Anàlisi de Components Principals

Introducció

El problema central de l'anàlisi de dades és la reducció de la dimensionalitat.

És a dir, si és possible descriure amb precisió els valors de les p variables per un petit subconjunt $r < p$ d'aquestes variables amb una pèrdua mínima d'informació.

Aquest és l'objectiu de l'anàlisi de components principals: donades n observacions de p variables, (taula de dades) s'analitza si és possible representar aquesta informació amb menys variables.

Introducció

Dit de forma més explícita, volem transformar les variables de la nostra taula de dades en unes noves variables anomenades components principals que siguin incorrelacionades entre sí i que siguin combinació lineal de les variables originals.

Si el nombre de variables noves és més petit que les variables originals, hi haurà una pèrdua d'informació.

Volem que aquesta pèrdua sigui mínima en el sentit de que les components principals heretin la màxima variabilitat de les variables originals.

Components principals

Anomenarem $X_1, \ldots, X_p$ a les nostres variables originals i $CP_1, \ldots, CP_r$ a les variables components principals on $r \leqslant p$.

Volem calcular una matriu Λ tal que:

$$\mathbf{CP} = \Lambda \mathbf{X},$$

on $\mathbf{CP} = (CP_1, \ldots, CP_r)^\top$, $\mathbf{X} = (X_1, \ldots, X_p)^\top$ i

$$\Lambda = \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{1p} \\ \alpha_{21} & \cdots & \alpha_{2p} \\ \cdots & \cdots & \cdots \\ \alpha_{r1} & \cdots & \alpha_{rp} \end{pmatrix}.$$

Components principals

Escrit en components:

$$\begin{aligned} CP_1 &= \alpha_{11}X_1 + \cdots + \alpha_{1p}X_p, \\ CP_2 &= \alpha_{21}X_1 + \cdots + \alpha_{2p}X_p, \\ \vdots &\quad \vdots \\ CP_r &= \alpha_{11}X_1 + \cdots + \alpha_{rp}X_p. \end{aligned}$$

Components principals

En la pràctica, sigui $\mathbf{X}$ la nostra matriu $n \times p$ que representa la taula de dades original on tenim n individus i p variables que suposarem centrada. O sigui, les mitjanes de les columnes de $\mathbf{X}$ són nul·les.

Volem obtenir una nova matriu $\mathbf{Y}$ $n \times r$ corresponent a les components principals tal que: $\mathbf{Y} = \mathbf{X} \cdot \Lambda^\top$.

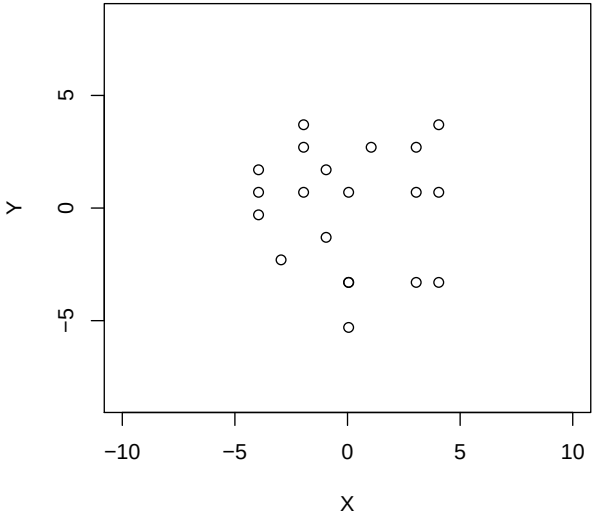
Escrit en components:

$$y_{ki} = x_{k1}\alpha_{i1} + \cdots + x_{kp}\alpha_{ip}, \text{ per } k = 1, \ldots, n, i = 1, \ldots, r.$$

Ens adonem que la matriu $\mathbf{Y}$ també serà centrada.

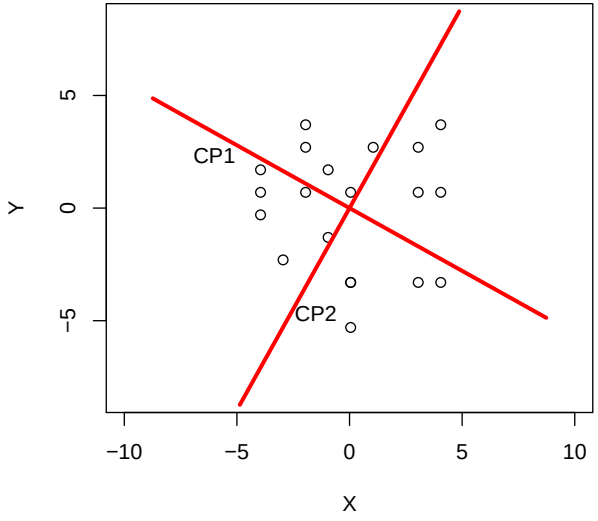
Interpretació geomètrica

Suposem que $p = 2$ i que el nostre "núvol" de punts de la nostra taula de dades és el que mostra la figura:



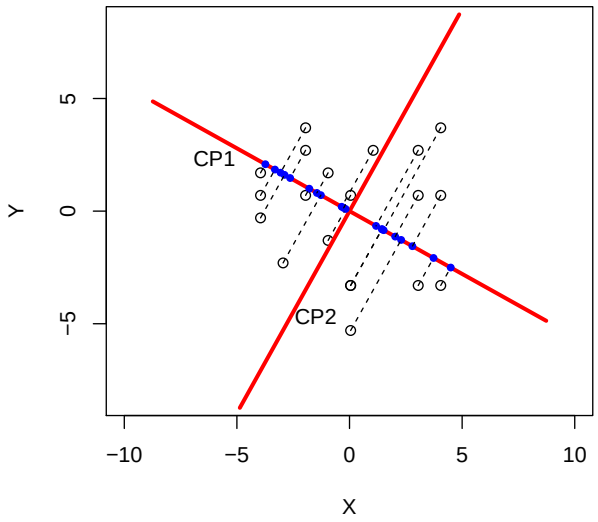
Interpretació geomètrica

A continuació mostram les dues components principals. O sigui, les direccions on les projeccions de les dades tenen màxima variabilitat:



Interpretació geomètrica

Si projectam en la direcció de la primera component, obtendrem les projeccions següents (punts blaus):



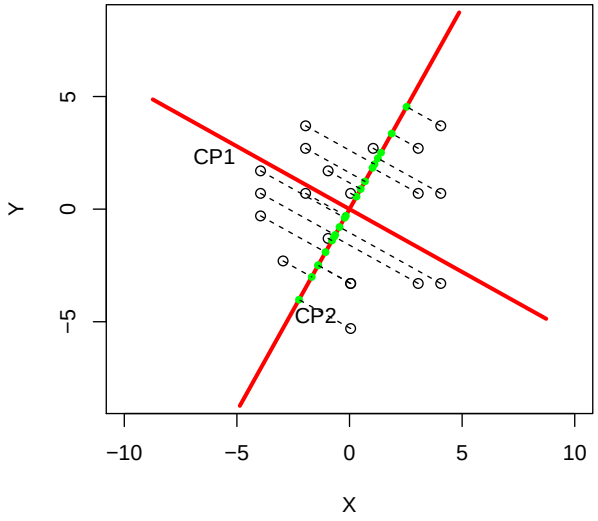
Interpretació geomètrica

Això significa que la variància dels punts blaus és màxima en el sentit de que si haguéssim escollit una altra direcció o una altra recta i haguéssim projectat sobre aquesta segona recta, la variància de les projeccions hagués estar menor.

