MNStatC Statistics July Problem

Cheenta Statistics and Analytics Department

Happy National Statistics Day

June 29, 2021

> Statistics must have a clearly defined purpose, one aspect of which is scientific advance and the other, human welfare and national development.
>
> *P.C.Mahalanobis*

1 Impact of Mahalanobis Distance on Multivariate Normal

1.1 Introduction

Just like Euclidean distance is defined as the shortest distance between two points on the Euclidean plane, Mahalanobis Distance is also a similar kind of measure which considers the impact of correlation if the point are correlated.

One problem with the Euclidean distance measure is that it does not take the impact of correlation (association) into account. But, often in statistical problem the variability and association of the points are of intrinsic importance.

Given, two vectors in $\underline{X}, \underline{Y} \in \mathbb{R}^n$, we have the Euclidean distance between $\underline{X}, \underline{Y}$, given by

$$d(\underline{X}, \underline{Y}) = \sqrt{(\underline{X} - \underline{Y})^\top (\underline{X} - \underline{Y})}.$$

But while modifying this idea of distance where each variate of $\underline{X}$ on $\underline{Y}$ follows some probability law, Mahalanobis's approach to scale the contribution of individual to the distance according to the variability of each variable.

It is a scale invariant metric and provides a measure of distance between a point $\underline{X} \in \mathbb{R}^n$ generated from a given n-variate probability distribution $f_{\underline{X}}$ with mean vector $\underline{\mu} = \mathbb{E}(\underline{X})$ and a variance - covariance matrix $\sum^{n\times n} = \mathbb{E}[(\underline{X} - \underline{\mu})(X - \underline{\mu})^\top]$.

Then

$$D(\underline{X}, \underline{\mu}) = \sqrt{(\underline{X} - \underline{\mu})^\top \Sigma^{-1} (\underline{X} - \underline{\mu})}$$

is defined as the Mahalanobis Distance. Observe if the variance - covariance matrix is an identity matix, $D(\underline{X}, \underline{Y}) = d(\underline{X}, \underline{Y})$.

1.2 Problem

- Can you elaborate what it implies geometrically when $\underline{X}$ in a random vector from n-variate Normal population?
- Further considering $\underline{X} \sim N_n(\underline{\mu}, \Sigma)$, show that the density at any point on $\mathbb{R}^n$ increases with the decrease in the Mahalanobis Distance. Express the geometric significance of the claim
- Given two random vectors, $\underline{X}, \underline{Y}$, when will you call them equally likely?