Chapter - 4

Statistical Techniques Used in Project Management

Project managers need to make sense on the many data that they have. Analytical tools are used in project management to achieve such need. Such tools are used to create a forecast of potential outcomes based on the variations present in the environmental and project variables. There are different types of analytical tools used and one of the most common tools is regression analysis.

1. MULTIPLE REGRESSIONS:

Concept: Multiple regressions is the most commonly utilized multivariate technique. It examines the relationship between a single metric dependent variable and two or more metric independent variables. Multiple regressions are often used as a forecasting tool. Goal is to use the linear composite of two or more continuous and/or categorical variables (predictors) to: 1) predict scores on a single continuous variable (criterion), or to 2) explain the nature of the single continuous criterion variable from what is known about the predictor variables. In prediction, the criterion is the main emphasis because decisions are made on its value, but often times, it is difficult to directly measure or obtain a subject's actual score on the criterion; therefore, it is important to estimate or predict one's criterion score based on the value of the predictor scores.

Analysis: The technique relies upon determining the linear relationship with the lowest sum of squared variances; therefore, assumptions of normality, linearity,

and equal variance are carefully observed. The beta coefficients (weights) are the marginal impacts of each variable, and the size of the weight can be interpreted directly.

Multiple Linear Regression Analysis consists of more than just fitting a linear line through a cloud of data points. It consists of 3 stages – (1) analysing the correlation and directionality of the data, (2) estimating the model, i.e., fitting the line, and (3) evaluating the validity and usefulness of the model.

Firstly, the scatter plots should be checked for directionality and correlation of data. Typically you would look at an individual scatter plot for every independent variable in the analysis.

The data is fit to run a multiple linear regression analysis.

However, most often data contains quite a large amount of variability (just as in the third scatter plot example) in these cases it is up for decision how to best proceed with the data.

The second step of multiple linear regressions is to formulate the model, i.e. that variable X1, X2, and X3 have a causal influence on variable Y and that their relationship is linear.

The third step of regression analysis is to fit the regression line.

Project implication: The regression analysis is a technique that involves examining the series of input variables in relations to the corresponding output results. This particular project management technique is used in establishing the statistical relationship between two variables. It is used to determine whether one variable, the independent variable, is used to predict the dependent variable. In regression analysis, the stronger the relationship is between the two variables, the greater the accuracy in predicting their relationship. Project managers can easily see the relationship between two variables by using a simple linear formula and plotting the results on a chart.

Using this tool provides project managers a clearer understanding on the relationship of two things. Used together with the other analytical tools, they can use information for the better decision-making.

2. STEPWISE REGRESSION :

Concept: In statistics, **stepwise regression** is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some pre-specified criterion. Usually, this takes the form of a sequence of *F*-tests or *t*-tests, but other techniques are possible, such as adjusted R^2, Akaike information criterion, Bayesian information criterion, Mallows's C_p, PRESS, or false discovery rate.

The frequent practice of fitting the final selected model followed by reporting estimates and confidence intervals without adjusting them to take the model building process into account has led to calls to stop using stepwise model building altogether or to at least make sure model uncertainty is correctly reflected.

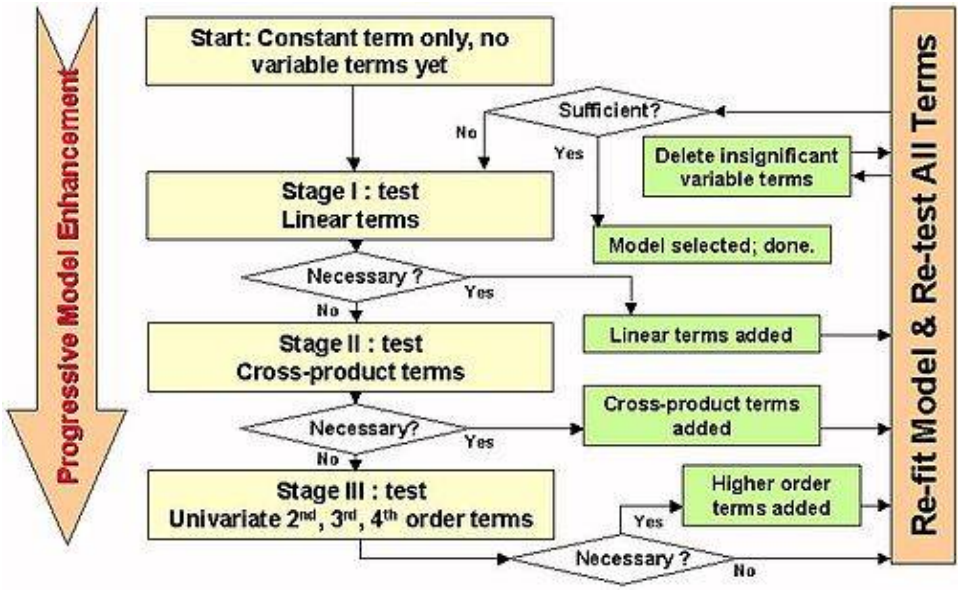
Analysis:

The main approaches are:

Forward selection, which involves starting with no variables in the model, testing the addition of each variable using a chosen model fit criterion, adding the variable (if any) whose inclusion gives the most statistically significant improvement of the fit, and repeating this process until none improves the model to a statistically significant extent.

Backward elimination, which involves starting with all candidate variables, testing the deletion of each variable using a chosen model fit criterion, deleting the variable (if any) whose loss gives the most statistically insignificant deterioration of the model fit, and repeating this process until no further variables can be deleted without a statistically significant loss of fit.

Bidirectional elimination, a combination of the above, testing at each step for variables to be included or excluded.



Project implication:

Stepwise regression analysis can successfully be applied for successful resource allocation and performance appraisal as well in the following way:

- Allocation of resources can be made in proportion with R2 value contributed by few selected and most effective variables retained at the last step.
- Performance of different factors in an organization can be filtered and rewards can be assigned as per the contribution (R2 value made by each of the factor) .
- The management issues and intervention points can be downsized by following stepwise elimination of trivial factors and at the last step landing on few selected critical factors.
- Project management can save cost, resource, time, opportunity and risk by downsizing the subjective variables into few dealable factors.
- By following stepwise regression a further filtration can also be possible between qualitative and quantitative variables.

3. PATH ANALYSIS

Concept: In statistics, path analysis is used to describe the directed dependencies among a set of variables. This includes models equivalent to any form of multiple regression analysis, factor analysis, canonical correlation analysis, discriminant analysis, as well as more general families of models in the multivariate analysis of variance and covariance analyses (MANOVA, ANOVA, ANCOVA).

In addition to being thought of as a form of multiple regression focusing on causality, path analysis can be viewed as a special case of structural equation modelling (SEM) – one in which only single indicators are employed for each of the variables in the causal model. That is, path analysis is SEM with a structural model, but no measurement model. Other terms used to refer to path analysis include causal modelling, analysis of covariance structures, and latent variable models.

Path analysis is considered by Judea Pearl to be a direct ancestor to the techniques of Causal inference.

