TÉCNICAS ESTADÍSTICAS MULTIVARIADAS PARA LA EXPLORACIÓN Y REDUCCIÓN DE DATOS

Centro de Investigación en Matemáticas A.C.

www.cimat.mx

AUTORES

M. en E. Sergio Martín Nava Muñoz
nava@cimat.mx

Dr. Jorge Raúl Pérez Gallardo
raul.perez@cimat.mx

Centro de Investigación en Matemáticas, A.C.
Unidad Aguascalientes
MÉXICO
2018

Unidad II Análisis de Componentes Principales

Antes de empezar:

A lo largo del material de estudio usted encontrará referencias visuales que le indicarán el tipo de texto que está leyendo.
A continuación, se las presentamos.

Importante:
Presentan definiciones en relación a los conceptos trabajados.

Ejemplos:
Con este ícono se destacan ejemplos o casos.

Integrando Ideas:
Son párrafos que sintetizan las ideas desarrolladas hasta ese momento.

Actividades optativas y sin entrega obligatoria:
Indican todas las actividades que usted podrá realizar a lo largo de cada unidad y que cuentan con la respuesta al final de la misma.

Es el ícono del software estadístico que se utilizará en este curso. Su presencia indica que es el momento de utilizarlo.

Objetivos:

Durante el trascurso de esta unidad se pretende que el participante logre:

- El participante será capaz de reconocer cuando el uso de componentes principales es adecuado.
- Determinará el número de componentes principales a extraer.
- Podrá evaluar si la reducción de dimensiones obtenida con componentes principales vale la pena.
- Podrá interpretar los componentes principales analizando los coeficientes obtenidos o en base a las salidas gráficas del análisis.
- Sabrá juzgar la utilidad de llevar a cabo un análisis de componentes principales en función del problema.

CONTENIDO

1. INTRODUCCIÓN 4

2. CONCEPTOS BÁSICOS 6

3. RAÍCES Y VECTORES CARACTERÍSTICOS 11

4. DEFINICIÓN DE LOS COMPONENTES PRINCIPALES 14

5. OBTENCIÓN ALGEBRAICA DE LOS COMPONENTES PRINCIPALES 15

6. PASOS PARA EFECTUAR UN ANÁLISIS DE COMPONENTES PRINCIPALES 18

6.1. Selección del número de Componentes Principales 19

6.1.1. Porcentaje de Variación Total Acumulada 19

6.1.2. Tamaño de la varianza 21

6.1.3. Método gráfico 22

6.2. Representación gráfica de los datos 23

6.2.1. Identificación de grupos 23

6.2.2. Identificación de outliers 23

7. CORRELACIONES ENTRE VARIABLES Y COMPONENTES PRINCIPALES .24

8. APLICACIONES DEL ANÁLISIS DE COMPONENTES PRINCIPALES 25

1. Introducción

En el desarrollo de una investigación es común que nos interese estudiar varias características de una población al mismo tiempo; esto es, aparte de que pueda interesarnos el comportamiento individual de cada característica o *variable*, también se quiere estudiar el comportamiento conjunto de todas las variables.

Como las poblaciones generalmente son grandes o infinitas, lo que se hace es tomar una muestra representativa y se miden sobre ella las variables de interés; si la población es pequeña, se podrán medir las variables de interés sobre cada miembro dicha población. En cualquier caso el resultado es una tabla o matriz de datos de n renglones por p columnas, donde el i-ésimo renglón son los valores de las p variables medidas sobre el i-ésimo individuo u observación, y la j-ésima columna son los valores de la j-ésima variable medida sobre los n individuos de la muestra; es por lo anterior que los renglones representan individuos y las columnas representan variables; por lo general se desea tener más individuos que variables ($p < n$). A continuación se presenta una matriz de datos con p variables medidas sobre n individuos.

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Por ejemplo, los individuos (renglones) pueden ser diferentes estudiantes, diferentes variedades de plantas, las mediciones de contaminantes en varias semanas consecutivas, distintas muestras de sedimento de una bahía o bien, distintas ciudades de un país. Por su parte, las variables (columnas) pueden ser: calificaciones de diferentes materias, diferentes características químicas de las plantas, diferentes contaminantes que afectan a una ciudad o bahía, distintos tipos de crímenes que se cometen en las ciudades o diferentes indicadores económicos.

Es claro que si observamos la matriz de datos a simple vista, es imposible darse cuenta de la estructura de los datos en conjunto, sobre todo si la matriz de datos es grande. El problema de entender cuál es la estructura inherente a los datos, cómo se relacionan las variables entre sí, cuáles variables son las más importantes, como se comportan los individuos con respecto a estas variables, etc., son algunas de las interrogantes que resolveremos por medio de la técnica en estudio.

Las técnicas estadísticas multivariadas se encargan del análisis de datos estadísticos que involucren a más de dos variables medidas sobre una muestra o población; el caso de dos variables es el más sencillo que se puede presentar y se puede resolver por metodologías propias.

Existen muchas técnicas estadísticas multivariadas las cuales permiten tener distintas perspectivas del comportamiento global de los datos, lo que puede llevar a conclusiones interesantes acerca de la población en estudio.

Los objetivos de estas técnicas se pueden clasificar en:

1) *Simplificar la estructura de los datos*; se trata de encontrar un conjunto mucho más pequeño de variables independientes que describan adecuadamente el comportamiento de los datos.
2) *Clasificar*; se intenta descubrir cómo se agrupan los individuos entre sí; el interés también puede ser observar la agrupación de las variables.
3) *Analizar la interdependencia*; consiste en estudiar la relación que se da entre las diferentes variables, sin tener necesariamente una conjetura previa de cómo es esa relación.
4) *Analizar la dependencia*; a diferencia del objetivo anterior aquí si interesa explicar el comportamiento de ciertas variables en función de otras.

De acuerdo al objetivo que se persiga en un estudio, existe una técnica multivariada más adecuada para resolver el problema; sin embargo, se pueden aplicar varias técnicas al mismo tiempo para poder llegar a conclusiones más consistentes acerca de las relaciones que ocurren en la matriz de datos.

El *Análisis de Componentes Principales*, ACP por sus siglas, es una de las técnicas multivariadas más difundidas y de mayor uso en la actualidad que permite establecer la estructura de un conjunto de datos multivariados obtenidos de una población cuya distribución de probabilidades no necesita ser conocida; el ACP es una técnica de *análisis de interdependencia*, ya que a todas las variables en estudio se les otorga igual valor a priori.

La técnica de componentes principales fue descrita por primera vez por Karl Pearson en 1901 aunque no propuso una forma práctica de implementar su procedimiento para más de 2 o 3 variables; en 1933 Harold Hotelling desarrolló este aspecto práctico de la técnica descubierta por Pearson; además, fue Hotelling quien le dio el nombre de *componentes* a los elementos encontrados en esta técnica, que hasta ese entonces se les llamaba *factores*.

El problema al que se enfrentaron los usuarios de esta técnica en ese entonces fue que las operaciones numéricas tenían que hacerlas a mano, y para un estudio de más de dos variables se volvía algo totalmente impráctico. Fue hasta que se puso de moda el uso de computadoras electrónicas cuando las técnicas multivariadas comenzaron a divulgarse velozmente en todo el mundo, esto fue a partir de los años 60's.

Desde sus orígenes, el ACP ha sido aplicado en situaciones muy variadas: en psicología, sociología, medicina, meteorología, geografía, ecología, agronomía, estudios de mercado, finanzas, etc.

El ACP permite:

- Estudiar la relación existente entre las variables medidas.

- Reducir la dimensión del problema generando nuevas variables que puedan expresar la información contenida en el conjunto original de datos, medida ésta en términos de variabilidad.
- Para aplicarse como paso previo a futuros análisis, en particular a aquellos que exijan como supuesto la independencia de las variables entre sí, como es el caso de Análisis de Regresión Lineal Múltiple, o puede emplearse como una técnica complementaria en el caso del Análisis Clúster.
- Eliminar, cuando sea posible, algunas de las variables originales si ellas aportan poca información, medida de nuevo en términos de variabilidad.

Las nuevas variables generadas se denominan *componentes principales* y poseen en algunos casos características deseables tales como independencia (si se asume multinormalidad) y en todos los casos no correlación; de aquí se infiere que si las variables originales no están correlacionadas, esta técnica no ofrece ninguna ventaja.

El ***Análisis de Componentes Principales*** se aplica cuando se dispone de un conjunto de datos multivariados y no se puede postular, sobre la base de conocimientos previos del universo en estudio, una estructura particular de las variables.

El ACP deberá ser aplicado cuando se desee conocer la relación entre los elementos de una población y se sospeche que en dicha relación influye de manera desconocida un conjunto de variables o propiedades de los elementos. Finalmente se recalca que el ACP es una técnica matemática que no requiere que el usuario especifique un modelo estadístico para explicar la estructura de error. En particular no se hace ningún supuesto acerca de la distribución probabilística de las variables originales, aunque también hay que reconocer que por lo general es más fácil interpretar los componentes principales cuando se ha hecho el supuesto de multinormalidad.

En estas notas se denotará a las matrices con letras mayúsculas negritas; en particular la matriz de datos se escribirá $\mathbf{X}$. A los vectores se les denotará con letras minúsculas negritas, o bien con letras griegas. A las variables originales las denotaremos por X_i donde i puede tomar el valor de 1, …, p; a los componentes principales (CP) los denotaremos por Z_i donde la i puede ser 1, …, p; las componentes numéricas tanto de matrices como de vectores se escriben con letras minúsculas estándar con los subíndices que sean necesarios.

2. Conceptos básicos

Uno de los conceptos fundamentales en el Análisis Multivariado es el de función de distribución multivariada; se denotará a la *variable aleatoria p-dimensional*, o *vector aleatorio*, por $\mathbf{x}$, donde

$$\mathbf{x} = [X_1, X_2, \ldots, X_p]^{\mathrm{T}}$$

y $X_1, X_2, \ldots, X_p$ son variables aleatorias univariadas; se define la *función de distribución* asociada al vector $\mathbf{x}$, en el caso discreto, por

$$\begin{aligned} F(\mathbf{x}^0) &= \Pr\,(\,\mathbf{x} = \mathbf{x}^0\,) \\ &= \Pr\,(X_1 = x_1{}^0,\, X_2 = x_2{}^0,\, \ldots,\, X_p = x_p{}^0) \end{aligned}$$

donde $\mathbf{x}^0 = [x_1{}^0, \ldots, x_p{}^0]^{\mathrm{T}}$; en el caso continuo, la función de distribución se define por

$$\begin{aligned} F(\mathbf{x}^0) &= \Pr\,(\,\mathbf{x} \leq \mathbf{x}^0\,) \\ &= \Pr\,(X_1 \leq x_1{}^0,\, X_2 \leq x_2{}^0,\, \ldots,\, X_p \leq x_p{}^0) \end{aligned}$$

La *función de densidad*, denotada por f, para el caso continuo, es

$$f\,(\mathbf{x}) = \frac{\partial^p F\,(\mathbf{x})}{\partial \mathbf{x}}$$

o equivalentemente, se puede escribir

$$F\,(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}} f\,(\mathbf{u})\,d\mathbf{u}$$

y para el caso discreto se reemplaza la integral por una sumatoria.

