Análisis de Componentes Principales (ACP)

Víctor Riquelme F.
DIM; FCFM; UCHILE.
vriquelme@dim.uchile.cl

Noviembre, 2008

Ingredientes

- El actor principal de este análisis será la matriz $X = (x_{ij})_{\substack{i=1\dots n \\ j=1\dots p}}$, donde las filas (x_i^t) representan las mediciones de las características de n individuos y las columnas (x^j) representan las mediciones de p variables.
- Se supondrá que las variables (columnas) están centradas $(\sum_{i=1}^{n} x_{ij} = 0 \ \forall j)$, y normalizadas $(\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1 \ \forall j)$.
- Si las variables no están centradas y normalizadas hay que centrarlas (a cada variable se le resta su media muestral) y normalizarlas (dividirlas por la desviación estandar muestral).

En $\mathbb{R}^p$

- En $\mathbb{R}^p$ se grafican los individuos (los ejes son las variables). Se quieren encontrar mejores ejes para una representación más "clara". Buscamos escribir $X = \sum_{i=1}^{r} c_i u_i^t$
- Aquí, $u_1 \in \mathbb{R}^p$ es el vector director de la "mejor recta que pasa por todos los puntos", y $c_1 \in \mathbb{R}^n$ es el vector cuya componente j es la proyección ortogonal del individuo j sobre la recta generada por u_1.
- u_i son los vectores propios de la matriz $V = \frac{1}{n} X^t X$ (esta es la matriz de varianza-covarianza empírica. Como los datos están centrados y normalizados, esta es la matriz de correlaciones). Esta matriz se puede escribir como $V = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^t$. Definimos l_i los valores propios asociados con u_i, tales que $l_1 \geq l_2 \geq \ldots l_r > 0$.
- Los vectores $u_1, \ldots, u_r$ son los ejes principales, y los c_i son las componentes principales.
- $$c_i = X u_i \quad ; \quad u_i = \frac{X^t c_i}{||X^t c_i||} \quad ; \quad ||c_k||^2 = l_k$$
- En el plano principal se grafican c_{1i} y c_{2i}, para cada individuo x_i.

En $\mathbb{R}^n$

- En $\mathbb{R}^n$ se grafican las variables. Los ejes son los individuos. También se buscan nuevos ejes para una mejor representación de las variables.
- Aquí los ejes principales son d_j, y las componentes principales son v_j. Aquí se mira la matriz X^t.
- $$d_j = \frac{c_j}{\sqrt{l_j}} \quad ; \quad v_j = \sqrt{l_j} u_j$$
- La componente k de v_j es igual a la correlación entre la variable x^k y el eje principal d_j.
- En el círculo de correlaciones se grafican $v_{1j} = \sqrt{l_1} u_{1j}$ y $v_{2j} = \sqrt{l_2} u_{2j}$ para cada variable x^j.
- La distancia al origen de una variable en el círculo de correlaciones mide la calidad de la representación de esa variable.
- El coseno del ángulo entre dos variables es igual a la correlación entre esas variables.

Issues

- *La variabilidad en la componente principal $i-$ésima es* $\frac{l_i}{\sum_{j=1}^{r} l_j}$.
- *La variabilidad acumulada en el plano principal es igual a* $\frac{l_1+l_2}{\sum_{j=1}^{r} l_j}$
- Si se quiere introducir una nueva observación x_{n+1}, su proyección en el eje principal k es igual a $x_{n+1}^t u_k$
- Si se introduce una nueva variable y, su proyección sobre la componente principal es igual a $\mathbb{C}orr(y, c_k)$.

