Analisis Multivariado 1 (Apunte basado en notas de clases del profesor Victor Yohai)

Andres Farall (afarall@hotmail.com) y Susana Sombielle (ssombielle@gmail.com)

June 24, 2011

1 Bibliografia

- Multivariate Observations, by G.A.F. Seber (Dificil).
- Applied Multivariate Data Analysis Volume 2, by J. D. Jobson (Facil).
- Multivariate Descriptive Statistical Analysis, by L. Lebart (intermedio).
- Análisis de datos multivariantes, by Daniel Peña (intermedio).
- The Elements of Statistical Learning, by T. Hastie, R Tibshirani and J. Friedman (intermedio).

2 Algunas convenciones, definiciones y propiedades

- Observamos simultaneamente d variables que conforman un vector $\mathbf{X}' = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_d}) \in \mathbb{R}^{\mathbf{d}}$ que posee una funcion de distribucion $F(\mathbf{X}, \theta)$.
- En una estructura matricial el primer indice identifica las filas y el segundo las columnas.
- El operador de trasposicion cumple con $(AB)' = B'A'$.
- Los vectores son vectores columna por defecto.

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \mathbf{x_n} \end{bmatrix}$$

- El operador esperanza aplicado a un vector aleatorio en $\mathbb{R}^d$ se define asi:

$$E(\mathbf{X}) = E\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{bmatrix} = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_d) \end{bmatrix}$$

- El operador esperanza aplicado a una matriz $A = \{a_{i,j}\}$ se define asi:

$$E(A) = \{e_{i,j}\} \quad con \quad e_{i,j} = E(a_{i,j})$$

- El operador Varianza de un vector aleatorio $\mathbb{R}^d$ de define asi:

$$VAR\,(\mathbf{X}) = \quad E\left[(X - E(\mathbf{X}))(X - E(\mathbf{X}))'\right] \quad = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & & \vdots \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & & \\ \vdots & & & \ddots & \\ \sigma_{d1} & \cdots & & & \sigma_{dd} \end{bmatrix}$$

- El operador Covarianza entre dos vectores aleatorios, $\mathbf{X}' = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_d}) \in \mathbb{R}^{\mathbf{d}}$ e $\mathbf{Y}' = (\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_k}) \in \mathbb{R}^{\mathbf{k}}$ se define asi

$$COV\,(\mathbf{X},\mathbf{Y}) = \quad E\left[(X - E(\mathbf{X}))(\mathbf{Y} - E(\mathbf{Y}))'\right] =$$

$$= \begin{bmatrix} cov(x_1, y_1) & cov(x_1, y_2) & cov(x_1, y_3) & \cdots & cov(x_1, y_k) \\ cov(x_2, y_1) & cov(x_2, y_2) & cov(x_2, y_3) & & \vdots \\ cov(x_3, y_1) & cov(x_3, y_2) & cov(x_3, y_3) & & \\ \vdots & & & \ddots & \\ cov(x_d, y_1) & \cdots & & & cov(x_d, y_k) \end{bmatrix}$$

- La matriz $A = \{a_{i,j}\}$ es simetrica $\iff a_{i,j} = a_{j,i}$.
- Dada $A = \{a_{i,j}\}$ simetrica, λ es autovalor y $\mathbf{b}$ es autovector correspondiente si $A\mathbf{b} = \lambda\mathbf{b}$.
- Dada $A \in \mathbb{R}^{d\times d}$ simetrica, se dice definida positiva si $\forall \mathbf{X} \in \mathbb{R}^{\mathbf{d}}, \mathbf{X}'A\mathbf{X} > 0$.
- Dada $A \in \mathbb{R}^{d\times d}$ simetrica, se dice semi-definida positiva si $\forall \mathbf{X} \in \mathbb{R}^{\mathbf{d}}, \mathbf{X}'A\mathbf{X} \geq 0$.
- Al valor escalar que se obtiene calculando $\mathbf{X}'A\mathbf{X}$ se lo llama forma cuadratica.
- $Tr\,(A + B) = Tr\,(A) + Tr\,(B)$ y $Tr\,(AB) = Tr\,(BA)$.
- Los autovalores no nulos de AB coinciden con los de BA. (Si las matrices son cuadradas, los nulos también coinciden).
- Sea A una matriz simétrica de $d \times d$. Todos sus autovalores son reales. Si llamamos $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ a estos autovalores, sucede que:
  - $tr\,(A) = \sum_{i=1}^{d} \lambda_i$
  - $|A| = \prod_{i=1}^{d} \lambda_i$

- $|I \pm A| = \prod_{i=1}^{d} (1 \pm \lambda_i)$
- $A \geq 0 \Leftrightarrow \lambda_i \geq 0 \ \ \forall i$.
- $A > 0 \Leftrightarrow \lambda_i > 0 \ \ \forall i$.
- $A \geq 0$ y $|A| \neq 0 \Rightarrow A > 0$.
- $A > 0 \Rightarrow A^{-1} > 0$.
- $A > 0 \Leftrightarrow$ existe $R \in \mathbb{R}^{d \times d}$ no singular tal que $A = RR' \Leftrightarrow$ existe una matriz ortogonal $B \in \mathbb{R}^{d \times d}$ tal que si $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_d)$ con $\lambda_i > 0 \ \ \forall i$ entonces $A = B\Lambda B'$ (es lo que se denomina *descomposición espectral* de A).
- $A \geq 0$ de rango $r \Leftrightarrow$ existe $R \in \mathbb{R}^{d \times d}$ de rango r tal que $A = RR' \Leftrightarrow$ existe una matriz ortogonal $B \in \mathbb{R}^{d \times d}$ tal que si $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_d)$ con $\lambda_i \geq 0 \ \forall i$ entonces $A = B\Lambda B'$.

- La matriz P de $d \times d$ se dice *de proyección* si es simétrica e idempotente (es decir, $P^2 = P$). Se cumple lo siguiente:
  - $rg(P) = r \Leftrightarrow \lambda_i = 1$ para $i = 1, \ldots, r$ y $\lambda_i = 0$ para $i = r+1, \ldots, d$. Entonces $P = \sum_{i=1}^{r} \mathbf{t}_i \mathbf{t}_i'$ para ciertos $\mathbf{t}_i$ ortonormales.
  - $rg(P) = tr(P)$.
  - $I - P$ también es de proyección.

- Sea X de $n \times p$ y de rango p. La matriz $P = X(X'X)^{-1} X'$ es una matriz de proyección.

- Vector de medias $\overline{\mathbf{X}} = \frac{\sum_{i=1}^{n} \mathbf{X_i}}{n}$

- Si $\mathbf{X}$ e $\mathbf{Y}$ son vectores aleatorios (no necesariamente de la misma dimensión) se puede ver que:
  - $COV(\mathbf{X}, \mathbf{Y}) = E(\mathbf{X}\mathbf{Y}') - E(\mathbf{X}) E(Y')$.
  - $COV(A\mathbf{X}, B\mathbf{Y}) = A\, COV(\mathbf{X}, \mathbf{Y})\, B'$.
  - Si $\mathbf{a}$ es un vector no aleatorio, $VAR(\mathbf{X} - \mathbf{a}) = VAR(\mathbf{X})$.
  - $VAR(A\mathbf{X}) = A\, VAR(\mathbf{X})\, A'$.

- Sea $\mathbf{X}_1, \ldots, \mathbf{X}_n$ una muestra de vectores aleatorios de dimensión d con varianza Σ y $\{a_i\}_{1 \leq i \leq n}, \{b_i\}_{1 \leq i \leq n}$ escalares no aleatorios. Se cumple:
  - $VAR\left(\sum_{i=1}^{n} a_i \mathbf{X}_i\right) = \left(\sum_{i=1}^{n} a_i^2\right) \Sigma$.
  - $COV\left(\sum_{i=1}^{n} a_i \mathbf{X}_i, \sum_{j=1}^{n} b_j \mathbf{X}_j\right) = \mathbf{O} \Leftrightarrow \sum_{i=1}^{n} a_i b_i = 0$.
  - Si $\mathbf{X} \sim (\mu, \Sigma)$ y A es simétrica, entonces $E(\mathbf{X}' A \mathbf{X}) = tr(A\Sigma) + \mu' A \mu$.

- Sea $\mathbf{X}_1, \ldots, \mathbf{X}_n$ una m.a. con parametros (μ, Σ). Se puede ver que:

- $E\left(\bar{\mathbf{X}}\right) = \mu$ y $VAR\left(\bar{\mathbf{X}}\right) = \Sigma/n$.
- $E\left(Q\right) = \left(n-1\right)\Sigma, \quad$ con $Q = \sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$

- Sean $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n$ vectores aleatorios i.i.d., la matriz de covarianza muestral $\hat{VAR}\left(\mathbf{X}\right) = \frac{\sum_{i=1}^{n}(\mathbf{X_i}-\overline{\mathbf{X}})(\mathbf{X_i}-\overline{\mathbf{X}})'}{n} = \frac{Q}{n} = S$.

3 Descomposicion Espectral

Dada $A \in \mathbb{R}^{n\times n}$ simetrica existen n autovalores $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_n$ y correspondientes autovectores $\mathbf{v}_1, \mathbf{v}_2 \ldots \mathbf{v}_n \in \mathbb{R}^n$ tales que forman una base ortonormal. Sean

$$V = [\mathbf{v_1}, \mathbf{v_2} \ldots \mathbf{v_n}] \text{ y } \Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & & \\ 0 & 0 & \lambda_3 & & \\ \vdots & & & \ddots & \\ 0 & & & & \lambda_n \end{bmatrix}$$

entonces:

$V'V = I = VV'$

$AV = V\Lambda \Longrightarrow AVV' = V\Lambda V' \Longrightarrow A = V\Lambda V' = \sum_{i=1}^{n} \lambda_i \mathbf{v_i}\mathbf{v_i'}$

3.1 Teorema de la descomposición espectral

Sea $A \in \mathbb{R}^{n\times n}$ con $A = A'$ (simérica) entonces:

1) Todos sus autovalores son reales $\lambda_i \in \mathbb{R}$, $i = 1...n$

2) Existen V (matriz ortogonal formada por los autovectores $\mathbf{v}_1 \ldots \mathbf{v}_n$ de A en sus columnas) y Λ (matriz diagonal formada por los autovalores $\lambda_1 \ldots \lambda_n$ de A) tales que

$$A = V\Lambda V' \text{ con } \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \\ \vdots & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix} \text{ y } V = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \\ \vdots & \vdots & & \vdots \end{bmatrix}$$

3) Otra forma (importante) de escribir a A es:

$$A = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i'$$

que equivale a sumar los productos externos de los autovectores, ponderados por sus autovalores

Demostración

Definimos la función $f : \mathbb{R}^n \to \mathbb{R}$

$$f\left(\mathbf{x}\right) = \mathbf{x}'A\mathbf{x} = \sum_{i=1}^{n}\sum_{j=i}^{n} a_{ij}\mathbf{x}_i\mathbf{x}_j \text{ con } \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \text{ y } a_{ij} = a_{ji}$$

forma cuadrática asociada a la matriz A

Definimos la función $g : \mathbb{R}^n \to \mathbb{R}$

$$g(\mathbf{x}) = \mathbf{x}'\mathbf{x} = \|\mathbf{x}\|^2 \qquad \text{la norma cuadrado del vector } \mathbf{x}$$

Queremos hallar el vector $\mathbf{x} \in \mathbb{R}^n$ que maximice f sujeto a $g(\mathbf{x}) = 1$

El gradiente de la función f es

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f=}{\partial xn} \end{bmatrix} = \begin{bmatrix} 2\sum_{j=1}^{n} a_{1j}x_j \\ \vdots \\ 2\sum_{j=1}^{n} a_{nj}x_j \end{bmatrix} = 2A\mathbf{x}$$

por lo que el gradiente de la función g es

$$\nabla g(\mathbf{x}) = \begin{bmatrix} \frac{\partial g}{\partial x_1} \\ \vdots \\ \frac{\partial g}{\partial xn} \end{bmatrix} = 2\mathbf{x}$$

Estamos en condiciones de aplicar el Teorema del Multiplicador de Lagrange, pues buscamos el máximo de una función $f : U \to \mathbb{R}$ (con U abierto) de clase C^k con $k \geq 1$, en la hiperficie constituida por $S = g^{-1}(1)$ imagen inversa de un valor regular $c = 1 \in \mathbb{R}$ de una función $g : U \to \mathbb{R}$ de clase C^k(cáscara de la bola unitaria en $\mathbb{R}^n$), que por ser un compacto sabemos que la función alcanza un máximo y que éste cumple con la condición necesaria de punto crítico, esto es

$$\nabla L(\mathbf{x}) = \nabla f(\mathbf{x}) - \lambda \nabla g(\mathbf{x}) = \mathbf{0} \to \nabla f(\mathbf{x}) = \lambda \nabla g(\mathbf{x}) \qquad (1)$$

$$g(\mathbf{x}) = 1 \qquad (2)$$

que resultan de derivar e igualar a 0 el lagrangiano

$$L(\mathbf{x}) = f(\mathbf{x}) - \lambda(g(\mathbf{x}) - 1)$$

donde (1) equivale a pedir que en el máximo la dirección de máximo crecimiento de la función debe ser perpendicular a la cáscara.

Reemplazando en (1) tenemos

$$\nabla f(\mathbf{x}) = \lambda \nabla g(\mathbf{x}) \leftrightarrow 2A\mathbf{x} = \lambda 2\mathbf{x} \longleftrightarrow A\mathbf{x} = \lambda \mathbf{x}$$

Así los puntos que satisfacen la condición necesaria son por definición los autovectores de la matriz A.

Por ser los autovalores las raices de la ecuación característica, tenemos n autovalores $\lambda_1 \ldots \lambda_n$, con sus n autovectores asociados $\mathbf{v}_1 \ldots \mathbf{v}_n$ Valuando la función f en el autovector $\mathbf{v}_i$ (elegido tal que $\|\mathbf{v}_i\| = 1$) obtenemos

$$f(\mathbf{v}_i) = \mathbf{v}_i' A \mathbf{v}_i = \mathbf{v}_i' \lambda_i \mathbf{v}_i = \lambda_i \mathbf{v}_i' \mathbf{v}_i = \lambda_i$$

De esta forma, como todos los autovalores pertenecen a la imagen de f, estos son reales, con lo que queda demostrado 1)

Por otro lado, dado que existe un máximo y como éste debe cumplir con la condición necesaria, los candidatos a máximo son los autovectores (con norma 1) $\mathbf{v}_1 \ldots \mathbf{v}_n$ y sus correspondientes imágenes ordenadas son $f(\mathbf{v}_i) = \lambda_1 \geq f(\mathbf{v}_2) = \lambda_2 \geq \ldots \geq f(\mathbf{v}_n) = \lambda_n$

Así

$\max_{\mathbf{x}} f(\mathbf{x}) = \lambda_1$ donde λ_1 es el mayor autovalor

$arg\max_{\mathbf{x}} f(\mathbf{x}) = \mathbf{v}_1$ donde $\mathbf{v}_1$ es el autovector asociado al mayor autovalor λ_1.

El mayor autovalor puede no ser único

Ahora buscamos el máximo de la función $f : E \to \mathbb{R}$ con $E = \{x \in \mathbb{R}^n : \mathbf{x}'\mathbf{v}_1 = 0\}$ abierto, en la hiperficie constituida por $S = g^{-1}(1)$ imagen inversa de un valor regular $c = 1 \in \mathbb{R}$ de una función $g : E \to \mathbb{R}$ (cáscara de la bola unitaria en el ortogonal a $\mathbf{v}_1$).

Nuevamente como estamos buscando el máximo de f en las condiciones del Teorema de Lagrange éste debe cumplir con la condición necesaria de punto crítico que es la definición de autovector

$$\nabla f(\mathbf{x}) = \lambda \nabla g(\mathbf{x}) \leftrightarrow A\mathbf{x} = \lambda\mathbf{x}$$

Así

$\max_{\mathbf{x},\ \mathbf{x}\ \prime\mathbf{v}_1=0} f(\mathbf{x}) = \lambda_2$ donde λ_2 es el segundo mayor autovalor

$arg\max_{\mathbf{x},\ \mathbf{x}\ \prime\mathbf{v}_1=0} f(\mathbf{x}) = \mathbf{v}_2$ donde $\mathbf{v}_2$ es el autovector asociado al autovalor λ_2.

El mecanismo se repite hasta el último autovector $\mathbf{v}_n$, con su autovalor asociado λ_n, el que debe ser el mínimo de la función f pues su imagen es la menor de todos los puntos críticos.

De esta manera los autovectores de la matriz A forman una base ortonormal

$$AV = V\Lambda \to AVV' = V\Lambda V' \to A = V\Lambda V'$$

pues $VV' = I$ con lo que queda demostrado 2)

Para demostar 3) expresamos la matriz A en función de sus elementos a_{ij}

$$\{a_{ij}\} = A = V\Lambda V' = \left\{\sum_{k=1}^{n} \lambda_k v_{ik} v_{kj}\right\} = \sum_{k=1}^{n} \lambda_k \{v_{ik} v_{kj}\} = \sum_{k=1}^{n} \lambda_k \mathbf{v}_k \mathbf{v}_k'$$

Ejemplo

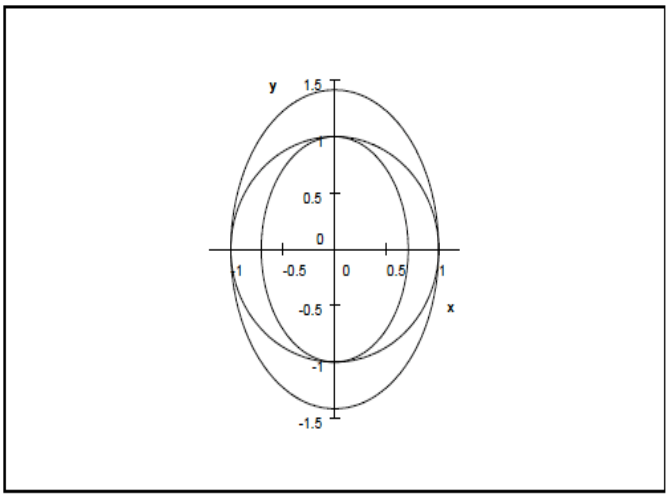
$$\text{Sea } A = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{4} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

entonces la función f es

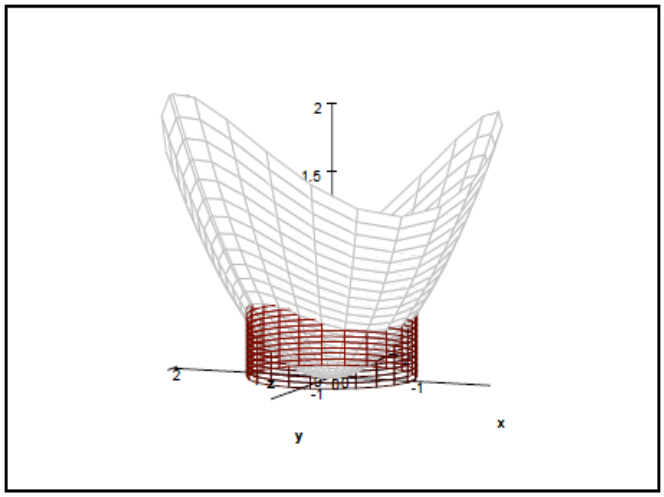
$$f(\mathbf{x}) = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{4} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \frac{1}{2}x^2 + \frac{1}{4}y^2$$

Sabemos que los puntos críticos son $\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ y $\mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ con sus valores de función $f(\mathbf{v}_1) = \lambda_1$ y $f(\mathbf{v}_2) = \lambda_2$.