La media de un vector aleatorio x, es un vector $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_p]$, tal que

$$\mu_i = E\{X_i\} = \int_{-\infty}^{\infty} x f_i\,(x)\,dx, \quad i = 1, \ldots, p$$

donde $f_i(x)$ denota la función de densidad de la variable univariada X_i; esta definición es para el caso continuo, para el caso discreto $E\{X_i\} = \Sigma x_i P_i\,(x)$, donde $P_i(x)$ es la función de distribución de probabilidades de la variable univariada X_i.

Si se tiene 2 variables aleatorias, X_i y X_j, la covarianza entre ellas, denotada por σ_{ij} se define por

$$\begin{aligned} Cov\,(X_i, X_j) &= E\,\{(X_i - \mu_i)\,(X_j - \mu_j)\} \\ &= E\,\{X_i X_j\} - \mu_i \mu_j = \sigma_{ij} \end{aligned}$$

Si se tienen p variables aleatorias, entonces habrá ½ p $(p-1)$ covarianzas; es conveniente escribir estas cantidades en forma de matriz; la *matriz de varianzas y covarianzas*, o *matriz de dispersión*, o simplemente *matriz de covarianzas*, es

$$\sum = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{pp} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix};$$

nótese que los términos de la diagonal son las varianzas de las variables consideradas, y que esta matriz es simétrica y positiva semidefinida. Una forma de escribir la matriz anterior es

$$\begin{aligned} \Sigma &= E\left\{(\mathbf{x}-\boldsymbol{\mu})\,(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\right\} \\ &= E\left\{\mathbf{x}\,\mathbf{x}^{\mathrm{T}}\right\} - \boldsymbol{\mu}\,\boldsymbol{\mu}^{\mathrm{T}} \end{aligned}$$

La *matriz de correlaciones* está dada por

$$\boldsymbol{\rho} = \begin{pmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{pp} & \rho_{p2} & \cdots & \rho_{pp} \end{pmatrix}$$

donde

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}};$$

obsérvese que la matriz $\boldsymbol{\rho}$ es simétrica y semipositiva definida. El vector $\boldsymbol{\mu}$, la matriz de covarianzas Σ, y la matriz de correlaciones $\boldsymbol{\rho}$ son parámetros poblacionales del vector aleatorio $\mathbf{x}$; para hallar sus estimaciones consideremos que se tienen n observaciones del vector aleatorio $\mathbf{x}$, es decir se tiene

$$\mathbf{X} = [\mathbf{x}_1,\ \mathbf{x}_2,\ \ldots,\ \mathbf{x}_n]^{\mathrm{T}}$$

donde cada $\mathbf{x}_i$ es un vector aleatorio, es decir,

$$\mathbf{x}_i = [x_{i1},\ x_{i2},\ \ldots,\ x_{ip}]$$

y la estimación del vector de medias es

$$\widehat{\boldsymbol{\mu}} = [\widehat{\mu}_1, \ldots, \widehat{\mu}_p]^{\mathrm{T}}$$

donde

$$\widehat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad j = 1, \ldots p;$$

La estimación de la matriz Σ está dada por

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{pp} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$

donde

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}}$$

$\sqrt{s_{ii}}$ y $\sqrt{s_{jj}}$ y son las desviaciones estándar muestrales de las variables X_i y X_j respectivamente; r_{ij} es la estimación de la correlación poblacional entre las variables X_i y X_j.

Es claro de las expresiones anteriores que $s_{ij} = s_{ji}$, y, análogamente $r_{ij} = r_{ji}$; además $r_{ii} = 1$, por lo que la matriz **D** queda de la forma

$$\mathbf{D} = \begin{pmatrix} 1 & r_{21} & \cdots & r_{p1} \\ r_{21} & 1 & \cdots & r_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{pp} & r_{p2} & \cdots & 1 \end{pmatrix};$$

nótese que la matriz **D** es simétrica, semipositiva definida.

La función de distribución multivariada más usada es la distribución normal multivariada (DNM); recuerde que una variable aleatoria normal univariada X, con media μ y varianza σ^2 tiene como función de densidad

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

y se escribe $X \sim \mathrm{N}(\mu, \sigma^2)$. En el caso multivariado, un vector aleatorio p-dimensional $\mathbf{x}$ tiene una distribución normal multivariada si su función de densidad conjunta es

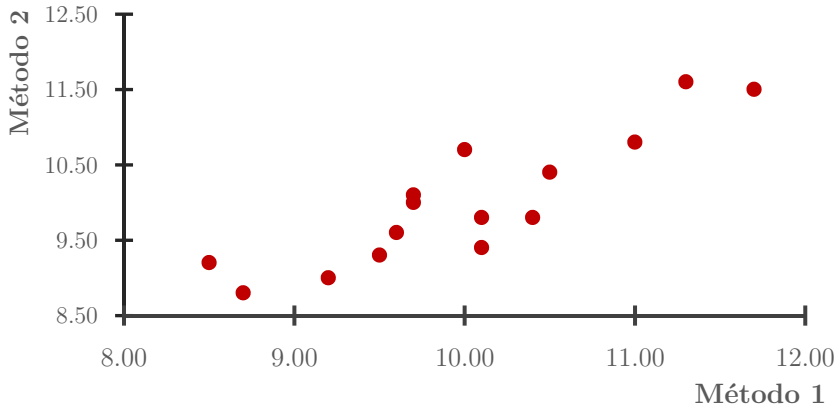
$$f(\mathrm{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\sum|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathrm{x}-\mu)' \sum^{-1} (\mathrm{x}-\mu)\right)$$

donde Σ es una matriz de tamaño $p \times p$ positiva definida; se puede comprobar que ésta es una función de densidad y que $\boldsymbol{\mu}$ es la media de $\mathbf{x}$ con esta función de densidad y que Σ es la matriz de covarianzas de $\mathbf{x}$; la notación para indicar que un vector aleatorio tiene una densidad normal p-variada con media $\boldsymbol{\mu}$ y matriz de covarianzas Σ es $\mathbf{x} \sim \mathrm{N_p}\,(\boldsymbol{\mu}, \Sigma)$.

Ejemplo 4.1. Control de calidad de químico. Se tiene un proceso en el que se lleva a cabo una prueba de control de calidad para la concentración de un componente químico en una solución, mediante dos métodos diferentes.

Observación	Método 1	Método 2
1	10.0	10.7
2	10.4	9.8
3	9.7	10.0
4	9.7	10.1
5	11.7	11.5
6	11.0	10.8
7	8.7	8.8
8	9.5	9.3
9	10.1	9.4
10	9.6	9.6
11	10.5	10.4
12	9.2	9.0
13	11.3	11.6
14	10.1	9.8
15	8.5	9.2

¿Qué se puede hacer con estos datos? las soluciones son inacabables. Una posibilidad sería calcular las diferencias en las concentraciones observadas y probar que la diferencia de medias es cero, usando la prueba t para diferencias apareadas. La técnica de análisis de varianza, trataría estos datos como una ANOVA de dos vías con métodos y corridas como factores.



La gráfica del ejemplo 4.1. sugiere el uso de regresión para determinar si es posible predecir el resultado de un método del otro. Sin embargo el requerimiento de que los dos métodos sean intercambiables significa que sean capaces de predecirse en cualquiera de las dos direcciones, lo que (usando mínimos cuadrados ordinarios) implicaría la utilización de dos ecuaciones. Las ecuaciones de mínimos cuadrados para predecir el método 1 del método 2, mientras que por otro lado una ecuación para predecir el método 2 del método 1.

Si se requiere una sola ecuación de predicción, uno podría invertir alguna de las ecuaciones de regresión, pero ¿cuál?, y ¿qué pasa con las consecuencias teóricas de hacer esto?

La línea que desarrolla este rol directamente se llama línea de regresión ortogonal que minimiza las desviaciones perpendiculares a la línea misma. La línea se obtiene por el método de componentes principales, de hecho es el primer componente principal.

Para ilustrar el método de componentes principales necesitamos obtener las medias, varianzas y covarianzas muestrales. Sea X_{1k} el resultado del método 1 para la k-ésima corrida y X_{2k} el resultado del método 2 corrida k.

$$\overline{X} = \begin{bmatrix} \overline{X}_1 \\ \overline{X}_2 \end{bmatrix} = \begin{bmatrix} 10.00 \\ 10.00 \end{bmatrix}$$

$$S = \begin{bmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{bmatrix} = \begin{bmatrix} .7986 & .6793 \\ .6793 & .7343 \end{bmatrix}$$

$$S_{ij} = \frac{n\sum X_{ik}X_{jk} - \sum X_{ik}\sum X_{jk}}{[n\,(n-1)]}$$

3. Raíces y vectores característicos

El método de componentes principales se basa en el resultado clave del álgebra de matrices: si **A** es una matriz $p \times p$ simétrica y no singular, tal como la matriz de covarianzas **S,** puede reducirse a una matriz diagonal **L** al pre-multiplicarla y pos-multiplicarla por una matriz ortogonal **U** tal que

$$\mathbf{U}^{\mathrm{T}}\mathbf{S}\mathbf{U} = \mathbf{L}$$

Los elementos de la diagonal de **L**, l_1, $l_{2,...}$, l_p se llaman las *raíces características*, *raíces latentes* o *eigenvalores* de **S**. Las columnas de **U**, u_1, u_2,..., u_p se llaman los *vectores característicos* o *eigenvectores* de **S**.

