Capítulo 2

Análisis de componentes principales

2.1. Teoría

Sea $\mathbf{X}$ una muestra de datos (de la variable/población $\vec{x} \rightsquigarrow \mathcal{N}(\vec{\mu}, \Sigma)$) donde el elemento x_{ij} corresponde al individuo i y a la variable continua j ($i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, p\}$). Sea Σ la matriz $p \times p$ de covarianzas de X (en rigor, sería la estimación máximo-verosímil de la Σ antes mencionada).

2.1.1. Objetivo

Nuestro objetivo es explicar el comportamiento conjunto de muchas variables a partir de unas pocas: reducción dimensional. Por comportamiento entendemos la variabilidad, medida mediante la varianza.
El análisis de componentes principales pretende hallar direcciones ortogonales (y, por tanto, linealmente independientes, luego estadísticamente independientes si hay gausianidad) en las que se maximice la varianza de $\mathbf{X}$.

2.1.2. Primera componente principal

Sea $\vec{a} \in \mathbb{R}^p$. Pretendemos que $y_1 := \vec{a}' (\vec{x} - \vec{\mu})$ (primera componente principal) tenga varianza máxima. Trabajaremos con variables centradas (con esperanza nula: $\mathcal{E}[y_1] = 0$) porque nos simplifican la notación y el tratamiento. Por las propiedades de la varianza,

$$\text{máx}\, \text{Var}[y_1] = \text{máx}\, \vec{a}' \, \Sigma \, \vec{a}$$

Si no se restringe $\vec{a}$ eso vale infinito, luego buscaremos sólo entre los $\vec{a}$ con norma uno.

$$\underset{\|\vec{a}\|=1}{\text{máx}}\ \vec{a}' \, \Sigma \, \vec{a}$$

Como Σ es simétrica, entonces existen Λ diagonal y $\mathbf{U}$ ortogonal tales que $\Sigma = \mathbf{U}\,\Lambda\,\mathbf{U}'$. Sean $\Lambda =: \mathrm{diag}[\lambda_1, \ldots, \lambda_p]$, para los que, sin pérdida de generalidad, supondremos $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$; como Σ es semidefinida positiva, entonces $\lambda_i \geq 0, \forall\, i$. Sea $\vec{b} := \mathbf{U}'\,\vec{a}$. Nótese que $\|\vec{b}\| = \|\vec{a}\|$ porque (intuitivamente) $\mathbf{U}$ representa una rotación y porque (analíticamente) $\|\vec{b}\|^2 = \vec{b}'\,\vec{b} = \vec{a}'\,\mathbf{U}\,\mathbf{U}'\,\vec{a} = \vec{a}'\,\mathbf{I}\,\vec{a} = \vec{a}'\,\vec{a}$.

$$\max_{\|\vec{a}\|=1} \vec{a}'\,\Sigma\,\vec{a} = \max_{\|\vec{b}\|=1} \vec{b}'\,\Lambda\,\vec{b} = \sum_{i=1}^{p} \lambda_i\, b_i^2 \leq \sum_{i=1}^{p} \lambda_1\, b_i^2 \leq \lambda_1 \sum_{i=1}^{p} b_i^2 = \lambda_1$$

donde $\lambda_1 = \max_i \lambda_i$. Entonces, escogiendo $\vec{a}$ como el vector propio (autovector) asociado al valor propio (autovalor) mayor de Σ, se alcanza la cota. (Es decir, se escoge $\vec{b} := (1, 0, \ldots, 0)$, suponiendo que la primera columna de $\mathbf{U}$ es aquel autovector.)

Nótese que el máximo se alcanza en distintos vectores. Dejando aparte el caso de que el autovalor máximo tenga multiplicidad mayor que 1, téngase en cuenta que, si $\vec{a}$ es una solución, también lo es $-\vec{a}$. Recuérdese esto a la hora de interpretar los signos de los coeficientes de las componentes principales.

