BITMONEY

# HERRAMIENTA:

# COMPONENTES PRINCIPALES

BITMONEY

# Autores

**M.I.C. Carlos Abraham Carballo Monsivais**

**I.S.C. Leticia Edith Trujillo Ballesteros**

**C.L.M.A. Sacbe García García**

BITMONEY

Hackathon Blockchain 2020

Elaboración 2020, Primera edición

BITMONEY

# Análisis de Componentes Principales

## Contenido

BITMONEY

# HERRAMIENTA. Análisis de Componentes Principales

## 1.1. Planteamiento

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. 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. 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.

El *Análisis de Componentes Principales*, ACP, 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.

BITMONEY

- Reducir la dimensión del problema cuidando 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 de Conglomerados o 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.

## 1.2. 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 **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 **x**, en el caso discreto, por

$$F(\mathbf{x}^0) \;=\; \mathrm{P}(\mathbf{x} = \mathbf{x}^0)$$

BITMONEY

$$= \mathrm{P}(X_1 = x_1{}^0, X_2 = x_2{}^0, \ldots, X_\mathrm{p} = x_\mathrm{p}{}^0)$$

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

$$F(\mathbf{x}^0) = \mathrm{P}(\mathbf{x} \leq \mathbf{x}^0)$$

$$= \mathrm{P}(X_1 \leq x_1{}^0, X_2 \leq x_2{}^0, \ldots, X_\mathrm{p} \leq x_\mathrm{p}{}^0)$$

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 dos variables aleatorias, $X_i$ y $X_j$, la covarianza entre ellas, denotada por $\boldsymbol{\sigma}_{ij}$ se define por

$$Cov(X_i, X_j) = E\{(X_i - \mu_i)(X_j - \mu_j)\}$$

$$= E\{X_i X_j\} - \mu_i \mu_j = \boldsymbol{\sigma}_{ij}$$

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

$$\Sigma = E\{(\mathbf{x} - \boldsymbol{\mu})\,(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\}$$

$$= E\{\mathbf{x}\,\mathbf{x}^{\mathrm{T}}\} - \boldsymbol{\mu}\,\boldsymbol{\mu}^{\mathrm{T}}$$

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 $\rho$ es simétrica y semipositiva definida. El vector $\boldsymbol{\mu}$, la matriz de covarianzas $\boldsymbol{\Sigma}$, y la matriz de correlaciones $\boldsymbol{\rho}$ son parámetros poblacionales del vector aleatorio $\mathbf{x}$. Para hallar sus estimaciones considere 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,

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

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 $\boldsymbol{\Sigma}$ está dada por

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

BITMONEY

donde

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

$\sqrt{s_{ii}}$ y $\sqrt{s_{jj}}$ 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 $\mu$ y varianza $\sigma^2$ 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 **x** tiene una distribución normal multivariada si su función de densidad conjunta es

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

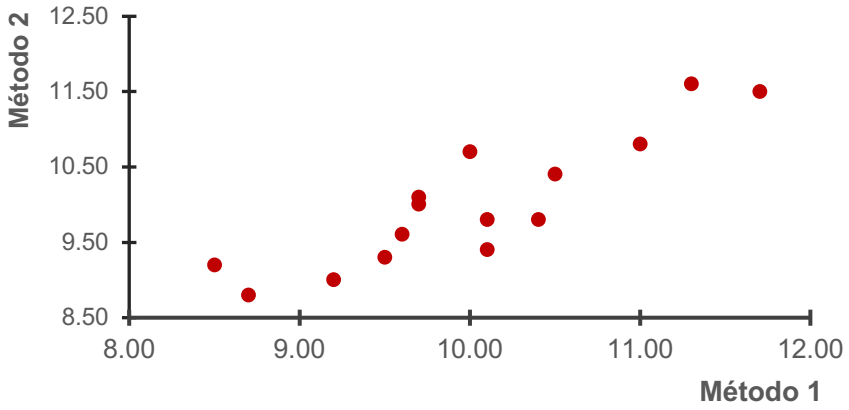
donde **Σ** es una matriz de tamaño $p \times p$ positiva definida. Se puede comprobar que ésta es una función de densidad y que **μ** es la media de **x** con esta función de densidad y que **Σ** es la matriz de covarianzas de **x**. La notación para indicar que un vector aleatorio tiene una densidad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

**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 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?

BITMONEY

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.

