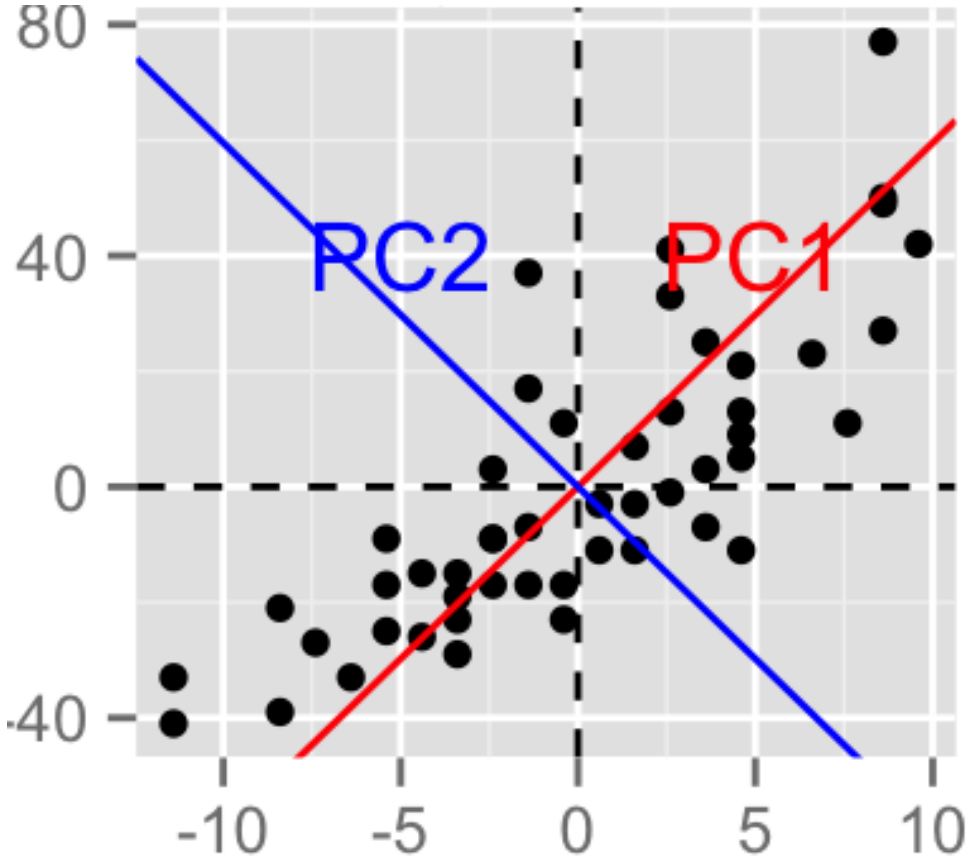
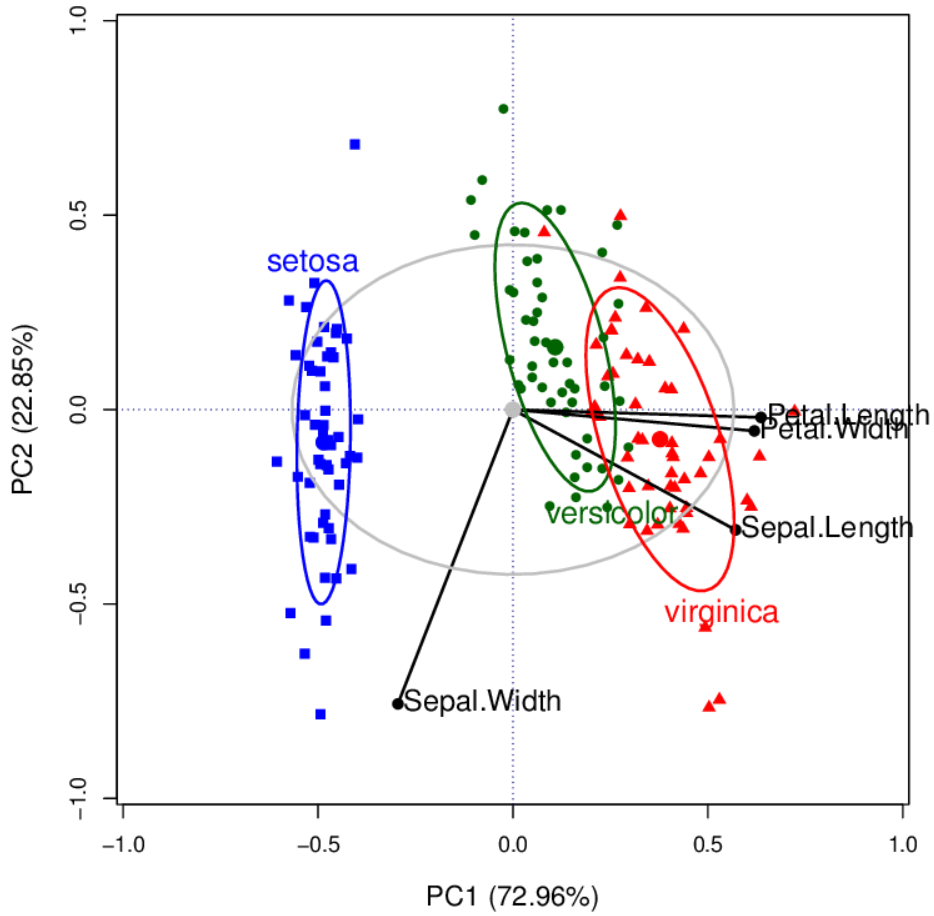
Análisis por Componentes Principales (PCA)





Oscar Centeno Mora

Índice

1 Introducción

2 Componentes Principales

3 Varianza explicada

4 Proyecciones

5 Aplicación en R

6 Otros elementos de análisis

Índice

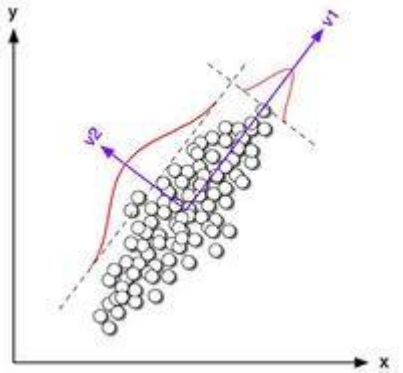
7 Conclusión

Índice

1 Introducción

Introducción

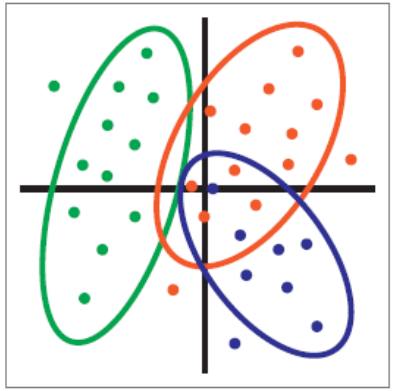
- El análisis por componente principales hace parte de un grupo de análisis descriptivos multidimensionales llamados métodos factoriales.



- Creados en los años 30 pero desarrollados en los años 60, se mejoran los aspectos geométricos y las representaciones gráficas del PCA.

- Son métodos descriptivos, no se apoyan en modelos probabilísticos sino más bien de un modelo geométrico para mejorar la representación multidimencional.

- A partir de una matriz rectangular de datos con *p* variables cuantitativas y *n* unidades, el análisis por componentes principales propone diversas representación geométricas para el entendimiento de los individuos y las variables.



- Lo que se busca es ver si existe una estructura, no conocida a priori, para el conjunto de casos y variables, y así mejorar la interpretación.

Introducción

- Se busca reconocer grupos de unidades con diferencias o similitudes, y así brindar una explicación a la data en el contexto que le compete.

- Para las variables buscamos ver cuáles están muy correlacionadas entre ellas, y, de forma contraria, cuáles no están correlacionadas con las otras. Para los individuos, queremos ver aglomeraciones de estos y ver dónde se distribuyen en un plano. Finalmente podemos ver la conjunción de variables e individuos.

- En la visualización de los individuos o unidades, se utiliza el concepto de las distancias entre los individuos.

- En la visualización de las variables, estas se llevan a cabo en función de sus correlaciones.

Introducción

- En el análisis por PCA, en la proyección de individuos y variables, se deben de tomar en cuenta las medidas de calidad de las representaciones: criterio global y criterios individuales.

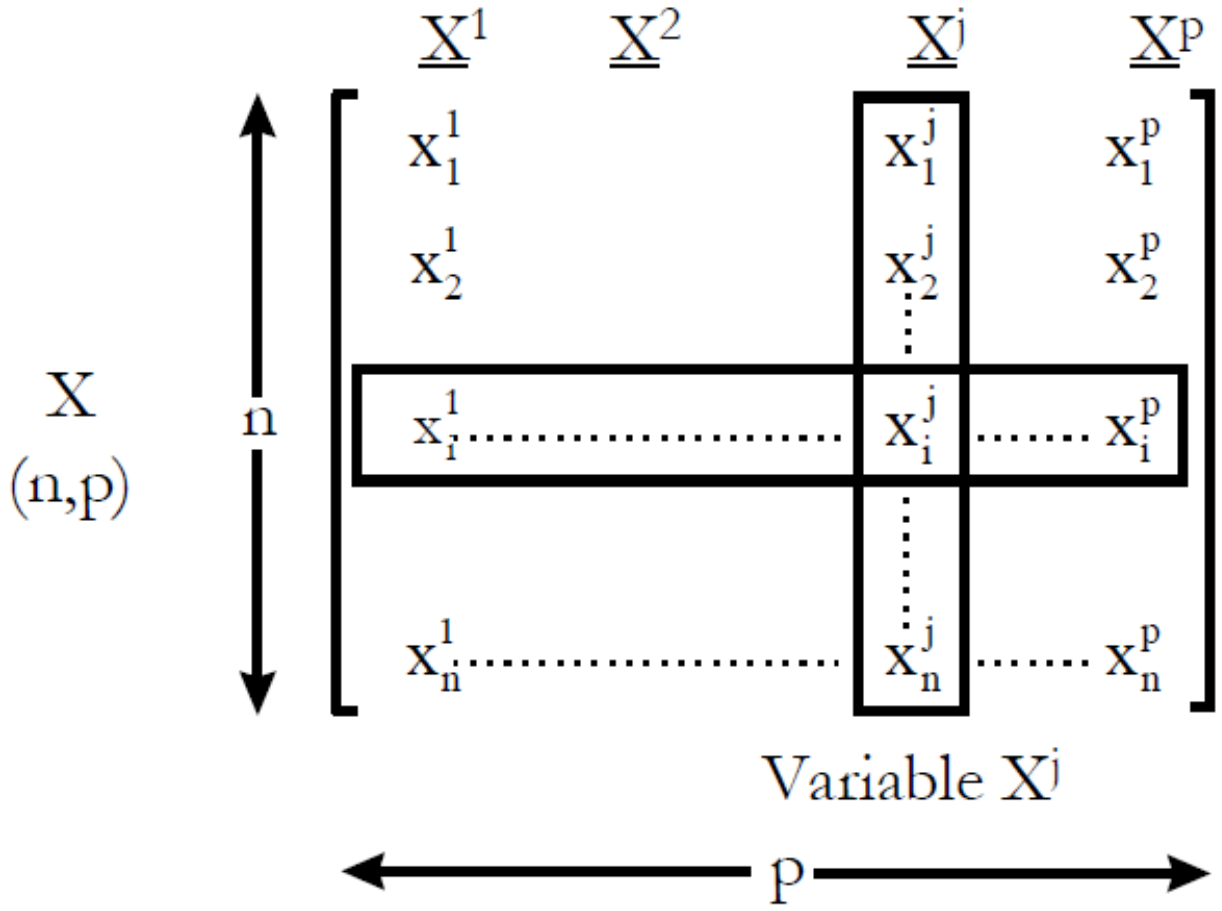
- A veces se recomienda nombrar a los nuevos ejes, y de ahí explicar la posición de los individuos. El presente curso no profundiza sobre este aspecto subjetivo de nombramiento.

- Después de la explicación del método, no se debe olvidar de donde provienen los datos utilizados, lo que representan y el significado para el contexto en causa. La interpretación de la data es lo más importante.

- Como todo método descriptivo, llevar a cabo un PCA no es un fin en sí. El PCA sirve para conocer mejor los datos, detectar valores sospechosos, y ayuda a formular hipótesis que se deben estudiar mediante modelos predictivos.

Introducción

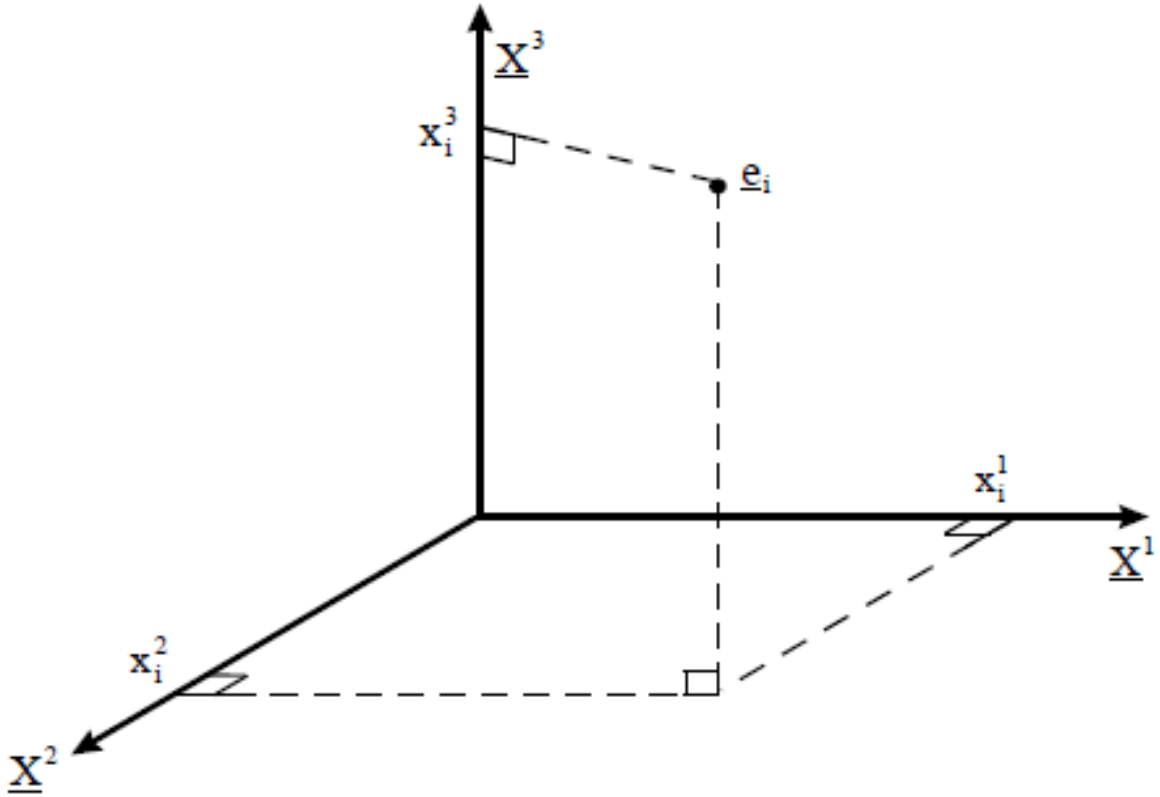
Sea una matriz de datos corresponde a *p* variables cuantitativas y *n* individuos.



INDIVIDUO = elemento del espacio R^p
VARIABLE = elemento del espacio R^n

Introducción

- Sea una representación de los individuos. Cada individuo denominado e_i, se puede asociar un punto en el eje R^p, o espacio de los individuos.
- Cada variable de la matriz X es asociada a un eje de R^p.