Els punts blaus representen les coordenades que tenen els punts de la nostra taula de dades si haguéssim agafat com eix d'abscisses, l'eix de la primera component CP_1.

Interpretació geomètrica

Si projectam en la direcció de la segona component, obtendrem les projeccions següents (punts verds):



Components principals

Condicions han de verificar les components principals:

Han d'esser incorrelades.

O sigui, $r_{CP_i,CP_j} = 0$ o si $\mathbf{y}_i$ i $\mathbf{y_j}$ són les columnes i i j de la matriu $\mathbf{Y}$, $r_{\mathbf{y}_i,\mathbf{y}_j} = 0$. Dit en altres paraules, la matriu de covariàncies o de correlacions de la taula de dades $\mathbf{Y}$ serà diagonal.

Components principals

Condicions han de verificar les components principals:

Les variàncies de les components principals han de decréixer.

O sigui,

$$\text{var}(CP_1) \geqslant \text{var}(CP_2) \cdots \geqslant \text{var}(CP_p).$$

D'aquesta forma, la component principal CP_1 serà la que tengui més variabilitat de totes i per tant, la més important, CP_2, la segona més important i així successivament.

ACP sobre la matriu de covariàncies

Considerem $\mathbf{X}$ $n \times p$ la nostre matriu de dades que suposarem centrada on tenim n individus i p variables. Si no ho fos, l'haurem de centrar.

Sigui $\mathbf{S}$ $p \times p$ la matriu de covariàncies de $\mathbf{X}$.

Recordem que $\mathbf{S}$ es calcula com:

$$\mathbf{S} = \frac{1}{n}\mathbf{X}^\top \cdot \mathbf{X}.$$

ACP sobre la matriu de covariàncies

Siguin $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$ els valors propis de la matriu $\mathbf{S}$ en ordre creixent.

Siguin $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$ els corresponents vectors propis que suposarem ortogonals i normalitzats. Això és, suposarem que són perpendiculars dos a dos i tenen norma euclídea unitat.

Sigui $\mathbf{V}$ la matriu de vectors propis que té els vectors anteriors per columnes.

Aleshores la matriu Λ és la transposada de la matriu $\mathbf{V}$:

$$\Lambda = \mathbf{V}^\top.$$

ACP sobre la matriu de covariàncies

Exemple

Considerem la següent matriu de dades:

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & 3 \\ 1 & 0 & 3 \\ 2 & 3 & 0 \\ 3 & 0 & 1 \end{pmatrix}$$

ACP sobre la matriu de covariàncies

Exemple

Com que la matriu no està centrada, primer la centram:

$$\begin{aligned}\tilde{\mathbf{X}} &= \mathbf{H}_4\mathbf{X} \\ &= \begin{pmatrix} 0.75 & -0.25 & -0.25 & -0.25 \\ -0.25 & 0.75 & -0.25 & -0.25 \\ -0.25 & -0.25 & 0.75 & -0.25 \\ -0.25 & -0.25 & -0.25 & 0.75 \end{pmatrix} \cdot \begin{pmatrix} 1 & -1 & 3 \\ 1 & 0 & 3 \\ 2 & 3 & 0 \\ 3 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} -0.750 & -1.500 & 1.250 \\ -0.750 & -0.500 & 1.250 \\ 0.250 & 2.500 & -1.750 \\ 1.250 & -0.500 & -0.750 \end{pmatrix}\end{aligned}$$

ACP sobre la matriu de covariàncies

Exemple

La matriu de covariàncies serà:

$$\mathbf{S} = \frac{1}{4}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} = \begin{pmatrix} 0.688 & 0.375 & -0.812 \\ 0.375 & 2.250 & -1.625 \\ -0.812 & -1.625 & 1.688 \end{pmatrix}$$

Els valors propis de la matriu de covariàncies anterior són:

$$3.830 \quad 0.774 \quad 0.021$$

ACP sobre la matriu de covariàncies

Exemple

Els corresponents vectors propis (per columnes) associats als vectors propis anteriors són:

$$\begin{matrix} -0.253 & 0.750 & 0.612 \\ -0.722 & -0.567 & 0.396 \\ 0.644 & -0.342 & 0.685 \end{matrix}$$

Ens adonem que els vectors anteriors són ortogonals i estan normalitzats.

ACP sobre la matriu de covariàncies

Exemple

Així la matriu que ens canviarà de variables originals a components principals serà la transposada dels vectors propis:

$$\Lambda = \begin{pmatrix} -0.253 & -0.722 & 0.644 \\ 0.750 & -0.567 & -0.342 \\ 0.612 & 0.396 & 0.685 \end{pmatrix}$$

Per tant, la matriu de dades en les noves variables serà:

$$\mathbf{Y} = \tilde{\mathbf{X}}\Lambda^\top = \tilde{\mathbf{X}}\mathbf{V} = \begin{pmatrix} 2.078 & -0.139 & -0.197 \\ 1.355 & -0.706 & 0.199 \\ -2.996 & -0.632 & -0.055 \\ -0.438 & 1.477 & 0.053 \end{pmatrix}$$

ACP sobre la matriu de covariàncies

Exemple

Si calculam la matriu de covariàncies de les components principals val:

$$\mathbf{S}_{CP} = \begin{pmatrix} 3.830 & 0.000 & -0.000 \\ 0.000 & 0.774 & 0.000 \\ -0.000 & 0.000 & 0.021 \end{pmatrix}$$

Ens adonam que és diagonal, per tant, les covariàncies entre variables diferents són nul·les i en la diagonal hi ha les variàncies de les components principals que estan en ordre creixent. Us sonen aquests valors?

ACP sobre la matriu de covariàncies

Exemple

Efectivament, són els valors propis de la matriu de covariàncies de les variables originals $\mathbf{S}$.

Comprovem que la variabilitat s'ha conservat.

La variabilitat de les variables originals serà la suma dels valors de la diagonal de la matriu $\mathbf{S}$:

$$0.688 + 2.25 + 1.688 = 4.625.$$

ACP sobre la matriu de covariàncies

Exemple

La variabilitat de les components principals serà la suma dels valors de la diagonal de la matriu $\mathbf{S}_{CP}$:

$$3.83 + 0.774 + 0.021 = 4.625.$$

Podem observar que les dues variabilitats coincideixen.

La primera component principal (1a. columna de la matriu **CP**) hereta el $\frac{3.83}{4.625} \cdot 100\% = 82.801\%$ de la variabilitat total. Les dues primeres components principals hereten el $\frac{3.83+0.774}{4.625} \cdot 100\% = 99.544\%$ de la variabilitat total.

ACP sobre la matriu de covariàncies

Exemple

Facem un exemple més complet. La taula següent ens dóna l'edat en dies (x_1), l'alçada al néixer en cm. (x_2), el seu pes en kg. en néixer (x_3) i l'augment en tant per cent del seu pes actual respecte el seu pes en néixer (x_4) de 9 nens i nenes recent nats.

