Joachim Krauth
University of Düsseldorf, Psychological Institute, FRG

Techniques of Classification in Psychology I: Factor Analysis, Facet Analysis, Multidimensional Scaling, Latent Structure Analysis

Krauth, J.: Techniques of classification in psychology I: Factor analysis, facet analysis, multidimensional scaling, latent structure analysis.
In: Int. Classif. 8 (1981) No. 2, p.126–132, 37 refs.

Classification problems are quite common in psychology, and many classification procedures were developed and applied in this specific branch of science. Some of these methods or groups of methods, respectively, are introduced here. The mostly used technique in psychology is factor analysis. We describe the underlying model of factor analysis, its indeterminacy problems, modifications and extensions, and give some hints for application. Next facet analysis is discussed, which is based on ideas from variance and factor analysis. Another procedure is multidimensional scaling (MDS), which can be used either as a preliminary stage for other classification procedures or as a classification method itself. A distinction is made between metric and nonmetric MDS-procedures, and between methods for proximity data, dominance data, profile data, and conjoint measurement data. Finally the models and applications of latent structure analysis are discussed. (Author)

1. Aims

In the following we try to give a review of those methods, which are used in psychology for the purpose of classification. It is self-evident that we cannot give a complete description of these techniques, since even a bibliography on this topic would comprise thousands of titles. Therefore we will give only a short introduction into each method.

One might ask, whether it makes sense to consider in particular classification techniques in psychology, since e.g. cluster analysis and discriminant analysis are well-known methods of numerical classification, which are used not only in psychology but in many other branches of science as well. But these methods were adopted for the specific needs of psychology and their behavior in the psychological context was investigated by many authors. Other methods, e.g. factor analysis, multidimensional scaling, latent structure analysis, facet theory, typal analysis etc., were developed primarily for use in psychology, and only in a second step their usefulness for other fields was discovered. Because of this we felt justified to restrict ourselves to classification methods used in psychology.

2. Concepts of classification

The term "classification" has been used in literature in more than one meaning and it seems necessary to clarify this matter before describing the different methods. If we have given a sample of n subjects and for each subject the values of p variables, we might suspect that the n subjects form certain distinct groups in the space of the p variables. If we assume the existence of such a structure, the groups are called *clusters* and methods of finding such groups are called *cluster analyses.* In (16) the term cluster analysis is reserved for the case, where every variable is recorded for each individual. If the division into subclasses is based on criteria, which may vary from one class to another, the term *classification analysis* is used.

In (17) the term classification analysis is used in both of the meanings above, and the term cluster analysis is used, if one tries to identify groups of variables instead of groups of subjects. Since in this latter case the aim is for the most part to relate the (observed) manifest variables to a smaller number of (not observed) latent variables, it might be better to use instead the term analysis of latent structure.

Up to now we assumed a hidden structure in the data, whichpermits a division of the data into disjoint clusters. But sometimes this assumption is unrealistic. It might be sensible to differentiate between people, which are more intelligent and others, which are less intelligent. But by all we know, "intelligence" corresponds to a continuous latent scale. The only way to distinguish between more and less intelligent people is to define a cut-off point, which is defined in such a way that people with an "intelligence score" below this point are called less intelligent, and people with a score above this point are called more intelligent. This kind of classification, where an artificial structure is imposed on the data, is called *dissection* in (17). Of course, dissection is neither restricted to only two groups nor to only one dimension.

In this context we should also discuss the concept of a *type.* Methods by which such types can be identified are called *typal analyses.* One definition of a type describes it as a subset of all those subjects, which form a group due to their similarity. Many authors identify therefore types with clusters. Other authors define types as classes, which are not necessarily disjoint. But there are many more definitions of types used in literature as well. E.g., in (4) no fewer than 45 semantic usages of type are given. But this list is by no means exhaustive. Therefore we believe that it is necessary to discuss typal analyses beside cluster analyses.

Another group of problems assumes that the existence of two or more populations is given and for each population a sample of subjects *(training sample).* The problem is to find a rule, which enables us to allot some new subject to the correct population. This is termed a problem of *allocation* or *discrimination.* If the *allocation rules* are given by functions based on weights for the different variables, the term *discriminant analysis* is used. There are several modifications of the discrimination problem. For example, the allocation of the training samples to the populations may not be known (cf. (5)). We may also consider the case, where the populations are known exactly, and the problem is to allot a new subject to one of the populations. This problem is called *pattern recognition.*

In (17) three reasons for discrimination are given. The first one is *lost information,* i.e. for some reason or other we know only the values of some variables for a subject but no longer to which class it belongs. A second reason

is *unattainable information*. This is the case in diagnostics, where we must diagnose a disease from certain equivocal symptoms. A last reason is *prediction*. Here we use observations at a previous point of time to discriminate between certain types of behavior in the future.

