From Complexity to Simplicity:
on the Application of Three Techniques for Multivariate Data Analysis

Sven Junghagen

WP 15/2000

December 2000

MPP Working Paper No.15/2000 ©
December 2000
ISBN: 87-90403-83-5
ISSN: 1396-2817

Department of Management, Politics and Philosophy
Copenhagen Business School
Blaagaardsgade 23B
DK-2200 Copenhagen N
Denmark
Phone: +45 38 15 36 30
Fax: +45 38 15 36 35
E-mail: as.lpf@cbs.dk
http://www.cbs.dk/departments/mpp

From Complexity to Simplicity: on the Application of Three Techniques for Multivariate Data Analysis

Sven Junghagen

PhD, Associate Professor
Copenhagen Business School
Department of Management, Politics and Philosophy

INTRODUCTION

The aim of this paper is to give an overview of three important techniques; factor, cluster and discriminant analysis. I find it necessary and more important to have an understanding of the basic assumptions and the underlying foundations of the methods rather than a thorough mathematical understanding of the algorithms. I will therefore account for a description of these methods based on a geometrical viewpoint and a more conceptual view rather than mathematical. A general description of the three techniques will be followed by a case, showing an application of all three techniques in the same study.

FACTOR ANALYSIS

THE USE AND OBJECTIVES OF FACTOR ANALYSIS

The social science researcher is very often confronted with a data set that contains a large number of variables as well as a large number of cases. This situation with a high level of complexity in raw data almost inevitably leads to complications in the analysis of the material. A suitable way to handle complex data sets is to use a factor analysis approach in order to reduce the complexity in data.

Factor analysis could be used for one or more of the following objectives (Hair et al, 1995; Wold, Esbensen, Geladi, 1987; Befring, 1994):

- Reduction of the data set into a smaller amount of variables with a minimum loss of information, which is the main purpose of all factor analysis techniques.
- Simplification of and overviewing the underlying structures of an empirical material.

- Modelling and identification of relationships and groupings among either variables or cases. The grouping of cases is similar to cluster analysis and is commonly referred to as Q-type factor analysis. In the case of grouping variables the technique is called R-type factor analysis.
- Selection of representative variables and construction of new variables, or factors, for use in a subsequent multivariate statistical analysis.

The objective for factor analysis is, as in all research, defined by the research problem at hand. All possible objectives are not to be considered at the same time, but can be considered in different steps of the analysis. Depending of the objective at hand, there are a few different sub techniques that can be used. The choice of technique is for instance depending on whether the analysis is exploratory or confirmatory or if the analysis should group variables or cases.

THE BASIC ASSUMPTIONS AND GEOMETRIC INTERPRETATION OF FACTOR ANALYSIS

The basic assumptions of factor analysis are more conceptual than statistical in their nature. A geometric interpretation of the method is suitable for comprehensive reasons. The initial solution of the factor analysis is the extraction of principal components or unrotated orthogonal factors. The principal component analysis, PCA, was first described with a geometric interpretation by Pearson (1901), who formulated the analysis as finding ”lines and planes of closest fit to systems of points in space”. This interpretation is based on the assumption that there are no dependent or independent variables, so that the ”best fit line” will be a consequence of the least squares of the perpendiculars from the system of points to the line. This minimisation of the squares of distances leads to that the best line of fit for a system of points goes through the centroid of this system.

Assume an example with three variables, x, y and z. Data collection with data in n cases has been conducted so there is a data set with three variables measured. Despite the less complex structure, with only three variables, it serves its purpose for explanation. The raw data can be plotted in a three dimensional scatterplot which is shown in Figure 1. The raw data scores forms a cluster of data in which the maximum variance in the three dimensions is searched for. This maximum variance is the cause of the first component, which is represented by the line in the data scores in Figure 1.

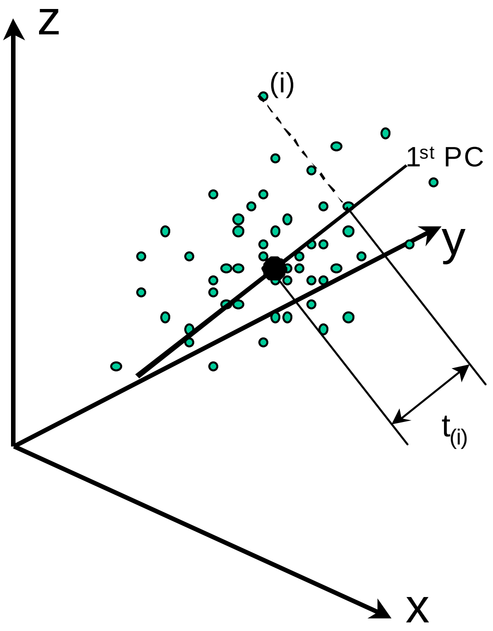

Figure 1 Scatterplot of three variables and a principle component (Source: Wold, Esbensen, Geladi, 1987)

After the extraction of the first component, a second component is extracted. This component is formed orthogonal, i.e. right-angled, to the first component (Pearson, 1901). The component scores of the cases are the orthogonal projection of the scores on the component line, by means of the distance to the origin. As seen in Figure 1, the score $t_{(i)}$ of case (i) is the projection on the first component. In order to achieve a geometric representation of data in the reduced dimensionality, one can make a scatterplot of the component scores. In this score plot the mutual connection between cases are shown and it is possible to identify groupings among cases by a visual examination of the plot. If there are distinctive differences among the groups of cases, there can be reason for separate analyses for subgroups (Hair et al, 1995).

In order to get a picture of the interdependent relationships between the variables in the data set, the factor loadings are computed. The factor loading of a variable can be defined as the Pearson correlation between the variable and the factor (Parasuraman, 1991). If the factor loading of a variable is squared, this indicates what percentage of the variance in the original variable that is explained by the factor (Hair et al, 1995). The relation between the factors and the original variables can be plotted in a loading plot, which will be discussed further in the next section.