La primera componente principal estará, por tanto, relacionada con el autovector asociado al mayor autovalor de $\mathbf{\Sigma}$. Para hallar la segunda componente principal, $y_2 := \vec{a}'(\vec{x} - \vec{\mu})$, donde $\vec{a}$ es distinto de la de antes, aparte de maximizar su varianza interesa que su covarianza con y_1 sea nula, de forma que incorpore información no redundante en la medida de lo posible. Sean

$$y_1 =: \vec{u_1}'\,(\vec{x} - \vec{\mu}) \qquad y_2 = \vec{a}'\,(\vec{x} - \vec{\mu})$$

Por ser centradas,

$$\begin{aligned}
\mathrm{Cov}\,[y1, y2] &= \mathcal{E}\,[y_1\, y_2] \\
&= \mathcal{E}\left[\{\vec{u_1}'\,(\vec{x} - \vec{\mu})\}\,\{\vec{a}'\,(\vec{x} - \vec{\mu})\}\right] \\
&= \vec{u}_1'\,\mathcal{E}\,[(\vec{x} - \vec{\mu})\,(\vec{x} - \vec{\mu})]\;\vec{a} \\
&= \vec{u}_1'\,\mathbf{\Sigma}\,\vec{a}
\end{aligned}$$

Entonces,

$$\begin{aligned}
0 &= \mathrm{Cov}\,[y1, y2] \\
&= \vec{u}_1'\,\mathbf{\Sigma}\,\vec{a} \\
&= \vec{u}_1'\,\mathbf{U}\,\mathbf{\Lambda}\,\mathbf{U}'\,\vec{a} \\
&= \vec{u}_1'\,(\vec{u}_1, \ldots, \vec{u}_p)\,\mathbf{\Lambda}\,\mathbf{U}'\,\vec{a} \\
&= (1, 0, \ldots, 0)\,\mathbf{\Lambda}\,\mathbf{U}'\,\vec{a} \\
&= (\lambda_1, 0, \ldots, 0)\,\mathbf{U}'\,\vec{a} \\
&= \lambda_1\,\vec{u}_1'\,\vec{a} \\
&\Rightarrow\; \vec{u}_1'\,\vec{a} = 0 \text{ pues } \lambda_1 > 0
\end{aligned}$$

Luego habrá que maximizar

$$\vec{a}' \, \mathbf{\Sigma} \, \vec{a} = \vec{a}' \, \mathbf{U} \, \mathbf{\Lambda} \, \mathbf{U}' \, \vec{a} = \vec{b}' \, \mathbf{\Lambda} \, \vec{b} \leq \sum_i \lambda_2 \, b_i^2 = \lambda_2$$

donde la desigualdad se justifica porque

$$\mathbf{U}' \, \vec{a} = \begin{pmatrix} \vec{u}_1' \\ \vec{u}_2' \\ \vdots \\ \vec{u}_p' \end{pmatrix} \vec{a} = \begin{pmatrix} \vec{u}_1' \, \vec{a} \\ \vec{u}_2' \, \vec{a} \\ \vdots \\ \vec{u}_p' \, \vec{a} \end{pmatrix} = \begin{pmatrix} \vec{0} \\ \vec{u}_2' \, \vec{a} \\ \vdots \\ \vec{u}_p' \, \vec{a} \end{pmatrix}$$

2.1.3. Siguientes componentes principales

De forma similar, se puede definir y_i (la i-ésima componente principal) $i \in \{1, \ldots, p\}$) a partir del autovector $\vec{u}_i$ asociado al autovalor λ_i, el i-ésimo ordenando de forma decreciente. Así, $y_i := \vec{u}_i \, (\vec{x} - \vec{\mu})$ y $\operatorname{Var}[y_i] = \lambda_i$. En forma matricial,

$$\vec{y} = \mathbf{U}' \, (\vec{x} - \vec{\mu})$$

que podemos invertir de la siguiente manera:

$$\vec{x} - \vec{\mu} = \mathbf{U} \, \vec{y} = y_1 \, \vec{u}_1 + \cdots + y_p \, \vec{u}_p$$

2.1.4. Elección del número de componentes

En la práctica nos quedaremos con q componentes ($q < p$), con lo cual la reconstrucción de $\vec{x}$ no será perfecta, pues se emplearía

$$\vec{x} - \vec{\mu} \approx \vec{x}_q - \vec{\mu}_q = \mathbf{U}_q \, \vec{y} = y_1 \, \vec{u}_1 + \cdots + y_q \, \vec{u}_q$$