ACP sobre la matriu de covariàncies

Exemple

x_1	x_2	x_3	x_4	Sexe
78	48.2	2.75	29.5	Nina
69	45.5	2.15	26.3	Nina
77	46.3	4.41	32.2	Nina
88	49	5.52	36.5	Nin
67	43	3.21	27.2	Nina
80	48	4.32	27.7	Nina
74	48	2.31	28.3	Nina
94	53	4.3	30.3	Nin
102	58	3.71	28.7	Nin

Hem afegit una variable més (sexe de l'infant). Ens demanem si aquestes 4 variables són capaces d'explicar o de predir la variable anterior.

ACP sobre la matriu de covariàncies

Exemple

La matriu de dades centrada seria:

$$\tilde{\mathbf{X}} = \begin{pmatrix} -3.000 & -0.578 & -0.881 & -0.056 \\ -12.000 & -3.278 & -1.481 & -3.256 \\ -4.000 & -2.478 & 0.779 & 2.644 \\ 7.000 & 0.222 & 1.889 & 6.944 \\ -14.000 & -5.778 & -0.421 & -2.356 \\ -1.000 & -0.778 & 0.689 & -1.856 \\ -7.000 & -0.778 & -1.321 & -1.256 \\ 13.000 & 4.222 & 0.669 & 0.744 \\ 21.000 & 9.222 & 0.079 & -1.556 \end{pmatrix}$$

ACP sobre la matriu de covariàncies

Exemple

La matriu de covariàncies serà:

$$\mathbf{S} = \frac{1}{9}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} = \begin{pmatrix} 119.333 & 43.133 & 6.148 & 10.878 \\ 43.133 & 17.193 & 1.148 & 1.169 \\ 6.148 & 1.148 & 1.111 & 2.422 \\ 10.878 & 1.169 & 2.422 & 8.818 \end{pmatrix}$$

Els valors propis de la matriu de covariàncies anterior són:

$$136.296 \quad 9.390 \quad 0.722 \quad 0.047$$

ACP sobre la matriu de covariàncies

Exemple

Els corresponents vectors propis (per columnes) associats als vectors propis anteriors són:

$$\begin{matrix} 0.935 & -0.031 & 0.250 & 0.248 \\ 0.340 & 0.342 & -0.665 & -0.571 \\ 0.047 & -0.241 & 0.572 & -0.782 \\ 0.084 & -0.908 & -0.411 & -0.016 \end{matrix}$$

Ens adonem que els vectors anteriors són ortogonals i estan normalitzats.

ACP sobre la matriu de covariàncies

Exemple

Així la matriu que ens canviarà de variables originals a components principals serà la transposada dels vectors propis:

$$\Lambda = \begin{pmatrix} 0.935 & 0.340 & 0.047 & 0.084 \\ -0.031 & 0.342 & -0.241 & -0.908 \\ 0.250 & -0.665 & 0.572 & -0.411 \\ 0.248 & -0.571 & -0.782 & -0.016 \end{pmatrix}$$

ACP sobre la matriu de covariàncies

Exemple

Les expressions de les components principals en funció de les variables originals són:

$$\begin{aligned} CP_1 &= 0.935X_1 + 0.34X_2 + 0.047X_3 + 0.084X_4 \\ CP_2 &= -0.031X_1 + 0.342X_2 - 0.241X_3 - 0.908X_4 \\ CP_3 &= 0.25X_1 - 0.665X_2 + 0.572X_3 - 0.411X_4 \\ CP_4 &= 0.248X_1 - 0.571X_2 - 0.782X_3 - 0.016X_4 \end{aligned}$$

ACP sobre la matriu de covariàncies

Exemple

Per tant, la matriu de dades en les noves variables serà:

$$\mathbf{Y} = \tilde{\mathbf{X}}\Lambda^\top = \tilde{\mathbf{X}}\mathbf{V} = \begin{pmatrix} -3.049 & 0.158 & -0.846 & 0.276 \\ -12.683 & 2.562 & -0.328 & 0.103 \\ -4.326 & -3.312 & 0.008 & -0.229 \\ 7.295 & -6.899 & -0.171 & 0.024 \\ -15.279 & 0.696 & 1.071 & 0.190 \\ -1.323 & 1.283 & 1.423 & -0.314 \\ -6.980 & 1.408 & -1.471 & -0.240 \\ 13.691 & 0.206 & 0.517 & 0.281 \\ 22.655 & 3.898 & -0.202 & -0.090 \end{pmatrix}$$

ACP sobre la matriu de covariàncies

Exemple

Si calculam la matriu de covariàncies de les components principals val:

$$\mathbf{S}_{CP} = \begin{pmatrix} 136.296 & -0.000 & 0.000 & 0.000 \\ -0.000 & 9.390 & -0.000 & -0.000 \\ 0.000 & -0.000 & 0.722 & 0.000 \\ 0.000 & -0.000 & 0.000 & 0.047 \end{pmatrix}$$

Igual que passava en l'exemple anterior, és diagonal, com esperàvem i en la diagonal hi surten els valors propis de la matriu de covariàncies de les dades originals.

ACP sobre la matriu de covariàncies

Exemple

Comprovem que la variabilitat s'ha conservat.

La variabilitat de les variables originals serà la suma dels valors de la diagonal de la matriu $\mathbf{S}$:

$$119.333 + 17.193 + 1.111 + 8.818 = 146.455.$$

La variabilitat de les components principals serà la suma dels valors de la diagonal de la matriu $\mathbf{S}_{CP}$:

$$136.296 + 9.39 + 0.722 + 0.047 = 146.455.$$

Podem observar que les dues variabilitats coincideixen.

ACP sobre la matriu de covariàncies

Exemple

La primera component principal (1a. columna de la matriu **CP**) hereta el $\frac{136.296}{146.455} \cdot 100\% = 93.064\%$ de la variabilitat total. Les dues primeres components principals hereten el $\frac{136.296+9.39}{146.455} \cdot 100\% = 99.475\%$ de la variabilitat total.

ACP sobre la matriu de covariàncies

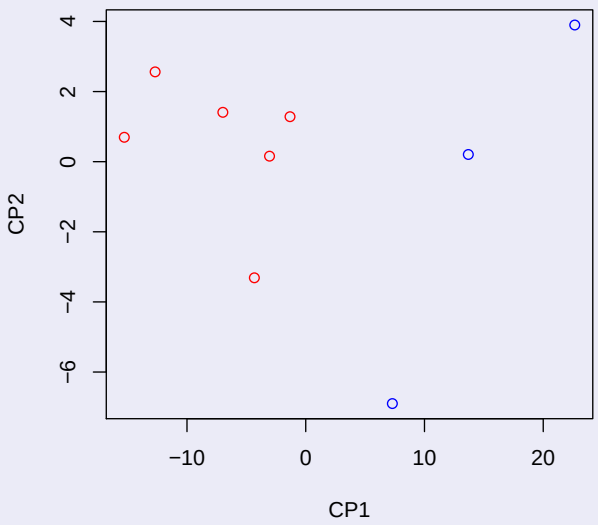
Exemple

Això ens fa pensar que si només tenim en compte les dues primeres components, podem fer un gràfic on hi estiguin representats tots els nens dibuixant de blau els nens i de vermell les nenes.

Comprovem que les dues primeres components separen bé els nens i les nens. Concloem que la nostra taula de dades "explica" la variable sexe.

