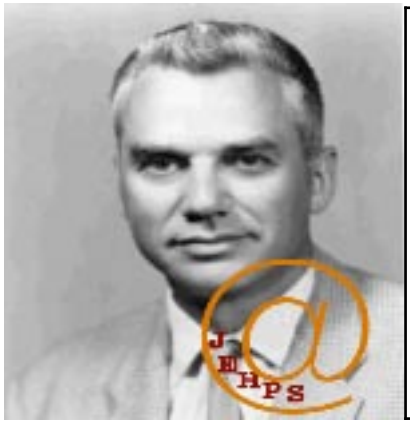

Journ@l Electronique d'Histoire des Probabilités et de la Statistique

Electronic Journ@l for History of Probability and Statistics

Vol 4, n°2; Décembre/December 2008

www.jehps.net

The biological stimulus to multidimensional data analysis

John GOWER[1]

Introduction

In the following, I shall try to summarise developments over the past 100 years that have led to the current interest in data analysis and more particularly in multidimensional data analysis (see the Conclusion section for a discussion of these terms). Necessarily, my views are coloured by my own experience but I hope that some points of general interest will emerge.
Developments may be divided into three periods: Prehistory (up to about 1950), The Early Computer Age (up to about 1980), and the current Later Computer Age. I shall concentrate on developments made during the first two periods and how they have influenced current practice. I conclude with some more general comments.

Prehistory

The first half of the twentieth century was an era in which many statistical methods and ideas were developed. We begin at the start of the 20th century, by when the Biometric School was well-established at University College, London, under Karl Pearson in the Galton Chair of Eugenics; a little later in 1911, he became the first U.K. professor of statistics. Pearson had a special interest in anthropometry and developed many ideas that remain in use or underpin more recent work. Of special interest to us are the Analysis of Mixtures (Pearson, 1894), Principal Components Analysis (1901) and the Coefficient of Racial Likeness (1926), later leading to Mahalanobis Distance (1936). A basic statistical tool was the multinormal distribution with its linear regressions and correlation parameters. Some early similarity coefficients belong to this period.

[1] Department of Mathematics, The Open University, Milton Keynes, MK7 6AA, U.K.

R. A. Fisher developed methods for analyzing field experiments based on the Gauss linear model, with dummy variables defining additive main effects and interactions. He also explored (Fisher and Mackenzie, 1923) the possibility of expressing two-way interaction multiplicatively; this, the first example of a biadditive model, required eigenvalue calculations. Much later, Fisher's multiplicative/biadditive model was taken up by plant breeders interested in genotype-environment interaction and the first steps taken into developing multiplicative models for higher order interactions.

Eigenvalue methods for determining optimal scores for the levels of categorical variables were developed independently by Fisher (1940) and Guttman (1941); these may be seen as forerunners of Multiple Correspondence Analysis (see Gower, 1990). Hirschfeld (1935), later better known as H. O. Hartley, asked what scores should be given to the levels of two categorical variables to maximize their correlation. This was an early manifestation of the Correspondence Analysis of a two-way contingency table and a precursor to Hotelling's (1936) Canonical Correlation Analysis. Canonical variate analysis, also a form of canonical analysis based on the two-sided eigenvalue problem, with links to discriminant analysis, Mahalanobis distance and, at a formal level, with canonical correlation, was suggested by C. R. Rao. Another Hotelling (1933) innovation was his reformulation of principal components analysis as a method for the factor analysis of a correlation matrix; in my opinion, this influential paper has been at the heart of much subsequent misunderstanding and confusion. Effectively, Karl Pearson had been concerned with approximating a data matrix $\mathbf{X}$ whereas Hotelling was concerned with approximating the derived matrix $\mathbf{X'X}$, give or take a normalisation/standardisation or two. The two problems are underpinned by shared algebra, leading to similar computations but the statistical interpretations are very different. The singular value decomposition plays an important part in unravelling the interrelations between all these methods and it does not help that the SVD of $\mathbf{X}$ is commonly computed by first calculating the eigenstructure of $\mathbf{X'X}$ (see Gower, 2006 for a detailed discussion). Although the SVD was known in an algebraic context from the 1880s, it was Eckart and Young (1936), publishing in the first issue of Psychometrika, who established its property, already implicit in Pearson's work, of giving a low-rank least-squares approximation to $\mathbf{X}$.