The basic factor analysis model, where the object is to represent a variable z_j in terms of underlying factors, could be expressed in the form:

$$z_j = a_{j1}F_1 + a_{j2}F_2 + \dots + a_{jm}F_m + d_jU_j \qquad (j = 1,2, \dots ,n),$$

where each of the n observed variables is described linearly in terms of *m* common factors and a unique factor. The common factors account for the correlations among the variables, while each unique factor accounts for the remaining variance of that variable, including error. The coefficients of the factors, a_j, are the factor loadings of the variable. (Harman, 1967)

FACTOR INTERPRETATION

The initial solution, the principal component analysis, achieves the objective of data reduction, since it is the best linear representation of the total variance of the data set. This solution can yet be inadequate to achieve a meaningful interpretation of the factors. The unrotated solution extract factors in the order of their importance. The first factor often tends to be a mean of all the included variables, and it accounts for the largest amount of variance in data. The following factors will then be based on the residual amount of variance. The effect of a rotation is a redistribution of the variance from the early factors to the later. In this way, a simpler and theoretically more meaningful solution is achieved. (Hair et al, 1995) In Figure 2 of factor rotation are shown.

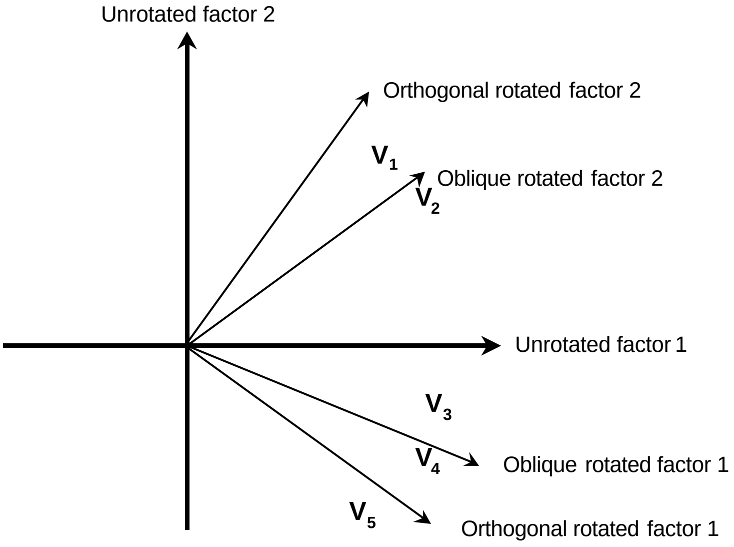

Figure2: Orthogonal factor rotation (Source: Hair et al, 1995)

Among the two rotation approaches, the orthogonal rotation is the more commonly used. The analytical techniques involved in the oblique rotation are not as well developed and are subject to controversy. There is however a problem concerned with the orthogonal rotation. It is not so realistic to assume that the underlying dimensions of an empirical material are uncorrelated.

According to the example, the oblique rotation gives a factor solution that is more correlated with original variables than the orthogonal rotation. On the basis of the loading plot, or the factor-loading matrix, which contains the same information in table form, the factors can be given theoretical meaning. When determining what the factor scores of a rotated factor solution is a measurement of, one must look at what is measured by such variables. These variables allow a definition (or naming) of the factor. What the variables measures, in combination with other variables, is what the factor measures. (Lindeman, Merenda, Gold, 1980)

QUALITY AND VALIDITY MEASURES FOR FACTOR MODEL EVALUATION

The first concern of quality control is the input in the model, i.e. the raw data measured. There are disagreements in the literature on the requirements of the raw data. There are statements that factor analysis requires data that are at least interval-scaled (e.g. Parasuraman, 1991). On the other hand some authors claim that factor analysis can be applied to any kind of data, even discontinuous and ordinal-scaled, as long as the analysis is used for descriptive purposes (Joliffe, 1986). There is obviously a variety of conceptions regarding the use of different scales in factor analysis, but it seems that a healthy scepticism and an awareness of the limitations of low-order-scaled variables is in order.

A number of measures are used in order to evaluate variables and factors included in the model. The first evaluation is concerned with the question weather or not the included variables can result in a relevant factor model. This evaluation is made before the factor extractions, in a correlation matrix where the observed correlation coefficients are compared with the partial correlation coefficients. The Kaiser-Meyer-Olkin (KMO) measure indicates if the correlations between pairs of variables can be explained by the other variables (Norusis, 1994). The KMO measure gives an indication if it is appropriate to proceed with a factor analysis. The KMO measure gives an indication of the overall reliability of the variables.

When it comes to an individual evaluation of variables included in the factor model, there are concepts like communality, uniqueness and error included determining the reliability of a variable. The communality of an original variable is defined as the proportion of variance in it, accounted for by all the extracted factors (Parasuraman, 1991). The uniqueness is the specific variance of the original variable, not explained by the factor. The resulting variance is based on error and can therefore be called the unreliability of the variable. The complement of this is the reliability and is a measure of the systematic component of the variable. In other words, the communality is less than or equal to the reliability, and equals the reliability when there is no uniqueness. (Harman, 1967, Hair et al, 1995)

The concept of eigenvalue; the total standardised variance accounted for by a factor (Parasuraman, 1991), can be used for a selection of the factors included in the model. A guideline frequently used is to include factors with an eigenvalue over 1. (Hair et al, 1995) An interpretation of this guideline is that one should not include a factor that contains less information than an original variable, since the standardised variance of the variable is 1.

