Mahalanobis Distance

G J McLachlan

Geoffrey John McLachlan obtained his BSc (Hons.), PhD and DSc degrees from the University of Queensland. He has been on the faculty of the Department of Mathematics, University of Queensland throughout his career and he is currently Professor of Statistics there. McLachlan's research interests have been in the field of multivariate analysis and he has contributed extensively in the areas of discriminant analysis, mixture models, cluster analysis and more recently to image analysis. His books (with K E Basford) on *Mixture Models; Inference and Applications to Clustering* (1988) and *Discriminant Analysis and Statistical Pattern Recognition* (1992) are standard references for these topics and he has more recently (1997) published (with T Krishnan) *EM Algorithm and Extensions*.

Introduction

Are the Anglo-Indians more similar to the upper castes of Bengal than to the lower castes? Did the Jabel Moya people (of Sudan) arise from dynastic and predynastic Egyptian and Nubian peoples or from people of the negroid stock? In the genus *Micraster* (heart-urchins of the chalk in England), did the species *M.corbovis* and *M.coranguinum* arise from the species *M. cortestudinarum*? How different are the metabolic characteristics of normal persons, chemical diabetics and overt diabetics as determined by a total glucose tolerance test and how to make a diagnosis? On the basis of remote sensing data from a satellite, how do you classify various tracts of land by vegetation type, rock type, etc.?

Answers to such questions are of importance in inference about interrelations between racial or ethnic groups or species and hypothesising about their origins and evolution, in developing methods for medical diagnosis, and in developing geographical information systems.

In order to answer questions of this sort, a measure of divergence or distance between groups in terms of multiple characteristics is used. The most often used such measure is the Mahalanobis distance; the square of it is called Mahalanobis Δ^2. Mahalanobis proposed this measure in 1930 (Mahalanobis, 1930) in the context of his studies on racial likeness. Since then it has played a fundamental and important role in statistics and data analysis with multiple measurements, has become an important piece in a statistician's repertoire and has found applications in many fields where classification, numerical taxonomy and statistical pattern recognition problems are encountered – from archaeology to medical diagnosis to remote sensing.

Craniometric and anthropological studies are the first field in which the generalised distance measure of Mahalanobis was applied and have since attracted the attention of many workers interested in the theory of multivariate methods and its manifold applications in various classification and statistical pattern recognition tasks.

Craniometric and anthropological studies are the first field in which the generalised distance measure of Mahalanobis was applied.

Definition of Mahalanobis Distance and Explanation

Suppose we have two distinct groups (populations) which we shall label as G_1 and G_2. For example, in some community, G_1 and G_2 might represent girls and boys, respectively or, in a medical diagnosis situation, normal and diseased people, respectively. Consider a number (say, p) of relevant characteristics of individuals in these groups. These characteristics or measurements, for example, may be on some physical characteristics such as height or weight, or on some medical features, such as blood pressure or heart rate. We let $\boldsymbol{X}$ denote a (random) vector that contains the measurements made on a given individual or entity under study.

Often in practice, we are interested in measuring and then summarizing the differences between groups, here G_1 and G_2. A common assumption is to take the p-dimensional random vector $\boldsymbol{X}$ as having the same variation about its mean within either group. Then the difference between the groups can be considered in terms of the difference between the mean vectors of $\boldsymbol{X}$, in each group relative to the common within-group variation.

A measure of this type is the Mahalanobis squared distance defined by

$$\Delta^2 = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2), \qquad (1)$$

where the superfix T denotes matrix transpose, Σ denotes the common (nonsingular) covariance matrix of $\boldsymbol{X}$ in each group G_1 and G_2. It can be seen that since Σ is a (nonsingular) covariance matrix, it is positive-definite and hence Δ is a metric.

If the variables in $\boldsymbol{X}$ were uncorrelated in each group and were scaled so that they had unit variances, then Σ would be the identity matrix and (1) would correspond to using the (squared) Euclidean distance between the group-mean vectors μ_1 and μ_2 as a measure of difference between the two groups. It can be seen that the presence of the inverse of the covariance matrix Σ of $\boldsymbol{X}$ in the quadratic form (1) is to allow for the different scales on which the variables are measured and for nonzero correlations between the variables. Notice that for standardising a variable X (that is, to make variance equal to one) we divide X by its standard deviation. The quadratic form (1) has the effect of transforming the variables to uncorrelated standardised variables $\boldsymbol{Y}$ and computing the (squared) Euclidean distance between the mean vectors of $\boldsymbol{Y}$ in the two groups.