ACP sobre la matriu de covariàncies

Exemple



Propietats de l'ACP sobre la matriu de covariàncies

Recordem que la matriu $\mathbf{S}$ és la matriu de covariàncies de les variables originals i $\mathbf{S}_{CP}$ és la matriu de covariàncies de les components principals.

La diagonal de la matriu $\mathbf{S}$ està formada per les variàncies de les variables originals s_i^2, $i = 1, \ldots, p$.

Definim la variància total de la nostra taula de dades com la suma de les variàncies o la traça de la matriu $\mathbf{S}$:

$$\text{Variància Total} = tr(\mathbf{S}) = \sum_{i=1}^{p} s_i^2.$$

Propietats de l'ACP sobre la matriu de covariàncies

- La variància de la i-èssima component principal és el valor propi i-èssim de la matriu de covariàncies $\mathbf{S}$: $\text{Var}(CP)_i = \lambda_i$.
- Es conserva la variància total. O sigui, la variància total de les variables originals i de les components principals és la mateixa:

$$\sum_{i=1}^{p} \text{var}(X_i) = \sum_{i=1}^{p} \text{var}(CP_i) = \sum_{i=1}^{p} \lambda_i.$$

Propietats de l'ACP sobre la matriu de covariàncies

- Les components principals són incorrelades. O, dit, en altres paraules, la seva matriu de covariàncies és diagonal:

$$\mathbf{S}_{CP} = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \lambda_p \end{pmatrix}$$

Propietats de l'ACP sobre la matriu de covariàncies

- Donada una taula de dades, definim la variància generalitzada com el determinant de la matriu de covariàncies. Aleshores, la variància generalitzada de les variables originals i de les components principals coincideix:

$$\det(\mathbf{S}) = \det(\mathbf{S}_{CP}) = \lambda_1 \cdots \lambda_p.$$

- La proporció de variància explicada per la component j-èssima és: $\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}$. Per tant, la variància explicada per les k primeres components val: $\frac{\sum_{i=1}^{k} \lambda_j}{\sum_{i=1}^{p} \lambda_i}$.

Propietats de l'ACP sobre la matriu de covariàncies

- Sigui $\mathbf{X}_i$ la i-èssima variable original. O sigui, la i-èssima columna de la matriu de dades $\mathbf{X}$. Sigui $\mathbf{CP}_j$ la j-èssima component principal. O sigui, la j-èssima columna de la matriu de dades $\mathbf{CP}$. Aleshores la covariància entre les variables (columnes) $\mathbf{X}_i$ i $\mathbf{CP}_j$ val:

$$\text{cov}(\mathbf{X}_i, \mathbf{CP}_j) = \lambda_j u_{ji},$$

  on u_{ji} és la i-èssima component del vector propi unitari $\mathbf{u}_j$ corresponent al valor propi λ_j.

- Seguin la mateixa notació anterior, la correlació entre $\mathbf{X}_i$ i $\mathbf{CP}_j$ val:

$$\text{cor}(\mathbf{X}_i, \mathbf{CP}_j) = \frac{\sqrt{\lambda_j} u_{ji}}{s_i}.$$

Propietats de l'ACP sobre la matriu de covariàncies

- En general si definim la matriu $\mathbf{S}_{X,CP}$ de components $s_{ij} = \text{cov}(\mathbf{X}_i, \mathbf{CP}_j)$, podem escriure:

$$\mathbf{S}_{X,CP} = \mathbf{V}\text{diag}(\lambda_1, \ldots, \lambda_p),$$

  on $\mathbf{V}$ és la matriu de vectors propis de la matriu de covariàncies $\mathbf{S}$.

- En general si definim la matriu $\mathbf{R}_{X,CP}$ de components $r_{ij} = \text{cor}(\mathbf{X}_i, \mathbf{CP}_j)$, podem escriure:

$$\mathbf{R}_{X,CP} = \mathbf{V}\text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p})\text{diag}\left(\frac{1}{s_1}, \ldots, \frac{1}{s_p}\right),$$

  on s_i és la desviació típica de la variable original $\mathbf{X}_i$.

Propietats de l'ACP sobre la matriu de covariàncies

La primera component principal seria la varietat lineal de dimensió 1 (una recta) que conserva la major variabilitat (anomenada inèrcia) del "núvol" de punts.

De la mateixa manera, les dues primeres components principals serien la varietat lineal de dimensió 2 (pla) que conserva la major variabilitat (anomenada inèrcia) del "núvol" de punts.

En general, les k primeres components principals serien la varietat lineal de dimensió k que conserva la major variabilitat (anomenada inèrcia) del "núvol" de punts.

Exemple

Exemple

Comprovem la relació anterior entre les variables originals i les components principals en l'exemple dels infants. Recordem que la matriu centrada de dades era:

$$\tilde{\mathbf{X}} = \begin{pmatrix} -3.0 & -0.6 & -0.9 & -0.1 \\ -12.0 & -3.3 & -1.5 & -3.3 \\ -4.0 & -2.5 & 0.8 & 2.6 \\ 7.0 & 0.2 & 1.9 & 6.9 \\ -14.0 & -5.8 & -0.4 & -2.4 \\ -1.0 & -0.8 & 0.7 & -1.9 \\ -7.0 & -0.8 & -1.3 & -1.3 \\ 13.0 & 4.2 & 0.7 & 0.7 \\ 21.0 & 9.2 & 0.1 & -1.6 \end{pmatrix}$$

Exemple

Exemple

La matriu de components principals era:

$$\mathbf{CP} = \begin{pmatrix} -3.049 & 0.158 & -0.846 & 0.276 \\ -12.683 & 2.562 & -0.328 & 0.103 \\ -4.326 & -3.312 & 0.008 & -0.229 \\ 7.295 & -6.899 & -0.171 & 0.024 \\ -15.279 & 0.696 & 1.071 & 0.190 \\ -1.323 & 1.283 & 1.423 & -0.314 \\ -6.980 & 1.408 & -1.471 & -0.240 \\ 13.691 & 0.206 & 0.517 & 0.281 \\ 22.655 & 3.898 & -0.202 & -0.090 \end{pmatrix}$$

Exemple

Exemple

La covariància entre les dues matrius anteriors val:

$$\text{cov}(\tilde{X}, \mathbf{CP}) = \begin{pmatrix} 127.503 & -0.289 & 0.180 & 0.012 \\ 46.349 & 3.210 & -0.480 & -0.027 \\ 6.397 & -2.263 & 0.413 & -0.036 \\ 11.426 & -8.524 & -0.296 & -0.001 \end{pmatrix}$$

Exemple

Exemple

Si fem $\mathbf{V}\text{diag}(\lambda_1, \ldots, \lambda_p)$ obtenim el mateix:

$$\begin{pmatrix} 0.935 & -0.031 & 0.250 & 0.248 \\ 0.340 & 0.342 & -0.665 & -0.571 \\ 0.047 & -0.241 & 0.572 & -0.782 \\ 0.084 & -0.908 & -0.411 & -0.016 \end{pmatrix} \cdot \begin{pmatrix} 136.296 & 0.000 & 0.000 & 0.000 \\ 0.000 & 9.390 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.722 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.047 \end{pmatrix}$$

$$= \text{cov}(\tilde{X}, \mathbf{CP})$$

Exemple

Exemple

La correlacions entre les dues matrius anteriors val:

$$\text{cor}(\tilde{X}, \mathbf{CP}) = \begin{pmatrix} 1.000 & -0.009 & 0.019 & 0.005 \\ 0.957 & 0.253 & -0.136 & -0.030 \\ 0.520 & -0.701 & 0.461 & -0.160 \\ 0.330 & -0.937 & -0.118 & -0.001 \end{pmatrix}$$