As in all statistical analysis, the sample size is of importance. A general rule for the sample size is to sample at least five cases for every variable measured. The researcher should always try to maximise the ratio between cases and variables, so that the factors extracted will be general and not sample specific. (Hair et al, 1995)

Much of the earlier discussion of validity and reliability is concerned with the internal validity of the model. If there is a need for external validity, the most direct method for validating the results is to use a confirmatory approach. This is done by a split sample, where one part of the sample is used to construct the model, and the other part is used to confirm it. In confirmatory factor analysis, there are a number of measures used in to establish the external validity, and the overall model fit. Residuals can be used as a criterion for validation. The residuals can be defined as the difference between the observed correlations in the initial correlation matrix and the reproduced correlations in the factor matrix. A factor model with good fit to data has residuals that are close to zero. (Malhotra, 1993, Bollen, 1989)

The likelihood-ratio χ^2-test in confirmatory factor analysis tests if the two matrices, the observed and the predicted, differ considerably. There are some criticisms against the χ^2-test, since it is very sensitive to sample size and will indicate significant results because of a large sample. The GFI; goodness-of-fit index, which is a non-statistical test, is much alike the χ^2-test without the consideration of degrees of freedom. (Hair et al, 1995)

CLUSTER ANALYSIS

THE BASIC PRINCIPLES AND USE OF CLUSTER ANALYSIS

Cluster analysis is a technique in which the objective is to identify groups of cases or, more seldom, groups of variables. The cluster analysis, when referred to as Q analysis, is in a way analogue to one of the variants of factor analysis, which will be discussed in a coming section of this paper. The groups resulting from cluster analysis is based on the characteristics given by the variables included in the data set. It might be obvious, but it is essential to understand that the relationship between cases in the data set is fully dependent upon the cluster variate. Cluster analysis does not offer an estimation of the variate, but uses the variate itself in the comparison of cases.

The ultimate objective of cluster analysis is to partition a set of objects into two or more groups based on the similarity of these objects. The aim is to maximise within-group similarity and to minimise between-group similarity. (Sharma, 1996, Hair et al, 1995) We can assume a simple example, in which we have data on a population based on two variables, e.g. education level and income. A scatter plot of this fictive data set is shown in Figure 3.

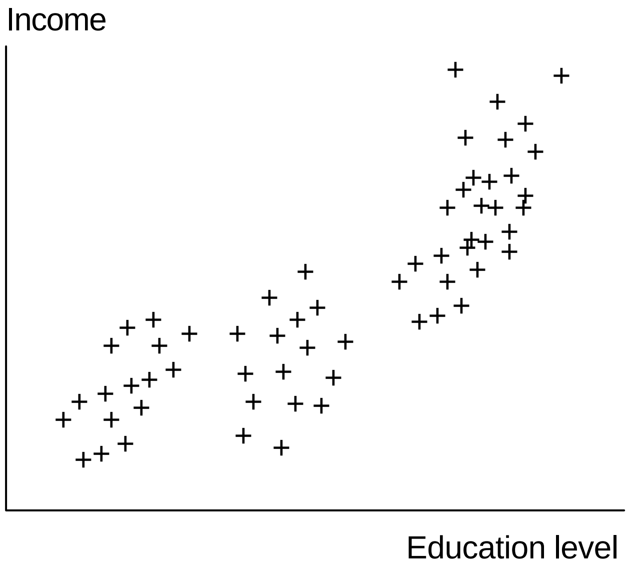

Figure 3: Scatter plot of education level and income; fictive example

In the scatter plot, one can see that there are three distinctive groups of cases, formed by similar patterns in the relationship between education level and income. The similarity within the groups is high, and the dissimilarity between the groups is also high. In this case it is possible to identify three clusters based on the cluster variate.

Similarity is a fundamental concept in cluster analysis. The range of methods involved in cluster analysis is all based on similarity measures. It is possible though, to measure similarity between objects in different ways. Correlational measures, distance measures, and association measures are the three dominant ways to measure similarity. Cluster analysis based on correlational measures is analogue to Q-type factor analysis, which will be discussed further on. Association measures are used when the cluster variate is based on nonmetric variables such as ordinal and nominal scaled variables. Distance measures are the most common used similarity measures and will be discussed below.

There are two general categories of cluster analysis; hierarchical and non-hierarchical cluster analysis. Both categories deal with the fundamental assumption of within-group and between-group similarities discussed above. The hierarchical procedures are starting with a situation were all cases forms their own cluster. The clustering procedure then identifies the two most similar cases, which form the first cluster. In the next stage, the two closest cases are grouped in to a cluster. This procedure continues until all cases are members of one group consistent of all cases in the analysis.

The non-hierarchical cluster analysis procedures do not use the build-up sequence as in hierarchical methods. These analysis procedures identify a number of clusters, where the actual number is predefined. The analyst does not only define the number of clusters wanted, but also a preliminary definition of the cluster characteristics. The cluster seed, a defined case that will be included in a cluster, does this. The choice of cluster seed cases is made upon experience, theoretical relevance, or an exploratory hierarchical cluster analysis.

As in all research, the problem at hand must define the proper method. It is therefore not possible to define the best procedures for cluster analysis; hierarchical or non-hierarchical methods. The hierarchical methods are not relying on definitions made by the researcher as much as the non-hierarchical methods. On the other hand, the hierarchical methods do have the disadvantage of not being able to repair mistakes made in an early stage of the analysis. Two cases that are members of the same cluster could in fact be better off in other clusters, when the other cases have been accounted for. As mentioned above, a combination of both methods could be a good solution.

DISTANCE MEASURES

The most commonly used distance measure is the Euclidean distance. This measure can be shown in an example with two variables, geometrically shown in 2.