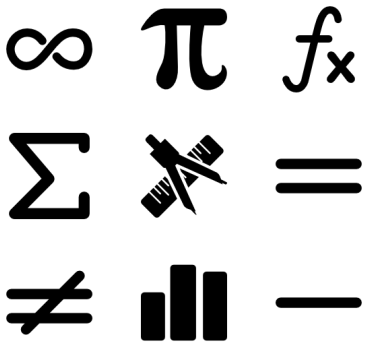


Imposible visualizar a partir de p>3.

Introducción

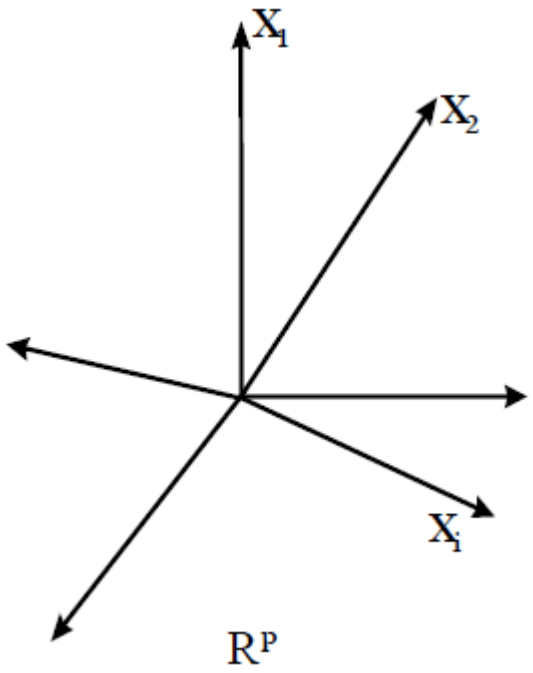
- Se busca representar a los *n* individuos o casos, para un sub espacio vectorial F_k de R_p de dimensión *k* (*k* pequeño como 2, 3,…; por ejemplo un plano).

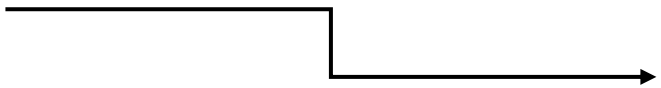
- En otras palabras, se busca definir *k* nuevas variables que son combinaciones lineares de las *p* variables iniciales, que en su conjunto tratarán de optimizar la perdida de información dado el proceso de reducción de variables.

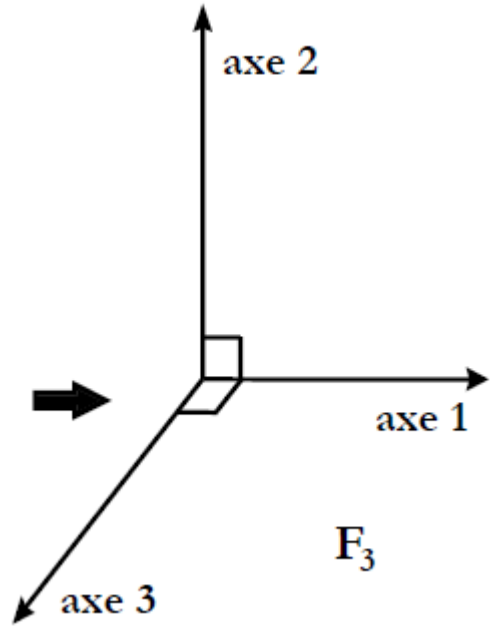


- Las nuevas variables se denominarán ***componentes principales***, los ejes serán llamados ***ejes principales***, y las formas lineales asociadas se llamarán ***factores principales***.

Introducción







Espacio vectorial inicial

Nuevo espacio vectorial con los ejes principales

Introducción

¿Cuántas dimensiones trabajaremos?

Índice

1 Introducción

2 Componentes Principales

Índice

1 Introducción

2 Componentes Principales

- Eigenvectores y eigenvalues
- Interpretación geométrica de los CP
- Cálculo de los componentes principale
- Escalado de las variables
- Reproducibilidad de los componentes
- Influencia de outliers

Componentes principales: eigenvectores y eigenvalues

Eigenvectores

Los *eigenvectors* son un caso particular de multiplicación entre una matriz y un vector. Obsérvese la siguiente multiplicación:

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} x \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4\, x \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

El vector resultante de la multiplicación es un múltiplo entero del vector original.

Los *eigenvectors* de una matriz son todos aquellos vectores que, al multiplicarlos por dicha matriz, resultan en el mismo vector o en un múltiplo entero del mismo.
Los *eigenvectors* tienen una serie de propiedades matemáticas específicas:

- Los *eigenvectors* solo existen para matrices cuadradas y no para todas. En el caso de que una matriz *n* x *n* tenga *eigenvectors*, el número de ellos es *n*.

Componentes principales: eigenvectores y eigenvalues

- Si se escala un *eigenvector* antes de multiplicarlo por la matriz, se obtiene un múltiplo del mismo *eigenvector*. Esto se debe a que si se escala un vector multiplicándolo por cierta cantidad, lo único que se consigue es cambiar su longitud pero la dirección es la misma.
- Todos los *eigenvectors* de una matriz son perpendiculares (ortogonales) entre ellos, independientemente de las dimensiones que tengan.

Dada la propiedad de que multiplicar un *eigenvector* solo cambia su longitud pero no su naturaleza de *eigenvector*, es frecuente escalarlos de tal forma que su longitud sea 1. De este modo se consigue que todos ellos estén estandarizados. A continuación se muestra un ejemplo:

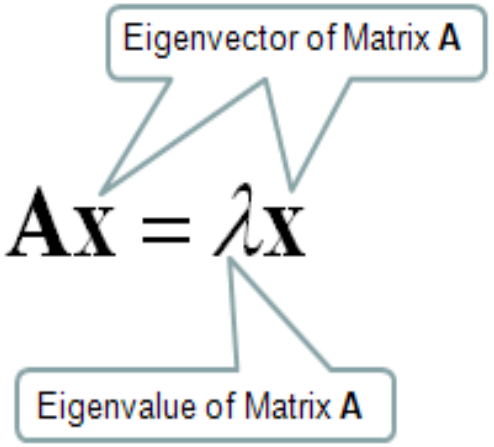
El eigenvector $\binom{3}{2}$ tiene una longitud de $\sqrt{3^2 + 2^2} = \sqrt{13}$. Si se divide cada dimensión entre la longitud del vector, se obtiene el *eigenvector* estandarizado con longitud 1.

$$\binom{3}{2} \div \sqrt{13} = \begin{pmatrix} 3/\sqrt{13} \\ 2/\sqrt{13} \end{pmatrix}$$

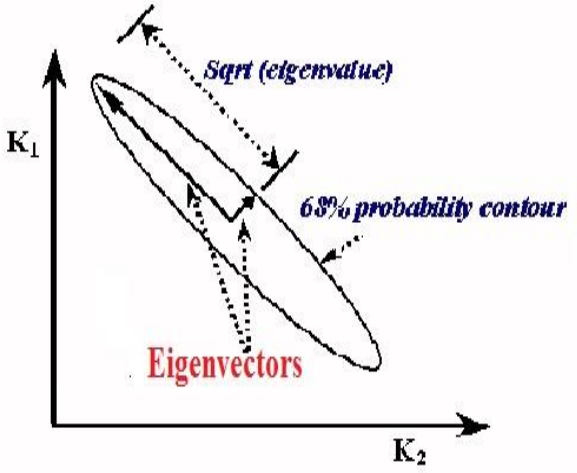
Componentes principales: eigenvectores y eigenvalues

Eigenvalue

Cuando se multiplica una matriz por alguno de sus *eigenvectors* se obtiene un múltiplo del vector original, es decir, el resultado es ese mismo vector multiplicado por un número. Al valor por el que se multiplica el *eigenvector* resultante se le conoce como *eigenvalue*. A todo *eigenvector* le corresponde un *eigenvalue* y viceversa.



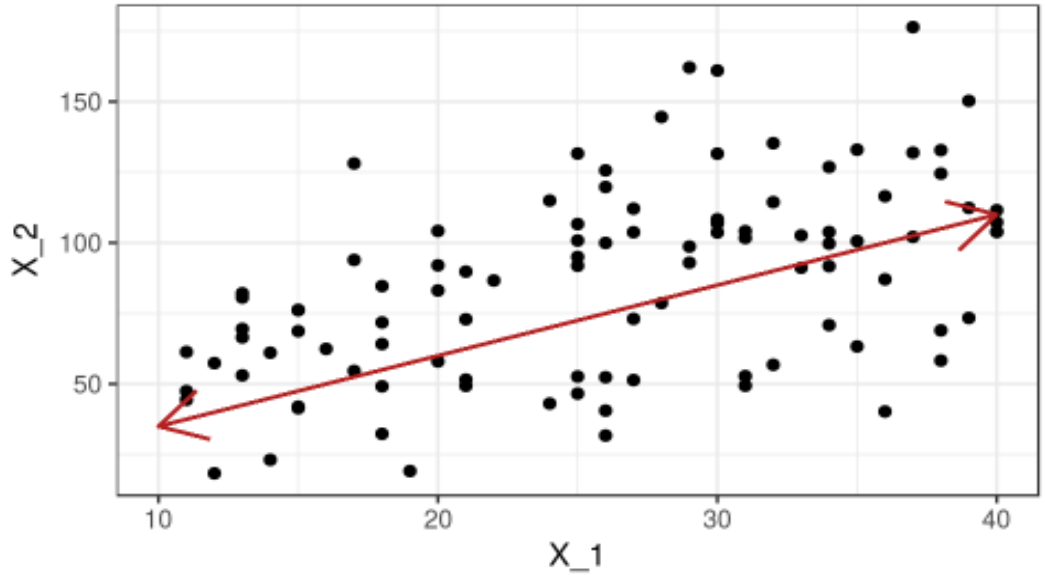
En el método PCA, cada una de las componentes se corresponde con un *eigenvector*, y el orden de componente se establece por orden decreciente de *eigenvalue*. Así pues, el primer componente es el *eigenvector* con el *eigenvalue* asociado más alto.



Componentes principales: interpretación geométrica

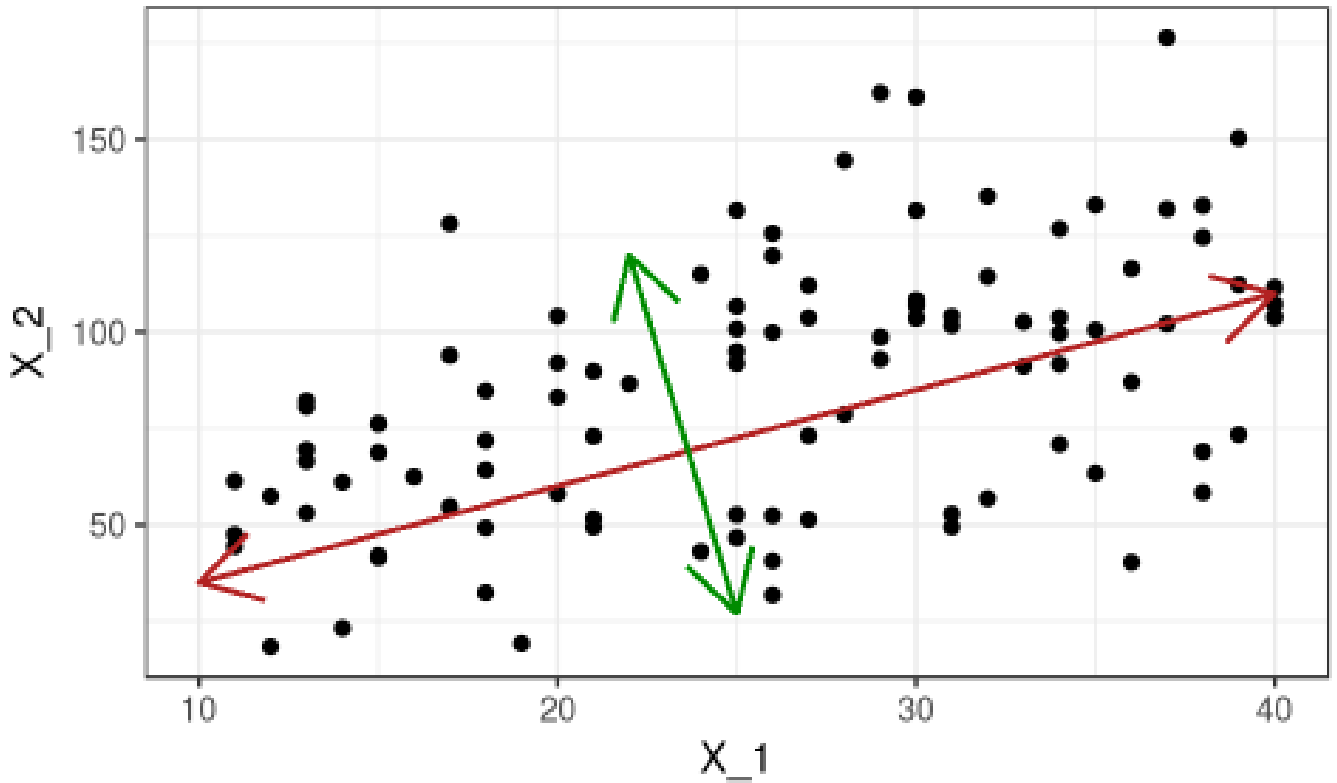
Interpretación geométrica de los componentes principales

Una forma intuitiva de entender el proceso de PCA consiste en interpretar las componentes principales desde un punto de vista geométrico. Supóngase un conjunto de observaciones para las que se dispone de dos variables ($\boldsymbol{X_1}, \boldsymbol{X_2}$). El vector que define en el primer componente principal ($\boldsymbol{Z_1}$) sigue la dirección en la que las observaciones varían más (linea roja). La proyección de cada observación sobre esa dirección equivale al valor del primer componente para dicha observación (*principal component scores*, z_{i1}).



Componentes principales: interpretación geométrica

El segundo componente ($\boldsymbol{Z_2}$) sigue la segunda dirección en la que los datos muestran mayor varianza y que no está correlacionada con la primera componente. La condición de no correlación entre componentes principales equivale a decir que sus direcciones son perpendiculares/ortogonales. Los componentes son ortogonales entre ellos en el PCA.



Componentes principales: cálculo

Cada componente principal (Z_i) se obtiene por combinación lineal de las variables. Se puede entender como nuevas variables obtenidas al combinar de una determinada forma las variables originales. El primer componente principal de un grupo de variables ($X_1, X_2, \dots, X_p$) es la combinación lineal normalizada de dichas variables que tiene mayor variancia:

$$Z_1 = \emptyset_{11}X_1 + \emptyset_{21}X_2 + \cdots + \emptyset_{p1}X_p$$

Que la combinación lineal sea normalizada implica que:

$$\sum_{j=1}^{p} \emptyset_{j1}^2 = 1$$

Los términos $\emptyset_{11}, \dots, \emptyset_{p1}$ reciben el nombre de cargas o loadings y son los que definen a la componente. $\emptyset_{11}$es el *loading* de la variable X_1 del primera componente principal. Los *loadings* pueden interpretarse como el *peso* / *importancia* / *ponderación* que tiene cada variable en cada componente y, por lo tanto, ayudan a conocer que tipo de información recoge cada una de las componentes.

Componentes principales: cálculo

Dado un set de datos X con *n* observaciones y *p* variables, el proceso a seguir para calcular la primera componente principal es:

- Centralización de las variables: se resta a cada valor la media de la variable a la que pertenece. Con esto se consigue que todas las variables tengan media cero.
- Se resuelve un problema de optimización para encontrar el valor de los *loadings* con los que se maximiza la varianza. Una forma de resolver esta optimización es mediante el cálculo de *eigenvector-eigenvalue* de la matriz de covarianzas.