In the following we describe some of the methods of classification, which are used in psychology. We did not try to organize the representation of these methods according to the classification concepts given above, since this leads to some difficulties caused by the fact that some methods can be used for different purposes. The procedures are therefore presented one after the other with reference to other methods if necessary.

3. Factor analysis

3.1 Fundamental principles

We start from a sample of n subjects. For each subject we measure p variables $X_1, \ldots, X_p$, the so-called *manifest* variables, e.g. p test scores. These form a p-dimensional column vector X. The hypothesis is that these p variables can be explained in principle by $q < p$ latent variables $Y_1, \ldots, Y_q$, the so-called *common factor scores*, which form the q-dimensional column vector Y. In this way the originally p-dimensional space is reduced to a q-dimensional space. One hopes that this reduction can be done in such a way that each latent variable can be explained by a group of certain manifest variables, and that these groups are more or less disjoint. In this way a classification of the manifest variables is derived. The relation between latent and manifest variables is given by linear equations

$$X_i = \mu_i + (\lambda_{i1} Y_1 + \lambda_{i2} Y_2 + \ldots + \lambda_{iq} Y_q) + E_i, \quad i = 1, \ldots, p.$$

By E_i we denote the unique part of X_i, which cannot be explained by the common factors. Another interpretation is that E_i is an error term. But this last interpretation becomes dubious, if there exists a latent factor score, which is related to only one of the manifest variables we have considered. In other words, if a certain manifest variable cannot be classified, i.e. related to a common factor score, it should be considered as a representative of a separate class.

By assuming

$$E[X_i] = \mu_i, \quad i = 1, \ldots, p$$

for the expectation of X_i we can without loss of generality take

$$E[Y_j] = E[E_i] = 0, \quad i = 1, \ldots, p, \quad j = 1, \ldots, q.$$

The coefficients $\lambda_{i1}, \ldots, \lambda_{iq}$ are constant parameters, which form the (pxq)-matrix Λ of *factor loadings*. Denoting by μ the column vector of $\mu_1, \ldots, \mu_p$ and by E the column vector of $E_1, \ldots, E_p$, we can combine the p linear equations into one matrix equation

$$X = \mu + \Lambda Y + E.$$

We assume that the errors are uncorrelated, i.e.

$$\mathrm{Cov}[E_i, Y_j] = 0, \quad i \neq j, \quad i = 1, \ldots, p, \quad j = 1, \ldots p,$$

and that the common factors and the errors are uncorrelated, i.e.

$$\mathrm{Cov}[E_i, E_j] = 0, \quad i = I, \ldots, p, \quad j = I, \ldots, q.$$

Then the covariance matrix Σ_X of the manifest variables, which contains in the main diagonal the variances of $X_1, \ldots, X_p$ and in the other cells the covariances, is given by

$$\Sigma_X = \Lambda \Sigma_Y \Lambda' + \Sigma_E$$

with the covariance matrix Σ_Y of the common factor scores and the covariance matrix Σ_E of the errors. By Λ' the transpose of Λ is denoted. The matrix Σ_E is a diagonal matrix of the variances of $E_1, \ldots, E_p$, since the covariances are zero. If we assume that Σ_Y equals the identity matrix I, this means that we have standardized common factor scores with variances equal to one, which are uncorrelated or *orthogonal*, respectively. This leads to

$$\Sigma_X = \Lambda \Lambda' + \Sigma_E.$$

The problem in factor analysis is to estimate the matrix Σ_X and the vector μ by means of the p-dimensional measurement vectors X for the n subjects, and then to factorize Σ_X according to the equation above. By this procedure we get estimates of the factor loadings Λ and can express each manifest variable as a linear combination of latent variables and vice versa. If we have found a solution Λ and multiply it by any non-singular orthonormal (qxq)-matrix, we get another solution. By means of such a *rotation* we can try to find a solution with a simple structure, i.e. with factor loadings, which are either near one or near zero, respectively. If Λ is a matrix of ones and zeros, we have a classification of the p manifest variables into q disjoint classes. It is obvious that we can achieve this aim even better by assuming a matrix Σ_Y different from I, i.e., if we allow for correlated or *oblique*, respectively, common factor scores.

3.2 Indeterminacy problems

In (7) and elsewhere it has been pointed out that one cannot expect unique solutions from factor analysis, i.e. the same factor analytic model can result in many totally different classifications of the variables. The first indeterminacy concerns the number q of common factors. Values of q, which are either larger or smaller, respectively, than the unknown number of "real" common factors, will lead to incorrect interpretations of the data. A second indeterminacy concerns the matrix Σ_E of error variances for a given number q of common factors. A third indeterminacy concerns the matrix Λ of factor loadings for given Σ_E and q. By using orthogonal and oblique rotation we can get infinitely many matrices of factor loadings leading to different classifications of the variables. Finally even for given parameters q, Σ_E and Λ there may exist different common factor scores Y.