Computational resources were very limited, so extensive use of eigenvalue and matrix inverse calculations was prohibitive. Consequently, much of the work developed in the first half of the twentieth century had little immediate impact on applications. Following the computer revolution, many of the methods developed in this period came into their own and are routinely used today, becoming the springboard for further advances.

Perhaps the greatest intellectual achievement of this period was, starting with the Wishart (1928) distribution of the sample covariance matrix of a multinormal distribution, the derivation of increasingly difficult functional forms for the distributions of other multivariate statistics, mostly based on eigenvalues. Impressive as these results are, they seem to have had little influence on subsequent developments, many arguing that they

provide tools for testing irrelevant hypotheses. In general, formal inferential methods have not found a place in most forms of data analysis.

The Early Computer Age

By the mid 1950s primitive electronic computers permitted the practical exploitation of the methods that had been suggested earlier. In 1955, I myself was fortunate to be appointed to a post at Rothamsted Experimental Station where an early electronic computer, the Elliott-NRDC 401 had recently been installed (see Gower, 1986). This small prototype machine had fewer than 2000 32-bit words of memory and a bizarre machine code; nevertheless it was capable of serious work. Part of the justification for getting this machine was to see if electronic computers might be useful in agricultural-research statistics. As part of this remit, Frank Yates, then Head of the Rothamsted Statistics Department, initiated what he called The Agricultural Statistics Research Service, which encouraged collaboration with scientists throughout Britain who might have a computational need. Applications should have agricultural or at least biological interest, but this restriction was interpreted very broadly. On my arrival I found that work was already in progress on a canonical variate analysis to use tooth measurements to discriminate between populations of fossil and modern apes. At that time, it was necessary to develop all one's own subroutines, including those for reading and printing numbers in decimal form, basic functions such as division, square root and logarithms, let alone major algorithms for evaluating matrix inverses and eigenstructure.

Taxonomy

Peter Sneath's (1958) use of the single-linkage algorithm for microbial hierarchical classification encouraged many young taxonomists, anxious to try out the new method. Several came to me through the Agricultural Statistics Research Service. The first was from the Low Temperature Research Institute (food storage temperatures), followed by several from the British Museum of Natural History, Kew, University of Oxford, and Linguistics from Uppsala. My first program for cluster analysis could cope with up to 32 units but only by packing four similarity coefficients into each of 128 32-bit words; this gave an accuracy of about 0.5per cent to each coefficient, good enough for most applications. Space was insufficient to store both the program for evaluating the similarities and the values themselves, so as soon as it was calculated, each similarity was output to paper tape. This tape was reread by a program that performed the actual cluster analysis and did ancillary things, such as printing trees and calculating average cluster values.

It was clear from the outset that one had to distinguish between characters that could be "present" or "absent" from characters that could have two (or more) comparable states such as "red" and "yellow". Further, "known absence" should be distinguished from "missing" or "not known". With slight modifications, the Jaccard and Simple Matching similarity coefficients could cope with the simpler situations but with the wide ranging set of taxonomic application areas, it soon became clear that these coefficients needed to be supplemented, by developing a general coefficient of similarity that could handle the many different types of measurement used by taxonomists. My colleagues, publishing in

taxonomic journals, wished to cite a paper describing the general similarity coefficient but I had written nothing, thinking that the topic was of insufficient interest, and it did not occur to me to seek joint authorship. Eventually, I derived some mildly interesting properties of the coefficient and in 1967 submitted a paper for publication (Gower, 1971); I cannot recall what prompted the four year delay but it reflects a more leisurely age, especially when one notes that the original work was done about 1960.

Plant ecologists too were ready to exploit computers. My main link was with W.T. Williams, then Professor of Botany in the University of Southampton, later to be Chief of the CSIRO Division of Tropical Agriculture. Working with N. J. Lambert, who was interested in the origins and ecology of the peat bogs of the Norfolk Broads in the east of England, Williams and Lambert (1959) had developed Association Analysis, using a chi-squared coefficient to measure the association between pairs of quadrats as a function of presence and absence of species. Later, Williams' main collaborative colleague on numerical ecology was G. N. Lance, who became Chief of the CSIRO Computing Division and set up one of the world's first large-scale (pan Australian) computing networks. Originally, ecologists were concerned with one-dimensional "ordination", identified with ecological clines, but were becoming interested in two-dimensional (or more) descriptions. Williams had noticed that a "Principal Components Analysis" of an association matrix (i.e. its eigenvectors) gave acceptable two-dimensional ordinations. The justification turned out to be that a distance proportional to the square-root of dissimilarity was being approximated, leading to Principal Coordinates Analysis (Gower, 1966); compare the parallel psychometric development from scaling to multidimensional scaling, with the development of Classical Scaling. An interesting reaction to the justification of what had been a purely empirical observation was the notion that whatever might be done by non-statisticians could expect eventual statistical validation.