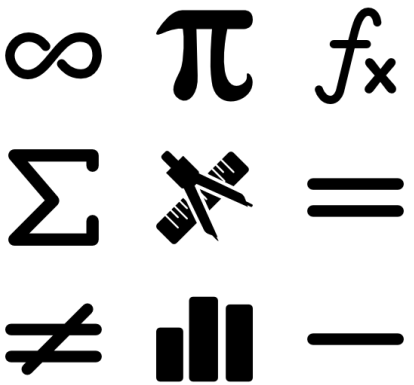
Una vez calculada la primera componente (Z_1) se calcula el segundo (Z_2) repitiendo el mismo proceso, pero añadiendo la condición de que la combinación lineal no pude estar correlacionada con el primer componente. Esto equivale a decir que Z_1 y Z_2 tienen que ser perpendiculares u ortogonales. El proceso se repite de forma iterativa hasta calcular todos las posibles componentes $(\min(n-1, p))$ o hasta que se decida detener el proceso. El orden de importancia de los componentes viene dado por la magnitud del ***eigenvalue*** asociado a cada ***eigenvector***.

Componentes principales: escalado de las variables

- Escalado de las variables es estandarizar las variables. ¿Sabemos a qué nos referimos con esto?

- El proceso de PCA identifica aquellas direcciones en las que la varianza es mayor.

- Como la varianza de una variable se mide en su misma escala elevada al cuadrado, si antes de calcular los componentes no se estandarizan todas las variables para que tengan media 0 y desviación estándar 1, aquellas variables cuya escala sea mayor dominarán al resto. De ahí que sea recomendable estandarizar siempre los datos.



Componentes principales: reproducibilidad de los componentes

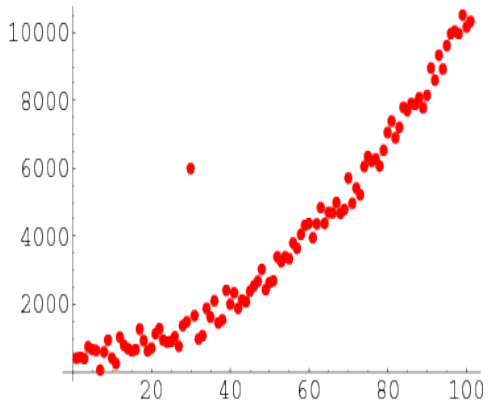
- El proceso de PCA genera siempre los mismos componentes principales independientemente del software utilizado, es decir, el valor de los *loadings* resultantes es el mismo. La única diferencia que puede darse es que el signo de todos los *loadings* esté invertido.

- Esto es así porque el vector de *loadings* determina la dirección del componente, y dicha dirección es la misma independientemente del signo (el componente sigue una línea que se extiende en ambas direcciones).

- Del mismo modo, el valor específico de los componentes obtenidos para cada observación (**PCS**: *principal component scores*) es siempre el mismo, a excepción del signo.

Componentes principales: influencia por los outliers

- Al trabajar con varianzas, el método PCA es altamente sensible a los *outliers*, por lo que es altamente recomendable analizar si los hay... La detección de valores atípicos con respecto a una determinada dimensión es algo relativamente sencillo de hacer mediante comprobaciones gráficas.



- Sin embargo, cuando se trata con múltiples dimensiones el proceso se complica. Por ejemplo, considérese un hombre que mide 2 metros y pesa 50 kg. Ninguno de los dos valores es atípico de forma individual, pero en conjunto se trataría de un caso muy excepcional.

- La distancia de *Mahalanobis* es una medida de distancia entre un punto y la media que se ajusta en función de la correlación entre dimensiones; permite encontrar potenciales *outliers* en distribuciones multivariantes.

Índice

1 Introducción

2 Componentes Principales

3 Varianza explicada

- Proporción de la varianza explicada
- Óptimo de componentes princilaes

Varianza explicada: porción de la varianza del PCA

Una de las preguntas más frecuentes que surge tras realizar un PCA es: ¿Cuánta información presente en el set de datos original se pierde al proyectar las observaciones en un espacio de menor dimensión? o lo que es lo mismo ¿Cuanta información es capaz de capturar cada una de las componentes principales obtenidas?

Para contestar a estas preguntas se recurre a la proporción de varianza explicada por cada componente principal.

Asumiendo que las variables se han normalizado para tener media cero, la varianza total presente en el set de datos se define como:

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j-1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$$

Varianza explicada: porción de la varianza del PCA

Y la varianza explicada para el m-ésimo componente estaría dado por:

$$\frac{1}{n}\sum_{i=1}^{n} z_{im}^2 = \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{p} \emptyset_{jm} x_{ij})^2$$

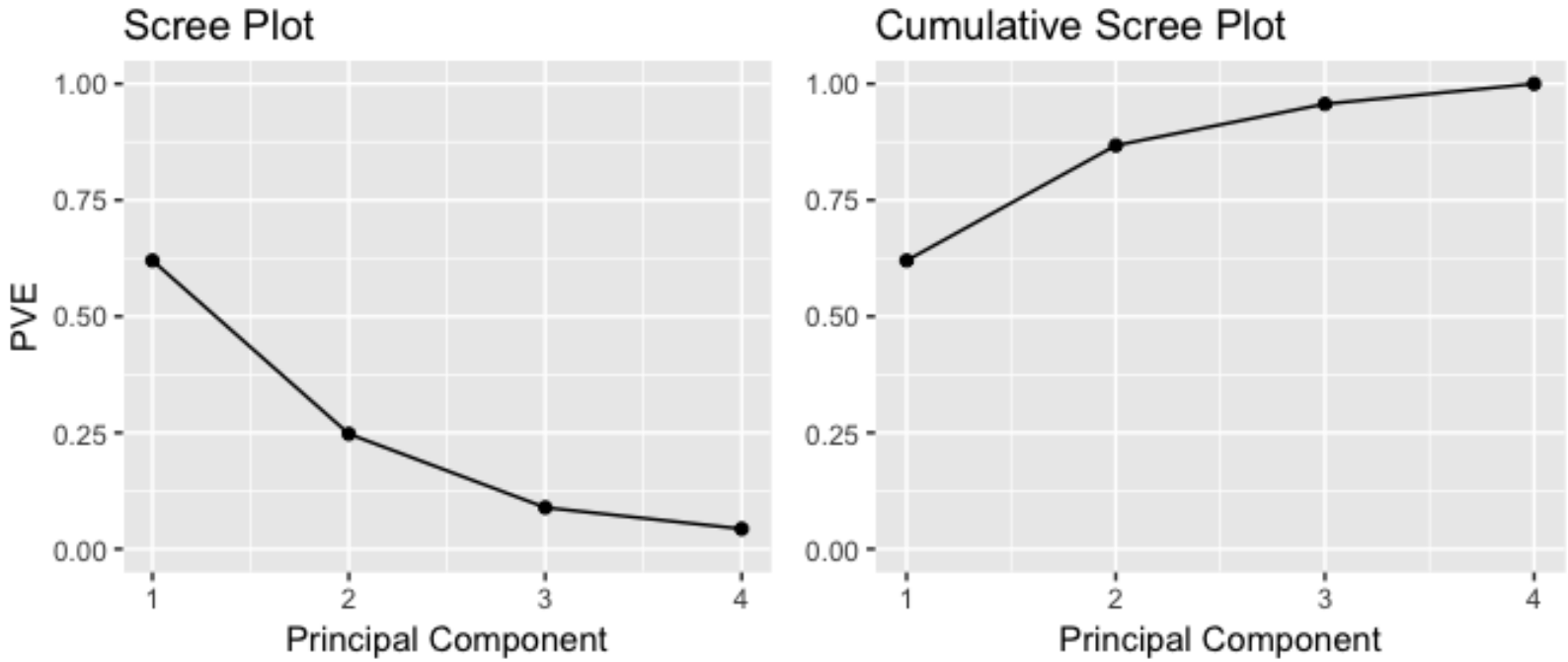
Por lo tanto, la proporción de varianza explicada para el m-ésimo componente se expresa:

$$\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{p} \emptyset_{jm} x_{ij}\right)^2}{\sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij}^2}$$

Tanto la proporción de varianza explicada como la proporción de varianza explicada acumulada son dos valores de gran utilidad a la hora de decidir el número de componentes principales a utilizar en los análisis posteriores. Si se calculan todas las componentes principales de un set de datos, entonces, aunque transformada, se está almacenando toda la información presente en los datos originales. El sumatorio de la proporción de varianza explicada acumulada de todas las componentes es siempre 1.

Varianza explicada: óptimo número de PC

Por lo general, dada una matriz de datos de dimensiones $n * p$, el número de componentes principales que se pueden calcular es como máximo de $n - 1$ o p (el menor de los dos valores es el limitante). Sin embargo, siendo el objetivo del PCA reducir la dimensionalidad, suelen ser de interés utilizar el número mínimo de componentes que resultan suficientes para explicar los datos. No existe una respuesta o método único que permita identificar cual es el número óptimo de componentes principales a utilizar. Una forma de proceder muy extendida consiste en evaluar la proporción de varianza explicada acumulada y seleccionar el número de componentes mínimo a partir del cual el incremento deja de ser sustancial. Se suele utilizar el gráfico de sedimentación (*scree plot*).



Índice

1 Introducción

2 Componentes Principales

3 Varianza explicada

4 Proyecciones

Índice

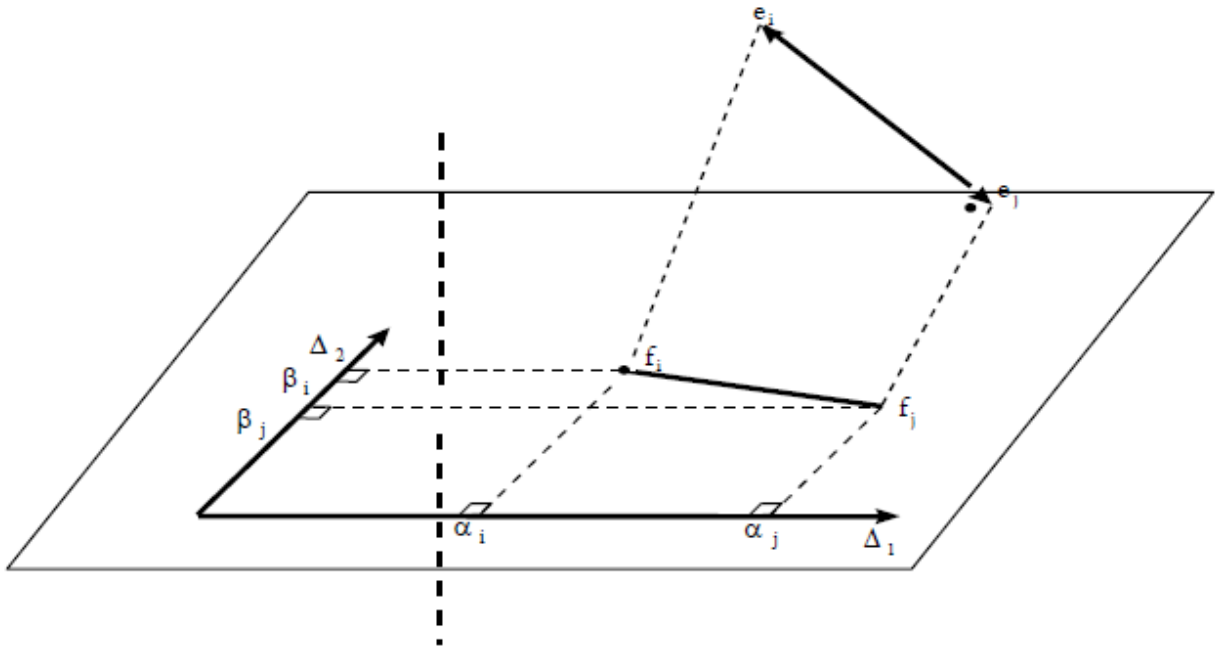
4

Proyecciones

- Individuos
- Variables
- Individuos y variables

Proyecciones: individuos

- En la definición de los ejes principales, F_k deberá ser ajustado lo mejor posible a la nube de los individuos: la suma de cuadrados de las distancias de los individuos a los F_k tiene que ser mínima. F_k es el sub espacio tal que la nube proyectada deba tener una inercia (dispersión) máxima. Los dos puntos anteriores se fundamentan sobre la noción de distancia y proyección ortogonal.



La distancia entre f_i y f_j es inferior o igual a la distancia entre e_i y e_j.

Proyecciones: individuos

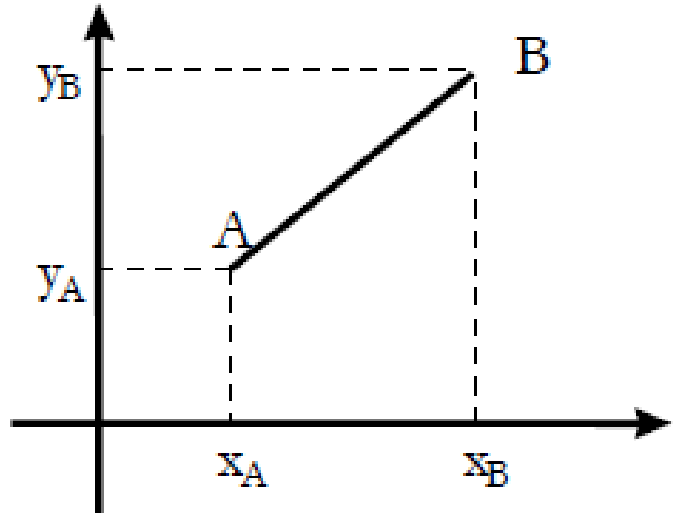
- En el espacio R^p de p dimensiones, se generaliza la siguiente noción: la distancia euclidiana entre dos individuos se expresa como:

$$e_i = \left(x_i^1\ x_i^2\ \ldots\ x_i^p\right) \qquad e_j = \left(x_j^1\ x_j^2\ \ldots\ x_j^p\right)$$

$$d^2\left(e_i, e_j\right) = \left(x_i^1 - x_j^1\right)^2 + \left(x_i^2 - x_j^2\right)^2 + \ \ldots \left(x_i^p - x_j^p\right)^2$$

- Por lo tanto

$$d^2\left(e_i, e_j\right) = \sum_{k=1}^{p}\left(x_i^k - x_j^k\right)^2$$



$$d^2\left(A, B\right) = \left(x_B - x_A\right)^2 + \left(y_B - y_A\right)^2$$

Proyecciones: individuos

- Cuando se tienen variables con diferentes unidades de medida, el problema se revuelve transformando los datos mediante el proceso de la estandarización.

- La observación X_i^k se remplaza por:

$$\frac{X_i^k - \overline{X}^k}{S_k}$$

- Nótese que al realizar la estandarización, suponemos la normalidad de los datos.

Proyecciones: individuos y la inercia total

- Para adecuar la distancia de las unidades de estudio, se utiliza la inercia. Esto es la suma ponderada de las distancias cuadradas de los individuos para el centro de gravedad g.

- La inercia mide la dispersión total de la nube de puntos.

- La inercia se expresa como: $$I_g = \sum_{i=1}^{n} \frac{1}{n} d^2(e_i, \underline{g})$$

O de forma general $$I_g = \sum_{i=1}^{n} p_i \, d^2(e_i, \underline{g})$$

$$\sum_{i=1}^{n} p_i = 1$$

Proyecciones: individuos y la inercia total

- La inercia es por lo tanto igual a la suma de las variancias de las variables estudiadas.

