JOURNAL OF CHEMOMETRICS

Published online in Wiley Online Library: 9 January 2015

(wileyonlinelibrary.com) DOI: 10.1002/cem.2692

The Mahalanobis distance and its relationship to principal component scores

by Richard G. Brereton

Journal of Chemometrics Columnist

The Mahalanobis distance is one of the most common measures in chemometrics, or indeed multivariate statistics. It can be used to determine whether a sample is an outlier, whether a process is in control or whether a sample is a member of a group or not. In the article on the chi-squared and multinormal distributions, we discussed the role of the Mahalanobis distance in the chi-squared distribution, although it is used in many other situations [1]. Although many people use this distance in chemometrics, it is poorly understood, especially its relationship to principal components analysis (PCA), another very common technique.

"The squared Mahalanobis distance is equal to the sum of squares of the scores of all non-zero standardised principal components."

1930s

The Mahalanobis distance was first proposed by the Indian statistician P. C. Mahanobis in 1936 [2]. The 1930s were important years for the development of multivariate concepts, primarily in biology, economics, and psychology, with many famous names such as R. A. Fisher and H. Hotelling active during these years and communicating together.

For traditional univariate statistics, it is usual to calculate the number of standard deviations an observation is from the centre of a dataset and use this value to determine various statistics about it. Extending to multivariate situation, Mahalanobis proposed a distance Δ from the centre of the data.

DEFINITION

The traditional mathematical definition is given by $\Delta^2 = (\boldsymbol{x} - \overline{\boldsymbol{x}})\ \boldsymbol{S}^{-1}(\boldsymbol{x} - \overline{\boldsymbol{x}})'$ where

- $\boldsymbol{x}$ is a row vector (in the notation of this article) consisting of the multivariate measurement for an observation; for example, if three variables are measured, it could be a row vector [1.5 0.62 7.15] where each number consists of a single measured variable for a single observation. Often, there will be many more variables; for example, we may be dealing with a spectrum or a chromatogram;
- $\overline{\boldsymbol{x}}$ is the mean of the sample we have measured; and
- $\boldsymbol{S}$ is the variance–covariance matrix of the sample; its diagonal elements consist of the variance of each variable and the off diagonals of the covariance.

A ′ represents a transpose; that is, the rows and columns are interchanged. In later articles, we will brush up on matrix algebra for those that are unfamiliar.

STANDARD DEVIATION

When there is only one variable, this simplifies to $\Delta^2 = (x - \overline{x})v^{-1}(x - \overline{x})'$, where v is the variance or $\Delta^2 = (x - \overline{x})^2/\mathrm{v}$, giving $\Delta = (x - \overline{x})\ /s$ where s is the standard deviation of the data. Hence, the Mahalanobis distance for one variable is the same as the number of standard deviations; an observation is away from the mean.

MANY VARIABLES

For more than one variable, the Mahalanobis distance can be visualised as the distance of a point from the centre of a dataset, onto an ellipse (if there are two variables) whose main direction is that of the data, as illustrated in Figure 1. The ellipse drawn is that of equal Mahalanobis distance so any point lying on that ellipse will be equally far from the centroid. The axes of the ellipse are scaled so that they are both of the same length for points at equal Mahalanobis distance from the centre.

There is however an alternative way of visualising the Mahalanobis distance, and that is by transforming the original data, as illustrated in Figure 2 for a dataset characterised by two variables.

- Centre the dataset, that is, subtract its mean. In the calculation, this is carried out by subtracting $\overline{\mathbf{x}}$.
- Then rotate the dataset. This is to make the variables independent, a condition for a chi-squared distribution.
- Finally, scale the axes so that the variance (and standard deviation) for each variable equals 1 so the sum of squares equals the number of rows, or observations, or sample size. For experts, the matrix $\boldsymbol{S}$ could be viewed as the product of two matrices, one for rotation and one for scaling, and so can its inverse. Once the variables are independent, the scaling matrix is a diagonal matrix. Rotations, which are also called transformations, will be discussed in later articles.

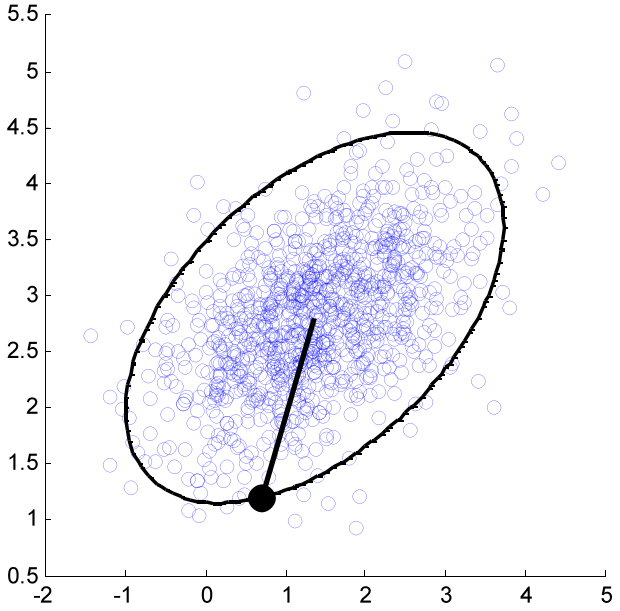

Figure 1. Mahalanobis distance of a point from its centroid.