Exemple

Exemple

Si fem $\mathbf{V}\text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p})\text{diag}\left(\frac{1}{s_1}, \ldots, \frac{1}{s_p}\right)$ obtenim el mateix, o sigui el producte de les tres matrius següents:

$$\mathbf{V} = \begin{pmatrix} 0.935 & -0.031 & 0.250 & 0.248 \\ 0.340 & 0.342 & -0.665 & -0.571 \\ 0.047 & -0.241 & 0.572 & -0.782 \\ 0.084 & -0.908 & -0.411 & -0.016 \end{pmatrix}$$

$$\text{diag}(\sqrt{\lambda}) = \begin{pmatrix} 11.675 & 0.000 & 0.000 & 0.000 \\ 0.000 & 3.064 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.850 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.216 \end{pmatrix}$$

Exemple

Exemple

$$\text{diag}(1/s) = \begin{pmatrix} 0.092 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.241 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.949 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.337 \end{pmatrix}$$

Exemple

Exemple

Vegem quin percentatge de variabilitat tenim si consideram només les primeres components principals:

Variables	Varianza Explicada
$\mathbf{CP}_1$	$136.296/146.455 = 0.931$
$\mathbf{CP}_{1,2}$	$145.686/146.455 = 0.995$
$\mathbf{CP}_{1,2,3}$	$146.408/146.455 = 0.9997$
$\mathbf{CP}_{1,2,3,4}$	1

En aquest exemple, si només tenguéssim en compte les dues primeres components explicaríem el 99.48% de la variabilitat total.

ACP sobre la matriu de correlacions

Per realitzar l'ACP sobre la matriu de correlacions, es fa de la mateixa manera que l'ACP sobre la matriu de covariàncies però en lloc de fer servir aquesta matriu es fa servir la matriu de correlacions $\mathbf{R}$.

O sigui, es calculen els valors propis λ_i de la matriu $\mathbf{R}$ juntament amb la matriu de vectors propis $\mathbf{V}$.

Això és equivalent a aplicar l'ACP sobre la matriu de covariàncies però en lloc de fer servir la matriu centrada original, es fa servir la matriu de dades tipificada $\mathbf{Z}$.

Per tant, totes les propietats enunciades sobre la matriu de covariàncies serien vàlides per la matriu de correlacions substituint simplement la matriu $\mathbf{S}$ per la matriu $\mathbf{R}$.

ACP sobre la matriu de correlacions

Exemple

Anem a repetir l'exemple dels infants però fent una ACP sobre la matriu de correlacions.
Recordem les dades:

x_1	x_2	x_3	x_4	Sexe
78	48.2	2.75	29.5	Nina
69	45.5	2.15	26.3	Nina
77	46.3	4.41	32.2	Nina
88	49	5.52	36.5	Nin
67	43	3.21	27.2	Nina
80	48	4.32	27.7	Nina
74	48	2.31	28.3	Nina
94	53	4.3	30.3	Nin
102	58	3.71	28.7	Nin

ACP sobre la matriu de correlacions

La matriu de dades tipificada serà:

$$\mathbf{R} = \begin{pmatrix} -0.275 & -0.139 & -0.836 & -0.019 \\ -1.099 & -0.791 & -1.405 & -1.096 \\ -0.366 & -0.598 & 0.739 & 0.891 \\ 0.641 & 0.054 & 1.792 & 2.339 \\ -1.282 & -1.393 & -0.400 & -0.793 \\ -0.092 & -0.188 & 0.654 & -0.625 \\ -0.641 & -0.188 & -1.254 & -0.423 \\ 1.190 & 1.018 & 0.635 & 0.251 \\ 1.922 & 2.224 & 0.075 & -0.524 \end{pmatrix}$$

ACP sobre la matriu de correlacions

Exemple

La matriu de correlacions $\mathbf{R}$ de la matriu $\mathbf{X}$ o la matriu de covariàncies de la matriu $\mathbf{Z}$ serà:

$$\mathbf{R} = \begin{pmatrix} 1.000 & 0.952 & 0.534 & 0.335 \\ 0.952 & 1.000 & 0.263 & 0.095 \\ 0.534 & 0.263 & 1.000 & 0.774 \\ 0.335 & 0.095 & 0.774 & 1.000 \end{pmatrix}$$

Els valors propis de la matriu de correlacions anterior són:

$$2.503 \quad 1.286 \quad 0.208 \quad 0.003$$

ACP sobre la matriu de correlacions

Exemple

Els corresponents vectors propis (per columnes) associats als vectors propis anteriors són:

$$\begin{matrix} -0.579 & 0.352 & -0.014 & 0.735 \\ -0.482 & 0.566 & 0.171 & -0.647 \\ -0.506 & -0.444 & -0.712 & -0.199 \\ -0.420 & -0.599 & 0.681 & -0.031 \end{matrix}$$

Ens adonem que els vectors anteriors són ortogonals i estan normalitzats.

ACP sobre la matriu de correlacions

Exemple

Així la matriu que ens canviarà de variables originals a components principals serà la transposada dels vectors propis:

$$\Lambda = \begin{pmatrix} -0.579 & -0.482 & -0.506 & -0.420 \\ 0.352 & 0.566 & -0.444 & -0.599 \\ -0.014 & 0.171 & -0.712 & 0.681 \\ 0.735 & -0.647 & -0.199 & -0.031 \end{pmatrix}$$

ACP sobre la matriu de correlacions

Exemple

Les expressions de les components principals en funció de les variables originals tipificades són:

$$\begin{aligned} CP_1 &= -0.579Z_1 - 0.482Z_2 - 0.506Z_3 - 0.42Z_4 \\ CP_2 &= 0.352Z_1 + 0.566Z_2 - 0.444Z_3 - 0.599Z_4 \\ CP_3 &= -0.014Z_1 + 0.171Z_2 - 0.712Z_3 + 0.681Z_4 \\ CP_4 &= 0.735Z_1 - 0.647Z_2 - 0.199Z_3 - 0.031Z_4 \end{aligned}$$

ACP sobre la matriu de correlacions

Exemple

Per tant, la matriu de dades en les noves variables serà:

$$\mathbf{Y} = \mathbf{Z}\Lambda^\top = \mathbf{ZV} = \begin{pmatrix} 0.657 & 0.207 & 0.563 & 0.056 \\ 2.189 & 0.446 & 0.134 & 0.018 \\ -0.249 & -1.329 & -0.017 & -0.058 \\ -2.287 & -1.941 & 0.316 & 0.007 \\ 1.949 & -0.587 & -0.476 & 0.063 \\ 0.075 & -0.054 & -0.922 & -0.057 \\ 1.274 & 0.478 & 0.582 & -0.087 \\ -1.606 & 0.563 & -0.124 & 0.082 \\ -2.002 & 2.216 & -0.056 & -0.024 \end{pmatrix}$$

ACP sobre la matriu de correlacions

Exemple

Si calculam la matriu de correlacions de les components principals val:

$$\mathbf{R}_{CP} = \begin{pmatrix} 1.000 & 0.000 & 0.000 & -0.000 \\ 0.000 & 1.000 & 0.000 & -0.000 \\ 0.000 & 0.000 & 1.000 & 0.000 \\ -0.000 & -0.000 & 0.000 & 1.000 \end{pmatrix}$$

Surt la matriu diagonal, fet que posa de manifest que les components principals són incorrelades.

