18)	0.14	0.12	(0.05)	(0.04)	(0.02)	(0.01)

Figura 4

Standard deviations:

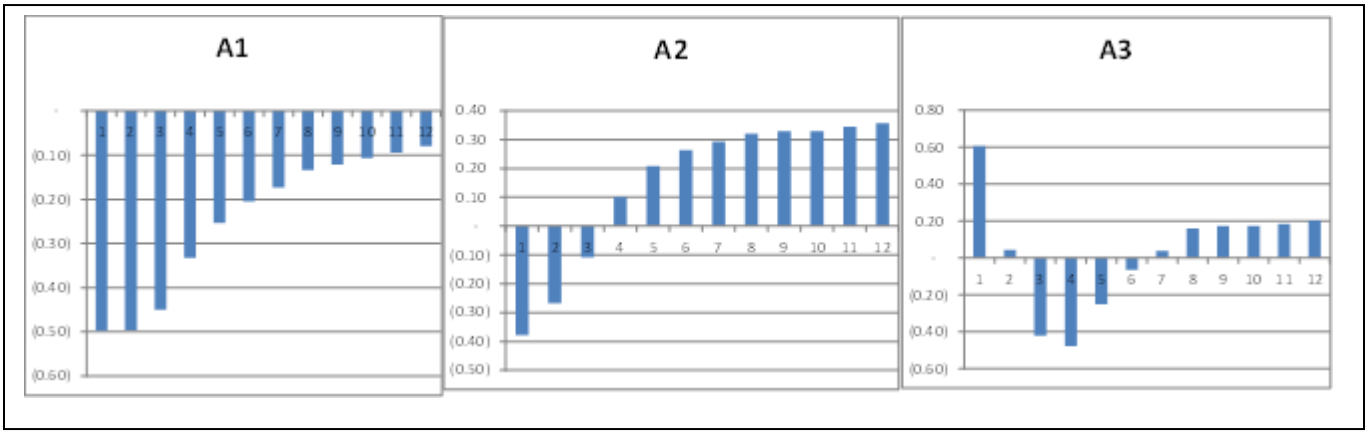
4.22	1.57	0.27	0.14	0.10	0.07	0.05	0.04	0.03	0.03	0.02	0.02

Variance

17.81	2.46	0.08	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00

La varianza de la primera componente representa el 87% de la varianza total, la de las dos primeras el 99.43% y el de las 3 primeras el 99.8%. Allí vemos que fácilmente podemos reducir el sistema de una dimensión 12 a quizás 1 o 2 dimensiones y aun así explicar un gran porcentaje de la varianza total. En general las 3 primeras componentes explican más del 90% de la varianza en los movimientos de las curvas de tasas de interés a nivel global. De allí que se haya querido interpretar cada componente como los movimientos que se dan en las curvas. En la Figura 5 podemos observar las gráficas de los 3 primeros vectores de la matriz A.

Figura 5



El panel izquierdo de la Figura 5 muestra los elementos del vector A1. Vemos que todos los valores son negativos. En el caso puntual de esta muestra tenemos que hubo una reducción en la tasa del Banco Central de Chile muy drástica (casi 800 puntos básicos en un período de tiempo muy corto). Por ello vemos que estos valores son negativos y más fuertes en los nodos cortos. Cuando se realiza este ejercicio en otras curvas generalmente se observa que los valores son positivos y aproximadamente de la misma magnitud. Es por ello que se dice que la primera componente principal está asociada con un movimiento paralelo de la curva de tasas de interés.

Vemos que los elementos de A2 son al principio negativos y luego positivos. Se dice que la segunda componente principal está asociada a los movimientos de pendiente (empinamiento o aplanamiento). Por último el vector A3 tiene unos valores positivos al principio, luego negativos en el medio y al final nuevamente positivos. La tercera componente está asociada con los movimientos de curvatura (butterflies).

Este primer análisis descriptivo permite interpretar un poco los datos pero hasta ahora carece de una utilidad práctica. La primera aplicación que vamos a ver del ACP es su uso como filtro. En ingeniería este método es bastante utilizado para eliminar ruido de un conjunto de datos y poder hacer lecturas y predicciones más fácilmente.