Para decidir el valor de q podemos tener en cuenta la proporción de varianza global explicada por las q primeras componentes principales, que sería

$$\frac{\operatorname{tr} \operatorname{Var}[\vec{x}_q]}{\operatorname{tr} \operatorname{Var}[\vec{x}]} = \frac{\operatorname{tr} \operatorname{Var}[(y)_{i=1}^q]}{\operatorname{tr} \operatorname{Var}[\vec{y}]_{i=1}^p} = \frac{\operatorname{tr} \operatorname{diag}[\lambda_1, \ldots, \lambda_q]}{\operatorname{tr} \mathbf{\Sigma}} = \frac{\sum_{i=1}^q \lambda_i}{\sum_{i=1}^p \lambda_i} = \frac{\sum_{i=1}^q \lambda_i}{\sum_{i=1}^p \sigma_i^2}$$

Para decidir q convendría tener en cuenta también:

- Si son muchas las componentes adicionales para conseguir un aumento notable de la proporción de variación explicada. Por ejemplo, si con dos componentes se explica el 60 % de la variabilidad, y hacen falta seis componentes para explicar el 70 %, entonces mejor quedarse sólo con las dos primeras componentes.
- Si las dos siguientes componentes por entrar explican cada una aproximadamente la misma cantidad de variabilidad, entonces cójanse las dos o ninguna. Meter sólo una de ellas sería arbitrario.

Otro criterio sería tener en cuenta la proporción de varianza explicada en la i-ésima variable original parcialmente reconstruida, que sería

$$\frac{\mathrm{Var}\left[x_{i;q}\right]}{\mathrm{Var}\left[x_i\right]} = \frac{\mathrm{Var}\left[\sum_{j=1}^{q} u_{ij}\, y_j\right]}{\sigma_{ii}} = \frac{\sum_{j=1}^{q} u_{ij}^2\ \mathrm{Var}\left[y_j\right]}{\sigma_{ii}} = \frac{\sum_{j=1}^{q} u_{ij}^2\, \lambda_j}{\sigma_{ii}}$$

La correlación entre variables originales y componentes principales se puede calcular como sigue:

$$\mathrm{Cor}\left[x_i, y_j\right] = \frac{\mathrm{Cov}\left[x_i, y_j\right]}{\sigma_i \sqrt{\lambda_j}} = \frac{u_{ij}\, \lambda_j}{\sigma_i \sqrt{\lambda_j}} = \frac{u_{ij}}{\sigma_i} \sqrt{\lambda_j}$$

pues

$$\mathrm{Cov}\left[\vec{x}, \vec{y}\right] = \mathcal{E}\left[\left(\vec{x} - \vec{\mu}\right) \vec{y}'\right] = \mathcal{E}\left[\left(\vec{x} - \vec{\mu}\right) \left(\mathbf{U}' \left(\vec{x} - \vec{\mu}\right)\right)'\right] =$$
$$= \mathcal{E}\left[\left(\vec{x} - \vec{\mu}\right) \left(\vec{x} - \vec{\mu}\right)' \mathbf{U}\right] = \mathcal{E}\left[\left(\vec{x} - \vec{\mu}\right) \left(\vec{x} - \vec{\mu}\right)'\right] \mathbf{U} = \Sigma\, \mathbf{U} = \mathbf{U}\, \Lambda$$

2.2. En R

Ulizaremos álgebra matricial. En concreto, `X` es una matriz `n`×`I` que representa los datos de `I` variables en `n` individuos; en las fórmulas siguientes, X se indicará también como una combinación de vectores columna $\vec{x}_1$, ..., $\vec{x}_I$. Considerése la matriz $\mathbf{H}$ siguiente:

$$\mathbf{H} \coloneqq \mathbf{I} - \frac{1}{n}\, \mathbf{J} \coloneqq \mathbf{I} - \frac{1}{n}\, \vec{1}\, \vec{1}'$$