Las *raíces características* pueden obtenerse de la solución de la siguiente ecuación, llamada *ecuación característica*:

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

donde $\mathbf{I}$ es la matriz identidad. Esta ecuación produce un polinomio de grado p en l del cual se obtienen l_1, $l_{2,...}$, l_p

Ejemplo 4.1. Control de calidad de químico. *Continuación*

Para este ejemplo, hay $p = 2$ variables y

$$\begin{aligned} |\mathbf{S}-\lambda\mathbf{I}| &= \begin{bmatrix} .7986 - l & .6793 \\ .6793 & .7343 - l \end{bmatrix} \\ &= .124963 - 1.53291l + l^2 \end{aligned}$$

los valores del que satisfacen esta ecuación son $l_1 = 1.4465$ y $l_2 = 0.0864$. Los vectores característicos pueden obtenerse de la solución de la ecuación

$$|\mathbf{S} - \lambda\,\mathbf{I}|\ t_i = 0$$

$$u_i = \frac{t_i}{\sqrt{t_i' t_i}}$$

para $i = 1, 2, ..., p$. Para este ejemplo, para $i = 1$

$$[\mathbf{S}-l\mathbf{I}]\,t_i = \begin{bmatrix} .7986 - 1.4465 & .6793 \\ .6793 & .7343 - 1.4465 \end{bmatrix} \begin{bmatrix} t_{11} \\ t_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Estas son dos ecuaciones con dos incógnitas. Para resolverle, hacemos $t_{11} = 1$

$$\begin{gathered} -0.6478 + 0.6793\ t_{21} = 0 \\ \Rightarrow\ \ t_{21} = 0.9538 \\ u_1 = \frac{t_1}{\sqrt{t_1' t_1}} = \frac{1}{\sqrt{1.9097}} \begin{bmatrix} 1 \\ 0.9538 \end{bmatrix} = \begin{bmatrix} 0.7236 \\ 0.6902 \end{bmatrix} \end{gathered}$$

Similarmente, utilizando l2 = 0.0864 entonces,

$$u_2 = \begin{bmatrix} -0.6902 \\ 0.7236 \end{bmatrix}$$

Así

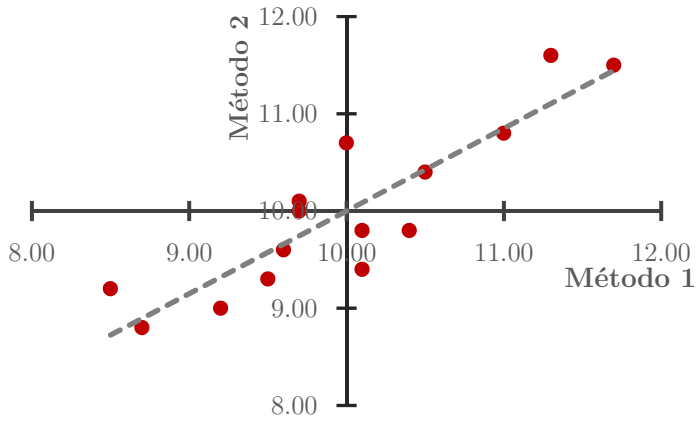
$$\mathbf{U} = \begin{bmatrix} u_1 & u_2 \end{bmatrix} = \begin{bmatrix} 0.7236 & -0.6902 \\ 0.6902 & 0.7236 \end{bmatrix}$$

que es ortogonal: $u_1'u_1 = 1, u_2'u_2 = 1, u_1'u_2 = 0$

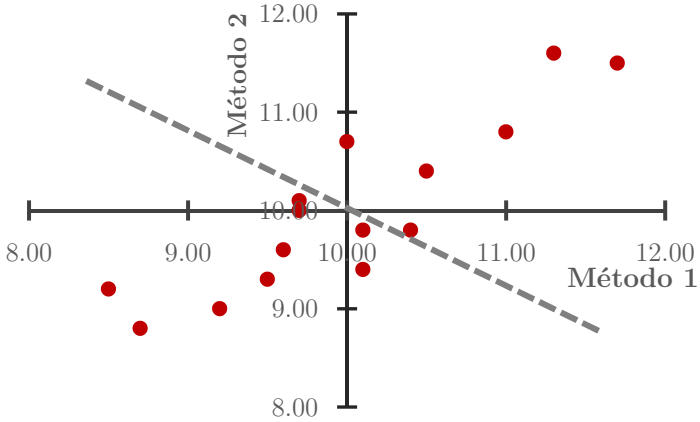
Mas aun

$$\mathbf{U'SU} = \begin{bmatrix} 0.7236 & 0.6902 \\ -0.6902 & 0.7236 \end{bmatrix} \begin{bmatrix} .7986 & .6793 \\ .6793 & .7343 \end{bmatrix} \begin{bmatrix} 0.7236 & -0.6902 \\ 0.6902 & 0.7236 \end{bmatrix}$$
$$= \begin{bmatrix} 1.4465 & 0 \\ 0 & .0864 \end{bmatrix} = \mathbf{L}.$$

Geométricamente, el procedimiento descrito no es más que una rotación del eje principal de las coordenadas originales x_1 y x_2 en la media como se muestra en la figura



Los elementos de los vectores característicos son los cosenos de las direcciones de los nuevos ejes con respecto a los viejos. Es decir, $u_{11} = 0.7236$ es el coseno del ángulo entre el eje x_1 y y el nuevo eje; $u_{21} = 0.6902$ es el coseno del ángu lo entre este nuevo eje y el eje x_2. El nuevo eje relacionado a $\mathbf{u}_1$ es la línea de regresión ortogonal que estábamos buscando. En la figura que se muestra a continuación se puede observar la misma relación pero con $\mathbf{u}_2$.



4. Definición de los Componentes Principales

Los ***componentes principales*** son nuevas variables formadas por combinaciones lineales normalizadas de las variables originales de mayor varianza posible y no correlacionadas entre sí.

Así, el *primer componente principal* es una combinación lineal de las variables X_1,..., X_p, es decir,

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$

de tal forma que la varianza de Z_1 sea lo más grande posible, sujeto a la normalización, es decir a la restricción de que

$${a_{11}}^2 + {a_{12}}^2 + \ldots + {a_{1p}}^2 = 1$$

esta condición se necesita para que no se pueda incrementar la varianza de Z_1 con solo aumentar cualquier coeficiente a_{1j} para $j = 1, \ldots, p$.

El *segundo componente principal* es de la forma

$$Z_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$$

donde la varianza de Z_2 es lo más grande posible sujeto a la restricción de que

$${a_{21}}^2 + {a_{22}}^2 + \ldots + {a_{2p}}^2 = 1$$

y además le pedimos que no este correlacionado con el primer componente principal.

El *tercer componente principal* es de la forma

$$Z_3 = a_{31}X_1 + a_{32}X_2 + \ldots + a_{3p}X_p$$

donde la varianza de Z_3 es lo más grande posible sujeto a la restricción de que

$${a_{31}}^2 + {a_{32}}^2 + \ldots + {a_{3p}}^2 = 1$$

y además no está correlacionado con ninguno de los dos componentes principales anteriores.

Análogamente podemos definir tantos componentes principales como variables tengamos, en este caso hasta p componentes principales.

Dadas las propiedades que tienen los CP, si las variables originales están muy correlacionadas, los primeros CP responderán por la mayoría de la variabilidad de los datos. Puede decirse que los CP son un reacomodo de la dependencia que existe entre las variables originales; es por esto que si las correlaciones entre las variables originales son pequeñas, no se gana nada haciendo ACP.

Los ***pesos*** o ***cargas*** del i-ésimo CP resultan ser los coeficientes del i-ésimo *eigenvector* de la *matriz covarianzas* o la *de correlaciones*, según con la que se esté trabajando, ordenados con respecto a sus eigenvalores respectivos de mayor a menor y el i-ésimo *eigenvalor* (ordenados como se indicó) representa la varianza del i-ésimo CP.

Nótese que algunos eigenvalores pueden ser cero, además no es posible obtener eigenvalores negativos ya que la matriz de correlaciones (o covarianzas) es simétrica y positiva definida.

5. Obtención algebraica de los Componentes Principales

Si bien en estas notas se desarrolla un enfoque algebraico para obtener los *Componentes Principales*, existen varias formas de lograr dicho objetivo.

Consideremos a la matriz de datos $\mathbf{X}$ de n individuos con p variables; se supondrá que las variables están correlacionadas entre sí. Si las variables se encuentran medidas en las mismas unidades y se puede suponer que tienen aproximadamente la misma varianza, entonces se toma $\mathbf{A} = \mathbf{S}$, es decir, $\mathbf{A}$ será la *matriz de covarianzas* de $\mathbf{X}$, de lo contrario se toma $\mathbf{A} = \mathbf{D}$, es decir, como la *matriz de correlaciones* de $\mathbf{X}$. En caso de que $\mathbf{A} = \mathbf{D}$ se supondrá que las variables X_i para $i = 1,\ldots, n$ han sido estandarizadas (tienen media cero y varianza uno) antes de efectuar el análisis.

Por definición, el primer CP es la combinación lineal normalizada

$$Z_1 = \sum_{i=1}^{p} a_{1i} X_i = \mathbf{a}_1' \mathbf{x}$$

donde

$$\mathbf{a'}_1 = [a_{11},\ a_{12},\ \ldots,\ a_{1p}]$$
$$\mathbf{x'} = [X_1,\ X_2,\ \ldots,\ X_p]$$

Tal que $Var\{Z_1\}$ es máxima; ahora,

$$Var\{Z_1\} = \mathbf{a'}_1\mathbf{A}\mathbf{a}_1$$

por lo que el problema consiste en maximizar $\mathbf{a'}_1\mathbf{A}\mathbf{a}_1$ bajo la restricción de que $\mathbf{a'}_1\mathbf{a}_1 = 1$; este problema se puede resolver utilizando *multiplicadores de Lagrange*, el lagrangiano es

$$\mathcal{L} = \mathbf{A}\mathbf{a}_1 - \lambda(\mathbf{a'}_1\mathbf{a}_1 - 1)$$

donde λ es el *multiplicador de Lagrange*, y para maximizar esta función se deriva con respecto a $\mathbf{a}_1$ y se iguala a cero, obteniéndose

$$\mathbf{A}\mathbf{a}_1 - \lambda\mathbf{a}_1 = 0$$

O equivalentemente

$$(\mathbf{A} - \lambda\mathbf{I}_p)\mathbf{a}_1 = 0$$

donde $\mathbf{I}_p$ es la *matriz identidad* de orden p; para que este sistema de ecuaciones tenga solución, se necesita que

$$\det(\mathbf{A} - \lambda\mathbf{I}_p) = 0$$

así que λ resulta ser un *eigenvalor* de la matriz $\mathbf{A}$ obteniendo la siguiente relación

$$\mathbf{A}\mathbf{a}_1 = \lambda\mathbf{a}_1$$

entonces $\mathbf{a}_1$ es el *eigenvector* correspondiente al *eigenvalor* λ; sin embargo $\mathbf{A}$ tiene p *eigenvalores*, ¿cual debemos tomar?, la respuesta se basa en lo siguiente, obsérvese que

$$\mathbf{a'}_1\mathbf{A}\mathbf{a}_1 = \mathbf{a'}_1(\lambda\mathbf{a}_1) = \lambda\ \mathbf{a'}_1\mathbf{a}_1 = \lambda$$

por lo que $Var\{Z_1\} = \lambda$; así, si se quiere maximizar $Var\{Z_1\}$, lo que se debe hacer es tomar λ como el mayor *eigenvalor* de la matriz $\mathbf{A}$, y a1 resulta ser el *eigenvector* correspondiente al *eigenvalor* λ.