Recordemos que el ACP se basa en la transformación lineal del conjunto original de datos X de la forma

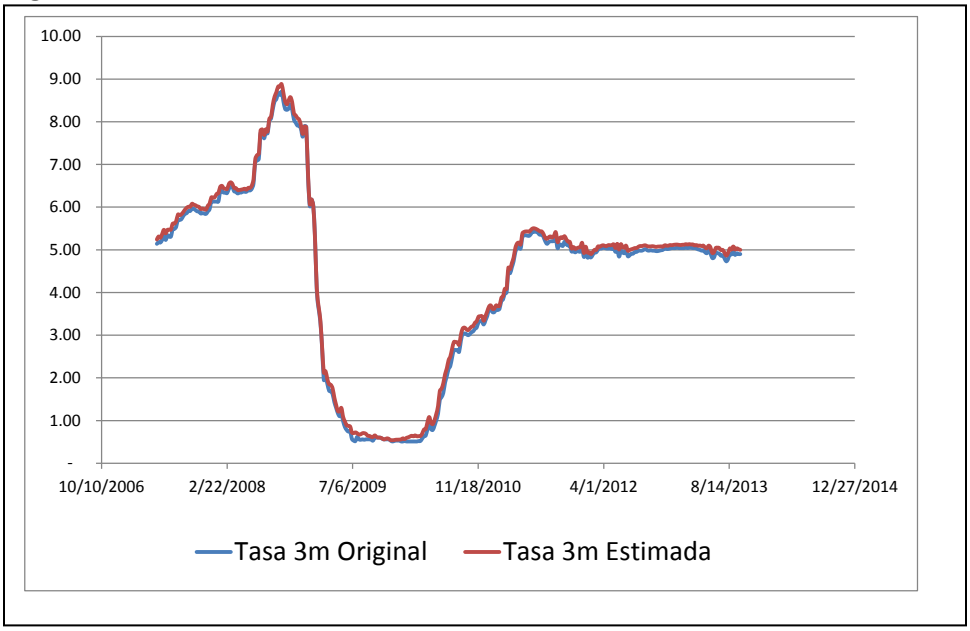
$$Y = A'X$$

Para aplicar la técnica de filtro debemos cambiar las filas de Y cuyas componentes no vamos a utilizar por valores nulos. A esta nueva matriz la llamaremos Y*. Esto se puede observar con detalle en el archivo de EXCEL adjunto “EJEMPLO PCA IRS CLP.xlsx” en la hoja que se llama PCA12dim. De esta forma podemos usar Y* para recobrar una aproximación del conjunto de datos original X. A esta aproximación la llamaremos X* y se calcula

$$X^* = Y^* A$$

En la Figura 6 podemos observar la evolución de la tasa de 3 meses original y la tasa de 3 meses estimada usando únicamente las 3 primeras componentes principales. Podemos observar que el ajuste es bastante bueno.

Figura 6



Una segunda aplicación del ACP es para la generación de escenarios y estimación de riesgos. Podemos por ejemplo analizar la exposición de un portafolio de renta fija ante los movimientos en las componentes principales. Como ya lo habíamos mencionado anteriormente las 3 primeras componentes principales se interpretan como movimientos de nivel, de pendiente y de curvatura.

Para poder utilizar el ACP en esta aplicación necesitamos construir una base de datos que contenga las diferencias entre las tasas de cada plazo en períodos consecutivos. De esta forma, en nuestro ejemplo construimos una base de datos con las variaciones semanales en las tasas de los IRS chilenos de diferentes plazos. Los elementos de los 3 primeros vectores obtenidos de la matriz A al igual que sus desviaciones estándar se pueden observar en la Figura 7

Por ejemplo la tasa de 3 meses que corresponde al primer renglón, se moverá - 0.24 puntos básicos cuando el factor 1 (o la primera componente principal) se mueva en una unidad. La tasa de 20 años (último renglón de la columna 1) por su parte se moverá -0.1 puntos básicos cuando el factor 1 cambie en 1 unidad. Las desviaciones estándar de los tres primeros factores (o componentes principales) son 44.5, 18.18 y 12.56 puntos básicos respectivamente. De esta manera un movimiento del primer factor equivalente a una desviación estándar causará que la tasa de 3 meses se mueva -10.49 puntos básicos (-0.24*44.5) y que la tasa de 20 años se mueva -4.52 puntos básicos (-0.1*44.5).