H es una matriz de centrado. Si premultiplicamos X por H, el resultado es una matriz con los datos centrados respecto al vector de medias muestrales:

$$\mathbf{H}\, \mathbf{X} = \left(\mathbf{I} - \frac{1}{n}\, \vec{1}\, \vec{1}'\right) \left[\vec{x}_1 \ \ldots\ \vec{x}_p\right] = \left(\mathbf{I} - \frac{1}{n}\, \vec{1}\, \vec{1}'\right) \left[\ldots\ \vec{x}_i\ \ldots\right] =$$

$$= \mathbf{X} - \vec{1} \left[\ldots\ \frac{1}{n}\, \vec{1}'\, \vec{x}_i\ \ldots\right] = \mathbf{X} - \vec{1} \left[\ldots\ \bar{x}_i\ \ldots\right] = \mathbf{X} - \vec{1}\, \vec{\bar{x}}'$$

H es idempotente: $H\, H = H$. Es fácil demostrarlo a partir de su definición. En su idempotencia radica el hecho de que centrar una variable ya centrada la deja como estaba.
Un ejemplo de cómo calcular lo de la sección anterior en R:

```
> X       <- as.matrix (iris[,1:4]) # as.matrix pues protesta %*% si no
> n       <- nrow (X)               # número de filas = tamaño muestral
> I       <- diag (n)               # matriz identidad
> J       <- matrix (1, n, n)
> H       <- I - J/n                # matriz idempotente de centrado
> nSigma  <- t(X) %*% H %*% X       # matriz de sumas de productos
> Sigma   <- nSigma / n             # matriz de covarianzas
> auto    <- eigen (Sigma)
> U       <- auto$vectors           # autovectores por columnas
> U                                 # coeficientes de las componentes
```

```
            [,1]        [,2]        [,3]       [,4]
[1,]  0.36138659 -0.65658877  0.58202985  0.3154872
[2,] -0.08452251 -0.73016143 -0.59791083 -0.3197231
[3,]  0.85667061  0.17337266 -0.07623608 -0.4798390
[4,]  0.35828920  0.07548102 -0.54583143  0.7536574
```

```
> Landa    <- diag (auto$values)     # matriz con autovalores
> Y        <- H %*% X %*% U          # puntuaciones de los individuos en
> head (Y)                           # las componentes principales
```

```
          [,1]       [,2]        [,3]         [,4]
[1,] -2.684126 -0.3193972  0.02791483  0.002262437
[2,] -2.714142  0.1770012  0.21046427  0.099026550
[3,] -2.888991  0.1449494 -0.01790026  0.019968390
[4,] -2.745343  0.3182990 -0.03155937 -0.075575817
[5,] -2.728717 -0.3267545 -0.09007924 -0.061258593
[6,] -2.280860 -0.7413304 -0.16867766 -0.024200858
```

```
> corXY    <- diag (1 / sqrt(diag(Sigma))) %*% U %*% sqrt(Landa)
> corXY                              # correlaciones variables/componentes
```

```
           [,1]       [,2]        [,3]        [,4]
[1,]  0.8974018 -0.3906044  0.19656672  0.05882002
[2,] -0.3987485 -0.8252287 -0.38363030 -0.11324764
[3,]  0.9978739  0.0483806 -0.01207737 -0.04196487
[4,]  0.9665475  0.0487816 -0.20026170  0.15264831
```

Otra manera de llegar a lo mismo, usando la función `princomp` de R[1]:

```
> X        <- iris[,1:4]
> n        <- nrow (X)
> análisis <- princomp (X)
> U        <- unclass (loadings (análisis))
> Landa    <- diag (análisis$sdev)
> Sigma    <- var (X) * (n-1)/n
> corXY    <- diag (1 / sqrt(diag(Sigma))) %*% U %*% Landa
```

(Téngase en cuenta que esta `Landa` contiene desviaciones típicas, mientras que Λ y la `Landa` del listado previo contenían los autovalores, es decir, varianzas.) Así, la correlación entre la segunda variable y la tercera componente sería

```
> corXY [2, 3]
```

```
[1] -0.3836303
```

[1] En R también existe `prcomp`, que utiliza un algoritmo alternativo para calcular las componentes. Una ventaja de `prcomp` es que permite calcular componentes principales incluso si hay más variables que observaciones.