Regresión

Ir a Resolucion Si se quiere hacer una regresión lineal de una nueva variable Y con respecto a las componentes principales,

$$Y = \beta_1 d_1 + \cdots + \beta_r d_r \qquad con \qquad \beta_j = \mathbb{C}orr(d_j, Y)$$
$$Y = \bar{\beta}_1 c_1 + \cdots + \bar{\beta}_r c_r \qquad con \qquad \bar{\beta}_i = \mathbb{C}orr(c_i, Y)/\sqrt{l_i}$$

Y con respecto a las otras variables

$$Y = \gamma_1 x^1 + \cdots + \gamma_p x^p \quad con \quad \gamma_i = \sum_{j=1}^{r} \bar{\beta}_j u_{ji}$$

Problema

Se consideran V_1, V_2, V_3 y V_4 cuatro variables obtenidas sobre 20 observaciones repartidas en 3 clases A, B y C:

Clase	V_1	V_2	V_3	V_4	Clase	V_1	V_2	V_3	V_4
C	45	25	30	160	C	60	27	13	350
C	40	30	30	200	B	38	37	25	240
C	32	32	36	210	B	35	38	27	220
C	35	28	37	250	B	22	38	40	180
C	50	33	17	260	A	18	33	49	190
B	55	45	0	300	B	15	39	46	185
B	58	35	7	320	A	20	40	50	300
C	62	28	10	310	A	25	35	40	220
B	48	32	20	280	A	22	33	45	225
B	52	34	14	300	C	32	26	42	150

Se efectuó un análisis en Componentes Principles sobre las variables V_1, V_2 y V_3 y se obtuvieron los siguientes resultados: matriz de correlaciones de V_1, V_2 y V_3, los valores propios y las correlaciones entre V_1, V_2 y V_3, y las dos primeras componentes principales $CP1$ y $CP2$; y el gráfico del plano principal.