Para obtener el segundo CP, recuerde que

$$Z_2 = \sum_{i=1}^{p} a_{2i}X_i = \mathbf{a}_2'\mathbf{x}$$

donde

$$\mathbf{a'}_2 = [a_{21},\ a_{22},\ \ldots,\ a_{2p}]$$

y queremos encontrar el vector $\mathbf{a}_2$ tal que maximice

$$Var\ \{Z_2\} = \mathbf{a'}_2\mathbf{A}\mathbf{a}_2$$

Sujeto a las restricciones

$$\mathbf{a'}_2\mathbf{a}_2 = 1 \text{ (normalidad)}$$
$$\mathbf{a'}_2\mathbf{a}_1 = 0 \text{ (ortogonalidad)}$$

la segunda condición se deriva de que se pide correlación cero entre los CPs, es decir, se quiere que

$$0 = \text{Cov(Z1, Z2)} = \text{Cov}(\mathbf{a'}_1\mathbf{x},\ \mathbf{a'}_2\mathbf{x}) = \mathbf{a'}_1\mathbf{A}\mathbf{a}_2 = \mathbf{a'}_2\mathbf{A}\mathbf{a}_1 = \mathbf{a'}_2\,(\lambda\mathbf{a}_1) = \lambda\mathbf{a'}_2\,\mathbf{a}_1$$

Ahora el lagrangiano es

$$\mathcal{L} = \mathbf{a'}_2\mathbf{A}\mathbf{a}_2 - \lambda(\mathbf{a'}_2\mathbf{a}_2 - 1) - \mu(\mathbf{a'}_2\,\mathbf{a}_1)$$

donde λ y μ son los *multiplicadores de lagrange*, y al derivar esta función con respecto a $\mathbf{a}_2$ e igualar a cero se obtiene

$$\mathbf{A}\mathbf{a}_2 - \lambda\mathbf{a}_2 - \frac{\mu}{2}\mathbf{a}_1 = 0$$

y se premultiplica ambos lados de esta ecuación por $\mathbf{a'}_1$; se obtiene

$$\mathbf{a}_1'\mathbf{A}\mathbf{a}_2 - \lambda\mathbf{a}_1'\mathbf{a}_2 - \frac{\mu}{2}\mathbf{a}_1'\mathbf{a}_1 = 0$$

se tiene que $\mathbf{a'}_1\mathbf{A} - \lambda\mathbf{a'}_1 = 0$, por lo que

$$\mathbf{a'}_1\mathbf{A}\mathbf{a}_2 - \lambda\mathbf{a'}_1\mathbf{a}_2 = 0$$

obteniendo

$$\frac{\mu}{2}\mathbf{a}_1'\mathbf{a}_1 = 0$$

Lo que implica que $\mu = 0$. Al sustituir lo anterior,

$$\mathbf{A}\mathbf{a}_2 - \lambda\mathbf{a}_2 = 0$$

por lo que

$$(\mathbf{A} - \lambda \mathbf{I}_p)\mathbf{a}_2 = 0$$

y siguiendo un razonamiento análogo al que se hizo para el primer CP, resulta que λ debe ser el segundo mayor *eigenvalor* de $\mathbf{A}$, y $\mathbf{a}_2$ debe ser el *eigenvector* correspondiente al *eigenvalor* λ.

Este procedimiento se puede seguir sucesivamente hasta hallar los p componentes principales y en todos los casos se tendrá que el *i*-ésimo CP es el *i*-ésimo *eigenvector*, ordenados decrecientemente de acuerdo a sus *eigenvalores* correspondientes, de la matriz $\mathbf{A}$.

Los resultados numéricos difieren si se utiliza la *matriz de correlaciones* o la *de covarianzas*; el escoger $\mathbf{D}$ implica considerar a todas las variables de igual importancia; sin embargo se debe recalcar que se pueden presentar situaciones en las que no es necesario estandarizar, por ejemplo si se tienen todas las variables en estudio de la misma clase: binarias, en la misma escala, porcentajes, o medidas en las mismas unidades y ordenes de magnitud.

6. Pasos para efectuar un Análisis de Componentes Principales

Los pasos para efectuar un *Análisis de Componentes Principales* son los siguientes:

1) Si las variables en estudio se miden en las mismas unidades, entonces el primer paso consiste en calcular la *matriz de covarianzas* de los datos; si las variables no tienen las mismas unidades entonces se necesita la *matriz de correlaciones*; para hallarla se estandarizan las variables originales para que tengan media cero y varianza uno y a éstas les calculamos su *matriz de dispersión*; algunos paquetes computacionales pueden encuentran directamente de los datos la *matriz de correlaciones*; a la matriz con la que se trabajará se le denota en estas notas $\mathbf{A}$.
2) Observe en la matriz $\mathbf{A}$ grupos de variables con correlaciones 'altas', si casi todas las correlaciones son 'pequeñas' entonces no tiene sentido aplicar un ACP.
3) Calcular los *eigenvalores* $\lambda_1,\ldots, \lambda_p$ y los correspondientes *eigenvectores* $\mathbf{a}_1,\ldots, \mathbf{a}_p$ de la matriz con la que se esté trabajando, ya sea la de covarianzas o la de correlaciones. Los coeficientes del *i*-ésimo componente principal están dados por $\mathbf{a}_i$, mientras que su varianza es λ_i, es decir

$$\begin{aligned} Z_i &= a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p \\ Var\,(Z_i) &= \lambda_i \end{aligned}$$

y recuérdese que los *eigenvalores* se ordenan de mayor a menor, así la varianza del primer CP es mayor a las varianzas de los demás CP´s.

4) Eliminar los CP´s que expliquen muy poco del problema en términos de variabilidad; para esto existen criterios para determinar con cuántos CP´s nos quedamos. (ver sección 6.1)
5) Observe los grupos de variables que se forman sugeridos por los componentes principales y considere si los componentes tienen alguna interpretación significativa.
6) Use las cargas de los componentes en estudios subsiguientes como una forma de reducir la dimensionalidad del problema.

Es importante observar que esta técnica no es independiente de la escala, es decir, si cambiamos la escala de medición de una de las variables, por decir, de centímetros a metros, entonces los resultados que obtenemos cambian ya que lo que hicimos fue cambiar una columna de la matriz de datos por lo que la *matriz de covarianzas*, o en su caso la *matriz de correlaciones*, también cambia y por consiguiente los *eigenvalores* y *eigenvectores* también van a cambiar y finalmente se obtiene un conjunto de CP's diferentes.

Si una de las variables tuviera varianza mucho mayor que las demás, entonces esta variable dominará en el primer CP basado en la *matriz de covarianzas*, no importando como sea la estructura de correlación entre las variables; ahora, si se escalan las variables para que tengan la misma varianza, por decir varianza uno, entonces el primer CP va a ser muy diferente al obtenido con la *matriz de covarianzas*; este problema se evita trabajando con la matriz de correlaciones para que todas las variables tengan la misma varianza.

Note que los primeros tres puntos anteriores son pasos algebraicos, fáciles de implementar en una computadora; en las siguientes secciones se describen los últimos tres puntos.

6.1. Selección del número de Componentes Principales

El paso cuatro habla de eliminar a algunos CP's; en esta sección se estudian dichos criterios para que la selección del número de CP's que permanecerán en el estudio no sea completamente arbitraria, aunque se debe reconocer que son métodos *ad hoc*.

6.1.1. Porcentaje de Variación Total Acumulada

Se sabe que la suma de las varianzas de las variables en estudio (estandarizadas o no, dependiendo si se trabaja con la *matriz de correlaciones* o *de covarianzas*, respectivamente) es igual a la suma de los *eigenvalores* de la matriz $\mathbf{A}$, es decir,

$$\sum_{i=1}^{p} \sigma_i^2 = \sum_{i=1}^{p} \lambda_i;$$

a su vez cada *eigenvalor* representa la varianza del CP correspondiente, por lo que una forma de medir la contribución a la variabilidad total del i-ésimo CP es calcular

$$\frac{\sigma_i^2}{\sum_{i=1}^{p}\sigma_i^2} = \frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i};$$

Al cociente se le conoce como ***proporción de la variabilidad total*** explicada por el i-ésimo CP.

El ***porcentaje de la variación total acumulada*** explicada por los primeros m CP ($m < p$) resulta de sumar las primeras m proporciones de la *variabilidad total* explicada por cada uno de dichos CP's multiplicada por 100, es decir

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_m}{\sum_{i=1}^{p}\lambda_i} \times 100;$$

si se trabaja con la *matriz de correlaciones*, nótese que el denominador de las expresiones anteriores es igual a p.

Para decidir qué valor debe tomar m en una situación particular, debemos examinar cuántos CP's es necesario considerar para que el *porcentaje de variación total acumulada* sea satisfactorio a nuestras necesidades; por lo general se considera que con lograr el 80% de la variación total es suficiente, aunque debe recalcarse que en algunos casos se puede requerir controlar más (o menos) la variabilidad total y ahí recae en la experiencia del investigador la decisión a tomar.

Ejemplo 4.1. Selección de CP's. En la siguiente tabla se ejemplifica tres casos en los que se tienen cinco variables en los cuales se obtuvieron los *eigenvalores* y se calculó el *porcentaje de variación total acumulada* (VTA) para cada caso.

Caso 1

CP's	Eigenvalor	% Var	VTA
1	1.75	0.35	
2	1.50	0.30	65%
3	1.40	0.28	93%
4	0.20	0.04	97%
5	0.15	0.03	100%

Caso 2

CP's	Eigenvalor	% Var	VTA
1	1.10	0.22	
2	1.05	0.21	43%
3	1.00	0.20	63%
4	0.95	0.19	81%
5	0.90	0.18	100%

Caso 3

CP's	Eigenvalor	% Var	VTA
1	3.75	0.75	
2	0.35	0.07	82%
3	0.35	0.07	89%
4	0.30	0.06	95%
5	0.25	0.05	100%

Si se busca retener un 80% de la variación total de los datos, en el *caso 1* se observa que los tres primeros componentes principales explican el 93% de la variabilidad total, y los últimos dos CP's casi no contribuyen a la variabilidad total; así, en este caso se puede considerar tomar los primeros 3 CP's y reducir de esta forma la dimensión del problema de 5 a 3.

En el *caso 2* es difícil decidir con cuántos CP's quedarse, ya que todas las proporciones de variabilidad explicada son muy parecidas, o si quedarse con la dimensión original del problema; esta decisión se puede facilitar considerando las correlaciones entre las variables dadas. Sentido estrictos, y considerando la condición impuesta, deberían considerarse los primeros 4 CP's aunque habría que estudiar la pertinencia de realizar la reducción de una sola dimensión.

En el *caso 3* se puede considerar que con un CP ya tenemos una buena parte de la variabilidad total (muy cercana al 80% deseado), además, si se observan las demás proporciones de variabilidad explicada, se puede ver que contribuyen muy poco a dicha variabilidad.