Figura 7

	A1	A2	A3
CHSWPC.Curncy	(0.24)	0.56	(0.07)
CHSWPF.Curncy	(0.30)	0.44	(0.09)
CHSWP1.Curncy	(0.35)	0.32	0.02
CHSWP2.Curncy	(0.36)	0.12	0.05
CHSWP3.Curncy	(0.34)	(0.04)	0.07
CHSWP4.Curncy	(0.32)	(0.15)	0.08
CHSWP5.Curncy	(0.31)	(0.22)	0.07
CHSWP7.Curncy	(0.27)	(0.27)	0.14
CHSWP8.Curncy	(0.27)	(0.26)	0.01
CHSWP10.Curncy	(0.26)	(0.26)	0.03
CHSWP15.Curncy	(0.25)	(0.26)	(0.00)
CHSWP20.Curncy	(0.10)	(0.15)	(0.97)
σ	44.50	18.18	12.56

Algunas personas utilizan esta información para calcular el Valor en Riesgo (VaR) de su portafolio de renta fija. Veamos un ejemplo sencillo de como aplicaría esto basado en un ejercicio del libro de Hull en la sección 16.9. Acá vamos a calcular la exposición a las 2 primeras componentes principales.

Supongamos que tenemos un portafolio que tiene exposición a las tasas de 1,2,3,4 y 5 años. Las exposiciones se expresan como la pérdida ante un movimiento hacia arriba en 1 punto básico en la tasa de interés de cada plazo y son las siguientes

Tasa 1 año	Tasa 2 años	Tasa 3 años	Tasa 4 años	Tasa 5 años
+10MM	+4MM	-8MM	-7MM	+2MM

La exposición al primer factor vendría dada por:

10*(-0.35) + 4*(-0.36) - 8*(-0.34) - 7*(-0.32) +2*(-0.31) = -0.63MM

La exposición al segundo factor vendría dada por:

10*(0.32) + 4*(0.12) - 8*(-0.04) - 7*(-0.15) +2*(-0.22) = 4.62MM

El cambio total en el valor del portafolio ΔP ante un cambio en los factores 1 y 2 vendría dado por:

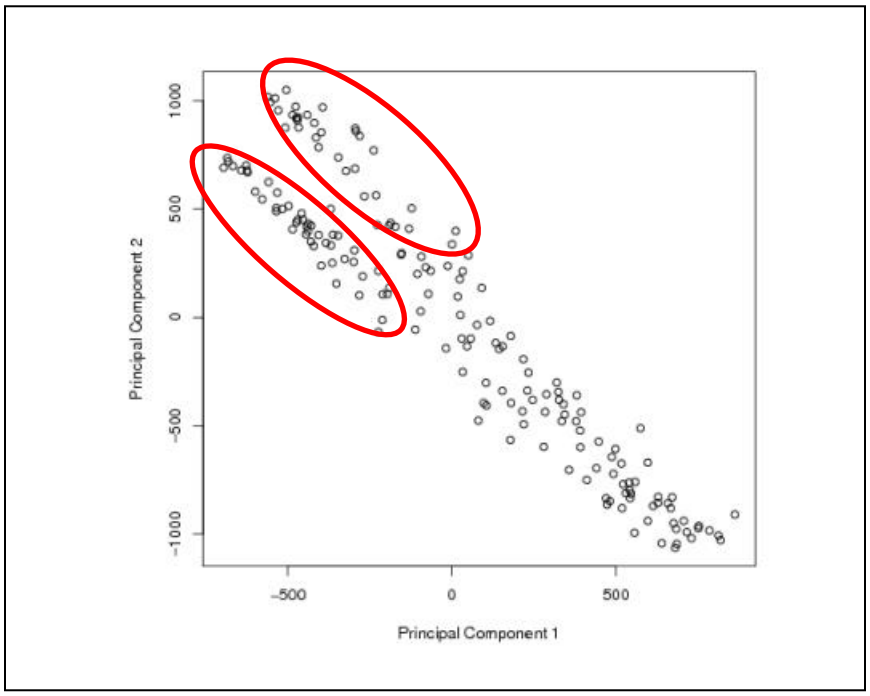
$\Delta P = -0.63f1 + 4.62f2$

Como los factores no están correlacionados podemos decir que 1 desviación estándar en la exposición del portafolio es $\sqrt{0.63^2 * 44.5^2 + 4.62^2 * 18.18^2} = 88.57$. EL VaR a 1 semana con un 99% de confianza sería entonces 88.57 * 2.33 = CLP 206.37 MM

Algunos problemas del ACP

Es importante anotar que las transformaciones lineales de variables aleatorias normales producen nuevamente una distribución normal. Dado que Λ es una matriz diagonal los componentes del vector Y son variables normales independientes. Aun cuando los datos no estén distribuidos normalmente se puede aplicar la descomposición de componentes principales. Los Y's resultantes van a estar no correlacionados pero no van a ser independientes. Estos problemas se pueden chequear fácilmente con una inspección visual realizando un gráfico de dispersión entre pares de componentes. Por ejemplo en la Figura 8 podemos observar una muestra que parece tener dos regímenes diferentes ya que muestran dos relaciones fuertes en diferentes momentos del tiempo

Figura 8



Referencias

- Jolliffe I. (2002). Principal Component Analysis. Springer
- HULL, J. (2006). Futures, Options and Other derivatives.