3.3 Some precautions

In (9) several hints are given, which can in certain circumstances prevent wrong interpretations of factor analytical results. First it is proposed that there should be at least three manifest variables for each common factor, since otherwise we cannot expect a "good" rotational solution. Second the manifest variables should not be factorially complex, i.e. each manifest variable should belong to a single common factor. Third it is better to

choose the number q of common factors too great than too small. If we have extracted more common factors than can be actually assumed, we can delete them from interpretation. Fourth the manifest variables used in factor analysis should be variables that are linearly independent, since otherwise the correlation coefficients are spurious. E.g. by using sums and differences of the manifest variables as additional variables we can produce artifacts.

As a fifth point the population on which the analysis is based, should be homogeneous, since otherwise we cannot expect good estimates of Σ_X and μ. Since the estimation of Σ_X is the basis of the whole analysis, we must try to get good estimates of the intercorrelations of the manifest variables. Conditions for this are beside others a large number n of subjects, very reliable measurements and approximately linear relations between these variables.

In (12) a four-stage approach for factor analysis was proposed in order to get reliable classifications. In the first stage *exploratory* factor analysis is used with the aim of deriving the probable number q of common factors. In a second stage the factors or classes, respectively, are given names that are based on theoretical arguments, and which are chosen in such a way that many researchers agree with respect to the manifest variables, which should have high factor loadings on the factor with a given name. In a third stage simple cross-validation on many data sets is done based on *confirmatory* factor analysis. This means that we test the goodness of fit of the factor structure, which we found in the first stage. The goodness of fit tests are based on the results for samples of subjects different from the sample used in the first stage and make the additional assumption that the vector X of manifest variables is multivariate normal with mean vector μ and covariance matrix Σ_X (cf. (I)). In the fourth stage a double cross-validation on samples different from those used in the first and third stage is performed. The goodness of fit of these data with respect to the factor structure found in stage one and the average parameter estimates found for the samples in stage three is then tested.

3.4 Modifications and extensions

The majority of factor analyses is done in the way we described in 3.1. The starting point is a data matrix of n rows corresponding to the subjects and p columns corresponding to the variables. In the so-called R-technique (cf. (3)), which we have considered up to now, the columns of the matrix are correlated and the correlation matrix is factorized. This leads to classes of variables. In the Q-technique (cf. (30)) the rows of the matrix are correlated and the corresponding correlation matrix is factorized. This yields classes of subjects, i.e. we get a classification of the subjects.

In contrast to the R- and Q-techniques, which involve a population of subjects, the P- and O-techniques are applied to a single subject. P-technique starts by measuring a set of variables on one subject and repeats these measurements on a sufficient number of occasions. Then the correlation matrix of the variables is factorized yielding classes of variables, which tend to change with time in the same way. By inverting P-technique in the same way, as we inverted R-technique to get Q-technique, we derive O-technique. This means that we correlate points of time or occasions instead of variables and get classes of occasions.

In S- and T-technique we consider only one variable but several subjects and several occasions. The correlation matrix of occasions leads to T-technique yielding classes of occasions, while S-technique, which is the inversion of T-technique, is based on the correlations of subjects and leads to classes of subjects.

All the modifications of factor analysis considered above are based on a two-dimensional data matrix, from which a covariance matrix is derived, which is then factorized. In (34) a three-mode factor analysis was considered, which can be used, if we have a three-dimensional data space, which results, e.g., if subjects, variables, and occasions are considered. Generalizations to more than three dimensions are obvious. An extension to four-mode matrices is given in (22). Of particular interest is the use of three-mode factor analysis for factors of change (cf. (33)). Nowadays multivariate change structures are also analysed by means of the so-called LISREL model of Jöreskog (cf. (21)). This modell as well as the so-called AVOCS-model are special cases of *general covariance structure analysis,* which is reviewed in (14). Common factor analysis is a special case of such models as well, as has been discussed in (2) in detail.

One of the basic assumptions of common factor analysis is the assumption of linear relations between manifest and latent variables, though it is obviously seldom realistic to expect linear relations. Even monotonic relations cannot be always assumed. In (25) an approach to this problem based on orthonormal polynomials is proposed.

3.4 Relationship to other procedures

The common origins of factor analysis and latent structure analysis are discussed in (8), while in (35) relations between three-mode factor analysis and multidimensional scaling are given. Empirical comparisons of factor analysis with cluster analysis are given in (27), while in (18) empirical comparisons with *order analysis* are discussed.