There are two main requirements for path analysis:

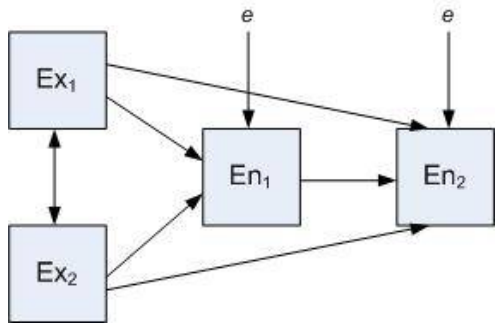
1. All causal relationships between variables must go in one direction only (you cannot have a pair of variables that cause each other)
2. The variables must have a clear time-ordering since one variable cannot be said to cause another unless it precedes it in time.

Components: Typically path analysis involves the construction of a path diagram in which the relationships between all variables and the causal direction between them are specifically laid out. When conducting path analysis one might first construct an input path diagram, which illustrates the hypothesized relationships. After statistical analysis has been completed, a researcher would then construct an output path diagram, which illustrates the relationships as they actually exist, according to the analysis conducted. Analysis: Typically, path

models comprise an inner and an outer model. Variables substantiating the outer model are called 'exogenous', variables constituting the inner model are referred to as 'endogenous'. Commonly, exogenous variables are those forming factors (factor analysis) or predictors in a regression (multiple regressions).

In the model below, the two exogenous variables (Ex_1 and Ex_2) are modelled as being correlated and as having both direct and indirect (through En_1) effects on En_2 (the two dependent or 'endogenous' variables/factors). In most real-world models, the endogenous variables may also be affected by variables and factors stemming from outside the model (external effects including measurement error). These effects are depicted by the "e" or error terms in the model.

Using the same variables, alternative models are conceivable. For example, it may be hypothesized that Ex_1 has only an indirect effect on En_2, deleting the arrow from Ex_1 to En_2; and the likelihood or 'fit' of these two models can be compared statistically.



Project implication: Path analysis can successfully be applied for successful resource allocation and performance appraisal as well in the following way:

- In the analytical aspect of social ecology, wherein, most of the variables have got inter and intra level interactions leading to direct, indirect and residual effect on surrounding ecosystem, the application of path analysis is most suitable,

- The hidden or implicit effect of any project component can be made tracable as well as dealable,
- The ecological performance and functions therein, the efficacy and proficiency of different ecological factors can be elucidated by calculating direct effect(beta value) and residual effect, the more the residual effect, the higher is the system error and operating non-compliances among and between system and criterion variables.
- Both genotypic and phenotypic contribution as well as expositions can be identified by applying the path analysis model. In most cases, the project are suffering from implicit or hidden factors, by applying path analysis, we can identify the hidden agenda and their respective contributions to the system behavior and function.
- Those who have highest indirect effect, it is expected that they have got associating properties and impact with other variables performing in an iso-chronous manner, so isolation of these factors will help either expedite (positively) or retarded. The function of other factors as already been included in project execution and monitoring strategy.

4. CANONICAL CORRELATION.

Concept: Canonical correlation is another extension of multiple regression where rather than using a single outcome variable Y, two or more Y variables are predicted by two or more predictor X variables. The purpose is usually not to predict Ys from Xs but to explain the relationship between the X and Y variable sets. The analysis can be run in both directions, Xs predicting Ys and Ys predicting Xs, but the researcher is generally concerned about Xs predicting Ys where the Y set is the new hypothesized inter-relationship and the X set are the traditional predictors seen in prior research.

In Multiple regression, several X variables (predictors) are independent and only one Y variable (criterion) is dependent. In canonical correlation, there are multiple sets of X variables and multiple sets of Y variables..

Analysis: Because canonical correlation involves multiple Xs and multiple Ys, shared variance exists along two or more dimensions or geometric axes (with one outcome variable in multiple regression, there is only one dimension or axis). In fact, the number of dimensions or axes is equal to the number of variables in the smaller of the two sets. For example, if the data had three X variables and four Y variables, then the number of dimensions would be three. If the data had two X variables and two Y variables, the number of dimensions would be two. Each dimension will be represented by two linear composites, one for the X set of variables and one for the Y set. The correlation between the two composites is called the *canonical correlation*, RC. It is analogous to the multiple R in multiple regression. The Rc2, also labeled as the *eigenvalue*, indicates the proportion of variance shared between the two composites. Like multiple regression where the predictor variable with the greatest degree of shared variance with Y is entered into the model first, the two sets of X and Y variables with the greatest degree of shared variance are entered into the canonical correlation model first. This canonical correlation of the X and Y composite variables is called the 1st *canonical* function. After this first canonical function extracts its proportion of shared variance then a second canonical correlation is computed from the remaining variance. Because the second function does not include any of the variance of the first function, the two sets are totally uncorrelated (orthogonal or independent). Each successive canonical function will be uncorrelated with all previous canonical functions. It is important to also note that because each successive canonical function has less and less remaining variance in which to operate, each successive RC gets smaller than the previous canonical correlations.

Interpretation involves five steps:

1) The overall significance of the model (null: explained variance = 0) is tested by Wilk's lambda. If the model is significant (explained variance > 0), then proceed.
2) Determine which canonical functions are important enough to keep. Two indices are considered simultaneously. A $\chi 2$ based on Wilks lambda is calculated for the total set of canonical functions with the null being that all of the RC = 0. If the chi-square is significant, then one can conclude that at least the first canonical correlation is significantly greater than 0. This is because the first set has the greatest RC value. The second chi-square tests the null that all of the remaining RC = 0. At the point the chi-square becomes non-significant, the remaining chi-squares will also be non-significant. As mentioned in previous sections, statistical significance is heavily influenced by sample size, so one must also look at the RC2 value of each canonical function. A general rule given by Pedhazur (1997) is to keep functions with RC2 > 10%.
3) As in other regression analyses, regression coefficients will be calculated for each variable in the X and Y set. These are named *canonical weights or coefficients*. These are standardized and their values will be recalculated for each successive canonical function. These can be interpreted as to their respective contribution but have the same interpretation confound seen in other regression coefficients.
4) Structure coefficients are more widely used for determining the importance of each variable. The structure coefficient is the correlation between a variable and its respective linear composite. For example, the structure coefficient for X1 is the correlation of X1 with the X linear composite. The squared structure coefficient is the amount of the linear composite variance that is explained by X1. Most researchers construct a table of structure coefficients much like the

table used for factor analysis. The columns represent the canonical functions retained and the rows represent each variable. The cell data represent the structure coefficients. Like factor analysis, the researcher then underscores the structure coefficients with values greater than .45 indicating that these have significant contribution to the variance of the canonical function. Furthermore, the variables which are retained provide a description of the dimension represented in that particular canonical function. These dimensions can even be named as factors are in factor analysis.

5) As discussed, the RC2 represents only the shared variance within a specific canonical function and does not represent the proportion explained in the total X and Y variance. A *redundancy index* is calculated for each canonical function to provide an estimate of the variance explained for the full model. A separate redundancy is calculated for the X set and the Y set. It is the average structure coefficient multiplied by the canonical correlation.