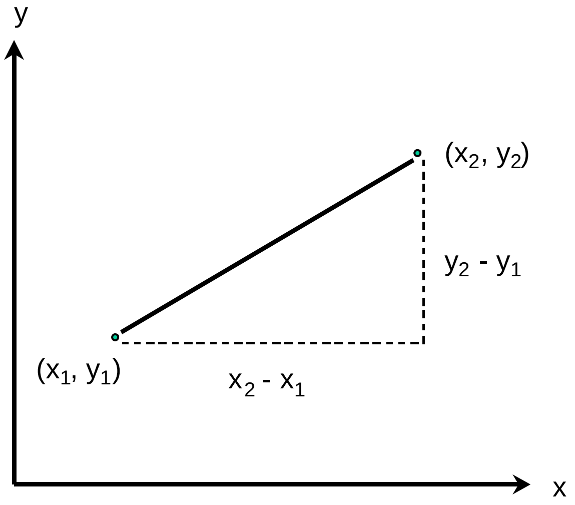

Figure 4. An example of Euclidean distance between two cases measured on two variables X and Y. (Source: Hair et al, 1995)

The distance between the first case, represented by the point (X_1, Y_1) and the second case, represented by the point (X_2, Y_2), is calculated as the length of the hypotenuse of a right-angled triangle.

$$D = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

Where D is the Euclidean distance between the cases. The Euclidean distance can also be squared, which in fact means that the distance value is the sum of the squared differences, without taking the square root. This will speed up the analysis somewhat, and is recommended for the centroid and the Ward's method of clustering.

The city-block approach is another distance measure that is different from the Euclidean distance. The city-block distance indicates the sum of the absolute differences of the variables. An underlying assumption of this measure is that the variables are uncorrelated to each other. This causes problems with this distance measure. Another problem is concerning the fact that it is the absolute difference between the variables that is in focus. Unstandardised variables will lead to that variables with a larger absolute variance will have greater impact on the analysis than variables with less absolute variance.

The weighting problem is solved with a third distance measure, the Mahalanobis distance, which is a form of Euclidean distance. This distance measure incorporates a standardisation of the variables with regards to their standard deviations. Furthermore, this measure also considers eventual intercorrelations among the variables. The Mahalanobis distance is in a way analogue to the R^2 in regression analysis. Since this measure considers weighting of variables and intercorrelation among variables, it can be recommended as a distance measure in cluster analysis.

A COMPARISON TO Q-TYPE FACTOR ANALYSIS

As mentioned above, cluster analysis can sometimes be compared with factor analysis. This is when factor analysis is conducted with the purpose to group cases, which is called Q factor analysis. The main difference between Q factor analysis and hierarchical cluster analysis is that the factor analysis deals with correlation between cases, when hierarchical cluster analysis deals with distances between cases.

This difference between the two methods can in fact lead to different results, which can be shown with an example. Assume three variables, measured among four respondents. The data set and score profiles are shown in Figure 5.

	Variables		
Respondent	1	2	3
A	7	6	7
B	6	7	6
C	4	3	4
D	3	4	3

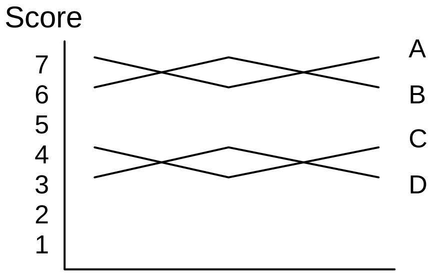

Figure 5. Comparison score profiles for Q-type factor analysis and cluster analysis (Source: Hair et al, 1995)

If a Q-type factor analysis is conducted, that would result in one group consisting of respondents A and C. The other group would contain respondents B and D. That is, because of the fact that the profiles of these respondents correlate. In a cluster analysis, where the grouping is carried out based on the distance between cases, the result would be different. In this case, the groups would consist of A and B versus C and D. This is because the cluster analysis groups the closest pairs. A proper method must therefore be chosen based on the problem at hand and a choice between correlations or distances as similarity definitions.

DISCRIMINANT ANALYSIS

THE BASIC ASSUMPTIONS AND USE OF DISCRIMINANT ANALYSIS

Discriminant analysis is a broad term which refers to a set of statistical activities that are used to explain the variation of a nominal or ordinal scaled dependent variable, by the variation of a set of interval scaled independent variables. Hence, the application and interpretation of discriminant analysis is very much the same as in multiple regression. The main difference is that the dependent variable is nominal or ordinal in discriminant analysis, while it should be at least interval in multiple regression.

Discriminant analysis is mainly used in two problem situations. In the first situation, the interest is focused to the ways in which groups differ. Another situation is when the cluster or class membership of an object is to be determined by other variables. In these applications, the classes are predefined and the number of categories in the dependent variable is equal to the number of predefined classes. (Sharma 1996, Hair et al, 1995) It is very likely that both applications are used in the same research situation.

The first basic assumption is that the data cases are members of two or more exclusive groups, in which the membership is defined by the dependent variable. In these groups the covariance matrices of the independent variables should be approximately equal. Another important assumption is that no independent variable can be a linear combination of another independent variable. The independent variables have to be measured with interval or ratio scales in order to use them in a proper mathematical way. Each group must be drawn from a population, which has a multivariate normal distribution. A final assumption is analogue to most multivariate techniques involving a variate; all relationships are linear. Non-linear relationships are not considered in discriminant analysis unless the variables are transformed in an appropriate way.

A violation of any of these assumptions may lead to problems in estimating the discriminant function. In a situation where the assumption concerning multivariate normality is violated, a logistic regression model can be more suitable. The logistic regression model is designed for a dichotomous dependent variable and is therefore suited for analysis of variables with a binomial distribution. A limitation of logistic regression in comparison with discriminant analysis is that the logistic regression can only handle a two-group analysis. The discriminant analysis can deal with a dependent categorical variable classifying cases into n groups.

A GEOMETRIC INTERPRETATION OF THE MODEL

Consider an example where there is a data set that includes cases, which are members of two distinctive groups, for instance users and nonusers of a product. Except the membership of the cases, there is also data for a number of other variables. The objective is now to use the independent variables to describe the difference between users and nonusers, group A and group B, and to design a model for prediction of future users or nonusers.