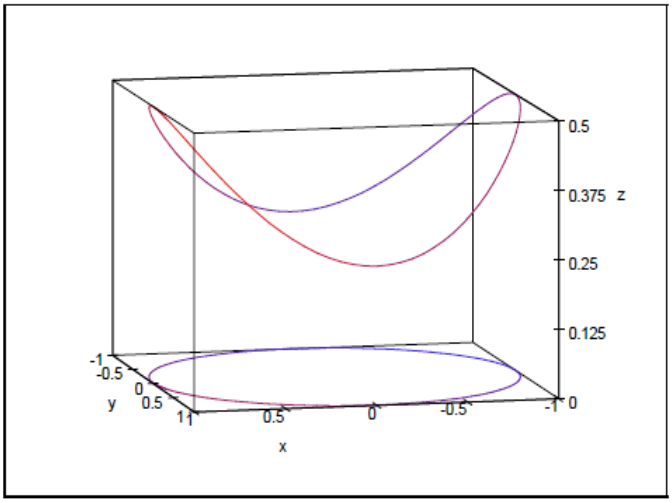
Las curvas de nivel $f(\mathbf{x}) = \frac{1}{2} = \lambda_1$, $f(\mathbf{x}) = \frac{1}{4} = \lambda_2$ y $x^2 + y^2 = 1$.



La función $z = f(\mathbf{x})$ y su intersección con $x^2 + y^2 = 1$.



La función $z = f(\mathbf{x})$ restringida al compacto $S = g^{-1}(1)$.



3.2 Raiz cuacrada de una matriz

Sea A semifedinida positiva, por la descomposicion espectral $A = B\Lambda B'$.

Sea $\Lambda^{1/2} = \begin{bmatrix} \lambda_1^{1/2} & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2^{1/2} & & & \\ 0 & 0 & \lambda_3^{1/2} & & \\ \vdots & & & \ddots & \\ 0 & & & & \lambda_p^{1/2} \end{bmatrix}$ entonces

$A = B\Lambda B' = B\Lambda^{1/2}\Lambda^{1/2}B' = (B\Lambda^{1/2})(\Lambda^{1/2}B') = (B\Lambda^{1/2})(B\Lambda^{1/2})' = CC'$

Asi la matriz C se llama raiz cuadrada de A, que no necesariamente es unica, por ejemplo si

$$R = B\Lambda^{1/2}B'$$

entonces

$$RR = B\Lambda^{1/2}B'B\Lambda^{1/2}B' = B\Lambda B' = A$$

por lo que la matriz $R \neq C$ es tambien una raiz cuadrada de A. Notar que $R = R'$ (simetrica).

4 Distribucion Normal Multivariada

Sea el vecor $\mathbf{X}' = (x_1, x_2, \ldots, x_d)$, $\mu \in \mathbb{R}^d$ y $\Sigma \in \mathbb{R}^{d\times d}$ definida positiva, entonces se dice que $\mathbf{X}$ sigue una distribucion normal multivariada $N_d(\mu, \Sigma)$ si la funcion de densidad de $\mathbf{X}$ es de la forma

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{X}-\mu)'\Sigma^{-1}(X-\mu)}$$

Asi, $E(\mathbf{X}) = \mu$ y $VAR(\mathbf{X}) = \Sigma$

Si $\mathbf{Y} = \mathbf{\Sigma^{-1/2}}(X - \mu)$ entonces $\mathbf{Y} \sim \mathbf{N}(0, I)$ por lo que $Y_1, Y_2 \ldots Y_n$ son v.a. $N(0,1)$ independientes.

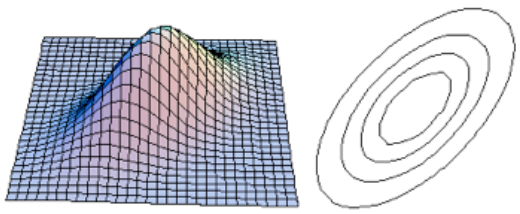
Propiedad importante: Sea $A \in \mathbb{R}^{hxd}$ con $rg(A) = h$ y $\mathbf{b} \in \mathbb{R}^h$ entonces

Si $\mathbf{Z} = A\mathbf{X} + \mathbf{b} \Longrightarrow \mathbf{Z} \sim N_h(A\mu + \mathbf{b}, A\Sigma A')$

Corolario importante: Tomando como $A = e_i' \in \mathbb{R}^{1\times d}$ los canonicos en $\mathbb{R}^d$ se deduce que las marginales de un vector normal son tambien variables (unidimensionales) normales.

Caso particular bivariado:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right)$$



4.1 Algunas Propiedades

- Sean $\mathbf{y}_i \sim N_d(\mu_i, \Sigma_i)$ independientes $(1 \leq i \leq n)$. Se prueba que $\sum_{i=1}^n a_i \mathbf{y}_i \sim N_d\left(\sum_{i=1}^n a_i\mu_i, \sum_{i=1}^n a_i^2\Sigma_i\right)$.
  - Sea $\mathbf{X}_1, \ldots, \mathbf{X}_n$ una m.a. de vectores $N_d(\mathbf{0}, \Sigma)$. Formemos la matriz $X' = (\mathbf{X}_1, \ldots, X_n) \in \mathbb{R}^{d\times n}$.
    - Si $\mathbf{a}$ de $n \times 1$ es un vector no aleatorio, entonces $X'\mathbf{a} \sim N_d\left(\mathbf{0}, \|\mathbf{a}\|^2 \Sigma\right)$.
    - Si $\{\mathbf{a}_1, \ldots, \mathbf{a}_r\}$ es un conjunto de vectores ortogonales no aleatorios, entonces los vectores aleatorios $\mathbf{u}_i = X'\mathbf{a}_i$ $(1 \leq i \leq r)$ son independientes.

* Si $\mathbf{b}$ de $d \times 1$ es un vector no aleatorio, entonces $X\mathbf{b} \sim N_n(\mathbf{0}, (\mathbf{b}'\Sigma\mathbf{b}) I_n)$. En particular, tomando $b = e_j$ el canonico j en $\mathbb{R}^d$, el vector conformado por todas las variables j-esimas de la muestra $\mathbf{x}^{(j)} \sim N_n(\mathbf{0}, \sigma_{jj} I_n)$, con $\Sigma = (\sigma_{ij})$.

Definición: Si las variables aleatorias $X_1, X_2, ..., X_n$ son i.i.d. $N_1\left(\mu_i, \sigma^2\right)$, entonces

$$U = \sum_{i=1}^{n} \frac{X_i^2}{\sigma^2} \sim \chi_n^2(\delta)$$

es decir que la distribución de la variable aleatoria U se denomina χ^2 *no central* con parámetro de centralidad $\delta = \sum_{i=1}^{n} \frac{\mu_i^2}{\sigma^2}$.

- Consideremos $\mathbf{X} \sim N_d(\mu, \Sigma)$.
  - Se prueba que $\mathbf{X}'\Sigma^{-1}\mathbf{X} \sim \chi_d^2(\delta)$ con $\delta = \mu'\Sigma^{-1}\mu$.
  - Si B es simétrica de rango k y $B\Sigma$ es idempotente, tambien se prueba que $\mathbf{x}'B\mathbf{x} \sim \chi_k^2(\delta)$ con $\delta = \mu'B\mu$.

4.2 Distribucion normal multivariada condicional

Sea el vecor $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2})$, $\mu \in \mathbb{R}^{d1+d2}$ y $\Sigma \in \mathbb{R}^{(d1+d2)\times(d1+d2)}$ tal que $\mathbf{X} \sim N_d(\mu, \Sigma)$, con

$\mu = (\mu_1, \mu_2)$ y $\Sigma = \begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{bmatrix}$

la distribucion de $\mathbf{X_2}$ condicional a que $\mathbf{X_1} = \mathbf{x}_1$ es

$\mathbf{X}_2|\mathbf{X}_1 = \mathbf{x}_1 \sim N_{d_2}\left(\Sigma_{2,1}\Sigma_{1,1}^{-1}(\mathbf{x}_1 - \mu_1) + \mu_2, \Sigma_{2,2} - \Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2}\right)$

4.3 La distribucion Wishart central

Decimos que $W \in \mathbb{R}^{d\times d}$ posee una distribucion Wishart central, $W \sim W_d(n, \Sigma)$ con Σ definida positiva, cuando $W = \mathbf{X}_1\mathbf{X}_1' + \mathbf{X}_2\mathbf{X}_2' + \cdots + \mathbf{X}_n\mathbf{X}_n'$, donde $\mathbf{X}_1, \mathbf{X}_2 \ldots \mathbf{X}_n$ son v.a.i. con distribucion $N_d(\mathbf{0}, \Sigma)$.

El hecho que Σ sea definida positiva implica $n \geq d$.

Es claro que

$$E(W) = n\Sigma$$

5 Teorema Central del Limite Multivariado

Sean $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n$ vectores aleatorios i.i.d. con $E(\mathbf{X}_i) = \mu$ y $VAR(\mathbf{X}_i) = \Sigma$ entonces

$$\mathbf{Z}_n = \sqrt{n}(\overline{\mathbf{X}} - \mu) \xrightarrow{D} N_d(\mathbf{0}, \Sigma)$$

es decir

$$F_{Z_n} \longrightarrow F_Z$$

donde $\mathbf{Z} \sim \mathbf{N_d}(0, \mathbf{\Sigma})$

6 Estimadores de Maxima-Verosimilitud de μ y Σ para el modelo $N_d(\mu, \Sigma)$.

Sea $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n$ una muestra aleatoria de vectores con distribucion $N_d(\mu, \Sigma)$, buscamos estimadores de μ y Σ por el metodo de maxima-verosimilitud, es decir, dada la muestra buscamos un vector $\hat{\mu}$ y una matriz $\hat{\Sigma}$ que maximicen la verosimilitud de la muestra.

Veremos que:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \mathbf{X_i}}{n} = \overline{\mathbf{X}}$$

$$\hat{\Sigma} = \frac{\sum_{i=1}^{n} (\mathbf{X_i} - \overline{\mathbf{X}})(\mathbf{X_i} - \overline{\mathbf{X}})'}{n} = \frac{Q}{n} = S$$

Demostracion:

$$L(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n, \mu, \Sigma) = \quad \prod_{i=1}^{n} f(\mathbf{X}_i, \mu, \Sigma) =$$

$$= \prod_{i=1}^{n} \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{X} - \mu)' \Sigma^{-1} (X - \mu)}$$

$$\ln L(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n, \mu, \Sigma) = \quad LL(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n, \mu, \Sigma) =$$

$$= -\frac{nd}{2} \ln(2\pi) - \frac{n}{2} \ln(|\Sigma|) - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{X_i} - \mu)' \Sigma^{-1} (\mathbf{X}_i - \mu)$$

Supongamos que conocemos el valor de Σ que maximiza $L(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n, \mu, \Sigma)$, buscamos maximizar la verosimilitud con respecto a μ. De esta forma solo la expresion

$-\frac{1}{2} \sum_{i=1}^{n} (\mathbf{X} - \mu)' \Sigma^{-1} (X_i - \mu)$

depende de μ. Asi buscamos el valor de μ que minimiza

$h(\mu) = \sum_{i=1}^{n} (\mathbf{X_i} - \mu)' \Sigma^{-1} (\mathbf{X}_i - \mu)$

pero

$$h(\mu) = \sum_{i=1}^{n} (\mathbf{X_i} - \mu)' \Sigma^{-1} (\mathbf{X}_i - \mu) = \quad \sum_{i=1}^{n} Tr \left\{ (\mathbf{X_i} - \mu)' \Sigma^{-1} (\mathbf{X}_i - \mu) \right\} \quad =$$

$$= \sum_{i=1}^{n} Tr \left\{ \Sigma^{-1} (\mathbf{X}_i - \mu)(\mathbf{X_i} - \mu)' \right\} \quad = Tr \left\{ \Sigma^{-1} \sum_{i=1}^{n} (\mathbf{X}_i - \mu)(\mathbf{X_i} - \mu)' \right\}$$

por otro lado

$$\sum_{i=1}^{n} (\mathbf{X}_i - \mu)(\mathbf{X_i} - \mu)' \quad = \quad \sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}} + \overline{\mathbf{X}} - \mu)(\mathbf{X_i} - \overline{\mathbf{X}} + \overline{\mathbf{X}} - \mu)' =$$

$$= \quad \sum_{i=1}^{n} \left\{ (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})' + (\overline{\mathbf{X}} - \mu)(\overline{\mathbf{X}} - \mu)' + (\mathbf{X_i} - \overline{\mathbf{X}})(\overline{\mathbf{X}} - \mu)' + (\mathbf{X_i} - \overline{\mathbf{X}})(\overline{\mathbf{X}} - \mu)' \right\} \quad =$$

$$= \quad \sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})' + \sum_{i=1}^{n} (\overline{\mathbf{X}} - \mu)(\overline{\mathbf{X}} - \mu)' \quad +0+0 =$$

$$= Q + n(\overline{\mathbf{X}} - \mu)(\overline{\mathbf{X}} - \mu)'$$

volviendo a $h(\mu)$

$$h(\mu) = Tr \left\{ \Sigma^{-1} \sum_{i=1}^{n} (\mathbf{X}_i - \mu)(\mathbf{X_i} - \mu)' \right\} \quad =$$

$$h(\mu) = Tr \left\{ \Sigma^{-1} \left\{ Q + n(\overline{\mathbf{X}} - \mu)(\overline{\mathbf{X}} - \mu)' \right\} \right\} =$$

$$= Tr(\Sigma^{-1} Q) + nTr \left\{ \Sigma^{-1} (\overline{\mathbf{X}} - \mu)(\overline{\mathbf{X}} - \mu)' \right\}$$

de esta forma, ya que $Tr(\Sigma^{-1}Q)$ es fijo (no depende de μ), la expresion a minimizar es

$$Tr \left\{ \Sigma^{-1} (\overline{\mathbf{X}} - \mu)(\overline{\mathbf{X}} - \mu)' \right\} = \quad Tr \left\{ (\overline{\mathbf{X}} - \mu)' \Sigma^{-1} (\overline{\mathbf{X}} - \mu) \right\} \quad =$$

$$= (\overline{\mathbf{X}} - \mu)' \Sigma^{-1} (\overline{\mathbf{X}} - \mu) \quad \geq \quad 0$$

por ser Σ definida positiva. Por lo que este termino es nulo $\Longleftrightarrow$ $\mu = \overline{\mathbf{X}}$, sin importar el valor de Σ que supusimos fijo.

Asi

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \mathbf{X_i}}{n} = \overline{\mathbf{X}}$$

Vimos que siendo Σ conocido el maximo de $L(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n, \mu, \Sigma)$ se alcanza tomando $\hat{\mu} = \overline{\mathbf{X}}$, reemplazando nos queda

$$\ln L(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n, \mu = \overline{\mathbf{X}}, \Sigma) = \quad LL(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n, \mu = \overline{\mathbf{X}}, \Sigma) =$$

$$= -\frac{nd}{2} \ln(2\pi) - \frac{n}{2} \ln(|\Sigma|) - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{X_i} - \overline{\mathbf{X}})' \Sigma^{-1} (\mathbf{X}_i - \overline{\mathbf{X}}) \quad =$$

recordando el desarrollo hecho para $h(\mu)$ nos queda

$$= -\tfrac{nd}{2}\ln(2\pi) - \tfrac{n}{2}\ln(|\Sigma|) - \tfrac{1}{2}Tr(\Sigma^{-1}Q) \quad =$$

$$= -\frac{nd}{2}\ln(2\pi) + \frac{n}{2}\ln(\left|\Sigma^{-1}\right|) - \frac{n}{2}Tr(\Sigma^{-1}\frac{Q}{n}) \quad =$$

$$= -\frac{nd}{2}\ln(2\pi) + \frac{n}{2}\ln(\left|\Sigma^{-1}\frac{Q}{n}\left[\frac{Q}{n}\right]^{-1}\right|) - \frac{n}{2}Tr(\Sigma^{-1}\frac{Q}{n}) \quad =$$

$$= -\frac{nd}{2}\ln(2\pi) + \frac{n}{2}\ln(\left|\Sigma^{-1}\frac{Q}{n}\right|) + \frac{n}{2}\ln(\left|\frac{Q}{n}^{-1}\right|) - \frac{n}{2}Tr(\Sigma^{-1}\frac{Q}{n}) \quad =$$

$$= -\frac{nd}{2}\ln(2\pi) + \frac{n}{2}\ln(\left|\Sigma^{-1}\frac{Q}{n}\right|) - \frac{n}{2}\ln(\left|\frac{Q}{n}\right|) - \frac{n}{2}Tr(\Sigma^{-1}\frac{Q}{n}) \quad = \quad g(\Sigma)$$

buscamos maximizar $g(\Sigma)$, proponemos $\Sigma = \frac{Q}{n}$, reemplazando queda

$$g(\Sigma = \frac{Q}{n}) = -\frac{nd}{2}\ln(2\pi) + \frac{n}{2}\ln(|I_d|) - \frac{n}{2}\ln(\left|\frac{Q}{n}\right|) - \frac{n}{2}Tr(I_d) \quad =$$

$$= -\frac{nd}{2}\ln(2\pi) - \frac{n}{2}\ln(\left|\frac{Q}{n}\right|) - \frac{nd}{2}$$

la estrategia ahora es ver que $g(\frac{Q}{n}) \geq g(\Sigma)$ para todo Σ semi-definido positivo, o lo que es lo mismo, $g(\frac{Q}{n}) - g(\Sigma) \geq 0$, veamos

$$g(\frac{Q}{n}) - g(\Sigma) = \quad -\tfrac{n}{2}\ln(\left|\tfrac{Q}{n}\right|) - \tfrac{nd}{2} \quad -\frac{n}{2}\ln(\left|\Sigma^{-1}\frac{Q}{n}\right|) + \frac{n}{2}\ln(\left|\frac{Q}{n}\right|) + \frac{n}{2}Tr(\Sigma^{-1}\frac{Q}{n})$$

$$= \quad -\frac{nd}{2} - \frac{n}{2}\ln(\left|\Sigma^{-1}\frac{Q}{n}\right|) + \frac{n}{2}Tr(\Sigma^{-1}\frac{Q}{n})$$

$$= \quad -\frac{nd}{2} - \frac{n}{2}\ln(\left|\Sigma^{-1}S\right|) + \frac{n}{2}Tr(\Sigma^{-1}S)$$

$$= \quad -\frac{nd}{2} - \frac{n}{2}\ln(|A|) + \frac{n}{2}Tr(A)$$

llamando $A = \Sigma^{-1}S$, llamemos $B = S^{1/2}\Sigma^{-1}S^{1/2}$, probaremos que:

- $Tr(A) = Tr(B)$
- $|A| = |B|$

- B es simetrica y definida positiva

$$Tr(A) = Tr(\Sigma^{-1}S) \quad Tr(\Sigma^{-1}S^{1/2}S^{1/2}) \quad = Tr(S^{1/2}\Sigma^{-1}S^{1/2}) = Tr(B)$$

$$|A| = \quad \left|\Sigma^{-1}S\right| = \quad \left|\Sigma^{-1}\right||S| = \frac{|S|}{|\Sigma|} = \frac{|S|^{1/2}|S|^{1/2}}{|\Sigma|} = \left|S^{1/2}\Sigma^{-1}S^{1/2}\right| = |B|$$

Asi, siendo $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_d$ los d autovalores positivos de B

$$g(\frac{Q}{n}) - g(\Sigma) \quad = \quad = -\frac{nd}{2} - \frac{n}{2}\ln(|A|) + \frac{n}{2}Tr(A) = -\frac{nd}{2} - \frac{n}{2}\ln(|B|) + \frac{n}{2}Tr(B)$$

$$= -\frac{nd}{2} - \frac{n}{2}\sum_{i=1}^{d}\ln(\lambda_i) + \frac{n}{2}\sum_{i=1}^{d} \quad \lambda_i \quad = \frac{n}{2}\left[\sum_{i=1}^{d}\lambda_i - d - \sum_{i=1}^{d}\ln(\lambda_i)\right] =$$

$$= \frac{n}{2}\sum_{i=1}^{d}\left[\lambda_i - \ln(\lambda_i) - 1\right] = \quad \frac{n}{2}\sum_{i=1}^{d}\left[w(\lambda_i)\right]$$

donde $w(x) = x - \ln(x) - 1$, siendo facil ver que $w(x) \geq 0, \forall x > 0$, por lo que

$$g(\tfrac{Q}{n}) - g(\Sigma) \geq 0$$

y

$$\hat{\Sigma} = \frac{\sum_{i=1}^{n}(\mathbf{X_i} - \overline{\mathbf{X}})(\mathbf{X_i} - \overline{\mathbf{X}})'}{n} = \frac{Q}{n} = S$$

expandiendo el estimador queda

$$\hat{\Sigma} = S = \frac{\sum_{i=1}^{n}(\mathbf{X_i} - \overline{\mathbf{X}})(\mathbf{X_i} - \overline{\mathbf{X}})'}{n} \quad =$$

$$= \quad \frac{1}{n}\left[\sum_{i=1}^{n}\mathbf{X_i}\mathbf{X_i'} - \sum_{i=1}^{n}\mathbf{X_i}\overline{\mathbf{X}}' - \sum_{i=1}^{n}\overline{\mathbf{X}}\mathbf{X_i'} + \sum_{i=1}^{n}\overline{\mathbf{X}}\overline{\mathbf{X}}'\right] =$$

$$= \quad \frac{1}{n}\left[\sum_{i=1}^{n}\mathbf{X_i}\mathbf{X_i'} - (\sum_{i=1}^{n}\mathbf{X_i})\overline{\mathbf{X}}' - \overline{\mathbf{X}}(\sum_{i=1}^{n}\mathbf{X_i'}) + n\overline{\mathbf{X}}\overline{\mathbf{X}}'\right] =$$

$$= \quad \frac{1}{n}\left[\sum_{i=1}^{n}\mathbf{X_i}\mathbf{X_i'} - n\overline{\mathbf{X}}\overline{\mathbf{X}}'\right] =$$

es el estimador de maxima-verosimilitud de Σ, calculando la esperanza tenemos

$$E(\hat{\Sigma}) = E(\frac{\sum_{i=1}^{n}(\mathbf{X_i}-\overline{\mathbf{X}})(\mathbf{X_i}-\overline{\mathbf{X}})'}{n}) = \frac{1}{n}E \quad \left[\sum_{i=1}^{n}\mathbf{X_i}\mathbf{X_i'} - n\overline{\mathbf{X}}\overline{\mathbf{X}}'\right] =$$

$$= \frac{(n-1)}{n}\Sigma$$

(ver ejercicio de la practica y siguiente *demostracion*).

