Introducción al Análisis Multivariado

SEMANA-3

Raúl Alberto Pérez

Universidad Nacional de Colombia, Escuela de Estadística, 2021-I

Distancias

Introducción

Muchas técnicas importantes del análisis multivariado se basan en el concepto de distancia.

Al medir distancias entre variables, se obtiene una idea de la proximidad entre ellas.

Definición

Dados dos vectores $\underline{\mathbf{x}}, \underline{\mathbf{y}} \in \mathbb{R}^p$, $\underline{\mathbf{x}} = \left(x_1, x_2, \ldots, x_p\right)$ y $\underline{\mathbf{y}} = \left(y_1, y_2, \ldots, y_p\right)$, la distancia euclídea entre ellos se define como sigue:

$$d\left(\underline{\mathbf{x}}, \underline{\mathbf{y}}\right) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_p - y_p)^2} = \|\underline{\mathbf{x}} - \underline{\mathbf{y}}\|$$

Desde el punto de vista estadístico, la distancia euclídea no es muy-satisfatoria, ya que cada coordenada está ponderada por un mismo factor de 1.

Cuando las coordenadas representan medidas sujetas a fluctuaciones aleatorias de diferentes magnitudes (por ejemplo: altura en metros, peso en kilogramos, distancia en kilómetros, etc), es preferible ponderar las coordenadas de acuerdo a su variabilidad.

Es usual usar ponderaciones pequeñas para las coordenadas sujetas a un alto grado de variabilidad.

Debido a esto es necesario desarrollar una distancia que tenga en cuenta la variabilidad y la dependencia (correlación) entre las variables.

Distancia Estadística

Dado un vector aleatorio p-dimensional $\underline{\mathbf{x}} = \left(X_1, X_2, \ldots, X_p\right)$ cuyas componentes X_i´s son independientes, se define la distancias estadística de $\underline{\mathbf{x}}$ al origen de coordenadas $\underline{\mathbf{0}} = (0, 0, \ldots, 0)$ en $\mathbb{R}^p$ como sigue:

$$d\left(\underline{\mathbf{0}}\ ,\ \underline{\mathbf{x}}\right) = \sqrt{\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}} + \ldots + \frac{x_p^2}{s_{pp}}}$$

Similarmente dados dos vectores aleatorios $\underline{\mathbf{x}} = \left(X_1, X_2, \ldots, X_p\right)$ y $\underline{\mathbf{y}} = \left(Y_1, Y_2, \ldots, Y_p\right)$ en $\mathbb{R}^p$ de la misma población, entonces la distancia estadística entre $\underline{\mathbf{x}}$ y $\underline{\mathbf{y}}$ está dada por:

$$d\left(\underline{\mathbf{x}}\ ,\ \underline{\mathbf{y}}\right) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}} + \ldots + \frac{(x_p - y_p)^2}{s_{pp}}}$$

Si $s_{11} = s_{22} = \cdots = s_{pp}$, entonces se puede usar la distancia euclídea con pesos-iguales.

Si las variables de los vectores anteriores no son independientes, entonces las expresiones anteriores no son adecuadas, por lo que hay que definir otras distancias que tengan en cuenta la correlación (o covarianzas) entre las variables.

Distancia de Mahalanobis

La distancia de Mahalanobis entre dos vectores aleatorios $\underline{\mathbf{x}} = \left(X_1, X_2, \ldots, X_p\right)$ y $\underline{\mathbf{y}} = \left(Y_1, Y_2, \ldots, Y_p\right)$ en $\mathbb{R}^p$ se define como sigue:

$$d(\underline{\mathbf{x}}\,,\,\underline{\mathbf{y}}) = (\underline{\mathbf{x}} - \underline{\mathbf{y}})^t \boldsymbol{\Sigma}^{-1} (\underline{\mathbf{x}} - \underline{\mathbf{y}}), \quad \text{si } \boldsymbol{\Sigma} \text{ es conocida,}$$

$$d(\underline{\mathbf{x}}\,,\,\underline{\mathbf{y}}) = (\underline{\mathbf{x}} - \underline{\mathbf{y}})^t S^{-1} (\underline{\mathbf{x}} - \underline{\mathbf{y}}), \quad \text{si } \boldsymbol{\Sigma} \text{ es desconocida}$$

La distancia de Mahalanobis entre un vector aleatorio $\underline{\mathbf{x}} = \left(X_1, X_2, \ldots, X_p\right)$ y su vector de medias $\underline{\mu} = \left(\mu_1, \mu_2, \ldots, \mu_p\right)$ en $\mathbb{R}^p$ se define como sigue:

$$d(\underline{\mathbf{x}}\ ,\ \underline{\mu}) = (\underline{\mathbf{x}} - \underline{\mu})^t \boldsymbol{\Sigma}^{-1} (\underline{\mathbf{x}} - \underline{\mu}), \quad \text{si } \boldsymbol{\Sigma} \text{ y } \underline{\mu} \text{ son conocidas}$$

$$d(\underline{\mathbf{x}}\ ,\ \overline{\mathbf{x}}) = (\underline{\mathbf{x}} - \overline{\mathbf{x}})^t S^{-1} (\underline{\mathbf{x}} - \overline{\mathbf{x}}), \quad \text{si } \boldsymbol{\Sigma} \text{ y } \underline{\mu} \text{ son desconocidas}$$

Ejemplo Práctico:

Los Biólogos **Grojan** y **Wirth** (1981) descubrieron dos nuevas especies de insectos **Ameroheleafasciata** (AF) y **Apseudofasciata** (APF). Puesto que las especies son similares en apariencia, resulta útil para el biólogo estar en capacidad de clasificar un especimen como AF o APF basado en características externas que son fáciles de medir. Entre alguna de las características que distinguen los AF de los APF, los biólogos reportan medidas de la longitud de las antenas (X) y la longitud de las alas (Y), ambas medidas en milímetros, de nueve insectos AF y de seis insectos APF.

Una de las preguntas que motivaron el estudio fue: ¿Será posible encontrar una regla que nos permita clasificar un insecto dado como AF o como APF, basados únicamente en las mediciones de las antenas y las alas? Rta: Si.
Los datos son los siguientes:

Datos del ejemplo

ESPECIE	X	Y
AF	1.38	1.64
AF	1.40	1.20
AF	1.24	1.72
AF	1.36	1.74
AF	1.38	1.82
AF	1.48	1.82
AF	1.54	1.82
AF	1.38	1.90
AF	1.56	2.08
APF	1.14	1.78
APF	1.20	1.86
APF	1.18	1.96
APF	1.30	1.96
APF	1.26	2.00
APF	1.28	2.00

Para el conjunto de datos dado, responda las siguientes preguntas:

1. Construya un gráfico de X e Y. Comente acerca de la apariencia de los datos.

2. Para cada grupo de individuos, calcule el vector de medias muestrales, la matriz de varianzas-covarianzas muestrales, la matriz de correlaciones muestrales, la varianza total y la varianza generalizada. Realice la descomposición espectral de cada una de las dos matrices de Var-Cov asociadas a cada grupo de individuos.

3. Calcule la distancia euclídea entre los vectores de media de cada grupo de individuos.

4. Calcule la distancia de Mahalanobis entre los vectores de media de cada grupo de individuos.

5. ¿Considera razonable usar la distancia de Mahalanobis en cada uno de los dos grupos?