Mahalanobis distance

What is the Mahalanobis distance?

The Mahalanobis distance (MD) is the distance between two points in multivariate space. In a regular Euclidean space, variables (e.g. x, y, z) are represented by axes drawn at right angles to each other; The distance between any two points can be measured with a ruler. For uncorrelated variables, the Euclidean distance equals the MD. However, if two or more variables are correlated, the axes are no longer at right angles, and the measurements become impossible with a ruler. In addition, if we have more than three variables, we can't plot them in regular 3D space at all. The MD solves this measurement problem, as it measures distances between points, even correlated points for multiple variables.

The Mahalanobis distance measures distance relative to the centroid — a base or central point which can be thought of as an overall mean for multivariate data. The centroid is a point in multivariate space where all means from all variables intersect. The larger the MD, the further away from the centroid the data point is.

Uses

The most common use for the Mahalanobis distance is to find multivariate outliers, which indicates unusual combinations of two or more variables. For example, it's fairly common to find a 6' tall woman weighing 185 lbs, but it's rare to find a 4' tall woman who weighs that much.

Formal Definition

The Mahalanobis distance between two objects is defined as:

$$d\ (Mahalanobis) = [(X_B - X_A)^T\ C^{-1}\ (X_B - X_A)]^{0.5}$$

Where:

X_A and X_B is a pair of objects, and C is the sample covariance matrix.

Another version of the formula, which uses distances from each observation to the central mean:

$$d_i = [(X_i - \bar{X})^T\ C^{-1}\ (X_i - \bar{X})]^{0.5}$$

Where:

X_i = an object vector

$\bar{X}$ = arithmetic mean vector

Disadvantages

Although Mahalanobis distance is included with many popular statistics packages, some authors question the reliability of results.

A major issue with the MD is that the inverse of the correlation matrix is needed for the calculations. This can't be calculated if the variables are highly correlated.