Las proporciones de varianzas explicadas en las variables originales por las q primeras componentes serían:

```
> q <- 2
> apply (corXY, 1, function (fila) sum (fila[1:q]^2))
```

```
[1] 0.9579017 0.8400028 0.9980931 0.9365937
```

2.3. Gráficos

Para representar gráficamente en dos dimensiones el resultado de un análisis de componentes principales, se suelen representar en los mismos ejes tanto los individuos de la muestra como las propias variables. Hay varios criterios posibles para elegir las escalas.

Directo. Para representar los individuos (ejes inferior e izquierdo), se utilizan sus proyecciones sobre las dos primeras componentes principales (`scores`). Para las variables (ejes superior y derecho), se utiliza la coordenada de los autovectores correspondiente a la componente en cuestión (`loadings`).

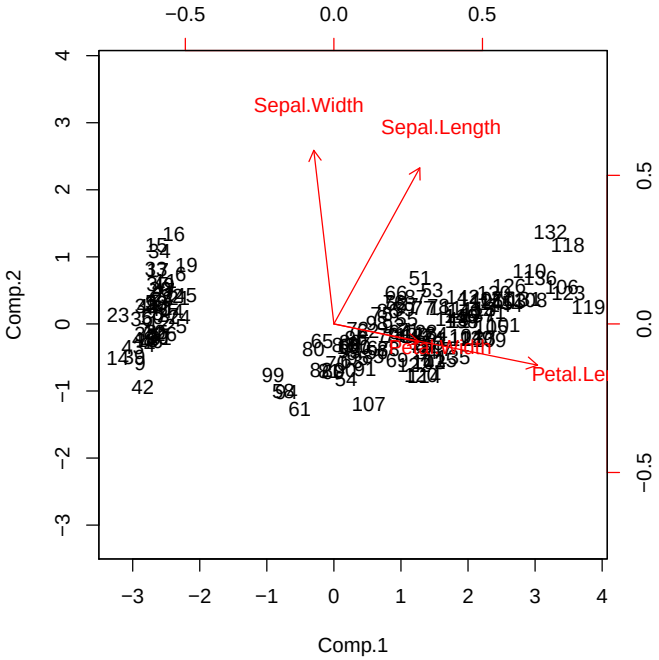
```
> i <- 42                                   # individuo 42 (p.ej.)
> j <- 2                                    # Sepal.Width  (p.ej.)
> análisis$scores[i,]
```

```
    Comp.1     Comp.2     Comp.3     Comp.4
-2.8493687 -0.9409606  0.3492304  0.3199875
```

```
> análisis$loadings[j,]
```

```
     Comp.1      Comp.2      Comp.3      Comp.4
-0.08452251  0.73016143 -0.59791083 -0.31972310
```

```
> biplot (análisis, scale=0)
```



De esta forma, los individuos aparecen en una proyección "real", y las variables aparecen como los autovectores unitarios proyectados.

Escalado. En este caso, las coordenadas para los individuos se tipifican dividiendo por la desviación típica de la componente correspondiente, es decir, por la raíz cuadrada correspondiente autovalor. Así, las coordenadas del individuo `i`-ésimo serían

```
análisis$scores[i,] / análisis$sdev
```

De esta manera, la nube de individuos se distribuye de manera más homogénea en el gráfico (se distorsiona la forma real para dar más visibilidad a las dimensiones con menos dispersión).
Para representar las variables, sus coordenadas son las del caso directo pero multiplicadas por las desviaciones típicas (raíces cuadradas de los autovalores) correspondientes. Esto destaca la dimensión asociada a la primera componente principal.
Por otro lado, de esta manera las escalas para individuos y para variables son similares.