6.1.2. Tamaño de la varianza

Otro método para seleccionar el número de CP's a considerar en el estudio fue propuesto por Kaiser en 1960. Se propone considerar a los CP´s cuyos *eigenvalores* respectivos sean mayores que uno; este criterio tiende a incluir muy pocos componentes cuando el número original de variables en estudio es inferior a 20.

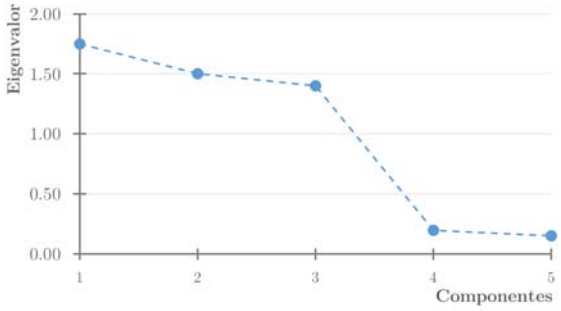
Ejemplo 4.1. Selección de CP's. *Continuación.*
Considerando las tablas anteriores, para el *caso 1* tenemos que, según el criterio de Kaiser, sólo se debería de considerar los tres primeros CP's los cuales explican el 93% de la variación total de los datos; para el segundo caso se tiene que los primeros tres CP's cumplen el criterio de Kaiser representando el 63% de la variación total acumulada. Para el último caso, un solo CP satisface el criterio impuesto y representa el 75% de la variación total acumulada.

6.1.3. Método gráfico

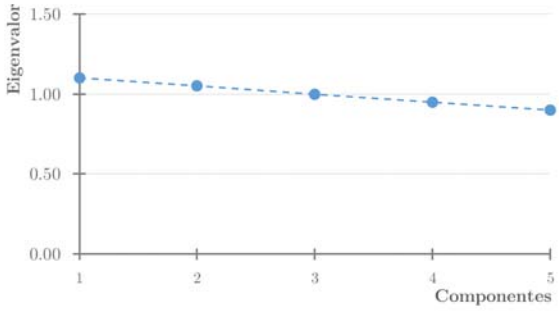
El método gráfico para seleccionar el número de CP's fue propuesto por Catell en 1966; este método consiste en graficar el *eigenvalor* por cada componente en un diagrama de dispersión, uniendo los puntos con una línea. La regla visual consiste en considerar aquellos CP's anteriores al punto de inflexión más pronunciado en la curva; a esta gráfica se le conoce como la *gráfica de ladera* o *scree plot*.

Ejemplo 4.1. Selección de CP's. *Continuación.*
Las siguiente gráfica de ladera presentan los casos analizados para determinar el número de CP's que deben considerarse. Los resultados de la inspección visual coinciden con los obtenidos en la subsección anterior para los *casos 1* y *caso 3*. La gráfica del *caso 2* no es posible distinguir el punto de inflexión.

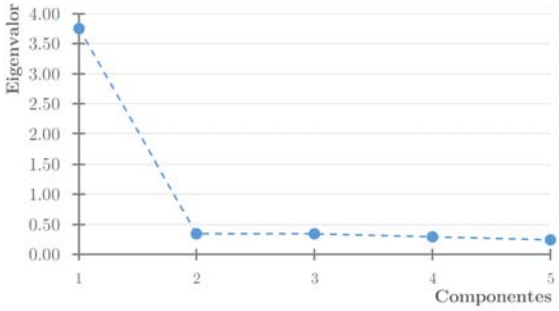
Caso 1



Caso 2



Caso 3



Existen otros métodos como el de validación cruzada y de correlaciones, para determinar el número de CP's con los que se debe quedar el analista. Más adelante se describirán las correlaciones entre los CP's y las variables originales y como éstas ayudaran a decidir el número de CP's a considerar.

Debe recalcarse que en la decisión final del número de componentes a considerar interviene la experiencia del experimentador y las necesidades del estudio; no se puede dar una fórmula mágica que resuelva todos los casos.

6.2. Representación gráfica de los datos

6.2.1. Identificación de grupos

El paso 5 mencionado en la sección 6 habla de observar grupos de variables; para hacer esto se considera la gráfica de los dos o tres primeros componentes principales, según la variabilidad total que expliquen. Para hacer esta gráfica se sustituyen los valores de $\{x_1, x_2, \ldots, x_n\}$ (n vectores aleatorios p-dimensionales, los renglones de la matriz $\mathbf{X}$) en los primeros dos (o tres) componentes principales, obteniéndose para cada individuo (renglón) dos (o tres) mediciones, representando los valores de los primeros dos (tres) CP's para dichos individuos; a estos valores se les llama *puntuaciones* o *scores*.

En la gráfica a considerar se grafican los n valores del primer CP contra los n valores del segundo CP; si se toman en cuenta tres CP's en lugar de dos, entonces se tendrá un gráfica en el espacio, en lugar del plano.

Si se transpone la matriz de datos $\mathbf{X}$, entonces el ACP puede servir para identificar grupos de individuos.

6.2.2. Identificación de *outliers*

Si se grafican los *scores* de los últimos dos componentes principales, esta gráfica puede ayudar a identificar *outliers*, o valores discrepantes; asimismo se puede utilizar la gráfica descrita anteriormente para identificar *outliers* (de los primeros componentes), aunque se sugiere que se use la gráfica con los últimos dos CP's.

Cuando se emplea la gráfica de los primeros dos CP's para reconocer *outliers*, se identifican a las observaciones que contribuyen a aumentar en alto grado la varianza y covarianza (o correlación), y si se usan los últimos dos CP's se identifican a las observaciones que contribuyen a aumentar la dimensión de los datos.

El primer caso se puede deber a la naturaleza de los datos y no necesariamente a un *outlier*; por ejemplo si una variable tiene varianza mucho mayor que las demás variables en estudio y no se estandarizan dichas variables, cuando se efectúe la gráfica de los dos primeros CP's se identificaran observaciones con valores grandes en la primera CP.

Si se supone normalidad multivariada de los datos se pueden usar gráficas de papel normal para detectar *outliers* en los primeros y en los últimos CP's

7. Correlaciones entre variables y Componentes Principales

Recuerde que la correlación entre la j–variable y el k–ésimo CP está dada por la fórmula

$$r_{jk} = \frac{cov\,(X_j, Z_k)}{\sqrt{var\,(X_j)\,\lambda_k}}$$

donde

$$\begin{aligned} cov\,(X_j, Z_k) &= cov\left(\sum_{i=1}^{p} a_{ji} Z_i, Z_k\right) \\ &= a_{jk} Var\{Z_k\} = a_{jk}\lambda_k \end{aligned}$$

así que

$$r_{jk} = \frac{a_{jk}\lambda_k}{\sqrt{s_{jj}\lambda_k}} = a_{jk}\sqrt{\frac{\lambda_k}{s_{jj}}}$$

y si se usa la matriz de correlaciones se tendrá

$$r_{jk} = a_{jk}\sqrt{\lambda_k}$$

Si se eleva al cuadrado r_{jk}, la cantidad obtenida se puede interpretar como una medida de asociación entre los componentes principales y las variables y es una manera de cuantificar la proporción de la variación total de una variable original (X_j) explicada por el componente k; es decir, para obtener la proporción de la variabilidad de la variable j debida al k–ésimo componente se calcula

$$r_{jk}^2 = \begin{cases} a_{jk}^2 \lambda_k s_{jj}^{-1} & \text{con covarianzas} \\ a_{jk}^2 \lambda_k & \text{con correlaciones} \end{cases}$$

Una propiedad importante de estos índices es

$$\sum_{k=1}^{p} r_{jk}^2 = 1$$

Una forma de medir cuál es la proporción de la varianza de cada variable original considerada después de reducir la dimensionalidad del problema seleccionando m CP's (mediante los métodos descritos anteriormente), es sumando las m primeras proporciones de la variación total de una variable original explicada por el componente k, $k = 1, \ldots, m$, es decir,

$$r_j^2 = \sum_{k=1}^{m} r_{jk}^2$$

y para cada caso se tendrá

$$r_j^2 = \begin{cases} \frac{1}{s_{jj}} \sum_{k=1}^{m} a_{jk}^2 \lambda_k & \text{con covarianzas} \\ \sum_{k=1}^{m} a_{jk}^2 \lambda_k & \text{con correlaciones} \end{cases}$$

mientras más se aproxime este índice a la unidad, mejor aproximación se tendrá.

8. Aplicaciones del Análisis de Componentes Principales

El ACP se puede considerar como un paso previo en algunas técnicas multivariadas, tal como se indica en el inciso 6) de la sección 6; las situaciones donde se puede emplear este método son las siguientes:

- *Regresión*; el uso de CP's para ajustar una ecuación de regresión fue propuesto por Kendall en 1957 para aquellas situaciones en las que las variables independientes presenten colinealidad.

- *Análisis de Conglomerados* o *Clúster Análisis*; la gráfica de los primeros dos o tres CP's puede ser útil para identificar posibles grupos de variables o individuos.

- *Análisis Discriminante*; se puede reducir la dimensionalidad del problema empleando sólo los componentes principales que contengan un porcentaje adecuado de variabilidad total acumulada.

- *Detección de Outliers*; en todas las técnicas multivariadas es necesario que no se tenga en los datos *outliers* o valores discrepantes; por medio del ACP se pueden identificar dichos valores discrepantes para eliminarlos del estudio o investigar cuál es la causa de dicho *outlier*.

Ejemplo 4.2. Medidas corporales de gorriones hembras. Después de una gran tormenta el 1 de febrero de 1898, se recolectaron algunos gorriones moribundos para el laboratorio de biología de la Universidad Brown, en Rhode Island. Días más tarde, alrededor de la mitad de los pájaros murieron y se vio una oportunidad para estudiar el efecto de la selección natural en los pájaros. Se tomaron ocho mediciones morfológicas a cada gorrión y su peso; los resultados para cinco de las variables medidas a las hembras (49 en total) se tienen en la tabla 1 al final de este capítulo.

Cuando se recolectaron los datos el principal interés era el de verificar la teoría de la selección natural de Darwin; sin embargo tomando estos datos

ahora podemos plantear la siguiente pregunta: *¿Los vivientes y no-sobrevivientes, muestran la misma variabilidad en las mediciones?*

Observemos que para este ejemplo el ACP es útil; para la matriz de datos de este ejemplo tenemos que la matriz de correlaciones está dada por

	X_1	X_2	X_3	X_4	X_5
X_1	1				
X_2	0.735	1			
X_3	0.662	0.674	1		
X_4	0.645	0.769	0.763	1	
X_5	0.605	0.529	0.526	0.607	1

recordemos que la matriz de correlaciones es una matriz simétrica, por lo que basta que especifiquemos los elementos que se encuentran por debajo (o arriba) de la diagonal principal. En este ejemplo en el que todas las variables fueron medidas en la misma escala (milímetros), podríamos trabajar con la matriz de covarianzas, en lugar de la de correlaciones.