Another very influential paper at this period, also emanating from the taxonomic world, was that of Sokal and Michener (1958), which introduced the Unweighted Pair-Group method of clustering. Gower (1967) compared this with other methods, including Association Analysis, showing that several methods were variants of progressively minimizing weighted sums-of squares within groups. Variant methods hinged on how similarity between groups was recalculated after one pair had merged. Several general formulae became available, controlled by parameters that allowed a whole range of clustering procedures to be accommodated within a single program. The Rothamsted programs soon fell into this class, thus allowing not only for a wide range of similarity coefficients but also a wide range of clustering algorithms. The situation was becoming anarchic, with taxonomists being able to select the combination of coefficient and cluster method that best satisfied their prejudices. So, it seemed welcome, if a little curious, when Jardine and Sibson (1967, 1971) showed that the single linkage method uniquely satisfied a set of plausible axioms. Disappointingly, although a nice result in mathematics, it seemed that the axioms were too stringent. Of concern to me was that the result implied that the single linkage method was acceptable, however unsuitable the data might be for hierarchical representation, a consequence that could be interpreted as a strong reason for not using single linkage. Actual data might very well conform to the axioms with a whole range of clustering methods, so we were not much further forward.

The statistical reaction

As we have seen, much of this early work in the general area of classification was initiated by non-statisticians and indeed by people who would not claim to have a mathematical background. It was not always well-received by professional statisticians, as might be perceived by reading Cormack's (1971) review of work to that date and the following discussion. In my opinion there were two general areas of potential misapprehension (i) the meaning to be given to the term classification and (ii) the role of probability. To the statistical world, ever since Fisher's formulation of discriminant analysis, classification meant assignment of samples to previously established classes and this was done in a stochastic framework that could be expressed (though not by Fisher himself) in terms of the probabilities of correct and incorrect classification. Like the statisticians, the taxonomic world is also interested in assignment to classes, termed by them identification, and ever since the 17th century had successfully used non-stochastic hierarchical identification keys for the purpose (see Gower and Payne, 1986 for a discussion of algorithms developed for identifying yeasts). However, taxonomists, unlike statisticians, were equally concerned with the establishment of the classes themselves – species, genera, families etc. - and reserved the term classification for this activity. Forming classifications of biological populations could be expressed in a non-stochastic framework by, so far as possible, choosing categorical characteristics that did not change within populations - easily achievable, except for closely related populations where quantitative variables become important and taking into account stochastic intra-population variation is unavoidable. To some statisticians, used to handling within-population variable quantitative variates, the proper approach to forming classes was that of disentangling mixtures of populations, first studied in univariate form by Pearson at the end of the nineteenth century; computers now allowed multivariate mixture problems to be tackled. Such problems certainly exist but they are not in the mainstream of what is meant by a taxonomic classification; further, they tend to have unpleasant statistical and computational properties.

As I see it, the concept of the multivariate data matrix tends to blur important distinctions that fuel confusion. To some, especially statisticians, a data matrix represents a sample of size n from a single multivariate distribution of p usually quantitative stochastic variables; to others, especially taxonomists, it represents n distinct populations described by p usually non-stochastic categorical variables. The intermediate situation where the n samples are divided into k recognized groups is very common; in the case of mixtures, the k groups are initially undifferentiated and have to be recognized. More elaborate structure might be imposed on the samples and we may also impose structure on the variables, but shall not pursue that here. A further source of confusion is between a data matrix, as just discussed, and a two-way table as used for example in correspondence analysis or when fitting a biadditive model. The root of the problem is that these things are all two-way arrays that cannot be distinguished by a computer and so are open to invalid methods of analyses. The computer, or package, user has to be aware of the different structures and what methods of analysis are appropriate to their data.