Project implication:

1. It helps in estimation of cost, time, resource and management impact as being contributed by score of causal variables in an isochronous manner.
2. Isolation and identification of unique combination between two sets of variables help in measuring the efficacy of time, cost, resource in terms of contributory factors.
3. It helps to delineate strategic interventions by reducing the collateral variables.

5. FACTOR ANALYSIS

Concept: The term factor analysis refers to a set of analytical techniques designed to reduce data into smaller, meaningful groups based upon their inter-correlations or shared variance. The assumption is that those items or variables that are correlated must be measuring a similar factor or trait or construct. In the case where only a few variables are used, the researcher may be able to determine

groupings by simply observing the content of each variable; however, for large data sets and/or more ambiguous items, this task would be formidable. Factor Analysis reduces the information in a model by reducing the dimensions of the observations. This procedure has multiple purposes. It can be used to simplify the data, for example reducing the number of variables in predictive regression models. If factor analysis is used for these purposes, most often factors are rotated after extraction. Factor analysis has several different rotation methods—some of them ensure that the factors are orthogonal. Then the correlation coefficient between two factors is zero, which eliminates problems of multi-co linearity in regression analysis.

Factor analysis is also used in theory testing to verify scale construction and operationalization. In such a case, the scale is specified upfront and we know that a certain subset of the scale represents an independent dimension within this scale. This form of factor analysis is most often used in structural equation modelling and is referred to as Confirmatory Factor Analysis. For example, we know that the questions pertaining to the big five personality traits cover all five dimensions N, A, O, and I. If we want to build a regression model that predicts the influence of the personality dimensions on an outcome variable, for example anxiety in public places, we would start to model a confirmatory factor analysis of the twenty questionnaire items that load onto five factors and then regress onto an outcome variable.

Factor analysis can also be used to construct indices. The most common way to construct an index is to simply sum up the items in an index. In some contexts, however, some variables might have a greater explanatory power than others. Also sometimes similar questions correlate so much that we can justify dropping one of the questions completely to shorten questionnaires. In such a case, we can use factor analysis to identify the weight each variable should have in the index.

Factor analysis is a data reduction technique that can reduce the number of items by grouping them and by examining the content of the items in each group one can determine the structure or composition of each group thereby giving a better explanation of the data. It is important to note that factor analysis is not used in prediction or explaining the relationship between different sets of variables, nor is it used to determine group differences. The goal is to explain the underlying structure or composition of the data; therefore we are dealing only with one set of variables.

Two types of factor analysis exist. The first, *exploratory factor analysis* (EFA) is used to explore or derive the underlying factor structure of a data matrix often without regard to theory. The purpose of EFA is to determine if underlying factors exist within a data set, and if so, what those factors are. The researcher does not necessarily need to have any expectations or theory beforehand—it can simply be "exploratory." Two types of EFA are commonly used. *Principle components analysis* (PCA) tries to account for all variance among the variables/items and so includes both shared variance and unique/error variance of each variable/item. *Principle factor analysis* (PFA) accounts for only shared variance of each variable/item. Currently, most journals prefer PFA.

The second, *confirmatory factor analysis* (CFA), is used to test a priori theory. The researcher specifies what factors exist and what variables/items constitute each factor and then "orders" these parameters into a data set to determine if indeed the factors and variables/items describe or fit the data. CFA is most commonly conducted through Structural Equation Modeling (SEM). The following discussion is on factor analysis pertaining to EFA.

The main goal of factor analysis is to explain as much variance as possible in a data set by using the smallest number of factors (groupings of variables based on high inter-correlations) and the smallest amount of items or variables within each factor. Inherently one balances explained variance with simplicity. Critical to this

technique is that one wants to ensure that the variance left out of the solution is primarily error variance. In EFA, the first factor derived is a linear composite of all variables/items such that it maximizes the amount of total variance explained (or extracted)—no other linear combination will extract as much variance. The proportion of variance extracted is called the *eigen value.* A second factor which is orthogonal to the first (uncorrelated) is then derived from the remaining variance, and its eigen value will be derived. Like the first factor, the second factor will be a linear composite of all variables/items. This process continues until all variance has been extracted, and the number of extractions is equal to the number of original variables/items.

Components:

Factor loadings: Commonality is the square of standardized outer loading of an item. Analogous to Pearson's r, the squared factor loading is the percent of variance in that indicator variable explained by the factor. To get the percent of variance in all the variables accounted for by each factor, add the sum of the squared factor loadings for that factor (column) and divide by the number of variables. (Note the number of variables equals the sum of their variances as the variance of a standardized variable is 1.) This is the same as dividing the factor's eigen value by the number of variables.

Interpreting factor loadings: By one rule of thumb in confirmatory factor analysis, loadings should be .7 or higher to confirm that independent variables identified a priori are represented by a particular factor, on the rationale that the .7 level corresponds to about half of the variance in the indicator being explained by the factor. However, the .7 standard is a high one and real-life data may well not meet this criterion, which is why some researchers, particularly for exploratory purposes, will use a lower level such as .4 for the central factor and .25 for other

factors. In any event, factor loadings must be interpreted in the light of theory, not by arbitrary cut off levels.

In oblique rotation, one gets both a pattern matrix and a structure matrix. The structure matrix is simply the factor loading matrix as in orthogonal rotation, representing the variance in a measured variable explained by a factor on both a unique and common contributions basis. The pattern matrix, in contrast, contains coefficients which just represent unique contributions. The more factors, the lower the pattern coefficients as a rule since there will be more common contributions to variance explained. For oblique rotation, the researcher looks at both the structure and pattern coefficients when attributing a label to a factor. Principles of oblique rotation can be derived from both cross entropy and its dual entropy.

Communality: The sum of the squared factor loadings for all factors for a given variable (row) is the variance in that variable accounted for by all the factors, and this is called the communality. The communality measures the percent of variance in a given variable explained by all the factors jointly and may be interpreted as the reliability of the indicator.

Spurious solutions: If the communality exceeds 1.0, there is a spurious solution, which may reflect too small a sample or the researcher has too many or too few factors.

Uniqueness of a variable: That is, uniqueness is the variability of a variable minus its communality.

Eigenvalues/characteristic roots: The eigenvalue for a given factor measures the variance in all the variables which is accounted for by that factor. The ratio of eigenvalues is the ratio of explanatory importance of the factors with respect to the variables. If a factor has a low eigenvalue, then it is contributing little to the explanation of variances in the variables and may be ignored as redundant with more important factors. Eigenvalues measure the amount of variation in the total sample accounted for by each factor.

Extraction sums of squared loadings: Initial eigenvalues and eigenvalues after extraction (listed by SPSS as "Extraction Sums of Squared Loadings") are the same for PCA extraction, but for other extraction methods, eigenvalues after extraction will be lower than their initial counterparts. SPSS also prints "Rotation Sums of Squared Loadings" and even for PCA, these eigenvalues will differ from initial and extraction eigenvalues, though their total will be the same.