Ahora se estiman 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*.

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

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

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

## 1.2.1. Más sobre valores y vectores propios

El *método de componentes principales* se basa en el resultado clave del álgebra de matrices: si **A** es una matriz *p* × *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 los *valores propios*, *valores característicos*, *raíces latentes* o *eigenvalores* de **S**. Las columnas de **U**, $u_1$, $u_2$, ..., $u_p$ se llaman los *vectores propios*, *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 **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 4.2. 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, $t_{11}$ = 1

$$-0.6478 + 0.6793\, t_{21} = 0$$

$$t_{21} = 0.9538$$

$$u_1 = \frac{t_1}{\sqrt{t_1' t_1}} = \frac{1}{\sqrt{1.9097}} \begin{bmatrix} 1 \\ 0.9538 \end{bmatrix} = \begin{bmatrix} 0.7236 \\ 0.6902 \end{bmatrix}$$

Similarmente, utilizando $l_2 = 0.0864$ entonces,

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

Así

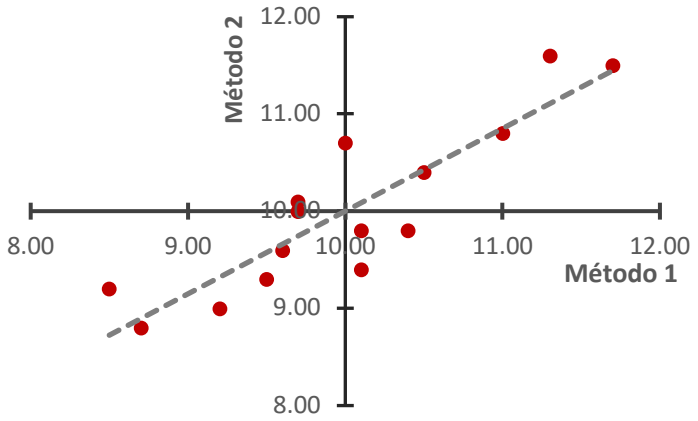
$$\mathbf{U} = \begin{bmatrix} u_1 & u_2 \end{bmatrix} = \begin{bmatrix} 0.7236 & -0.6902 \\ 0.6902 & 0.7236 \end{bmatrix}$$

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

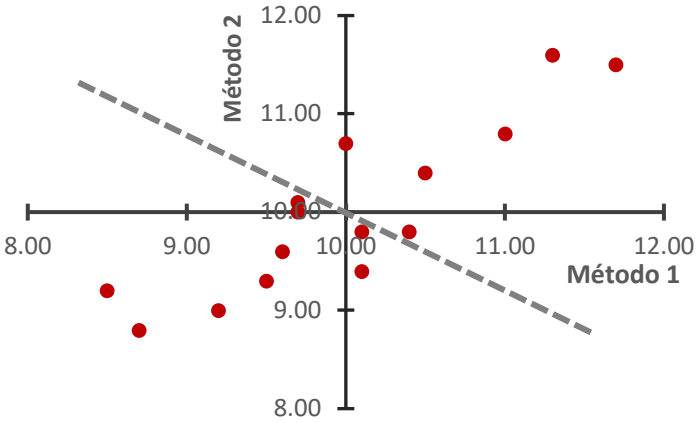
Mas aun

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

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 ángulo 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 $\mathbf{u}_2$.




## 1.3. Definición y obtención 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.

BITMONEY

Así, el *primer componente principal* es una combinación lineal de las variables $X_1, \ldots, X_p$, 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 esté 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 componentes principales (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 *vector propio* de la *matriz covarianzas* o la *de correlaciones*, según con la que se esté trabajando, ordenados con respecto a sus *valores propios* respectivos de mayor a menor y el *i*-ésimo *valores propios* (ordenados como se indicó) representa la varianza del *i*-ésimo CP.