Por ultimo calculemos el valor de la (log) verosimilitud en el maximo

$$\ln L(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n, \mu = \overline{\mathbf{X}}, S) = \quad LL(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n, \mu = \overline{\mathbf{X}}, \Sigma = S) =$$

$$= -\frac{nd}{2}\ln(2\pi) - \frac{n}{2}\ln(|S|) - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{X_i} - \overline{\mathbf{X}})'S^{-1}(\mathbf{X}_i - \overline{\mathbf{X}}) \quad =$$

$$= -\frac{nd}{2}\ln(2\pi) - \frac{n}{2}\ln(|S|) - Tr\left(\frac{1}{2}\sum_{i=1}^{n}(\mathbf{X_i} - \overline{\mathbf{X}})'S^{-1}(\mathbf{X}_i - \overline{\mathbf{X}})\right) \quad =$$

$$= -\frac{nd}{2}\ln(2\pi) - \frac{n}{2}\ln(|S|) - \frac{1}{2}\sum_{i=1}^{n}Tr\left((\mathbf{X_i} - \overline{\mathbf{X}})'S^{-1}(\mathbf{X}_i - \overline{\mathbf{X}})\right) \quad =$$

$$= -\frac{nd}{2}\ln(2\pi) - \frac{n}{2}\ln(|S|) - \frac{1}{2}\sum_{i=1}^{n}Tr\left(S^{-1}(\mathbf{X_i} - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})'\right) \quad =$$

$$= -\frac{nd}{2}\ln(2\pi) - \frac{n}{2}\ln(|S|) - \frac{1}{2}Tr\left(S^{-1}\sum_{i=1}^{n}(\mathbf{X_i} - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})'\right) \quad =$$

$$= -\frac{nd}{2}\ln(2\pi) - \frac{n}{2}\ln(|S|) - \frac{1}{2}Tr\left(S^{-1}nS\right) \quad =$$

$$= -\frac{nd}{2}\ln(2\pi) - \frac{n}{2}\ln(|S|) - \frac{n}{2}Tr\left(I_d\right) \quad =$$

$$= -\frac{nd}{2}\ln(2\pi) - \frac{n}{2}\ln(|S|) - \frac{nd}{2}$$

6.1 Propiedades estadisticas de los estimadores maximo-verosimiles

Veremos que:

- $\overline{\mathbf{X}}$ y Q son independientes
- $\overline{\mathbf{X}}$ se distribuye como $N_d(\mu, \frac{\Sigma}{n})$, o sea, $\sqrt{n}(\overline{\mathbf{X}} - \mu) \sim N_d(\mu, \Sigma)$
- Q se distribuye como $W_d(n-1, \Sigma)$

Para demostrar estas propiedades alcanza con hacerlo para el caso particular $\mu = \mathbf{0}$.

Lema previo: Sea $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n$ una muestra aleatoria de vectores con distribucion $N_d(\mathbf{0}, \Sigma)$, sea $B \in \mathbb{R}^{n \times n}$ una matriz ortogonal. Podemos pensar en la matriz aleatoria $X \in \mathbb{R}^{n \times d}$

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & & & \\ x_{3,1} & 0 & x_{3,3} & & \\ \vdots & & & \ddots & \\ x_{n,1} & & & & x_{n,d} \end{bmatrix} = \begin{bmatrix} \mathbf{X'}_1 \\ \mathbf{X'_2} \\ \vdots \\ \mathbf{X'_n} \end{bmatrix} =$$

$$= \begin{bmatrix} \mathbf{X}^{(1)} & \mathbf{X}^{(2)} & \mathbf{X}^{(3)} & \cdots & \mathbf{X}^{(d)} \end{bmatrix}$$

sean $\mathbf{Y}^{(i)} = B\mathbf{X}^{(i)}$ para $1 \leq i \leq d$ nuevos vectores columna, que forman una matriz $Y = BX$

$$Y = \begin{bmatrix} y_{1,1} & y_{1,2} & y_{1,3} & \cdots & y_{1,d} \\ y_{2,1} & y_{2,2} & & & \\ y_{3,1} & 0 & y_{3,3} & & \\ \vdots & & & \ddots & \\ y_{n,1} & & & & y_{n,d} \end{bmatrix} = \begin{bmatrix} \mathbf{Y'_1} \\ \mathbf{Y'_2} \\ \vdots \\ \mathbf{Y'_n} \end{bmatrix} =$$

$$= \begin{bmatrix} \mathbf{Y}^{(1)} & \mathbf{Y}^{(2)} & \mathbf{Y}^{(3)} & \cdots & \mathbf{Y}^{(d)} \end{bmatrix}$$

queremos ver que $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \cdots, \mathbf{Y}_n$ es una muestra aleatoria (i.i.d.) de vectores con distribucion $N_d(\mathbf{0}, \Sigma)$. Tomemos un $\mathbf{Y_i}$ cualquiera, es claro que sus d componntes son normales, ya que las componentes de todas las $\mathbf{Y}^{(1)}\mathbf{Y}^{(2)}\mathbf{Y}^{(3)} \cdots \mathbf{Y}^{(d)}$ son normales. A su vez todas las componentes tienen media 0 ya que todas las componentes de las $\mathbf{Y}^{(1)}\mathbf{Y}^{(2)}\mathbf{Y}^{(3)} \cdots \mathbf{Y}^{(d)}$ tienen esperanza nula. Solo faltaria demostrar que son independientes y con matriz de covarianzas Σ. Veamos como es la matriz de covarianzas de los vectores columna $\mathbf{X}^{(1)} \quad \mathbf{X}^{(2)} \quad \mathbf{X}^{(3)} \quad \cdots \quad \mathbf{X}^{(d)}$.

$$COV(\mathbf{X^{(i)}}, X^{(j)}) \quad = \quad E(\mathbf{X^{(i)}} X'^{(j)}) = A = \{a\}_{k,l} \in \mathbb{R}^{n \times n}$$

$$a_{kl} \quad = E(x_{k,i}, x_{l,j}) \quad = \begin{cases} 0 & k \neq l \\ \sigma_{ij} & k = l \end{cases}$$

asi $A = \sigma_{i,j} I_n$.

Veamos como es la matriz de covarianzas de los vectores columna $\mathbf{Y}^{(1)}\mathbf{Y}^{(2)}\mathbf{Y}^{(3)} \cdots \mathbf{Y}^{(d)}$

$$COV(\mathbf{Y^{(i)}}, \mathbf{Y}^{(j)}) \quad = \quad BE(\mathbf{X^{(i)}} \mathbf{X'}^{(j)})B' = BAB' = B\sigma_{i,j} I_n B' = \sigma_{i,j} I_n BB' = \sigma_{i,j} I_n = A$$

Pot lo que $COV(\mathbf{Y'_i}) = \Sigma$ para $1 \leq i \leq d$ y ocurre que $\mathbf{Y_i}$ es independiente de $\mathbf{Y_j}$si $i \neq j$.

Retomando la demostracion, sea $B \in \mathbb{R}^{n\times n}$ una matriz ortogonal cuyo primer vector fila es

$$\mathbf{b_1'} = \left[\begin{array}{ccccccc} \frac{\mathbf{1}}{\sqrt{\mathbf{n}}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{\mathbf{1}}{\sqrt{\mathbf{n}}} \end{array} \right]$$

asi, dada la muestra aleatoria de vectores $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n$ y la nueva muestra aleatoria de vectores $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \cdots, \mathbf{Y}_n$ provenientes de los vectores $\mathbf{Y}^{(1)}\mathbf{Y}^{(2)}\mathbf{Y}^{(3)}\cdots\mathbf{Y}^{(d)}$ generados por la transformacion $\mathbf{Y}^{(i)} = B\mathbf{X}^{(i)}$, vemos que

$$\mathbf{Y_1} = \mathbf{b}_1'X = \left[\begin{array}{cccccc} \mathbf{b}_1'\mathbf{X^{(1)}} & \mathbf{b}_1'\mathbf{X^{(2)}} & \mathbf{b}_1'\mathbf{X^{(3)}} & \cdots & \mathbf{b}_1'\mathbf{X^{(d)}} \end{array} \right]$$

$$= \left[\begin{array}{cccccc} \frac{1}{\sqrt{n}}\sum_{i=1}^n x_{i,1} & \frac{1}{\sqrt{n}}\sum_{i=1}^n x_{i,2} & \frac{1}{\sqrt{n}}\sum_{i=1}^n x_{i,3} & \cdots & \cdots & \frac{1}{\sqrt{n}}\sum_{i=1}^n x_{i,d} \end{array} \right]$$

$$= \sqrt{n}\overline{\mathbf{X}} \sim N(\mathbf{0}, \Sigma)$$

pues $\mathbf{Y_1} \sim N(\mathbf{0}, \Sigma)$.
Por otro lado

$$Y'Y = X'B'BX = X'X$$

$$Y'Y = \sum_{i=1}^n \mathbf{Y_i}\mathbf{Y_i'} = Y_1Y_1' + \sum_{i=2}^n \mathbf{Y_i}\mathbf{Y_i'}$$

$$= n\overline{XX}' + \sum_{i=2}^n \mathbf{Y_i}\mathbf{Y_i'}$$

despejando, nos queda

$$\sum_{i=2}^n \mathbf{Y_i}\mathbf{Y_i'} = \sum_{i=1}^n \mathbf{Y_i}\mathbf{Y_i'} - n\overline{XX}' = \sum_{i=1}^n \mathbf{X_i}\mathbf{X_i'} - n\overline{XX}' = Q$$

y como $\mathbf{Y}_2, \mathbf{Y}_3, \cdots, \mathbf{Y}_n$ son $n-1$ vectores aleatorios normales independientes, entonces

Q se distribuye como $W_d(n-1, \Sigma)$

por ende se tiene que

$$E(Q) = (n-1)\Sigma$$

y el estimador de la matriz de varianzas y covarianzas cumple

$$E(S) = E(\tfrac{Q}{n}) = \frac{(n-1)}{n}\Sigma$$

La independencia entre $\overline{\mathbf{X}}$ y Q surge claramente de la independencia entre $\mathbf{Y}_1$ y el resto de los vectores $\mathbf{Y}_2, \mathbf{Y}_3, \cdots, \mathbf{Y}_n$.

6.2 Estimador insesgado de Σ

Dado que $E(S) = \frac{(n-1)}{n}\Sigma$, un estimador insesgado seria $\frac{n}{n-1}S = \frac{Q}{n-1} = S^*$

7 Estadistico de Hotelling

El estadistico de Hotelling (teórico) se obtiene mediante una forma cuadratica que combina un vector normal $\mathbf{V} \sim N_d(\mathbf{0}, \Sigma)$ y una matriz con distribucion Wishart $W = W_d(n, \Sigma)$, de la siguiente manera:

$$T^2_{d,n} = \mathbf{n}\mathbf{V}'W^{-1}\mathbf{V} \sim H(d,n)$$

Un teorema dificil muestra que $\frac{n-d+1}{dn}T^2_{d,n-1} \sim F_{d,n-d+1}$.

El estadistico T^2 de Hotelling se utiliza para testear hipotesis de media de una poblacion normal multivariada, o como resultado del T.C.L. multivariado, para testear medias de poblaciones no normales pero con muestras suficientemente nomerosas.

Si tenemos $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n$ una muestra aleatoria de vectores con distribucion $N_d(\mu, \Sigma)$, donde μ y Σ son desconocidos, y deseamos testear $H_0 : \mu = \mu_0$, el estadistico propuesto es el siguiente:

$$T^2_{d,n-1} = n(\overline{\mathbf{X}} - \mu_\mathbf{0})'S^{*-1}(\overline{\mathbf{X}} - \mu_\mathbf{0}) = n(n-1)(\overline{\mathbf{X}} - \mu_\mathbf{0})'Q^{-1}(\overline{\mathbf{X}} - \mu_\mathbf{0}) =$$

$$= (n-1)\left[\sqrt{n}(\overline{\mathbf{X}} - \mu_\mathbf{0})\right]' Q^{-1}\left[\sqrt{n}(\overline{\mathbf{X}} - \mu_\mathbf{0})\right]$$

y se nota de la siguiente manera $T^2_{d,n-1} \sim H(d, n-1)$.

El test se rechaza cuando $T^2_{d,n-1} > H_{\alpha,d,n-1}$ o equivalentemente cuando $\frac{n-d}{d(n-1)}T^2 > F_{\alpha,d,n-d.}$

8 El test de Hotelling como interseccion de infinitos tests univariados (tecnica canonica)

Sea $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n$ una muestra aleatoria de vectores con distribucion $N_d(\mu, \Sigma)$, con μ y Σ desconocidos, y deseamos testear $H : \mu = \mu_0$ versus $K : \mu \neq \mu_0$

El estadistico propuesto es el siguiente:

$$T^2_{d,n-1} = n(\overline{\mathbf{X}} - \mu_\mathbf{0})'S^{*-1}(\overline{\mathbf{X}} - \mu_\mathbf{0})$$

y rechazamos H cuando $T^2_{d,n-1} > T^2_{d,n-1,\alpha}$

Podemos derivar este mismo test basandonos en tests univariados, sea

$$H_\beta : \beta'\mu = \beta'\mu_0$$

versus

$$K_\beta : \beta'\mu \neq \beta'\mu_0$$

de esta forma

$$H = \bigcap^{\beta \neq 0} H_\beta$$

y

$$K = \bigcup^{\beta \neq 0} K_\beta$$

llamemos $Z_{\mathbf{i}}^\beta = \beta'\mathbf{X}_i$ para $1 \leq i \leq n$, asi $Z_{\mathbf{i}}^\beta \sim N_1(\beta'\mu, \beta'\mathbf{\Sigma}\beta)$, usamos el estadistico univariado

$$t_\beta \quad = \quad \sqrt{n}\frac{(\bar{z_\beta} - \beta'\mu_0)}{s_\beta}$$

con $s_\beta^2 = \sum_{i=1}^n \frac{(z_i^\beta - \bar{z}_\beta)^2}{n-1} = \beta' S^* \beta$ (ejercicio).

Rechazamos H_β cuando $|t_\beta| > t_{\alpha/2,n-1}$ o lo que es lo mismo si $t_\beta^2 > t_{\alpha/2,n-1}^2$.

Primero veamos que el estadistico puede reexpresarse convenientemente en terminos de la media de la muestra original

$$t_\beta^2 \quad = \quad n\frac{\left\{\beta'(\bar{\mathbf{X}} - \mu_\mathbf{0})\right\}^2}{\beta' S^* \beta}$$

donde, recordemos

$$S^* \quad = \frac{\sum_{i=1}^n (\mathbf{X_i} - \overline{\mathbf{X}})(\mathbf{X_i} - \overline{\mathbf{X}})'}{n-1}$$

para rechazar H basta con que algun t_β^2 sea grnade (mayor que el punto critico $t_{\alpha/2,n-1}^2$), asi que podemos definir el estadistico

$$\underset{\beta}{max} \quad (t_\beta^2)$$

probaremos que

$$\max_\beta(t_\beta^2) \quad = \quad T_{d,n-1}^2$$

Demostracion:

Si llamemos $\delta = \sqrt{n}(\bar{\mathbf{X}} - \mu)$ entonces el estadistico $t_\beta^2 = \frac{(\beta'\delta)^2}{\beta' S^* \beta}$ veremos que

$$\max_\beta(t_\beta^2) \quad = \quad \delta' S^{*-1} \delta$$

alcanzando el máximo en el vector

$$\beta \quad = \quad cS^{*-1}\delta$$

para cualquier $c \in \mathbb{R}$.
Usando la desigualdad de Cauchy–Schwarz

$$\left(\mathbf{X}'\mathbf{Y}\right)^2 \quad \leq \quad \|\mathbf{X}\|^2 \|\mathbf{Y}\|^2$$

donde la igualdad se cumple cuando ambos vectores tienen la misma direccion, es decir

$$\mathbf{Y} \quad = \quad c\mathbf{X}$$

Sabiendo que S^* es simetrica y definida positiva entonces podemos escribir $S^* = RR$ (con R simetrica) , asi

$$t_\beta^2 \quad = \quad \frac{(\beta'\delta)^2}{\beta' S^* \beta} = \frac{(\beta' R R^{-1}\delta)^2}{\beta' S^* \beta} = \frac{\left[(R'\beta)'\left(R^{-1}\delta\right)\right]^2}{\beta' S^* \beta} \leq$$

$$\leq \quad \frac{\|R'\beta\|^2 \left\|R^{-1}\delta\right\|^2}{\beta' S^* \beta} = \frac{(R'\beta)'(R'\beta)\left(R^{-1}\delta\right)'\left(R^{-1}\delta\right)}{\beta' S^* \beta} =$$

$$= \quad \frac{\beta' RR\beta\delta'(R^{-1})'R^{-1}\delta}{\beta' S^* \beta} =$$

$$= \quad \frac{\beta' S^* \beta\delta'(R^{-1})'R^{-1}\delta}{\beta' S^* \beta} =$$

$$= \delta'(R')^{-1}R^{-1}\delta \quad =$$

$$= \delta' R^{-1}R^{-1}\delta \quad =$$

$$= \delta'(RR)^{-1}\delta \quad = \delta'(S^*)^{-1}\delta$$

asi

$$t_\beta^2 \qquad \leq \delta'(S^*)^{-1}\delta = \left(\sqrt{n}(\bar{\mathbf{X}} - \mu)\right)' S^{*-1}\left(\sqrt{n}(\bar{\mathbf{X}} - \mu)\right) =$$

$$= n(\bar{\mathbf{X}} - \mu)' S^{*-1}(\bar{\mathbf{X}} - \mu) = T_{d,n-1}^2$$

la igualdad se cumple si

$$R\beta = cR^{-1}\delta$$

$$R^{-1}R\beta = cR^{-1}R^{-1}\delta$$

$$\beta = cS^{*-1}\delta$$

por lo que

$$\max_{\beta}(t_{\beta}^2) = \delta' S^{*-1}\delta = T_{d,n-1}^2$$

9 Regiones de Confianza para μ

Las regiones de confianza son el equivalente multivariado a los intervalos de confianza univariados.