Gower and Ross (1998) discuss further the relationships between stochastic and non-stochastic classification methods. The lack of a stochastic element was not the only criticism made by statisticians but also the nature of the methods based on algorithmic recipes with no clearly stated objective model, let alone a criterion for fitting it. This is not only a feature of classification analysis but is to be found in other areas of data analysis to which I shall return below.

What is classification for?

Both identification trees and taxonomic classifications were expressed as tree structures. A source of much, sometimes heated, discussion among the taxonomists themselves, divided into the pheneticists and the cladists, was to what extent these tree structures corresponded to evolutionary trees; special methods were developed for constructing evolutionary trees, some of a stochastic nature and some not. Of course, it would be nice if phenetic, cladistic and identification trees were all the same and, indeed, it is very likely that the three are far from independent; things that look alike often, but not necessarily, share a common evolutionary lineage; things sharing a common lineage or which look alike will share characters useful for identification. Fundamentally, pheneticists and cladists had different objectives and it surprised me that it was not accepted that these might demand different classifications. What really underlies the phenetic-cladistic controversy is that there is more than one reason for forming a classification. When one moves beyond biological taxonomy, reasons for classification can have no concern with evolutionary desiderata. One good general reason for forming classes is so that they have the property that assigning new objects to these classes is optimal in some manner. Gower (1974) developed the non-stochastic method of Maximal Predictive Classification to form classes with the property that with assignment of an object to a class, one could predict correctly the greatest number of properties of that object; the class predictors are optimal for identification. Similarly, it can be shown that classes given by mixture models are optimal for future (stochastic) discrimination. In this way, forming classes and assigning to classes may be closely interlinked.

This dual relationship between class-formation and class-assignment seems also to be at the root of controversies over the role of correlation between variables in classification problems. Broadly speaking, a set of highly correlated variables are redundant in discriminatory problem; any one will do as well as the others, so all but one may be discarded. Yet, it is the very correlation between variables that leads us to form classes, because it is natural to put things together that share many properties. Indeed, this is the notion behind Maximal Predictive Classes that successfully permit the prediction of sets of associated characters. Thus, it seems to me that when forming classes we need to recognize correlated sets but once classes are formed, independent characters are better for assignment. An early manifestation of this issue was the controversy over the relative merits of Pearson‘s Coefficient of Racial Likeness, which ignores correlation, and Mahalanobis’ D-squared, which downgrades correlated characters (see, Fisher, 1936). The whole issue is complicated by when and how to differentiate between inter and intra class correlation.

Although taxonomists form classes of populations on the basis of non-stochastic variables, there is no reason why statisticians should not form classes from populations described by stochastic variables but this problem seems to have been neglected. When forming classes of populations, the degree of overlap between populations may not be the only relevant information. To see this, consider three, say, populations, two with similar means, very different from the mean of the third population. If the third population is much more variable than the other two, then it might be regarded as being stochastically more similar to the first two than each of these is to the other. Nevertheless, it may be better to place the first two populations together in a class separately from the third. In the limit, when there is no overlap, assignment to a population may be made with certainty but still there may be good reasons for grouping populations with similar, though non-overlapping, measurements.

Measures of fit

Few would hold that the impressive work of the 1930s and 1940s on the functional forms of the distributions of various multivariate statistics has been of much value in classification work, or in other areas of data analysis, except possibly in the field of discriminant analysis. In practice, the common tools for assessing results are the bootstrap, jackknife and permutation tests. These are all appropriate when the milieu is of a stochastic nature, but as we have seen in taxonomy, often it is not. Then we have two possibilities (a) there may be a stochastic element in the choice of *characters*, rather than *samples*, used as the basis of a classification and (b) we may be more concerned with approximation rather than with stochastic error. Different choices of similarity coefficients will give different results and the effects of such choices might usefully be investigated by bootstrap-like techniques, now omitting variables rather than samples. Approximation, is a familiar and useful idea from 18th century mathematics, in which polynomials (Legendre, Laguerre,…) are used to approximate more complicated mathematical functions. The difference between the exact function and the polynomial approximation is systematic and nobody would claim it to have any stochastic interpretation. It seems to me that many taxonomic classifications and matrix approximations are often better interpreted in this way. For example, depending on the type of data matrix (see above), a two-dimensional PCA approximation, using Eckart and Young's theorem, may be sometimes better interpreted as an approximation to an exact fit in higher dimensional space, rather than one of minimizing a sum-of-squares of stochastic residuals. I think we should more often consider returning to the 18th century. One might add here that sensitivity analysis, the effect of outliers and robustness of fit are not necessarily the prerogatives of stochastic formulations.

