cases. Nonlinear renewal theory has been applied to approximate the properties of several sequential tests and estimates (*see* SEQUENTIAL ANALYSIS). The recent monograph by Woodroofe [15] and text by Siegmund [12] describe the development of nonlinear renewal theory and its applications to statistics. They include references to statistical applications. See also Lalley [11].

References

[1] Blackwell, D. (1948). *Duke Univ. Math. J.*, **15**, 145–150.

[2] Blackwell, D. (1953). *Pacific J. Math.*, **3**, 315–320.

[3] Chow, Y. S., Hsiung, C., and Lai, T. L. (1979). *Ann. Prob.*, **7**, 304–318.

[4] Erdös, P., Feller, W., and Pollard, H. (1949). *Bull. Amer. Math. Soc.*, **55**, 201–204.

[5] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. 2. Wiley, New York.

[6] Hagwood, C. (1980). *Commun. Statist.*, **A9**, 1677–1698.

[7] Hagwood, C. and Woodroofe, M. (1982). *Ann. Prob.*, **10**, 844–848.

[8] Lai, T. L. and Siegmund, D. (1977). *Ann. Statist.*, **5**, 946–954.

[9] Lai, T. L. and Siegmund, D. (1979). *Ann. Statist.*, **7**, 60–76.

[10] Lalley, S. (1972). *Commun. Statist.*, **1**, 193–206.

[11] Lalley, S. (1983). *Zeit. Wahrscheinlichkeitsth.*, **63**, 293–321.

[12] Siegmund, D. (1984). *Sequential Analysis*. SIAM, Philadelphia, PA.

[13] Takahashi, H. and Woodroofe, M. (1981). *Commun. Statist.*, **A10**, 2113–2135.

[14] Woodroofe, M. (1976). *Ann. Prob.*, **4**, 67–80.

[15] Woodroofe, M. (1982). In *Sequential Analysis*, Regional Conference Series in Applied Mathematics No. 39. SIAM, Philadelphia.

(RANDOM WALK
RENEWAL THEORY
REPEATED SIGNIFICANCE TESTS)

MICHAEL B. WOODROOFE

NONMETRIC DATA ANALYSIS

Nonmetric data analysis in its broader sense refers to a set of models and techniques for analysis of nonmetric data. Nonmetric data here refer to nominal or ordinal data (*see* NOMINAL DATA *and* ORDINAL DATA) as opposed to metric data, which refer to interval or ratio data [17] (*see* MEASUREMENT STRUCTURES AND STATISTICS). Nonmetric data (sometimes called qualitative or categorical data) are obtained in a variety of ways. For example, in attitude surveys, the respondent may be asked to endorse attitude statements with which he or she agrees. In some mental tests, the examinee either passes or fails test items. In consumer research, the subject may be asked to rank-order food products according to preference. In multidimensional scaling*, stimulus confusion data which are used as (inverse) ordinal measures of subjective distances between the stimuli, may be taken. In some instances metric data may be "discretized" for the purpose of data analysis.

Methods to analyze nonmetric data may be classified into two major approaches. One is *quantitative analysis of qualitative data* [23], and the other is *parametric approaches to nonmetric scaling* [18–20]. The first approach is primarily descriptive but is more general in its applicability. Nonmetric data analysis, in its narrower sense, usually refers to this first approach. The second approach is less general but is more powerful in situations for which particular models are intended. For other approaches to nonmetric data analysis, see related entries listed after the references.

The essential idea behind the first approach is that nonmetric data are nonlinear transformations of metric data. Thus if an appropriate transformation is applied, the transformed data may be analyzed by a "quantitative" model. Unlike other methods that require data transformations, a specific transformation to be applied does not have to be predetermined in this approach. Both the best data transformation and the best parameter estimates of models are obtained on the basis of a single optimization criterion.

Let y_i denote the ith original observation. This y_i is assumed to be quantified a priori. For example, $y_i = 1$ or 0 depending on whether person i passes or fails a certain test

item, or $y_1 = 2$, $y_2 = 3$, and $y_3 = 1$, if the y_i are rank-ordered and it is observed that $y_2 > y_1 > y_3$. The numbers are assigned and interpreted "nonmetrically." That is, for nominal data, only identity or nonidentity of the numbers (i.e., for any two numbers, $a = b$ or $a \neq b$) is meaningful, whereas for ordinal data, ordinal properties of the numbers (i.e., for $a \neq b$, either $a < b$ or $a > b$) are also meaningful. However, in either case neither the difference nor the ratio of two numbers is meaningful. The y_i is transformed by function f, and $f(y_i)$, the transformed data, is fitted by model $g(X_i, \alpha)$, where X_i is some auxiliary information about i (if there is any), and α is a vector of unknown parameters. Both f and g are real-valued functions, possibly defined only at discrete values of their arguments. The problem is to find f and g such that an overall discrepancy between $f(y_i)$ and $g(X_i, \alpha)$, $i = 1, \ldots, I$ is a minimum. More specifically, define a least-squares* loss function,

$$\text{Stress} = \sum_{i=1}^{I} (f(y_i) - g(X_i, \alpha))^2.$$

This criterion is minimized with respect to both f and α under some appropriate normalization restriction.