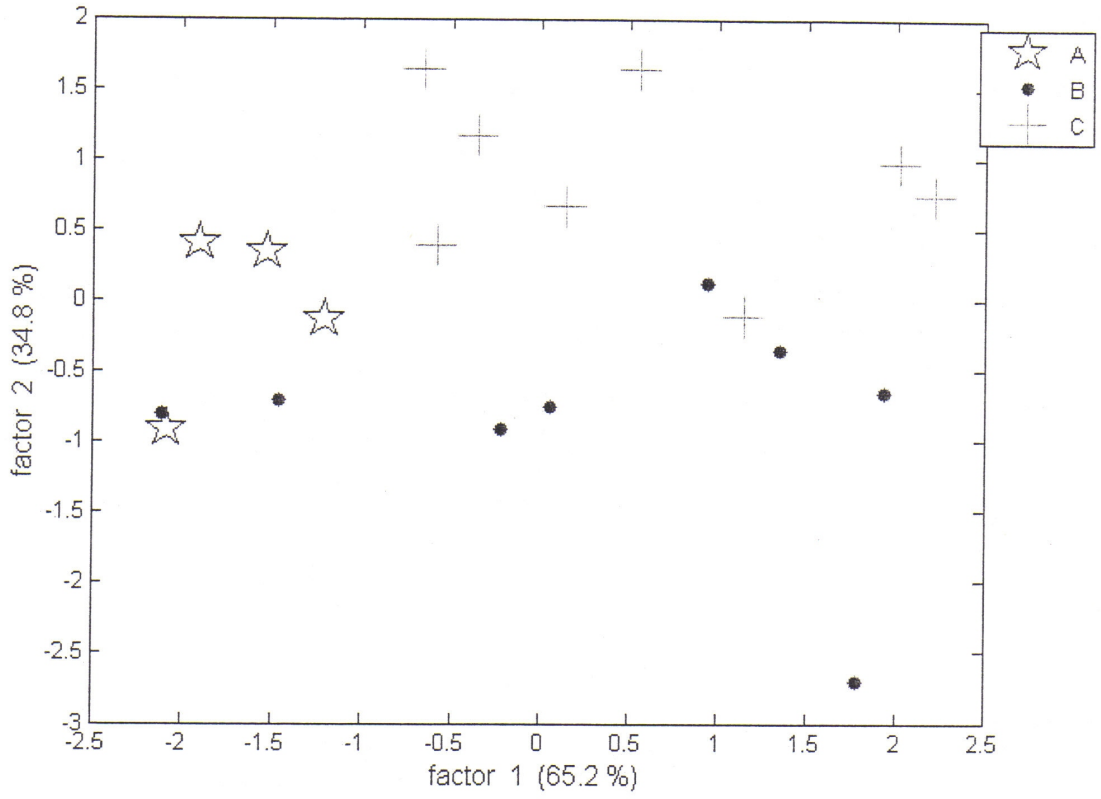
Ir a Resolucion

Matriz Correlaciones

	V_1	V_2	V_3
V_1	1.00	−0.26	−0.94
V_2	−0.26	1.00	−0.09
V_3	−0.94	−0.09	1.00

	Media	Desviación	$Factor1$	$Factor2$
VP			1.956	1.044
%Acum. de VP			65.20	100
V_1	38.20	14.98	0.997	0.076
V_2	33.40	5.18	−0.189	−0.982
V_3	28.40	14.50	−0.962	0.273

Plano Principal



Preguntas

1. Dibuje el círculo de correlaciónes y coméntelo.
2. Justifique la calidad de la representación y comente el plano principal.
3. Proponga un indicador único que representa a las observaciones. ¿A qué criterio corresponde? Dé los valores del indicador sobre las 20 oservaciones.
4. Se quiere efectuar la regresión de V_4 sobre V_1, V_2 y V_3. ¿Qué problema numérico va a encontrar?
5. La correlación de la variable V_4 con la primera componente principal C_1 es 0.69 y con la segunda es -0.29. Ubique V_4 en el círculo de correlaciones. Comente. Dé el coeficiente de correlación múltiple de V_4 sobre V_1, V_2 y V_3.

Resolución

Parte 1

Ir a Tablas En la primera tabla se encuentra la matriz de correlaciones de las variables V_1, V_2 y V_3.
En la segunda tabla se encuentran los valores propios, y los vectores v_1 y v_2.
Notar que los vectores que aparecen como C_1 y C_2 no tienen norma 1, pero al dividirlos por la raiz del respectivo valor propio quedan normalizados. Por eso se puede identificar que con los v_i y no los u_i. En el círculo de correlaciones se grafican entonces los puntos $(0.997, 0.076)$ para V_1, $(-0.189, -0.982)$ para V_2 y $(-0.962, 0.273)$ para V_3. Sacar conclusiones!!!.

Parte 2

En las dos componentes principales C_1 y C_2 se encuentra toda la variabilidad, así que prácticamente no se necesitarían las otras componentes principales para explicar a las variables. La representación es buena
Los elementos del grupo A son cercanos entre si, los del grupo B también son cercanos entre si, pero los del grupo C estan muy dispersos entre si. El cálculo de las componentes principales es

$$c_i = Xu_i = Xv_i/\sqrt{l_i}$$

Parte 3

Para poder hacer un índice único, lo mejor es proyectar todo sobre una sola componente principal, la que tiene mayor variabilidad, o sea, la primera. Así:

$$\begin{aligned} c_1 &= \frac{1}{\sqrt{1.956}}(0.997V_1 - 0.189V_2 - 0.962V_3) \\ &= 0.7129V_1 - 0.1351V_2 - 0.6878V_3 \end{aligned}$$

El índice aumenta bastante con V_1, disminuye bastante con V_3 y también disminuye, pero no tanto, con V_2.

Los resultados para c_1 son:

(0.5051	0.1273	-0.5974	-0.3907	1.1434
1.8619	1.9483	2.1904	0.9322	1.3546
1.9791	0.0764	-0.1903	-1.4331	-1.9157
-2.0822	-2.0520	-1.2064	-1.5333	-0.7176)

El individuo 8 es el con índice mayor (2.1904), y el individuo 16 es el con menor índice (-2.0822).