- Then the two new variables, which we will call t, are independent and normalised. If the distribution of squares of each variable can be characterised by a chi-squared distribution, so can their sum as discussed in the chi-squared and multinormal distributions [1].
- The Mahalanobis distance from the centre is now given by $\Delta^2 = t_1^2 + t_2^2$, and if the original dataset is multinormal, the Mahalanobis distances form a chi-squared distribution.

PRINCIPAL COMPONENTS ANALYSIS

Many readers of these articles will have heard of principal components analysis (PCA). For those that are unfamiliar with the technique, it will be discussed at length in later articles. However, from the discussion in the preceding text, it should be evident that the squared Mahalanobis distance is equal to the sum of squares of the scores of all non-zero standardised principal components. This is a bit of a mouthful but putting it algebraically

$$\Delta^2 = t_1^2 + t_2^2 + \ldots + t_k^2$$

if there are k non-zero principal components.

- A traditional limitation of the Mahalanobis distance is that it cannot be calculated if the number of variables exceeds the sample size; this is because of a problem with the inverse of $\boldsymbol{S}$.
- Many in chemometrics go to great lengths to overcome this limitation. In many modern datasets, the number of variables far exceeds the sample size; for example, they may be spectroscopic wavelengths or Gas Chromatography Mass Spectrometry (GCMS) peaks. Often approaches like variable selection are used.
- However using the alternative computational method for calculating the Mahalanobis distance, this is not a limitation. The number of non-zero PCs is never more than the smaller of the number of observations or variables.

There are of course many other diverse reasons for using the Mahalanobis distance measure besides the chi-squared criterion, and we will be using this measure and comparing to other distances in different contexts in future articles.

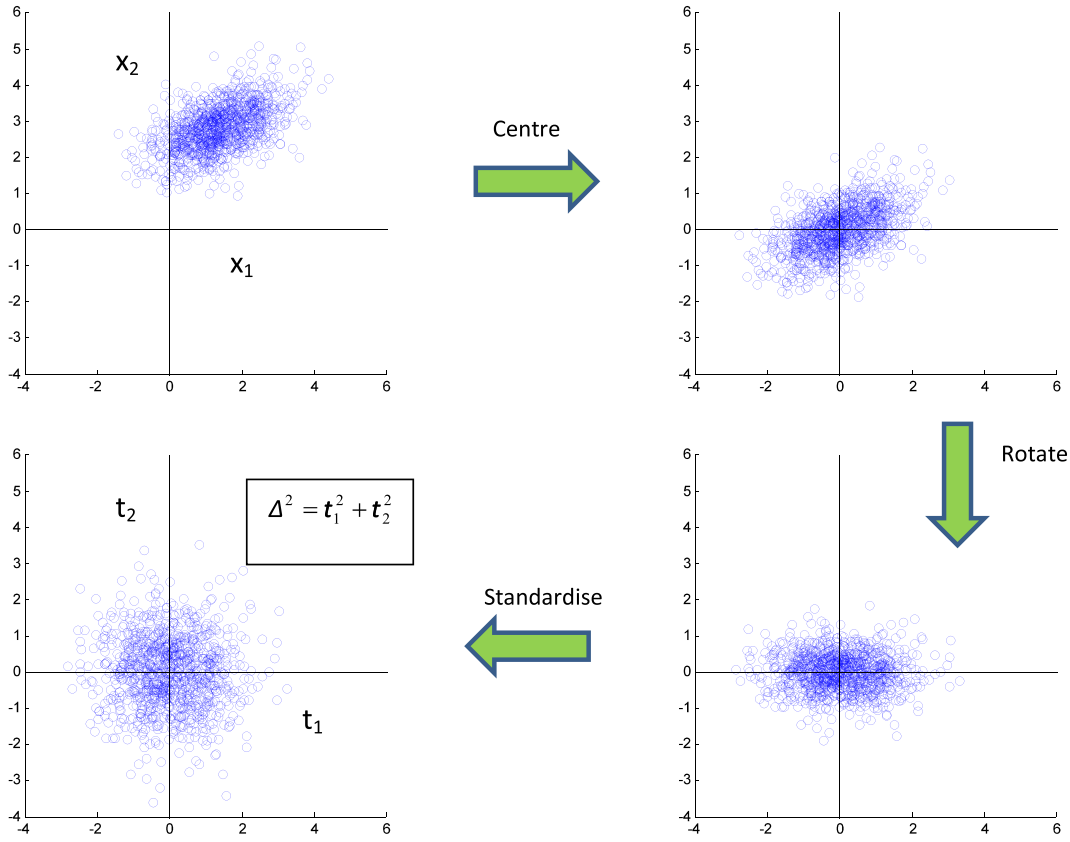

Figure 2. Steps that can be used for determining the Mahalanobis distance.