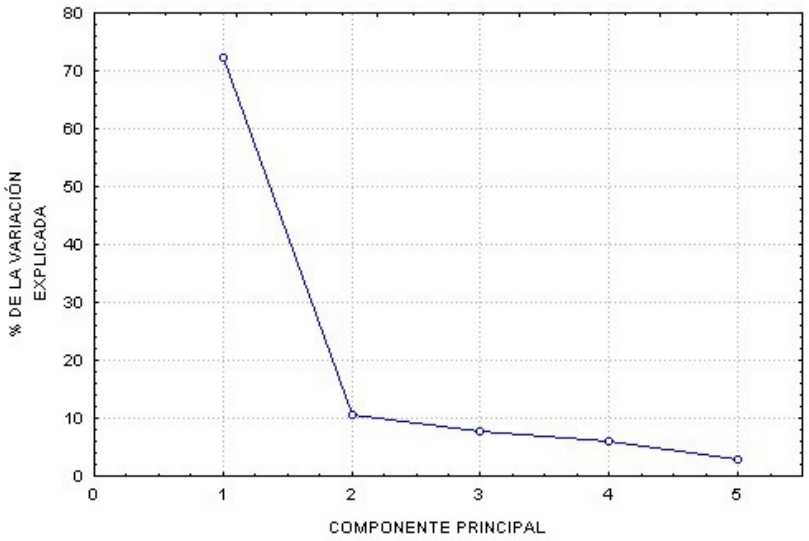
De la matriz de correlaciones podemos observar que las correlaciones entre las variables en estudio son altas (recordemos que las correlaciones varían entre -1 y 1), por lo que tiene sentido efectuar un ACP para establecer las relaciones entre las variables.

Para efectuar el paso 3) de la sección 6 se calculan los *eigenvalores* y *eigenvectores* de la matriz de correlaciones; esto se puede hacer mediante algún paquete matricial como GAUSS, el MATLAB o el S-plus. Los *eigenvalores* y *eigenvectores* se muestran en la siguiente tabla

Componente	Eigenvalor	Eigenvectores				
		X_1	X_2	X_3	X_4	X_5
1	3.616	0.452	0.462	0.451	0.471	0.398
2	0.532	0.051	-0.300	-0.325	-0.185	0.877
3	0.386	-0.690	-0.341	0.455	0.411	0.179
4	0.302	-0.420	0.548	-0.606	0.388	0.609
5	0.165	0.374	-0.530	-0.343	0.652	-0.192

De la tabla anterior se pude observar que la suma de los *eigenvalores* es 5, por lo que el primer componente principal explica el 72.3% de la variabilidad total, el segundo explica el 10.6%, el tercero explica el 7.7%, el cuarto el 6.0% y el quinto componente explica el 3.3%. Claramente se tiene que el primer componente principal explica mucho más, en términos de variabilidad, que

los demás componentes; sin embargo no es fácil decidir si debemos quedarnos con un componente (72.3% de variabilidad acumulada), con dos (82.9%) o con tres (90.6%).



Basados en la gráfica de ladera anterior se podría considerar tomar únicamente al primer CP. El primer componente principal es:

$$Z_1 = 0.452X_1 + 0.462X_2 + 0.451X_3 + 0.471X_4 + 0.398X_5$$

donde X_1, ..., X_5 son las variables estandarizadas. Este componente se puede interpretar como un promedio de las variables estandarizadas, y a este promedio se le puede asociar el concepto de tamaño; así, se puede interpretar el primer CP como un índice del tamaño de los gorriones; por tanto el 72.3% de la variabilidad de los datos se puede atribuir a las diferencias en el tamaño de los gorriones.

El segundo CP es:

$$Z_2 = -0.051X_1 + 0.3X_2 + 0.325X_3 + 0.185X_4 - 0.877X_5$$

este componente se puede interpretar como un contraste entre las variables X_2 (*longitud del ala*), X_3 (*longitud de pico y cabeza*) y X4 (*longitud del húmero*) contra X_5 (*longitud del esternón*). Entonces Z_2 es pequeña si X_2, X_3 y X_4 son pequeñas pero X_5 es grande, y Z_2 será grande si X_2, X_3 y X_4 son grandes y X_5 es pequeña; así, Z_2 se puede interpretar como un índice de la diferencia en la forma de los gorriones. Nótese que el bajo coeficiente de X_1 hace que no afecte la interpretación de este componente.

Para hallar los scores se multiplican las cargas de los componentes por los valores de cada variable (estandarizada) para cada gorrión; así, para el primer gorrión se tiene

$$\begin{aligned} Z_1 &= 0.452(-0.542) + 0.462(0.725) + 0.451(0.177) + 0.471(0.055) + \\ &\quad 0.398(-0.33) \\ &= 0.064 \end{aligned}$$

recuerde que como se trabajó con la matriz de correlaciones es necesario estandarizar primero, es decir, para el primer gorrión

$$\begin{aligned} x_1 &= \frac{156 - 157.98}{3.654} = -0.524 \\ x_2 &= \frac{245 - 241.327}{5.068} = 0.725 \\ x_3 &= \frac{31.6 - 31.459}{0.795} = 0.177 \\ x_4 &= \frac{18.5 - 18.469}{0.564} = 0.055 \\ x_5 &= \frac{20.5 - 20.827}{0.991} = -0.330 \end{aligned}$$

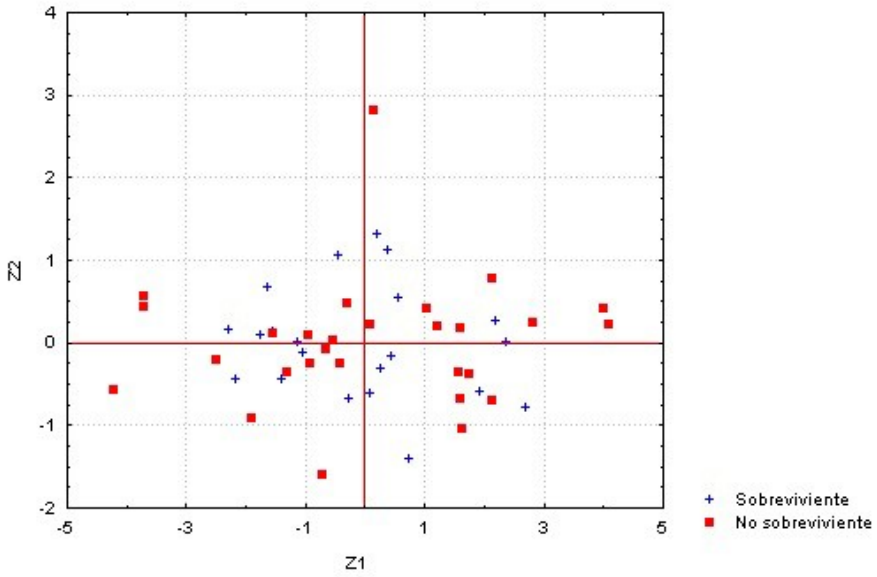
y para el segundo gorrión se tiene

$$\begin{aligned} \text{Z2} &= -0.051(-0.542) + 0.3(0.725) + 0.325(0.177) + 0.185(0.055) \\ &\quad -0.877\ (-0.33) \\ &= 0.602 \end{aligned}$$

este procedimiento se puede seguir para cada gorrión.

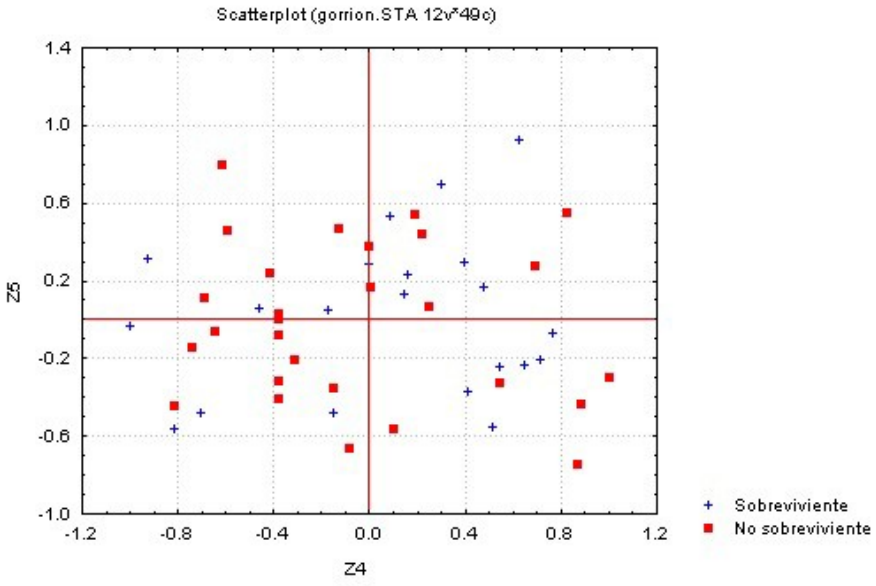
Se sabe que los primeros 21 gorriones de la tabla de datos sobrevivieron, mientras que los restantes 28 murieron; una pregunta de interés es saber si existe alguna diferencia morfológica entre los gorriones que vivieron y los que murieron; para esto se hace la gráfica de los *scores* del primer contra el segundo CP, y recuerde que estas dos componentes explican el 82.9% de la variabilidad total en los datos.

De la gráfica que se muestra más abajo se deduce que los gorriones con valores extremos en el primer componente principal murieron; se podría suponer que lo mismo ocurre para el segundo componente, aunque esto no es muy claro.



Una interpretación de lo anterior pudiera ser que el hecho de tener valores extremos en el primer componente principal influye para que sea más probable que un gorrión muera; es decir, dentro de los gorriones hembras, los gorriones pequeños y los grandes son los que tuvieron más problemas para sobrevivir.

Si se grafica los últimos dos componentes principales, esta gráfica puede indicar la existencia de *outliers*.



En esta gráfica no se observa ningún comportamiento anómalo, por lo que no se tiene evidencia para sospechar de algún *outlier*.

En la siguiente tabla se muestran las correlaciones entre los componentes principales y las variables.

	CP_1	CP_2	CP_3	CP_4	CP_5
X_1	0.8591	0.0370	-0.4292	-0.2309	0.1517
X_2	0.8779	-0.2184	-0.2117	0.3009	-0.2150
X_3	0.8567	-0.2366	0.2825	-0.3329	-0.1390
X_4	0.8951	-0.1346	0.2254	0.2132	0.2643
X_5	0.7562	0.6390	0.1109	0.0378	-0.0781

y para obtener la proporción de variabilidad en las variables originales explicadas por los CP's, se eleva cada término de la tabla anterior al cuadrado, y se obtiene

	CP_1	CP_2	CP_3	CP_4	CP_5
X_1	0.7380	0.0013	0.1842	0.0533	0.0230
X_2	0.7707	0.0476	0.0448	0.0905	0.0462
X_3	0.7339	0.0559	0.0798	0.1108	0.0193
X_4	0.8012	0.0181	0.0508	0.0454	0.0698
X_5	0.5718	0.4083	0.0122	0.0014	0.0060

De la tabla anterior se puede interpretar que el primer CP es el componente que mejor explica a las 5 variables; para la variable 5 esto no es muy claro, pues el segundo componente también aporta una buena proporción de variabilidad.