In *Figure* 1, we have plotted two univariate normal densities with means $\mu_1=0$ and $\mu_2=1$ and common variance σ^2 for two values of σ^2, ($\sigma^2=1$ and 1/4), corresponding to a Mahalanobis distance of $\Delta=1$ and 2, respectively. Although the two means μ_1

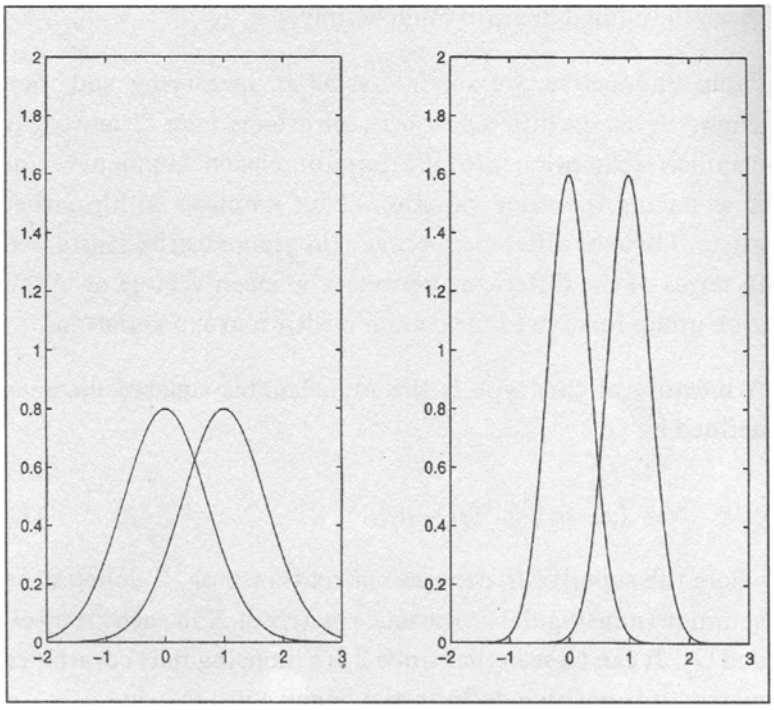

Figure 1. Representation of Mahalanobis distance for the univariate case.

and μ_2 of these two densities are the same Euclidean distance (one) apart in each case, the Mahalanobis distance Δ in the second case is twice that in the first case, reflecting less overlap between the two densities (and hence larger Mahalanobis distance between the two corresponding groups) in the second case due to their smaller common variance.

Mahalanobis Δ^2 is mainly used in classification problems, where there are several groups and the investigation concerns the affinities between groups.

Uses of Δ^2

Mahalanobis Δ^2 is mainly used in classification problems, where there are several groups and the investigation concerns the affinities between groups. The object of the study may be to form clusters of members who are similar to each other, perhaps in a hierarchical scheme.

Another situation in which Mahalanobis Δ^2 is relevant is in the problem of pattern recognition or discriminant analysis, where a formula is developed on the basis of knowledge of μ_1, μ_2 and Σ in order that a new element can be classified (assigned, identified or recognised) into one of these two groups with as little chance of error as possible under the circumstances. In a medical diagnosis problem, you may consider the knowledge mentioned above to come out of past medical records and experience, and the new element being a new patient who has to be diagnosed (that is, classified into either the 'normal' or the 'diseased' group). As is evident from the figures above, when there is overlap between the two groups, this diagnosis cannot be carried out without error and the formula attempts to diagnose with as little error as possible (optimal discriminant function).

This error depends on the overlap and can be measured in terms of Δ^2 in certain cases. If the distributions of $\boldsymbol{X}$ are multivariate normal in each group with a common covariance matrix and if the two groups appear in equal proportions, then the optimal discriminant function for assigning a new element has an overall error rate equal to $\Phi(-1/2\Delta)$, where $\Phi(\cdot)$ denotes the standard normal distribution function (see McLachlan, 1992). This error rate is equal to 0.3085 and 0.1587 for $\Delta=1$ and 2, respectively,

Very large numbers of measures of similarity between groups have been proposed, but the Mahalanobis Δ^2 has been found to be the most suitable in a majority of applications.

representing two densities that are closely ($\Delta=1$) and moderately ($\Delta=2$) separated in terms of the Mahalanobis distance between them.

Very large numbers of measures of similarity between groups have been proposed, and about thirty are known in the literature. But the Mahalanobis Δ^2 has been found to be the most suitable in a majority of applications. It is now known that many standard distance measures such as Kolmogorov's variational distance, the Hellinger distance, Rao's distance, etc., are increasing functions of Mahalanobis distance under assumptions of normality and homoscedasticity and in certain other situations.

Mahalanobis proposed an axiom for the validity of the use of Δ^2 in classification problems. It is called dimensional convergence. It can be shown that $\Delta_p^2 \leq \Delta_\infty^2$, $\Delta_p^2 \to \Delta_\infty^2$ as $p \to \infty$. The axiom states that a suitable choice of p can be made if and only if Δ_∞^2 is finite.

