EXTENSIONS OF FRACTILE GRAPHICAL ANALYSIS TO HIGHER DIMENSIONAL DATA

P.C. MAHALANOBIS
INDIAN STATISTICAL INSTITUTE

Tech. Report No. 7/66 15 February 1966

Research And Training School
Indian Statistical Institute
203 Barrackpore Trunk Road, Calcutta 35, India

EXTENSIONS OF FRACTILE GRAPHICAL ANALYSIS TO HIGHER DIMENSIONAL DATA

by

P. C. Mahalanobis
Indian Statistical Institute, Calcutta

INTRODUCTION

The technique of fractile graphical analysis (FGA), in the form of bivariate graphs, was developed and applied in the analysis of economic data in a series of papers (Mahalanobis, 1958a, 1958b, 1960, 1962). The wide applicability of this method was demonstrated in recent papers by Linder (1963), Rhea Das (1960a,b, 1964) and others. Some conjectures made by the author were studied in a number of theoretical papers by Kawada (1961), Kitagawa (1960), Mitrafanava (1961), Sethuraman (1961, 1963) and Takeuchi (1961). The computational aspects through the use of electronic computers were examined by Roy and Kalyanasundaram (1963). The object of the present paper is to provide further theoretical foundations of FGA and extend it to higher dimensional data.

Genesis of the problem. Let us suppose that we wish to study the differences in the distribution of per capita consumption of cereals between two different regions or over time in a single region. It is a common practice in such cases to compare the mean values, standard deviations and other measures characterising the distributions. Such overall comparisons may not be meaningful or completely informative specially when differences in the consumption of cereals are not the same for all 'comparable sub-sets' of the populations under comparison (in two regions or at different points of time). Thus it may be pertinent to ask whether there is any differential increase in the consumption of cereals between the 'poor' and 'rich' sections of the populations. Such a problem lead us to define comparable subsets populations such as poor, rich and so on. For this purpose we use a suitable concomitant variable such as per capita income of an individual as an indication of economic states. We note that absolute values of income at different times or in different regions are not however, comparable and therefore comparable subsets cannot be defined as groups of individuals having the same per capita income. But groups of individuals having the same relative economic status. in the two populations, as defined by ranks with respect to income within a population, might be comparable. In other situations there may be other ways of defining comparable subsets of the populations. After choosing a number of comparable subsets on the basis of a concomitant variable, we examine the difference in the distributions of a main variable for every pair of comparable subsets. Fractile graphical analysis is a convenient technique by which the desired comparisons can be made through appropriate graphs drawn on the basis of sample data. The actual computations involved when there are one main variable and one concomitant variable are briefly explained in section 2 and the generalizations to cases with more than one concomitant variable are discussed in other sections.

2. BIVARIATE FRACTILE GRAPH

A fractile graph in the case of two dimensional data in defined as follows. Let (Y, X) denote the two variables and $(Y_1, X_1), \ldots, (Y_N, X_N)$, be N independent observations. One particular variate, say X, is selected for ranking the observations in ascending order. Replacing the X values by ranks and arranging the observations in ascending order of X, we obtain

$$(Y_{(1)}, 1), (Y_{(2)}, 2), \ldots, (Y_{(N)}, N) \tag{2.1}$$

where $Y_{(i)}$ is the Y value associated with the X value of rank i.
Now divide the observations (2.1) into a chosen number, g, of groups such that each group consists of $h = N/g$ consecutive observations. These are called fractile groups. The ith fractile group represented by $[i]$ consists of the observations

$$(Y_{(ih)}, ih), (Y_{(ih+1)}, ih+1), \ldots, (Y_{(ih+h-1)}, ih+h-1) \tag{2.2}$$

which are replaced by the pair

$$(Y_{[i]}, i) \tag{2.3}$$

where i represents the ith fractile group and $Y_{[i]}$ is any statistic (such as the mean, median, maximum etc.) based on the Y values of the observations in $[i]$,

$$Y_{(ih)}, \ldots, Y_{(ih+h-1)}. \tag{2.4}$$

The g pairs

$$(Y_{[1]}, 1), \ldots, (Y_{[g]}, g) \tag{2.5}$$

provide the fractile graph, by plotting $Y_{[i]}$ against i, $i = 1, \ldots, g$ and joining the successive points by straight lines.