En la siguiente tabla se muestra la variación total (en porcentajes) explicada por los primeros componentes para cada variable

	CP_1	CP_2	CP_3
X_1	74	74	92
X_2	77	82	86
X_3	73	79	87
X_4	80	82	87
X_5	57	98	99

Como conclusión acerca de la dimensionalidad de este problema, se debe tener en cuenta el objetivo, si el interés es cubrir adecuadamente todas las variables, sería conveniente considerar 3 CP's; si se tiene un particular interés en la quinta variable, entonces basta con los primeros dos CP's. A través de esta tabla se observa que no es adecuado considerar sólo un componente, como lo sugería la gráfica de ladera.

Si se considera adecuado tomar dos CP's, los análisis hechos acerca del agrupamiento de los gorriones que sobrevivieron en la gráfica de los primeros dos CP's toman relevancia y serían las conclusiones a las que se llegaría.

Ejemplo 4.3. Caracterización de la producción lechera. Los datos que se usan en este ejemplo se derivan de un proyecto de la Universidad Francisco Miranda y corresponden al distrito Federación del Estado Falcón, en Venezuela.
Con el objeto de conocer la situación del sector lechero en ese distrito, se realizó una encuesta durante la cual se visitó a productores en su finca. Se reunió información acerca de una serie de variables que influyen en la producción total por finca y la productividad por finca y vaca y se efectuó un ACP. En la siguiente tabla se sintetizan los valores promedio y las desviaciones para cada variable; los índices de maquinaria y los de instalaciones se calcularon teniendo en cuenta el costo de reposición de maquinaria y equipo y su estado al momento de la visita. El índice sanitario se calculó ponderando el costo de las vacunaciones preventivas y del tratamiento curativo brindado por cada productor.

	Código	Promedio	Desviación estándar
Superficie por finca	SUP	383.51	328.25
Numero de vacas	VACA	100.13	73.23
Índice sanitario	SANI	18.17	4.00
Índice de instalaciones	INST	71.16	16.43
Índice de maquinaria	MAQ	12.38	21.57

Se calculó la matriz de covarianzas con las seis variables utilizando los datos originales; la matriz resultante es

	SUP	VACA	SANI	INST	MAQ	PROM
SUP	10,7746.30					
VACA	11,875.63	5,362.88				
SANI	268.34	24.44	15.97			
INST	389.19	156.36	10.64	269.93		
MAQ	1,704.49	293.25	14.02	34.59	465.30	
PROM	-25.30	-31.44	0.76	3.64	2.68	2.25

Las varianzas de las variables SUP y VACA son mucho mayores que las otras varianzas; esto se debe a que sus magnitudes son también mayores (ver los

promedios). También se puede distinguir que las covarianzas son altas, por lo que se supone que existe dependencia entre las variables, y por tanto tiene sentido efectuar un ACP.

En la siguiente tabla se presentan los *eigenvalores* y la proporción de la variación total explicada por cada uno de los componentes, usando la matriz de covarianzas

Componentes	Eigenvalor	Var. explicada	Var. acumulada
1	109,135.00	95.85	95.85
2	4009.67	3.52	99.37
3	439.17	0.39	99.76
4	261.79	0.23	99.99
5	14.74	0.01	100.00
6	1.91	0.01	100.00

y la tabla con los *eigenvectores*, o sea los componentes principales, es la siguiente

Variable	CP_1	CP_2	CP_3	CP_4	CP_5	CP_6
SUP	0.9934	-0.1142	-0.0126	-0.0015	-0.0023	-0.0003
VACA	0.1137	0.9926	-0.0331	0.0254	0.0032	0.0076
SANI	0.0025	-0.0014	0.0265	-0.0335	0.9982	-0.0424
INST	0.0037	0.0299	0.1443	-0.9882	-0.0376	-0.0149
MAQ	0.0159	0.0275	0.9885	0.1462	-0.0216	-0.0070
PROM	-0.0003	-0.0070	0.0104	-0.0154	0.0417	0.9989

De esta tabla se puede ver que el primer CP refleja la variación en la superficie de las fincas, el segundo en el número de vacas por finca, el tercero en el índice de maquinaria, el cuarto en el índice de instalaciones, el quinto en el índice sanitario y el sexto en el promedio de leche por vaca.

En la siguiente tabla se muestran las correlaciones entre las variables originales y los componentes principales

Variable	CP_1	CP_2	CP_3	CP_4	CP_5	CP_6
SUP	0.9956	0.0005	0.0000	0.0000	0.0000	0.0000
VACA	0.2631	0.7366	0.0000	0.0000	0.0000	0.0000
SANI	0.0427	0.0005	0.0184	0.0184	0.9199	0.0002
INST	0.0055	0.0133	0.9471	0.9471	0.0001	0.0000
MAQ	0.0593	0.0652	0.0120	0.0120	0.0000	0.0000
PROM	0.0044	0.0874	0.0276	0.0276	0.0114	0.8485

Al analizar esta tabla se observa la misma situación anterior del ejemplo anterior. El 99.56% de la variación de la variable SUP queda explicada por el primer CP y el 0.05% restante lo explica el segundo CP, así que si se escogen estos dos CPs, se habrá explicado al 100% la variación de la variable SUP.

El 26.31% de la variable VACA está sintetizado en el primer CP y el 73.66% en el segundo; al considerar a estos dos CP's se conocerá el 99.99% de la variación de VACA.

Sin embargo el aporte en las otras cuatro variables será muy pequeño y por ende la variación explicada por los dos primeros CP's será muy baja: 4.32% para SANI, 1.88% para INST, 12.45% para MAQ y 9.18% para PROM.

Esta situación, frecuente cuando se trabaja con variables que se expresan en distintas unidades y cuyos recorridos de valores difieren en uno o más órdenes de magnitud, puede influir de manera apreciable en la interpretación de resultados.

Se recomienda examinar el efecto que tiene sobre el análisis la estandarización de los datos originales; no hacerlo implica suponer que la variable que asume valores absolutos más grandes es la que influye en el análisis.

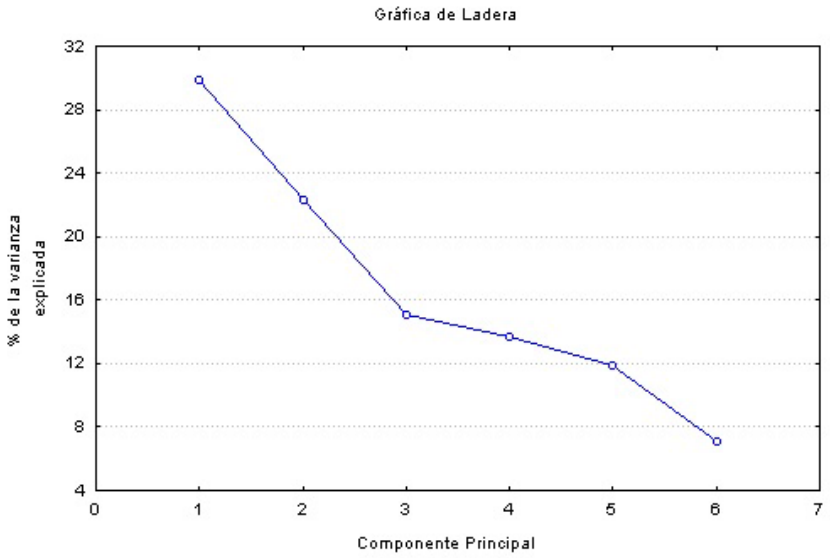
La matriz de correlación para estos datos es la siguiente

	SUP	VACA	SANI	INST	MAQ	PROM
SUP	1.0000					
VACA	0.4940	1.0000				
SANI	0.2046	0.0835	1.0000			
INST	0.0722	0.1300	0.1620	1.0000		
MAQ	0.2407	0.1856	0.1627	0.0976	1.0000	
PROM	-0.0514	-0.2865	0.1265	0.1477	0.0829	1.0000

En la siguiente tabla se presenta los *eigenvalores* y la varianza explicada y acumulada total

Componentes	Eigenvalor	Var. explicada	Var. acumulada
1	1.79412	29.90	29.9
2	1.34091	22.53	52.25
3	0.90316	15.05	67.30
4	0.82129	13.69	80.99
5	0.71244	11.88	92.87
6	0.42805	7.13	100.00

En este caso el primer componente sólo sintetiza el 29.9% y los dos primeros el 52.25%; es necesario considerar hasta 4 CP's para obtener el 81% de la variación y hasta el quinto para 93%. El criterio de Kaiser sugiere en esta situación tomar los dos primeros componentes; asimismo, la gráfica de ladera también sugiere considerar los dos o tres primeros CP's.



En la siguiente tabla se muestran los *eigenvectores* de la matriz de correlaciones

Variable	CP_1	CP_2	CP_3	CP_4	CP_5	CP_6
SUP	0.5866	0.0999	-0.1344	-0.0883	-0.5341	-0.5787
VACA	0.5692	0.3415	0.1840	0.0625	-0.1361	0.7093
SANI	0.3247	-0.4176	-0.0818	-0.7734	0.3267	0.0925
INST	0.2418	-0.4118	0.7935	0.2820	0.1554	-0.1927
MAQ	0.3961	-0.2286	-0.5297	0.5397	0.4671	-0.0283
PROM	-0.1053	-0.6908	0.1767	0.1389	0.5892	0.3376

Los coeficientes del primer CP indican que es un promedio entre todas las variables, con ponderación ligeramente mayor para SUP y VACA; el coeficiente de PROM es negativo, así que disminuirá el valor del primer componente si aumenta la producción de leche por vaca. Sin embargo su contribución es pequeña y valores elevados del primer componente estarán asociados a valores elevados de las variables SUP, VACA, SANI, INST y MAQ.

Las fincas con mayores rebaños lecheros, la menor productividad por vaca, las peores condiciones sanitarias y deficiencias en instalaciones y equipo serán las que tengan valores mayores en la segunda CP.

Si consideramos únicamente dos CP's, se puede decir que las mejores fincas serán aquellas que tengan valores elevados para la primera CP y valores pequeños para la segunda CP.

En la siguiente tabla se muestra la variación explicada por los primeros dos CP's para cada variable en porcentaje

Variable	Var. Explicada (2 CP's)
SUP	62
VACA	74
SANI	42
INST	33
MAQ	35
PROM	65

Si se contara con los datos, se debería de efectuar una gráfica de los *scores* del primer CP contra los *scores* del segundo CP para tratar de identificar grupos con problemas en las variables planteadas. Antes de hacer interpretaciones se debe calcularlos.

