Componentes Principales

Oscar Centeno Mora

Introducción

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

- Creados en los años 30 y desarrollados en los años 60 donde se mejore los aspectos geométricos y las representaciones gráficas.

- Son métodos descriptivos, no se apoyan en modelos probabilísticos sino más bien de un modelo geométrico.

- A partir de una matriz rectangular de datos con *p* variables cuantitativas y *n* unidades, el análisis por componentes principales propone diversas representación geométricas de estos dos.

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

Introducción

- Lo que se busca es ver si se pueden buscar grupos de unidades con diferencias o similitudes.

- Para las variables buscamos ver cuáles están muy correlacionadas entre ellas, y, de forma contraria, cuales no están correlacionadas con las otras.

- 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 se deben de tomar en cuenta siempre las medidas de calidad de las representaciones: criterio global y criterios individuales.

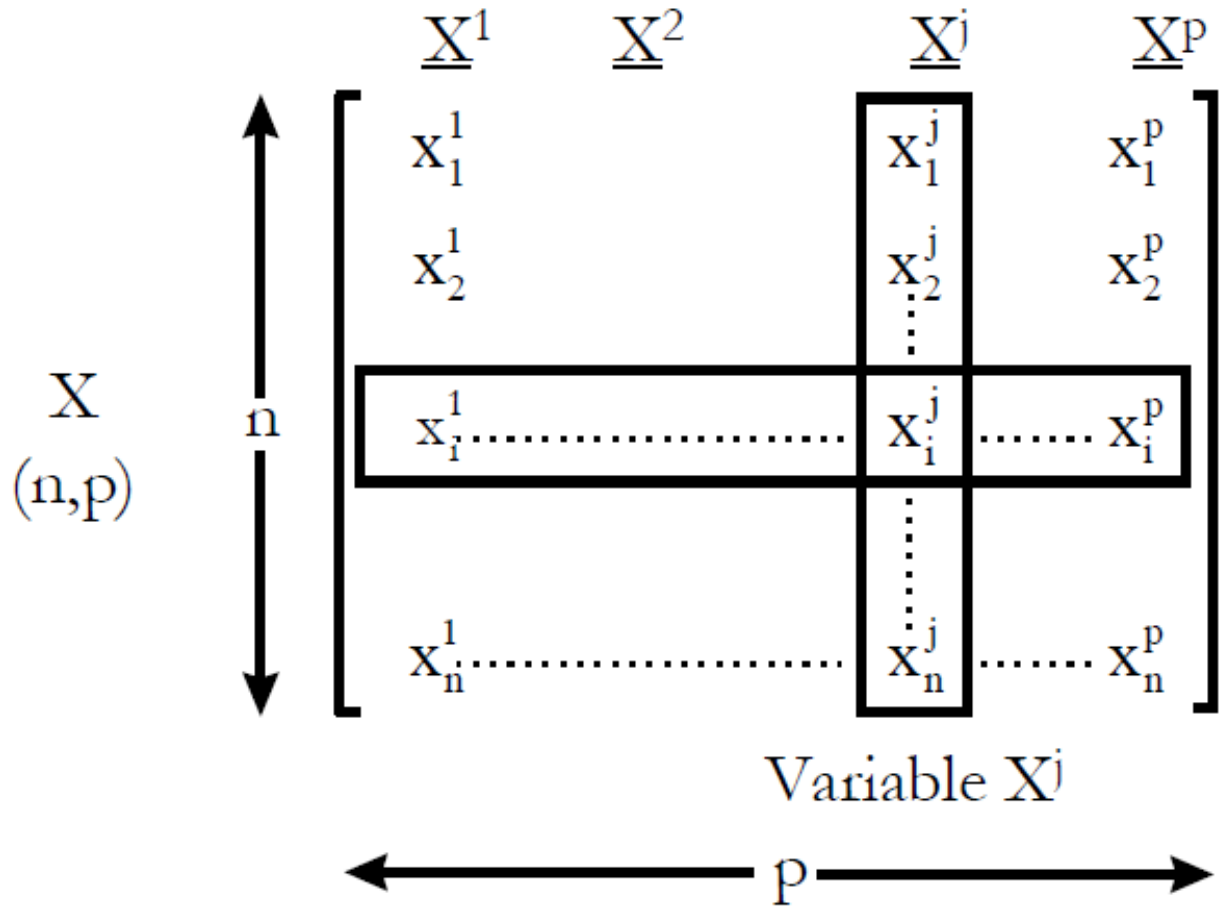
- A veces se recomienda nombrar a los nuevos ejes, y de ahí explicar la posición de los individuos.