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

Como se mencionó en la sección anterior al obtener los valores y vectores propios se generan proyecciones de los datos y de las variables en otros ejes representados por los componentes principales. Como se muestra en la **Figura *1***, las ecuaciones resultantes (*Z*) para cada componente permiten proyectar a los datos en el nuevo espacio. Lo anterior es de gran ayuda, ya que permite mediante una herramienta gráfica visualizar la estructura que guardan los datos multivariados reduciendo la dimensión original.

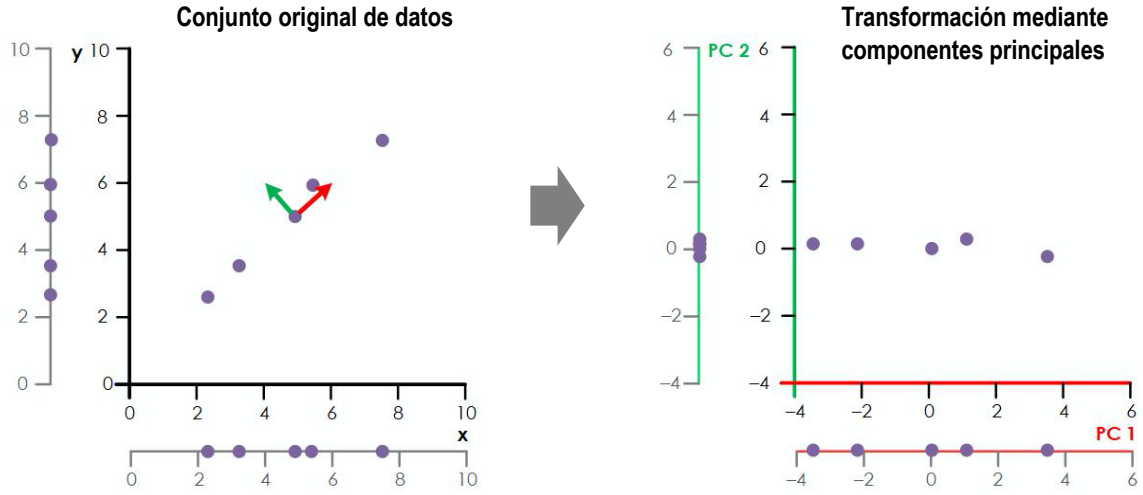



**Figura 1: Visualización de transformación mediante componentes principales.**

## 1.3.1. 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.

Considérese a la matriz de datos **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 **A** = **S**, es decir, **A** será la *matriz de covarianzas* de **X**, de lo contrario se toma **A** = **D**, es decir, como la *matriz de correlaciones* de **X**. En caso de que **A** = **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

BITMONEY

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

donde

$$\mathbf{a}_1{}^{\mathrm{T}} = [a_{11}, a_{12}, \ldots, a_{1p}]$$

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

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

$$Var\ \{Z_1\} = \mathbf{a}_1{}^{\mathrm{T}}\mathbf{A}\mathbf{a}_1$$

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

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

donde $\lambda$ 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 $\lambda$ resulta ser un *valor propio* de la matriz **A** obteniendo la siguiente relación

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

entonces $\mathbf{a}_1$ es el *vector propio* correspondiente al *valor propio* $\lambda$; sin embargo **A** tiene *p valores propios*, ¿cuál se debe tomar?, la respuesta se basa en lo siguiente, obsérvese que

$$\mathbf{a}_1{}^{\mathrm{T}}\mathbf{A}\mathbf{a}_1 = \mathbf{a}_1{}^{\mathrm{T}}\ (\lambda\mathbf{a}_1) = \lambda\ \mathbf{a}_1{}^{\mathrm{T}}\mathbf{a}_1 = \lambda$$

por lo que $Var\{Z_1\} = \lambda$; así, si se quiere maximizar $Var\{Z_1\}$, lo que se debe hacer es tomar $\lambda$ como el mayor *valor propio* de la matriz **A**, y $\mathbf{a}_1$ resulta ser el *vector propio* correspondiente al *valor propio* $\lambda$.