Sea $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n$ una muestra aleatoria de vectores con distribucion $N_d(\mu, \Sigma)$, buscamos una region (aleatoria) $RC(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n) \subseteq \mathbb{R}^d$ que satisfaga la siguiente propiedad:

$$P_{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n}\left[\mu \in RC(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n)\right] = 1 - \alpha$$

recordemos que el estadistico T^2 cumple

$$P\left[T^2 \leq T_{d,n-1,\alpha}^2\right] = P\left[n(\overline{\mathbf{X}} - \mu)' S^{*-1}(\overline{\mathbf{X}} - \mu) \leq T_{d,n-1,\alpha}^2\right] = 1 - \alpha$$

parece natural porponer la siguiente region

$$RC = \left\{\mu : n(\overline{\mathbf{X}} - \mu)' S^{*-1}(\overline{\mathbf{X}} - \mu) \leq T_{d,n-1,\alpha}^2\right\}$$

esta region conforma un elipsoide en $\mathbb{R}^d$.

10 Intervalos de Confianza Simultaneos

Sea $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n$ una muestra aleatoria de vectores con distribucion $N_d(\mu, \Sigma)$, con μ y Σ desconocidos, y tomemos una combinacion lineal

$$\mathbf{Z}_{\beta} = \beta' \mathbf{X}$$

asi $Z_{\mathbf{i}}^{\beta} \sim N_1(\beta'\mu, \beta' \mathbf{\Sigma} \beta)$, y buscamos un intervalo de confianza para

$$E\left(Z_{\beta}\right) = \beta' E(\mathbf{X}) = \beta'\mu = \gamma_{\beta}$$

el intervalo univariado es

$$IC_\beta \quad = \quad \left(\bar{\mathbf{Z}}_\beta - \frac{t_{\alpha/2,n-1}S_\beta}{\sqrt{n}}; \bar{\mathbf{Z}}_\beta + \frac{t_{\alpha/2,n-1}S_\beta}{\sqrt{n}}\right)$$

donde

$$S_\beta^2 \quad = \frac{\sum_{i=1}^n (\mathbf{Z_i} - \overline{\mathbf{Z}})^\mathbf{2}}{n-1}$$

El intervalo de confianza verifica

$$P(\gamma_\beta \in IC_\beta) \quad = \quad 1-\alpha$$

Sin embargo, si interesa buscar intervalos de confianza para mas de una combinacion lineal

$$E\left(Z_\beta^i\right) \qquad = \beta_\mathbf{i}' E(\mathbf{X}) = \beta_\mathbf{i}' \mu = \gamma^\mathbf{i}{}_\beta$$

para $i \in C$ con $\#C > 1$ y asegurarse que todos ellos satisfagan simultaneamente la misma probabilidad de cobertura, es decir

$$P(\underset{\mathbf{i}\in\mathbf{C}}{\cap} \gamma_\beta^\mathbf{i} \in IC_\beta^i) \quad = \quad 1-\alpha$$

hay al menos dos alternativas

10.1 Metodo de Bonferroni (pocas combinaciones lineales)

Supongamos que la cantidad de combinaciones es $K = \#C$

$$P(\underset{\mathbf{i}\in\mathbf{C}}{\cap} \gamma_\beta^\mathbf{i} \in IC_\beta^i) \quad = \quad 1 - P(\underset{\mathbf{i}\in\mathbf{C}}{\cup} \gamma_\beta^\mathbf{i} \notin IC_\beta^i) \geq$$

$$\geq \quad 1 - \sum_{i\in C} P(\gamma_\beta^\mathbf{i} \notin IC_\beta^i) = 1 - K\alpha$$

por lo que si disminuimos la probabilidad de no cobertura de α a α/K garantizamos que

$$P(\underset{\mathbf{i}\in\mathbf{C}}{\cap} \gamma_\beta^\mathbf{i} \in IC_\beta^i) \quad \geq \quad 1-\alpha$$

el inconveniente de esta alternativa consiste en que es excesivamente conservadora para cada intervalo individual, pues

$$P(\gamma_\beta \in IC_\beta) \quad = \quad 1-\alpha/K$$

exige a los intervalos ser demasiado grandes.

10.2 Metodo simultaneo (muchas combinaciones lineales)

Vamos a demostrar que si definimos el intervalo

$$IC_\beta^S \quad = \quad \left(\gamma : \frac{\sqrt{n}(\overline{Z}_\beta - \gamma)}{S_\beta} \leq \sqrt{T^2_{d,n-1,\alpha}}\right)$$

cumple con

$$P(\underset{\beta\in\mathbb{R}^{\mathbf{d}}}{\cap}\gamma_\beta \in IC_\beta^S) \quad = \quad 1-\alpha$$

Demostracion:

$$P(\underset{\beta\in\mathbb{R}^{\mathbf{d}}}{\cap}\gamma_\beta \in IC_\beta^S) \quad = \quad P\left(\underset{\beta\in\mathbb{R}^{\mathbf{d}}}{\cap}\left\{\frac{\sqrt{\mathbf{n}}(\overline{\mathbf{Z}}_\beta - \gamma)}{\mathbf{S}_\beta} \leq \sqrt{\mathbf{T^2_{d,n-1},\alpha}}\right\}\right) =$$

$$P\left(\underset{\beta\in\mathbb{R}^{\mathbf{d}}}{\cap}\left\{\frac{\mathbf{n}(\gamma - \mathbf{Z}_\beta)^{\mathbf{2}}}{\mathbf{S}^{\mathbf{2}}_\beta} \leq \mathbf{T^2_{d,n-1},\alpha}\right\}\right) = P\left(\max_\beta(t^2_\beta) \leq T^2_{d,n-1,\alpha}\right) \quad =$$

$$= P\left(n(\overline{\mathbf{X}} - \mu)' S^{*-1}(\overline{\mathbf{X}} - \mu) \leq T^2_{d,n-1,\alpha}\right) \quad = \quad 1-\alpha$$

11 El test de Hotelling como Test de cociente de MaximaVerosimilitud (sin demostracion).

El estadistico de Hotelling tambien puede ser derivado del cociente de Maxima Verosimilitud para el vector μ de medias.

Sea $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n$ una muestra aleatoria de vectores con distribucion $N_d(\mu, \Sigma)$, con μ y Σ desconocidos, y deseamos testear $H : \mu = \mu_0$ versus $K : \mu \neq \mu_0$, por definicion el estadistico del cociente de Maxima Verosimilitud es

$$CV(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n) \quad = \quad \frac{\max_{\mu,\Sigma} L(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n, \mu, \Sigma)}{\max_{\Sigma} L(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n, \mu_{\mathbf{0}}, \Sigma)}$$

Valores grandes del estadistico (mayores a 1) muestran evidencia en contra de la hipotesis H y el test rechaza cuando el estadistico satisface, para alguna constante K_α debidamente elgida

$$CV(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n) \quad > \quad K_\alpha$$

Luego de algunos calculos puede verse que

$$CV(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n) \quad = \quad g(n(\overline{\mathbf{X}} - \mu_{\mathbf{0}})' S^{*-1}(\overline{\mathbf{X}} - \mu_{\mathbf{0}})) = g(T^2_{d,n-1})$$

donde $g()$ es una funcion monotona creciente. Por lo tanto ambos estadisticos, el de Hotelling y el de CV, son equivalentes.

12 Analisis de Perfiles

El problema de manera formal puede plantearse del siguinte modo:

Si tenemos $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n$ una muestra aleatoria de vectores con distribucion $N_d(\mu, \Sigma)$, donde μ y Σ son desconocidos, y deseamos testear

$H_0 : A\mu = \mathbf{b}$ donde $A \in \mathbb{R}^{k \times d}$, $\mathbf{b} \in \mathbb{R}^{\mathbf{k}}$ y $Rg(A) = k$

es decir

$$\begin{array}{ll} H_0 : & \mathbf{a}_1'\mu = \mathbf{b_1} \\ & \mathbf{a}_2'\mu = \mathbf{b_2} \\ & \mathbf{a}_3'\mu = \mathbf{b_3} \\ & \quad \vdots \\ & \mathbf{a}_k'\mu = \mathbf{b_k} \end{array}$$

donde la matriz

$$A = \left| \begin{array}{l} \mathbf{a}_1' \dots \\ \mathbf{a}_2' \dots \\ \mathbf{a}_3' \dots \\ \quad \vdots \\ \mathbf{a}_k' \dots \end{array} \right| \text{ y el vector } \mathbf{b} = \left| \begin{array}{l} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_k \end{array} \right|$$

podemos realizar las siguientes transformaciones lineales a los vectores originales

$\mathbf{Y}_1 = A\mathbf{X}_1 \dots \mathbf{Y}_i = A\mathbf{X}_i \dots \mathbf{Y}_n = A\mathbf{X}_n$

asi $\mathbf{Y}_i \sim N_k(A\mu, A\Sigma A')$ y la hipotesis a testear se convierte en $H_0 : \mu_Y = \mathbf{b}$ con $\mu_{\mathbf{Y}} = \mathbf{A}\mu$, por lo que podemos usar el estadistico

$$T^2_{d,n-1} = n(\overline{\mathbf{Y}} - \mathbf{b})' S_Y^{*-1} (\overline{\mathbf{Y}} - \mathbf{b})$$

12.1 Algunos ejemplos

13 Comparacion de dos muestras independientes

Sea $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_m$ una muestra aleatoria de vectores con distribucion $N_d(\mu_{\mathbf{X}}, \Sigma_X)$ y sea $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \cdots, \mathbf{Y}_n$ otra muestra aleatoria (independiente de la anterior) de vectores con distribucion $N_d(\mu_{\mathbf{Y}}, \Sigma_Y)$, con $\mu_{\mathbf{X}}$, $\mu_{\mathbf{Y}}$, Σ_X y Σ_Y desconocidos, y deseamos testear $H : \mu_{\mathbf{X}} = \mu_Y$ versus $K : \mu_{\mathbf{X}} \neq \mu_Y$, una hipotesis mas general seria

$H_0 : \mu_X - \mu_Y = \mathbf{b}$ versus $K : \mu_X - \mu_Y \neq \mathbf{b}$ con $\mathbf{b}$ conocido.

Hay que separar el problema en dos casos:

- Las matrices de covarianza coinciden ($\Sigma_X = \Sigma_Y$).
- Las matrices de covarianza son distintas ($\Sigma_X \neq \Sigma_Y$).

13.1 Igual matriz de covarianzas

Sea $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_m$ una muestra aleatoria de vectores con distribucion $N_d(\mu_{\mathbf{X}}, \Sigma)$ y sea $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \cdots, \mathbf{Y}_n$ otra muestra aleatoria (independiente de la anterior) de vectores con distribucion $N_d(\mu_{\mathbf{Y}}, \Sigma)$, con $\mu_{\mathbf{X}}$, $\mu_{\mathbf{Y}}$ y Σ desconocidos, y deseamos testear:

$H_0 : \mu_X - \mu_Y = \mathbf{b}$ versus $K : \mu_X - \mu_Y \neq \mathbf{b}$ con $\mathbf{b}$ conocido.

Los estimadores de maxima verosimilitud en este caso son

$\hat{\mu}_X = \frac{\sum_{i=1}^{n} \mathbf{X_i}}{m} = \overline{\mathbf{X}}$

$\hat{\mu}_Y = \frac{\sum_{i=1}^{n} \mathbf{y_i}}{n} = \overline{\mathbf{Y}}$

$\hat{\Sigma} = \frac{\sum_{i=1}^{m} (\mathbf{X_i} - \overline{\mathbf{X}})(\mathbf{X_i} - \overline{\mathbf{X}})' + \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{Y_i} - \overline{\mathbf{Y}})(\mathbf{Y_i} - \overline{\mathbf{Y}})'}{n+m} = \frac{Q_X + Q_Y}{n+m} = \frac{Q}{n+m} = S$

un estimador insesgado de Σ es

$\frac{\sum_{i=1}^{m} (\mathbf{X_i} - \overline{\mathbf{X}})(\mathbf{X_i} - \overline{\mathbf{X}})' + \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{Y_i} - \overline{\mathbf{Y}})(\mathbf{Y_i} - \overline{\mathbf{Y}})'}{n+m-2} = \frac{Q_X + Q_Y}{n+m-2} = \frac{Q}{n+m-2} = S^*$

llamemos $\delta = \mu_X - \mu_Y$ al parametro de interes

asi las hipotesis quedan

$H : \delta = \mathbf{b}$ versus $K : \delta \neq \mathbf{b}$

un estimador del parametro de interes es $\hat{\delta} = \hat{\mu}_{\mathbf{X}} - \hat{\mu}_{\mathbf{Y}} = \bar{X} - \overline{\mathbf{Y}}$ con esperanza y varianza

$$E(\hat{\delta}) \quad = \quad \delta = \mu_X - \mu_Y$$

$$VAR(\hat{\delta}) \quad = \quad \frac{\Sigma}{m} + \frac{\Sigma}{n} = \frac{(m+n)\Sigma}{mn}$$

asi, bajo H

$$\sqrt{\frac{mn}{m+n}}(\hat{\delta} - \mathbf{b}) \quad \sim \quad N_d(\mathbf{0}, \Sigma)$$

y

$$Q \quad \sim \quad W_d(m+n-2, \Sigma)$$

de esta forma, analogamente al caso de una sola muestra, puede definirse el estadistico de Hotelling

$$T^2_{d,m+n-2} \quad = \quad \frac{mn}{m+n}(\hat{\delta} - \mathbf{b})' S^{*-1} (\hat{\delta} - \mathbf{b}) =$$

$$= \quad \left[\sqrt{\frac{mn}{m+n}} (\hat{\delta} - \mathbf{b}) \right]' \left(\frac{Q}{m+n-2} \right)^{-1} \left[\sqrt{\frac{mn}{m+n}} (\hat{\delta} - \mathbf{b}) \right]$$

y se rechaza H cuando $T^2_{d,m+n-2} > T^2_{d,m+n-2,\alpha}$

13.2 Matrices de covarianzas distintas

Sea $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_m$ una muestra aleatoria de vectores con distribucion $N_d(\mu_{\mathbf{X}}, \Sigma_X)$ y sea $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \cdots, \mathbf{Y}_n$ otra muestra aleatoria (independiente de la anterior) de vectores con distribucion $N_d(\mu_{\mathbf{Y}}, \Sigma_Y)$, con $\mu_{\mathbf{X}}$, $\mu_{\mathbf{Y}}$, Σ_X y Σ_Y desconocidos, y deseamos testear $H : \mu_X - \mu_Y = \mathbf{b}$ versus $K : \mu_X - \mu_Y \neq \mathbf{b}$ con $\mathbf{b}$ conocido.

Este problema puede ser considerado como la version multivariada del problema de Behrens–Fisher. Para un resumen de algunas soluciones propuestas vease *SOME ASPECTS OF MULTIVARIATE BEHRENS-FISHER PROBLEM* de *Junyong Park* y *Bimal Sinha*. Proponemos, analogamente al caso de muestras con identica matriz de varianzas y covarianzas, el estadistico

$$T^2_{d,m+n-2} = \frac{mn}{m+n}(\hat{\delta} - \mathbf{b})' S^{*-1}(\hat{\delta} - \mathbf{b}) =$$

$$= \left[\sqrt{\frac{mn}{m+n}}(\hat{\delta} - \mathbf{b})\right]' \left(\frac{Q}{m+n-2}\right)^{-1} \left[\sqrt{\frac{mn}{m+n}}(\hat{\delta} - \mathbf{b})\right]$$

y se rechaza H cuando $T^2_{d,m+n-2} > F_{T^2_{d,m+n-2}}(1-\alpha)$, siendo $F_{T^2_{d,m+n-2}}$ la funcion de distribucion de la variable aleatoria $T^2_{d,m+n-2}$ bajo la hipotesis nula H. El punto de corte $F_{T^2_{d,m+n-2}}(1-\alpha)$ puede ser estimado mediante la tecnica de Bootstrap Parametrico (PB). Vease *A parametric bootstrap solution to the MANOVA under heteroscedasticity* de *K. Krishnamoorthy* y *Fei Lu.*

13.2.1 Comportamiento asintotico.

Supongamos que $\frac{n}{m+n} \longrightarrow \lambda$, notemos que

$$E(\hat{\delta}) = \delta = \mu_X - \mu_Y$$

asi bajo H

$$E\left(\sqrt{\frac{mn}{m+n}}\left(\hat{\delta} - \mathbf{b}\right)\right) = \mathbf{0}$$

por otro lado

$$VAR\left(\sqrt{\frac{mn}{m+n}}\hat{\delta}\right) = \left(\tfrac{mn}{m+n}\right) \left(\frac{\Sigma_X}{m} + \frac{\Sigma_Y}{n}\right) =$$

$$= \frac{n\Sigma_X}{m+n} + \frac{m\Sigma_Y}{m+n} \longrightarrow \lambda\Sigma_X + (1-\lambda)\Sigma_Y = \Sigma$$

y

$$S^* = \tfrac{Q_X+Q_Y}{n+m-2} = \tfrac{(m-1)S^*_X+(n-1)S^*_Y}{n+m-2} \longrightarrow (1-\lambda)\,\Sigma_X + \lambda\Sigma_Y$$

sabemos que

$$\frac{mn}{m+n}(\hat{\delta}-\mathbf{b})'\Sigma^{-1}(\hat{\delta}-\mathbf{b}) \sim \quad \chi^2_d$$

entonces si $\lambda = \frac{1}{2}$, de tal forma que $\lambda = 1-\lambda$, y bajo la hipotesis H

$$T^2_{d,m+n-2} \quad = \quad \frac{mn}{m+n}(\hat{\delta}-\mathbf{b})'S^{*-1}(\hat{\delta}-\mathbf{b}) \xrightarrow{D} \chi^2_d$$

si queremos un estadistico con un comportamiento 'razonable' para cualquier λ podriamos definir

$$S^P = \frac{(n-1)S^*_X+(m-1)S^*_Y}{n+m-2} \longrightarrow \lambda\Sigma_X + (1-\lambda)\,\Sigma_Y$$

de esta forma nos aseguramos que

$$\frac{mn}{m+n}(\hat{\delta}-\mathbf{b})'S^{P-1}(\hat{\delta}-\mathbf{b}) \xrightarrow{D} \chi^2_d$$

14 Distancia de Mahalanobis

Sean $\mathbf{X_1} \in \mathbb{R}^{\mathbf{d}}$ y $\mathbf{X_2} \in \mathbb{R}^{\mathbf{d}}$ dos observaciones multivariadas y Σ una matriz simetrica definida positiva (generalmente la matriz de varianzas-covarianzas), se define la distancia de Mahalanobis entre $\mathbf{X_1}$ y X_2 a