Many methods developed by those with a biological background were of an algorithmic nature with no clearly expressed model or loss-function to define the best fit. To provide a more formal framework for classification, the ultrametric property of trees, and later the notion of additive trees, were proposed as models to be fitted by least squares. In the early 70's, noting that ultrametric trees could be embedded in Euclidean space, I measured fit by the best least-squares rotation of a tree to the unconstrained Euclidean representation of the data (later reported by Gower and Banfield, 1975). Ideally, one

would like to minimise this criterion but this seems an even more intractable problem than the more popular alternative of minimising the ultrametric stress criterion.

Procrustes analysis, asymmetry and biplots

Using rotations to measure the fit of trees was the start of my interest in Procrustes analysis and it seemed to be a dead-end. Then, Alan Bilsborough from Cambridge came to me with data on populations of fossil skulls, based on sets of variables defined by different regions of the skull, including the jaw bone. The different parts of the skull each gave a canonical variate analysis whose configurations represented the populations. These configurations could be compared in pairs by Procrustes analysis to give an association matrix for the different skull regions. In turn, these could be analysed by MDS or other established methods to map how each part of the skull contributed to the separation between populations (Gower, 1971). Later, it became apparent that Procrustean methods were useful for comparing the performance of judges' assessments of meat carcasses, a problem I had been consulted on by the Meat Research Institute. I developed Generalised Procrustes Analysis to compare the simultaneous performance of several judges (Gower, 1975). This methodology is particularly useful when different judges use different variables of their own choice, so-called free choice profiling. Procrustes Analysis is now important in the analysis of biological shape and many further applications and developments have emerged (see Gower and Dijksterhuis, 2006 for further references).

Multidimensional scaling and hierarchical classification methods may be viewed as methods for approximating a symmetric matrix of distances, dissimilarities or similarities. Square non-symmetric tables with rows and columns labelled by different modes of similar concepts (e.g. import/export, father/son,…) are common. Often asymmetry was ignored by analysing the symmetric elements $a_{ij} + a_{ji}$ but could there be useful information in the skew-symmetric component $a_{ij} - a_{ji}$? Gower (1977) and Constantine and Gower (1978) developed the necessary algebra and associated visualisations for a least-squares approximation; Gower (1980) gave some biological applications. The non-Euclidean geometry developed to interpret skew-symmetry in terms of triangular areas is interesting. Other methods are available for analysing asymmetry, including those that ignore the square nature of the table, but I think that different mechanisms often underlie symmetry and departures from symmetry, so it can be useful to separate the components and then examine whether there may be links between the two parts.

I had long been aware of Ruben Gabriel's (1971) work on biplots for PCA and biadditive models. In PCA, this allows information on the variables to be represented as vectors, together with the usual display of the approximate relative distances between the samples. I had always regarded PCA as a multidimensional scaling method, so it was natural to consider how similar information could be supplied for other MDS methods. Gower and Harding (1988) showed how this could be done with nonlinear biplots for quantitative variables in classical scaling, which like PCA is a projection method; Gower (1992) showed how categorical variables could be included and later extended the

methodology to general methods of metric and nonmetric multidimensional scaling (see chapter 3 of Gower and Hand, 1996). It became apparent that biplot axes were axes that could be calibrated in the same way as familiar coordinate axes, the various generalisations depending on notions of generalisations of coordinate axes to calibrated nonlinear and labelled convex category-regions. For categorical variables, the calibrations are nominal, possibly ordered on a linear or nonlinear axis. These generalised axes are used by finding the nearest marker or label to a sample point, which of course gives normal projection for axes that form a continuum and, in particular, orthogonal projection for linear axes. It turns out that in approximations, separate sets of axes are required for predicting the values to be associated with a sample from those for interpolating, or positioning, a new sample. In exact representations both sets of axes coincide, as in classical usage. In PCA and Correspondence Analysis, based as they are on the SVD, it is usual to interpret displays by using the inner product but the use of projection onto calibrated axes amounts to the same thing and is easier to use – it also allows the axes to be moved around for convenience whereas inner products depend on a fixed origin.