Factor scores (also called component scores in PCA): are the scores of each case (row) on each factor (column). To compute the factor score for a given case for a given factor, one takes the case's standardized score on each variable, multiplies by the corresponding loadings of the variable for the given factor, and sums these products. Computing factor scores allows one to look for factor outliers. Also, factor scores may be used as variables in subsequent modeling. (Explained from PCA not from Factor Analysis perspective).

Analysis: It involves three steps:

1) The researcher determines the number of factors to keep. Several decision rules can be employed but generally keep those factors with eigen values greater than 1.0. The reason is that any given variable has a variance equal to 1.0 (since variables are standardized, the std. dev. = 1.0, and variance is the std. dev. Squared) which means that eigen values should explain more variance than at least one variable/item. In a satisfactory EFA, the total variance of the retained eigen values should be greater than 70%.

2) Once the number of factors has been determined, the researcher then determines which variables/items “load” on each factor. This is determined by the coefficient of the variable/item. Because each factor is a linear composite of all variables/items, each variable/item will have a different coefficient for each factor. This coefficient is actually a structure coefficient since it is the correlation between the variable/item and the factor; however, it is called the factor loading or the

factor structure coefficient. The higher the coefficient, the greater the variable/item's contribution to the factor, and the square of the coefficient is the amount of variance of the factor that is explained by the variable/item. In general, the desired outcome is that each variable/item will have a large loading on only one factor and small loadings on the remaining factors.

Before actually determining the composition of each factor, the axes which represent factor dimensions can be rotated geometrically so that the new set of axes are positioned closer to their respective factor variables/items. These axes are similar to those in a Cartesian coordinate system. For example, in a two-factor model, every variable/item will have a factor loading value for Factor I and for Factor II. Most will have a relatively high loading value (>.5) on one factor and relatively small loading on the other factor (< .5). Each factor is an axis in space and the factor loading of a variable/item is the coordinate value on that axis. If variable X1 has a factor loading of .63 on FI, and .22 on FII, then its coordinate point is (.63, .22). Also, the distance of X1 from the origin can be calculated by the Pythagorean theorem (add .63 squared and .22 squared and take the square root). By rotating the axes, one can increase the value along the FI dimension and decrease the FII dimension value as long as the distance of X1 to the origin remains unchanged. This strengthens the association of X1 with FI while decreasing its association to FII. This makes interpretation of the factor structure (composition) simpler. Usually when rotating the axes, the researcher assumes that the factors are truly orthogonal and so an orthogonal rotation is used, that is the axes remain at 90 degrees to one another. This is easy to visualize in a two-factor model. If F1 is rotated 35 degrees clockwise, FII must also be rotated 35 degrees clockwise. However, sometimes the factors are somewhat correlated and so the rotation that occurs is based upon the correlation between the factors. The cosine of the angle between the axes is equal to the correlation. By the way, this rule

holds true when orthogonal rotation is used because the angles are at 90 degrees and the cosine of 90 degrees is 0.

3) With the output arranged in a table where the columns are the factors and the rows are the variables/items, one then highlights or underscores the loadings under each factor that are greater than .50. Once all contributing variables on all factors have been identified, then one must determine the content of each factor and assign an appropriate name for each. This is done by analyzing the content or general theme of the variable/items that are highlighted.

Factor analysis has been implemented in several statistical analysis programs since the 1980s:

BMDP, JMP (statistical software), Python: module Scikit-learn, R (with the base function *factanal* or *fa* function in package **psych**). Rotations are implemented in the *GP Arotation* R package., SAS (using PROC FACTOR or PROC CALIS), SPSS, Stata.

Project implication: 1.Factor analysis is commonly used in biology, psychometrics, personality theories, marketing, product management, operations research, and finance.

2. It is an effective tool for the estimation of possibility of combination and performance of resource variable.

6. PRINCIPAL COMPONENT ANALYSIS:

Concept: Principal component analysis (PCA) involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called *principal components*. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

Objectives of principal component analysis

Traditionally, principal component analysis is performed on the Covariance matrix or on the Correlation matrix. These matrices can be calculated from the data matrix. The covariance matrix contains scaled sums of squares and cross products. A correlation matrix is like a covariance matrix but first the variables, i.e. the columns, have been standardized. We will have to standardize the data first if the variances of variables differ much, or if the units of measurement of the variables differ. The mathematical technique used in PCA is called eigen analysis: we solve for the eigen values and eigenvectors of a square symmetric matrix with sums of squares and cross products. The eigenvector associated with the largest eigen value has the same direction as the first principal component. The eigenvector associated with the second largest eigen value determines the direction of the second principal component. The sum of the eigen values equals the trace of the square matrix and the maximum number of eigenvectors equals the number of rows (or columns) of this matrix.

Analysis: There are two methods to help you to choose the number of components to keep. Both methods are based on relations between the eigen values.

Plot the eigen values, If the points on the graph tend to level out (show an "elbow"), these eigen values are usually close enough to zero that they can be ignored.

Limit the number of components to that number that accounts for a certain fraction of the total variance.

Project implementation: 1.Discovers or reduces the dimensionality of the data set.

2. It helps to identify new meaningful underlying variables.

7. DISCRIMINANT FUNCTION ANALYSIS

Concept: Discriminant function analysis is used to determine which variables discriminate between two or more naturally occurring groups. For example, an educational researcher may want to investigate which variables discriminate between high school graduates who decide (1) to go to college, (2) to attend a trade or professional school, or (3) to seek no further training or education. For that purpose the researcher could collect data on numerous variables prior to students' SSLC. After SSLC, most students will naturally fall into one of the three categories. *Discriminant Analysis* could then be used to determine which variable(s) are the best predictors of students' subsequent educational choice.

Stepwise Discriminant Analysis

Probably the most common application of discriminant function analysis is to include many measures in the study, in order to determine the ones that discriminate between groups. For example, an educational researcher interested in predicting high school students' choices for further education would probably include as many measures of personality, achievement motivation, academic performance, etc. as possible in order to learn which one(s) offer the best prediction. Put another way, we want to build a "model" of how we can best predict to which group a case belongs. In the following discussion we will use the term "in the model" in order to refer to variables that are included in the prediction of group membership, and we will refer to variables as being "not in the model" if they are not included.

Forward stepwise analysis. In stepwise Discriminant function analysis, a model of discrimination is built step-by-step. Specifically, at each step all variables are reviewed and evaluated to determine which one will contribute most to the discrimination between groups. That variable will then be included in the model, and the process starts again.

Backward stepwise analysis. One can also step backwards; in that case all variables are included in the model and then, at each step, the variable that contributes least to the prediction of group membership is eliminated. Thus, as the result of a successful Discriminant function analysis, one would only keep the "important" variables in the model, that is, those variables that contribute the most to the discrimination between groups.

***F* to enter, *F* to remove.** The stepwise procedure is "guided" by the respective *F* to enter and *F* to remove values. The *F* value for a variable indicates its statistical significance in the discrimination between groups, that is, it is a measure of the extent to which a variable makes a unique contribution to the prediction of group membership.