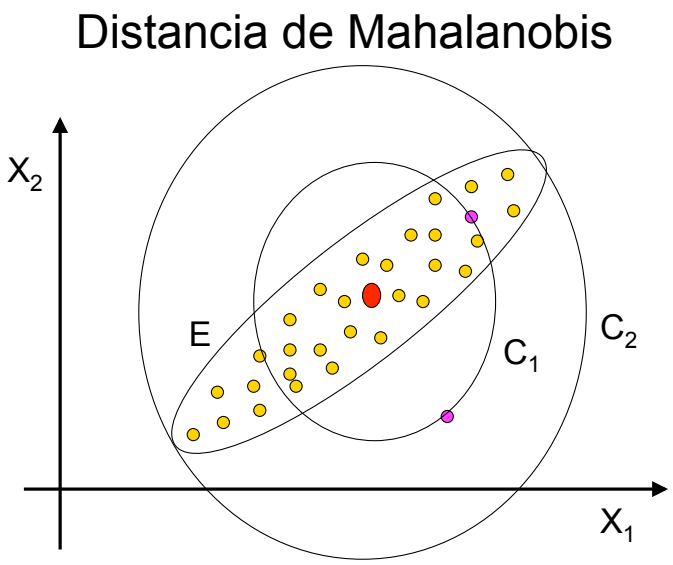
$$DM_\Sigma(\mathbf{X_1},\mathbf{X_2}) \quad = \quad (\mathbf{X_1}-\mathbf{X_2})'\Sigma^{-1}(\mathbf{X_1}-\mathbf{X_2})$$

es equivalente a la distancia euclidea cuadrada de las observaciones transformadas $\mathbf{Z}_1 = \Sigma^{-1/2}\mathbf{X_1}$ e $\mathbf{Z}_2 = \Sigma^{-1/2}\mathbf{X_2}$, es decir

$$DM_\Sigma(\mathbf{X_1},\mathbf{X_2}) \quad = \quad (\mathbf{X_1}-\mathbf{X_2})'\Sigma^{-1}(\mathbf{X_1}-\mathbf{X_2}) =$$

$$= \quad (\mathbf{X_1}-\mathbf{X_2})'\Sigma^{-1/2}\Sigma^{-1/2}(\mathbf{X_1}-\mathbf{X_2}) =$$

$$= \quad (\mathbf{\Sigma^{-1/2}}(\mathbf{X_1}-\mathbf{X_2}))'(\Sigma^{-1/2}(\mathbf{X_1}-\mathbf{X_2})) = D^2(\mathbf{Z_1},\mathbf{Z_2})$$



14.1 Descomposicion de la Distancia de Mahalanobis

Let T^2 be the Mahalanobis distance, the MYT decomposition is

$$T^2 = T_1^2 + T_{2/1}^2 + T_{3/1,2}^2 + T_{4/1,2,3}^2 + \ldots + T_{p/1,2,3,4,\ldots,p-1}^2$$

where the first and last term are

$T_1^2 = \frac{(x_1 - \overline{x}_1^2)2}{S_1^2}$ and $T_{p/1,2,3,4,\ldots,p-1}^2 = \frac{(x_p - \overline{x}_{p/1,2,3,4,\ldots,p-1}^2)^2}{S_{p/1,2,3,4,\cdots,p-1}^2}$

We define the contribution of variable p to the distance as

$$C_p = \frac{T_{p/1,2,3,4,\ldots,p-1}^2}{T^2}$$

because all terms are positive, the contribution satisfies

$$0 \leq C_p \leq 1$$

From now on we will call

If in a certain day happens that T^2 is big, and $T_{p/1,2,3,4,\ldots,p-1}^2$ is close to 1, then the station/variable p is wrong, because

T_{j/A_j}^2 is small

for all $j \in \{1, 2, 3, 4, \ldots, p-1\}$ and $A_j \subseteq \{1, 2, 3, 4, \ldots, j-1, j+1, \ldots, p-1\}$

The p relevant decompositions are

$$T^2 = T_2^2 + T_{3/2}^2 + T_{4/2,3}^2 + T_{5/2,3,4}^2 + \ldots + T_{1/2,3,4,\ldots,p}^2$$

$$T^2 = T_1^2 + T_{3/1}^2 + T_{4/1,3}^2 + T_{5/1,3,4}^2 + \ldots + T_{2/1,3,4,\ldots,p}^2$$

$$T^2 = T_1^2 + T_{2/1}^2 + T_{4/1,2}^2 + T_{5/1,2,4}^2 + \ldots + T_{3/1,2,4,\ldots,p}^2$$

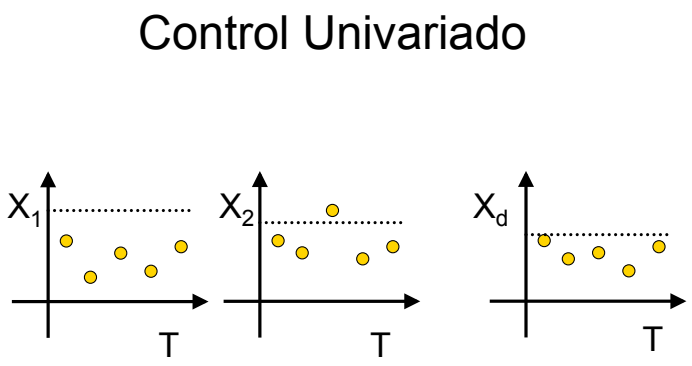
$$\vdots$$

$$T^2 = T_1^2 + T_{2/1}^2 + T_{3/1,2}^2 + T_{4/1,2,3}^2 + \ldots + T_{p-1/1,2,3,4,\ldots,p-2,p}^2$$

$$T^2 = T_1^2 + T_{2/1}^2 + T_{3/1,2}^2 + T_{4/1,2,3}^2 + \ldots + T_{p/1,2,3,4,\ldots,p-1}^2$$

14.2 Aplicacion a control de procesos

El control de procesos variable por variable no resulta



15 Inferencia sobre Matrices de Covarianza - Test de independencia por bloques

Sea $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n$ una muestra aleatoria de vectores con distribucion $N_d(\mu, \Sigma)$, particionamos el vector $\mathbf{X}$ de la siguiente manera

$$\mathbf{X} = \begin{bmatrix} \mathbf{X^{(1)}} \\ \mathbf{X^{(2)}} \end{bmatrix}$$

donde

$$dim(\mathbf{X^{(1)}}) = d_1$$
$$dim(\mathbf{X^{(2)}}) = \mathbf{d_2}$$

por lo que

$$\mu = \begin{bmatrix} \mu^{(\mathbf{1})} \\ \mu^{(\mathbf{2})} \end{bmatrix}$$

y

$$\mathbf{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

La hipotesis que queremos testear es:

H_0 : independencia entre $X^{(1)}$ y $X^{(2)}$, es decir

$H_0 : \Sigma_{12} = \Sigma'_{21} = 0_{d_1 \times d_2}$

15.1 Test del Cociente de Maxima Verosimilitud

Por definicion el estadistico del cociente de Maxima Verosimilitud es

$$CV(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n) \quad =$$

$$= \quad \frac{\max_{\mu,\Sigma} L(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n, \mu, \Sigma)}{\max_{\mu_1 \Sigma_{11} \mu_2 \Sigma_{22}} L(\mathbf{X^{(1)}}_1, \cdots, \mathbf{X^{(1)}}_n, \mu_{\mathbf{1}}, \Sigma_{11}) L(\mathbf{X^{(1)}}_1, \cdots, \mathbf{X^{(1)}}_n, \mu_{\mathbf{1}}, \Sigma_{11})}$$

Recordemos que

$$\ln L(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n, \mu, \Sigma) = \quad LL(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n, \mu, \Sigma) =$$

$$= -\tfrac{nd}{2} \ln(2\pi) - \tfrac{n}{2} \ln(|\Sigma|) - \tfrac{1}{2} \textstyle\sum_{i=1}^{n} (\mathbf{X_i} - \mu)' \Sigma^{-1} (X_i - \mu)$$

el maximo de la log verosimilitud se obtiene tomando

$\mu = \frac{\sum_{i=1}^{n} \mathbf{X_i}}{n} = \overline{\mathbf{X}}$

$\Sigma = \frac{\sum_{i=1}^{n} (\mathbf{X_i} - \overline{\mathbf{X}})(\mathbf{X_i} - \overline{\mathbf{X}})'}{n} = \frac{Q}{n} = S$

asi, el maximo es

$$\max_{\mu,\Sigma} LL(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n, \mu, \Sigma) \quad = \quad -\frac{nd}{2} \ln(2\pi) - \frac{n}{2} \ln(|S|) - \frac{nd}{2}$$

asi

$$\max_{\mu,\Sigma} L(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n, \mu, \Sigma) \quad = \quad = C\, |S|^{-\frac{n}{2}}$$

volviendo al estadistico

$$CV(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n) \quad = \quad \frac{C\, |S|^{-\frac{n}{2}}}{C_1\, |S_1|^{-\frac{n}{2}}\, C_2\, |S_2|^{-\frac{n}{2}}} = \frac{C}{C_1 C_2} \left[\frac{|S|}{|S_1|\, |S_2|} \right]^{-\frac{n}{2}}$$

$$= C^* \left[\frac{|S_1|\, |S_2|}{|S|} \right]^{\frac{n}{2}}$$

que es equivalente a

$$CV(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n) \quad = \quad \frac{|S_1|\, |S_2|}{|S|}$$

este estadistico tiene una interpretacion interesante en terminos de la varianza generalizada de los vectores, y el test rechaza si el estadistico es lo suficientemente grande, es decir si

$$CV(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n) \quad = \quad \frac{|S_1|\, |S_2|}{|S|} > K_\alpha$$

con un poco mas de algebra matricial puede demostrarse que

$$CV(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n) \quad = \quad \frac{|S_1|\,|S_2|}{|S|} = \frac{|S_2|}{|S_{2/1}|}$$

donde

$$S_{2/1} \quad = S_{22} - S_{21} S_{11}^{-1} S_{12}$$

es un estimador de la matriz de covarianzas de la distribucion condicional de $\mathbf{X^{(2)}}$ dado $\mathbf{X^{(1)}}$.

Para llevar a cabo el test es conveniente utilizar la siguinte version del estadistico

$$CV^* \quad = \quad \frac{|S|}{|S_1|\,|S_2|} = \frac{|S_{2/1}|}{|S_2|}$$

bajo H_0 el estadistico CV^* sigue una distribucion $U(d_2, d_1, n - d_{1-1})$. Vease pagina 43 del libro Multivariate Observations (Seber).

15.2 Test derivado del principio de union e interseccion

Definamos las siguintes variables aleatorias (unidimensionales)

$$U_a \quad = \quad \mathbf{a}'\mathbf{X}^{(1)}$$

$$V_b \quad = \quad \mathbf{b}'\mathbf{X}^{(2)}$$

podemos pensar en la siguiente hipotesis unidimensional

H_{ab} :$\rho(U_a, V_b) = 0$

es claro que la hipotesis

$H_0 : \Sigma_{12} = \Sigma_{21}' = 0_{d_1 \times d_2}$

es equivalente a

$H_0 : \rho(U_a, V_b) = 0 \; \forall a \in \mathbb{R}^{d_1}, b \in \mathbb{R}^{d_2}$
esta correlacion (univariada) puede ser estimada por

$$\hat{\rho}(U_a, V_b) \quad = \hat{\rho}_{a,b} = \quad \frac{\sum_{i=1}^n (U_{a,i} - \bar{U}_a)(V_{b,i} - \bar{V}_b)}{\left[\sum_{i=1}^n (U_{a,i} - \bar{U}_a)^2 \sum_{i=1}^n (V_{b,i} - \bar{V}_b)^2\right]^{1/2}}$$

el test univariado rechaza cuando

$$\hat{\rho}_{a,b} \quad > \quad K_\alpha$$

En cuanto a la hipotesis multivariada $H_0 : \Sigma_{12} = \Sigma_{21}' = 0_{d_1 \times d_2}$ se rechaza cuando

$$\sup_{a,b} \hat{\rho}^2_{a,b} \quad > \quad K^*_\alpha$$

reemplazando los terminos univariados por los multivariados queda

$$\hat{\rho}_{a,b} = \quad \frac{\sum_{i=1}^n (U_{a,i} - \bar{U}_a)(V_{b,i} - \bar{V}_b)}{\left[\sum_{i=1}^n (U_{a,i} - \bar{U}_a)^2 \sum_{i=1}^n (V_{b,i} - \bar{V}_b)^2\right]^{1/2}} =$$

$$= \quad \frac{\sum_{i=1}^n (\mathbf{a}'\mathbf{X}^{(1)} - \mathbf{a}'\bar{\mathbf{X}}^{(1)})(\mathbf{b}'\mathbf{X}^{(2)} - \mathbf{b}'\bar{\mathbf{X}}^{(2)})}{\left[\sum_{i=1}^n (\mathbf{a}'\mathbf{X}^{(1)} - \mathbf{a}'\bar{\mathbf{X}}^{(1)})^2 \sum_{i=1}^n (\mathbf{b}'\mathbf{X}^{(2)} - \mathbf{b}'\bar{\mathbf{X}}^{(2)})^2\right]^{1/2}} =$$

$$= \quad \frac{\sum_{i=1}^n \mathbf{a}'(\mathbf{X}^{(1)} - \bar{\mathbf{X}}^{(1)})(\mathbf{X}^{(2)} - \bar{\mathbf{X}}^{(2)})'\mathbf{b}}{\left[\sum_{i=1}^n \left[\mathbf{a}'(\mathbf{X}^{(1)} - \bar{\mathbf{X}}^{(1)})\right]^2 \sum_{i=1}^n \left[\mathbf{b}'(\mathbf{X}^{(2)} - \bar{\mathbf{X}}^{(2)})\right]^2\right]^{1/2}} =$$

$$= \quad \frac{\mathbf{a}'Q_{1,2}\mathbf{b}}{\left[(\mathbf{a}'Q_{1,1}\mathbf{a})\,(\mathbf{b}'Q_{2,2}\mathbf{b})\right]^{1/2}}$$

asi, la maximizacion de $\hat{\rho}_{a,b} = \frac{\mathbf{a}'Q_{1,2}\mathbf{b}}{[(\mathbf{a}'Q_{1,1}\mathbf{a})(\mathbf{b}'Q_{2,2}\mathbf{b})]^{1/2}}$ es un problema de optimizacion de formas cuadraticas que, despues de algo de algebra, queda

$$\sup_{a,b} \hat{\rho}^2_{a,b} \quad = \quad \theta_1$$

donde $\theta_1, \theta_2, \cdots, \theta_d$ son los autovalores de la matriz $Q_{22}^{-1}Q_{21}Q_{11}^{-1}Q_{12}$, por lo que el test se rechaza si

$$\sup_{a,b} \hat{\rho}^2_{a,b} = \theta_1 \quad > \quad K^*_\alpha$$

la distribucion de θ_1 no tiene expresion en forma cerrada. Cuantiles utiles de la distribucion de $\sup_{a,b} \hat{\rho}^2_{a,b} = \theta_1$ se pueden encontrar en el apendice D.14 de Multivariate Observations (Seber).

16 Componentes Principales (PCA)

16.1 Componentes Principales como metodo de captacion de maxima variabilidad.

En este primer enfoque podemos pensar a Componentes Principales como una tecnica que busca reexpresar un fenomeno en dimension grande (d) en otro de dimension menor de modo tal de preservar (captar) la mayor variabilidad (informacion) posible.

Sea $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_n$ una muestra aleatoria de vectores en $\mathbb{R}^d$, con $\mathbf{E}(\mathbf{X}) = \mu$ y $VAR(\mathbf{X}) = \Sigma$ que por simplicidad de exposicion supondremos conocidos. Sin perdida de generalidad supondremos que

$\mu = 0$, y tomemos k combinaciones lineales arbitrarias que llamaremos componentes

$$\begin{aligned}
\mathbf{Y_1} &= \mathbf{a_1'X} \\
\mathbf{Y_2} &= \mathbf{a_2'X} \\
&\vdots \\
\mathbf{Y_k} &= \mathbf{a_k'X}
\end{aligned}$$

que pueden calcularse matricialmente asi

$$\mathbf{Y} = \mathbf{X}A$$

donde A es la matriz formada por los vectores columnas $\mathbf{a_1} \dots \mathbf{a_k}$. De esta forma convertimos la muestra original de n vectores en $\mathbb{R}^d$ en otra muestra de n vectores en $\mathbb{R}^k$,i.e. $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \cdots, \mathbf{Y}_n$.

Pofriamos definir a estas componentes con la finalidad de maximizar la varianza (univariada) de cada una de ellas, asi

$$\begin{aligned}
\mathbf{a_1} &= \underset{\|\mathbf{a}\|=1}{argmax} VAR(\mathbf{a}'X) \\
\mathbf{a_2} &= \underset{\|\mathbf{a}\|=1 \; y \; a'\Sigma a_1=0}{argmax} VAR(\mathbf{a}'X) \\
\mathbf{a_3} &= \underset{\|\mathbf{a}\|=1 \, , \, a'\Sigma a_1=0 \; y \; a'\Sigma a_2=0}{argmax} VAR(\mathbf{a}'X) \\
&\vdots \\
\mathbf{a_k} &= \underset{\|\mathbf{a}\|=1 \, , \, a'\Sigma a_1=0 \, , \, a'\Sigma a_2=0 \, \dots \, a'\Sigma a_{k-1}=0}{argmax} VAR(\mathbf{a}'X)
\end{aligned}$$

las restricciones $a'\Sigma a_i$ corresponden a pedir que $COV(Y, Y_i) = COV(\mathbf{a'X}, \mathbf{a_i'X}) = \mathbf{a'\Sigma a_i} = \mathbf{0}$. Bajo normalidad esta restriccion implicaria independencia (de las componentes). Claramente si la matriz $\Sigma = cI_d$ (para algun c) la restriccion equivale a pedir ortogonalidad entre las combinaciones.

Podemos reescribir el problema asi

$$\mathbf{a_1} = \underset{\|\mathbf{a}\|=1}{argmax} \mathbf{a}'\Sigma\mathbf{a}$$

$$\mathbf{a_2} = \underset{\|\mathbf{a}\|=1 \; y \; a'\Sigma a_1=0}{argmax} \mathbf{a}'\Sigma\mathbf{a}$$

$$\mathbf{a_3} = \underset{\|\mathbf{a}\|=1 \; , \; a'\Sigma a_1=0 \; y \; a'\Sigma a_2=0}{argmax} \mathbf{a}'\Sigma\mathbf{a}$$

$$\vdots$$

$$\mathbf{a_k} = \underset{\|\mathbf{a}\|=1 \; , \; a'\Sigma a_1=0 \; , \; a'\Sigma a_2=0 \; \ldots \; a'\Sigma a_{k-1}=0}{argmax} \mathbf{a}'\Sigma\mathbf{a}$$

Asi, siendo $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_d$ los d autovalores positivos de Σ y $t_1, t_2, t_3, \ldots, t_d$ sus d correspondientes autovectores, veremos que

$$\mathbf{t_1} = \underset{\|\mathbf{a}\|=1}{argmax} \mathbf{a}'\Sigma\mathbf{a}$$

Demostracion:

Tomemos una forma cuadratica arbitraria $\mathbf{a}'\Sigma\mathbf{a}$ y veamos que su valor esta acotado superiormente por $\mathbf{t_1}'\Sigma\mathbf{t_1} = \lambda_1$.