- Si denominamos V la matriz de variancias y covariancias:

$$V = \begin{pmatrix} s_1^2 & s_{12} \cdots\cdots s_{1p} \\ \vdots & s_2^2 \quad \vdots \\ s_{p1} & s_p^2 \end{pmatrix}$$

$$I_g = \sum_{i=1}^{p} s_i^2$$

$$I_g = \mathrm{Tr}(V)$$

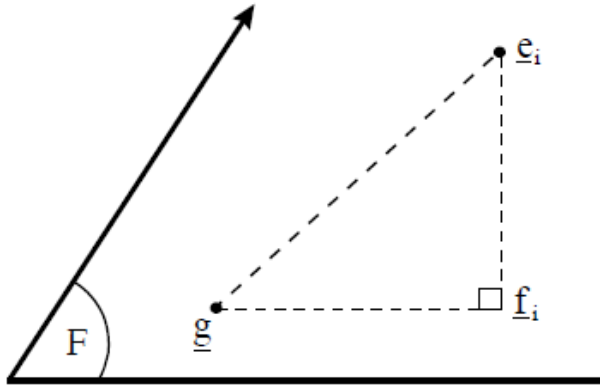
- En el caso donde las variables son estandarizadas, la variancia de cada variable es de 1.

- Por lo tanto, en el caso anterior, la inercia es igual a p (total de variables).

Proyecciones: individuos y la inercia total

- Sea F es un sub espacio de R^P
- fi la proyección ortogonal de ei sobre F se representa como

$$\|\underline{e}_i - \underline{g}\|^2 = \|\underline{e}_i - \underline{f}_i\|^2 + \|\underline{f}_i - \underline{g}\|^2 \qquad \forall i = 1 \ldots n$$



Proyección ortogonal de la nube sobre el sub espacio

Proyecciones: individuos y la inercia total

- Se busca para F que sea mínimo: $\sum_{i=1}^{n} p_i \left\| \underline{e}_i - \underline{f}_i \right\|^2$

- Lo que equivaldría que, según el teorema de Pitágoras a maximizar la siguiente expresión:

$$\sum_{i=1}^{n} p_i \left\| \underline{f}_i - \underline{g} \right\|^2$$

- De lo anterior se obtiene:

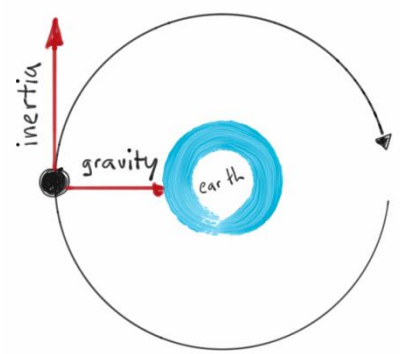
$$\left\| \underline{e}_i - \underline{g} \right\|^2 = \left\| \underline{e}_i - \underline{f}_i \right\|^2 + \left\| \underline{f}_i - \underline{g} \right\|^2 \qquad \forall i = 1 \ldots n$$

- Y por lo tanto:

$$\underbrace{\sum_{i=1}^{n} p_i \left\| \underline{e}_i - \underline{g} \right\|^2}_{\text{Inercia total}} - \underbrace{\sum_{i=1}^{n} p_i \left\| \underline{e}_i - \underline{f}_i \right\|^2}_{\text{Minimizar distancia entre individuos y la proyección}} = \underbrace{\sum_{i=1}^{n} p_i \left\| \underline{f}_i - \underline{g} \right\|^2}_{\text{Maximizar inercia de la nube proyectada}}$$

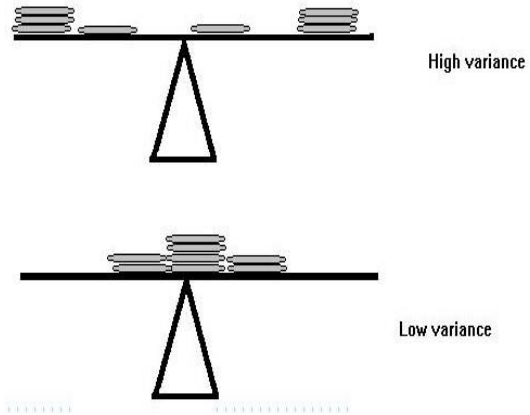
Proyecciones: individuos y la inercia total

- La búsqueda de los ejes con el máximo de inercia equivale a la construcción de nuevas variables con máxima variancia (las variables están asociadas a los nuevos ejes mediante combinaciones lineales).



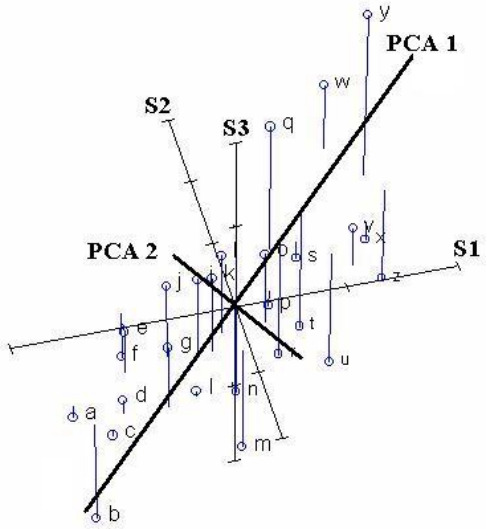
- En otros términos, realizamos un cambio del sistema de referencia en el espacio R^p de manera de colozarse en un nuevo sistema de representación donde el primer eje aporta la mayor cantidad posible de inercia total a la nube, el segundo eje la mayor cantidad de inercia no tomada en cuenta en el primer eje, y así consecutivamente.

- Esta reorganización se apoya sobre la diagonalización de la matriz de variancia y covariancia.



Proyecciones: individuos y la inercia total

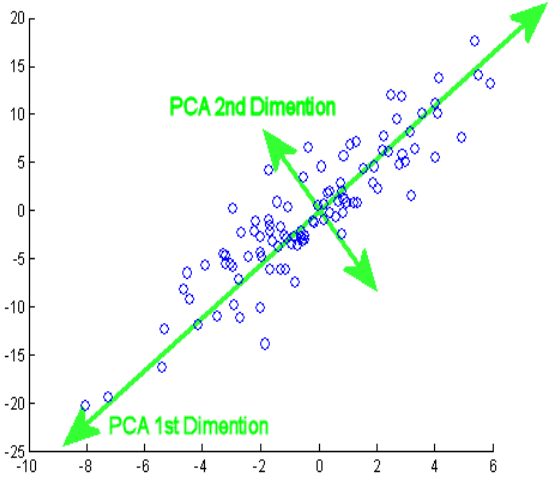
- Sobre los ejes principales, denominamos ejes principales de inercia a los ejes de direccionan los vectores propios de V normados en 1.



- En total hay *p*.

- El primer eje está asociado a la mayor cantidad de valores propios. Lo notamos como u^1.

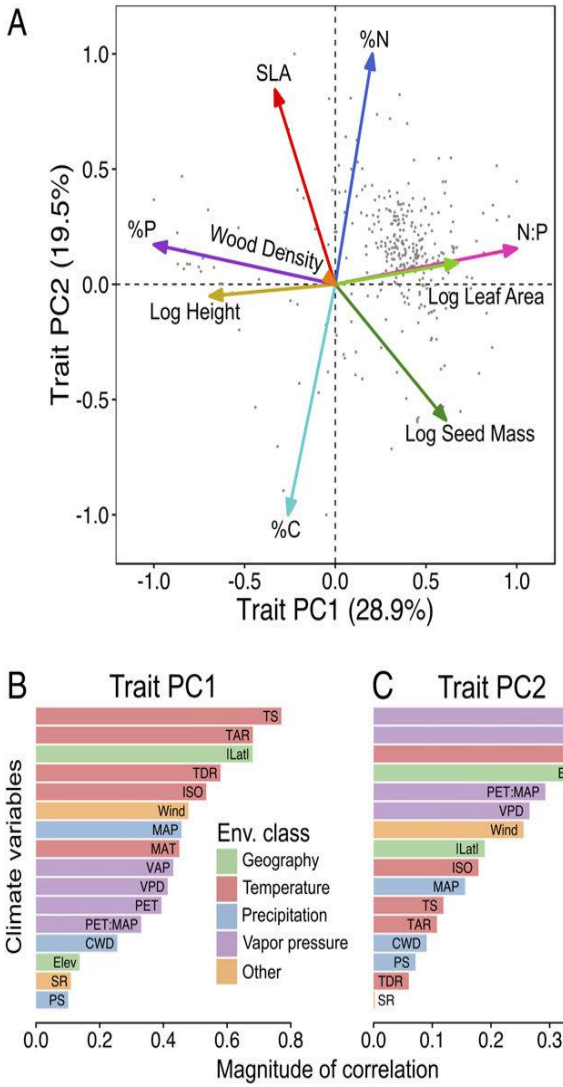
- El segundo eje está asociado a la mayor cantidad de valores propios. Lo notamos como u^2.



Proyecciones: individuos y la inercia total

- A cada eje está asociado una variable llamada componente principal.
- El componente c^1 es el vector que contiene las coordenadas de las proyecciones de los individuos sobre el eje 1.
- El componente c^2 es el vector que contiene las coordenadas de proyecciones de los individuos sobre el eje 2.
- Para obtener dichas coordenadas, cada componente principal es una combinación lineal de las variables iniciales.

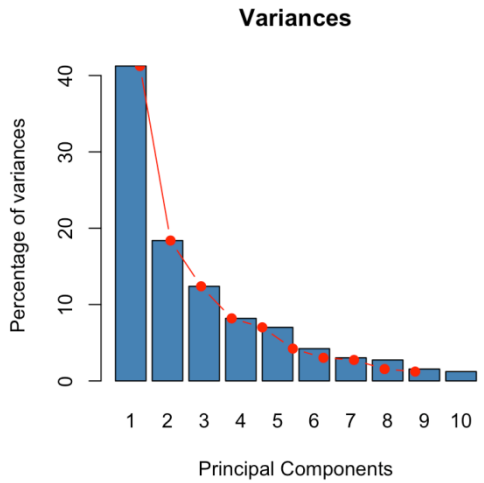
$$\underline{c}^1 = u_1^1 \underline{x}^1 + u_2^1 \underline{x}^2 + \ldots u_p^1 \underline{x}^p$$



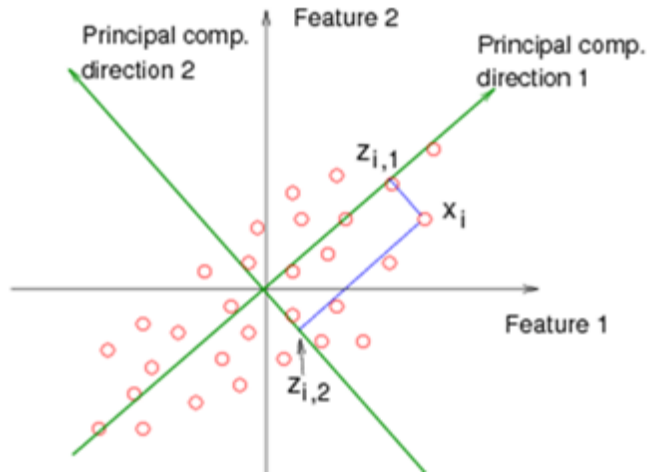
Proyecciones: individuos y la inercia total

1. La variancia de un componente principal es igual a la inercia contribuida al eje principal que le es asociado.

1ère composante	$\mathbf{c^1}$	variance : λ_1
2ème composante	$\mathbf{c^2}$	variance : λ_2
3ème composante	$\mathbf{c^3}$	variance : λ_3



2. Les componentes principales no están correlacionados entre ellos. En efecto, los ejes asociados son ortogonales



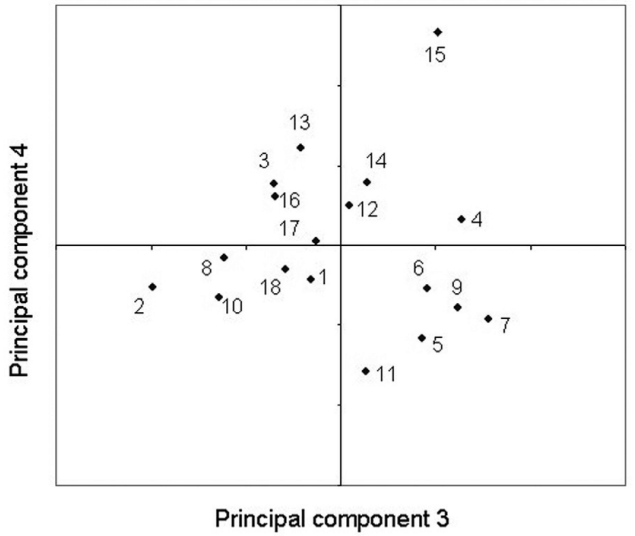
Proyecciones: individuos y la inercia total

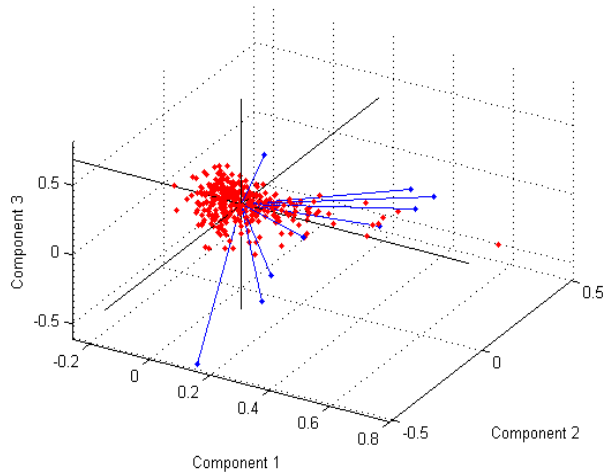
- Los *j* componentes principales brindan la información para ubicar a los n individuos sobre el los *j* ejes principales.

- El vector que permite ubicar a los individuos se presenta como:

$$\underline{c}^j = \begin{pmatrix} c_1^j \\ c_2^j \\ \vdots \\ c_n^j \end{pmatrix}$$

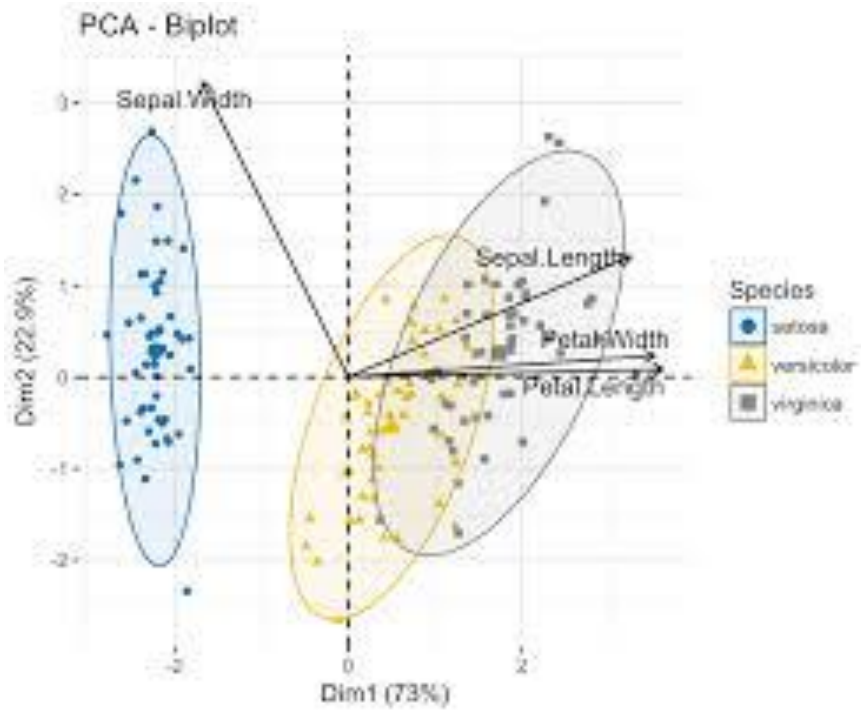
- En la práctica, si se desearía una representación plana de los individuos, lo mejor sería utilizar los dos primeros componentes principales.