Los valores de c_2 son

(1.6521	0.6757	0.3637	1.1578	-0.0801
-2.6456	-0.6014	0.8047	0.1544	-0.3146
1.0392	-0.7565	-0.9256	-0.7556	0.3384
-0.8724	-0.9644	-0.1698	0.2861	1.6138)

$$u_1 = (0.7129, -0.1351, -0.6878)^t$$

$$u_2 = (0.0744, -0.9611, 0.2672)^t$$

Esto servirá para calcular cosas.

Parte 4

Ir a Regresion Para hacer la regresión de V_4 con respecto a V_1, V_2 y V_3, se aplica la fórmula:

$$\begin{aligned} V_4 &= V_1\left(\frac{\mathbb{C}orr(c_1,V_4)}{\sqrt{l_1}}u_{11}+\frac{\mathbb{C}orr(c_2,V_4)}{\sqrt{l_2}}u_{21}\right) \\ &\quad +V_2\left(\frac{\mathbb{C}orr(c_1,V_4)}{\sqrt{l_1}}u_{12}+\frac{\mathbb{C}orr(c_2,V_4)}{\sqrt{l_2}}u_{22}\right) \\ &\quad +V_3\left(\frac{\mathbb{C}orr(c_1,V_4)}{\sqrt{l_1}}u_{13}+\frac{\mathbb{C}orr(c_2,V_4)}{\sqrt{l_2}}u_{23}\right) \\ &= V_1\left(0.4695\times 0.7129-0.2739\times\right) \\ &\quad +V_2\left(0.4695\times -0.1351-0.2739\times -0.9611\right) \\ &\quad +V_3\left(0.4695\times -0.6878-0.2739\times 0.2672\right) \\ &= 0.3143V_1+0.1998V_2-0.3961V_3 \end{aligned}$$

Para el cálculo anterior, V_4 hay que centrarlo y normalizarlo (esta puede ser la complicación, la diferencia de unidades).$media(V_4)=242.5$; $desv.estandar(V_4)=56.2694$.

Parte 5

Si X es la matriz normalizada con las variables (V_1,V_3,V_3), y $\hat{\beta}=(0.3143,0.1998,-0.3961)^t$ (los coeficientes de la regresión), entonces

$$\begin{aligned} R^2 &= 1-\frac{(V_4-X\hat{\beta})^t(V_4-X\hat{\beta})}{V_4^tV_4} \\ &= 0.5167 \\ R &= 0.7188 \end{aligned}$$

Recuerdo: La fórmula original para el R de la regresion es

$$R^2=1-\frac{(y-Z\hat{\beta})^t(y-Z\hat{\beta})}{\sum_{i=1}^n(y_i-\bar{y})^2}$$

Extra

¿Qué pasaría si se introdujese una nueva observación? ¿Se podría decir algo acerca de a que grupo pertenece?
Sea la observación $x_{21} = (19, 42, 42, 250)^t$. El cálculo de sus proyecciones sobre los ejes principales será (usando el análisis de componentes principales dado para las tres variables). Para introducir este nuevo dato se debe centrar y normalizar. Esto se realiza usando los datos que ya se tenian para las variabes V_1, V_2 y V_3 (o sea, sus medias y sus desviaciones estandar):

$$x = \left(\frac{19 - 38.2}{14.98}, \frac{42 - 33.4}{5.18}, \frac{42 - 28.4}{14.5} \right)^t = (-1.2817, 1.6602, 0.9379)^t$$

Entonces, las componentes principales de x serán las proyecciones sobre u_1 y u_2:

$$\begin{aligned} F_1 &= x^t u_1 \\ &= (-1.2817, 1.6602, 0.9379) \cdot (0.7129, -0.1351, -0.6878) \\ &= -1.7832 \\ F_2 &= x^t u_2 \\ &= (-1.2817, 1.6602, 0.9379) \cdot (0.0744, -0.9611, 0.2672) \\ &= -1.4404 \end{aligned}$$

La nueva observación tiene alta chance de pertenecer al grupo A (Pero no estamos seguros por el extraño comportamiento de los individuos del grupo C).