ANEXO 1

Para el caso de p=2 la forma de calcular las componentes principales es la siguiente:

1. Realizar una transformación lineal de nuestro conjunto de datos de la siguiente manera

$$Y = A' X$$

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \end{bmatrix}$$

X es nuestra matriz original de datos. En nuestro ejemplo X es un vector de 2 filas (la tasa de 2 años es una fila y la tasa de 5 años es otra fila).

$$Y = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1n} \\ Y_{21} & Y_{22} & \cdots & Y_{2n} \end{bmatrix} \begin{matrix} \rightarrow \text{Y1} \\ \rightarrow \text{Y2} \end{matrix}$$

En este caso A es una matriz ortogonal (recordemos que esta matriz es una rotación del sistema de coordenadas). Esta matriz satisface la ecuación A'A =I donde I es la matriz identidad

$$A = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix} \rightarrow \text{A1}$$

2. Calcular la matriz de varianzas y covarianzas de X. A esta la llamaremos Σ. En este caso particular Σ es una matriz 2x2. La matriz de varianzas y covarianzas de Y viene dada por $A'\Sigma A$. Además sabemos que la matriz de varianzas y covarianzas tiene una descomposición espectral de la forma $\Sigma = A\Lambda A'$. Esto hace que $A'\Sigma A = A' A\Lambda A'A = \Lambda$. La descomposición de componentes principales es un caso especial de la descomposición espectral en donde estamos buscando que Y1 y Y2 tengan la varianza más alta posible y además que sean linealmente independientes. Por lo tanto la matriz de varianzas y covarianzas de Y va a ser una matriz diagonal cuyos componentes corresponden a los vectores propios al cuadrado (varianzas). Recordemos que a Y1 y Y2 los denominamos componentes principales y a los valores a_{ij} los llamamos cargas.

3. EL siguiente paso es buscar A1 de forma tal que la varianza sea máxima. Esto es Max ($A1'\Sigma A1$). Lógicamente no existe un valor finito de A1 para que esto suceda por lo que tenemos que necesariamente imponer una restricción sobre A1. La restricción utilizada es $A1'A1 = 1$. Así entonces en SOLVER buscamos maximizar $Z = A1'\Sigma A1$ sujeto a $A1'A1 = 1$. Esto lo hacemos cambiando las celdas donde están los componentes de A1'.

4. Luego hacemos buscamos A2 de forma tal que la varianza sea máxima. Nuevamente tenemos que maximizar Z = A2'ΣA2 sujeto a A2'A2 = 1. En este caso tenemos una restricción adicional debido a que A1 y A2 deben ser ortogonales. Esta restricción la podemos escribir fácilmente haciendo A1'A2 = 0.

5. Lo último que debemos hacer es ordenar los vectores A1 y A2 dependiendo del tamaño de su varianza. El primer vector es aquel que tiene la mayor varianza.

El caso de p=2 está ilustrado en el archivo de EXCEL adjunto "EJEMPLO PCA IRS CLP.xlsx" en la hoja que se llama PCA2dim.

Para el caso general de p realizamos un procedimiento similar pero no es tan sencillo realizarlo en EXCEL como explicamos anteriormente. Sin embargo existen programas que ya tienen programado el algoritmo. Por ejemplo el programa R que es de uso gratuito tiene la función prcomp (X) que permite calcular las componentes principales para una base de datos X. Adjuntamos los archivos EjemploPCA.RData, y IRSCLP_BD.csv y PCA.txt para hacer el ejemplo en R. El archivo IRSCLP_BD.csv contiene la base de datos con las tasas de interés de los IRS de Chile de diferentes plazos con un total de 333 observaciones. El archivo PCA.txt contiene el código para correr el archivo en R. Los archivos deben ponerse en la carpeta mis documentos.

Para el caso de p general simplemente cambiamos las matrices X, A y Y por:

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pn} \end{bmatrix} \begin{matrix} \rightarrow X_1 \\ \\ \\ \rightarrow X_p \end{matrix} \qquad Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{p1} & y_{p2} & \cdots & y_{pn} \end{bmatrix} \begin{matrix} \rightarrow Y_1 \\ \\ \\ \rightarrow Y_p \end{matrix}$$

$$A = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{p1} \\ a_{12} & a_{22} & \cdots & a_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1p} & a_{2p} & \cdots & a_{pp} \end{bmatrix} \rightarrow A_1$$