PROPERTIES

The Mahalanobis distance has a number of interesting properties. For a given dataset (or training set), the sum of squares of the Mahalanobis distance of all observations, or rows in a data matrix, usually equals the product of the number of variables times the number of observations. So if the sample size is 50, and there are three variables, the sum of the 50 squared Mahalanobis distances will usually be 150. This is because the sum of squares each normalised principal component equals the sample size. This can be verified numerically.

CALCULATION

In Excel, the Mahalanobis distance is a bit awkward to calculate.

- Consider a set of 50 observations, characterised by two variables, in cells A1:B50.
- Calculate the mean of the dataset, a row vector. Place AVERAGE(A1:A50) in cell A52 and a similar calculation for column B in cell B52.
- The variance–covariance matrix can be calculated by using the VAR.P and COVARIANCE.P functions. It is usual but not essential to use the population statistics, as discussed in the next article.
- This is a matrix of consisting of two rows and two columns in this case. Place it in cells A54 to B55.
- You can calculate the inverse of this matrix in cells A57 to B58. Use the MINVERSE matrix function in Excel.
- To simplify, you can use the Name Manager (which is found in the 'Formulas' tab of Excel 2010 and higher), to call the mean 'm' and the inverse of the variance–covariance matrix 's'.
- Then in cell D1, calculate the squared Mahalanobis distance of sample 1 using the matrix syntax MMULT(MMULT((A1:B1-m),s),TRANSPOSE(A1:B1-m)) and then copy this down to cell D50.
- The 50 squared distances will be in cells D1 to D50. You can check that their sum is equal to 100 in this case. The nature of the data is irrelevant providing that the two columns are not correlated.

PRINCIPAL COMPONENT SCORES

It is then also possible to perform PCA. It is assumed that the reader has access to a way of generating numerical output that can be output into Excel. If you are unable to do this at this stage, skip this aspect.

- Centre the data. Some PCA algorithms do this automatically; you can tell by checking the average scores of the PCs, these should be 0 (or a very small number).
- Calculate the scores of the first two components, placing them in cells F1 to G50.
- Make sure that each component is normalised. Some methods such as singular value decomposition do this automatically, others such as NIPALS do not. If you are uncertain of your algorithm, just check the sums of squares, for the column F1 to F50, corresponding to the first PC, just calculate SUMSQ(F1:F50).
- If the column sums of squares equals 50 (so that its variance is equal to 1), then you are finished, if not, divide the scores by the standard deviation of the corresponding column, so for example a new number can be entered in cell J1 equalling F1/STDEVP(F$1:F$50); note that you should use the Excel definition of the population standard deviation. (There are several alternatives to this method, which the reader is left to figure out).
- Now just calculate the sum of the squared normalised scores for each sample, so for observation 1, if the standardised scores are in cells F1 and G1, this is SUMSQ(F1:G1).
- If carried out correctly, this should equal the squared Mahalanobis distance.

Of course, there are other and computationally somewhat simpler ways of checking this in packages such as Matlab, but for the purpose of this article, we primarily illustrate the steps in Excel. In fact, Excel is a good way of learning and consequently understanding methods because one can see all the numerical steps and check sums as one goes on. There are many outstanding Matlab programmers in chemometrics, but not every one of them would realise the connection between the Mahalanobis distance and principal components analysis, despite using both on a regular basis, possibly because it is possible to skip numerical stages when using Matlab.

There are several on-line resources that can help mainly with calculations and further insight. Dwinnel [3] presents a Matlab-oriented article. Wicklin's article [4] although written by SAS institute, requires no knowledge of SAS to understand and is a helpful discussion.

Richard G. Brereton

School of Chemistry, University of Bristol, Cantock's Close, Bristol, BS8 1TS, UK

REFERENCES

1. Brereton R.G. (2014) The chi squared and multinormal distributions. *J. Chemometrics*, DOI: 10.1002/cem.2680
2. Mahalanobis PC. On the generalised distance in statistics. *Proceedings of the National Institute of Science of India*, 1936; **2**: 49–55.
3. Dwinnell W. Data Mining in Matlab, http://matlabdatamining.blogspot.co.uk/2006/11/mahalanobis-distance.html
4. Wicklin R. What is Mahalanobis Distance? The DO loop, SAS online resources http://blogs.sas.com/content/iml/2012/02/15/what-is-mahalanobis-distance/