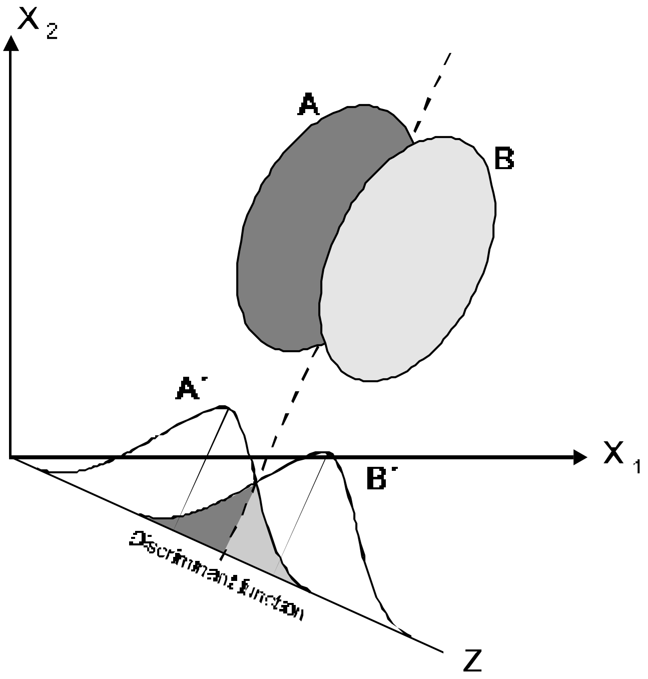

Figure 6. A two-group discriminant analysis (Source: Hair et al, 1995)

In Figure 6 the example is shown with two independent variables, in order to simplify the geometric interpretation. The two ellipses are representing the association in a scatterplot of variable X_1 and X_2. The discriminant function Z is sought where the two independent variables will separate the two groups as much as possible. The shaded area, which represents the overlap between the univariate distributions A´ and B´, will be minimised.

The discriminant function Z derived from the discriminant analysis can be expressed as follows:

$$Z = W_1X_1 + W_2X_2 + ... + W_nX_n$$

Where Z = discriminant score;
W_i = discriminant weight for variable i
X_i = independent variable i

The discriminant function can be used to determine the differences between the groups, for that purpose the discriminant weights are used. The variables with a high relative weight will contribute more to the discriminating power of the function than the variables with a smaller discriminant weight. In the case of classification, the function is used so that a new case will have a calculated discriminant score Z. A critical value of Z determines the cutting score between the two groups.

Another way to interpret the discriminant function is to use the discriminant loadings, which measure the linear correlation between the discriminant function and each of the independent variables. This method can be compared with use of factor loadings in factor analysis. A result of an interpretation of loadings is the possibility to label functions in the same way as labelling factors, so that discriminant functions can be given names with a conceptual and theoretical meaning.

STEPWISE SELECTION OF VARIABLES

The stepwise selection of variables is a way to ensure that the ”accurate” variables are included in the model. There are a couple of ways to perform the stepwise selection. In the forward stepwise procedure, the variable with the greatest univariate discrimination is selected. The procedure then searches for a variable that will provide with the best discrimination together with the first variable. This procedure continues until all variables are included, or until no additional variable will make a significant contribution to the model.

The backward selection is the reverse of the forward procedure. Here, all variables are included initially. The ”worst” variable is identified and cast out. This method can be more risky, when a variable in fact can be cast out and therefore reduce the discriminating power of another variable.

Stepwise procedures produce an optimal set of variables, but not a maximal set. To obtain a maximal set, all combinations of variables have to be taken into consideration.

A situation involving an analysis of more than two groups most often demands more than one discriminant function to discriminate groups. A first visual interpretation of the discriminant analysis can be made in a *territorial map*, which is showing how groups are distributed according to discriminant functions. A fictive example with three groups is shown in Figure 7.

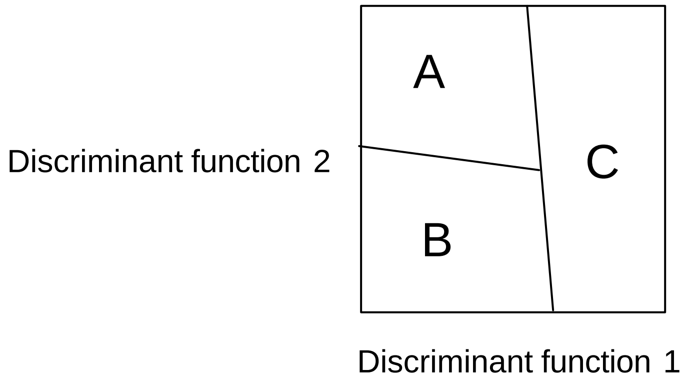

Figure 7. An example of a territorial map with two discriminant functions and three groups

The interpretation of this territorial map is that *function 1* first of all contributes to separate group *C* form group *A* and *B*. *Function 2* will then provide the information about how to separate group *A* from group *B*. If the discriminant functions have been labelled with a conceptual meaning in an interpretation of loadings, the resulting discussion will be fruitful for practical and theoretical use.

CASE: APPLICATION OF THREE TECHNIQUES

CASE SETTING

This case will show an approach for analysis used in order to classify SME's according to strategic dispositions to information technology among managers in these firms (Junghagen 1998; 1999). It is basically an approach taking its departure in a measurement of attitudes, behaviour, perceptions and individual characteristics among managers. An initial assumption is that it will be possible to make a classification based on how SME's view information technology in relation to their way of doing business.

RESEARCH DESIGN AND APPROACH

The analysis is carried out in a number of steps, involving three techniques; factor, cluster and discriminant analysis. The reason for using factor analysis is to reduce a fairly complex set of data into a set of factors, making more conceptual sense than the original manifest variables. These variables are operationalised from theoretical concepts in the research model, shown in Figure 8.

