ANALYSING MULTIDIMENSIONAL MEASUREMENT COMPARISON
DATA CONTAINING OCCASIONAL ERRATIC POINTS

J C Medley and N M Ridler
National Physical Laboratory
56 DRA, St Andrews Road, Malvern, Worcs WR14 3PS, UK

Abstract
This paper proposes a method for analysing multidimensional measurement comparison data containing occasional erratic data points. The statistics described exhibit robustness to the effects of such points, while being relatively simple to understand. They are introduced pictorially to emphasise their multidimensional "spatial" nature before a more formal mathematical description is given.

Introduction
Measurement comparisons are used in many scientific disciplines to evaluate the performance of measurement systems and techniques. For example, regular comparisons of RF and microwave parameter measurements are undertaken by the members of ANAMET, a network analyser metrology club run by NPL, to improve the members' confidence in their ability to make good quality measurements.

One of the main tasks for the organiser of a comparison is to analyse the measurement data and produce a summary so that individual participants can effectively assess their own performance in the exercise. Such analyses have traditionally used Gaussian statistics, however measurement comparisons sometimes have erratic data points in which case median based statistics may be of more use [1].

The data analysis problem is more complicated if the measurements are of multidimensional quantities. This paper describes statistics which extend the concept of a median in order to establish a summary value (a best estimate of location) and a measure of spread for such measurement data.

Estimating location
To illustrate the problem of multidimensional location estimation, a two-dimensional example will be discussed. Consider the two two-dimensional data sets shown in Figures 1A and 1B. The aim in each case is to find a single point in the plane which best represents the data set.

In set A the data is quite tightly grouped. Set B is identical but for one data point that has been moved to an outlying position away from the main grouping.

Figure 1: Example data sets
It is quite easy to judge by eye the position of a summary point for data set A - most such estimates will lie close to the cross in Figure 2A. For measurement comparisons, an estimate is required which is not unduly affected by outlying values. Hence in Figure 2B, the cross representing the summary value for data set B has not moved significantly from that for data set A.

The estimator used to calculate a summary value in each of these examples was the spatial median.

Figure 2: Locations estimated by the spatial median

---

1 N M Ridler is employed by Assessment Services Limited, and is working under contract at the National Physical Laboratory.
Spatial Median

The spatial median of a set of vectors $X_1, \ldots, X_n$ is the point $\mu$ that minimises the sum of the absolute deviations (distances) between $\mu$ and the $n$ data points. i.e.

$$\sum_{j=1}^n |X_j - \mu|$$

is minimised when $\mu$ is the spatial median. With respect to the two examples, the sum of the lengths of the lines joining the spatial median to the data points in each of the Figures 2A and 2B is a minimum.

Some properties of the spatial median

A comprehensive description of the properties of the spatial median can be found in [2]. The most important properties for the current application are given below.

i. The spatial median is a robust estimator of location, exhibiting similar behaviour to the conventional one-dimensional median.

ii. The spatial median is unique if the data has two or more dimensions [3]. In one dimension it is not unique if the number of data points is even. The conventional choice in this case is to take the point halfway between the two most central points as the median.

iii. The spatial median is defined without reference to any particular coordinate system, only requiring a consistent definition of distance. This is an advantage over techniques which depend on the coordinate system used. For example with complex data, the points represented by a) median real, median imaginary; and b) median magnitude, median phase; will in general be different.

Estimating spread

In keeping with the multidimensional median estimate of location, the median absolute deviation is utilised as a measure of spread.

Median absolute deviation

The median absolute deviation (MAD) of a set of vectors $X_1, \ldots, X_n$ from the point $\mu$ is given by

$$MAD(X_1, \ldots, X_n; \mu) = median\{ |X_i - \mu| ; i = 1, \ldots, n \}.$$  

The MAD can be taken as the radius of the circle or (hyper)sphere centred on $\mu$ that contains half of the data points (see Figures 3A and 3B). It can be used to judge whether the difference between a participant's value and the summary value is significant. As with the spatial median, the MAD is not unduly affected by the outlying value in data set B.

![Figure 3: Half of the data points lie within a radius of MAD of the spatial median.](image)

Implementation

The spatial median and MAD have been implemented as spreadsheet macros, the former using the algorithm described in [4]. They have been used to analyse scattering parameter data in a comparison organised for ANAMET [5] (although the MAD was used in a slightly different way) and will be used in further comparisons.

Summary

Statistics have been presented for the calculation of a summary value and measure of spread for multidimensional data from a measurement comparison. The statistics are resistant to the effects of occasional erratic data points and yet relatively simple to understand.

References


