V-3-3. Mahalanobis' Distance

In the multivariance analysis, we often discuss similarity of among data. Generally, we consider similarity as distance between data in multidimensional space. If the similarity of two data is high the point of the data will be plotted in the neighborhood in multidimensional space. However, when they has correlation and variance we have to modify the data considering the correlation and variance.

When the data a and b is expressed as vector $\boldsymbol{a}$ and $\boldsymbol{b}$,

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

The distance is $|\mathbf{a} - \mathbf{b}|$

$$\boldsymbol{a} - \boldsymbol{b} = \begin{pmatrix} a_1 - b_1 \\ a_2 - b_2 \\ a_3 - b_3 \end{pmatrix}$$

$$|\mathbf{a} - \mathbf{b}| = \sqrt{(\boldsymbol{a} - \boldsymbol{b})^T (\boldsymbol{a} - \boldsymbol{b})}$$

$$= \sqrt{(a_1 - b_1 \quad a_2 - b_2 \quad a_3 - b_3) \begin{pmatrix} a_1 - b_1 \\ a_2 - b_2 \\ a_3 - b_3 \end{pmatrix}}$$

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2}$$

When the vectors are transformed by matrix $\boldsymbol{A}$. The distance between $\boldsymbol{Aa}$ and $\boldsymbol{Ab}$ is as follow.

$$|\boldsymbol{Aa} - \boldsymbol{Ab}| = |\boldsymbol{A}(\boldsymbol{a} - \boldsymbol{b})|$$

$$= \sqrt{\big(\boldsymbol{A}(\boldsymbol{a} - \boldsymbol{b})\big)^T \boldsymbol{A}(\boldsymbol{a} - \boldsymbol{b})}$$

$$\sqrt{(a_1 - b_1 \quad a_2 - b_2 \quad a_3 - b_3)\boldsymbol{A}^T\boldsymbol{A} \begin{pmatrix} a_1 - b_1 \\ a_2 - b_2 \\ a_3 - b_3 \end{pmatrix}}$$

$$|\mathbf{A}(\mathbf{a} - \mathbf{b})|^2 = (a_1 - b_1 \quad a_2 - b_2 \quad a_3 - b_3)\boldsymbol{A}^T\boldsymbol{A} \begin{pmatrix} a_1 - b_1 \\ a_2 - b_2 \\ a_3 - b_3 \end{pmatrix}$$

$$|(\mathbf{a} - \mathbf{b})\mathbf{A}|^2 = (\mathbf{a} - \mathbf{b})^T \boldsymbol{A}^T \boldsymbol{A} (\mathbf{a} - \mathbf{b})$$

Here we denote $\mathbf{a} - \mathbf{b} = \boldsymbol{x}$

$$\boldsymbol{x}^T \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{x}$$

When

$$\boldsymbol{A} = \boldsymbol{\Sigma}^{-\frac{1}{2}}$$

$\boldsymbol{\Sigma}$: variancce covatiance matrix

$$|\boldsymbol{Ax}| = \sqrt{\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x}\right)^T \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x}}$$

$$= \sqrt{\boldsymbol{x}^T\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\right)^T \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x}} = \sqrt{\boldsymbol{x}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x}} = \sqrt{\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}$$

$$(\because \boldsymbol{\Sigma}^{-\frac{1}{2}} \text{ is symmetric. } \left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\right)^T = \boldsymbol{\Sigma}^{-\frac{1}{2}})$$

This is Mahalanobis' Distance $D_{\boldsymbol{a-b}} = \sqrt{(\boldsymbol{a}-\boldsymbol{b})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{a}-\boldsymbol{b})}$

$(\boldsymbol{\Sigma}: variance\ covariance\ matrix)$

Formula 70

Mahalanobis' Distance is compensated distance by variance and covariance and Mahalanobis distance is useful tool to consider similarity of data when the data has correlation.