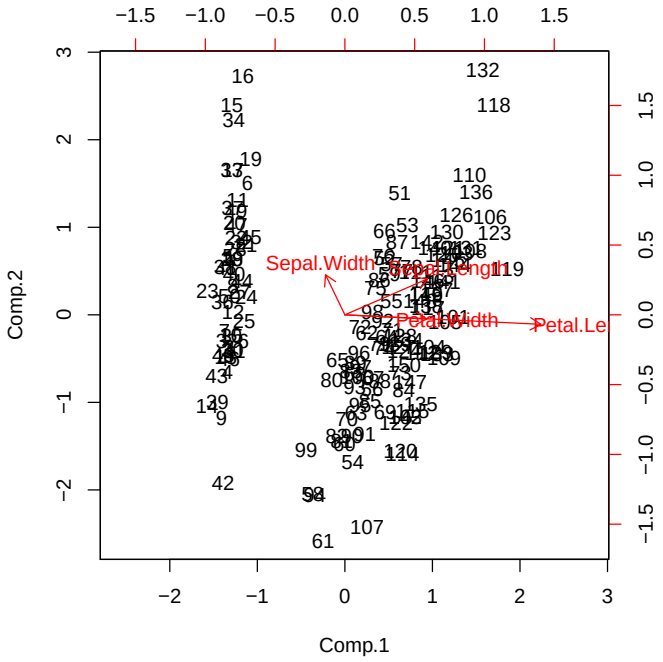
```
> análisis$scores[i,] / análisis$sdev

   Comp.1    Comp.2    Comp.3    Comp.4
-1.390341 -1.916528  1.252953  2.079587
```

```
> análisis$loadings[j,] * análisis$sdev
```

```
     Comp.1      Comp.2      Comp.3      Comp.4
-0.17322071  0.35848840 -0.16665321 -0.04919602
```

```
> biplot (análisis, pc=TRUE)
```



Recontraescalado. Si no se pone la opción `pc` a `TRUE`, se reescalan las coordenadas de individuos y variables dividiendo y multiplicando respectivamente por la raíz cuadrada del tamaño muestral. Es el criterio por omisión en R.

```
> n
```

```
[1] 150
```

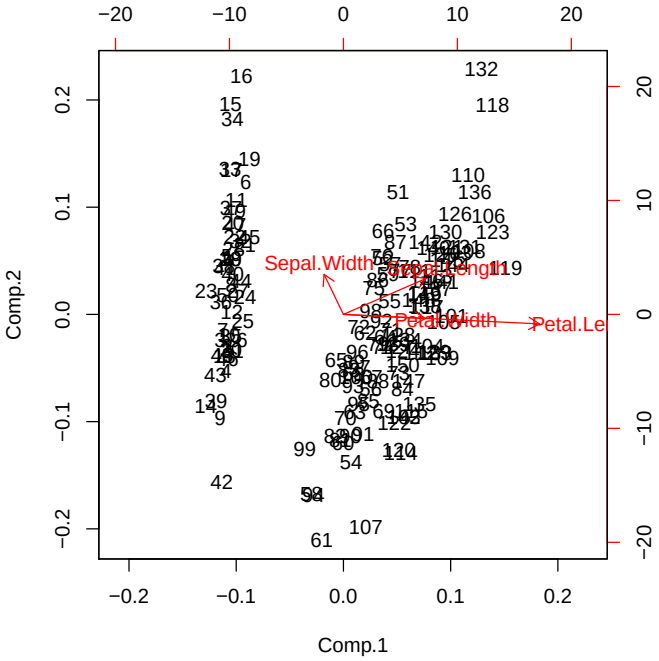
```
> análisis$scores[i,] / análisis$sdev / sqrt(n)
```

```
    Comp.1     Comp.2     Comp.3     Comp.4
-0.1135208 -0.1564839  0.1023032  0.1697976
```

```
> análisis$loadings[j,] * análisis$sdev * sqrt(n)
```

```
    Comp.1     Comp.2     Comp.3     Comp.4
-2.1215118  4.3905683 -2.0410766 -0.6025257

> biplot (análisis)
```



2.4. Ejemplo de cabezas de rugbi

1. Descargar el fichero desde `http://carleos.epv.uniovi.es/~carleos/adatos`
2. Descriptivos 1D.
3. Gráficos de caja.
4. Análisis de componentes principales de todas las variables salvo V1.
5. Gráfico de sedimentación (codo).
6. Matriz de covarianzas o de correlaciones
7. Interpretar los coeficientes de las componentes (¿dolicocéfalos?).