In (36) the method of covariance selection, which was developed in (6), is proposed as an alternative for factor analysis. Just as in common factor analysis the starting point is a data matrix for n subjects and p variables. It is observed that the elements of the inverse of the covariance matrix of the p variables, which are called *concentrations,* are multiples of the corresponding partial correlation coefficients. This means that a concentration of zero is equivalent with a zero partial correlation coefficient of the corresponding pair of variables. By means of an iterative process based on the assumption of a multivariate normal distribution that covariance selection model is sought, i.e. that concentration matrix with a given pattern of zeros, which fits the data in a best way. This yields a classification of the p variables into subgroups of variables. Variables in different subgroups are either identical or conditionally independent.

4. Facet analysis

In (10) the principles of facet theory are discussed. First hypotheses about the general properties of the subjects

are formulated, the so-called *facets*. For example, one can consider the two facets I = intellectual abilities with five elements and C = content with three elements. By the Cartesian product IC is meant the set of ordered pairs ic, where i is an element of I and c an element of C. In our example we have 5 x 3 = 15 of these *structuples* (= element combinations). For each structuple ic an item is constructed and for each item a score Y_{pic} is observed for each subject p of a population. Following (26) it is assumed that Y_{pic} is a linear function of a set of latent variables given by

$$Y_{pic} = \mu_{ic} + \sigma_{ic} S_p + \rho_{ic} R_{pi} + \gamma_{ic} G_{pc} + \rho\gamma_{ic} RG_{pic} + E_{pic}.$$

In this equation μ_{ic} is the mean score of the item in the population of subjects. The parameters $\sigma_{ic}, \rho_{ic}, \gamma_{ic}$, and $\rho\gamma_{ic}$ are loadings specific to the structuple ct. The other terms denote latent random variables. S_p is the score on a general latent variable, R_{pi} and G_{pc} are the scores on latent variables specific to the element i of the facet I and the element c of the facet C, and RG_{pic} is the score on a latent variable specific to the combination of i and c. By E_{pic} an error score is denoted. Since RG_{pic} and E_{pic} cannot be separated without replications, the last two terms can be combined into one term E_{pic}. In (26) other restrictions of the parameters are considered as well. Here and in (15) covariance structure analysis is used for performing facet analyses. In (29) the so-called *smallest space analysis* of Guttman (11) is used for testing the structural hypothesis. In particular the items are classified into subsets and order relations among subsets of items are established.

5. Multidimensional scaling

5.1 Definition

Following (31) a multidimensional scale is defined in the following way. We assume that we have an *empirical relational system,* i.e. a set of empirical objects on which a set of m empirical relations is defined. Further we assume the existence of an r-dimensional *numerical vector relational system,* i.e. a set of r-dimensional vectors with components that are real numbers and a set of m relations on the vectors. A *multidimensional scale* is then defined as an r-dimensional homomorphism that maps the empirical relational system onto a subsystem of the numerical relational system. An r-dimensional homomorphism assigns to each empirical object a numerical vector in such a way that whenever an empirical relation holds, the corresponding numerical relation holds as well, and that whenever an empirical relation does not hold, the corresponding numerical relation does not hold.

5.2 Relation to classification

Multidimensional scaling or MDS can be used in classification in different ways. On the one hand MDS can be necessary for the performance of cluster or typal analyses, which in most cases are based on r-dimensional data vectors for each object. In particular, if the data are not given in the form of real-valued vectors, but e.g. in form of preference data, an MDS must precede any cluster analysis. On the other hand MDS can be used directly as a method of classification. Since the dimension r of the space into which the empirical objects are mapped is not known, this number, which corresponds to the number of factors in factor analysis, must be determined. The dimensions, which are identified in MDS, correspond to the factors in factor analysis. If the objects are subjects, one gets results similar to Q-factor analysis, while in case of objects, which are variables, results similar to R-factor analysis are derived. The scores of the objects on the latent dimensions correspond to factor loadings. Objects with high scores with respect to a certain dimension form a class.

5.3 Types of data

It is possible to divide the different MDS-procedures into four classes according to the data. These may be proximity data, dominance data, profile data or conjoint measurement data. We follow here closely the classification given in (28).

5.3.1 Proximity data

In most cases an (nxn)-matrix is given for n objects. Each cell of this matrix contains a measure of proximity between two objects, e.g. a measure of similarity or a measure of dissimilarity. This measure may be given on a numerical or on a merely ordinal scale. Sometimes the cell entries in the main diagonal describing the proximity of identical objects are missing and for symmetrical measures of proximity even the entries of the triangular above-diagonal half can be missing. It is assumed that the proximity data are monotonically related to distances in some underlying latent space, where the monotonic function is decreasing in the case of measures of similarity and increasing for measures of dissimilarity.