Analysis: Interpreting a Two-Group Discriminant Function

In the two-group case, Discriminant function analysis can also be thought of as (and is analogous to) multiple regression. If we code the two groups in the analysis as *1* and *2*, and use that variable as the dependent variable in a multiple regression analysis, then we would get results that are analogous to those we would obtain via *Discriminant Analysis*. In general, in the two-group case we fit a linear equation of the type:

Group = a + b1*x1 + b2*x2 + ... + bm*xm

Where *a* is a constant and *b1* through *bm* are regression coefficients. The interpretation of the results of a two-group problem is straightforward and closely follows the logic of multiple regression: Those variables with the largest (standardized) regression coefficients are the ones that contribute most to the prediction of group membership.

Interpreting the Discriminant functions. As before, we will get *b* (and standardized *beta*) coefficients for each variable in each Discriminant (now also called *canonical*) function, and they can be interpreted as usual: the larger the

standardized coefficient, the greater is the contribution of the respective variable to the discrimination between groups. However, these coefficients do not tell us between which of the groups the respective functions discriminate. We can identify the nature of the discrimination for each discriminant (canonical) function by looking at the means for the functions across groups. We can also visualize how the two functions discriminate between groups by plotting the individual scores for the two Discriminant functions

Significance of discriminant functions. One can test the number of roots that add *significantly* to the discrimination between group. Only those found to be statistically significant should be used for interpretation; non-significant functions (roots) should be ignored.

Classification (Predictive Discriminant Analysis) Another major purpose to which discriminant analysis is applied is the issue of predictive classification of cases. Once a model has been finalized and the discriminant functions have been derived, how well can we *predict* to which group a particular case belongs?

***A priori* and *post hoc* predictions.** Before going into the details of different estimation procedures, we would like to make sure that this difference is clear. Obviously, if we estimate, based on some data set, the discriminant functions that best discriminate between groups, and then use the *same* data to evaluate how accurate our prediction is, then we are very much capitalizing on chance. In general, one will *always* get a worse classification when predicting cases that were not used for the estimation of the discriminant function. Put another way, *post hoc* predictions are always better than *a priori* predictions. Therefore, one should never base one's confidence regarding the correct classification of future observations on the same data set from which the discriminant functions were derived; rather, if one wants to classify cases predicatively, it is necessary to collect new data to "try out" (cross-validate) the utility of the discriminant functions.

Classification functions. These are not to be confused with the discriminant functions. The classification functions can be used to determine to which group each case most likely belongs. There are as many classification functions as there are groups. Each function allows us to compute

classification scores for each case for each group, by applying the formula:

Si = ci + wi1*x1 + wi2*x2 + ... + wim*xm

In this formula, the subscript *i* denotes the respective group; the subscripts *1, 2, ..., m* denote the *m* variables; *ci* is a constant for the *i*'th group, *wij*is the weight for the *j*'th variable in the computation of the classification score for the *i*'th group; *xj*is the observed value for the respective case for the *j*'th variable. *Si* is the resultant classification score. We can use the classification functions to directly compute classification scores for some new observations.

Classification of cases. Once we have computed the classification scores for a case, it is easy to decide how to classify the case: in general we classify the case as belonging to the group for which it has the highest classification score. Thus, if we were to study high school students' post-school career/educational choices (e.g., attending college, attending a professional or trade school, or getting a job) based on several variables assessed one year prior to graduation, we could use the classification functions to predict what each student is most likely to do after SSLC. However, we would also like to know the *probability* that the student will make the predicted choice. Those probabilities are called *posterior* probabilities, and can also be computed. However, to understand how those probabilities are derived, let us first consider the so-called *Mahalanobis* distances.

Mahalanobis distances. In general, the Mahalanob is distance is a measure of distance between two points in the space defined by two or more correlated variables. For example, if there are two variables that are uncorrelated, then we could plot points (cases) in a standard two-dimensional scatter plot; the Mahalanobis distances between the points would then be identical to the Euclidean

distance; that is, the distance as, for example, measured by a ruler. If there are three uncorrelated variables, we could also simply use a ruler (in a 3-D plot) to determine the distances between points. If there are more than 3 variables, we cannot represent the distances in a plot any more. Also, when the variables are correlated, then the axes in the plots can be thought of as being *non-orthogonal*; that is, they would not be positioned in right angles to each other. In those cases, the simple Euclidean distance is not an appropriate measure, while the Mahalanobis distance will adequately account for the correlations.

Mahalanobis distances and classification. For each group in our sample, we can determine the location of the point that represents the means for all variables in the multivariate space defined by the variables in the model. These points are called group *centroids*. For each case we can then compute the Mahalanobis distances (of the respective case) from each of the group centroids. Again, we would classify the case as belonging to the group to which it is closest, that is, where the Mahalanobis distance is smallest.

Posterior classification probabilities. Using the Mahalanobis distances to do the classification, we can now derive probabilities. The probability that a case belongs to a particular group is basically proportional to the Mahalanobis distance from that group centroid. In summary, the posterior probability is the probability, based on our knowledge of the values of other variables, that the respective case belongs to a particular group.

Summary of the prediction. A common result that one looks at in order to determine how well the current classification functions predict group membership of cases is the *classification matrix*. The classification matrix shows the number of cases that were correctly classified (on the diagonal of the matrix) and those that were misclassified.

Project implementation: It helps isolating and measuring efficiency of critical variables in making its difference in performance, product and proficiency.

8. CLUSTER ANALYSIS

Concept: The purpose of cluster analysis is to reduce a large data set to meaningful subgroups of individuals or objects. The division is accomplished on the basis of similarity of the objects across a set of specified characteristics. Outliers are a problem with this technique, often caused by too many irrelevant variables. The sample should be representative of the population, and it is desirable to have uncorrelated factors. This is a great tool for market segmentation.

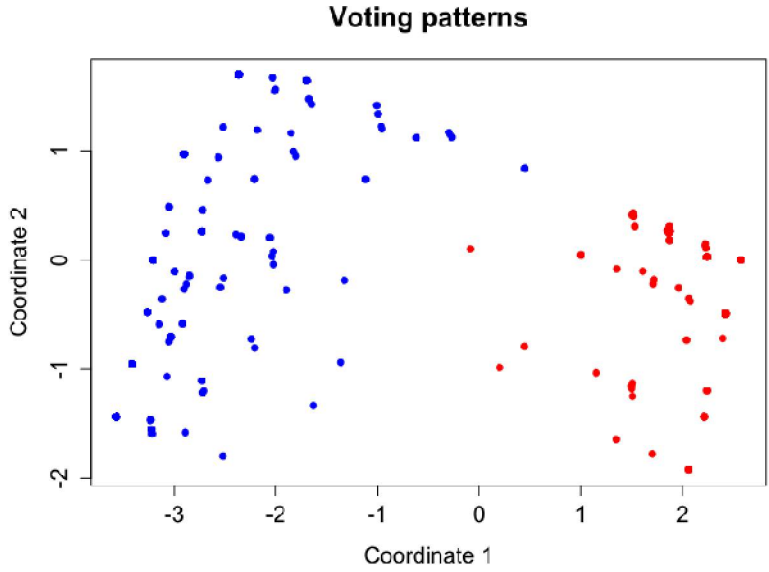
Analysis: There are three main clustering methods: hierarchical, which is a treelike process appropriate for smaller data sets; non-hierarchical, which requires specification of the number of clusters a priori; and a combination of both. There are four main rules for developing clusters: the clusters should be different, they should be reachable, they should be measurable, and the clusters should be profitable (big enough to matter).