General forms of f must be consistent with nonmetric properties of the data. That is, f must be such that the basic properties of nonmetric data are preserved through the transformation. (Such transformations are called admissible transformations.) This implies that f must be monotonic (order preserving) when the data are ordinal, and it must be one to one (identity preserving) when the data are nominal. Within the admissible types of transformations, a specific form of f is determined that minimizes Stress. For a given g the best monotonic transformation is obtained by Kruskal's [10] least-squares monotonic regression algorithm, and the best one-to-one transformation, by least-squares nominal transformation [5]. Since f is determined in such a way that it is closest to g among all admissible transformations, it may be considered to possess the same scale level as model g, provided that model g is appropriate for the data. The scale level of a model is the type of admissible transformations by which defining properties of the model are not destroyed. For example, if g is a distance model, which is a ratio model since the defining properties of the distance (the metric axioms) are preserved by multiplying the distance by a positive constant, f is also considered ratio at least approximately. The nonmetric data are, so to speak, "scaled up" by f to g.

Similarly, specific models (g) to be fitted depend on the nature of the data. For example, if the data are similarity data (*see* MULTIDIMENSIONAL SCALING), a distance model may be employed. If the data are conjoint data (*see* MEASUREMENT STRUCTURES AND STATISTICS), an additive model may be appropriate. Other models that may be fitted include linear regression* models, bilinear models (principal components* and factor analysis* models), and a variety of distance models including the Minkowski and the weighted distance models [3, 9] and the unfolding model [4] (*see* MULTIVARIATE ANALYSIS *and* MULTIDIMENSIONAL SCALING). Whichever model is chosen, model parameters are determined in such a way that Stress is a minimum. For a given f, least-squares estimates of model parameters are obtained as if the current f were metric data.

To illustrate, consider the situation in which ordinal data are analyzed by the regression model. Such a situation arises, for example, when we wish to find out why some cars are regarded as more desirable than others, based on various attributes (e.g., gas mileage) of cars and a preference ranking among them. Let y_i be the ith observation on the dependent variable (the preference rank of the ith car) and X_i the corresponding observations on the independent variables (the values of the attributes). The dependent variable (y_i) is monotonically transformed (so that if $y_i > y_j$, then $f(y_i) \geq f(y_j)$), and the regression coefficients (α) are estimated in such a way that Stress is a minimum. Two algorithms are currently in use for minimizing Stress with respect to f and α. One is the steepest descent algorithm (*see also* OPTIMIZATION *and* SADDLE-POINT

APPROXIMATIONS) used originally by Kruskal [10] for his nonmetric multidimensional scaling. The other is the alternating least squares (ALS) algorithm developed by Young, de Leeuw, and Takane. (This work is summarized in Young [23].) In the steepest descent algorithm, f, which minimizes Stress for a fixed g, is expressed as a function of $g(\alpha)$ and then substituted in Stress. The Stress, which is now expressed as a function of α only, is then minimized with respect to α. In the ALS algorithm, LS estimates of f and g are obtained alternately with one of them fixed while the other is updated. This algorithm is monotonically convergent.

The origin of the quantitative analysis of qualitative data can be traced back to Guttman's scale analysis [8]. This method is still widely used and has regained considerable theoretical interest in recent years [6, 15] (*see* CORRESPONDENCE ANALYSIS). Coombs' unfolding analysis [4] is important in that it was the first to suggest the possibility of recovering metric information from nonmetric data. The current trend in the quantitative analysis of qualitative data began with Shepard's [16] and Kruskal's [9] landmark work on nonmetric multidimensional scaling. Following their work, it was soon realized that models other than the distance model could be fitted to nonmetric data in a similar manner, and several fitting procedures were developed along this line [9, 22]. More recently the ALS algorithm was proposed as a unified algorithmic framework for the quantitative analysis of qualitative data; this has considerably widened the scope of models that can be fitted [6, 23]. For a list of currently available procedures, see Young [23].

In the parametric approaches to nonmetric scaling, nonmetric data are viewed as incomplete data. That is, a complete metric process is supposed to underlie the nonmetric data generation process, but the metric information is assumed to be lost when the observations are made, leaving only ordinal or nominal information in the observed data. Thus, if this information reduction mechanism can be captured in a model, the metric information may be recovered from the nonmetric data by working backward from the data.