Sometimes only the measures of proximity between two different sets of n or m objects, respectively, are given. This yields an (nxm)-data matrix. This matrix can be regarded as a corner submatrix of the complete (n+m) x (n+m) proximity matrix. It is assumed that the two sets of n and m points are embedded in the same space in such a way that for any object in one set the given measures of proximity between that object and all objects in the other set are monotonically related to the corresponding distances of the point corresponding to the one object to all points in the other set.

It is possible to consider even more incomplete proximity matrices by considering only comparisons of certain pairs of objects. An example is the case, where the pairs under consideration are linked by pairs with one object in common.

5.3.2 Dominance data

For n objects an (nxn)-matrix is given. Each cell contains a measure of the extent to which the row object dominates the column object. This measure can take the purely dichotomous form, if we record only whether one object dominates the other one. It is assumed that each object is represented by a score on a unidimensional scale in such a way that, if object i dominates object j, then object i has a higher score than object j.

In the multidimensional case we consider m of these (nxn)-dominance matrices. Each matrix describes the manifest dominance structure of n objects under differ-

ent conditions, e.g. for m different subjects. It is assumed that each object can be represented by a point in a space of two or more dimensions. The conditions are represented by directions in this space. If an object i dominates an object j in a particular matrix, then the point for object i falls beyond the point for object j in the corresponding direction.

5.3.3 Profile data

Here n objects and m variables are considered yielding an (nxm)-matrix. The entries give the measured values of the objects with respect to the variables. The entries of a row are interpreted as a *profile* of the object in question. By using a measure of profile similarity or dissimilarity any matrix of profile data can be transformed into a proximity or dominance matrix. It is assumed that the objects can be represented as n points in a space, and that by some rule the profile of an object determines the position of the corresponding point in the space.

5.3.4 Conjoint measurement data

An (nxm)-matrix is given, where the rows correspond to n levels of one variable and the m columns to m levels of another variable. An entry of this matrix describes the magnitude of the effect, which results for the combination of the corresponding row and column level. It is assumed that the levels of one variable can be represented as points on one unidimensional scale and the levels of the other variable as points on another unidimensional scale. This is done in such a way that each entry is a simple function of the scaled values associated with the levels in question. While matrices of proximity, dominance or profile data are always two-way matrices, matrices of conjoint measurement data can immediately by extended to more than two variables, which jointly contribute to an effect.

5.4 A selection of MDS-procedures

A distinction is made between metric and nonmetric MDS-procedures. A *nonmetric* procedure yields results that are invariant under monotonic transformations of the data. This means that nonmetric procedures are appropriate for merely ordinal data. *Metric* procedures assume interval scaled data, i.e. we have only invariance of the results with respect to linear transformations.

5.4.1 Methods for proximity data

In classical metric MDS (cf. (32)) the proximity data are related to the distances of points by means of a function of a specified form. Indirect methods assume that the proximity data arise from comparisons of subjective magnitudes, which are normally distributed. This yields estimates of the distances. Direct methods assume that the proximity data are directly related to the distances by a specified monotonic function. The first step is in both cases the estimation of the distances by means of the proximity data. In a second step these estimates are used to determine dimensionality, and in a third step the coordinates for the points in space are derived.

In nonmetric MDS proximity data are related to distances by a monotonic function. Using an initial configuration of points the coordinates are iteratively adjusted in such a way that a measure of goodness of fit is minimized.

In (19) the so-called *stress* is defined as a measure of goodness of fit in the following way. First the proximity measures δ, which are assumed to measure dissimilarity of the objects, are ranked in strictly ascending order. It is assumed that the dimension of the space, in which the objects are to be represented, is given by t. Now suppose that n objects are represented by n points in the t-dimensional space, i.e. by a so-called *configuration*. For each pair of points a distance d can be calculated. Now for each distance d a number $\hat{d}$ is sought, such that the $\hat{d}$'s have the same ranking as the dissimilarities δ. This is done in such a way that the square root of

$$\Sigma(d-\hat{d})^2/\Sigma d^2,$$

where the summation is over all pairs of objects, is minimized. The result is the so-called *stress* for a fixed configuration. The next step is to find that configuration with minimal stress. In further steps that dimension t and that distance measure is looked for, which yield the smallest stress values. In this way it is possible to identify the structure of the space, which fits the data best in the sense of stress.

5.4.2 Methods for dominance data

In the metric case each dominance matrix corresponds to an axis in a latent space and the projections of the points on this axis are related to the entries of the dominance matrix in question. In the nonmetric case only the order of the projections is considered. Methods for determining the number of dimensions and the rank order of the projections on the axes are called *unfolding*-techniques (cf. (13), (37)).