The graph so obtained is represented by G_g. The fractile graph for the entire population obtained by applying the procedure indicated to all members of the population is denoted by Γ_g. As the sample size increases G_g will provide a consistent estimator of Γ_g.

Separation between graphs. To compare two fractile graphs arising out of independent samples from two different populations it is necessary to know how different graphs of parallel samples from the same population can be. For this purpose we divide each sample into two independent halves. The graphs of the two half samples from the first population are denoted by G_{g1} and G_{g2} and that of the entire sample by G_g. Similarly we have the corresponding graphs G'_{g1}, G'_{g2} and G'_g for a sample from the second population. We then choose a measure of separation $||A - B||$ between any two graphs A and B. To test the significance of the observed difference between G_g and G'_g we may use a statistic of the type

$$M = 4 \sqrt{\frac{n\,n'}{n+n'}}\ ||G_g - G'_g|| \div \left(\sqrt{n}\ ||G_{g1} - G_{g2}|| + \sqrt{n'}\ ||G'_{g1} - G'_{g2}|| \right) \qquad \ldots \quad (2.6)$$

An overall measure of separation proposed in the earlier papers is the area between graphs. The exact distribution of M is unknown, but as a first approximation M^2 may be considered as a variance ratio on (g-1) and 2(g-1) d.f.

The purpose of drawing the fractile graphs is not, however, to end up in an overall comparison as provided by a statistic of the type (2.6). The object is to compare the graphs at each fractile point or at sets of consecutive fractile points and draw inferences. If necessary, a test of the type (2.6) may be used for chosen sections of the graphs to examine the significance of the observed differences.

In practice a significance test is hardly necessary when there is a clear separation of the graphs indicated by the fractile points for the two halves of one sample being completely above or below those for the other sample. The chance of clear separation at any fractile point when there is no real difference is 1/6, which is reduced to 1/20 if instead of two we have three parallel samples from each population. There is, thus, a slight advantage in splitting each sample into three sets and drawing three parallel graphs for each sample.

It may be readily seen that F.G.A. is different from the concepts of Lorenz curve or specific concentration curve. A wrong statement to the effeet that a fractile graph is derivable from a Lorenz curve or a specific concentration appears in a paper by Swamy (1963) in Econometrika. On the other hand, as shown by Iyengar and Bhattacharya (1965) a fractile graph is more general and more useful than concentration curves. Further fractile graphs can be drawn in situations where concentration curves are not meaningful as demonstrated in a test for normality by Linder (1963).

3. EXTENSION TO MULTIVARIATE CASE

Let us consider observations N from a (k + 1) variate distribution and denote the variates by $(Y, X_1, \dots, X_k)$, where $X_1, \dots, X_k$ are in the nature of k concomitant variables. There are several natural ways of defining comparable subsets. We shall consider some of them.

Hierarchical group ranking of observations. An observation is said to belong to the group ranking $[i_1, \ldots, i_k]$ if it is in the i_k-th fractile when ranked on X_k considering only the observations belonging to the group ranking $[i_1, \ldots, i_{k-1}]$, which is similarly defined with respect to $[i_1, \ldots, i_{k-2}]$, and so on. The group ranking $[i_1]$ simply defines the i_1-th fractile when ranked on X_1 only. If the number of fractiles considered for X_j is g_j, then the number of observations in the group $[i_1, \ldots, i_m]$ is

