Usando las primeras dos columnas de Γ en la página 38, la primera y segunda componentes principales para los datos del billetes del banco suizo son

$$y_1 = 0.044x_1 + 0.112x_2 + 0.139x_3 + 0.768x_4 + 0.202x_5 - 0.579x_6$$

$$y_2 = 0.011x_1 + 0.071x_2 + 0.066x_3 - 0.563x_4 + 0.659x_5 - 0.489x_6$$

x_1 = longitud del billete.

x_2 = ancho del billete (a la izquierda).

x_3 = ancho del billete (a la derecha).

x_4 = distancia de la figura en el billete al borde inferior del billete

x_5 = distancia de la figura en el billete al borde superior del billete.

x_6 = longitud de la diagonal del billete

De forma que la primera componente y_1, esencialmente corresponde a la diferencia entre el ancho del borde inferior del billete y la longitud de la diagonal

La segunda componente principal y_2, corresponde a la diferencia entre el ancho del borde superior del billete y la suma de: el ancho del borde inferior y la longitud de la diagonal en los billetes.

Una medida que nos dice qué tan bien explican la variabilidad las primeras q componentes principales, está dada por

$$\psi_q \equiv \frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{p} \lambda_j} = \frac{\sum_{j=1}^{q} VAR(Y_j)}{\sum_{j=1}^{p} VAR(Y_j)}$$

La figura D muestra una gráfica de las parejas ordenadas $\left(i, \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}\right)$; $i=1,2,\ldots,6$.

Podemos darnos cuenta de que la 1ª componente principal explica un 66 % de la variabilidad y las primeras dos componentes explican 88 %. . En algunos libros se sugiere seleccionar $i_0 \in \{1,2,..,6\}$ en donde la gráfica de

La figura D[1] presenta un "codo" o se dobla. Este valor de i_0 es el número de componentes principales que se usarán para representar a los datos originales $\mathbb{X}$.

El siguiente teorema resultará de interés

TEOREMA Sea $\mathbb{X}$ un vector aleatorio de dimensión $p \times 1$ tal que $\mathbb{E}(\mathbb{X}) = \mu$ y $VAR(\mathbb{X}) = \Sigma$. Sea $\mathbb{Y}$ el vector de componentes principales de $\mathbb{X}$, entonces

(1) $COV(\mathbb{X}, \mathbb{Y}) = \Gamma \Lambda$,

donde Γ es la matriz de dimensiones $p \times p$ cuyas columnas son los vectores propios de Σ (en la descomposición de Jordan de Σ) y Λ es una matriz diagonal (de dimensiones $p \times p$) que tiene los valores propios de Σ, $\lambda_1, \ldots, \lambda_p$ como elementos de la diagonal.

(1) Scree Plot

(2) La correlación $\rho_{X_iY_j}$ entre la variable X_i y la componente principal Y_j está dada por

$$\rho_{X_iY_j} = \gamma_{ij}\left(\frac{\lambda_j}{\sigma^2_{X_iX_i}}\right)^{1/2}, \quad \text{.....} \quad (\text{rho})$$

donde $\sigma^2_{X_iX_i} = \mathrm{VAR}(X_i) = \Sigma_{ii}$ la entrada i,i de la matriz Σ y γ_{ij} es la entrada i,j de la matriz Γ.

(☆) Sea γ_i la columna i en Γ, notemos que $\sum_{j=1}^{p} \lambda_j \gamma_{ij}^2 = \gamma_i' \Delta \gamma_i$ = elemento (i,i) de la matriz $\Gamma \Delta \Gamma' = \Sigma$, entonces tenemos

$$\sum_{j=1}^{p} \rho^2_{X_iY_j} = \sum_{j=1}^{p} \gamma_{ij}^2 \frac{\lambda_j}{\Sigma_{ii}} = \frac{\sum_{j=1}^{p} \gamma_{ij}^2 \lambda_j}{\Sigma_{ii}}$$

$$= \frac{\Sigma_{ii}}{\Sigma_{ii}} = 1 .$$

La anterior propiedad nos inspira a interpretar la "correlación" $\rho^2_{X_iY_j}$ como "la proporción de de la varianza de la variable X_i explicada por la **j**-ésima componente principal Y_j".

Siguiendo estas **ideas**, el porcentaje de la varianza de X_i explicado por las primeras q componentes principales $Y_1, \ldots, Y_q$ está dado por $\sum_{j=1}^{q} \rho^2_{X_iY_j}$.

Como todas las cantidades mencionadas son poblacionales, necesitamos definir como calcular la correlación en la ecuación (rho) **para** unos datos $\mathbf{x}_1, \ldots, \mathbf{x}_n$, esta correlación muestral está dada por

$$r_{X_iY_j} = g_{ij}\left(\frac{l_j}{\hat{\Sigma}_{ii}}\right)^{1/2}.$$

Notemos que podemos argumentar como en (☆) pero con cantidades muestrales:

$\sum_{j=1}^{p} l_j g_{ij}^2 = \mathbf{g}_i' \mathbf{L} \mathbf{g}_i$ = **elemento** (i,i) de la matriz $\mathbf{G}\mathbf{L}\mathbf{G}' = \hat{\Sigma}$ ($\mathbf{g}_i$ = columna i en $\mathbf{G}$),