- 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.

- 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 inferenciales.

El PCA: los datos

- Los datos corresponden a variables cuantitativas observadas para n individuos.

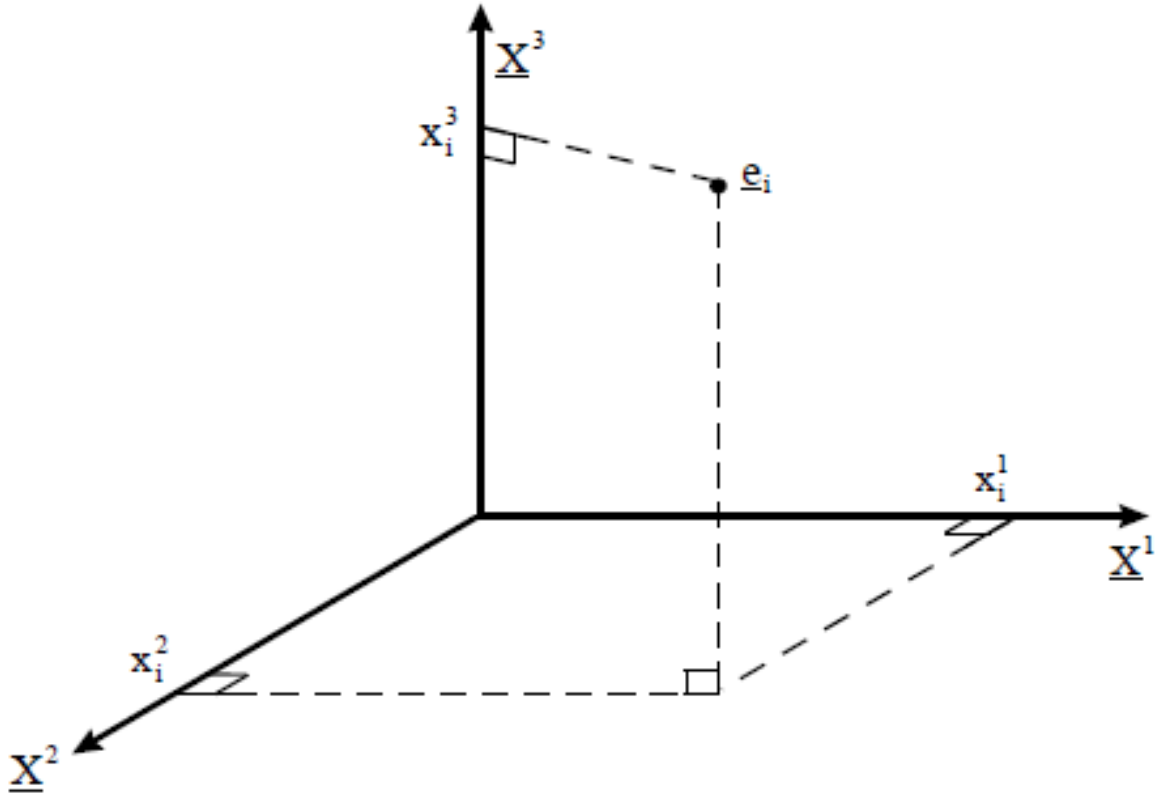


Individuo e'_i

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

El PCA: los datos

- Se trata de visualizar una nube de 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.

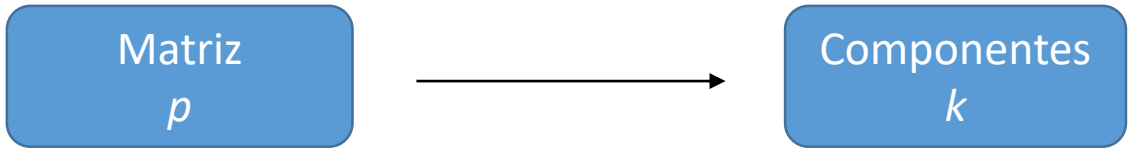


Imposible visualizar a partir de p>3.