Influences on the Present

Rather than trying to describe present-day data analysis, I shall make some general remarks on how I see the work prior to 1980 has influenced current trends.

As computers developed, they became more reliable and they could handle larger sets of data with more samples and more variables. Increasing speeds encouraged the development of complex models that hitherto had been beyond reach, together with iterative algorithms for fitting them, and allowed intensive data sampling methods, such as the bootstrap and jackknife, to be used on a routine basis. Further, newly available computer graphics technology made it possible to develop helpful graphical displays. Personal computers arrived in the 1980s and we were freed from the sometimes stultifying effects of the centralised computer service with its batch mode operation and efforts to control computer use. The door was open to universal access to computing and computing costs became so cheap that interactive usage was within grasp. I like to think that we have returned to the days when Fisher could say he learned all his statistics at the calculator.

Data analysis

Data Analysis was born from this computing revolution. Data Analysis, including Exploratory, Confirmatory and Initial Data Analysis, were terms introduced by John Tukey to describe techniques made practicable by the increased computing power of the 1960s. Tukey's Data Analysis was concerned mainly with small samples and with few variables. To some extent, coining the new term Data Analysis was a reaction to using the term Statistics, which had been subsumed by Mathematical Statistics, itself becoming increasingly remote from analysing data. Nevertheless, I regard Data Analysis as a synonym for Statistics. Multivariate Data Analysis was concerned with the analysis of many variables, though there was disagreement as to whether the variables concerned had to be random variables or not. Thus, some did not regard Multiple Regression as a multivariate method because in the Gauss Linear model, only the dependent variable is a

random variable. Similarly, most analyses of tabulated data based on Linear Models or Generalised Linear Models would not be regarded as multivariate methods. Neither would biadditive models or Correspondence Analysis, even though both give multidimensional representations of their respective models. Yet PCA, which uses very similar algebra, is generally regarded as a genuine multivariate method as is its categorical variable equivalent, Multiple Correspondence Analysis. To cut through these semantic quibbles, I prefer to join those who refer to Multidimensional Data Analysis to distinguish that part of statistics which has been the topic of this paper and, I believe, this entire publication. Analyse de Donnees cannot be translated as Data Analysis as it seems to be a special part of Multidimensional Data Analysis/Statistics concerned especially with Clustering methods and Correspondence Analysis - both of a two-way contingency table and Multiple Correspondence Analysis of several categorical variables. My preference is to use the term Statistics whenever possible, but when we especially want to refer to multidimensional displays, as in my title, then use Multidimensional Data Analysis.

Complex models

One thing that has emerged in recent years is that there is now an increase in the types of term that may be included in models. Thus, we may wish to include distances or ultrametrics or any other function that has attractive interpretational properties. Often, such terms are included in stand alone form but hybrid models have been around for over twenty years though they await adoption by GLMs or GAMs, or other recently developed classes of model. Specified transformations (e.g. the link function of GLMs) are included in these models but in nonmetric MDS the transformations are derived by the methodology, trading reduced dimensionality for a complicated transformation. One manifestation of the increase in complexity, is the way interactions may be modelled. Additive interaction parameters have long been included in linear models and we have seen that multiplicative biadditive interactions were considered, but little-used, as long ago as the 1920s. Nowadays, triple-product (or higher order) interactions may be fitted in a variety of forms. Unfortunately, from the statistical outlook, such model terms are not readily combined with main effect and two-factor interaction terms because they are rarely consistent with simple linear reparamaterisations, as was already shown by Gower (1977). If one holds with the view that high order interactions usually imply the occurrence of lower-order interaction and main effects, then triple products are seriously flawed to be considered as valid interaction terms. Although triple-product models may sometimes be useful in stand-alone form, they do not have the same status in more extended models as do additive interaction and biadditive parameters. Another difficulty is that it is quite difficult to interpret triple product terms. Extensions to three-way interaction is just one example where generalisation of well-understood simple models may be less useful than at first sight; my inclination is to keep things simple.

Apart from allowing increased model complexity, modern computing power allows one to handle vast amounts of data; hence the interest in large data sets and in data-mining. Automatic instrumentation ensures that there is no lack of large data sets. This is something new but I sometimes think that we are getting perilously close to the search for the philosophers stone. If only we could find the right recipe, great truths would be

revealed by analysing vast masses of data. My feeling is that it is at least as important to be careful about the quality of the data, something that seems to be more neglected than in the past, except perhaps in the conduct of clinical trials.