Project implementation: It helps estimating customer conglomeration, group behaviour and standardizing effective group number and performance.

9. MULTIDIMENSIONAL SCALING (MDS)

Concept: The purpose of MDS is to transform consumer judgments of similarity into distances represented in multidimensional space. This is a decompositional approach that uses perceptual mapping to present the dimensions. As an exploratory technique, it is useful in examining unrecognized dimensions about products and in uncovering comparative evaluations of products when the basis for comparison is unknown. Typically there must be at least four times as many objects being evaluated as dimensions. It is possible to evaluate the objects with nonmetric preference rankings or metric similarities (paired comparison) ratings. Kruskal's Stress measure is a "badness of fit" measure; a stress percentage of 0

indicates a perfect fit, and over 20% is a poor fit. The dimensions can be interpreted either subjectively by letting the respondents identify the dimensions or objectively by the researcher. **Multidimensional scaling** (**MDS**) is a means of visualizing the level of similarity of individual cases of a dataset. It refers to a set of related ordination techniques used in information visualization, in particular to display the information contained in a distance matrix. It is a form of non-linear dimensionality reduction. An MDS algorithm aims to place each object in *N*-dimensional space such that the between-object distances are preserved as well as possible. Each object is then assigned coordinates in each of the *N* dimensions. The number of dimensions of an MDS plot *N* can exceed 2 and is specified a priori. Choosing *N*=2 optimizes the object locations for a two-dimensional scatter plot.



Analysis: There are several steps in conducting MDS research:

Formulating the problem – What variables do you want to compare? How many variables do you want to compare? What purpose is the study to be used for?

Obtaining input data – For example, :- Respondents are asked a series of questions. For each product pair, they are asked to rate similarity (usually on a 7-

point Likert scale from very similar to very dissimilar). The first question could be for Coke/Pepsi for example, the next for Coke/Hires rootbeer, the next for Pepsi/Dr Pepper, the next for Dr Pepper/Hires rootbeer, etc. The number of questions is a function of the number of brands and can be calculated as where Q is the number of questions and N is the number of brands. This approach is referred to as the "Perception data : direct approach". There are two other approaches. There is the "Perception data : derived approach" in which products are decomposed into attributes that are rated on a semantic differential scale. The other is the "Preference data approach" in which respondents are asked their preference rather than similarity.

Running the MDS statistical program – Software for running the procedure is available in many statistical software packages. Often there is a choice between Metric MDS (which deals with interval or ratio level data), and Nonmetric MDS(which deals with ordinal data).

Decide number of dimensions – The researcher must decide on the number of dimensions they want the computer to create. The more dimensions, the better the statistical fit, but the more difficult it is to interpret the results.

Mapping the results and defining the dimensions – The statistical program (or a related module) will map the results. The map will plot each product (usually in two-dimensional space). The proximity of products to each other indicate either how similar they are or how preferred they are, depending on which approach was used. How the dimensions of the embedding actually correspond to dimensions of system behavior, however, are not necessarily obvious. Here, a subjective judgment about the correspondence can be made .

Test the results for reliability and validity – Compute R-squared to determine what proportion of variance of the scaled data can be accounted for by the MDS procedure. An R-square of 0.6 is considered the minimum acceptable

level. An R-square of 0.8 is considered good for metric scaling and .9 is considered good for non-metric scaling. Other possible tests are Kruskal's Stress, split data tests, data stability tests (i.e., eliminating one brand), and test-retest reliability.

Report the results comprehensively – Along with the mapping, at least distance measure (e.g., Sorenson index, Jaccard index) and reliability (e.g., stress value) should be given. It is also very advisable to give the algorithm (e.g., Kruskal, Mather), which is often defined by the program used (sometimes replacing the algorithm report), if you have given a start configuration or had a random choice, the number of runs, the assessment of dimensionality, the Monte Carlo method results, the number of iterations, the assessment of stability, and the proportional variance of each axis (r-square).

Project implementation:

It helps reduction of multidimensionality of critical factors in project management and its subsequent standardization for visualizing small group homogeneity.

10. CORRESPONDENCE ANALYSIS

Concept: This technique provides for dimensional reduction of object ratings on a set of attributes, resulting in a perceptual map of the ratings. However, unlike MDS, both independent variables and dependent variables are examined at the same time. This technique is more similar in nature to factor analysis. It is a compositional technique, and is useful when there are many attributes and many companies. It is most often used in assessing the effectiveness of advertising campaigns. It is also used when the attributes are too similar for factor analysis to be meaningful. The main structural approach is the development of a contingency (crosstab) table. This means that the form of the variables should be nonmetric. The model can be assessed by examining the Chi-square value for the model. Correspondence analysis is difficult to interpret, as the dimensions are a combination of independent and dependent variables.

Analysis: Correspondence analysis (CA) or reciprocal averaging is a multivariate statistical technique proposed by Herman Otto Hartley (Hirschfeld) and later developed by Jean-Paul Benzécri. It is conceptually similar to principal component analysis, but applies to categorical rather than continuous data. In a similar manner to principal component analysis, it provides a means of displaying or summarising a set of data in two-dimensional graphical form.

All data should be nonnegative and on the same scale for CA to be applicable, keeping in mind that the method treats rows and columns equivalently. It is traditionally applied to contingency tables — CA decomposes the chi-squared statistic associated with this table into orthogonal factors. Because CA is a descriptive technique, it can be applied to tables whether or not the statistic is appropriate.

Project implementation: It helps in conjoint analysis of performing variables in a project management scenario for ensuring rationalization of number of factors and variables through a concurrent estimation.

11. CONJOINT ANALYSIS

Concept: Conjoint analysis is often referred to as "trade-off analysis, " since it allows for the evaluation of objects and the various levels of the attributes to be examined. It is both a compositional technique and a dependence technique, in that a level of preference for a combination of attributes and levels is developed. A part-worth, or utility, is calculated for each level of each attribute, and combinations of attributes at specific levels are summed to develop the overall preference for the attribute at each level. Models can be built that identify the ideal levels and combinations of attributes for products and services.

'**Conjoint analysis**' is a survey-based statistical technique used in market research that helps determine how people value different attributes (feature, function, benefits) that make up an individual product or service.