Para obtener el segundo CP, recuérdese que

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

BITMONEY

donde

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

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

$$Var\ \{Z_2\} = \mathbf{a}_2^{\mathrm{T}}\mathbf{A}\mathbf{a}_2$$

Sujeto a las restricciones

$$\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 = 1 \text{ (normalidad)}$$

$$\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_1 = 0 \text{ (ortogonalidad)}$$

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

$$0 = Cov(Z_1, Z_2) = Cov(\mathbf{a}_1^{\mathrm{T}}\mathbf{x}, \mathbf{a}_2^{\mathrm{T}}\mathbf{x}) = \mathbf{a}_1^{\mathrm{T}}\mathbf{A}\mathbf{a}_2 = \mathbf{a}_2^{\mathrm{T}}\mathbf{A}\mathbf{a}_1 = \mathbf{a}_2^{\mathrm{T}}\ (\lambda\mathbf{a}_1) = \lambda\mathbf{a}_2^{\mathrm{T}}\ \mathbf{a}_1$$

Ahora el lagrangiano es

$$\mathcal{L} = \mathbf{a}_2^{\mathrm{T}}\mathbf{A}\mathbf{a}_2 - \lambda(\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 - 1) - \mu(\mathbf{a}_2^{\mathrm{T}}\ \mathbf{a}_1)$$

donde $\lambda$ y $\mu$ 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 $\mathbf{a}_1^{\mathrm{T}}$; se obtiene

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

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

$$\mathbf{a}_1^{\mathrm{T}}\mathbf{A}\mathbf{a}_2 - \lambda\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_2 = 0$$

obteniendo

$$\frac{\mu}{2}\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 = 0$$

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 *valor propio* de **A**, y $\mathbf{a}_2$ debe ser el *vector propio* correspondiente al *valor propio* 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 *vector propio*, ordenados decrecientemente de acuerdo a sus *valores propios* correspondientes, de la matriz **A**.

Los resultados numéricos difieren si se utiliza la *matriz de correlaciones* o la *de covarianzas*; el escoger **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 órdenes de magnitud.

### 1.3.2. 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

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

BITMONEY

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 es sumando las *m* primeras proporciones de la variación total de una variable original explicada por el componente *k, k* = 1, …, *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.

## 1.4. Selección del número de componentes principales

Una de las preguntas que frecuentemente surge es si son necesario considerar en un ACP todos los *p* componentes principales. En esta sección se estudian los criterios más empleados 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*.

### 1.4.1. 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 *valores propios* de la matriz **A**, es decir,

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

a su vez cada *valor propio* 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

BITMONEY

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

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, se debe examinar cuántos CP's es necesario considerar para que el *porcentaje de variación total acumulada* sea satisfactorio a las necesidades del analista. 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 4.3. 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 *valores propios* 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**

BITMONEY

| 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 tres CP's y reducir de esta forma la dimensión del problema de cinco a tres.

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 estricto, y considerando la condición impuesta, deberían considerarse los primeros cuatro 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.

### 1.4.2. 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 *valores propios* 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 4.4. Selección de CP's**. (*Continuación*)

Considerando las tablas anteriores, para el *caso 1* se tiene 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.

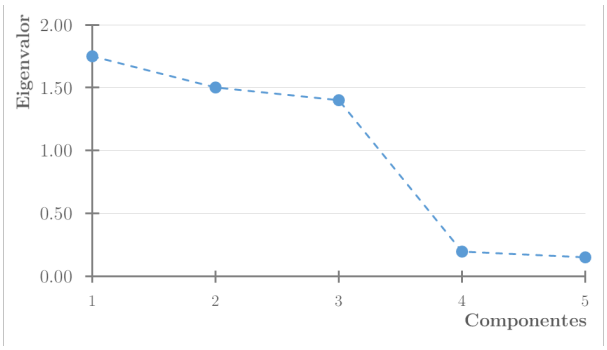
## 1.4.3. Método gráfico

El método gráfico para seleccionar el número de CP's fue propuesto por Catell en 1966. Este método consiste en graficar el *valor propio* 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*.