ACP sobre la matriu de correlacions

Exemple

La primera component principal (1a. columna de la matriu **CP**) hereta el $\frac{2.503}{4} \cdot 100\% = 62.575\%$ de la variabilitat total. Les dues primeres components principals hereten el $\frac{2.503+1.286}{4} \cdot 100\% = 94.732\%$ de la variabilitat total.

ACP sobre la matriu de correlacions

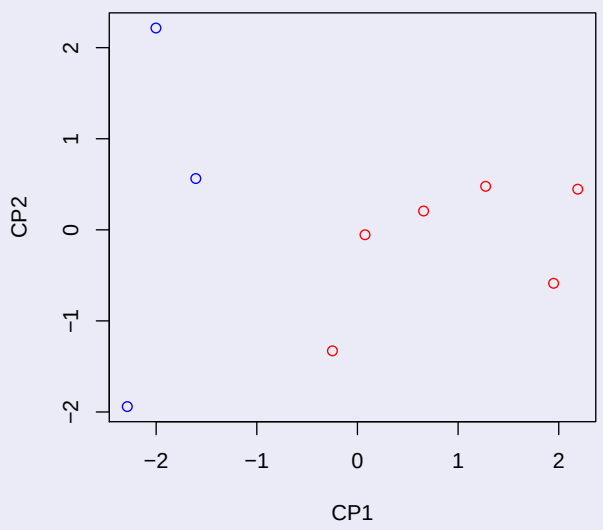
Exemple

Això ens fa pensar que si només tenim en compte les dues primeres components, podem fer un gràfic on hi estiguin representats tots els nens dibuixant de blau els nens i de vermell les nenes.

Comprovem que les dues primeres components separen bé els nens i les nens. Concloem que la nostra taula de dades "explica" la variable sexe.

ACP sobre la matriu de correlacions

Exemple



Etapes d'un ACP

- Primera etapa: decidir si es realitza l'ACP damunt les dades brutes centrades (matriu de covariàncies) o sobre les dades tipificades (matriu de correlacions).
  - Quan les variables originals **X** estan en unitats distintes, convé aplicar l'ACP de correlacions. Si estan en les mateixes unitats, ambdues alternatives són vàlides.
  - Si les diferències entre les variàncies són informatives i volem tenir-les en compte en l'anàlisi, no hem d'estandaritzar les variables i aplicar l'ACP de covariàncies.

Etapes d'un ACP

- Segona etapa: reducció de la dimensionalitat. Hem de decidir quantes components retenim. La quantitat de variància retenguda serà:

Comp.	Valor propi	Quantitat retinguda
Cp_1	λ_1	$\lambda_1/\sum_{i=1}^{p}\lambda_i$
Cp_2	λ_2	$(\lambda_1+\lambda_2)/\sum_{i=1}^{p}\lambda_i$
Cp_3	λ_3	$(\lambda_1+\lambda_2+\lambda_3)/\sum_{i=1}^{p}\lambda_i$
...	...	...
Cp_p	λ_p	$(\lambda_1+\ldots+\lambda_p)/\sum_{i=1}^{p}\lambda_i=1$

Etapes d'un ACP

- Segona etapa: Per decidir el nombre de components retingudes, hi ha dos mètodes:
  - Seleccionar components fins cobrir una proporció determinada de variància, com el 80% o el 90%.
  - Mètode de la mitjana aritmètica. Se retenen totes aquelles components $\mathbf{CP}_i$ que compleixin que $\lambda_i \geqslant \overline{\lambda} = \frac{\sum_{i=1}^{p}\lambda_i}{p}$. En el cas de l'ACP de correlacions, la condició anterior és $\lambda_i \geqslant 1$.

Etapes d'un ACP

Exemple

En l'exemple dels infants,

- si aplicam el primer mètode per decidir el nombre de components retingudes, fent l'ACP de correlacions, si només elegim la primera component, ja cobrim el 62.575% de la variància total. Si elegim les dues primeres, cobrim el 94.732% de la variància total.
- si aplicam el mètode de la mitjana aritmètica, hauríem de retenir dues components ja que els valors propis de la matriu de correlacions $\mathbf{R}$ eren
  2.503 1.286 0.208 0.003 .

Descomposició en valors singulars

Donada una matriu de dades $\mathbf{X}$ de dimensions $n \times p$, on $n \geqslant p$ i de rang p, es pot descompondre en producte de tres matrius:

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top,$$

on

- $\mathbf{U}$ és una matriu ortogonal $n \times p$ que té per columnes els p vectors propis de la matriu $\mathbf{X}\mathbf{X}^\top$ associats als p valors propis no nuls.
- Σ és una matriu diagonal $p \times p$ que té per diagonal les arrels quadrades dels valors propis de la matriu $\mathbf{X}^\top\mathbf{X}$.
- $\mathbf{V}$ és una matriu ortogonal $p \times p$ que té per columnes els vectors propis de la matriu $\mathbf{X}^\top\mathbf{X}$ associats als p valors propis no nuls.

Descomposició en valors singulars (SVD)

Exemple

Considerem la matriu $\mathbf{X}$ com la matriu de dades centrada de l'exemple dels infants.

La matriu $\mathbf{X}^\top\mathbf{X}$ val:

$$\mathbf{X}^\top\mathbf{X} = \begin{pmatrix} 1074.000 & 388.200 & 55.330 & 97.900 \\ 388.200 & 154.736 & 10.330 & 10.521 \\ 55.330 & 10.330 & 9.995 & 21.795 \\ 97.900 & 10.521 & 21.795 & 79.362 \end{pmatrix}$$

Descomposició en valors singulars (SVD)

Exemple

La matriu $\mathbf{X}\mathbf{X}^\top$ val: (mostram només les 4 primeres columnes, pensau que és 10×10)

$$\mathbf{X}^\top\mathbf{X} = \begin{pmatrix} 10.113 & 39.380 & 12.598 & -23.179 \\ 39.380 & 167.536 & 46.359 & -110.134 \\ 12.598 & 46.359 & 29.739 & -8.715 \\ -23.179 & -110.134 & -8.715 & 100.843 \\ 45.840 & 195.231 & 63.759 & -116.437 \\ 2.945 & 19.570 & 1.557 & -18.757 \\ 22.683 & 92.594 & 25.578 & -60.387 \\ -42.070 & -173.254 & -59.972 & 98.371 \\ -68.311 & -277.281 & -110.903 & 138.396 \end{pmatrix}$$

Descomposició en valors singulars (SVD)

Exemple

Els valors propis de la matriu $\mathbf{X}^\top\mathbf{X}$ són:

$$1226.665 \quad 84.511 \quad 6.498 \quad 0.419$$

Per tant, la matriu Σ serà:

$$\Sigma = \begin{pmatrix} 35.024 & 0.000 & 0.000 & 0.000 \\ 0.000 & 9.193 & 0.000 & 0.000 \\ 0.000 & 0.000 & 2.549 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.647 \end{pmatrix}$$