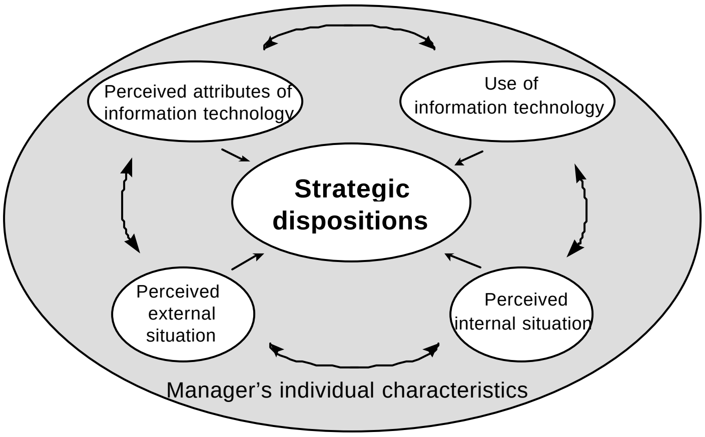
Perceived attributes of information technology
Use of information technology
Strategic dispositions
Perceived external situation
Perceived internal situation
Manager's individual characteristics

Figure 8. Research model in the case

The concepts in this model are based on a thorough theoretical discussion on possible parameters that can affect the strategic dispositions to information technology. Each and one of these concepts are latent variables, i.e. not measurable directly, causing a problem. The solution to this problem is to operationalise the concepts into items to be used in the research instrument, the questionnaire.

This model is not used later in the analysis to form hypotheses to be tested, but as a conceptual framework building the base for the research instrument. When data is collected, the first step is to perform an exploratory, or descriptive, factor analysis to find underlying structures in data.

An example of these factor analyses is shown in Table 1. The information in the table is showing the loadings for each and one of the variables in the three extracted factors. Only values above 0,40 are included, in order to simplify reading of the table. What can be seen here is that the factor model is very good, with a KMO=0,92. The criterion for selection of factors is that these should have an eigenvalue above 1.

The interpretation of this table show that it is possible to label factors according to the theoretical discussion building the research model. Take for example the second factor, *Complexity*, which by definition is a concept accounting for the perceived difficulty to understand and use a certain technology (Rogers, 1995). All items loading in this factor are indicating different aspects of this theoretical concept and the labelling of this factor is therefore rather straightforward.

Table 1. Factor loadings for analysis of perceived attributes of information technology.

	Factors		
Items	Advantage and compatibility	Complexity	Visibility
Information technology helps to improve decision making in my firm.	0.71879		
I cannot understand a thing about information technology.	-0.51766	0.53396	
I have noticed other firms using information technology in a successful manner.	0.50245		0.44789
Information technology means that my employees work faster.	0.71184		
I think I have a good knowledge of the implications of information technology. *	-0.40235	0.41641	-0.41692
Information technology does not lead to work improvement in our firm. *	0.80623		
Information technology is a suitable way of supporting work in our firm.	0.81682		
Information technology improves the quality of our work.	0.85337		
An investment in new information technology would not be hard to adapt to our present practices of work.	0.46239		
Information technology facilitates work for my employees.	0.81149		
Information technology systems are not suited for our way to work. *	0.79652		
An investment in information technology usually just leads to an increase in costs that will never pay off.	0.60222		
A new information technology solution would demand extensive educational efforts in the firm.		0.83213	
It is easy to make information technology work in a firm. *		0.52197	-0.40132
We have good possibilities of testing information technology solutions before purchasing.			0.84621

Kaiser-Meyer-Olkin Measure of Sampling Adequacy = 0.92. 43 percent residuals with absolute value > .05

* These items are loading negatively related to theoretical perceived attributes, so scales are inverted. A positive factor loading should hence be interpreted as a negative correlation between the original item and the factor.

The same kind of analysis is then carried out for all of the fields of interest that are assumed to have an influence on the strategic dispositions to information technology. This is then resulting in a significant lower amount of factors than original variables. These factors, making more conceptual sense than original variables, are now used to group objects into clusters, using cluster analysis.

The number of cluster chosen is determined by an iterative evaluation where discriminant analysis is used for evaluation of identified cluster solutions. For each and one of the cluster solutions, the predictive ability of a discriminant solution is used. A hierarchical method according to Ward's method is used with squared euclidean distance. The final number of clusters chosen is six clusters.

In order to characterise these groups discriminant analysis is used. In this case, a stepwise procedure is used to include variables, with a solution resulting in five discriminant functions. A first evaluation of the discriminant functions can be seen in Table 2. The table is showing that the first two functions are the most influential, with some influence from the third function. Evaluating the eigenvalue, in the same sense as in factor analysis can state this.

Table 2. Discriminant functions, eigenvalues and explained variance.

Function	Eigenvalue	Percent of variance	Cumulative percent
1	2,66	47,0	47,0
2	1,12	19,8	66,8
3	0,97	17,2	84,1
4	0,49	8,6	92,7
5	0,41	7,3	100,0

A visual evaluation of the two first functions can be seen in Figure 9, showing the relation between functions and groups.