Conclusion

New fields of study and technological advances have always stimulated statistical developments. Indeed, this is the background to much of what has been described above. It continues today as new forms of measurement are developed including, among others, microarrays, DNA sequences, spatial data, pixels, satellite imagery, tomography, shape. Although not a form of measurement, the technology of the VDU screen has been a valuable addition to the armoury of statisticians. It may be used to display long established forms of diagrams as well as stimulating new forms. The use of colour enhances the possibilities. Fast computing speeds allow dynamic graphics, showing how configurations change as parameter values change. The VDU screen has revolutionised interactive computing, especially in our field of multidimensional data analysis. But the news is not all good. Software packages bring new methodology within the reach of all. This is good but on reading research papers written by the users of these packages one wonders whether researchers always understand what methods are suitable for their research and whether they know how to interpret the computer output. We have a major communications problem.

There was a time when a problem was modelled in algebraic form, some criterion such as maximum likelihood or least squares specified to fit the model and then a good algorithm found to do the computations. The properties of the model, parameter estimates, the fitting criterion and of the algorithm could be studied. With luck, the problem would have a solution in closed form meaning that it could be fitted in terms of known functions. A known function was one whose properties had been studied and which could be computed and either tabulated or computed in terms of other known functions. The classical problem of this kind is given by the linear model, fitted by least-squares with matrix inversion algorithms; this has been studied since the late eighteenth century, and continues to be a source of interesting research problems. We have already remarked how things changed from the very beginning of the computer age. Commonly, no model was specified and, as with clustering algorithms, only an algorithm was specified, the whole notion of closed form solutions becoming irrelevant. As more and more complex algorithms were developed, themselves using other possibly recently developed algorithms, the class of available functions greatly increased. With few exceptions the properties of these algorithmically defined functions were little known and they could not be regarded as known functions. Even when a function had an explicit form, little was known of its properties – did it have a unique solution or were there multiple roots, was it continuous etc. When only an algorithm was specified, a new form of research developed with the objective of trying to determine what objective function might be suggested by the algorithm. When an objective function is available, one can develop different algorithms for its calculation and study their properties such as speed and accuracy. Many new functions are confounded with software and it seems to me that much work is needed to evaluate and assess their properties.

My closing paragraphs may appear to have a negative aspect, drawing attention to problems. I am aware that in the twenty-first century we are not supposed to have problems, merely challenges. Perhaps, some of my problems may be regarded as challenges but I suspect that others really are problems

References

Cormack, R. M. (1971). A review of classification (with discussion). *Journal of the Royal Statistical Society,* A, **134**, 321 – 367.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika* **1**, 211 – 218.

Fisher, R. A. and Mackenzie, W. A. (1923). Studies in crop variation. II. The manurial response of different potato varieties. *Journal of Agricultural Science* I, **13**, 311 – 320.

Fisher, R. A. (1936). CRL and the future of craniometry. *Journal of the Royal anthropological Institute*. **66**, 57 – 63.

Fisher, R. A. (1940). The precision of discriminant functions. *Annals of Eugenics* **10**, 422 – 429.

Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**, 325 – 338.

Gower, J.C. (1966). A Q-technique for the calculation of canonical variates. *Biometrika* **53**, 588 – 589.

Gower, J.C. & Ross, G.J.S. (1969). Minimum spanning trees and single linkage cluster analysis. *Applied Statistics* **18**, 54 – 64.

Gower, J.C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* **27**, 857 – 871.

Gower, J.C. (1967). A comparison of some methods of cluster analysis. *Biometrics* **23**, 627 – 637.

Gower, J.C. (1971). Statistical methods for comparing different multivariate analyses of the same data. In: *Mathematics in the Archaeological and Historical Sciences*. Eds. J.R. Hodson, D.G. Kendall & P. Tautu. Edinburgh: Edinburgh University Press. 138 – 149.

Gower, J.C. (1974). Maximal predictive classification. *Biometrics* **30**, 643 – 654

Gower, J.C. (1975). Relating classification to identification. In: *Biological identification with computers. Systematics Association Special Volume No. 7.* Ed, R.J. Pankhurst. London: Academic Press, 251 – 263

Gower, J.C. & Banfield, C.F. (1975). Goodness of fit criteria for hierarchical classifications and their empirical distributions. In: *Proceedings of the Sixth International Biometric Conference, 1974*. Eds.: L.C.A. Corsten & T. Postelnicu. Bucharest: Editura Academiei Republicii Socialiste Romania. 374 – 361.