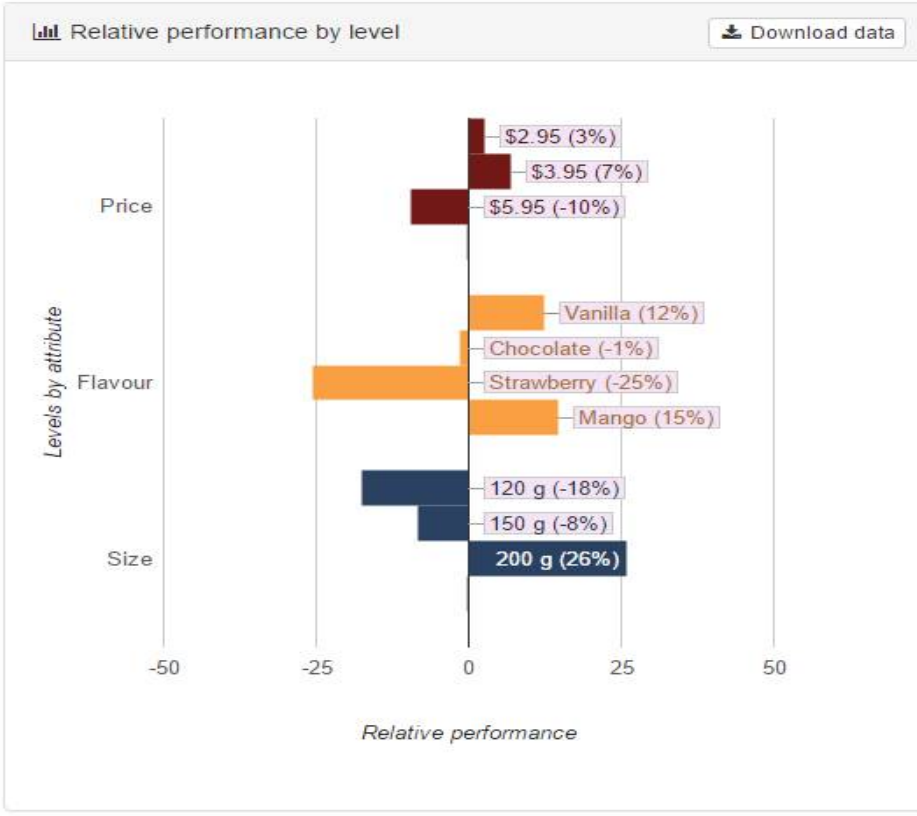
The objective of conjoint analysis is to determine what combination of a limited number of attributes is most influential on respondent choice or decision making. A controlled set of potential products or services is shown to survey respondents and by analyzing how they make preferences between these products, the implicit valuation of the individual elements making up the product or service can be determined. These implicit valuations (utilities or part-worths) can be used to create market models that estimate market share, revenue and even profitability of new designs.

Conjoint originated in mathematical psychology and was developed by marketing professor Paul E. Green at the Wharton School of the University of Pennsylvania. Other prominent conjoint analysis pioneers include professor V. "Seenu" Srinivasan of Stanford University who developed a linear programming (LINMAP) procedure for rank ordered data as well as a self-explicated approach, Richard Johnson who developed the Adaptive Conjoint Analysis technique in the 1980sand Jordan Louviere (University of Iowa) who invented and developed choice-based approaches to conjoint analysis and related techniques such as best–worst scaling.

Today it is used in many of the social sciences and applied sciences including marketing, product management, and operations research. It is used frequently in testing customer acceptance of new product designs, in assessing the appeal of advertisements and in service design. It has been used in product positioning, but there are some who raise problems with this application of conjoint analysis.

Conjoint analysis techniques may also be referred to as multiattribute compositional modelling, discrete choice modelling, or stated preference research, and is part of a broader set of trade-off analysis tools used for systematic analysis of decisions. These tools include Brand-Price Trade-Off, Simalto, and

mathematical approaches such as AHP, evolutionary algorithms or rule-developing experimentation.



Analysis: Depending on the type of model, different econometric and statistical methods can be used to estimate utility functions. These utility functions indicate the perceived value of the feature and how sensitive consumer perceptions and preferences are to changes in product features. The actual estimation procedure will depend on the design of the task and profiles for respondents, in the type of specification, and the scale of measure for preferences (it can be ratio, ranking, choice) which can have a limited range or not. For rated full profile tasks, linear regression may be appropriate, for choice based tasks, maximum likelihood estimation, usually with logistic regression are typically used. The original

methods were monotonic analysis of variance or linear programming techniques, but contemporary marketing research practice has shifted towards choice-based models using multinomial logit, mixed versions of this model, and other refinements. Bayesian estimators are also very popular. Hierarchical Bayesian procedures are nowadays relatively popular as well.

Project implementation:

1. Market research

One practical application of conjoint analysis in business analysis is given by the following example: A real estate developer is interested in building a high rise apartment complex near an urban Ivy League university. To ensure the success of the project, a market research firm is hired to conduct focus groups with current students. Students are segmented by academic year (freshman, upper classmen, graduate studies) and amount of financial aid received. Study participants are given a series of index cards. Each card has 6 attributes to describe the potential building project (proximity to campus, cost, telecommunication packages, laundry options, floor plans, and security features offered). The estimated cost to construct the building described on each card is equivalent. Participants are asked to order the cards from least to most appealing. This forced ranking exercise will indirectly reveal the participants' priorities and preferences. Multi-variate regression analysis may be used to determine the strength of preferences across target market segments.

2. Litigation

Federal courts in the United States have allowed expert witnesses to use conjoint analysis to support their opinions on the damages that an infringer of a patent should pay to compensate the patent holder for violating its rights. Nonetheless, legal scholars have noted that the Federal Circuit's jurisprudence on the use of conjoint analysis in patent-damages calculations remains in a formative stage.

ISBN: 978-93-85822-78-0

3. It helps in accreditation and analysis of different values, virtues of a project in an isochronous manner.

4. It helps eliminating inconsistency arising out of mutually conflicting properties and helps to attain a level of concurrent compliances.