As an example, let us discuss Thurstone's [2, 21] classical pair comparison model. In a pair comparison experiment, stimuli are presented in pairs, and the subject is asked to choose one member of a pair according to some prescribed criterion. The data are a collection of partial rank orders. Suppose stimuli i and j are compared in a particular trial. It is hypothesized that each stimulus, upon presentation, generates a latent metric process that varies randomly from trial to trial. Let X_i and X_j denote the random variables for the latent processes of stimuli i and j, respectively. For simplicity let us assume that $X_i \sim N(\mu_i, \frac{1}{2})$ and $X_j \sim N(\mu_j, \frac{1}{2})$. (The μ_i and μ_j represent the mean subjective values of the two stimuli. The variances of X_i and X_j are assumed to be equal, but their size can be arbitrarily set.) It is assumed that stimulus i is chosen when $X_i > X_j$ and stimulus j is chosen when $X_i < X_j$. Under the distributional assumptions on X the probability (p_{ij}) of stimulus i over stimulus j can be stated as

$$p_{ij} = \phi(\mu_i - \mu_j),$$

where ϕ is the distribution function of the standard normal distribution*. The likelihood* of observed data is stated as a function of parameters in the latent processes. For computational convenience, ϕ may be replaced by the logistic distribution* [14]. In any case μ_i and μ_j may be estimated to maximize p_{ij} if in fact stimulus i is chosen over stimulus j.

This basic principle can be extended in various ways. Suppose that μ_i represents a combined effect of one or more factors. It may then be appropriate to characterize the μ_i by an additive function of these factors. Pair comparisons of such μ_i provide the data for additive conjoint analysis [19]. As another example, suppose two pairs of stimuli are presented and the subject is asked to choose a more similar pair (this method is called the method of tetrads, which involves

pair comparisons of two similarities). Then stimulus (dis)similarities may be represented by a distance model, and then they are subject to pair comparisons. Nonmetric multidimensional scaling (in the sense of the second approach) is feasible with the pair comparison data [20]. As in the first approach, various other models may be fitted in a way that is consistent with the nature of the data.

Another line of extension is possible with regard to the kinds of judgments that are made. Stimuli may be rank-ordered. They may be rated on a categorical rating scale. A choice may be required among several comparison stimuli. In each case a specific model of information reduction mechanism (similar to that used in pair comparison situation) may be built into parameter estimation procedures. Then essentially the same analysis can be done as in the pair comparison case. Such procedures have been developed for similarity ratings [18], for similarity rankings [20], and for additivity analysis of rating and ranking data [19].

The history of the parametric approaches to nonmetric scaling is even older than the quantitative analysis of qualitative data. Thurstone's pair comparison model was originally proposed in the 1920s [21]. A similar model was developed in mental testing situations [13] in the early fifties. Around the same time, latent structure analysis* [12] was proposed, which accounts for observed response patterns to items by hypothesizing *latent structures*. (Again, conceptually, this is very similar to Thurstone's approach.) See Andersen [1], Bock and Jones [2], and Goodman [7] for recent developments in these models. More recently, Takane [18–20] has developed the conceptual framework for the parametric approaches to nonmetric scaling that is presented here.

References

[1] Andersen, E. B. (1980). *Discrete Statistical Models with Social Science Applications*. North-Holland, Amsterdam. (An excellent treatment of exponential family distributions for discrete data analysis.)

[2] Bock, R. D. and Jones, L. V. (1968). *The Measurement and Prediction of Judgment and Choice*. Holden-Day, San Francisco. (A comprehensive statistical treatment of Thurstonian scaling.)

[3] Carroll, J. D. and Chang, J. J. (1970). *Psychometrika*, **35**, 283–319. (A proposal of individual differences model in MDS.)

[4] Coombs, C. H. (1964). *A Theory of Data*. Wiley, New York.

[5] De Leeuw, J., Young, F. W., and Takane, Y. (1976). *Psychometrika*, **41**, 471–503. (The first account of the ALS algorithm.)

[6] Gifi, A. (1981). *Non-linear Multivariate Analysis*. Department of Data Theory, University of Leiden, The Netherlands. (An account of Guttman's scale analysis by ALS.)

[7] Goodman, L. A. (1978). *Analyzing Qualitative/Categorical Data*. Abt Associates, Cambridge, MA. (Recent developments in latent structure analysis.)

[8] Guttman, L. (1941). In *The Prediction of Personal Adjustment*, P. Horst, ed. Social Science Research Council.