5.4.3 Methods for profile data

In the metric case factor analysis can be used. In the nonmetric case one can use nonmetric factor analysis (20). This latter method assumes that each of the m values in a profile is a monotonic function of the coordinates of a latent space.

5.4.4 Methods for conjoint measurement data

It is assumed that each entry of the data matrix can be represented as a simple function of its latent row and column values. The latent values are estimated by an iterative adjustment of values for each of the rows and columns, such that a measure of overall departure from the model is minimized.

6. Latent structure analysis

6.1 Model

The *general latent structure model* can be formulated in the following form (8). We assume n manifest (observed) variables $X_1, \ldots, X_n$ and $m < n$ latent (not observed) variables $Y_1, \ldots, Y_m$, and a conditional probability distribution $f(x_1, \ldots, x_n | y_1, \ldots, y_m)$ of the X's for given Y's. The function f is a probability density for continuous X's and a set of probabilities for discrete X's. In a similar way we denote the marginal distribution of the Y's by $g(y_1, \ldots, y_m)$. Then we get the marginal distribution of the X's by

$$h(x_1, \ldots, x_n) = \int f(x_1, \ldots, x_n | y_1, \ldots, y_m) \, g(y_1, \ldots, y_m) \, dy_1 \ldots dy_m .$$

For a subject with the values $x_{01}, \ldots, x_{0n}$ for the manifest variables we get by means of Bayes' formula the distribution for its latent values by

$$k(y_1, \ldots, y_m | x_{01}, \ldots, x_{0n}) = f(x_{01}, \ldots, x_{0n} | y_1, \ldots, y_m) \, g(y_1, \ldots, y_m) \, h(x_{01}, \ldots, x_{0n}) .$$

Using the maximum likelihood approach we can assign a subject with the values $(x_{01}, \ldots, x_{0n})$ of the manifest variables to that point $(y_{01}, \ldots, y_{0m})$ in the latent space for which the function k is maximum.

In the first equation above at most the distribution h of the manifest variables is known, while f and g are unknown and cannot be deduced from h. Therefore we need additional assumptions about the form of f and g. A general assumption is that the *axiom of local independence* holds, i.e. for given values of the latent variables the manifest variables are independent. This yields

$$f(x_1, \ldots, x_n | y_1, \ldots, y_m) = f_1(x_1 | y_1, \ldots, y_m) \ldots f_n(x_n | y_1, \ldots, y_m) .$$

The functions on the right side of the equation are called *trace functions* or *trace lines,* respectively, and it is assumed that a least their form is known though there may be unknown parameters in these functions.

6.2 General procedure

In (23) a scheme of 9 steps is discussed, in which way a latent structure analysis can be performed. The first step concerns the selection and specification of the model. Here assumptions about the form of the trace lines and of the function g, i.e. about the distribution of the latent variables are made. In a second step the so-called *accounting equations,* which relate the distribution of the manifest variables with that of the latent variables, are formulated for the specific model. In a third step the conditions of *reducibility* are studied. These result from the fact that in the more complex models we have more accounting equations than latent parameters. The system of accounting equations can only be solved exactly, if the additional equations hold for the data as well. In a fourth step the *identifiability* of the latent parameters must be checked. The accounting equations are only in part independent, since many manifest parameters can be calculated from other manifest parameters. Therefore one cannot always be sure that it is possible to identify all latent parameters by means of the observed data.

The fifth step concerns the *identification* of the model, i.e. the accounting equations are solved for the latent parameters. In the sixth step the model is *fitted* to the data. In this step, which can be used for concrete data instead of the fifth step, we allow for random fluctuations of the data. In step number seven the goodness of fit is tested, i.e. it is investigated how well the data agree with the fitted model. In the eighth step the *recruitment pattern* is determined. For each *response pattern* of manifest variables there exists a recruitment pattern of the probabilities that a subject with this response pattern is located in a certain region of the latent space. In the last step the *classification* of the subjects is performed. If, e.g., the latent variables are discrete, and if the maximum likelihood approach is used, i.e. if each subject is assigned to that point of the latent space for which the probability is maximum, then we have a classification of the subjects to a set of disjoint latent classes. Another interpretation of this step is that *scale values* are assigned to the subjects.

Sometimes one is also interested in the dual problem of scaling or classifying the manifest variables. In this case scale values are assigned to the manifest variables indicating how much a single manifest variable contributes to the scale values of the subjects or how much a single manifest variable discriminates between subjects located at different points of the latent space, respectively.