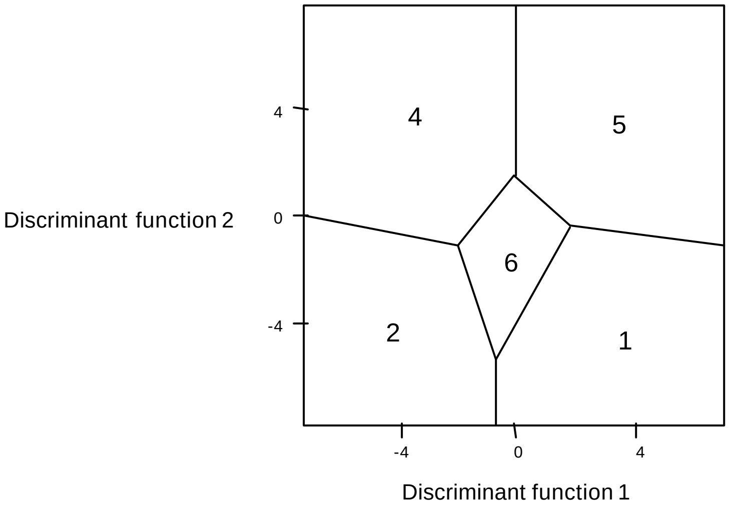

Figure 9. Territorial map for the two first discriminant functions.

Since this cluster solution involves six clusters, it is hard to use this two-function map alone. More discriminant functions are needed, even though there is an indication that the two first functions can discriminate between groups 1, 2, 4 and 5. Group 6 seems to be “in the middle”, but that is an illusion based on the fact that this map is a two-dimensional representation of a multidimensional system of discriminant functions. In the map, the cut is showing a picture where all other functions are at zero. Group 6 is actually lying behind the field of group 1, but in this particular cross-section it can not be seen. In order to get around this problem, we will have a look at the centroids for each and one of the discriminant functions within the different groups. The centroids for the groups in this case can be seen in Table 3.

Table 3. Centroids for discriminant functions within groups.

	Discriminant functions				
Group	1	2	3	4	5
1	1,43	-1,69	-0,61	0,96	-1,39
2	-2,65	-1,55	-1,31	-1,43	-0,03
3	1,44	1,01	-3,33	0,86	1,19
4	-2,00	0,94	0,26	0,46	-0,23
5	1,70	0,93	0,15	-0,74	-0,31
6	0,44	-0,77	0,79	0,17	0,65

In the table of centroids the reason is shown, for why we can not see group 3 in Figure . Even though one can see that group three is in the higher right corner of the map, the centroid for the third function is extremely low in comparison to other groups. That explains why we can not see group 3, it is just not touching this cross-section.

The next step is to identify the conceptual meaning of the discriminant functions. In Table 4, a structure matrix is shown.

Table 4. Structure matrix for discriminant functions. Correlations between functions and variables.

	Discriminant functions				
Variables	1	2	3	4	5
Level of use	0.65*	-0.13	0.05	-0.06	0.27
Advantage and compatibility	0.58*	0.13	0.09	0.24	-0.01
Information intensity	0.47*	0.38	0.01	-0.07	-0.08
Education level of manager	0.36*	-0.02	-0.02	0.20	0.03
Business customers	0.20*	-0.15	-0.11	0.07	0.05
Visibility	0.14*	0.07	-0.04	-0.05	-0.12
Standardisation	0.14	-0.32*	0.30	0.32	-0.28
Wish for change	0.11	0.24*	0.07	-0.04	-0.01
Exports to the EU	0.14	-0.24*	0.19	0.10	0.04
Urban firms	0.04	0.13*	-0.07	-0.12	0.07
Manager's age	-0.05	-0.12*	0.00	0.09	0.00
Risk aversion	-0.10	0.14	0.47*	-0.41	-0.18
Complexity	-0.10	-0.13	0.42*	0.11	-0.06
Centralisation	-0.15	0.09	-0.35*	-0.04	-0.27
Locus of control	0.24	-0.28	-0.28*	-0.14	-0.14
Routine purchase	0.01	-0.10	0.23*	-0.16	-0.06

* Denotes largest absolute correlation between each variable and any discriminant function.

The structure matrix can be compared to a factor-loading table. It is showing the correlation between each and one of the discriminant functions and the independent variables included in the function.

Since we now have six groups, we have five functions and we have the information on what variables that significantly are contributing to the discriminant functions, it is possible to define the groups. In this case, the six groups can be described as follows:

- ***Industrial bureaucracies***, i.e. relatively large firms with a high degree of formalisation and standardisation. Information technology is used to a large extent and both dynamics and uncertainty avoidance are low. Managers are not striving for change and development of the firm. These firms have a high degree of customer complexity and the relation dependency is hence low. The main motive for using information technology seems to be a structural influence, external as well as internal. These structures are also rather stabile.
- ***Sustenance firms***, i.e. relatively small firms, with a high degree of formalisation and standardisation. Information technology is used to a low degree and both dynamics and uncertainty avoidance are low. There are a lot of similarities between this group and the industrial bureaucracies, except for the size of the firm. Motives for using information technology

seems to be the same, namely a structural influence characterised by stability.

- ***Professional service firms***, i.e. small firms with extremely entrepreneurial managers and a high level of information technology use. The firms in this group are mostly within the professional service sector, e.g. marketing consultants, software consultants and other knowledge-intensive services. Information technology is perceived as a natural component of the firm's core competencies and is mainly used to support high dynamics.
- ***Local growth firms***, i.e. small firms that are rapidly growing and dynamic. Information technology is not used to any wider extent and perceived attributes of information technology are characterised by low visibility and a high degree of complexity. Perceived advantage and compatibility is high, however, in spite of a generally low degree of use in the group. Uncertainty avoidance is high, and entrepreneurial tendencies among managers are not especially significant, with an exception of the wish for change. The customer base is mostly local with a high level of customer complexity. It seems that overall strategies as well as the use of information technology are mainly formed in an adaptive mode.
- ***Industrial adhocracies***, i.e. generally relatively large firms, with a sophisticated use of information technology and a high level of information intensity. There is a low level of standardisation and formalisation among these firms. Decision-making is very much decentralised. Managers in this group are characterised by a high level of wish for change, as well as competitive edge. Dynamics as well as uncertainty avoidance are high within these firms. Their high level of technical sophistication combined with a low degree of standardisation and formalisation lead to the label industrial adhocracies.
- ***Subcontractors***, i.e. relatively large firms with a high degree of uncertainty avoidance and a low degree of dynamics, are generally sophisticated users of information technology. A key characteristic is their high dependency on a relation to a single customer. Managers in this group do not show entrepreneurial tendencies to any wider extent. Customer complexity is relatively low and relation dependency is high. The use of information technology seems mainly to be motivated by maintenance of strong ties to a single customer, and therefore the group is labelled subcontractors.

