Sea $m_{rs} = \sum_{j=1}^{p} \left| x_{rj} - x_{sj} \right|$ la distancia de Manhattan donde P= (x_{r1}, x_{r2},…, x_{rp}) y Q= (x_{s1}, x_{s2},…, x_{sp}) son dos observaciones de una matriz de datos X $_{(nxp)}$.

Probar que es una métrica.

Simetría: d(P,Q) = d(Q,P)

$$m_{rs} = \sum_{j=1}^{p} |Xrj - Xsj|$$

$$m_{sr} = \sum_{j=1}^{p} |Xsj - Xrj|$$

…

Sabemos que: |a|=|-a|

$$|x_{rj}-x_{sj}|=|-(x_{rj}-x_{sj})| \Rightarrow |x_{rj}-x_{sj}|=|x_{sj}-x_{rj}|$$

Tomando sumatorias:

$$\sum_{j=1}^{p} |Xrj - Xsj| = \sum_{j=1}^{p} |Xsj - Xrj|$$

$$\mathbf{m_{rs} = m_{sr}}$$

No negatividad: d(P,Q)>0 , si P ≠ Q

Si P ≠ Q, $\Rightarrow X_{rj} \neq X_{sj}$

$\Rightarrow |Xrj - Xsj| \neq 0$

$\Rightarrow \mathbf{m_{rs} > 0}$

Identidad: d(P,P)=0

$$m_{rr} = \sum_{j=1}^{p} |Xrj - Xrj| = 0$$

$$\Rightarrow |Xrj - Xrj| = 0 \ \forall j$$

Definición: d(P,Q) = 0, si P=Q

Si P=Q $\Rightarrow x_{rj} = x_{sj}, \ \forall \ j$

$$d(P,Q) = m_{rs} = \sum_{j=1}^{p} \sqrt{|Xrj - Xrj|}$$

$$\Rightarrow m_{rs} = 0 \Rightarrow d(P,Q) = 0$$

Desigualdad triangular: $d(P,R)+d(R,Q) \geq d(P,Q)$

Sea $R= (x_{h1}, x_{h2},\ldots, x_{hp})$

Tal que: $|x_{rj}-x_{sj}| = |x_{rj} -x_{hj}+x_{hj}-x_{sj}|$

$\Rightarrow |x_{rj}-x_{sj}| = |x_{rj} -x_{hj}+x_{hj}-x_{sj}| \leq |x_{rj}-x_{hj}|+|x_{hj}-x_{sj}|$

$\Rightarrow \mathbf{m_{rs} \leq m_{rh} + m_{hs}}$

CONCLUSIÓN: $\mathbf{m_{rs}}$ es una métrica

Sea $\mathbf{W}(nxp) = \left(\boldsymbol{w}_i\right)$ tal que $\boldsymbol{w}_i = \mathbf{S}^{-1/2}(\mathbf{x}_i - \bar{x})$. Encontrar la expresión matemática para la matriz de covarianzas.

Analizamos $\boldsymbol{w}_i$:

$$\boldsymbol{w}_i = \mathbf{S}^{-1/2}\,(\mathbf{x}_i - \bar{x})$$

(px1) (pxp) (px1)

$$\boldsymbol{w}'_i = (\mathbf{x}_i - \bar{x})'\,\mathbf{S}^{-1/2}$$

(1xp) (1xp) (pxp)

$$\mathbf{W}_{(nxp)} = \begin{pmatrix} (\mathbf{x}_1 - \bar{x})'\,\mathbf{S}^{-1/2} \\ (\mathbf{x}_2 - \bar{x})'\,\mathbf{S}^{-1/2} \\ \vdots \\ (\mathbf{x}_n - \bar{x})'\,\mathbf{S}^{-1/2} \end{pmatrix} \begin{matrix} (1xp) \\ (1xp) \\ \\ (1xp) \end{matrix}$$

$$\mathbf{W} = \begin{pmatrix} (\mathbf{x}_1 - \bar{x})' \\ (\mathbf{x}_2 - \bar{x})' \\ \vdots \\ (\mathbf{x}_n - \bar{x})' \end{pmatrix} \mathbf{S}^{-1/2} = \left(\mathbf{I}_n - \frac{1}{2}\mathbf{J}_n\right)\mathbf{X}\mathbf{S}^{-1/2} = \mathbf{H}\mathbf{X}\mathbf{S}^{-1/2} \Rightarrow \mathbf{W} = \mathbf{H}\mathbf{X}\mathbf{S}^{-1/2}$$

$$\Rightarrow \mathbf{S}_W = \frac{1}{n}\mathbf{W}'\mathbf{H}\mathbf{W} = \frac{1}{n}\mathbf{S}^{-1/2}\mathbf{X}'\mathbf{H}'\mathbf{H}\mathbf{H}\mathbf{X}\mathbf{S}^{-1/2} = \mathbf{I}_p \Rightarrow \mathbf{S}_W = \mathbf{I}_p$$