Sample Version of the Mahananobis Distance

In practice, the means μ_1 and μ_2 and the common covariance matrix Σ of the two groups G_1 and G_2 are generally unknown and must be estimated from random samples of sizes n_1 and n_2 from G_1 and G_2, yielding sample means $\bar{x}_1$ and $\bar{x}_2$ and (bias-corrected) sample covariance matrices $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$. The common covariance matrix Σ can then be estimated by the pooled estimate,

$$\boldsymbol{S}=\{(n_1-1)\boldsymbol{S}_1 + (n_2-1)\boldsymbol{S}_2\}/N,$$

where $N=n_1+n_2-2$. The sample version of the Δ^2 is denoted by D^2 and is given by

$$D^2=(\bar{x}_1-\bar{x}_2)^{\mathrm{T}}\,\boldsymbol{S}^{-1}(\bar{x}_1-\bar{x}_2).$$

Although D^2 is the sample Mahalanobis distance, it is usually referred to simply as the Mahalanobis distance, with Δ being referred to then as the population or true Mahalanobis distance. This usage is henceforth adopted here. The Mahalanobis distance

D^2 is known to overestimate its population counterpart Δ^2.

In situations where D^2 is used, there is a problem of its sampling variation. Thus for many rigorous applications of D^2, a knowledge of the sampling distribution of D^2 is needed. When Mahalanobis D^2 was proposed, the methodology for the derivation of sampling distributions was not as advanced as it is now and it required the mathematical prowess and ingenuity of R C Bose, S N Roy and C R Rao to resolve such statistical problems associated with the estimated D^2.

The method of derivation of the sampling distribution of D^2 is the same as that of what is known as Hotelling's T^2 since $T^2=kD^2$, where $k=(n_1n_2)/(n_1+n_2)$; see Rao (1973a). It follows under the assumption of normality that cD^2 is distributed as a noncentral F-distribution with p and $N-p+1$ degrees of freedom and noncentrality parameter $c\Delta^2$, where $c=k(N-p+1)/(pN)$. It can be used obviously to test the null hypothesis that $\Delta^2=0$, or equivalently, $\mu_1=\mu_2$. It also forms the basis of tests for the equality of two group means given that some covariates have also been observed on the individuals.

The Mahalanobis distance is thus a very useful statistic in multivariate analysis. It can be used, also for example, to test that an observed random sample $x_1, \ldots, x_n$ is from a multivariate normal distribution. We can proceed to test this null hypothesis by forming the Mahalanobis (squared) distances $D_1^2, \ldots, D_n^2$, where $D_j^2=(x_j-\bar{x})^T S^{-1}(x_j-\bar{x})$, and $\bar{x}$ and S denote respectively the sample mean and the (bias-corrected) sample covariance matrix of the n obervations in the observed sample. Then under the null hypothesis the D_j^2 should be distributed independently (approximately), with a common distribution that can be approximated by a chi-squared distribution with p degrees of freedom. Alternatively, we can form the modified Mahalanobis distances $d_1^2, \ldots, d_n^2$, where

$$d_j^2=(x_j-\bar{x}_{(j)})^T S_{(j)}^{-1}(x_j-\bar{x}_{(j)}),$$

The Mahalanobis distance is thus a very useful statistic in multivariate analysis.

where $\bar{x}_{(j)}$ and $S_{(j)}$ denote respectively the sample mean and (bias-corrected) sample covariance matrix of the $n-1$ observations after the deletion of x_j, $(j=1,\ldots,n)$. In this case, the d_j^2 can be taken to be approximately independent with the common distribution of qd_j^2 given exactly by a F-distribution with p and $n-p-1$ degrees of freedom, where $q=(n-1)(n-p-1)/\{(pn)(n-2)\}$. The values of D^2 or d_j^2 can be used also to assess whether x_j is an outlier.

Suggested Reading

[1] S Das Gupta, The evolution of the D^2-statistic of Mahalanobis, *Sankhya*, A55, 442–459, 1993.

[2] S Das Gupta, Mahalanobis distance, In P Armitage and T Colton (Eds), *Encyclopedia of Biostatistics*, Wiley, New York, 2369–2372, 1998.

[3] J K Ghosh and P P Majumdar, Mahalanobis, Prasanta Chandra, In P Armitage and T Colton (Eds), *Encyclopedia of Biostatistics*, Wiley, New York, 2372–2375, 1998.

[4] P C Mahalanobis, On tests and measures of group divergence, *Journal of the Asiatic Society of Bengal*, 26, 541–588, 1930.

[5] G J McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, New York, Wiley, 1992.

[6] C R Rao, *Linear Statistical Inference and its Applications*, Second Edition, Wiley, New York, 1973a.

[7] C R Rao, Prasanta Chandra Mahalanobis, 1893–1972, *Biographical Memoirs of Fellows of the Royal Society*, 19, 455–492, 1973b.

Further Suggested Reading

[1] P C Mahalanobis, A statistical study of the Chinese head, *Man in India*, 8, 107–122, 1928.

[2] P C Mahalanobis, On the generalised distance in statistics, *Proceedings of the National Institute of Sciences of India*, 2, 49–55, 1936.

[3] P C Mahalanobis, Normalisation of statistical variates and the use of rectangular coordinates in the theory of sampling distributions, *Sankhyā*, 3, 35–40, 1937.

[4] P C Mahalanobis, D N Majumdar and C R Rao, Anthropometric survey of the United Provinces, 1941: A statistical study, *Sankhyā*, 9, 89–324, 1949.

Address for Correspondence
G J McLachlan
Department of Mathematics
The University of Queensland
Brisbane, Qld. 4907
Australia.