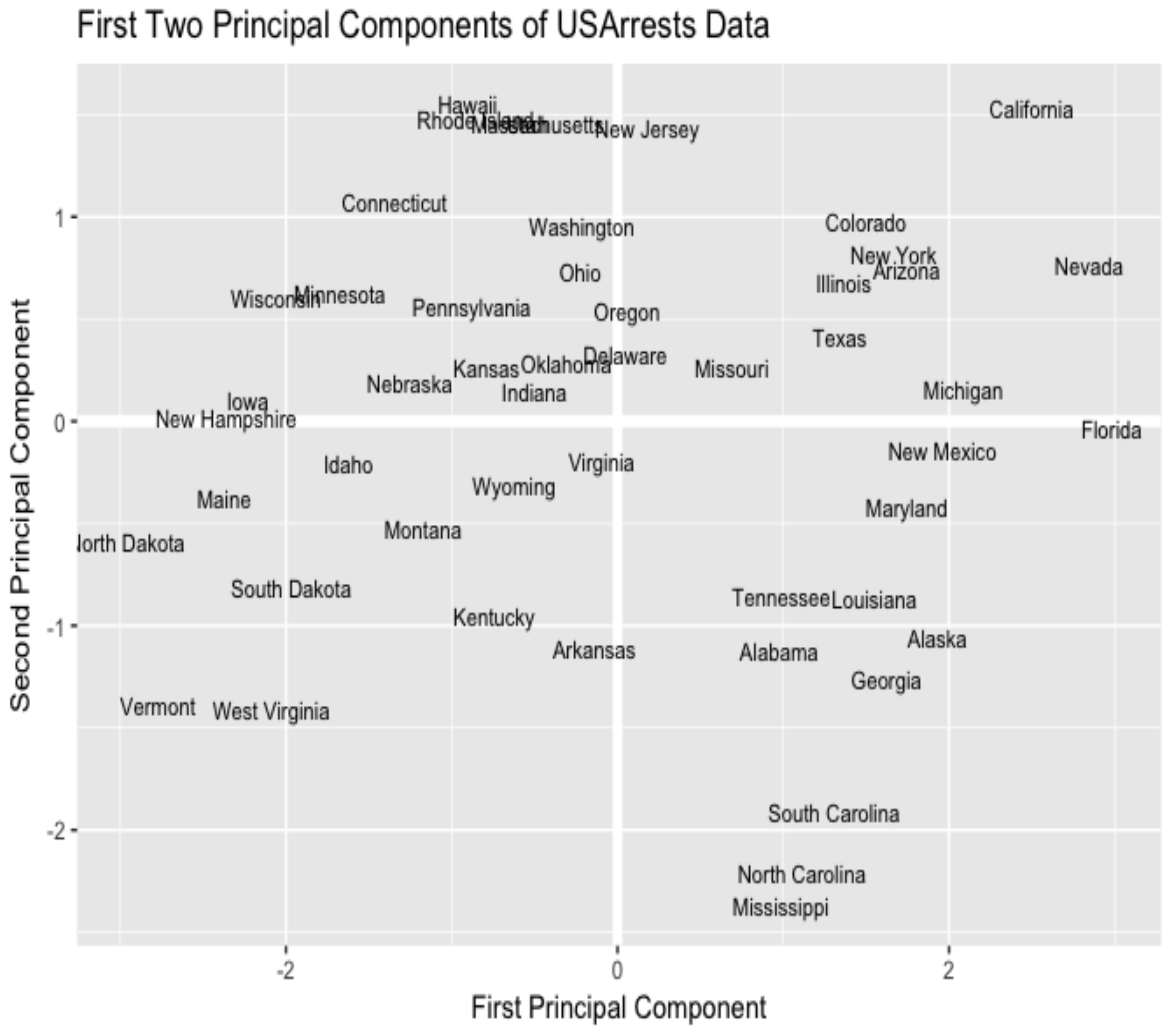
Proyecciones: representación de los individuos

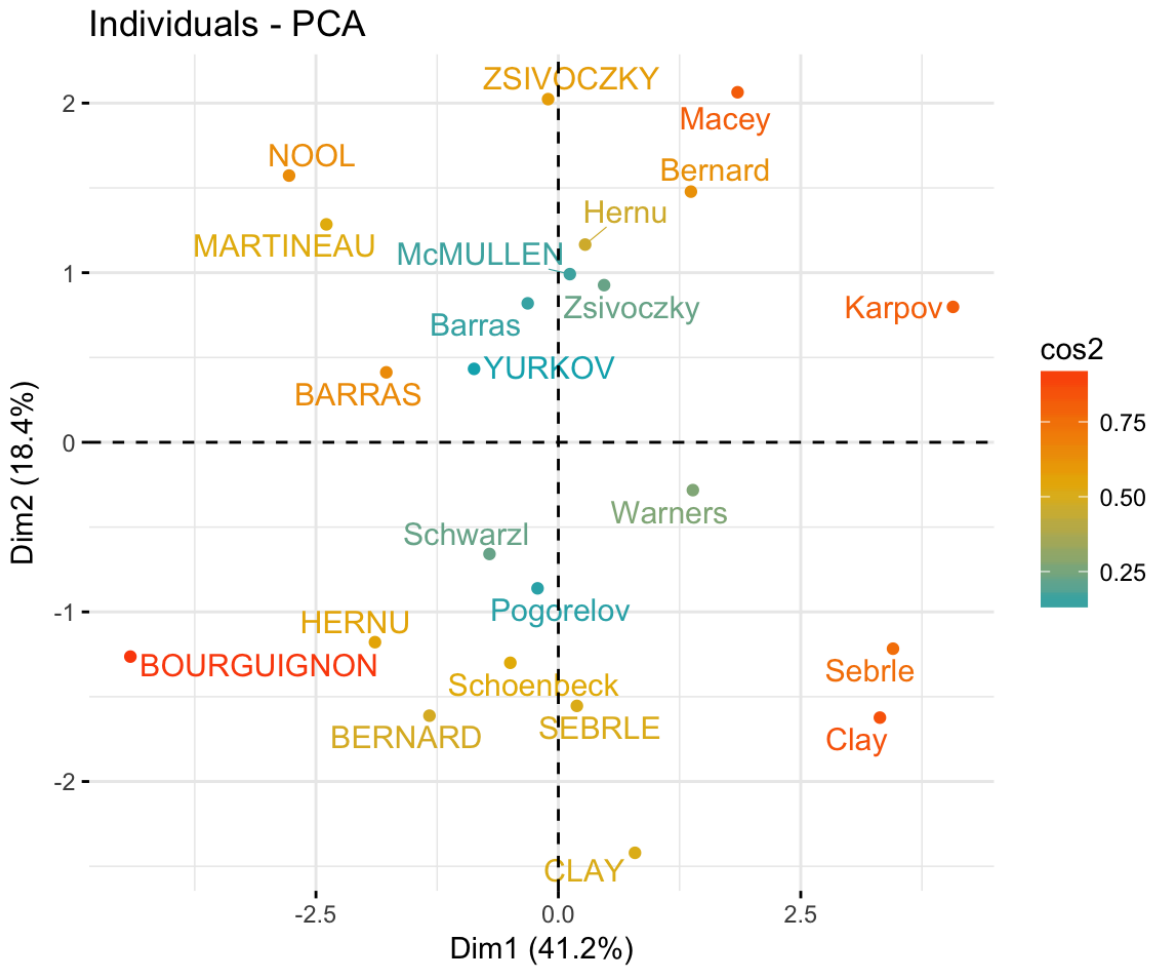
- EL objetivo es ver patrones a aglomeraciones de los casos. Es importante, de ser posible, nombrar al eje principal para mejorar la calidad del análisis.
- A veces, los casos muy aparte del resto de los individuos, son complicados de describir o comprender. Se les suele dar la denominación de valor atípico.



Importante: se debe de tener cuidado con la calidad de la representación de cada individuo.

Proyecciones: representación de los individuos

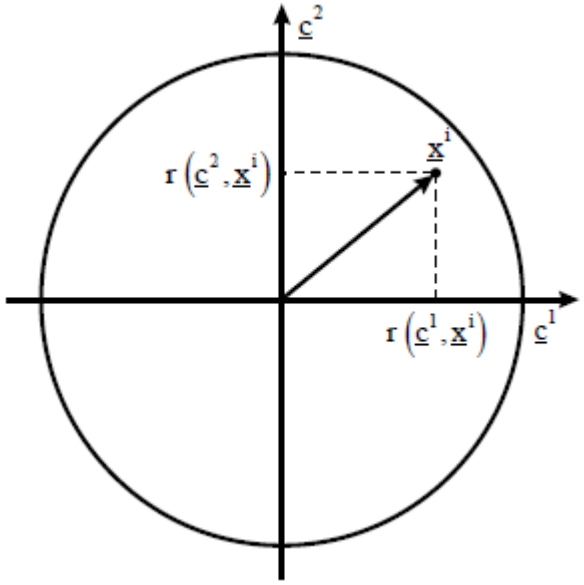




Proyecciones: las variables

- Las *proximidades* entre los componentes principales y las variables iniciales son medidas por la covariancia, y sobre todo con las correlaciones.

- El coeficiente de correlación lineal entre c_j y x_i se denota como $r(c^j,x^i)$



Circulo de correlaciones

El PCA: interpretación de la proximidad entre variables

- Se utiliza un producto escalar entre variables que permiten asociar a los parámetros conocidos: desviación estándar, coeficiente de correlación lineal en las representaciones geométricas (se parte de que las variables están centradas).

$$\left\langle \underline{x}^i, \underline{x}^j \right\rangle = \frac{1}{n} \sum_{k=1}^{n} x_k^i \; x_k^j$$

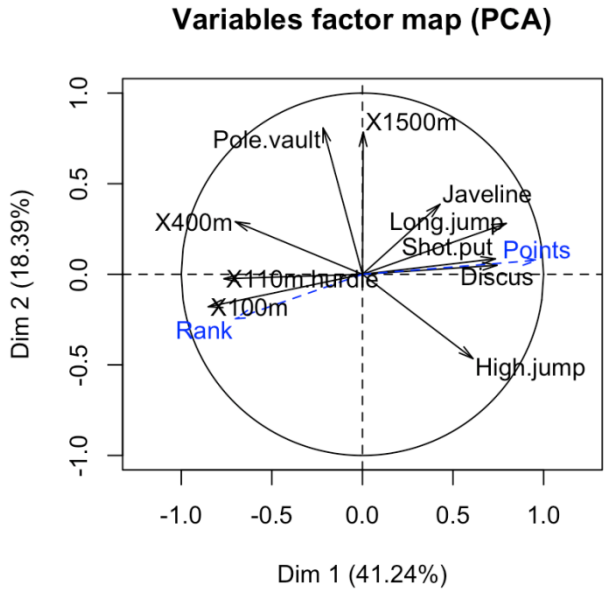
- Las ecuaciones restantes son:

$$\left\langle \underline{x}^i, \underline{x}^j \right\rangle = \mathrm{Cov}\left(\underline{x}^i, \underline{x}^j\right)$$

$$\left\| \underline{x}^i \right\|^2 = s_i^2$$

$$\left\| \underline{x}^i \right\|^2 = \left\langle \underline{x}^i, \underline{x}^i \right\rangle = \frac{1}{n} \sum_{k=1}^{n} \left(x_k^i\right)^2$$

$$\left\| \underline{x}^i \right\| = s_i$$

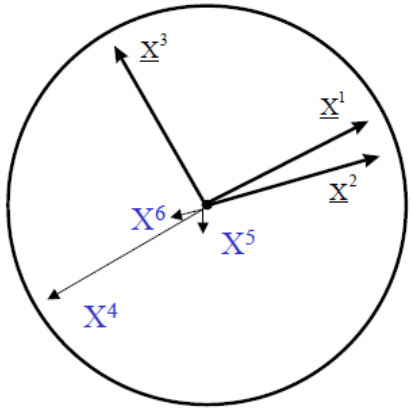


El PCA: coeficiente de correlación lineal

- Para asociar a las variables en el circulo de relación, el coseno del alguno formado por las variables Xi y Xj es el coeficiente de correlación lineal de las dos variables.

$$\cos\left(\widehat{\underline{X}^i, \underline{X}^j}\right) = \frac{\left\langle \underline{x}^i, \underline{x}^j \right\rangle}{\left\| \underline{X}^i \right\| \left\| \underline{X}^j \right\|} = \frac{\mathrm{Cov}\left(\underline{X}^i, \underline{X}^j\right)}{s_i \, s_j} = r\left(\underline{X}^i, \underline{X}^j\right)$$

- Una interpretación según el circulo es la siguiente:

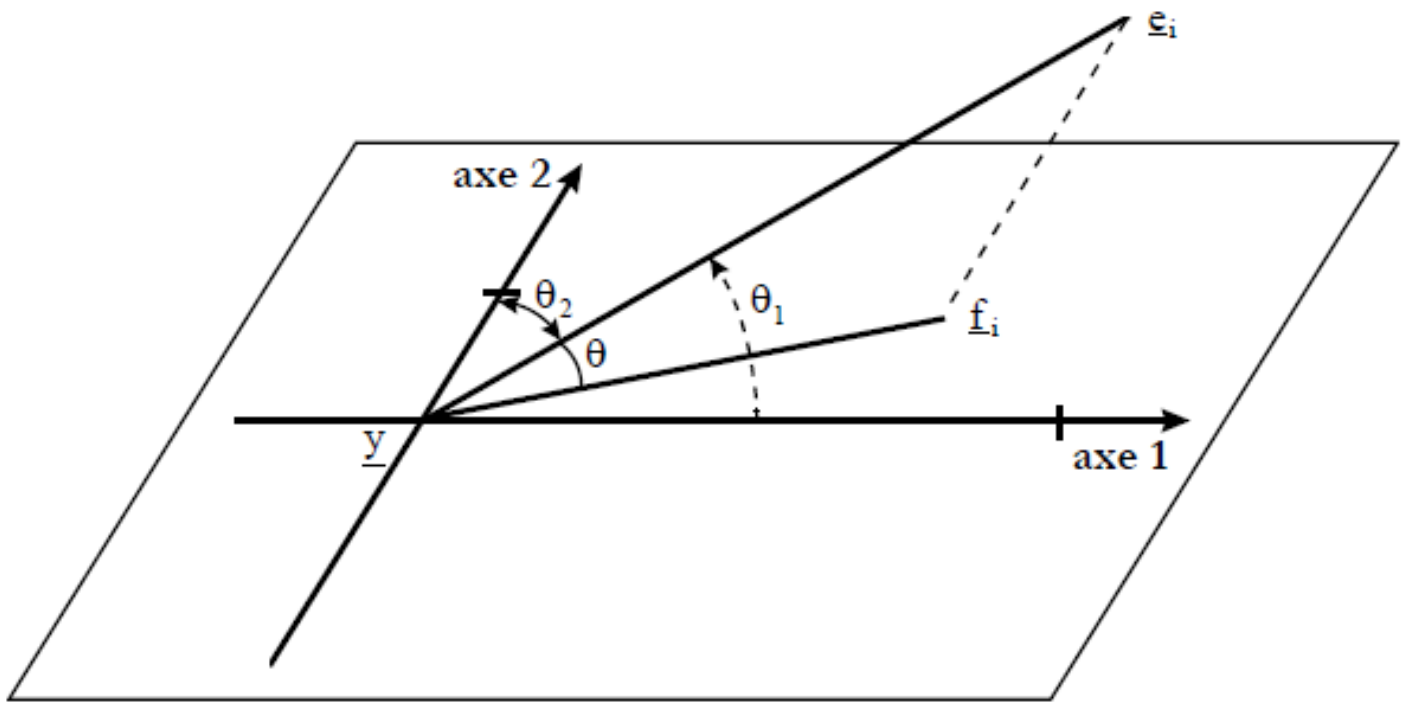


X^1 y X^2 poseen una correlación cercana a 1.