El PCA: objetivo

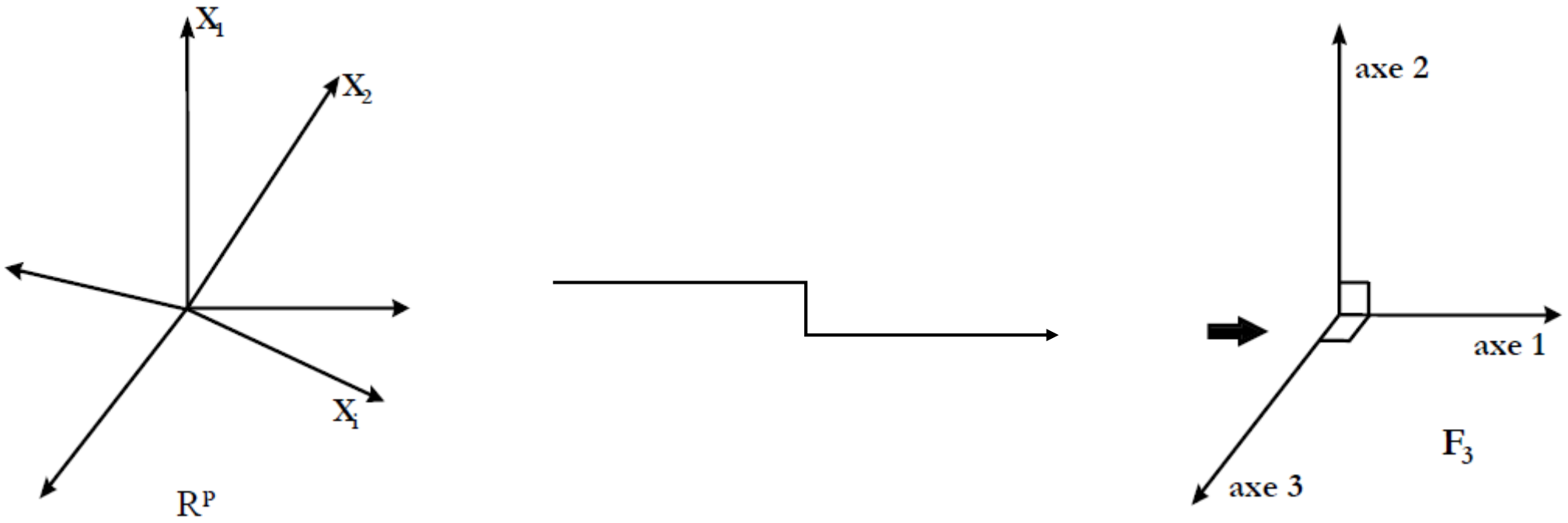
- Se busca representar a los n individuos o casos, para un sub espacio vectorial F_k de R_p de dimensión *k* (*k* pequeño como 2, 3,…; por ejemplo un plano).

- 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 como ***componentes principales***, los ejes serán llamados ***ejes principales***, y las formas lineares asociadas se llamarán ***factores principales***.

