TRUSS NETWORK ANALYSIS

T. V. Sathianandan and K. G. Mini
Fishery Resources Assessment Division
ICAR- Central Marine Fisheries Research Institute

27

Classification problems exists in numerical taxonomy in biology and many other branches of Science. The interest here is to classify objects into one of many existing classes and is based on measurements taken on a set of characteristics (called variables). Hence classification is a multivariate problem which can be divided into two broad categories.

- We have multiple measurements data from a number of individuals belonging to known groups. Also we have data collected on individuals whose group membership is not known and is to be determined using the measurements made on them. This problem in statistical terminology comes under Discriminant Analysis.
- Another type is the case when the groups are them selves unknown and a primary purpose of the analysis is to find groups so that those belonging to same group are similar than those belonging to different groups. This in statistics come under the heading of cluster analysis or pattern recognition.

Cluster Analysis:

This involves the search through multivariate data for observations that are similar enough to each other to be usefully identified as part of a common cluster. Clusters consist of observations that are close together and that the clusters themselves are separated. If each observation is associated with only one cluster, then the clusters form a partition of the data. Finding the partition into clusters is not always easy. There are numerous methods for clustering. Some methods of making clusters starts with models like mixture models of clusters. Examples of application of cluster analysis are studying genetic diversity within and between populations of and endangered fish species, clustering species of bees into higher-level taxonomic groups, developing clusters of patients based on physiological variables, constructing a speaker-independent word recognition system etc. Numerical methods of clustering with out any model can be into three major types; *hierarchical, partitioning* and *over lapping*.

Principal Component Analysis

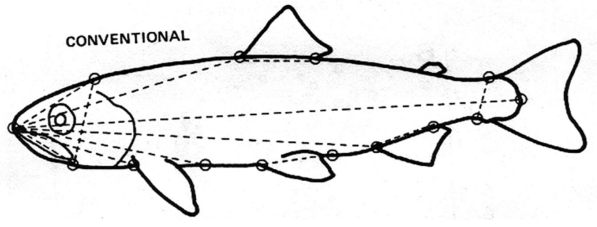
In principal component analysis we have a sample of observations taken on a set of variables and the objective is to find linear combinations of the variables so that the first linear combination accounts for maximum possible variation in the data, the second linear combination accounts for the next highest possible variation and so on. By this we get another set of transformed variables which are linear combinations of the original variables and they new set will have the property that by considering few of them we will be able to explain a major portion of the variability in the population. The approach in principal component analysis is to reduce dimensions by calculating the eigen values and eigen vectors of the covariance or correlation matrix and project the data orthogonally into the space spanned by the eigen vectors belonging to the largest eigen values. These projections are interesting due to the following reasons

- If projection is an aggregate of several clusters, then these can become individually visible only if the separation between clusters is larger than the internal scatter of the clusters. Thus, if there are only a few clusters, the leading principal axes will tend to pick projections with good separations.
- It tend to act as a variation reducing technique relegating most of the random noise to the trailing components and collecting the systematic structure into the leading ones.

Suppose that we have measurements on k variables $x_1, x_2, \cdots, x_k$ made on n individuals. Then we have *n* x *k* matrix of data and we can work out means for these variables which we can treat as a mean vector of length *k*. Also we can compute the variance covariance matrix **S** matrix using this data set. This matrix will be then used to compute the k principal components, say $z_i = a_{1i}x_1 + a_{2i}x_2 + \cdots + a_{ki}x_k$ for $i = 1,2,\cdots,k$ and the amount of variation explained by each of them will be available as $\lambda_1, \lambda_2, \cdots, \lambda_k$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$.

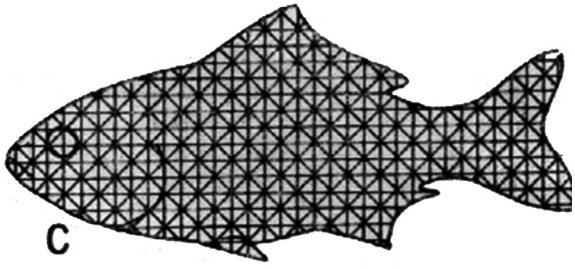
Truss Network Analysis

In systematics the interest is often in quantifying differences in form among different species or conspecific populations. When these are studied using conventional measurements (shown below) the amount of information available for analysis are repetitious and lack variation in oblique directions.