6.3 Modifications

By making specific assumptions with respect to the trace lines and the latent distribution g we derive submodels of the general latent structure model (cf. (24)). The best known and mostly used model is the *latent class model.* This assumes only one discrete latent variable Y. The values of Y are called *latent classes.* It is assumed that g corresponds to a multinomial distribution, and that the trace lines are the conditional probabilities that the manifest variables take certain values, if the subject belongs to a certain latent class.

The *latent polynomial model* assumes one continuous latent variable Y, trace lines that are polynomials and in most cases that g corresponds to a beta distribution. In particular, linear and quadratic trace lines were considered. For polynomial trace lines and a discrete distribution g the so-called *located class model* results. By assuming certain exponential functions for the trace lines and a uniform distribution on the interval between zero and one or alternatively the beta distribution for g the *latent content model* is derived. If the trace lines are certain step functions, the *latent distance model* results. If g corresponds to a standard normal distribution, and the trace lines are densities of normal distributions with unknown parameters, we get the *classical test theory model.* For logistic trace lines the probabilistic test theory models of Rasch and Birnbaum result.

In the *latent profile model* a discrete latent variable Y is assumed and continuous manifest variables but no specific assumptions with respect to the trace lines. For each manifest variable the existence of a *latent profile* is assumed, which is defined as the vector of the conditional expectations of that variable for given latent classes.

References:

(1) Anderson, T.W., Rubin, H.: Statistical inference in factor analysis. In: Neyman, J. (Ed.): Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability V. Berkeley: University of California Press. 1956. p. 111–150.

(2) Bentler, P.M.: Multistructure statistical model applied to factor analysis. In: Multivar. Behav. Res. 11 (1976) No. 1, p. 3–25.

(3) Cattell, R.B.: Factor analysis. New York: Harper 1952. 462 p.

(4) Cattell, R.B.: Personality and motivation structure and measurement. Yonkers-on-Hudson, N.Y.: World Book Co. 1957. 948 p.