Descomposició en valors singulars (SVD)

Exemple

La matriu $\mathbf{U}$ serà la següent matriu 10×4:

$$\mathbf{U} = \begin{pmatrix} -0.087 & 0.017 & 0.332 & 0.426 \\ -0.362 & 0.279 & 0.129 & 0.159 \\ -0.124 & -0.360 & -0.003 & -0.354 \\ 0.208 & -0.750 & 0.067 & 0.037 \\ -0.436 & 0.076 & -0.420 & 0.294 \\ -0.038 & 0.140 & -0.558 & -0.485 \\ -0.199 & 0.153 & 0.577 & -0.371 \\ 0.391 & 0.022 & -0.203 & 0.434 \\ 0.647 & 0.424 & 0.079 & -0.140 \end{pmatrix}$$

Descomposició en valors singulars (SVD)

Exemple

La matriu $\mathbf{V}$ serà la següent matriu 4×4:

$$\mathbf{U} = \begin{pmatrix} 0.935 & -0.031 & -0.250 & 0.248 \\ 0.340 & 0.342 & 0.665 & -0.571 \\ 0.047 & -0.241 & -0.572 & -0.782 \\ 0.084 & -0.908 & 0.411 & -0.016 \end{pmatrix}$$

Es pot comprovar que $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top$.

Relació ACP amb SVD

Considerem una matriu de dades $\mathbf{X}$ $n \times p$ que pot ésser centrada (ACP de covariàncies) o tipificada (ACP de correlacions).

Si considerem la seva SVD, $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top$, tenim que les components principals, $\mathbf{Y}$, valen $\mathbf{CP} = \mathbf{U}\Sigma$.

La prova és molt senzilla. Recordem que les components principals valien: $\mathbf{CP} = \mathbf{XV}$, on $\mathbf{V}$ era la matriu de vectors propis de la matriu de covariàncies $\mathbf{S} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$. Ara bé, aquesta matriu coincidirà amb la matriu de vectors propis de la matriu $\mathbf{X}^\top\mathbf{X}$ ja que els vectors propis de la matriu anterior i de la matriu de covariàncies $\mathbf{S}$ són els mateixos.
Per tant,

$$\mathbf{Y} = \mathbf{XV} = \mathbf{U}\Sigma\mathbf{V}^\top\mathbf{V} = \mathbf{U}\Sigma,$$

ja que la matriu $\mathbf{V}$ és ortogonal.

Relació ACP amb SVD

Teorema

El producte escalar de dues files de la matriu de dades $\mathbf{X}$ *coincideix amb el producte escalar de dues files de la matriu de components principals* $\mathbf{Y}$.

Prova. El producte escalar de dues files de la matriu $\mathbf{X}$ ve donada per la matriu $\mathbf{XX}^\top$ però:

$$\mathbf{XX}^\top = \mathbf{YV}^\top\mathbf{VY}^\top = \mathbf{YY}^\top,$$

aquesta última matriu ens dóna el producte escalar de dues files de la matriu de components principals.

Relació ACP amb SVD

Teorema

La distància euclídea de dues files de la matriu $\mathbf{X}$ *coincideix amb la distància euclídea de la matriu de components principals* $\mathbf{Y}$. *O sigui, la distància entre els individus respecte les variables originals i respecte les components principals se conserva.*

Prova. Siguin $\mathbf{f}_i$ i $\mathbf{f}_j$ dues files de la matriu $\mathbf{X}$ i siguin $\mathbf{g}_i$ i $\mathbf{g}_j$ dues files de la matriu $\mathbf{Y}$. Aleshores:

$$\|\mathbf{f}_i - \mathbf{f}_j\|^2 = (\mathbf{f}_i - \mathbf{f}_j)^\top(\mathbf{f}_i - \mathbf{f}_j) = \mathbf{f}_i^\top\mathbf{f}_i - 2\mathbf{f}_i^\top\mathbf{f}_j + \mathbf{f}_j^\top\mathbf{f}_j.$$

Ara bé, fent servir el teorema anterior tenim que:

$$\mathbf{f}_i^\top\mathbf{f}_i = \mathbf{g}_i^\top\mathbf{g}_i,\ \mathbf{f}_i^\top\mathbf{f}_j = \mathbf{g}_i^\top\mathbf{g}_j,\ \mathbf{f}_j^\top\mathbf{f}_j = \mathbf{g}_j^\top\mathbf{g}_j.$$

Concloem, doncs: $\|\mathbf{f}_i - \mathbf{f}_j\|^2 = \|\mathbf{g}_i - \mathbf{g}_j\|^2$.

Relació ACP amb SVD

Teorema

Sigui $\mathbf{X}$ *la nostra matriu de dades originals centrada o tipificada. Sigui* $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top$ *la seva SVD. Considerem la matriu* $\mathbf{V}_2 = \Sigma\mathbf{V}^\top$*. Aleshores el producte escalar de dues columnes de la matriu* $\mathbf{X}$ *i de dues columnes de la matriu* $\mathbf{V}_2$ *és el mateix.*

Prova. El producte escalar de dues columnes de la matriu $\mathbf{X}$ ho dóna la matriu $\mathbf{X}^\top\mathbf{X}$. Tendrem:

$$\mathbf{X}^\top\mathbf{X} = \mathbf{V}\Sigma\mathbf{U}^\top\mathbf{U}\Sigma\mathbf{V}^\top = \mathbf{V}\Sigma(\mathbf{V}\Sigma)^\top = \mathbf{V}_2^\top\mathbf{V}_2,$$

matriu que dóna el producte escalar de dues columnes de la matriu $\mathbf{V}_2$.

Relació ACP amb SVD

Teorema

Sigui $\mathbf{X}$ *la nostra matriu de dades originals centrada o tipificada. Sigui* $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top$ *la seva SVD. Considerem la matriu* $\mathbf{V}_2 = \Sigma\mathbf{V}^\top$*. Aleshores la correlació entre dues columnes de la matriu* $\mathbf{X}$ *i la correlació entre dues columnes de la matriu* $\mathbf{V}_2 = \Sigma\mathbf{V}^\top$ *és la mateixa. O sigui, la correlació entre dues variables originals (dues columnes de la matriu* $\mathbf{X}$*) i dues columnes de la matriu* $\mathbf{V}_2$ *és la mateixa.*

Per la prova, basta aplicar el teorema anterior.

Biplots

Per representar el resultat d'un ACP gràficament es fa servir el que s'anomena un biplot.

Un biplot és un gràfic bidimensional on es representa en el mateix gràfic els individus i les variables originals.

El biplot només té significat si la variabilitat explicada per les dues primeres components principals és alta, posem d'un 85% o més.

En un biplot hi ha 4 eixos coordenats: dos fan referència als individus (eixos d'abaix i de l'esquerra) i dos fan referència a les variables originals. (eixos de dalt i de la dreta)

Biplots

Els eixos que fan referència als individus són els corresponents a les dues components principals.

Per tant, les coordenades dels individus seran els valors de les dues primeres components estandaritzades (ja veurem què significa això amb un exemple).

Hem vist que la distància euclídea entre els valors originals dels individus (files de la matriu de dades $\mathbf{X}$) i els valors d'aquests mateixos individus respecte les components principals se conserva. Això significa que si les dues primeres components principals expliquen molta variabilitat, la distància entre individus que es veurà al biplot serà aproximadament la distància entre els individus segons les variables originals.