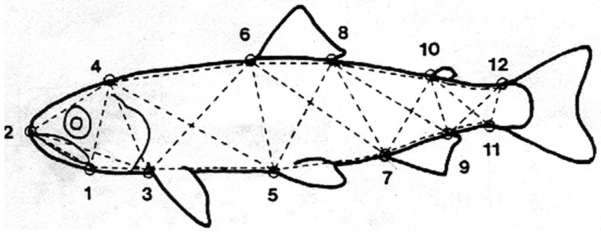
There are several biases and weaknesses inherent in traditional character set used to study stock differences in systematics.

- They tend to be in one direction only (longitudinal) lacking information of depth and breadth.
- Coverage is highly uneven both by region and orientation.
- Some landmarks like tip of the snout and posterior end of vertebral column are used repeatedly.
- Many landmarks are external rather than anatomical and their placement may not be homologous placement may not be homologous from form to form.
- Many measurements extends over much of the body.
- When measurements are taken on soft bodied organisms, the amount of distortion due to preservation can not be easily estimated.



The most ideal measurements which overcomes these problems is as in the picture C.

Truss is a geometric protocol for character selection which largely overcomes the disadvantages of conventional data sets and it leads to certain style of analysis. In *truss* system, homologous landmarks on the boundary of the form are divided into two tiers and paired. The distance measures connect these landmarks into an over determinate truss network which is a series of quadrilaterals each having internal diagonals. Each quadrilateral shares one side with each succeeding and preceding quadrilaterals (see figure).

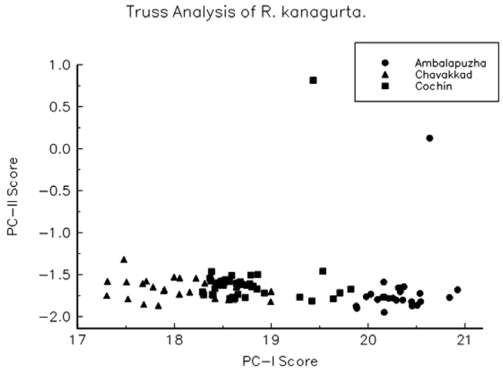


The following are the properties of a truss network measurements.

- It enforces systematic coverage across the form
- It exhaustively and redundantly archives the form
- The degree of measurement error in data can be measured and corrected
- Forms may be standardized to one or more common reference sizes by representing measured distances on some composite measure of body size and reconstructing the form using the distance values predicted at some standard body size.

- Principal components can be given geometrical interpretations. Component scores are measures of configuration while loadings are descriptors of shape change.
- Composite mapped forms are suitable for biorthogonal analysis of shape differences between forms.

In the analysis of multivariate data collected through truss network measurements the concept is that ***size*** and ***shape*** are the two factors which account for the association among the distance measures. *Size* is not considered as a single variable but as a factor which is obtained as a linear combination of the distance measures. *Shape* is considered as the geometry of the organism after information about position, scale and orientation has been removed. The *shape* discriminator should be independent of *size*, for it to be free from the effect of growth. Principal component (PC) analysis which does not require any prior information about groups is used in the analysis of truss data. A logarithmic transformation is first applied to the measurements before performing the PC analysis to reduce variance due to size variation and also because according to an allometric model diverse distance measures relate log linearly in a homogeneous population. The first component factor of the PC analysis is then interpreted as size component (which is not fully free from shape) and subsequent component factors are designated as shape variable (not fully free from size). Then a plot of the first principal component scores against the second principal component scores will more of less show clustering for different groups. The percentage of variation explained by this two factors also should be considered before making conclusions.



Suggested Reading

Anon. 1989. Discriminant Analysis and Clustering. *Stat. Sci.*, 4(1):34-69.

Huber, P.J.(1985). Projection Pursuit. *The Annals of Statistics*, 13(2):435-475.

Humphries, J.M. *et. al.* 1981. Multivariate Discrimination by shape in relation to size. *Syst. Zool.*, 30:291-308.

Morrison, D.F. 1990. Multivariate Statistical Methods. McGraw-Hill, New York. Strauss, R.E. and Bookstein. 1982. The truss: body form reconstruction in morphometrics. *Syst. Zool.*, 31:113-135.