Se trata de estudiar la estructura de ventas de una empresa que distribuye 10 productos (i=1,…,n) en 8 mercados (j=1,…,p). Sabiendo que k_{ij} representa el valor de las ventas del producto i en el mercado j en una tabla de contingencia T(i,j), verifique que $d^2(i,i')$ es una medida de distancia. Tal que,

$$d^2(i,i') = \sum_j \left(\frac{f_{ij}}{f_i\sqrt{f_j}} - \frac{f_{i'j}}{f_{i'}\sqrt{f_j}}\right)^2$$

$f_i=k_i/k$, $f_{ij}=k_{ij}/k$, $k=\Sigma_{i,j}k_{ij}$, $k_i=\Sigma_j k_{ij}$, $k_j=\Sigma_i k_{ij}$.

k_{ij} frecuencia absoluta

f_{ij} frecuencia relativa

Tabla de Contingencia

i \ j	1	2	…	8
1	f_{11}	f_{12}		$f_{1,8}$
2	f_{21}	f_{22}		$f_{2,8}$
…				
10	$f_{10,1}$	$f_{10,2}$		$f_{10,8}$

La misma expresión de distancia:

$$d^2(i,i') = \sum_j \frac{1}{f_j}\left[\frac{f_{ij}}{f_i} - \frac{f_{i'j}}{f_{i'}}\right]^2$$

Simetría: d(i,i')=d(i',i)

$$d(i,i') = \sqrt{\sum_j \frac{1}{f_j}\left[\frac{f_{ij}}{f_i} - \frac{f_{i'j}}{f_{i'}}\right]^2}$$

$$= \sqrt{\sum_j \frac{1}{f_j}\left[-\left(\frac{f_{i'j}}{f_{i'}} - \frac{f_{ij}}{f_i}\right)\right]^2}$$

$$= \sqrt{\sum_j \frac{1}{f_j}\left[\frac{f_{i'j}}{f_{i'}} - \frac{f_{ij}}{f_i}\right]^2}$$

$$= d(i',i)$$

No negatividad: d(i,i')>0

Si i≠i' ⇒ para alguna j:

$$\frac{f_{ij}}{f_i} \neq \frac{f_{i'j}}{f_{i'}}$$

$$\Rightarrow \quad \frac{f_{ij}}{f_i} - \frac{f_{i'j}}{f_{i'}} \neq 0$$

$$\Rightarrow \quad \frac{1}{f_j}\left[\frac{f_{ij}}{f_i} - \frac{f_{i'j}}{f_{i'}}\right]^2 > 0$$

$$\Rightarrow \quad d(i,i') > 0$$

Identidad: d(i,i')=0, entonces, i=i'

$$d(i,i')=0 \Rightarrow d^2(i,i') = \sum_j \frac{1}{f_j}\left[\frac{f_{ij}}{f_i} - \frac{f_{i'j}}{f_{i'}}\right]^2 = 0$$

$$\Rightarrow \frac{f_{ij}}{f_i} - \frac{f_{i'j}}{f_{i'}} = 0$$

$$\Rightarrow \frac{f_{ij}}{f_i} = \frac{f_{i'j}}{f_{i'}}$$

$$\Rightarrow i = i'$$

CONCLUSIÓN: **d(i,i')** es una distancia

Demostrar que para un conjunto de datos se cumple lo siguiente:

$$\frac{1}{np}\sum_{j=1}^{n}(x_j-\bar{x})'S^{-1}(x_j-\bar{x})=1$$

Analizando los términos de la sumatoria:

$$\frac{1}{np}\sum_{i=1}^{n}(X_i-\bar{X})'S^{-1}(X_i-\bar{X})$$

(1xp) (pxp) (px1)

(Variables aleatorias univariadas)

$$\frac{1}{np}\sum_{i=1}^{n}(X_i-\bar{X})'S^{-1}(X_i-\bar{X}) = \frac{1}{np}\sum_{i=1}^{n} Tr[(X_i-\bar{X})'S^{-1}(X_i-\bar{X})]$$