Hence, the analysis is completed. A classification is made without an a priori definition of groups, but based on the approached shown in Figure 10.

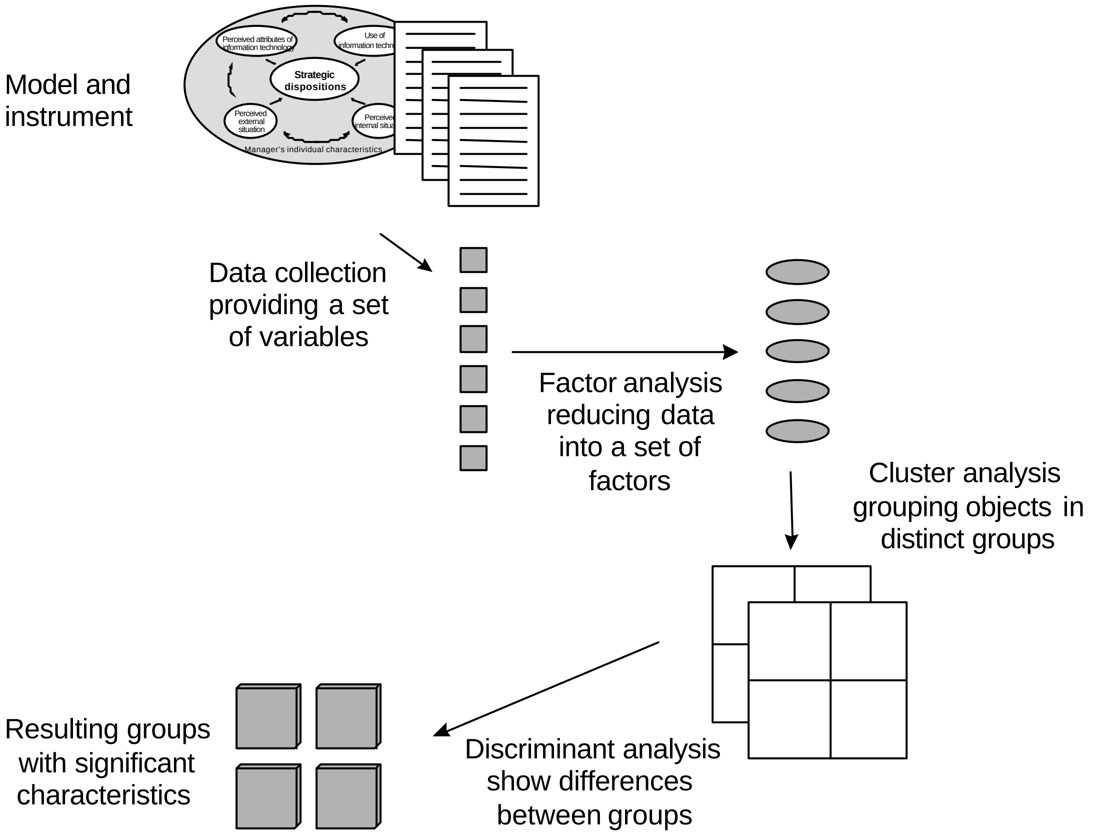
Model and instrument
Perceived attributes of information technology
Use of information techn
Strategic dispositions
Perceived external situation
Perceived internal situa
Manager's individual characteristics
Data collection providing a set of variables
Factor analysis reducing data into a set of factors
Cluster analysis grouping objects in distinct groups
Discriminant analysis show differences between groups
Resulting groups with significant characteristics

Figure 10. The overall approach in the case.

REFERENCES

Befring, Edvard, **Forskningsmetodik och statistik.** Lund. Studentlitteratur. 1994.

Bollen, Kenneth A, **Structural equations with latent variables**. New York. Wiley. 1989.

Hair Jr, J F; Anderson, R E; Tatham, R L; Black, W C, **Multivariate Data Analysis with Readings**. 4th ed. Englewood Cliffs. Prentice Hall. 1995.

Harman, Harry H, **Modern Factor Analysis**. Chicago. The University of Chicago Press. 1967.

Joliffe, I T, **Principal Component Analysis**. New York. Springer-Verlag. 1986.

Junghagen, Sven, **Strategiska Förhållningssätt till Informationsteknik i Små Företag.** Umeå. Umeå Universitet. 1998.

Junghagen, Sven, **Nyttan av IT – i småföretagarens ögon**. Stockholm. NUTEK. 1999.

Lindeman, R H; Merenda, P F, Gold R Z, **Introduction to Bivariate and Multivariate Analysis**. Glenview. Scott, Foresman and Company. 1980.

Malhotra, Naresh K, **Marketing Research - an applied orientation.** Englewood Cliffs. Prentice Hall. 1993.

Norusis, Marija J, **SPSS Professional statistics 6.1**. Chicago. SPSS Inc. 1994.

Parasuraman, A, **Marketing Research**. 2d ed. Reading. Addison-Wesley Publishing Company. 1991.

Pearson, Karl, *On lines and planes of closest fit to systems of point in space*, **The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science.** 6th series Vol 2 No 11. 1901.

Sharma, Subhash, **Applied Multivariate Techniques**. New York. Wiley. 1996.

Wold, S; Esbensen, K; Geladi P, *Principal Component Analysis*. **Chemometrics and Intelligent Laboratory Systems**. Iss 2/87 p 37-52. 1987.