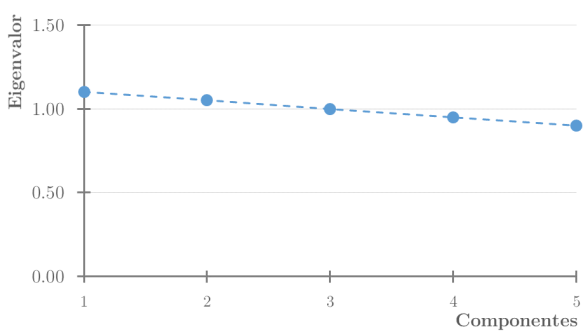
**Ejemplo 4.5. Selección de CP's**. (Continuación)

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.

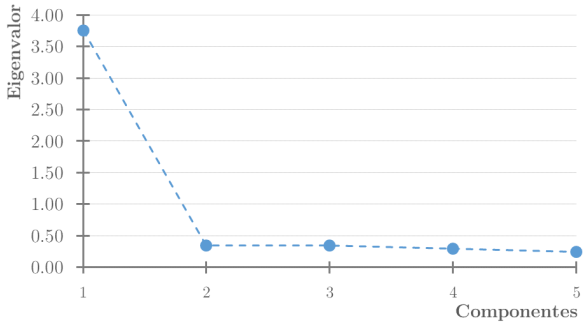
**Caso 1**




**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.

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.

## 1.5. Representación gráfica

Una forma de representar las proyecciones de cada observación y las variables al mismo tiempo en una dimensión más pequeña es mediante el *gráfico biplot*. Este grafico fue introducido por Gabriel en 1971 y constituye una opción de representación visual de la estructura de datos multivariantes. El término "*bi*" hace referencia a la posibilidad de representar, en el mismo gráfico, tanto las variables como los individuos. Las variables son representadas mediante vectores (flechas) mientras que las observaciones aparecen similar al gráfico de dispersión anterior. Como se muestra en la Figura 2 lo que busca este tipo de representación gráfica es visualizar la estructura de los datos proyectándolos en un plano fácil de interpretar. Para lograr dicha representación es necesario reducir la dimensionalidad de la base de datos utilizando los *vectores propios* asociados a los componentes seleccionados. Al realizar la reducción se está dispuesto a sacrificar un porcentaje de información (variablidad) de los datos.

Los gráficos biplot son importantes ya que la lectura del comportamiento de los datos multivariados se basa en conceptos geométricos sencillos:

1. El grado de similitud entre individuos bajo estudio es inverso a la distancia entre los mismos.
2. De cierta forma, la longitud y los ángulos de las flechas que representan las variables se interpretan en términos de variabilidad y variabilidad respectivamente. Es decir, mientras más grande el tamaño de la flecha más variabilidad es representado por los componentes graficados. Si el espacio de separación de dos flechas (variables) indica la covarianza, o correlación según la matriz empleada. Ángulos muy pequeños indican una correlación alta positiva entre las variables, un ángulo cerca a los 90° significa una correlación nula entre las variables y, un ángulo que sobrepasa los 90° es muestra de una correlación negativa.
3. La relación entre individuos y variables se interpretan en términos de las proyecciones de los individuos (puntos) sobre los vectores (flechas).

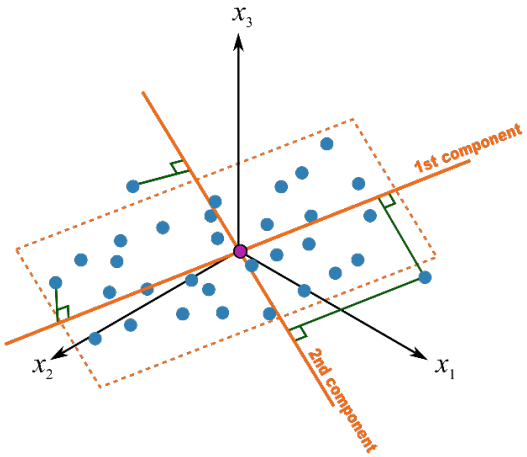


**Figura 2: Esquema de la proyección de individuos y variables en un plano**

Un ejemplo de lo mencionado en el párrafo anterior se aprecia en la Figura 3. Por lo que, para la presentación de los resultados de una investigación donde se haya aplicado alguna técnica multivariada ya sea para conocer la estructura de los datos o una reducción de dimensiones, el gráfico biplot permite obtener conclusiones generales y es más fácil de entender para aquellas personas no expertas en el uso de técnicas multivariadas.

