This Week's Citation Classic

CC/NUMBER 12
MARCH 24, 1980

Gower J C. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**:325-38, 1966. [Rothamsted Experimental Station, Harpenden, Herts., England]

Values dj: are a measure of 'distance' between individuals i and j. Conditions are given for points P_i to exist where the distances A (P_i, P_j) = d_{ij}. The method finds an approximate configuration Q_i, in a few dimensions, such that A(Q_i,Q_j) ≈ d_{ij}. Principal components, canonical variates, and factor analysis appear as special cases. [The *SCI*® indicates that this paper has been cited over 215 times since 1966.]

John C. Cower
Department of Statistics
Rothamsted Experimental Station
Harpenden, Herts., AL5 2JQ
England

October 15, 1979

"Multivariate data typically record the values of several chosen characteristics for each of many samples. Classical statistical methods usually operate on matrices of coefficients-of-association between pairs of characteristics (R-matrices). But, in the early 1960s, matrices of coefficients-of-association between pairs of samples (Q-matrices) arose naturally in taxonomic and ecological studies. New analytical techniques were developed (e.g., cluster analysis), but there was also a tendency to use the old R-matrix methods on Q-matrices which, although mathematically dubious, suggested biologically acceptable interpretations. The more important eigenvectors of the Q-matrix would be taken, as in principal components analysis (PCA), and the set of ith components of each vector treated as a coordinates of a point P_i, as when plotting factor loadings in factor analysis (FA). Similar samples tend to give neighbouring points, and dissimilar samples distant points. Although heuristically satisfying, the mathematical justification for such procedures needed investigation.

"When using a distance matrix (d_{ii} = O) or similarity matrix (d_{ii} = 1), I showed that although the heuristic methods had an acceptable theoretical least-squares basis they could be improved. Conditions for the equivalence of Q and R techniques were found and the classical multivariate methods of PCA, canonical variates analysis and, less successfully, FA were shown to be special cases of a more general formulation. An obvious relationship to PCA suggested that the new method be termed 'principal coordinates analysis' (PCO); this name caught on, especially in the biological literature. All this satisfactorily clarified things and unified several apparently little-related problems, but the success of PCO had one unforeseen consequence. Because one heuristic method had been justified, there was a tendency to discount mathematical objections to some very strange procedures, apparently in the belief that any heuristic method eventually achieves mathematical respectibility. PCO was quickly incorporated into computer programs (primarily designed for taxonomic studies) and naturally was cited by scientists using these programs. However, I think the main reason for its frequent citation is that it produces scatter diagrams allowing visual inspection and display of the (taxonomic) relationships. Such diagrams are basic to all data analysis but the high dimensional samples of multivariate data present special problems that PCO substantially overcomes.

"The development of PCO nicely illustrates some of the vagaries of scientific communication. Only after the paper was accepted for publication did a chance discovery of Torgerson's parallel work on multidimensional scaling (MDS) in psychology enable me to insert a last minute reference.[1] MDS, itself founded on earlier work by Young & Householder,[2] revealed that the problem had its origins in the educational/psychometric area as early as 1938, but statisticians not working in psychology (including myself) were totally unaware of this work. Similarly writers in the current psychological literature are seemingly unaware of much of the biometric literature. With such compartmentalisation it is no surprise that scientific ideas can be spread in unlikely ways, as when it took a visitor from the US to introduce PCO to an entomologist colleague at Rothamsted.

"Despite new, apparently less restrictive, methods such as non-metric MDS (which sometimes uses PCO to initiate its iterative algorithm), the older method retains its popularity because of its robustness and because, unlike the newer methods, it has a simply-computed globally optimum solution."

1. **Torgerson W S.** *Theory and methods of scaling.* New York: Wiley, 1958. 460 p.
2. **Young G & Householder A S.** Discussion of a set of points in terms of their mutual distances. *Psychometrika* **3**:19-22, 1938.