X^1 y X^3 poseen una correlación cercana a 0.

El PCA: criterios individuales

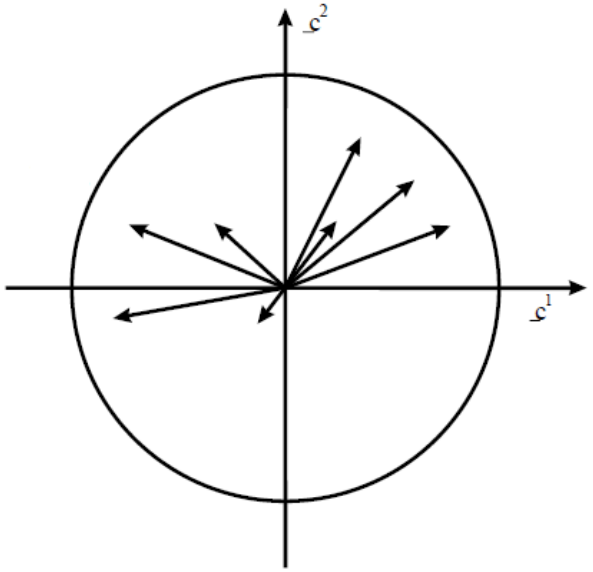
- Se utiliza los cosenos cuadrados:



$$\cos^2\theta = \cos^2\theta_1 + \cos^2\theta_2$$

El PCA: representación de las variables

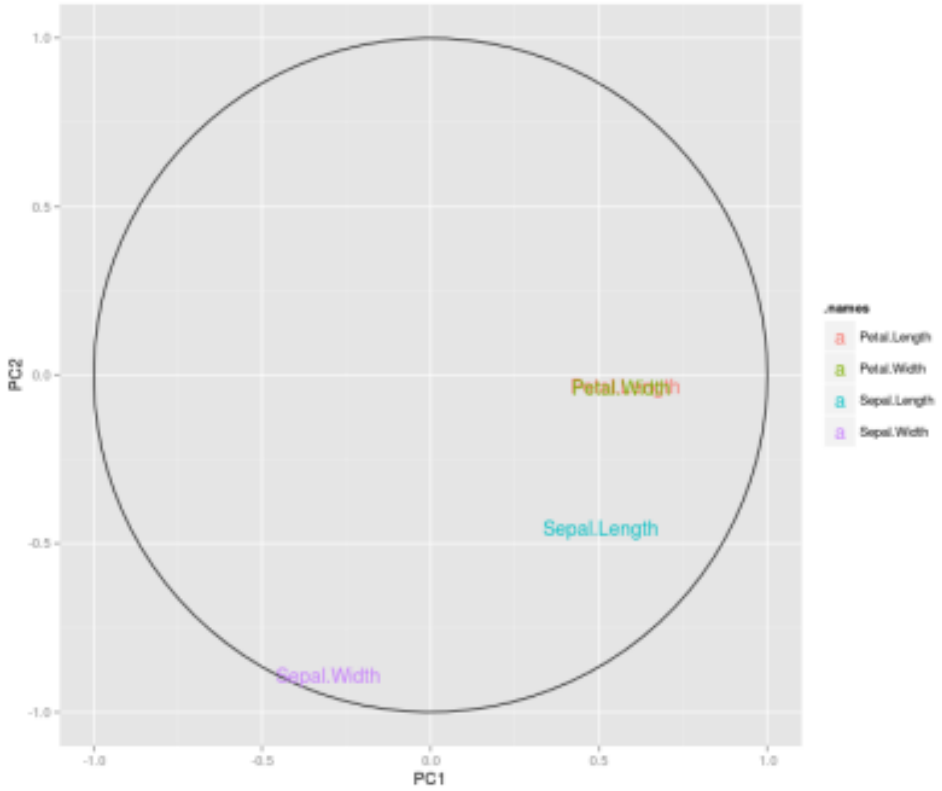
- El círculo de las correlaciones es la proyección de la nube de las variables sobre el plano de los componentes principales. Las variables que están bien representadas son aquellas que están cercanas al círculo. Caso contrarias, aquellas cerca del origen están mal representadas. La correlación puede verse también cómo una medida de calidad.

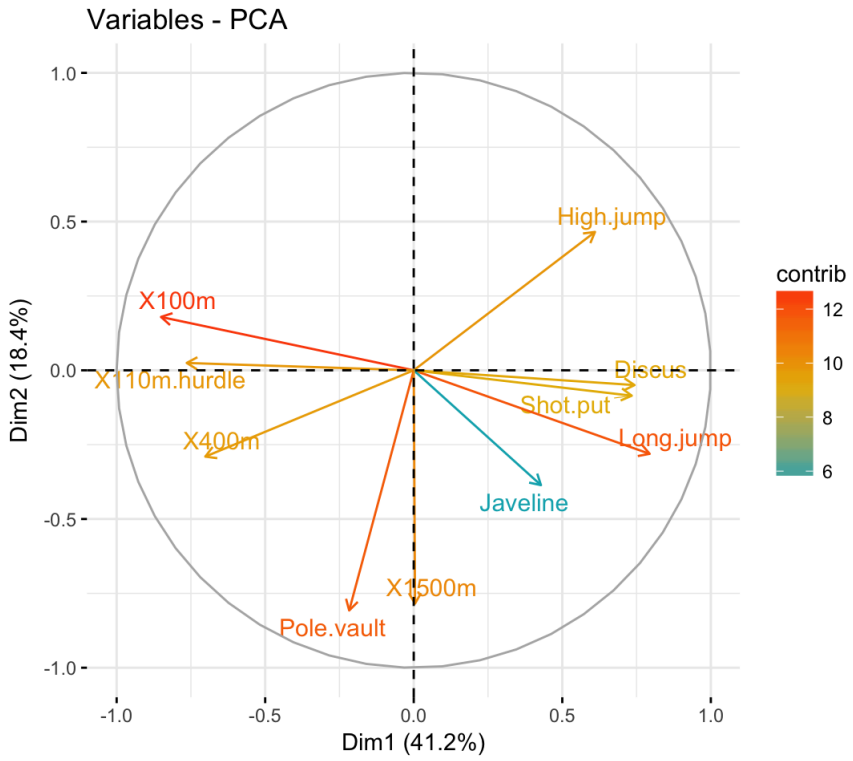


Correlación = coseno

El PCA: representación de las variables

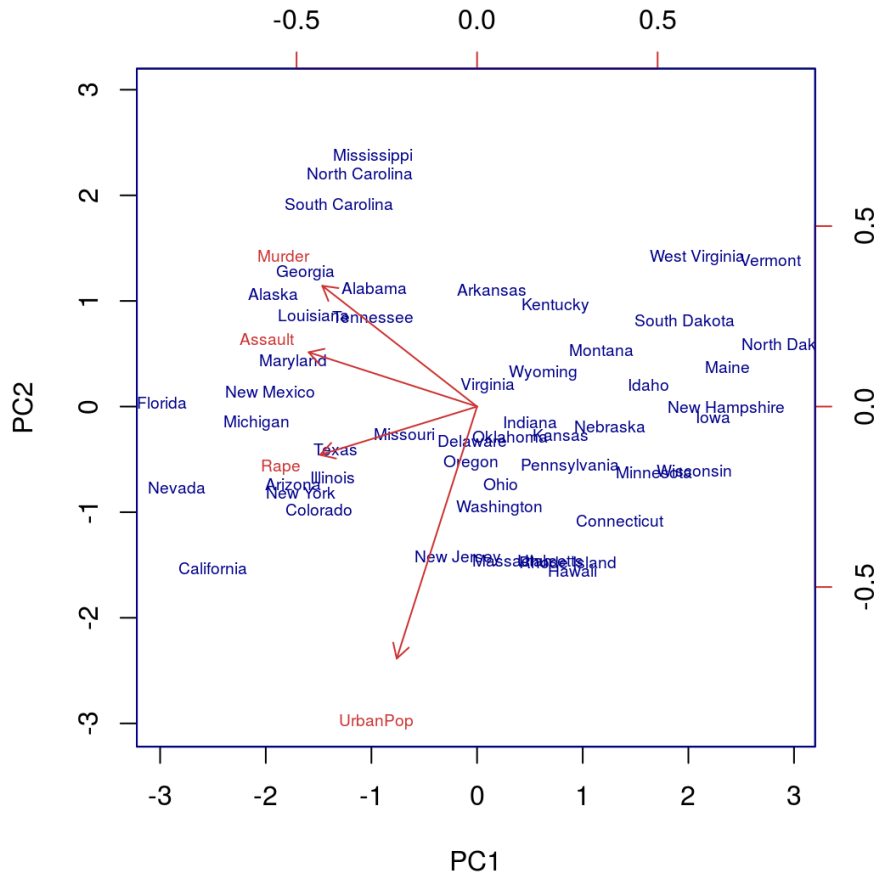
- Los ejemplos se encuentran en el laboratorio y en los siguientes enlaces:

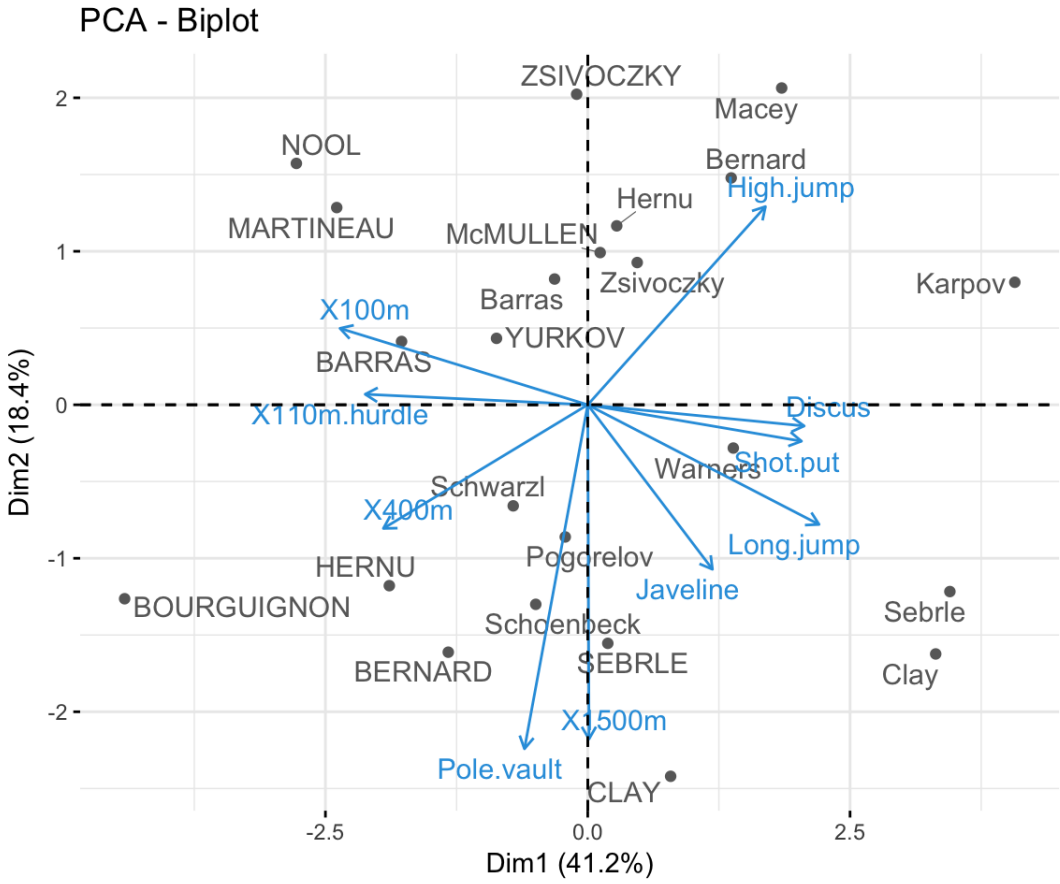




El PCA: representación individuos y variables

Sin entrar en detalles de cálculo, se pueden sobre poner ambas representaciones para conocer una posible relación entre individuos – variables.





Índice

1 Introducción

2 Componentes Principales

3 Varianza explicada

4 Proyecciones

5 Aplicación en R

Índice

5

Aplicación en R

- Librerías y funciones
- Elementos importantes
- Variancia explicada
- Proyección individuos
- Proyección variables
- Proyección indi y variables

Aplicación en R: funciones y librerías

- Para la aplicación del PCA en R, se pueden utilizar diversas librerías y funciones. Se exponen algunos ejemplos:

Las funciones y librerías parar realizar el PCA en R pueden ser:

1. prcomp() (stats)

```
pca1 = prcomp(USArrests, scale. = TRUE)
```

2. princomp() (stats)

```
pca2 = princomp(USArrests, cor = TRUE)
```

3. PCA() (FactoMineR)

```
pca3 = PCA(USArrests, graph = FALSE)
```

4. dudi.pca() (ade4)

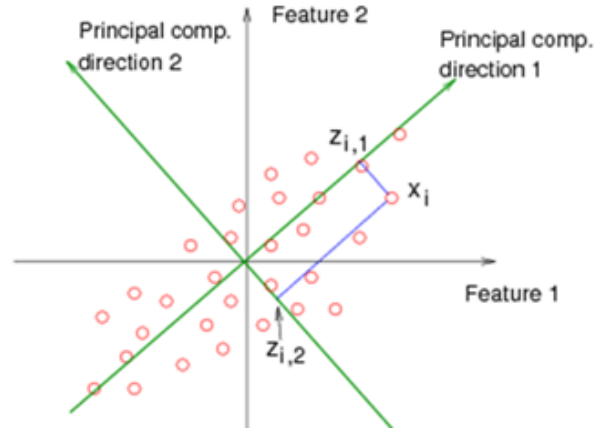
```
pca4 = dudi.pca(USArrests, nf = 5, scannf = FALSE)
```

5. acp() (amap)

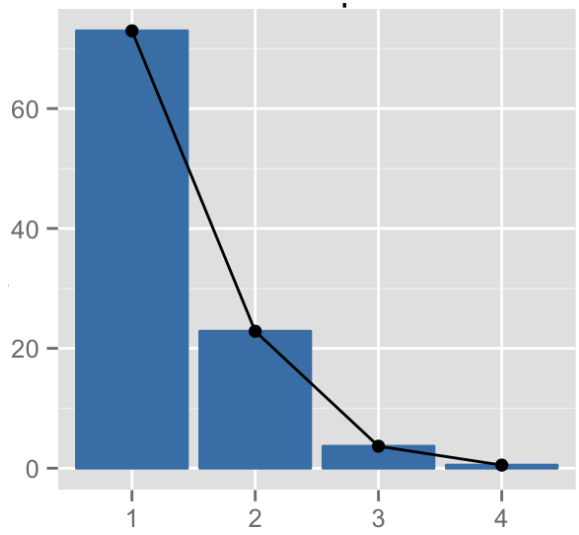
```
pca5 = acp(USArrests)
```

Aplicación en R: elementos de extración

- En el análisis por PCA, estos son los 5 elementos que nos va a interesar rescatar:

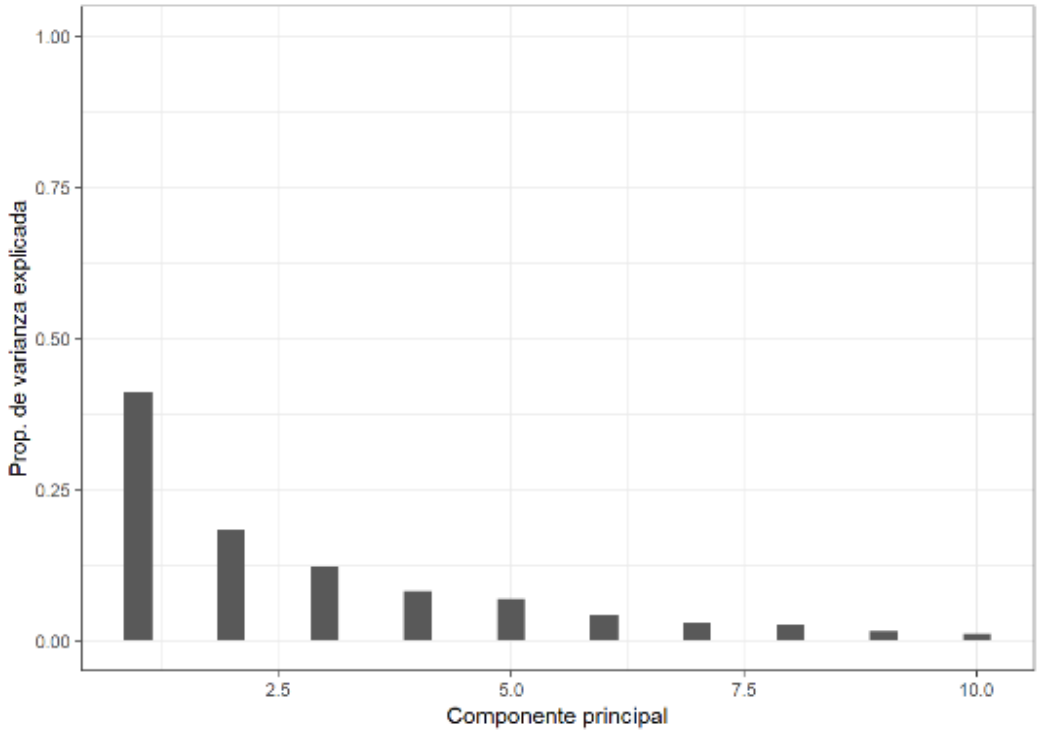


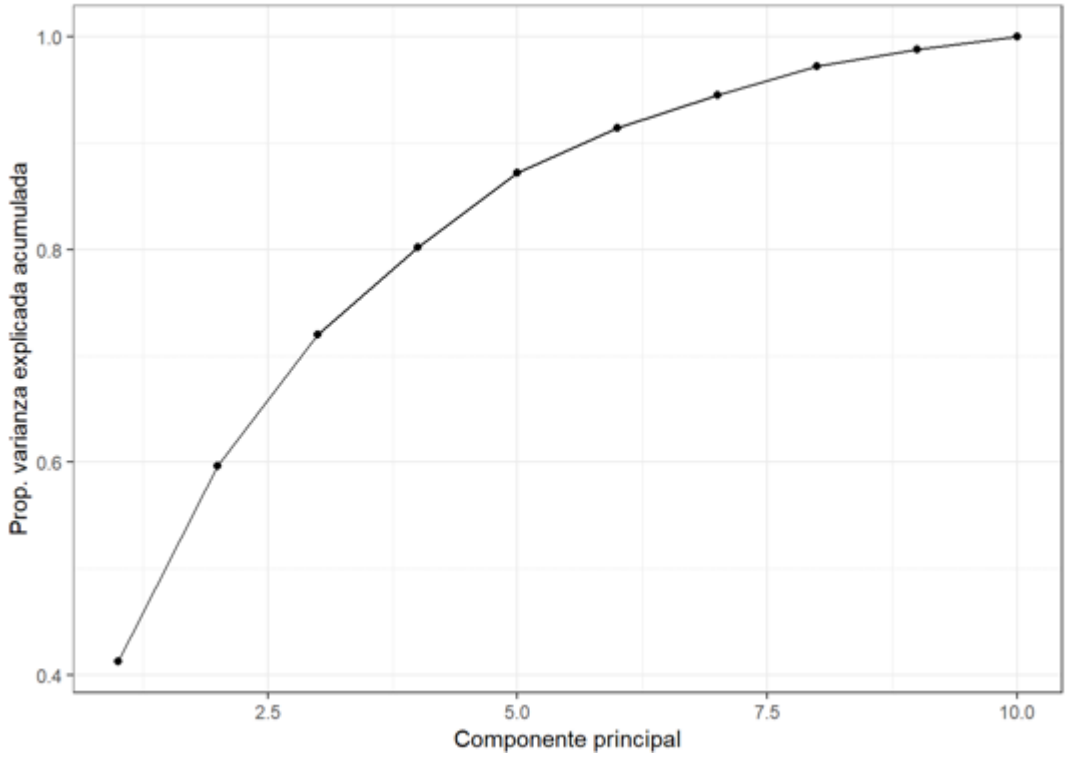
1. Los Eigenvalues y eigenvectors
2. Los loadings o cargas
3. El gráfico de sedimentación (scree plot)
4. La información de individuos y las variables
   i. Coordenadas
   ii. Contribuciones
   iii. Calidad de la representación



Aplicación en R: variancia explicada

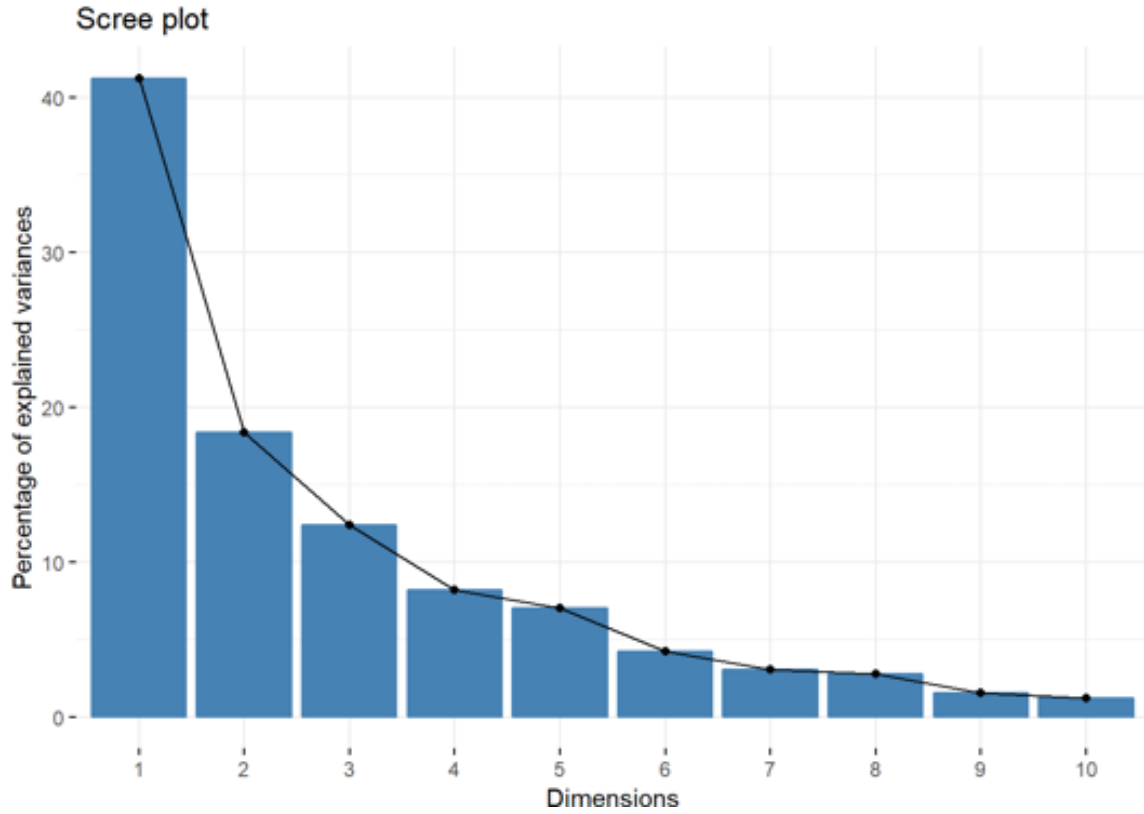
- Es importante antes de analizar las proyecciones conocer la variabilidad explicada por el PCA. Se puede hacer mediante la proporción (porcentaje) de la variancia explicada.





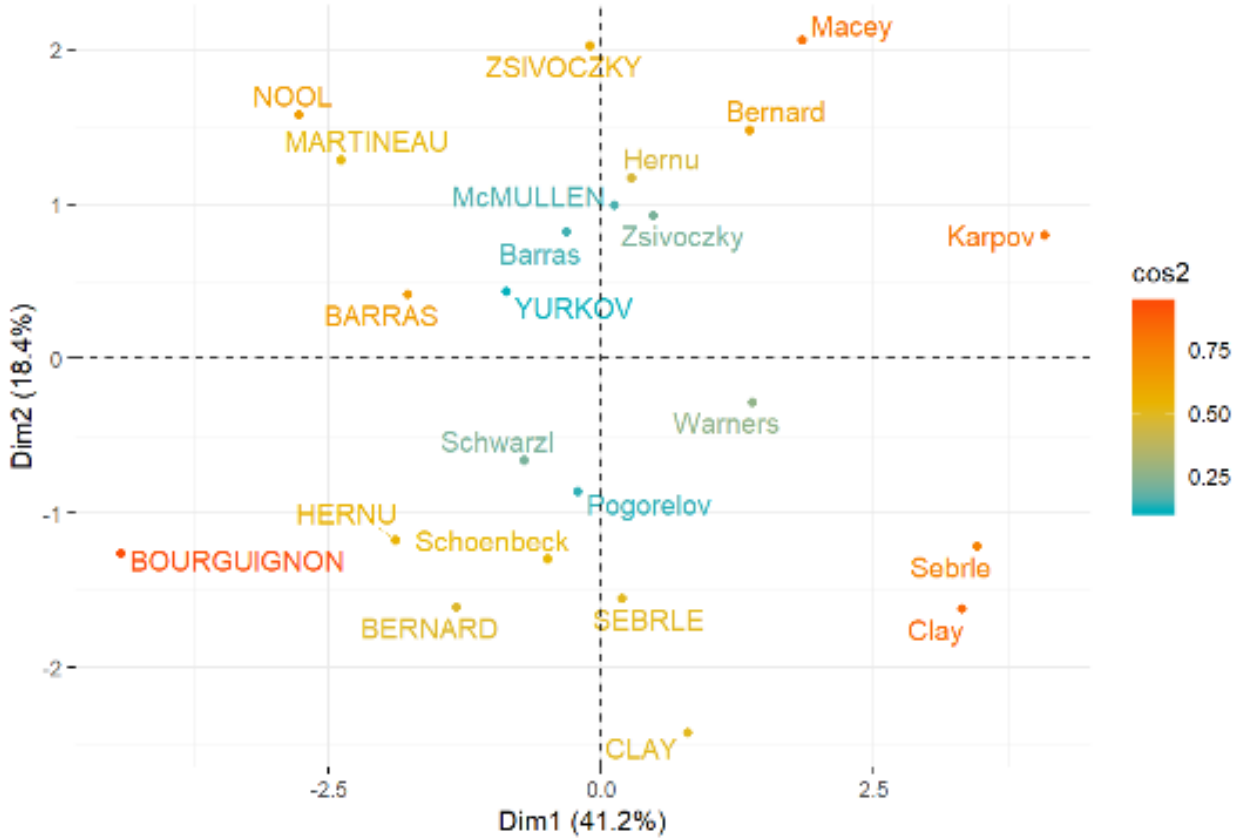
Aplicación en R: variancia explicada

- Otra forma es mediante el gráfico de sedimentación (scree plot). Sobre este, será analizado con mayor detalle en el tema de FA.



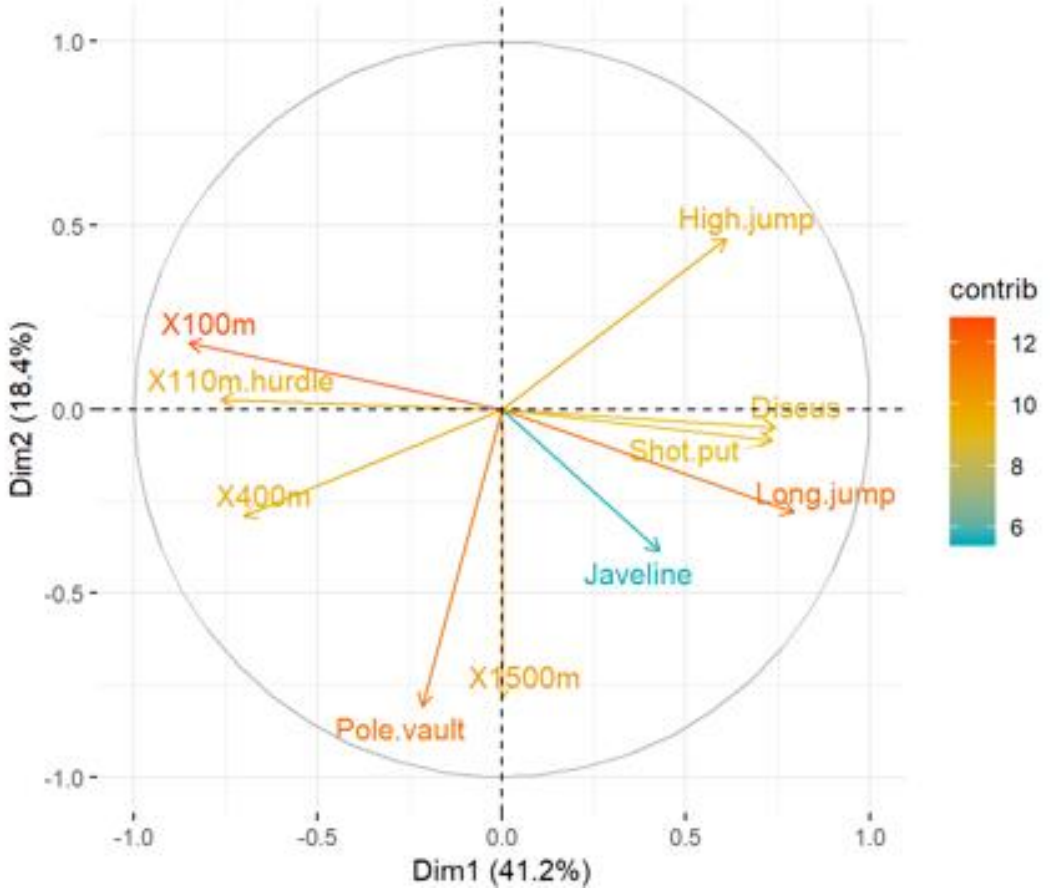
Aplicación en R: proyección de individuos

El primer análisis del PCA es conocer la relación entre los individuos. Acá nos preguntamos: ¿que relación o aglomeración vemos en las unidades de estudio? Para el presente caso, ¿qué se podría decir para la proyecciones de los individuos?