El PCA: nuevo espacio vectorial



Espacio vectorial inicial

Nuevo espacio vectorial con los ejes principales

Los individuos

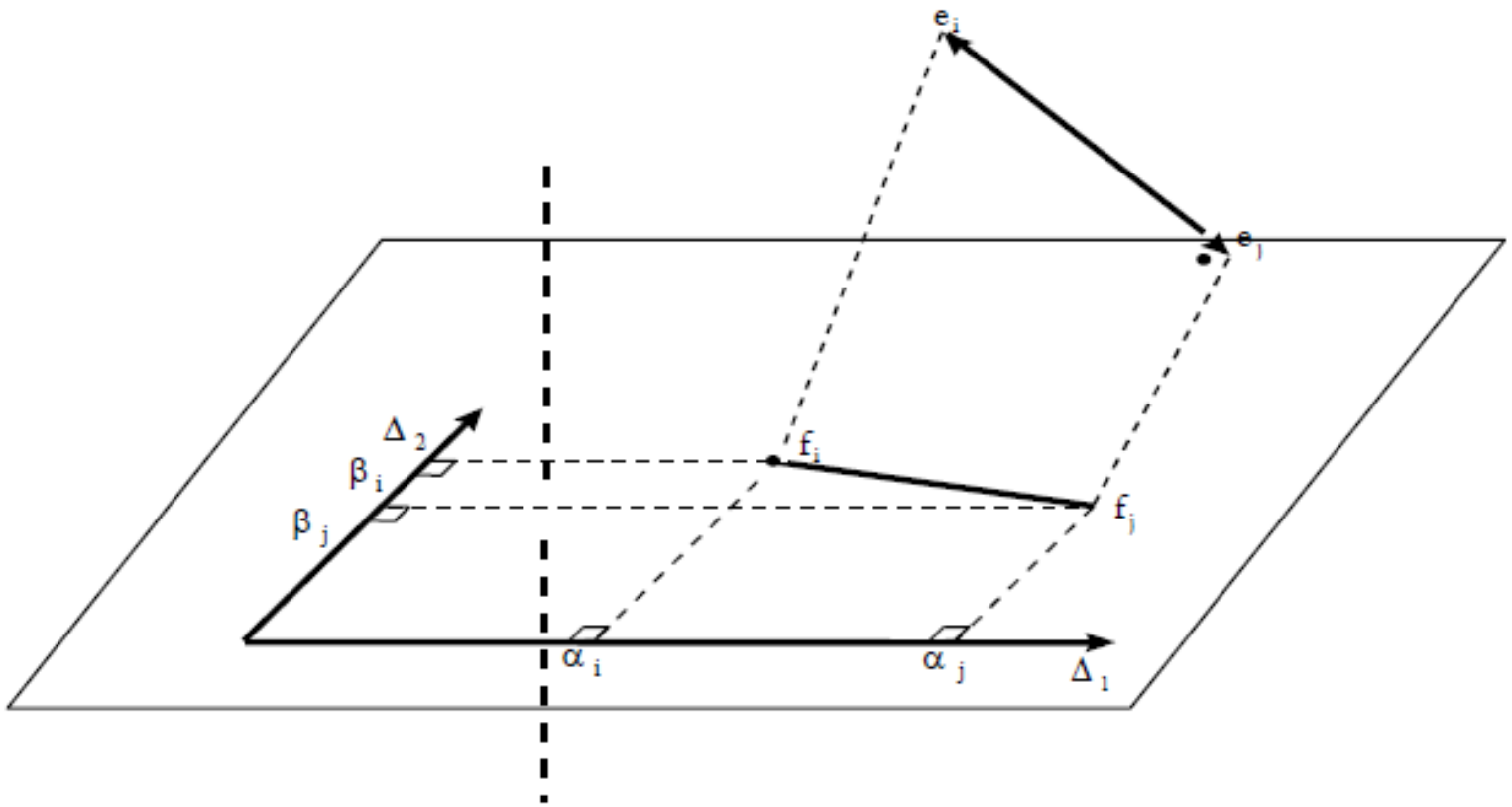
El PCA: la menor perdida posible de información

- 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.

El PCA: la menor perdida posible de información



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

El PCA: la elección de distancia entre los individuos

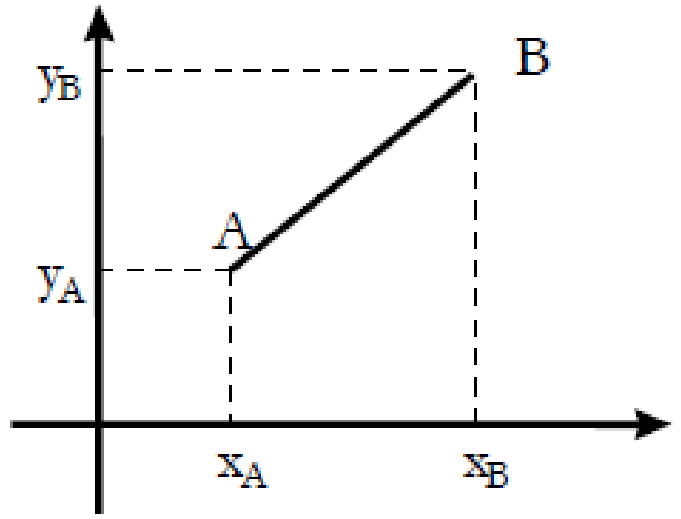
- En el espacio R^p de p dimensiones, se generaliza la siguiente noción: la distancia euclidiana entre dos individuos se expresa como:

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

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

- Por lo tanto

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



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

El PCA: la elección de distancia entre los 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.

El PCA: la inercia total

- La inercia 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

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

El PCA: 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:

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

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

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

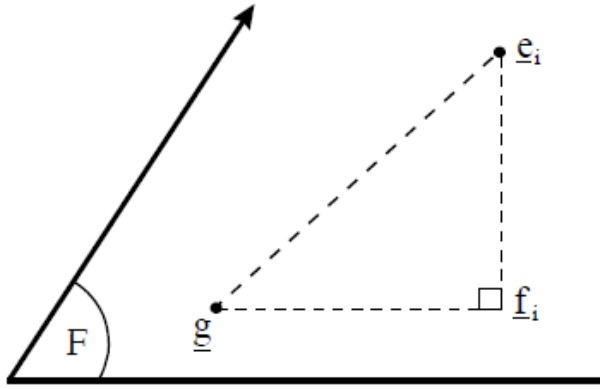
- En el caso donde las variables son estandarizadas, la variancia de cada variable es de 1.

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

El PCA: sobre la equivalencia en la perdida de información

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

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



Proyección ortogonal de la nube sobre el sub espacio

El PCA: sobre la equivalencia en la perdida de información

- 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:

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

El PCA: solución al problema originado

- 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.

El PCA: solución al problema originado

- 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.

El PCA: solución al problema originado

- 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.

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

El PCA: propiedad de los componentes principales

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

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

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

El PCA: representación de los individuos

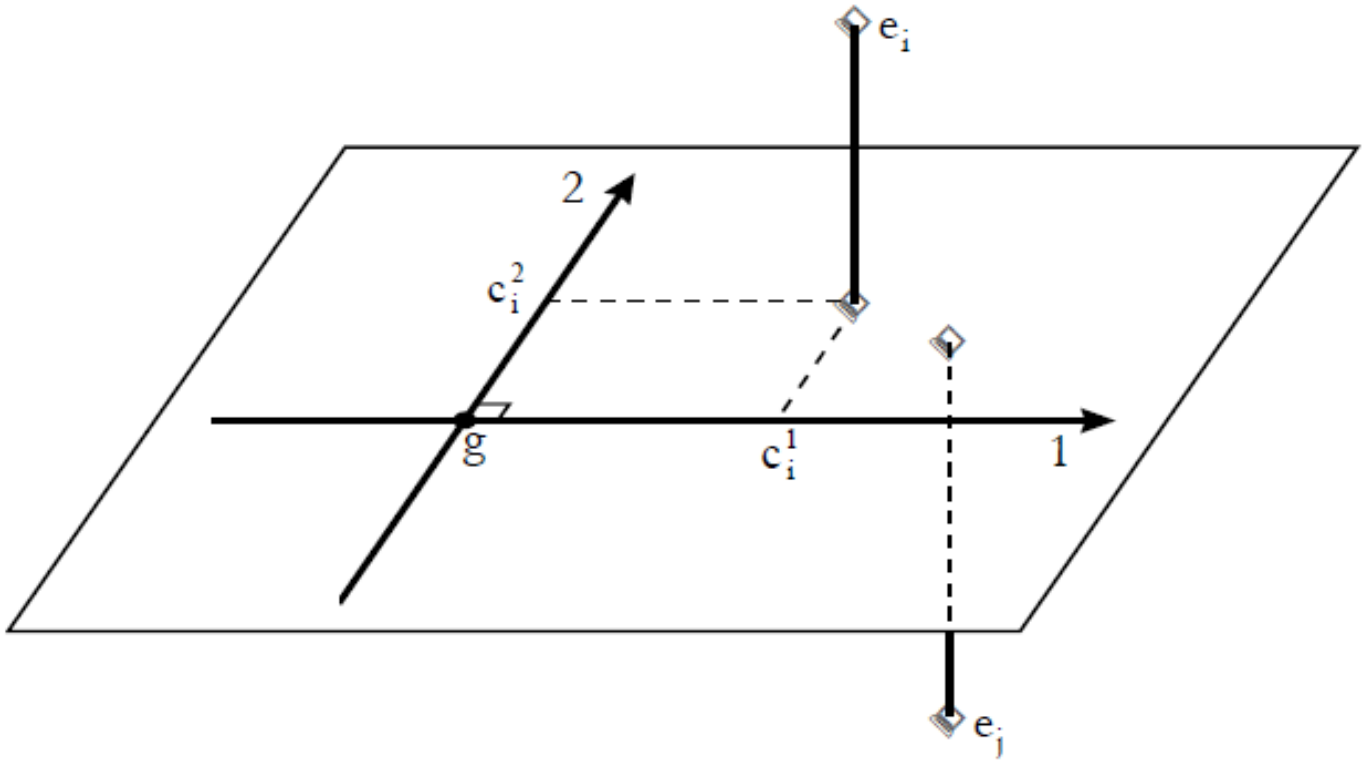
- Los *j* componentes principales brindan la información para ubicar a los n individuos sobre el los *j* ejes principales.