(5) Das Gupta, S.: Theories and methods in classification: A review. In: Cacoullos, T. (Ed.): Discriminant analysis and applications. New York: Academic Press 1973. p. 77–137.
(6) Dempster, A.P.: Covariance selection: In: Biometrics 28 (1972) No. 1, p. 157–175.
(7) Elffers, H., Bethlehem, J., Gill, R.: Indeterminacy problems and the interpretation of factor analysis results. In: Statist. Neerlandica 32 (1978) No. 4, p. 181–199.
(8) Fielding, A.: Latent structure models. In: O'Muircheartaigh, C.A., Payne, C. (Eds.): Exploring data structures. New York: Wiley 1977. p. 125–157.
(9) Guilford, J.P.: When not to factor analyze. In: Psychol. Bull. 49 (1952) No. 1, p. 26–37.
(10) Guttman, L.: What lies ahead for factor analysis? In: Educ. Psychol. Meas. 18 (1958) No. 3, p. 497–515.
(11) Guttman, L.: A general nonmetric technique for finding the smallest coordinate space for a configuration of points. In: Psychometrika 33 (1968) No. 4, p. 469–506.
(12) Hattie, J.: A four-stage factor analytic approach to studying behavioral domains. In: Appl. Psychol. Meas. 5 (1981) No. 1, p. 77–88.
(13) Heiser, W.J., De Leeuw, J.: Multidimensional mapping of preference data. In: Math. Sci. Humaines 73 (1981) No. 1, p. 36–96.
(14) Jöreskog, K.G.: Analysis of covariance structures. In: Scand. J. Statist. 8 (1981) No. 1, p. 65–92.
(15) Kelderman, H., Mellenbergh, G.J., Elshout, J.J.: Guilford's facet theory of intelligence: An empirical comparison of models. In: Multivar. Behav. Res. 16 (1981) No. 1, p. 37–61.
(16) Kendall, M.G.: The basic problems of cluster analysis. In: Cacoullos, T. (Ed.): Discriminant analysis and applications. New York: Academic Press 1973. p. 179–191.
(17) Kendall, M.G., Stuart, A.: The advanced theory of statistics. Vol. 3. Design and analysis, and time-series. London: Griffin 1966. 552 p.
(18) Krus, D.J., Weiss, D.J.: Empirical comparison of factor and order analysis on prestructured and random data. In: Multivar. Behav. Res. 11 (1976) No. 1, p. 95–104.
(19) Kruskal, J.B.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. In: Psychometrika 29 (1964) No. 1, p. 1–27.
(20) Kruskal, J.B., Shepard, R.N.: A nonmetric variety of linear factor analysis. In: Psychometrika 39 (1974) No. 2, p. 123–157
(21) Labouvie, E.W.: The study of multivariate change structures: A conceptual perspective. In: Multivar. Behav. Res. 16 (1981) No. 1, p. 23–35.
(22) Lastovicka, J.L.: The extension of component analysis to four-mode matrices. In: Psychometrika 46 (1981) No. 1, p. 47–57.
(23) Lazarsfeld, P.F.: Latent structure analysis. In: Koch, S. (Ed.): Psychology: A study of a science. New York: Mc Graw-Hill 1959. p. 476–535.
(24) Lazarsfeld, P.F., Henry, N.W.: Latent structure analysis. Boston: Houghton Mifflin 1968. 294 p.
(25) McDonald, R.P.: A general approach to nonlinear factor analysis. In: Psychometrika 27 (1962) No. 4, p. 397–416.
(26) Mellenbergh, G.J., Kelderman, H., Stijlen, J.G., Zondag, E.: Linear models for the analysis and construction of instruments in a facet design. In: Psychol. Bull. 86 (1979) No. 4, p. 766–776.
(27) Morf, M.E., Miller, C.M., Syrotuik, J.M.: A comparison of cluster analysis and Q-factor analysis. In: J. Clin. Psychol. 32 (1976) No. 1, p. 59–64.
(28) Shepard, R.N.: A taxonomy of some principal types of data and of multidimensional methods for their analysis. In: Shepard, R.N., Romney, A.K., Nerlove, S.B. (Eds.): Multidimensional scaling. Vol. I. New York: Seminar Press 1972. p. 21–47.
(29) Shye, S.: Achievement motive: A faceted definition and structural analysis. In: Multivar. Behav. Res. 13 (1978) No. 3, p. 327–346.
(30) Stephenson, W.: The study of behaviour. Q-technique and its methodology. Chicago: University of Chicago Press 1953. 376 p.
(31) Suppes, P., Zinnes, J.L.: Basic measurement theory. In: Luce, R.D., Bush, R.R., Galanter, E. (Eds.): Handbook of Mathematical Psychology. Vol. I. New York: Wiley 1963. p. 1–76.
(32) Torgerson, W.S.: Multidimensional scaling: I. Theory and method. In: Psychometrika 17 (1952) No. 4, p. 401–419.
(33) Tucker, L.R.: Implications of factor analysis of three-way matrices for measurement of change. In: Harris, C.W. (Ed.): Problems in measuring change. Madison: University of Wisconsin Press 1963. p. 122–155.
(34) Tucker, L.R.: Some mathematical notes on three-mode factor analysis. In: Psychometrika 31 (1966) No. 3, p. 279–311.
(35) Tucker, L.R.: Relations between multidimensional scaling and three-mode factor analysis. In: Psychometrika 37 (1972) No. 1, p. 3–27.
(36) Wermuth, N., Hodapp, V., Weyer, G.: Die Methode der Kovarianzselektion als Alternative zur Faktorenanalyse, dargestellt an Persönlichkeitsmerkmalen. In: Z. exp. angew. Psychol. 23 (1976) No. 2, p. 320–338.
(37) Zinnes, J.L., Griggs, R.A.: Probabilistic, multidimensional unfolding analysis. In: Psychometrika 39 (1974) No. 3, p. 327–350.

Call For Papers COLING 82

The Ninth Conference on Computational Linguistics will be held July 5th–10th, 1982 in Prague, Czechoslovakia. It is sponsored by the International Committee on Computational Linguistics in association with: Linguistic Institute of L. Štúr, Slovak Academy of Science, Bratislava and Faculty of Mathematics and Physics, Charles University, Prague.

Papers are invited for presentation especially from the following domains:

- theories, methods and problems of computational linguistics
- relations of computational linguistics to computer science, mathematics, linguistics, artificial intelligence, etc.
- representation of knowledge and inferencing as they relate to language understanding
- applications of natural language processing:

Authors wishing to present a paper should submit 4 copies of a 3 to 4 page summary, double spaced, by December 1st, 1981, to COLING 82 MFF UK, Linguistics, Malostranské n. 25, 118 00 PRAGUE 1, CZECHOSLOVAKIA.

NATO Advanced Study Institute on Numerical Taxonomy

This Institute will be held from 4 July to 16 July 1982 at Bad Windsheim, Federal Republic of Germany. It will present a review of the entire field of numerical taxonomy ranging from systematic theory through methodology and covering phenetic approaches as well as phylogenetic inference. Morning sessions will present state-of-the-art review lectures by a carefully chosen panel of international experts; the afternoons will present recent research results alternating with panel discussions at the end of each subject segment of the ASI; several evening lectures will round out the program. The intended audience will be largely postdoctoral with a few predoctoral participants.

Full particulars of the program are available from the Institute Director, Dr. Robert R. Sokal, Department of Ecology and Evolution, State University of New York at Stony Brook, Long Island, N.Y. 11794, USA. Tel.: 516-246-6162.