Biplots

Les coordenades corresponents a les variables originals (columnes de la matriu de dades $\mathbf{X}$) les ens dóna les dues primeres components en columnes de la matriu $\mathbf{V}_2 = \Sigma\mathbf{V}^\top$ ja que hem vist que la correlació entre les variables originals i les columnes de la matriu anterior és la mateixa.

Biplots

El gràfic de les variables originals es va mitjançant vectors. La interpretació que s'ha de fer és la següent: si l'angle entre dues variables originals és petit, significa que el cosinus d'aquest angle serà gran però aquest cosinus és la correlació entre les dues variables. Per tant, hi haurà molta correlació entre les variables. En canvi, si l'angle entre les dues variables està proper a un angle recte, la correlació entre aquestes variables és quasi nul·la.

Les llargades dels vectors són les variàncies de les variables originals. Per tant, com més llarg tengui un vector, la variable corresponent tendrà més dispersió.

Biplots

Exemple

Considerem l'exemple dels infants fent l'ACP de covariàncies. Recordem que les dues primeres components principals explicaven el 99.475% de la variabilitat total.
Les components principals eren les següents:

$$\mathbf{Y} = \begin{pmatrix} -3.049 & 0.158 & -0.846 & 0.276 \\ -12.683 & 2.562 & -0.328 & 0.103 \\ -4.326 & -3.312 & 0.008 & -0.229 \\ 7.295 & -6.899 & -0.171 & 0.024 \\ -15.279 & 0.696 & 1.071 & 0.190 \\ -1.323 & 1.283 & 1.423 & -0.314 \\ -6.980 & 1.408 & -1.471 & -0.240 \\ 13.691 & 0.206 & 0.517 & 0.281 \\ 22.655 & 3.898 & -0.202 & -0.090 \end{pmatrix}$$

Biplots

Exemple

Quan representem els individus (nens i nenes) en el biplot representarem les dues primeres columnes de la matriu anterior tipificada.
Això significa que dividirem cada element de la matriu per la norma euclídea de la columna. Així la norma euclídea de la primera columna de la matriu anterior val:

$$\sqrt{3.049^2 + \cdots + 22.655^2} = 35.024.$$

Fent el mateix amb la segona columna,

$$\sqrt{0.158^2 + \cdots + 3.898^2} = 9.193.$$

Biplots

Exemple

Les coordenades dels nens i nenes en el biplot seran:

$$\begin{pmatrix} -3.049 & 0.158 \\ -12.683 & 2.562 \\ -4.326 & -3.312 \\ 7.295 & -6.899 \\ -15.279 & 0.696 \\ -1.323 & 1.283 \\ -6.980 & 1.408 \\ 13.691 & 0.206 \\ 22.655 & 3.898 \end{pmatrix} \begin{pmatrix} \frac{1}{35.024} & 0 \\ 0 & \frac{1}{9.193} \end{pmatrix} =$$

Biplots

Exemple

Les coordenades dels nens i nenes en el biplot seran:

$$\begin{pmatrix} -0.087 & 0.017 \\ -0.362 & 0.279 \\ -0.124 & -0.360 \\ 0.208 & -0.750 \\ -0.436 & 0.076 \\ -0.038 & 0.140 \\ -0.199 & 0.153 \\ 0.391 & 0.022 \\ 0.647 & 0.424 \end{pmatrix}$$

Biplots

Exemple

Passem ara a calcular les coordenades de les variables originals.

Recordem que la matriu de vectors propis de la matriu de covariàncies $\mathbf{S}$ era:

$$\begin{pmatrix} 0.935 & -0.031 & 0.250 & 0.248 \\ 0.340 & 0.342 & -0.665 & -0.571 \\ 0.047 & -0.241 & 0.572 & -0.782 \\ 0.084 & -0.908 & -0.411 & -0.016 \end{pmatrix}$$

Biplots

Exemple

Si fem la SVD de la matriu de dades centrada $\tilde{\mathbf{X}}$, la matriu Σ serà:

$$\Sigma = \begin{pmatrix} 35.024 & 0.000 & 0.000 & 0.000 \\ 0.000 & 9.193 & 0.000 & 0.000 \\ 0.000 & 0.000 & 2.549 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.647 \end{pmatrix}$$

Per tant, la matriu $\mathbf{V}_2 = \Sigma\mathbf{V}^\top$ serà:

$$\mathbf{V}_2 = \begin{pmatrix} 32.764 & 11.910 & 1.644 & 2.936 \\ -0.283 & 3.143 & -2.216 & -8.345 \\ 0.636 & -1.694 & 1.459 & -1.047 \\ 0.161 & -0.370 & -0.507 & -0.010 \end{pmatrix}$$

Biplots

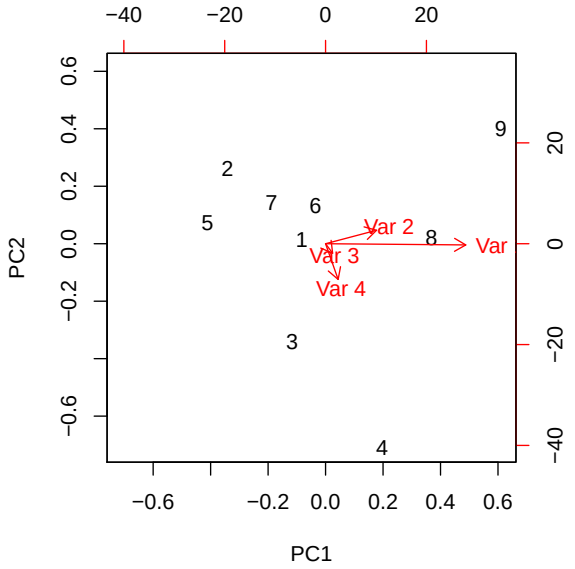
Exemple

Per tant, les coordenades de les 4 variables originals són:

- $\mathbf{x}_1$: Edat en dies. (32.764 -0.283)
- $\mathbf{x}_2$: Alçada en néixer. (11.910 3.143)
- $\mathbf{x}_3$: Pes en néixer (1.644 -2.216)
- $\mathbf{x}_4$: Augment en tant per cent del seu pes actual respecte del seu pes en néixer (2.936 -8.345)

Biplots

A continuació mostram el biplot:



Biplots

Comprovem les coordenades dels punts i de les variables.

Veiem també que entre les parelles de variables $(\mathbf{x}_1, \mathbf{x}_2)$ i $(\mathbf{x}_3, \mathbf{x}_4)$ hi ha bastanta correlació i entre les parelles $(\mathbf{x}_2, \mathbf{x}_3)$ i $(\mathbf{x}_2, \mathbf{x}_4)$ n'hi ha poca com podem comprovar si calculem la correlació entre les variables originals:

$$\begin{pmatrix} 1.000 & 0.952 & 0.534 & 0.335 \\ 0.952 & 1.000 & 0.263 & 0.095 \\ 0.534 & 0.263 & 1.000 & 0.774 \\ 0.335 & 0.095 & 0.774 & 1.000 \end{pmatrix}$$

Observem que les variables amb més dispersió són l'edat de l'infant i l'alçada en néixer.