$\Rightarrow \quad = \frac{1}{np}\sum_{i=1}^{n} Tr[S^{-1}(X_i - \bar{X})(X_i - \bar{X})']$ (propiedad de traza)

$= \frac{1}{np} Tr[\sum_{i=1}^{n} S^{-1}\frac{n}{n}(X_i - \bar{X})(X_i - \bar{X})']$

$= \frac{1}{np} Tr[nS^{-1}\sum_{i=1}^{n}\frac{1}{n}(X_i - \bar{X})(X_i - \bar{X})']$

$= \frac{1n}{np} Tr[S^{-1}S] = \frac{1n}{np} Tr[I_{pxp}] = \frac{1n}{np} p = 1$

$\Rightarrow \quad \frac{1}{np}\sum_{i=1}^{n}(X_i - \bar{X})'S^{-1}(X_i - \bar{X}) = 1$

Sea $\mathbf{D}^2_{ij}$ el cuadrado de la distancia de Mahalanobis entre las observaciones $\mathbf{x}_i$, $\mathbf{x}_j \in R^p$ y $\mathbf{S}=(1/n)\mathbf{X'HX}$ la matriz de covarianzas a partir de la muestra $\mathbf{X}_{(nxp)}$.

Demuestre que $\sum_i \sum_j \mathbf{D}^2_{ij} = 2n^2p$.

De los datos del problema:

$$D_{ij}^2 = (X_i - X_j)' S^{-1} (X_i - X_j) \qquad \text{(i)}$$

Sumamos y restamos $\bar{X}$ en cada paréntesis

$$D_{ij}^2 = [X_i - \bar{X}]' S^{-1} [X_i - \bar{X}] + [X_j - \bar{X}]' S^{-1} [X_j - \bar{X}] - 2 [X_i - \bar{X}]' S^{-1} [(X_j - \bar{X}]$$

Sea:

$$g_{ii} = [X_i - \bar{X}]' S^{-1} [X_i - \bar{X}]$$
$$g_{jj} = [X_j - \bar{X}]' S^{-1} [X_j - \bar{X}]$$
$$g_{ij} = [X_i - \bar{X}]' S^{-1} [(X_j - \bar{X}]$$

} en (i)

$$\Rightarrow \quad D_{ij}^2 = g_{ii} + g_{jj} - 2\, g_{ij} \qquad \text{(ii)}$$

Analizamos cada componente:

$$g_{ii} = [X_i - \bar{X}]' S^{-1} [X_i - \bar{X}]$$

(1xp) (pxp) (px1)

(Variable aleatoria univariada)

$$\Rightarrow \quad g_{ii} = tr[X_i - \bar{X}]' S^{-1} [X_i - \bar{X}]$$

$$g_{ii} = tr\,[S^{-1}\,[X_i - \bar{X}][X_i - \bar{X}]'\,]$$ (Propiedad de traza)

Sumatoria sobre i:

$$\sum_i g_{ii} = \sum_i tr[S^{-1}[X_i - \bar{X}][X_i - \bar{X}]'\,]$$

$$\sum_i g_{ii} = tr[S^{-1}\sum_i [X_i - \bar{X}][X_i - \bar{X}]'\,]$$ (iii)

Sabemos que:

$$S = \frac{1}{n}\sum_i [X_i - \bar{X}][X_i - \bar{X}]'$$ en (iii)

$$\sum_i g_{ii} = tr[S^{-1}nS]$$

$$\Rightarrow \quad \sum_i g_{ii} = n * tr[S^{-1}S]$$

$$\Rightarrow \quad \sum_i g_{ii} = n * tr[I_p]$$

$$\Rightarrow \quad \sum_i g_{ii} = np$$

Sumando respecto a j:

$$\sum_{j,i} g_{ii} = n^2 p \qquad \text{(iv)}$$

Análogamente respecto a g_{jj}:

$$\sum_{i,j} g_{jj} = n^2 p \qquad \text{(v)}$$

Para g_{ij}:

$$g_{ij} = [X_i - \bar{X}]' S^{-1} [(X_j - \bar{X}]$$

$$\sum_{i,j} g_{ij} = \sum_{i,j} [X_i - \bar{X}]' S^{-1} [(X_j - \bar{X}] = 0 \qquad \text{(vi)}$$

De (iv), (v) y (vi) en (ii):

$$D_{ij}^2 = n^2 p + n^2 p - s(0)$$

$$\Rightarrow \quad D_{ij}^2 = 2n^2 p$$

Gracias!!!