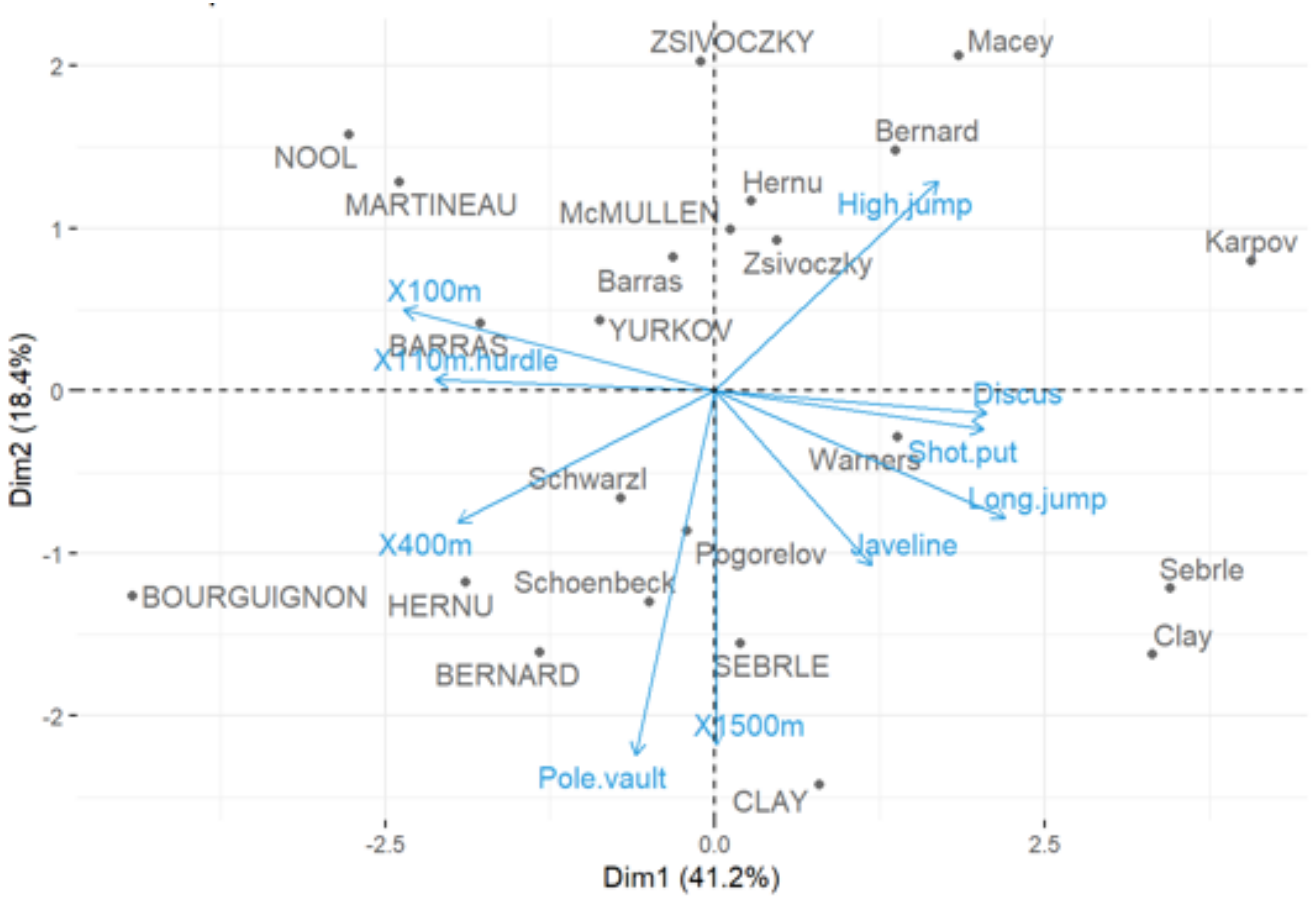
Aplicación en R: proyección de variables

Luego, interesa conocer el análisis de las variables. Se estudia el círculo de correlación y se establece relaciones y disimilitud entre las variables. Para el presente caso, ¿qué podríamos decir sobre el análisis de las variables?



Aplicación en R: proyección individuos - variables

Finalmente, se une la información de los individuos y las variables en una sola proyecciones, para determinar que individuos perteneces o tienen una mejor representación de acuerdo a una determinada variable. ¿Cómo podríamos entonces interpretar esto en nuestro caso?



Índice

1 Introducción

2 Componentes Principales

3 Varianza explicada

4 Proyecciones

5 Aplicación en R

6 Otros elementos de análisis

Índice

6

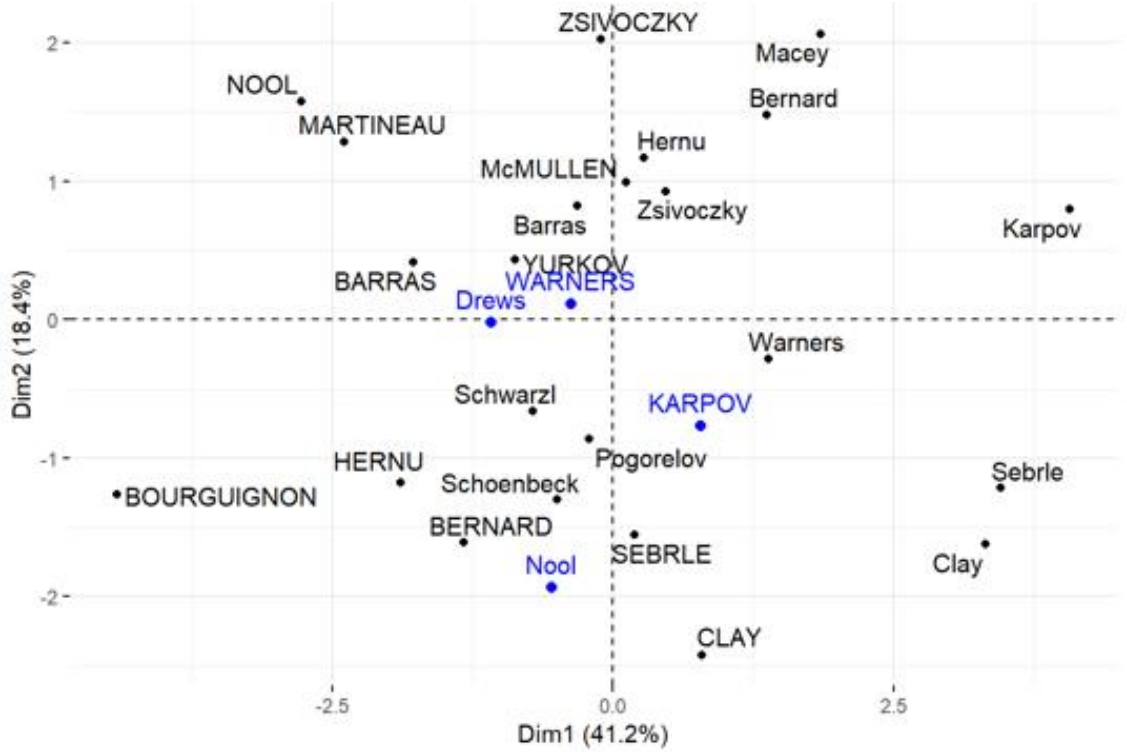
Otros elementos de análisis

Agregar individuos

Agregar variables

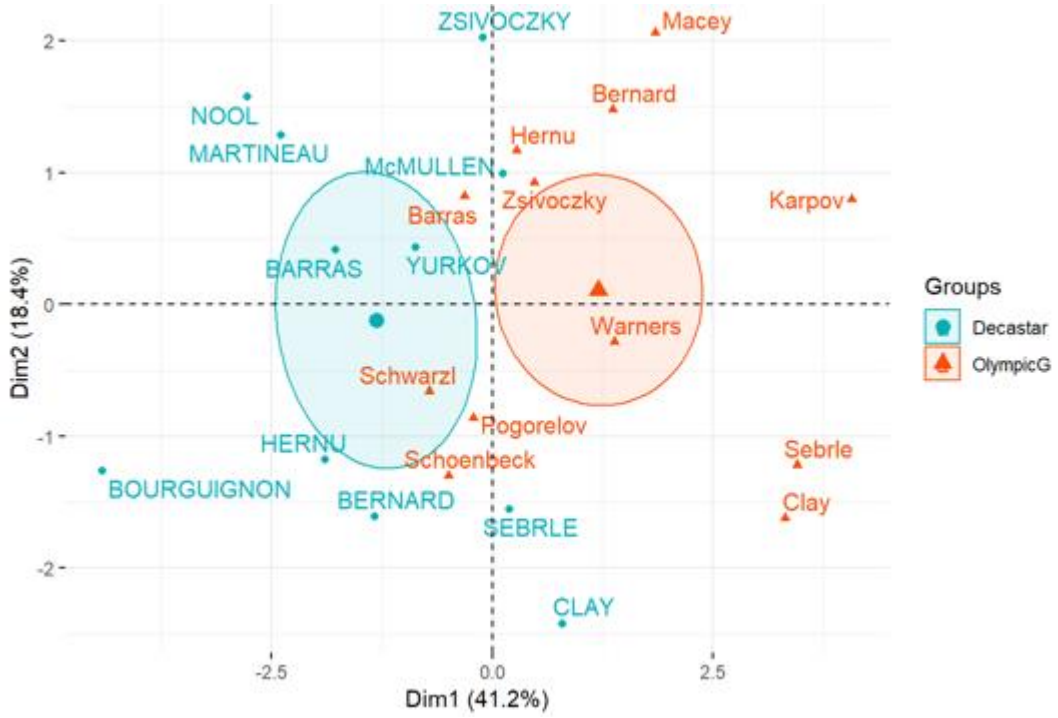
Otros elementos de análisis: agregar individuos

En la formación de la proyección de individuos, es posible que luego se tomaron otras medidas y se tiene nuevos casos. Estos podrían ser entonces agregados al análisis.



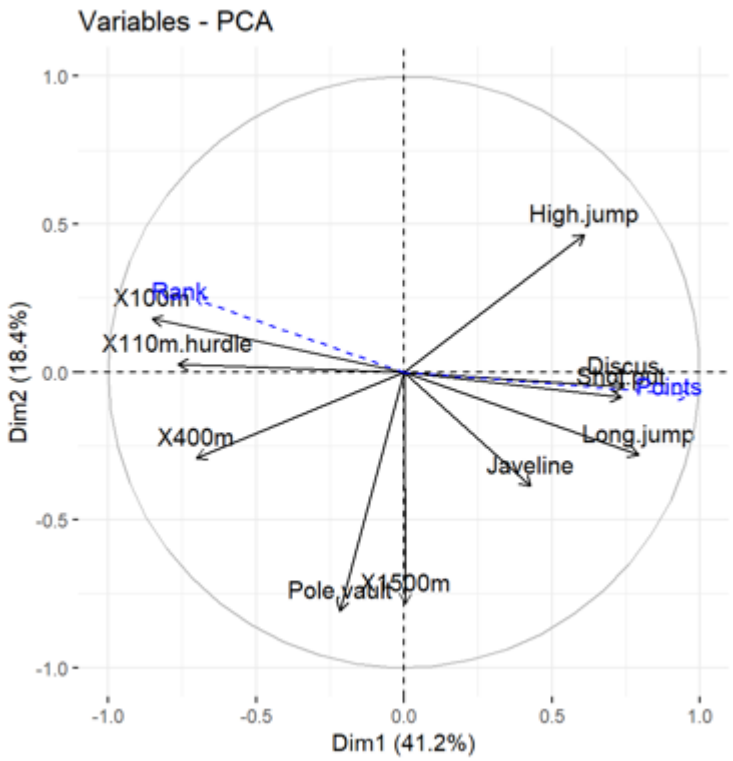
Otros elementos de análisis: agregar variable categórica

Aunque el análisis por PCA se utiliza para analizar ante todo variables cuantitativas o continuas, se podría agregar variables categóricas para ganar cierta. Esta tiene mucho sentido al querer separa los individuos según cierta categoría (ver la proyección de los individuos).



Otros elementos de análisis: agregar variable continua

Asimismo, ante nuevas variables continuas o cuantitativas, estas podrían ser agregadas al análisis del circulo de variables.

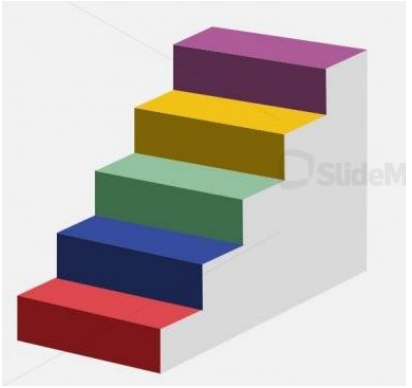


Índice

7 Conclusión

Conclusión: resumen del procedimiento por PCA

- En la realización del análisis por PCA, se deberían de llevar a cabo las siguientes etapas:

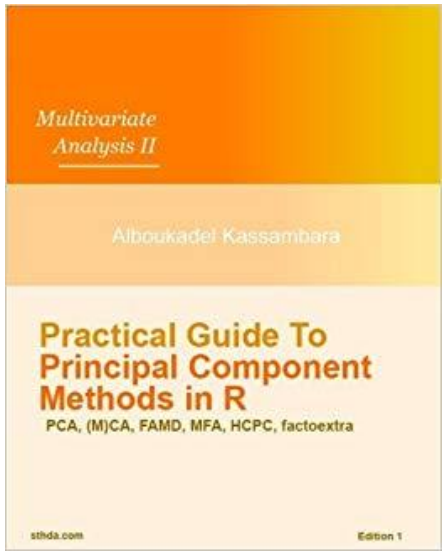


1. Análisis descriptivo de las variables
   i. Correlación y asociación
   ii. Distribución de variables (búsqueda de outliers)
   iii. Otros gráficos univariados y multivariados
2. Varianza explicada del modelo por PCA
3. Proyecciones
   i. Individuos
   ii. Variables
   iii. Individuos – variables
4. Interpretación de los resultados, sobre todo en las proyecciones
5. Conclusión del análisis de PCA al problema en cuestión

Conclusión: material por consultar

Se puede consultar los siguientes enlaces:

https://uc-r.github.io/pca
https://rpubs.com/Joaquin_AR/287787
https://www.rdocumentation.org/packages/FactoMineR/versions/1.41/topics/PCA
https://statquest.org/2017/11/27/statquest-pca-in-r-clearly-explained/
http://www.math.u-bordeaux.fr/~mchave100p/wordpress/wp-content/uploads/2013/10/ACP-TP.pdf
https://www.datacamp.com/community/tutorials/pca-analysis-r
https://www.r-bloggers.com/computing-and-visualizing-pca-in-r/
http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/
https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/
https://datascienceplus.com/principal-component-analysis-pca-in-r/
http://www.gastonsanchez.com/visually-enforced/how-to/2012/06/17/PCA-in-R/



También la lectura del libro: "*Practical guide to Principal Component Methods in R*" de Alboukadel Kassambara.

Conclusión

- El presente capítulo presentó el Análisis por Componentes Principales (PCA) como una forma de interpretar los datos en la reducción de variables.

- El objetivo del PCA, más allá de estudiar la variabilidad del modelo, se centra a partir de las proyecciones el poder entender tanto la relación de los individuos (unidades de estudio), las variables y de individuos – variables.

- El PCA realiza un supuesto muy grande, y es la ortogonalidad o la independencia de sus componentes, lo cuál no siempre es cierto.

- El método de Análisis de Factores (FA) será visto posteriormente como una forma de estudiar mejor la estructura de los datos, así para mejorar la interpretación en la reducción de la dimensionalidad de los datos.


The End