Ejemplo 4.4. Empleo de países Europeos. Considere la tabla 2 que viene al final de este capítulo. En esta tabla se presentan los porcentajes de gente empleada en nueve sectores de la industria en Europa; la matriz de correlaciones se presenta en la siguiente tabla

	AGR	MIN	FAB	DB	CON	SIND	FIN	SS	TC
AGR	1								
MIN	0.036	1							
FAB	-0.671	0.445	1						
DB	-0.400	0.406	0.385	1					
CON	-0.538	-0.026	0.495	0.060	1				
SIND	-0.737	-0.397	0.204	0.202	0.356	1			
FIN	-0.220	-0.443	-0.156	0.110	0.016	0.366	1		
SS	-0.747	-0.281	0.154	0.132	0.158	0.572	0.108	1	
TC	-0.565	0.157	0.351	0.375	0.338	0.188	-0.246	0.568	1

donde AGR=agricultura, MIN=minería, FAB=fabricas, DB=transporte de bienes, CON=construcción, SIND=servicios industriales, FIN=finanzas, SS=servicios sociales y TC=comunicaciones y transporte.

De esta matriz se puede observar que no hay correlaciones muy altas, por lo que debemos esperar que se necesitarán varios componentes principales para explicar una variabilidad aceptable de los datos.

Los *eigenvalores* y la varianza acumulada se presentan en la siguiente tabla

Componentes	Eigenvalor	Var. explicada	Var. acumulada
1	3.487	38.74 %	38.74 %
2	2.130	23.66 %	62.40 %
3	1.099	12.21 %	74.61 %
4	0.995	11.05 %	85.66 %
5	0.543	6.03 %	91.69 %
6	0.383	4.25 %	95.94 %
7	0.226	2.51 %	98.45 %
8	0.137	1.52 %	99.97 %

nótese que el último *eigenvalor* es cero (no aparece en la tabla).

Siguiendo el criterio del 80% de variabilidad acumulada, se puede considerar conservar los primeros cuatro CP's; sin embargo, también se puede considerar que sólo los primeros dos CP´s son importantes, pues son claramente mayores que uno, siguiendo el criterio de Kaiser, y si consideramos la gráfica de ladera, podremos observar que después los primeros cuatro *eigenvalores* son substancialmente mayores a los otros cinco, así que se debería de considerar a los primeros cuatro CP's.

Para la decisión final del número de componentes a considerar se debe tener en cuenta los objetivos del estudio; en este caso por simplicidad se conservarán los dos primeros CP's.

El primer CP es

$$Z_1 = 0.52(\text{AGR}) + 0(\text{MIN}) - 0.35(\text{FAB}) - 0.26(\text{DB}) - 0.33(\text{CON}) \\ -0.38(\text{SIND}) - 0.07(\text{FIN}) - 0.39(\text{SS}) - 0.37(\text{TC})$$

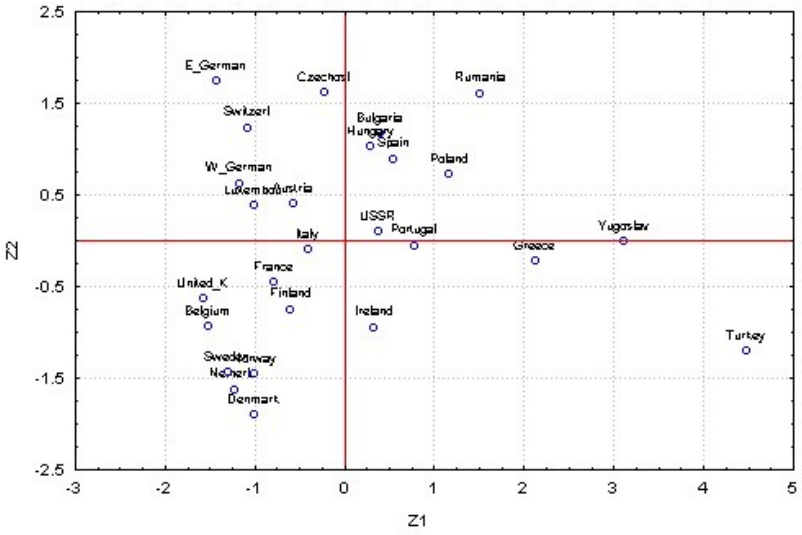
recuerde que como se trabajó con la matriz de correlaciones, las variables en Z1 deben estar estandarizadas; este componente se puede interpretar como

un contraste entre agricultura y fábricas, transporte de bienes, construcción, servicios industriales, servicios sociales y comunicaciones y transportes.

El segundo componente es

$$\begin{aligned} Z2 = {} & 0.05(\text{AGR}) + 0.62(\text{MIN}) + 0.36(\text{FAB}) + 0.26(\text{DB}) - 0.05(\text{CON}) \\ & -0.35(\text{SIND}) - 0.45(\text{FIN}) - 0.22(\text{SS}) + 0.20(\text{TC}) \end{aligned}$$

el cual contrasta los coeficientes de minería y fabricas con los de servicios industriales y finanzas. En la siguiente figura se muestra los 26 países en una gráfica de dispersión de Z_1 contra Z_2;



La mayoría de las democracias del Oeste se agruparon con valores bajos para Z_1 y Z_2. Irlanda, Portugal, España y Grecia tienen altos valores de Z_1 y los países comunistas, a excepto Yugoslavia, tienen valores altos para Z_2.

Tablas de Datos

Ave	L	A	C	H	E	CS	Ave	L	A	C	H	E	CS
ave01	156	245	31.6	18.5	20.5	1	ave26	160	250	31.7	18.8	22.5	0
ave02	154	240	30.4	17.9	19.6	1	ave27	155	237	31.0	18.5	20.0	0
ave03	153	240	31.0	18.4	20.6	1	ave28	157	245	32.2	19.5	21.4	0
ave04	153	236	30.9	17.7	20.2	1	ave29	165	245	33.1	19.8	22.7	0
ave05	155	243	31.5	18.6	20.3	1	ave30	153	231	30.1	17.3	19.8	0
ave06	163	247	32.0	19.0	20.9	1	ave31	162	239	30.3	18.0	23.1	0
ave07	157	238	30.9	18.4	20.2	1	ave32	162	243	31.6	18.8	21.3	0
ave08	155	239	32.8	18.6	21.2	1	ave33	159	245	31.8	18.5	21.7	0
ave09	164	248	32.7	19.1	21.1	1	ave34	159	247	30.9	18.1	19.0	0
ave10	158	238	31.0	18.8	22.0	1	ave35	155	243	30.9	18.5	21.3	0
ave11	158	240	31.3	18.6	22.0	1	ave36	162	252	31.9	19.1	22.2	0
ave12	160	244	31.1	18.6	20.5	1	ave37	152	230	30.4	17.3	18.6	0
ave13	161	246	32.3	19.3	21.8	1	ave38	159	242	30.8	18.2	20.5	0
ave14	157	245	32.0	19.1	20.0	1	ave39	155	238	31.2	17.9	19.3	0
ave15	157	235	31.5	18.1	19.8	1	ave40	163	249	33.4	19.5	22.8	0
ave16	156	237	30.9	18.0	20.3	1	ave41	163	242	31.0	18.1	20.7	0
ave17	158	244	31.4	18.5	21.6	1	ave42	156	237	31.7	18.2	20.3	0
ave18	153	238	30.5	18.2	20.9	1	ave43	159	238	31.5	18.4	20.3	0
ave19	155	236	30.3	18.5	20.1	1	ave44	161	245	32.1	19.1	20.8	0
ave20	163	246	32.5	18.6	21.9	1	ave45	155	235	30.7	17.7	19.6	0
ave21	159	236	31.5	18.0	21.5	1	ave46	162	247	31.9	19.1	20.4	0
ave22	155	240	31.4	18.0	20.7	0	ave47	153	237	30.6	18.6	20.4	0
ave23	156	240	31.5	18.2	20.6	0	ave48	162	245	32.5	18.5	21.1	0
ave24	160	242	32.6	18.8	21.7	0	ave49	164	248	32.3	18.8	20.9	0
ave25	152	232	30.3	17.2	19.8	0							

País	AGR	MIN	FAB	DB	CON	SIND	FIN	SS	TC
Belgium	3.3	0.9	27.6	0.9	8.2	19.1	6.2	26.6	7.2
Denmark	9.2	0.1	21.8	0.6	8.3	14.6	6.5	32.2	7.1
France	10.8	0.8	27.5	0.9	8.9	16.8	6	22.6	5.7
W. Germani	6.7	1.3	35.8	0.9	7.3	14.4	5	22.3	6.1
Ireland	23.2	1	20.7	1.3	7.5	16.8	2.8	20.8	6.1
Italy	15.9	0.6	27.6	0.5	10	18.1	1.6	20.1	5.7
Luxembourg	7.7	3.1	30.8	0.8	9.2	18.5	4.6	19.2	6.2
Netherlands	6.3	0.1	22.5	1	9.9	18	6.8	28.5	6.8
United Kingdom	2.7	1.4	30.2	1.4	6.9	16.9	5.7	28.3	6.4
Austria	12.7	1.1	30.2	1.4	9	16.8	4.9	16.8	7
Finland	13	0.4	25.9	1.3	7.4	14.7	5.5	24.3	7.6
Greece	41.4	0.6	17.6	0.6	8.1	11.5	2.4	11	6.7
Norway	9	0.5	22.4	0.8	8.6	16.9	4.7	27.6	9.4
Portugal	27.8	0.3	24.5	0.6	8.4	13.3	2.7	16.7	5.7
Spain	22.9	0.8	28.5	0.7	11.5	9.7	8.5	11.8	5.5
Sweden	6.1	0.4	25.9	0.8	7.2	14.4	6	32.4	6.8
Switzerland	7.7	0.2	37.8	0.8	9.5	17.5	5.3	15.4	5.7
Turkey	66.8	0.7	7.9	0.1	2.8	5.2	1.1	11.9	3.2
Bulgaria	23.6	1.9	32.3	0.6	7.9	8	0.7	18.2	6.7
Czechoslovakia	16.5	2.9	35.5	1.2	8.7	9.2	0.9	17.9	7
E. Germany	4.2	2.9	41.2	1.3	7.6	11.2	1.2	22.1	8.4
Hungary	21.7	3.1	29.6	1.9	8.2	9.4	0.9	17.2	8
Poland	31.1	2.5	25.7	0.9	8.4	7.5	0.9	16.1	6.9
Rumania	34.7	2.1	30.1	0.6	8.7	5.9	1.3	11.7	5
USSR	23.7	1.4	25.8	0.6	9.2	6.1	0.5	23.6	9.3
Yugoslavia	48.7	1.5	16.8	1.1	4.9	6.4	11.3	5.3	4