Veamos primero que $\mathbf{t_1}'\Sigma\mathbf{t_1} = \lambda_1$

$$\mathbf{t_1}'\Sigma\mathbf{t_1} = \mathbf{t_1}'\lambda_1\mathbf{t_1} = \lambda_1\mathbf{t_1}'\mathbf{t_1} = \lambda_1 \|t_1\|^2 = \lambda_1$$

Ahora veamos que $\mathbf{a}'\Sigma\mathbf{a} \leq \lambda_1$ para cualquioer $\mathbf{a}$. Los autovectores forman una base, por lo que el vector $\mathbf{a}$ puede escribirse, para d escalares convenientemente elegidos $y_1 \ldots y_d$, asi

$$\mathbf{a} = \sum_{i=1}^{d} \mathbf{y_i}\mathbf{t_i} \Longrightarrow \|\mathbf{a}\|^2 = \left(\sum_{i=1}^{d} \mathbf{y_i}\mathbf{t_i}\right)' \left(\sum_{i=1}^{d} \mathbf{y_i}\mathbf{t_i}\right)$$

$$= \left(\sum_{i=1}^{d} \mathbf{y_i}\mathbf{t_i}'\right) \left(\sum_{i=1}^{d} \mathbf{y_i}\mathbf{t_i}\right)$$

$$= \sum_{i=1}^{d}\sum_{j=1}^{d} \mathbf{y_i}\mathbf{y_j}\mathbf{t_i}'t_j$$

$$= \sum_{i=1}^{d}\sum_{j=1}^{d} \mathbf{y_i}\mathbf{y_j}I(i=j) = \sum_{j=1}^{d} \mathbf{y_i^2}$$

asi, pedir $\|\mathbf{a}\|^2 = 1 \Longrightarrow \sum_{j=1}^{d} \mathbf{y_i^2} = 1$

ahora

$$\mathbf{a}'\Sigma\mathbf{a} = \left(\sum_{i=1}^{d}\mathbf{y_i}\mathbf{t_i}\right)' \Sigma \left(\sum_{i=1}^{d}\mathbf{y_i}\mathbf{t_i}\right)$$

$$= \sum_{i=1}^{d}\sum_{j=1}^{d}\mathbf{y_i}\mathbf{y_j}\mathbf{t_i'}\Sigma t_j$$

$$= \sum_{i=1}^{d}\sum_{j=1}^{d}\mathbf{y_i}\mathbf{y_j}\mathbf{t_i'}\lambda_j t_j$$

$$= \sum_{i=1}^{d}\sum_{j=1}^{d}\mathbf{y_i}\mathbf{y_j}\lambda_j I(i=j)$$

$$= \sum_{i=1}^{d}\mathbf{y^2_i}\lambda_i \leq \sum_{i=1}^{d}\mathbf{y^2_i}\lambda_1 = \lambda_1\sum_{i=1}^{d}\mathbf{y^2_i} = \lambda_1$$

Un resultado interesante a ser remarcado es el siguiente: Pedir covarianza nula de las componentes Y_i con Y_j implica que $\mathbf{a_i'\Sigma a_j} = 0$ pero como los vectores $\mathbf{a}_i$ y $\mathbf{a}_j$ resultan los autovectores $\mathbf{t_i}$ y $\mathbf{t_j}$ sucede que $\mathbf{a_i'\Sigma a_j} = \mathbf{t_i'\Sigma t_j} = \mathbf{t_i'}\lambda_\mathbf{j}\mathbf{t_j} = \lambda_\mathbf{j}\mathbf{t_j'}\mathbf{t_j}$ entonces $\mathbf{a_i'\Sigma a_j} = 0$ debido a que $\mathbf{t_i't_j} = 0$.

Veamos ahora que lo mismo ocurre para el autovector k-esimo, es decir que si imponemos las restricciones

$$\mathbf{a}'\Sigma\mathbf{t_1} = 0 \Longrightarrow \mathbf{a}'\mathbf{t_1} = 0$$

$$\mathbf{a}'\Sigma\mathbf{t_2} = 0 \Longrightarrow \mathbf{a}'\mathbf{t_2} = 0$$

$$\vdots$$

$$\mathbf{a}'\Sigma\mathbf{t_{k-1}} = 0 \Longrightarrow \mathbf{a}'\mathbf{t_{k-1}} = 0$$

entonces

$$\mathbf{t_k} = \underset{\|\mathbf{a}\|=1}{argmax}\ \mathbf{a}'\Sigma\mathbf{a}$$

primero veamos que las anteriores restricciones imponen al vector la condicion de generarse solo en terminos de los autovectores $\mathbf{t_k}\ldots\mathbf{t_d}$

$$\mathbf{a}'\mathbf{t_j} = 0 \Leftrightarrow \left(\sum_{i=1}^{d}\mathbf{y_i}\mathbf{t_i}\right)'\mathbf{t_j} = 0 \Leftrightarrow$$

$$\Leftrightarrow \quad \sum_{i=1}^{d} \mathbf{y_i t_i'} t_j = 0 \Leftrightarrow y_j \left\| t_j \right\|^2 + \sum_{i \neq j} \mathbf{y_i t_i'} t_j = 0 \Longrightarrow y_j = 0$$

y esto se cumple para $1 \leq j \leq k-1$, asi

$$\mathbf{a} \quad = \quad \sum_{i=k}^{d} \mathbf{y_i t_i}$$

veamos

$$\mathbf{a'\Sigma a} \quad = \quad \left(\sum_{i=k}^{d} \mathbf{y_i t_i} \right)' \Sigma \left(\sum_{i=k}^{d} \mathbf{y_i t_i} \right)$$

$$= \quad \sum_{i=k}^{d} \sum_{j=k}^{d} \mathbf{y_i y_j t_i'} \Sigma t_j$$

$$= \quad \sum_{i=k}^{d} \sum_{j=k}^{d} \mathbf{y_i y_j t_i'} \lambda_j t_j$$

$$= \quad \sum_{i=k}^{d} \sum_{j=k}^{d} \mathbf{y_i y_j} \lambda_j I(i=j)$$

$$= \quad \sum_{i=k}^{d} \mathbf{y^2_i} \lambda_i \leq \sum_{i=k}^{d} \mathbf{y^2_i} \lambda_k = \lambda_k \sum_{i=1}^{d} \mathbf{y^2_i} = \lambda_k$$

Resumiendo, las componentes principales resultan ser

$$\mathbf{Y_1} \quad = \quad \mathbf{t_1' X}$$

$$\mathbf{Y_2} \quad = \quad \mathbf{t_2' X}$$

$$\vdots$$

$$\mathbf{Y_k} \quad = \quad \mathbf{t_k' X}$$

cada una de las cuales es una variable aleatoria (unidimensional) con varianza

$$VAR(Y_1) \quad = VAR \quad (\mathbf{t_1' X}) = \mathbf{t_1' \Sigma t_1} = \lambda_1$$

$$VAR(Y_2) \quad = VAR \quad (\mathbf{t_2'X}) = \mathbf{t}_2'\Sigma\mathbf{t}_2 = \lambda_2$$

$$\vdots$$

$$VAR(Y_k) \quad = VAR \quad (\mathbf{t_k'X}) = \mathbf{t}_k'\Sigma\mathbf{t}_k = \lambda_k$$

$$\vdots$$

$$VAR(Y_d) \quad = VAR \quad (\mathbf{t_d'X}) = \mathbf{t}_d'\Sigma\mathbf{t}_d = \lambda_d$$

y con covarianzas nulas, es decir $COV(Y_j, Y_i) = 0$ para $i \neq j$. Resulta de importancia el siguiente cociente

$$Proporcion\ de\ varianzas(k) \quad = \quad \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{d} \lambda_i}$$

es claro que

$$0 \leq \quad \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \quad \leq 1$$

y nos gustaria hallar un valor pequenio de k (i.e. $k = 2$) que de una proporcion de varianzas alta (i.e. ≥ 0.8). De ocurrir esto, y recordando el Teorema de la Descomposicion Espectral tenemos

$$\Sigma = T\Lambda T' = \sum_{i=1}^{d} \lambda_i \mathbf{t_i t_i'} \quad \approx \quad \sum_{i=1}^{k} \lambda_i \mathbf{t_i t_i'}$$

De esta forma, la Proporcion de varianzas antes definida puede ser considerada tanto como:

- La proporcion de de las sumas de las varianzas retenidas por las primeras k componentes principales. Esto solo involucra a las varianzas.
- La bondad de aproximacion de la matriz de varianzas-covarianzas producida por los k autovectores principales. Esto ultimo involucra a las covarianzas.

Dado que usualmente la matriz Σ es desconocida, las componentes principales se definen en funcion de una estimacion de la misma $\widehat{\Sigma} = S^*$. De esta manera todas las propiedades relacionadas con varianzas y covarianzas se cumplen a nivel muestral (no necesariamente poblacional).

16.2 Algunas propiedades necesarias

Propiedad 1: Sea $L \subset \mathbb{R}^d$ un subespacio de dimension k, sea $\{\mathbf{C_1}, C_2, \cdots, C_k\}$ una base ortonormal de L y formemos la matriz $C = [\mathbf{C_1}, C_2, \cdots, C_k] \in \mathbb{R}^{d \times k}$. Se cumple lo siguiente:

$$P(\mathbf{X}, L) \quad = \quad CC'\mathbf{X}$$

es decir, la proteccion del vector $\mathbf{X}$ sobre el subespacio L esta dada por $CC'\mathbf{X}$.

Demostracion:

Alcanza con probar

1. $P(\mathbf{X}, L) \in L$
2. $\mathbf{X} - P(\mathbf{X}, L) \perp L$

Empecemos por el punto 2 y veamos que $\mathbf{X} - P(\mathbf{X}, L)$ es ortogonal a todos los vectores que conforman la base de L

$$C'\left(\mathbf{X} - P(\mathbf{X}, L)\right) \quad = \quad C'\left(\mathbf{X} - CC'\mathbf{X}\right) = C'\mathbf{X} - C'CC'\mathbf{X} = \mathbf{C'X} - \mathbf{C'X} = \mathbf{0}$$

para el punto 1 veamos que $P(\mathbf{X}, L)$ es una combinacion lineal de los vectores que conforman la base de L

$$P(\mathbf{X}, L) \quad = \quad CC'\mathbf{X} = CU = \sum_{i=1}^{k} \mathbf{C_i} u_i$$

Propiedad 2: Sea la matriz $Q \in \mathbb{R}^{d \times d}$ y la matriz $C \in \mathbb{R}^{d \times k}$, con $k < d$, tal que $C'C = I_k$. Se cumple lo siguiente:

$$\sum_{i=1}^{k} \lambda_i(C'QC) \leq \sum_{i=1}^{k} \lambda_i(Q)$$

Demostracion:

Sean $\lambda_1 \geq \lambda_2 \geq \lambda_3 \leq \ldots \geq \lambda_d$ los d autovalores positivos de Q y $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \ldots, \mathbf{b}_d$ sus d correspondientes autovectores, armemos la matriz B

$$B \quad = \quad [\mathbf{b_1}, \mathbf{b}_2, \mathbf{b}_3, \cdots, \mathbf{b}_d] \in \mathbb{R}^{d \times d}$$

que cumple $B'B = BB' = I_d$ por ser ortogonal, y por la descomposicion espectral

$$Q \quad = \quad B\Lambda B'$$

con

$$\Lambda \quad = \quad \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & & \\ 0 & 0 & \lambda_3 & & \\ \vdots & & & \ddots & \\ 0 & & & & \lambda_d \end{bmatrix}$$

asi

$$C'QC \quad = \quad C'B\Lambda B'C = D'\Lambda D$$

llamando $D \in \mathbb{R}^{d\times k}$ a la matriz $B'C$, siendo ademas que la matriz $D'\Lambda D$ es simetrica, vemos que

$$\sum_{i=1}^{k} \lambda_i(C'QC) \quad = \quad Tr(C'QC) = Tr(D'\Lambda D) =$$

por definicion de traza y llamando $E = \{e_{ij}\} = D'\Lambda D = (\Lambda^{1/2}D)'(\Lambda^{1/2}D) = H'H$

$$= \sum_{i=1}^{k} e_{ii} \quad = \sum_{i=1}^{k}\sum_{j=1}^{d} h_{ji}^2 \quad = \sum_{i=1}^{k}\sum_{j=1}^{d} \lambda_j d_{ji}^2 = \sum_{j=1}^{d} \lambda_j \sum_{i=1}^{k} d_{ji}^2 = \sum_{j=1}^{d} \lambda_j f_j$$

$$= \quad \sum_{j=1}^{k} \lambda_j f_j + \sum_{j=k+1}^{d} \lambda_j f_j \quad \leq \sum_{j=1}^{k} \lambda_j f_j + \lambda_k \sum_{j=k+1}^{d} f_j =$$

$$= \sum_{j=1}^{k} \lambda_j f_j + \lambda_k \left(\sum_{j=1}^{d} f_j - \sum_{j=1}^{k} f_j \right) = \quad \sum_{j=1}^{k} \lambda_j f_j + \lambda_k \left(k - \sum_{j=1}^{k} f_j \right) =$$

$$= \quad \sum_{j=1}^{k} \lambda_j f_j + \lambda_k \left(\sum_{j=1}^{k} 1 - f_j \right) \quad \leq \sum_{j=1}^{k} \lambda_j \left(f_j + 1 - f_j \right) = \sum_{j=1}^{k} \lambda_j$$

faltaba ver que $\sum_{j=1}^{d} f_j = k$

$$\sum_{j=1}^{d} f_j \quad = \quad \sum_{j=1}^{d}\sum_{i=1}^{k} d_{ji}^2 = Tr\left(D'D\right) = Tr(C'BB'C) = Tr(C'C) = Tr(I_k)$$

16.3 Componentes Principales como una tecnica de proyeccion ortogonal

Demostraremos que Componentes Principales puede ser visto como una tecnica para hallar un subespacio de dimension pequenia (idealmente de dimension no mayor a 2) que se halle suficientemente cerca de las observaciones (o de los vectores aleatorios).

Dado un vector aleatorio $\mathbf{X} \in \mathbb{R}^d$ con $E(\mathbf{X}) = \mathbf{0}$ y $VAR(\mathbf{X}) = \Sigma$ buscamos un subespacio L de dimension $k \ll d$ que cumpla con la siguiente propiedad:

$$\min_{L} \quad E\left(\|\mathbf{X} - P(\mathbf{X}, L)\|^2 \right)$$

donde por $P(\mathbf{X}, L)$ se entiende la proyeccion ortogonal del vector $\mathbf{X}$ en el subespacio L.

Por ortogonalidad

$$\|\mathbf{X} - P(\mathbf{X}, L)\|^2 \quad = \quad \|\mathbf{X}\|^2 - \|P(\mathbf{X}, L)\|^2$$

asi que basta com buscar L que satisfaga

$$\max_{L} \quad E\left(\|P(\mathbf{X}, L)\|^2\right)$$

podemos escribir la matriz de proyeccion $H \in \mathbb{R}^{d\times d}$ como una transformacion lineal

$$P(\mathbf{X}, L) \quad = \quad H\mathbf{X}$$

donde la matriz H por ser de proyeccion debe ser simetrica, idempotente y de rango k, podemos descomponer la matriz de proyaccion asi

$$H \quad = \quad CC'$$

la matriz $C \in \mathbb{R}^{d\times k}$ posee rango k y cumple ademas que $C'C = I_k$. Veamos primero una cota superior para $E\left(\|P(\mathbf{X}, L)\|^2\right)$, de modo tal de saber que tan bien el subespacio (deterministico) L puede aproximar al vector aleatorio $\mathbf{X}$.

$$E\left(\|P(\mathbf{X}, L)\|^2\right) \quad = \quad E\left(\|H\mathbf{X}\|^2\right) = E\left((H\mathbf{X})'(H\mathbf{X})\right) =$$

$$= \quad E\left(\mathbf{X}'H'H\mathbf{X}\right) = E\left(\mathbf{X}'H\mathbf{X}\right) = E\left(\mathbf{X}'CC'\mathbf{X}\right) =$$

$$E\left(Tr\left(\mathbf{X}'CC'\mathbf{X}\right)\right) \quad = \quad E\left(Tr\left(\mathbf{C}'\mathbf{X}\mathbf{X}'C\right)\right) = Tr\left(E\left(\mathbf{C}'\mathbf{X}\mathbf{X}'C\right)\right) =$$

$$= \quad Tr\left(C'E\left(\mathbf{X}X'\right)C\right) = Tr\left(C'\Sigma C\right) = \sum_{i=1}^{k}\lambda_i(C'\Sigma C) \leq \sum_{i=1}^{k}\lambda_i(\Sigma)$$

esta ultima acotacion la hemos probado en la seccion anterior.

Veamos ahora que si tomamos como generadores del subespacio L a los autovectores asociados a los k mayores autovalores de Σ (llamemos L^* a este subespacio) alcanzamos la cota y por ende satisface la propiedad buscada ($\max_{L} E\left(\|P(\mathbf{X}, L)\|^2\right)$). Asi, sea

$$T_k \quad = \quad [\mathbf{t_1}, t_2, t_3, \cdots, t_k] \in \mathbb{R}^{d\times k}$$

la matriz cuyas columnas son los autovectores citaos y

$$L^* \quad = \quad \langle \mathbf{t_1}, \mathbf{t}_2, \mathbf{t}_3, \cdots, \mathbf{t}_k \rangle$$

el espacio generado, siendo la proyeccion del vector $\mathbf{X}$

$$P(\mathbf{X}, L^*) \quad = \quad T_k T_k' \mathbf{X}$$

la esperanza de la norma cuadrada de la proyeccion es

$$E\left(\|P(\mathbf{X}, L^*)\|^2\right) \quad = \quad E\left(\|T_k T_k' \mathbf{X}\|^2\right) = E\left(\left(T_k T_k' \mathbf{X}\right)' \left(T_k T_k' \mathbf{X}\right)\right) =$$

$$= \quad E\left(X' T_k T_k' T_k T_k' \mathbf{X}\right) \quad = E\left(X' T_k T_k' \mathbf{X}\right) =$$

$$E\left(Tr\left(\mathbf{X}' T_k T_k' \mathbf{X}\right)\right) \quad = \quad E\left(Tr\left(T_k{}' \mathbf{X}\mathbf{X}' T_k\right)\right) = Tr\left(E\left(T_k{}' \mathbf{X}\mathbf{X}' T_k\right)\right) =$$

$$= \quad Tr\left(T_k' E\left(\mathbf{X}X'\right) T_k\right) = Tr\left(T_k' \Sigma T_k\right) = \sum_{i=1}^{k} \lambda_i (T_k' \Sigma T_k)$$

recordando que

$$\Sigma \quad = \quad T \Lambda T' \Longrightarrow T_k' \Sigma T_k = T_k' T \Lambda T' T_k = \Lambda_k$$

siendo $\Lambda_k \in \mathbb{R}^{k \times k}$ la matriz diagonal de los primeros k autovalores de Σ. De esta forma

$$E\left(\|P(\mathbf{X}, L^*)\|^2\right) \quad = \quad \sum_{i=1}^{k} \lambda_i (\Lambda_k) = \sum_{i=1}^{k} \lambda_i$$

es decir que alcanza la cota superior y por ende es el maximo.

Calculemos por ultimo el valor esperado de la norma cuadrada de la diferencia entre el vector $\mathbf{X}$ y su proyeccion

$$E\left(\|\mathbf{X} - P(\mathbf{X}, L^*)\|^2\right) \quad = \quad E\left(\|\mathbf{X}\|^2 - \|P(\mathbf{X}, L^*)\|^2\right) = E\left(\|\mathbf{X}\|^2\right) - \sum_{i=1}^{k} \lambda_i$$

calculando

$$E\left(\|\mathbf{X}\|^2\right) \quad = \quad E\left(\mathbf{X}'\mathbf{X}\right) = Tr\left(E\left(\mathbf{X}'\mathbf{X}\right)\right) = E\left(Tr\left(\mathbf{X}'\mathbf{X}\right)\right) = E\left(Tr\left(\mathbf{X}\mathbf{X}'\right)\right) =$$

$$= \quad Tr\left(E\left(\mathbf{X}\mathbf{X}'\right)\right) \quad = Tr(\Sigma) = \sum_{i=1}^{d} \lambda_i$$

resultando

$$E\left(\|\mathbf{X} - P(\mathbf{X}, L^*)\|^2\right) \quad = \quad \sum_{i=1}^{d} \lambda_i - \sum_{i=1}^{k} \lambda_i = \sum_{i=k+1}^{d} \lambda_i$$

Seria deseable que exista un k chico, mucho mas pequenio que d, que tenga una distancia esperada pequeña entre el vector $\mathbf{X}$ y su proyeccion.