- 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.

El PCA: representación de los individuos



- Se debe de tener cuidado con la calidad de la representación de cada individuo.

El PCA: 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.

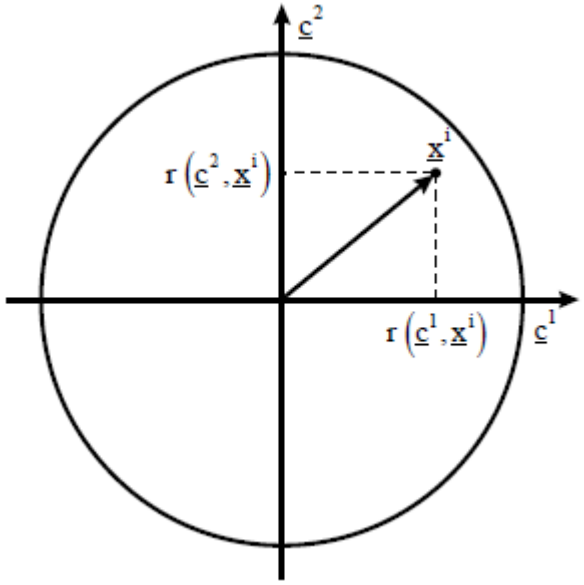
- A veces, los casos muy aparte del resto de los individuos, son complicados de describir o comprender. Se les suele dar la denominación de valor atípico.

Las variables

El PCA: la representación de 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:

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

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

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

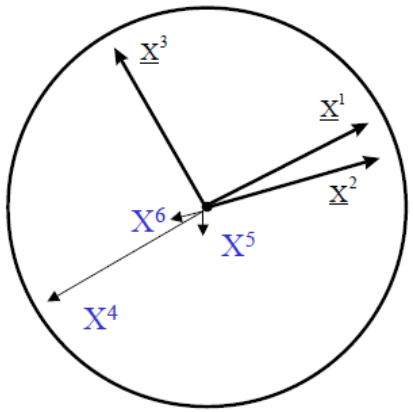
$$\left\| \underline{x}^{i} \right\| = s_{i}$$

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.

Validación de los resultados

El PCA: validación de la representación gráfica

- Para conocer la validez de los resultados globales, se pueden utilizar criterios de la parte explicada por la inercia para el eje *i*.

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \ldots \lambda_p}$$

- Para el caso de los dos primeros componentes, utilizaríamos el siguiente criterio.

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{p} \lambda_i}$$

- Este criterio, muy amenudeo expresado en porcentaje, mide el grado de reconstitución de los cuadrados según la distancia.