[9] Kruskal, J. B. (1964). *Psychometrika*, **29**, 1–27. (The first theoretically rigorous nonmetric MDS.)

[10] Kruskal, J. B. (1964). *Psychometrika*, **29**, 115–129.

[11] Kruskal, J. B. (1965). *J. R. Statist. Soc. B*, **27**, 251–265. (An application of the monotonic regression to additivity analysis.)

[12] Lazarsfeld, P. F. and Henry, N. (1968). *Latent Structure Analysis*. Houghton Mifflin, Boston.

[13] Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Earlbaum, Hillsdale, NJ. (An up-to-date illustration of latent trait test theory.)

[14] Luce, R. D. (1959). *Individual Choice Behavior*. Wiley, New York. (An axiomatic choice model and its mathematical properties.)

[15] Nishisato, S. (1980). *Analysis of Categorical Data*. University of Toronto Press. (The first English text on Guttman's scale analysis and its multidimensional extension, called dual scaling or correspondence analysis.)

[16] Shepard, R. N. (1962). *Psychometrika*, **27**, 125–140 and 219–246. (The first work on nonmetric MDS ever published.)

[17] Stevens, S. S. (1951). In *Handbook of Experimental Psychology*, S. Stevens, ed. Wiley, New York.

[18] Takane, Y. (1981). *Psychometrika*, **46**, 9–28.

[19] Takane, Y. (1982). *Psychometrika*, **47**, 225–241.

[20] Takane, Y. and Carroll, J. D. (1981). *Psychometrika*, **46**, 389–405. (References 18–20 describe parametric approaches to nonmetric scaling for different models and data.)

[21] Thurstone, L. L., (1959). *The Measurement of Values*. University of Chicago Press, Chicago. (A collection of his works.)

[22] Young, F. W. (1972). In *Multidimensional Scaling*, Vol. 1, R. Shepard et al., eds. Seminar Press, New York. (Polynomial conjoint scaling. An extension of Kruskal's algorithm to other models.)

[23] Young, F. W. (1981). *Psychometrika*, **46**, 357–388. (The most up-to-date account of the quantitative analysis of qualitative data. An excellent bibliography on this approach.)

(COMPONENT ANALYSIS
CORRESPONDENCE ANALYSIS
LATENT STRUCTURE ANALYSIS
MEASUREMENT STRUCTURES AND STATISTICS
MULTIDIMENSIONAL SCALING
NOMINAL DATA
OPTIMIZATION IN STATISTICS
ORDINAL DATA
REGRESSION (various entries))

YOSHIO TAKANE

NONOBSERVABLE ERRORS

In the general linear model*

$$Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{Y}$ is a $k \times 1$ vector of *observed* sample values the random component $\boldsymbol{\epsilon}$ is often referred to as nonobservable errors.

(GENERAL LINEAR MODEL)

NONPARAMETRIC CLUSTERING TECHNIQUES

Write N for the number of clusters in a set of multivariate observations; given N, numerous clustering techniques estimate the cluster membership of each observation. Most of these techniques lack a statistical basis, making determination of N problematical.

One statistical formalization of the clustering problem assumes the data come from a mixture* of normal distributions. This assumption allows determination of N using a likelihood* or other statistical criterion since, under the assumption, N equals the number of component distributions. Several current clustering algorithms use this approach; see, e.g., Lennington and Rassbach [3]. The normality assumption is frequently violated, making interpretation of the resulting clusters difficult.

A generalization of the normal mixture model supposes the observations arise from a mixture of unspecified distributions [2, p. 205]. Based on this supposition, the clustering problem reduces to obtaining a nonparametric estimate of the underlying density function.

One nonparametric density estimate uses the equal cell histogram. Given a threshold, the clusters are the connected regions above the threshold level. No theoretically defined threshold currently exists, although some authors suggest the expected value of the density given a uniform distribution over the range of the observations. Goldberg and Shlien [1] apply this technique to obtain a preliminary clustering of LANDSAT data. Each observation consists of four measurements in the range from 0 to 127; the number of cells equals the number of possible combinations, 64^4, and the threshold value is the average number of observations per nonempty cell. All contiguous cells with density above the threshold are connected, and then all cells with density below the threshold are joined to the nearest connected set; N is the number of connected sets.

An improved estimate can be obtained by allowing the data to determine the cells, as in ref. 6, in which Wong partitions the data space into k regions, for k between N and the number of observations, obtaining a density estimate inversely proportional to the volume of the regions. The k regions are the partition of the data space minimizing the within region sum of squares of the observations and correspond to the clusters found by the k-means clustering algorithm. This set of estimates is then used to assign the observations in each region to the appropriate cluster.

An alternative approach to nonparametric density estimation and hence to the problem of estimating N and cluster assignment, uses