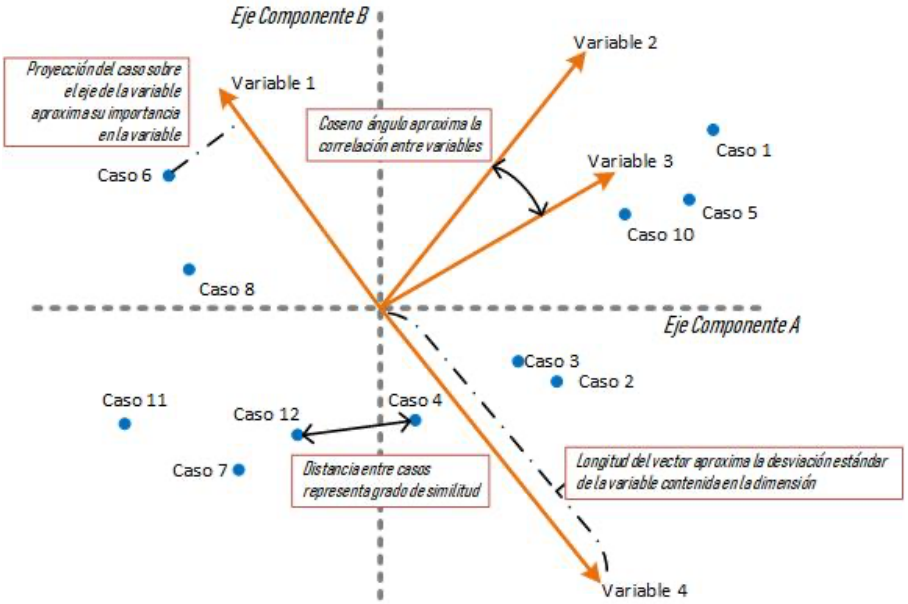



**Figura 3: Interpretación básica de las principales relaciones en el gráfico biplot**

## 1.5.1. Identificación de grupos

La representación gráfica de los componentes permite, entre otras cosas, observar la formación de 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 $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ ($n$ 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, se indican 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 una 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.

BITMONEY

### 1.5.2. 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

## 1.6. Pasos para realizar 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 hallar la matriz de correlaciones se estandarizan las variables originales para que tengan media cero y varianza uno y a éstas les calcula su *matriz de dispersión.* Algunos paquetes computacionales pueden encontrar directamente de los datos la *matriz de correlaciones.* A esta matriz se le denota como **A**.
2) Observe en la matriz **A** si existen grupos de variables con correlaciones "altas", si casi todas las correlaciones son "pequeñas" entonces no tiene sentido aplicar un ACP.
3) Calcular los *valores propios* $\lambda_1,\ldots,\lambda_p$ y los correspondientes *vectores propios* $\mathbf{a}_1,\ldots,\mathbf{a}_p$ 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 $\lambda_i$, es decir

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p$$

$$Var\,(Z_i) = \lambda_i$$

recuérdese que los *valores propios* 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 utilice los criterios de la sección 4.4. para determinar cuántos CP´s se deben considerar.
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 se cambia la escala de medición de una de las variables, por decir, de centímetros a metros, entonces los resultados que se obtienen cambian ya que se cambió una columna de la matriz de datos por lo tanto la *matriz de covarianzas*, o en su caso la *matriz de correlaciones*, también cambia y por consiguiente los *valores propios* y *vectores propios* 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 del procedimiento son pasos algebraicos, fáciles de implementar en una computadora. En los últimos el criterio que tome el analista es importante.

Revise el material adicional que se pone a disposición para revisar algunos problemas prácticos donde el ACP fue relevante. Así mismo, se presentan ejemplos desarrollados donde se indican los comandos a ejecutar para aplicar esta técnica utilizando el paquete estadístico R. Puede consultar los siguientes enlaces para profundizar el uso de esta herramienta:

- ➔ https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c
- ➔ https://towardsdatascience.com/the-mathematics-behind-principal-component-analysis-fff2d7f4b643

# Índice de Figuras