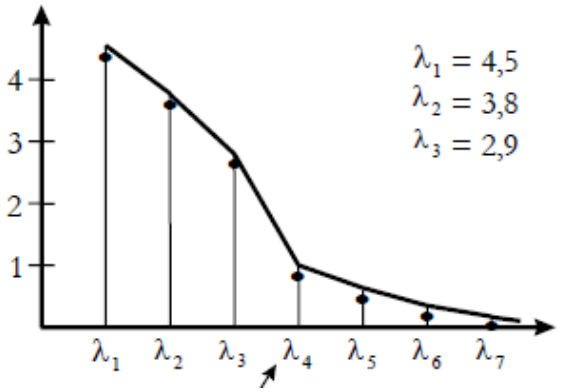
- Entre mayor correlación entre las variables, la reducción de variables será más contundente.

El PCA: cantidad de ejes

- Para determinar la cantidad de ejes, se puede utilizar el porcentaje de inercia requerido (a priori), o dividir la inercia total según la cantidad de variables iniciales.

- Otro método es dibujar un histograma con todos las contribuciones de todas las inercias, y ver donde es que se da un punto de corte. Este método tiene por representación el gráfico de sedimentación.

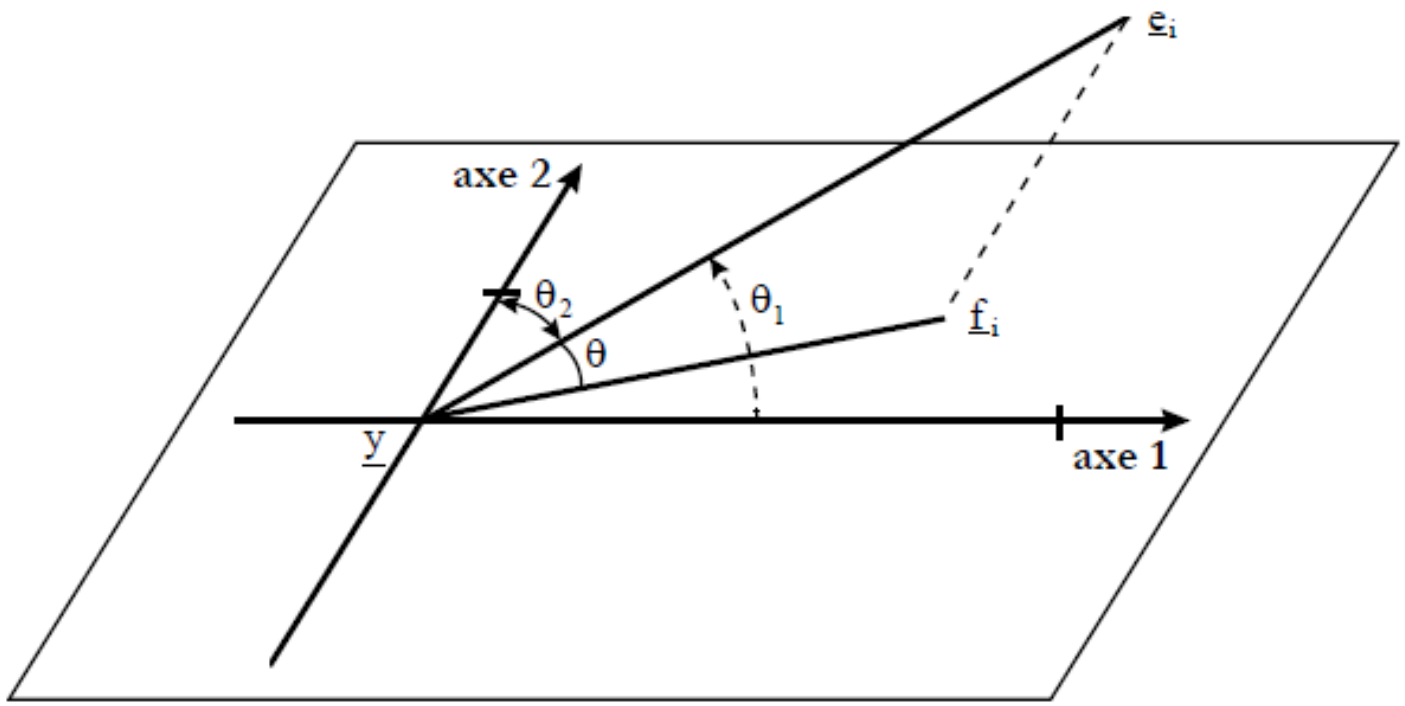
- En este se conservan los ejes o el número de componentes situados antes del corte.