$$h_m = \frac{N}{g_1 \ldots g_m}, \quad m = 1, \ldots, k.$$

Since i_n can take the g_m values, $1, \ldots, g_m$, there are $g_1, \ldots, g_k$ possible group rankings and the vectors defining them are lattice points in a k dimensional Euclidean space. Every observation can be uniquely identified by a group ranking. Let $Y_{[i_1 \ldots i_k]}$ be a statistic, such as the mean, median etc., of the Y values of the h_k observations belonging to the group ranking $[i_1, \ldots, i_k]$. We thus have a (k+1) dimensional vector

$$(Y_{[i_1 \ldots i_k]}, i_1, \ldots, i_k) \qquad (3.1)$$

as a generalization of (2.3). There are $N/h_k = g_1 \cdots g_k = g$ vectors of fractile points (3.1) which can be plotted in a $(k + 1)$ dimensional space.

As in the bivariate case we need to define a measure of separation between two sets of fractile points corresponding to two different samples (either parallel or from different sources). Let us represent the fractile points of another sample by

$$(Y'_{[i_1 \cdots i_k]}, i_1, \ldots, i_k).$$

We suggest two ways of defining a graphical measure. It may be seen that by fixing the coordinates $i_1, \ldots, i_{k-1}$, we have the conditional bivariate graphs generated by the pairs $(Y_{[i_1 \cdots i_k]}, i_k = 1, \ldots, g_k$. The area between such bivariate graphs for fixed $i_1, \ldots, i_{k-1}$ can be measured, which is represented by

$$a_{i_1 \cdots i_{k-1}}$$

The separation between the two sets of fractile points in the $(k + 1)$ dimensional space may be defined by

$$||G_g - G'_g|| = \Sigma\, a_{i_1 \cdots i_{k-1}} \tag{3.2}$$

where the summation in (3.2) is over all the combinations of $(i_1,\ldots,i_{k-1})$.

As in the bivariate case let $||G_{g1} - G_{g2}||$ be the separation between two parallel halves of one sample of size n and $||G'_{g1} - G'_{g2}||$, that between two parallel halves of another sample of size n' and $||G_g - G'_g||$, that between two samples. Then to test the hypothesis that the samples belong to two different populations we may use the statistic

$$M = \frac{4\sqrt{\frac{n\,n'}{n+n'}}\,||G_g - G'_g||}{\sqrt{n}\,||G_{g_1} - G_{g_2}|| + \sqrt{n'}\,||G'_{g_1} - G'_{g_2}||} \tag{3.3}$$

Then M^2 may be used approximately as a variance ratio on $k(g_k - 1)$ and $2k\,(g_k - 1)$ d.f.

Another approach is to consider a <u>simplex</u> consisting of 2^k points

$$(i_1 + j_1,\ i_2 + j_2\ ,\ \ldots,\ i_k + j_k)$$

with $(i_1,\ldots,\ i_k)$ as one corner, where each j_m can take two possible values 0 or 1 and graduate the corresponding Y points by a ruled surface, i.e., fit an equation of the type

$$Y = a + \Sigma\, b_i\, x_1 + \Sigma\,\Sigma\, c_{ij}\, x_i\, x_j + \ldots + m\, x_1 \ldots x_k$$

The local surfaces fitted to all the simplexes define a surface over the $g_1 \ldots g_k$ lattice points, which may be called a fractile surface. For two samples, there are two fractile surfaces and the volume between them, which is the natural extension of the area in the bivariate case, may be defined as the distance $||G_g - G'_g||$ between two fractile graphs, G_g and G'_g In this case M^2, where M is as defined in (3.3) with the new definition of separation as the volume between the fitted ruled surfaces $(g_1-1) \ldots (g_k-1)$ and $2(g_1-1) \ldots (g_k-1)$ d.f.

In the method of hierarchical ranking and construction of fractile surface, the results depend on the order in which the concomitant variates are considered. But there may be a natural way of deciding the order of the variates in practice for obtaining comparable subsets.