16.4 Componentes Principales como metodo para resumir (comprimir) informacion.

Tenemos un vector $\mathbf{X}$ de dimension d y queremos reemplazrlo por otro vector $\mathbf{Y}$ de dimension mas pequenia (k) de modo tal que la perdida de informacion sea la menor posible. En terminos formales:

Dado un vector aleatorio $\mathbf{X} \in \mathbb{R}^d$ con $E(\mathbf{X}) = \mathbf{0}$ y $VAR(\mathbf{X}) = \Sigma$ buscamos una funcion h $h : \mathbb{R}^k \longrightarrow \mathbb{R}^d$ lineal y una funcion g de reduccion de dimension $g : \mathbb{R}^d \longrightarrow \mathbb{R}^k$ que cumplan con la siguiente propiedad:

$$\min_{h \in \mathbb{H}, g} \quad E\left(\|\mathbf{X} - h(g(\mathbf{X}))\|^2\right)$$

cuando $\mathbb{H} = \{h(\mathbf{y}) = A\mathbf{y}\}$, con $A \in \mathbb{R}^{d \times k}$ e $\mathbf{y} \in \mathbb{R}^k$, es la clase de funciones lineales, la solucion al problema son las componentes principales.

Demostracion:

Dados $\mathbf{X} \in \mathbb{R}^{\mathbf{d}}$ y $A \in \mathbb{R}^{d \times k}$, para todo $\mathbf{y} \in \mathbb{R}^k$ se cumple que

$$\|\mathbf{X} - A\mathbf{y}\|^2 \quad \geq \quad \|\mathbf{X} - P(L_A, \mathbf{X}))\|^2$$

donde L_A es el subespacio generado por las columnas de la matriz A, de esta forma

$$E(\|\mathbf{X} - Ag(\mathbf{X})\|^2) \quad \geq \quad E(\|\mathbf{X} - P(L_A, \mathbf{X}))\|^2) \geq E(\|\mathbf{X} - P(L^*, \mathbf{X}))\|^2) =$$

$$= \quad E(\|\mathbf{X} - T_k T_k' \mathbf{X}\|^2)$$

donde la ultima desigualdad fue demostrada en la seccion anterior. Asi el minimo se alcanza en

$$g(\mathbf{X}) \quad = \quad T_k' \mathbf{X}$$

y

$$h(\mathbf{y}) \quad = \quad T_k \mathbf{y}$$

16.5 El espacio de las Componentes Principales.

Hemos visto, siguiendo los enfoques anteriores, que de ser posible, puede representarse una parte importante de la informacion en un subespacio de dimension menor al original. Parece razonable expresar las observaciones originales (de dimension d) en las k coordenadas sugeridas por el metodo de PCA. Asi definamos

$$\mathbf{Y} \quad = \quad T_k' \mathbf{X}$$

este vector $\mathbf{Y}$ pertenece a un espacio mas chico ($\mathbb{R}^k$) y puede representarse con tan solo k coordendas. Una propiedad importante que

posee esta representacion es la de preservar las distancias del espacio original. Mas especificamente:

$$d\left(P(\mathbf{X_1}, L^*), P(\mathbf{X_2}, L^*)\right)^2 = \left\|P(\mathbf{X_1}, L^*) - P(\mathbf{X_2}, L^*)\right\|^2 =$$

$$= \left\|T_k T_k' \mathbf{X_1} - T_k T_k' \mathbf{X_2}\right\|^2 = \left\|T_k T_k' \left(\mathbf{X_1} - \mathbf{X_2}\right)\right\|^2 =$$

$$= \left(T_k T_k' \left(\mathbf{X_1} - \mathbf{X_2}\right)\right)' \left(T_k T_k' \left(\mathbf{X_1} - \mathbf{X_2}\right)\right) =$$

$$= \left(\mathbf{X_1} - \mathbf{X_2}\right)' T_k T_k' T_k T_k' \left(\mathbf{X_1} - \mathbf{X_2}\right) =$$

$$= \left(\mathbf{X_1} - \mathbf{X_2}\right)' T_k T_k' \left(\mathbf{X_1} - \mathbf{X_2}\right) =$$

$$= \left(\mathbf{T_k'X_1} - T_k'\mathbf{X_2}\right)' \left(T_k'\mathbf{X_1} - T_k'\mathbf{X_2}\right) =$$

$$= \left\|T_k'\mathbf{X_1} - T_k'\mathbf{X_2}\right\|^2 =$$

$$= \left\|\mathbf{Y_1} - \mathbf{Y_2}\right\|^2 = d\left(\mathbf{Y_1}, \mathbf{Y_2}\right)^2$$

16.6 Las Componentes Principales desde una perspectiva geometrica de Rotacion-Reflexion y Truncamiento del espacio original.

Por ultimo las componentes principales pueden ser vistas como el resultado de una rotacion-reflexion del espacio original, seguida de un truncamiento (reduccion de dimension) de los ejes rotados de menor varianza. Sea

$$\mathbf{Z} = T'\mathbf{X}$$

veamos que el vector $\mathbf{Z} \in \mathbb{R}^{\mathbf{d}}$ es el resultado de rotar o reflejar (transformacion rigida o isometria) al vector $\mathbf{X}$ de modo tal de alinear las direcciones principales con los ejes canonicos. La matriz T, que define una tranformacion lineal de $\mathbb{R}^d$ en $\mathbb{R}^d$, induce una rotacion o reflexion pues es una matriz ortogonal de determinante de modulo uno ($|T| = 1$ o $|T| = -1$) . A su vez,

$$\mathbf{Y} = T_k'\mathbf{X} = \begin{bmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \ddots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \end{bmatrix} T'\mathbf{X} = [\mathbf{I_{k\times k}} \mathbf{0_{k\times d-k}}]\mathbf{Z}$$

Las coordenadas truncadas (de $k+1$ a d) son las de menor varianza en el espacio rotado. Se puede ver entonces que el vector de las componentes principales resulta de rotar o reflejar en terminos de las direcciones principales y truncar las direcciones no principales (secundarias).

16.7 Biplots

Un biplot es un grafico, generalmente bidimensional ($k = 2$), capaz de representar con un cierto grado de aproximacion tanto a las observaciones multidimensionales (dimension d mayor a 2) como asi tambien a las d variables. Las observaciones se representan mediante puntos en el grafico, mientras que las variables se representan mediante flechas.

16.7.1 Representacion de las n observaciones

Ya hemos visto que las primeras k componentes principales son

$$\begin{aligned} \mathbf{Y_1} &= \mathbf{t_1'X} \\ \mathbf{Y_2} &= \mathbf{t_2'X} \\ &\vdots \\ \mathbf{Y_k} &= \mathbf{t_k'X} \end{aligned}$$

es asi que la observacion i-esima puede representarse en $\mathbb{R}^k$ mediante las coordenadas $(\mathbf{Y^i}_1, Y_2^i, \ldots, Y_k^i)$, con

$$\begin{aligned} \mathbf{Y_1^i} &= \mathbf{t_1'X^i} \\ \mathbf{Y_2^i} &= \mathbf{t_2'X^i} \\ &\vdots \\ \mathbf{Y_k^i} &= \mathbf{t_k'X^i} \end{aligned}$$

Esta representacion posee la ventaja de que pese a estar en una dimension mas chica ($k < d$) preserva razonablemente bien las distancias originales entre observaciones, pues como vimos anteriormente

$$d\left(\mathbf{X_1}, X_2\right)^2 \approx d\left(P(\mathbf{X_1}, L^*), P(\mathbf{X_2}, L^*)\right)^2 \quad = \quad d\left(\mathbf{Y_1}, \mathbf{Y_2}\right)^2$$

donde $\mathbf{X_1}$ y $\mathbf{X_2}$ son dos observaciones cualesquiera en el espacio original.

16.7.2 Representacion de las d variables

Definamos, para $i = 1 \dots d$, al vector $\mathbf{v_i} \in \mathbb{R}^{\mathbf{k}}$ del siguiente modo

$$\mathbf{v_i} \quad = \quad \begin{bmatrix} \sqrt{\lambda_1} t_{i1} \\ \vdots \\ \sqrt{\lambda_k} t_{ik} \end{bmatrix}$$

donde t_{ij} es el elemento correspondiente a la fila i-esima columna j-esima de la matriz $T_k = \{t_{ij}\} \in \mathbb{R}^{d \times k}$. Veremos que este vector $\mathbf{v_i}$ es una "buena representacion" k dimensional de la vairable (original) i-esima. Es decir:

- El angulo formado entre los vecores $\mathbf{v_i}$ y $\mathbf{v_k}$ es una buena aproximacion a la correlacion existente entre la variable i y la variable k.
- El modulo del vector $\mathbf{v_i}$ es una buena aproximacion de la varianza de la variable i-esima.

Recordando la descomposicion espectral y suponiendo que las primeras k componentes explican una proporcion importante de la suma de varianzas, tenemos

$$\Sigma = T \Lambda T' = \sum_{i=1}^{d} \lambda_i \mathbf{t_i} \mathbf{t_i'} \quad \approx \quad \sum_{i=1}^{k} \lambda_i \mathbf{t_i} \mathbf{t_i'}$$

la covarianza entre la variable i-esima y la variable j-esima se puede escribir

$$\sigma_{ij} = \sum_{h=1}^{d} \lambda_h t_{ih} t_{jh} \quad \approx \quad \sum_{h=1}^{k} \lambda_h t_{ih} t_{jh} = \sum_{h=1}^{k} \left(\sqrt{\lambda_h} t_{ih} \right) \left(\sqrt{\lambda_h} t_{jh} \right) =$$

$$= \quad \mathbf{v_i'} \mathbf{v_j}$$

por lo que podemos aproximar la covarianza como

$$\sigma_{ij} \quad \approx \quad \mathbf{v_i'} \mathbf{v_j}$$

y la varianza de la variable i-esima como

$$\sigma_{ii} \quad \approx \quad \mathbf{v_i'} \mathbf{v_i} = \|\mathbf{v_i}\|^{\mathbf{2}}$$

pero por otro lado sabemos que

$$\mathbf{v_i'} \mathbf{v_j} \quad = \quad \|\mathbf{v_i}\| \, \|\mathbf{v_j}\| \, cos(\alpha_{ij})$$

donde α_{ij} denota el angulo formado entre los vectores $\mathbf{v_i}$ y $\mathbf{v_j}$. Por lo tanto

$$cos(\alpha_{ij}) \quad = \quad \frac{\mathbf{v_i'} \mathbf{v_j}}{\|\mathbf{v_i}\| \, \|\mathbf{v_j}\|} \approx \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}} \sqrt{\sigma_{jj}}} = cor(x_{,i} \, x_j)$$

16.7.3 Relacion entre observaciones y variables

Seria deseable que exista una relacion grafica entre observaciones y variables, es decir, que aquellas observaciones "mas alineadas" a ciertas variables, reflejen valores importantes de esas observaciones en esas variables. Mas especificamente nos gustaria que la proyeccion ortogonal de una observacion (punto del biplot) en un vector (flecha del biplot) sea una buena aproximacion del valor original que esa observacion tiene en la coordenada correspondiente de la variable original.

Tomemos una observacion cualquiera $\mathbf{X}$ en el espacio original ($\mathbb{R}^d$), proyectemosla en el subespacio generado por las direcciones principales ($P(\mathbf{X}, L^*)$) y veamos como podemos expresar su coordenada i-esima en funcion del vector $\mathbf{Y}$ (punto del biplot) y del vector $\mathbf{v_i}$ (flecha del biplot).

$$P(\mathbf{X}, L^*) \quad = \quad T_k T_k' \mathbf{X} = T_k \mathbf{Y}$$

llamemos $P(\mathbf{X}, L^*)^i$ a la coordenada i-esima del vector $P(\mathbf{X}, L^*)$, entonces

$$\mathbf{P(X, L^*)^i} \quad = \quad \sum_{h=1}^{k} t_{ih} \mathbf{y_h} = \sum_{h=1}^{k} \sqrt{\lambda_h} t_{ih} \frac{\mathbf{y_h}}{\sqrt{\lambda_\mathbf{h}}} = \mathbf{v_i'} \begin{bmatrix} \frac{Y_1}{\sqrt{\lambda_1}} \\ \vdots \\ \frac{Y_k}{\sqrt{\lambda_k}} \end{bmatrix} =$$

$$= \mathbf{v}_i' \left(\Lambda_k^{-1/2} \mathbf{Y} \right) = \mathbf{v}_i' \mathbf{Y^s}$$

de esta manera

$$\mathbf{P(X, L^*)^i} \quad = \quad \mathbf{v}_i' \mathbf{Y^s} = \|\mathbf{v_i}\| \, \|\mathbf{Y^s}\| \, cos(\alpha_{\mathbf{v_i Y^s}})$$

La felicidad no es completa ! Si en lugar del vector $\mathbf{Y^s}$ tuviesemos al vector $\mathbf{Y}$, el biplot constituido con los puntos $\mathbf{Y}$ y las flechas $\mathbf{v_i}$ poseeria las siguientes propiedades:

- Las distancias entre obsvaciones en el espacio original se aproximan por las distancias entre puntos del biplot.
- Las correlaciones entre variables se aproximan por los angulos entre flechas del biplot.
- Las coordenadas de las observaciones en las variables originales pueden aproximarse por la proyeccion de los puntos en lss flechas del biplot.

Dado que la relacion entre observaciones y variables se obtiene solo con $\mathbf{Y^s}$ (y no con $\mathbf{Y}$), se dispone de dos posibilidades:

- Realizar el biplot con $\mathbf{Y^s}$ y perder la interpretacion de distancias entre observaciones.
- Realizar el biplot con $\mathbf{Y}$ y perder la relacion entre observaciones y variables.

Veamos por ultimo que la primer opcion (trabajar con $\mathbf{Y^s}$) brinda, sin embargo, una interpretacion muy util en terminos de distancias entre puntos, es decir, las distancias representadas en el biplot entre los puntos $\mathbf{Y^s}$ aproximan a las distancias de Mahalanobis de las observaciones originales. Veamos

$$d\left(\mathbf{Y_1^s}, \mathbf{Y_2^s}\right)^2 \quad = \quad \left\|\mathbf{Y}_1^s - \mathbf{Y}_2^s\right\|^2 =$$

$$= \quad \left\|\Lambda_k^{-1/2}\mathbf{Y_1} - \Lambda_k^{-1/2}\mathbf{Y_2}\right\|^2 \quad = \left\|\Lambda_k^{-1/2}T_k'\mathbf{X_1} - \Lambda_k^{-1/2}T_k'\mathbf{X_2}\right\|^2 = \left\|\Lambda_k^{-1/2}T_k'\left(\mathbf{X_1} - \mathbf{X_2}\right)\right\|^2 =$$

$$= \quad \left(\Lambda_k^{-1/2}T_k'\left(\mathbf{X_1} - \mathbf{X_2}\right)\right)' \left(\Lambda_k^{-1/2}T_k'\left(\mathbf{X_1} - \mathbf{X_2}\right)\right) \quad =$$

$$= \quad \left(\mathbf{X_1} - \mathbf{X_2}\right)' T_k \Lambda_k^{-1/2}\Lambda_k^{-1/2}T_k'\left(\mathbf{X_1} - \mathbf{X_2}\right) \quad =$$

$$= \quad \left(\mathbf{X_1} - \mathbf{X_2}\right)' T_k \Lambda_k^{-1}T_k'\left(\mathbf{X_1} - \mathbf{X_2}\right) \quad =$$

y viendo que $T_k'T = \begin{bmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \ddots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \end{bmatrix} = [I_{k\times k}0_{k\times d-k}]$

$$= \quad \left(\mathbf{X_1} - \mathbf{X_2}\right)' T_k T_k' T \Lambda^{-1} T' T_k T_k'\left(\mathbf{X_1} - \mathbf{X_2}\right) \quad =$$

$$= \quad \left(\mathbf{X_1} - \mathbf{X_2}\right)' T_k T_k' \Sigma^{-1} T_k T_k'\left(\mathbf{X_1} - \mathbf{X_2}\right) \quad =$$

$$= \quad \left(T_k T_k'\mathbf{X_1} - T_k T_k'\mathbf{X_2}\right)' \Sigma^{-1}\left(T_k T_k'\mathbf{X_1} - T_k T_k'\mathbf{X_2}\right) \quad =$$

$$= \quad \left(P(\mathbf{X_1}, L^*) - P(\mathbf{X_2}, L^*)\right)' \Sigma^{-1}\left(P(\mathbf{X_1}, L^*) - P(\mathbf{X_2}, L^*)\right) \quad =$$

$$= DM_\Sigma \quad \left(P(\mathbf{X_1}, L^*), P(\mathbf{X_2}, L^*)\right)$$

17 Ejercicio de Componentes Principales

Basado en el conjunto de datos "crimen.csv", que contiene informacion de tasas delictivas, para cada uno de los estados de USA, medidas en un periodo de tiempo, se pide realizar un Analisis de Componentes Principales del mismo.

Los datos conforman una matriz de 50 filas (estados) y 9 columnas (variables). Las variables se detallan a continuacion:

- STATEN: Nombre del estado de EEUU.

- STATE: ID del estado de EEUU.
- MURDER: Tasa de asesinatos por cada 100000 habitantes.
- RAPE: Tasa de violaciones por cada 100000 habitantes.
- ROBBERY: Tasa de robos por cada 100000 habitantes.
- ASSAULT: Tasa de ataques violentos por cada 100000 habitantes.
- BURGLARY: Tasa de robo de casas por cada 100000 habitantes.
- LARCENY: Tasa de hurtos por cada 100000 habitantes.
- AUTO: Tasa de robo de automotores por cada 100000 habitantes.

Se solicita realizar las siguientes consignas:

1. Calcular la matriz de covarianzas.
2. Calcular los autovectores (e interpretar) de la matriz de covarianzas.
3. Calcular los autovalores (e interpretar) de la matriz de covarianzas.
4. Calcular las proyecciones de las 50 observaciones (estados) en el espacio generado por las primras dos direcciones principales.
5. Calcular las coordenadass (scores) de las 50 observaciones (estados) de las primeras dos direcciones principales.
6. Elegir una cantidad conveniente de factores ($k = 2$) y realizar dos Biplots, uno con las observaciones estandarizadas y el otro con las observaciones sin estandarizar. Interpretar el mismo y confrontar con los datos originales.
7. Evaluar, desde el punto de vista matricial, la aproximacion que se produce con las primeras dos direcciones principales. Es la bondad de la aproximacion uniforme ?
8. Calcular la matriz de correlaciones y repetir el analisis anterior.
9. Comparar el analisis basado en la matriz de covarianzas con el de la matriz de correlaciones. Cual le parece mas razonable y por que ?

18 La Descomposicion en Valores Singulares (SVD)

Habiendo visto previamente la descomposicion espectral, la Descomposicion en Valores Singulares puede ser considerada como una generalizacion de la primera. La SVD posee la ventaja de poder relacionar dos conjuntos de variables ($\mathbf{X}$ e $\mathbf{Y}$) mediante la factorizacion de la matriz de covarianzas $COV(\mathbf{X}, Y)$.

18.1 Teorema de la Descomposicion en Valores Singulares

Sea la matriz $M \in \mathbb{R}^{n \times p}$, que sin perdida de generalidad supondremos que satisface $n \geq p$. La misma puede ser factorizada de la siguiente manera

$$M = U\Lambda V'$$

donde, $U \in \mathbb{R}^{n \times n}$ es una matriz ortogonal, $\Lambda \in \mathbb{R}^{n \times p}$ es una matriz diagonal y $V \in \mathbb{R}^{p \times p}$ es una matriz ortogonal.