Gower, J.C. (1975). Generalized Procrustes analysis. *Psychometrika* **40**, 33 – 51.

Gower, J.C. (1977). The analysis of three-way grids. In: *Dimensions of Intra Personal Space, Vol. 2: The Measurement of Intra Personal Space by Grid Technique*. Ed.: P. Slater. Chichester: J. Wiley & Sons. 163 – 173.

Gower, J.C. (1977). The analysis of asymmetry and orthogonality. In: *Recent Developments in Statistics*. Eds.: J. Barra et al. Amsterdam: North Holland Press. 109 – 123.
Constantine, A.G. & Gower, J.C. (1978). Graphical representation of asymmetry. *Applied Statistics* **27**, 297 – 304.
Gower, J.C. (1980). Problems in interpreting asymmetrical chemical relationships. In: Chemosystematics: Principles and Practice. *Systematics Association Special Volume No. 16.* Eds.: F. Bisby J.C. Vaughan & C.A. Wright. London: Academic Press, 399 – 409.
Gower, J.C. & Payne, R. W. (1986). On identifying yeasts and related problems. In: *The Statistical Consultant in Action*. Eds. D.J. Hand & B.S. Everitt. Cambridge: Cambridge University Press. 108 – 120.
Gower, J.C. & Harding, S.A. (1988). Nonlinear biplots. *Biometrika* **75**, 445– 455.
Gower, J.C. (1990). Fisher's optimal scores and multiple correspondence analysis. *Biometrics* **46**, 947– 961.
Gower, J.C. (1992). Generalized biplots. *Biometrika* **79**, 475 – 493.
Gower, J.C. and Hand, D. J. (1996) *Biplots*. London: Chapman and Hall, 277 + xvi pp.
Gower, J.C. and Ross, G. J. S. (1998). Nonprobabilistic Classification. In: *Advances in Data science and Classification*. Eds. A. Rizzi, M. Vichi and H.-H. Bock. Springer, Berlin, 21 – 28.
Gower, J.C. and Dijksterhuis, G. (2004). *Procrustes Problems*. (Oxford Statistical Science Series No.30). Oxford: Oxford University Press. 233 + xiv pp.
Gower, J. C. (2006). Divided by a Common Language. In: *Multiple Correspondence Analysis and Related Methods*, Eds. Joerg Blasius and Michael Greenacre. Boca Raton, Florida, Chapman and Hall. 77 – 105.
Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In: *The Prediction of Personal Adjustment*, P. Horst et al. (eds), 319 – 348. Bulletin No. 48. New York: The Social Science Res. Counc.
Hirschfeld, H.O. (1935). A connection between correlation and contingency. *Proceedings of the Cambridge Philosophical Society* **31**, 520 – 524.
Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**, 417 – 441, 498 – 520.
Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* **28**, 321 – 377.
Jardine, N. and Sibson, R. (1968). The construction of hierarchic and non-hierarchic classifications. *Computer Journal*, **11**, 177 – 184.
Jardine, N. and Sibson, R. (1971). *Mathematical taxonomy*, New York: Wiley.
Mahalanobis, P. C. (1936) On the generalized distance in statistics. *Proceedings of the National Institute of Science India*. **12**, 49 – 55.
Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Phlosophical Transactions of the Royal Society, Series A,* **185**, 71 – 110.
Pearson, K. (1901). On lines and planes of closest fit to a system of points in space. *Philosophical Magazine*, **2**, 6th Series, 557 – 572.
Pearson, K. (1926). On the coefficient of racial likeness. *Biometrika,* **18**¸105 - 117.
Sneath, P.H. (1957). The application of computers to taxonomy. *Journal of General Microbiology*, **17**, 1409 – 1438.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships, **38**, *University of Kansas Scientific Bulletin*, 201 – 226.
Williams, W. T. and Lambert, J. M. (1959). Multivariate methods in plant ecology. I. Association analysis in plant communities. *Journal of Ecology*, **47**, 83 – 101.
Wishart, J. (1928). The generalized product moment distribution in samples from a normal multivariate distribution. *Biometrika*, **20A**, 32 – 52, (correction **20A**, 424)