Ranking by individual variables. We define an observation $(x_1,\ldots,x_k)$ to belong to the group ranking

$$(i_1, \ldots, i_k)$$

if x_1 belongs to the i_1-th fractile group when ranked on X_1 alone; x_2 belongs to the i_2-th fractile group when ranked on X_2 alone and so on. In such a case the number of observations will not be the same for each group ranking as in the earlier methods. But the procedure is independent of the order in which the concomitant variables are considered unlike in the earlier method. Since each fractile point is not determined on the same number of observations measures of separation such as volume between fractile surfaces will not have nice properties.

Reduction to a single concomitant. Another approach is to reduce the k dimensional concomitant variable to one dimension by considering a suitable function, in which case the theory of bivariate graphs applies. The reduction to one variable can be achieved by constructing a discriminant function; based on the concomitant variables only, between samples to be compared. The discriminant function provides the best separation with respect to the concomitant variables and ranking based on the discriminant function removes the differences due to the concomitant variables.

Practical applications of higher dimensional fractile graphical analysis are being considered and it is hoped to discuss them elsewhere.

In conclusion I wish to thank Dr. C. R. Rao for the useful discussions I had with him during the preparation of this paper.

REFERENCES

1. Das, Rhea S.(1960a) : Applications of fractile graphical analysis to psychometry I.- Item analysis. *Psychological Studies*, 5, 11-18.

2. __________ and Sharma,K.N.(1960b) : Applications of fractile graphical analysis to psychometry II.- Reliability. *Psychological Studies*, 5, 71-77.

3. __________ (1964) : Item analysis by probit and fractile graphical methods. *British Journal of Statistical Psychology*, 17,, 51-64.

4. Iyengar, N.S. and Bhattacharya, N. (1965): Some observations on fractile graphical analysis. *Econometrika*, 33, 644-645.

5. Kawada Yukiyosi (1961) : Some remarks concerning the expection of the error area in fractile analysis. *Sankhyā*, Series A, 23, 155-160.

6. Kitagawa, T. (1960) : Sampling distributions of statistics associated with a fractile graphic method. *Bull. Math. Statist.* Fukuoka, Japan, 9, 10-42.

7. Linder, A.(1963) : Convocation address. Indian Statistical Institute.

8. Mahalanobis, P.C. (1958) : A method of fractile graphical analysis with some surmises of results. *Transactions of the Bose Research Institute*, Calcutta, 22, 223-230.

9. ______________ Lectures in Japan : Fractile graphical analysis. Indian Statistical Institute.

10. __________(1960) : A method of fractile graphical analysis, *Econometrica*, 28, 325-351.

11. Mahalanobis, P.C. (1962) : A preliminary note on the consumption of cereals in India. Bull. Int. Stat. Inst., 39, 53-76.

12. Mahalanobis, P.C. and Lahiri, D.B. (1961) : Analysis of errors in censuses and surveys with special reference to experience in India, Sankhyā, Series A, 23; 325-358.

13. Mitrofanova; N.M. (1961) : On some problems of fractile graphical analysis. Sankhyā, Series A, 23, 145-154.

14. Parthasarathy, K.R. and Bhattacharya, P.K. (1961) : Some limit theorems in regression theory. Sankhya, Series A, 23, 91-102.

15. Roy, J. and Kalyanasundaram, G. (1963) : Punched card processing of sample survey data for fractile graphical analysis. Contributions to Statistics, (Volume presented to Professor P.C.Mahalanobis on the occasion of his 70th birthday).

16. Sethuraman, J. (1961) : Some limit distributions connected with fractile graphical analysis. Sankhyā, Series A, 23, 79-90.

17. __________ (1963) : Fixed interval analysis and fractile analysis. Contributions to Statistics, (Volume presented to Professor P.C.Mahalanobis on the occasion of his 70th birthday), Volume, 449-470.

18. Swamy, S. (1963) : Notes on fractile graphical analysis. Econometrika, 31, 551-554.

19. Takeuchi, K. (1961) : On some properties of error area in the fractile graph method. Sankhyā, Series A, 23, 65-78.