En este caso elegiríamos 3 componentes.

El PCA: criterios individuales

- Se utiliza los cosenos cuadrados:



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

El PCA: criterios individuales

- Para ***cada*** individuo, la calidad de la representación se define por el cuadrado del coseno del ángulo entre el eje de proyección y el vector e_i. Entre más cerca es cercana a 1, mejor es la calidad de la representación.

- En general, las calidades en la representación están dadas eje a eje. Para tener la calidad de la representación en un plano, se suma los criterios correspondientes a los ejes estudiados.

- Este criterio no tiene sentido para los casos cercanos del origen.

- Si se detectan individuos con cosenos cuadrados débiles, se debe tener cuenta de su distancia al origen antes de pensar sobre una mala representación.

El PCA: criterios individuales

- Es útil también calcular para cada eje la contribución brindada para los diversos individuos al eje.

- Consideremos el *k*-ésimo componente principal C^k, sea C^k_i el valor del componente para el individuo i. Este en termino de ***inercia*** sería:

$$\sum_{i=1}^{n} \frac{1}{n} \left(c_i^k\right)^2 = \lambda_k$$

- La ***contribución*** del individuo ei al componente n° k se definiría por la ecuación:

$$\frac{\frac{1}{n}\left(c_i^k\right)^2}{\lambda_k}$$

El PCA: criterios individuales

- No es aconsejable que un individuo tenga una contribución excesiva, dado que provoca la inestabilidad.

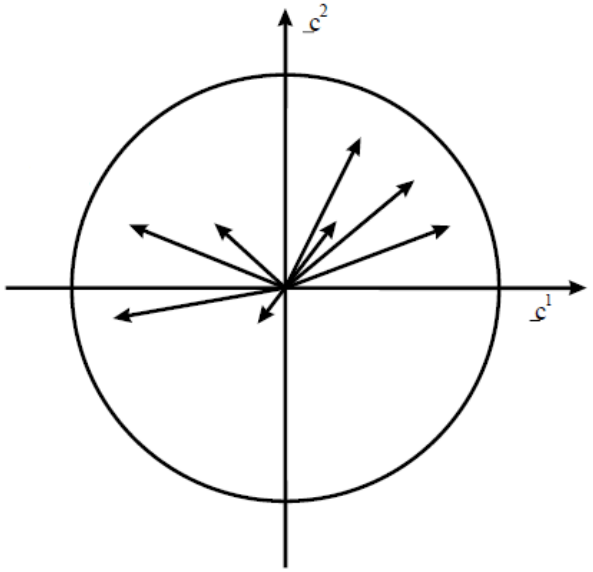
- En dicho caso se prefiere eliminar al o los casos que poseen contribución importantes.

- Este caso se puede presentar en las encuestas por muestreo.

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.



Correlación = coseno

Ejemplos ilustrativos

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

- http://www.sthda.com/english/wiki/principal-component-analysis-in-r-prcomp-vs-princomp-r-software-and-data-mining
- http://www.sthda.com/english/wiki/ade4-and-factoextra-principal-component-analysis-r-software-and-data-mining#at_pco=smlwn-1.0&at_si=58d852c3ce898442&at_ab=per-2&at_pos=0&at_tot=1
- http://www.sthda.com/english/wiki/principal-component-analysis-how-to-reveal-the-most-important-variables-in-your-data-r-software-and-data-mining#at_pco=smlwn-1.0&at_si=58d2e52d6d391656&at_ab=per-2&at_pos=0&at_tot=1
- http://www.sthda.com/english/wiki/factominer-and-factoextra-principal-component-analysis-visualization-r-software-and-data-mining

THE END