La matriz U es una matriz ortogonal de rango n conformada por vectores columna que reciben el nombre de Vectores Singulares a Izquierda

$$U = [\mathbf{u_1}, \mathbf{u}_2, \mathbf{u}_3, \cdots, \mathbf{u}_n] \in \mathbb{R}^{n \times n}$$

que cumple $U'U = UU' = I_n$ por ser ortogonal.

La matriz V es tambien una matriz ortogonal de rango p conformada por vectores columna que reciben el nombre de Vectores Singulares a Derecha

$$V = [\mathbf{v_1}, \mathbf{v}_2, \mathbf{v}_3, \cdots, \mathbf{v}_p] \in \mathbb{R}^{p \times p}$$

que cumple $V'V = VV' = I_n$ por ser ortogonal.

La matriz Λ es una matriz diagonal con elementos no negativos $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_p \geq 0$, llamados Valores Singulares

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \ddots & & & \\ 0 & 0 & \lambda_p & & \\ \vdots & & & 0 & \\ 0 & & & & 0 \end{bmatrix}$$

18.2 Algunas propiedades importantes de la SVD

18.2.1 La SVD como metodo de maximizacion de formas bilineales

Sea la matriz $M \in \mathbb{R}^{n \times p}$ con $n \geq p$, y la SVD de la misma

$$M = U\Lambda V'$$

donde, $U \in \mathbb{R}^{n \times n}$ es una matriz ortogonal, $\Lambda \in \mathbb{R}^{n \times p}$ es una matriz diagonal y $V \in \mathbb{R}^{p \times p}$ es una matriz ortogonal. Puede demostrrse que

$$(\mathbf{u_1}, \mathbf{v}_1) = \underset{\|\mathbf{u}\|=1 \; y \; \|\mathbf{v}\|=1}{argmax} \mathbf{u}'M\mathbf{v} \quad \text{con} \quad \mathbf{u_1}'M\mathbf{v_1} = \lambda_1$$

$$(\mathbf{u_2}, \mathbf{v}_2) = \underset{\|\mathbf{u}\|=1, \, \|\mathbf{v}\|=1, \, \mathbf{u}'\mathbf{u}_1=0 \, \text{y} \;\; \mathbf{v}'\mathbf{v}_1 = 0}{argmax} \mathbf{u}'M\mathbf{v} \quad \text{con} \quad \mathbf{u_2}'M\mathbf{v_2} = \lambda_2$$

$$\vdots$$

$$(\mathbf{u_p}, \mathbf{v}_p) \quad = \quad \underset{\|\mathbf{u}\|=1,\, \|\mathbf{v}\|=1,\, \mathbf{u'u_i}=0 \text{ y } \mathbf{v'v}_i=0}{argmax} \quad \mathbf{u}'M\mathbf{v} \quad \text{con} \quad \mathbf{u_p}'M\mathbf{v_p} = \lambda_p$$

para $i = 1, 2, \ldots p-1$.

Asi los valores singulares $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_d$ son los valores maximos de las formas bilineales y estos maximos se obtienen evaluando la forma en los vectores singulares $\mathbf{u_1}, \mathbf{u}_2, \mathbf{u}_3, \cdots, \mathbf{u}_p$ y $\mathbf{v_1}, \mathbf{v}_2, \mathbf{v}_3, \cdots, \mathbf{v}_p$.

18.2.2 La SVD como metodo de aproximacion de matrices

Dada la matriz no necesariamente simetrica $M \in \mathbb{R}^{n \times p}$ de rango p, con $n \geq p$, y la SVD de la misma

$$M \quad = \quad U\Lambda V'$$

puede demostrarse que la matriz C de rango $k \leq p$ que mejor aproxima a M es

$$C = \sum_{i=1}^{k} \lambda_i \mathbf{u_i} \mathbf{v_i'} \approx M = \sum_{i=1}^{p} \lambda_i \mathbf{u_i} \mathbf{v_i'}$$

Mas especificamente

$$C \quad = \quad \underset{A:rango(A)=k}{argmin} \|A - M\|_F$$

donde $\|.\|_F$ denota la norma Frobenius, es decir

$$\|A - M\|_F \quad = \quad \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{p} (a_{ij} - m_{ij})^2}$$

con $A = \{a_{ij}\}$ y $M = \{m_{ij}\}$.

18.3 Ejercicio teorico basado en la Descomposicion en Valores Singulares

Sean dos vetores aleatorios $\mathbf{X} = [x_1, x_2, \ldots, x_n]$ e $\mathbf{Y} = [y_1, y_2, \ldots, y_p]$, usando la Descomposicion en Valores Singulares, caracterice y describa la relacion entre $\mathbf{X}$ e $\mathbf{Y}$.

Sugerencia: Inspirese en el analisis realizado bajo la tecnica de Componentes Principales.

19 Nociones de Varianzas Generalizadas

19.1 La traza de la matriz de Varianzas-Covarianzas

Una nocion razonable de variabilidad generalizada empirico es la de calcular un promedio muestral de las distancias euclideas cuadradas al centro.

$$V = \frac{1}{n}\sum_{i=1}^{n} d^2(\mathbf{X_i}, \bar{\mathbf{X}})$$

que por ser un escalar

$$= Tr \left(\frac{1}{n}\sum_{i=1}^{n} d^2(\mathbf{X_i}, \bar{\mathbf{X}})\right)$$

$$= Tr\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X_i} - \bar{\mathbf{X}})'(\mathbf{X_i} - \bar{\mathbf{X}})\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} Tr\left((\mathbf{X_i} - \bar{\mathbf{X}})'(\mathbf{X_i} - \bar{\mathbf{X}})\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} Tr\left((\mathbf{X_i} - \bar{\mathbf{X}})(\mathbf{X_i} - \bar{\mathbf{X}})'\right)$$

$$= Tr\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X_i} - \bar{\mathbf{X}})(\mathbf{X_i} - \bar{\mathbf{X}})'\right)$$

$$= Tr\,(S)$$

19.2 El determinante de la matriz de Varianzas-Covarianzas

20 Teoria estadistica de la decision

La Teoria Estadistica de la deision es un enfoque general que permite pensar a los metodos de regresion y de clasificacion en un mismo marco. Empecemos con una variable "a ser explicada" (Y) cuantitativa.

20.1 Variable Y continua

Sea un vector aleatorio $\mathbf{X} \in \mathbb{R}^d$ de variables explicativas, y sea $Y \in \mathbb{R}$ la variable "a ser explicada". La funcion de densidad conjunta de $\mathbf{X}$ e Y es $f(\mathbf{X}, Y)$. Buscamos una funcion $g(\mathbf{X})$ que prediga a Y dado un valor de $\mathbf{X} = \mathbf{x}$. Necesitamos definir una funcion de perdida que

mida la distancia de $g(\mathbf{X})$ a Y de modo tal de penalizar los errores de prediccion

$$L(Y, g(\mathbf{X}))$$

siendo la funcion de perdida clasica la funcion de perdida cuadratica

$$L(Y, g(\mathbf{X})) \quad = \quad [Y - g(\mathbf{X})]^2$$

Un critrio razonable para elegir la funcion g es pedir minimizar el valor esperado de la funcion de perdida, es decir, el error esperado de prediccion es en este caso el Error Cuadratico de Prediccion de la funcion g

$$ECP(g) = E\left(L(Y, g(\mathbf{X}))\right) \quad = \quad E\left[Y - g(\mathbf{X})\right]^2 = \int \left[Y - g(\mathbf{X})\right]^2 f(\mathbf{X}, Y) =$$

$$= E_{\mathbf{X}} E_{Y/\mathbf{X}}([Y - g(\mathbf{X})]^2 \mid \quad \mathbf{X})$$

como en la practica el vector $\mathbf{X}$ esta fijo ($\mathbf{X} = \mathbf{x}$), tiene sentido condicionar a $\mathbf{X}$, por lo que la funcion buscada $\hat{g}(\mathbf{X})$ es la que minimiza, para cada $\mathbf{x}$, la esperanza en Y(dado $\mathbf{X} = \mathbf{x}$)

$$\hat{g}(\mathbf{x}) = \underset{c}{argmin} E_{Y/\mathbf{X}}([Y - c]^2 \mid \quad \mathbf{X} = \mathbf{x})$$

y el valor c que satisface esto es la esperanza condicional

$$\hat{g}(\mathbf{x}) = E_Y(Y \mid \quad \mathbf{X} = \mathbf{x})$$

pues la esperanza condicional es el valor que minimiza el error cuadratico esperado. A esta funcion se la denomina genericamente Funcion de Regresion.

20.2 Variable Y categorica

Supondremos ahora que la variable Y puede tomar los valores $Y = 1, 2, \ldots, k$, es decir, existen k poblaciones a las que pueden pertenecer las observaciones. Asi las funciones predictoras $f : \mathbb{R}^d \to \{1, 2, \ldots, k\}$ toman valores enteros. En este caso la funcion de densidad conjunta del vector $\mathbf{X}$ e Y es

$$f(\mathbf{X}, Y) = Pr(\mathbf{X} = \mathbf{x}, Y = i) = Pr(\mathbf{X} = \mathbf{x} | Y = i) Pr(Y = i) = \quad f_i(\mathbf{X})\pi_i$$

con $i \in \{1, 2, \ldots, k\}$,donde

$f_i(\mathbf{X})$ es la funcion de densidad de las variables explicativas de la poblacion i-esima.

π_i la probabilidad, a priori, que un elemento pertenezca a la poblacion i-esima.

Hay que definir una funcion de perdida conveniente para este problema. En general podriamos pensar que si la funcion g predice el verdadero valor de Y la perdida debiera ser 0, y el error para los demas casos dependera del verdadero valor $Y = i$ y del valor predicho $g(\mathbf{X}) = j$. En los problemas de clasificacion a la funcion de perdida se la conoce en general cmo funcion de costos $C(.)$ (de mala clasificacion). Esta funcion es aleatoria, pues depende tanto del vector $\mathbf{X}$ como de Y, asi definimos

$$L(Y, g(\mathbf{X})) \quad = \quad \sum_{i=1}^{k}\sum_{j=1}^{k} C(j|i) I(Y = i) I(\mathbf{X} \in R_j)$$

Es fundamental notar que la sentencia $g(\mathbf{X}) = j$ es equivalente a $I(\mathbf{X} \in R_j) = 1$, de esta manera la definicion de un metodo de clasificacion ($g(.)$) establece una particion del espacio $\mathbb{R}^d$ en k regiones $\Re = (R_1, R_2, \ldots, R_k)$ tales que $\mathbb{R}^d = \underset{i}{\cup} R_i$ con $R_i \cap R_j = \emptyset$ para $i \neq j$.

Calculemos ahora el error esperado de prediccion

$$EP(g) = E\left(L(Y, g(\mathbf{X}))\right) \quad = \quad E\left[\sum_{i=1}^{k}\sum_{j=1}^{k} C(j|i) I(Y = i) I(\mathbf{X} \in R_j)\right] =$$

$$\sum_{i=1}^{k}\sum_{j=1}^{k} C(j|i) E\left[I(Y = i) I(\mathbf{X} \in R_j)\right] \quad = \quad \sum_{i=1}^{k}\sum_{j=1}^{k} C(j|i) Pr\left[Y = i, \mathbf{X} \in R_j\right] =$$

$$\sum_{i=1}^{k}\sum_{j=1}^{k} C(j|i) Pr\left(\mathbf{X} \in R_j | \mathbf{Y} = i)\right) Pr(\mathbf{Y} = i) \quad = \quad \sum_{i=1}^{k}\sum_{j=1}^{k} C(j|i) \int \ldots \int_{\mathbf{x} \in R_j} f_i(\mathbf{x}) d\mathbf{x} \pi_i =$$

$$\int \ldots \int_{\mathbf{x} \in \mathbb{R}} \sum_{j=1}^{k} I(\mathbf{X} \in R_j) \sum_{i=1}^{k} C(j|i) \pi_i f_i(\mathbf{x}) d\mathbf{x} = \quad \int \ldots \int_{\mathbf{x} \in \mathbb{R}} \sum_{j=1}^{k} I(\mathbf{X} \in R_j) h_j(\mathbf{x}) d\mathbf{x}$$

donde $h_j(\mathbf{x}) = \sum_{i=1}^{k} C(j|i) \pi_i f_i(\mathbf{x}) = \sum_{i \neq j} C(j|i) \pi_i f_i(\mathbf{x})$ solo depende de $\mathbf{x}$. Es importante notar que el metodo de clasificacion se halla determinado por la expresion $I(\mathbf{X} \in R_j)$, que define la particion. Por lo tanto

$$EP(g) = E\left(L(Y, g(\mathbf{X}))\right) \quad = \quad \int \ldots \int_{\mathbf{x} \in \mathbb{R}} \sum_{j=1}^{k} I(\mathbf{X} \in R_j) h_j(\mathbf{x}) d\mathbf{x}$$

Ahora vamos a ver que para encontrar la particion que minimice esta expresion alcanza con elegir, para cada $\mathbf{x}$, la region (j) que minimice $h_j(\mathbf{x})$, es decir que la *regla* de clasificacion optima sera

$$\mathbf{x} \in R_j \quad \Leftrightarrow \quad h_j(\mathbf{x}) = \underset{i \in \{1,2,\ldots,k\}}{min} h_i(\mathbf{x})$$

recordando que

$$h_j(\mathbf{x}) \quad = \quad \sum_{i \neq j} C(j|i)\pi_i f_i(\mathbf{x})$$

que puede ser interpretado como el costo esperado de clasificar mal en la poblacion j una obsrvacion con valores observados $\mathbf{x}$.

Veamos que si elegimos las regiones (i.e. la particion) siguiendo la *regla* anterior entonces el Error Esperado de Prediccion es minimo, o sea

$$Regla \quad \Rightarrow \quad Min\, EP$$

Equivalentemente, basta con demostrar que si tomamos una particion (que llamaremos P_2) que no minimiza el Error Esperao de Prediccion entonces la regla anterior no se cumple, asi

$$\overline{Min\, EP} \quad \Rightarrow \quad \overline{Regla}$$

Supongamos que las densidades poblaciones $f_i(\mathbf{x})$ (para $i = 1 \ldots K$) son continuas, y sean dos metodos de clasificacion distintos $g^{p_1}()$ y $g^{p_2}()$ que inducen respectivamente las particiones p_1 y p_2. Y supongamos que $EP\left(g^{p_1}\right) < EP\left(g^{p_2}\right)$, es decir

$$\int \ldots \int_{\mathbf{x}\in\mathbb{R}} \sum_{j=1}^{k} I(\mathbf{X} \in R_j^{p_1}) h_j(\mathbf{x}) d\mathbf{x} \quad = \int \ldots \int_{\mathbf{x}\in\mathbb{R}} H^1(\mathbf{x}) d\mathbf{x} <$$

$$< \quad \int \ldots \int_{\mathbf{x}\in\mathbb{R}} \sum_{j=1}^{k} I(\mathbf{X} \in R_j^{p_2}) h_j(\mathbf{x}) d\mathbf{x} = \int \ldots \int_{\mathbf{x}\in\mathbb{R}} H^2(\mathbf{x}) \mathbf{dx}$$

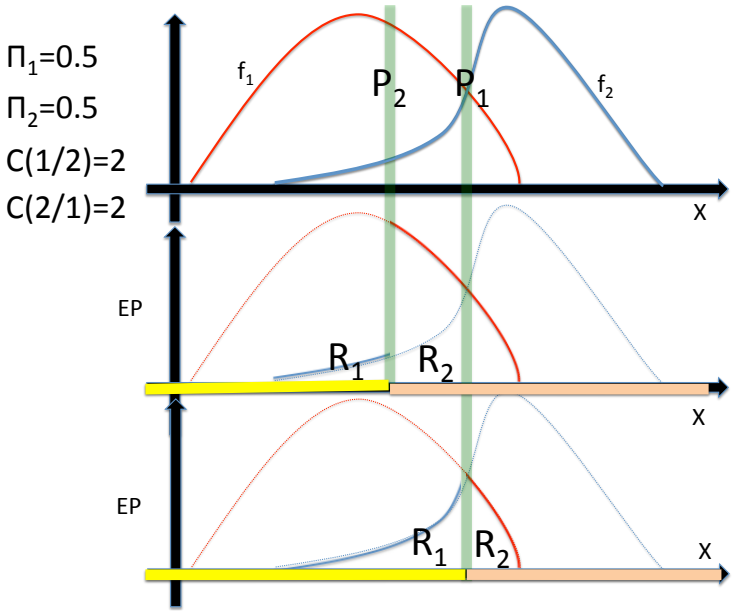
entonces, como las $h_j(\mathbf{x})$ son continuas (pues las $f_i(x)$ lo son), existe una bola $B_\epsilon(\mathbf{x_0}) \subset \mathbb{R}^d$ que satisface

$$\int \ldots \int_{\mathbf{x}\in\mathbf{B}_\epsilon(\mathbf{x_0})} H^1(\mathbf{x}) \mathbf{dx} \quad < \quad \int \ldots \int_{\mathbf{x}\in B_\epsilon(\mathbf{x_0})} H^2(\mathbf{x}) \mathbf{dx}$$

por lo que debe existir un $\mathbf{x_1} \in B_\epsilon(\mathbf{x_0})$ que cumple $H^1(\mathbf{x_1}) < H^2(\mathbf{x_1})$ o, lo que es lo mismo

$$h_{j^*}(\mathbf{x_1}) \quad < \quad h_{j^{**}}(\mathbf{x_1})$$

para algun par de regiones $R_{j^*}^{p_1}$ y $R_{j^{**}}^{p_1}$ de las respectivas particiones. Lo cual muestra que para la particion (metodo) p_2 no se cumple la *regla*.



De esta manera parece razonable asignar (clasificar) a una observacion con valores $\mathbf{x}$ en aquella poblacion que brinde el menor costo esperado de mala clasificacion. Es importante notar que si disponemos de un criterio (como el recien mencionado) para clasificar una observacion cualquiera ($\mathbf{x}$) entonces se dispone de un metodo (general) de clasificacion, es decir, de una particion del espacio de covariables.

20.2.1 Costos iguales de mala clasificacion

Veamos el caso particular en el que los costos de mala clasificacion son iguales, es decir, sin importar de que poblacion venga la observacion ni a que poblacion (erronea) se la asigne, el costo de mala clasificacion es $C(j/i) = c$ para todo $i \neq j$. Asi

$$h_j(\mathbf{x}) \quad = \quad c \sum_{i \neq j} \pi_i f_i(\mathbf{x})$$

y buscar el j que minimice $h_j(\mathbf{x})$ es equivalente a buscar el j que maximice la expresion $\pi_j f_j(\mathbf{x})$ pues

$$c\pi_j f_j(\mathbf{x}) \quad = \quad c \sum_{i=1}^{k} \pi_i f_i(\mathbf{x}) - h_j(\mathbf{x}) = ck - h_j(\mathbf{x})$$

donde $\sum_{i=1}^{k} \pi_i f_i(\mathbf{x}) = k$ es constante en j. Pero a su vez maximizar $\pi_j f_j(\mathbf{x})$ es equivalente a maximizar

$$\frac{\pi_j f_j(\mathbf{x})}{\sum_{i=1}^{k} \pi_i f_i(\mathbf{x})} \quad = \quad P(Y = j | \mathbf{X} = \mathbf{x})$$

que recibe la denominacion de Clasificador de Bayes, pues clasifica una observacion en aquela poblacion que maximiza la probabilidad a posteriori de pertenencia.

21 Analisis Discriminante (Clasificacion)

21.1 Metodo linealde Fisher (LDA)

21.2 Metodo cuadratico de Fisher (QDA)

21.3 Regresion Logistica

21.4 Arboles de Clasificacion

21.5 Vecinos mas Cercanos (KNN)

22 Otras tecnicas multivariadas

22.1 Segmentacion

22.2 Analisis Factorial

22.3 Analisis de Correspondencia

22.4 Reglas de Asociacion