12. STRUCTURAL EQUATION MODELLING

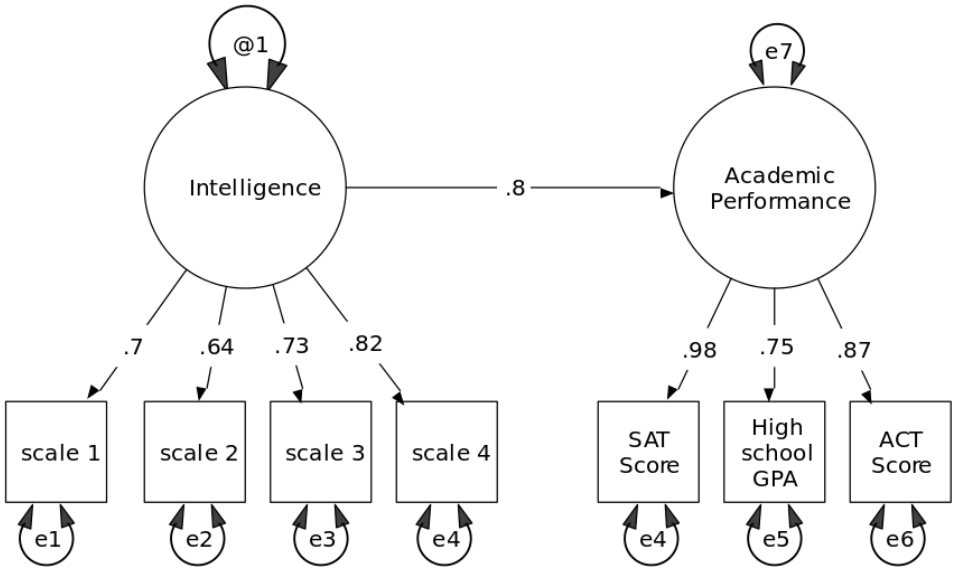
Concept: Unlike the other multivariate techniques discussed, structural equation modelling (SEM) examines multiple relationships between sets of variables simultaneously. This represents a family of techniques, including LISREL, latent variable analysis, and confirmatory factor analysis. SEM can incorporate latent variables, which either are not or cannot be measured directly into the analysis. For example, intelligence levels can only be inferred, with direct measurement of variables like test scores, level of education, grade point average, and other related measures. These tools are often used to evaluate many scaled attributes or to build summated scales.

Structural equation modelling (**SEM**) includes a diverse set of mathematical models, computer algorithms, and statistical methods that fit networks of constructs to data. SEM includes confirmatory factor analysis, path analysis, partial least squares path modelling, and latent growth modelling. The concept should not be confused with the related concept of structural models in econometrics, nor with structural models in economics. Structural equation models are often used to assess unobservable 'latent' constructs. They often invoke a measurement model that defines latent variables using one or more observed variables, and a structural model that imputes relationships between latent variables. The links between constructs of a structural equation model may be estimated with independent regression equations or through more involved approaches such as those employed in LISREL.

Use of SEM is commonly justified in the social sciences because of its ability to impute relationships between unobserved constructs (latent variables) from observable variables. To provide a simple example, the concept of human intelligence cannot be measured directly as one could measure height or weight. Instead, psychologists develop a hypothesis of intelligence and write measurement instruments with items (questions) designed to measure intelligence according to their hypothesis. They would then use SEM to test their hypothesis using data gathered from people who took their intelligence test. With SEM, "intelligence" would be the latent variable and the test items would be the observed variables.

A simplistic model suggesting that intelligence (as measured by four questions) can predict academic performance (as measured by SAT, ACT, and high school GPA) is shown above (top right). In SEM diagrams, latent variables are commonly shown as ovals and observed variables as rectangles. The diagram above shows how error (e) influences each intelligence question and the SAT, ACT, and GPA scores, but does not influence the latent variables. SEM provides numerical estimates for each of the parameters (arrows) in the model to indicate the strength of the relationships. Thus, in addition to testing the overall theory, SEM therefore allows the researcher to diagnose which observed variables are good indicators of the latent variables.

Various methods in structural equation modelling have been used in the sciences, business, and other fields. Criticism of SEM methods often addresses pitfalls in mathematical formulation, weak external validity of some accepted models and philosophical bias inherent to the standard procedures.



Analysis: Model specification

Two main components of models are distinguished in SEM: the *structural model* showing potential causal dependencies between endogenous and exogenous variables, and the *measurement model* showing the relations between latent variables and their indicators. Exploratory and confirmatory factor analysis models, for example, contain only the measurement part, while path diagrams can be viewed as SEMs that contain only the structural part.

In specifying pathways in a model, the modeller can posit two types of relationships: (1) *free* pathways, in which hypothesized causal (in fact counterfactual) relationships between variables are tested, and therefore are left 'free' to vary, and (2) relationships between variables that already have an estimated relationship, usually based on previous studies, which are 'fixed' in the model.

A modeller will often specify a set of theoretically plausible models in order to assess whether the model proposed is the best of the set of possible models. Not only must the modeller account for the theoretical reasons for building the model

as it is, but the modeller must also take into account the number of data points and the number of parameters that the model must estimate to identify the model. An identified model is a model where a specific parameter value uniquely identifies the model, and no other equivalent formulation can be given by a different parameter value. A data point is a variable with observed scores, like a variable containing the scores on a question or the number of times respondents buy a car. The parameter is the value of interest, which might be a regression coefficient between the exogenous and the endogenous variable or the factor loading (regression coefficient between an indicator and its factor). If there are fewer data points than the number of estimated parameters, the resulting model is "unidentified", since there are too few reference points to account for all the variance in the model. The solution is to constrain one of the paths to zero, which means that it is no longer part of the model.

Estimation of free parameters

Parameter estimation is done by comparing the actual covariance matrices representing the relationships between variables and the estimated covariance matrices of the best fitting model. This is obtained through numerical maximization via expectation–maximization of a *fit criterion* as provided by maximum likelihood estimation, quasi-maximum likelihood estimation, weighted least squares or asymptotically distribution-free methods. This is often accomplished by using a specialized SEM analysis program, of which several exist.

Assessment of model and model fit

Having estimated a model, analysts will want to interpret the model. Estimated paths may be tabulated and/or presented graphically as a path model. The impact of variables is assessed using path tracing rules.

ISBN: 978-93-85822-78-0

It is important to examine the "fit" of an estimated model to determine how well it models the data. This is a basic task in SEM modelling: forming the basis for accepting or rejecting models and, more usually, accepting one competing model over another. The output of SEM programs includes matrices of the estimated relationships between variables in the model. Assessment of fit essentially calculates how similar the predicted data are to matrices containing the relationships in the actual data.

Formal statistical tests and fit indices have been developed for these purposes. Individual parameters of the model can also be examined within the estimated model in order to see how well the proposed model fits the driving theory. Most, though not all, estimation methods make such tests of the model possible.

Of course as in all statistical hypothesis tests, SEM model tests are based on the assumption that the correct and complete relevant data have been modelled. In the SEM literature, discussion of fit has led to a variety of different recommendations on the precise application of the various fit indices and hypothesis tests.

There are differing approaches to assessing fit. Traditional approaches to modelling start from a null hypothesis, rewarding more parsimonious models (i.e. those with fewer free parameters), to others such as AIC that focus on how little the fitted values deviate from a saturated model (i.e. how well they reproduce the measured values), taking into account the number of free parameters used. Because different measures of fit capture different elements of the fit of the model, it is appropriate to report a selection of different fit measures. Guidelines (i.e., "cut off scores") for interpreting fit measures, including the ones listed below, are the subject of much debate among SEM researchers..

Some of the more commonly used measures of fit include:
Chi-squared

A fundamental measure of fit used in the calculation of many other fit measures. Conceptually it is a function of the sample size and the difference between the observed covariance matrix and the model covariance matrix.

Akaike information criterion (AIC)

A test of relative model fit: The preferred model is the one with the lowest AIC value.

where k is the number of parameters in the statistical model, and L is the maximized value of the likelihood of the model.

Root Mean Square Error of Approximation (RMSEA)

Fit index where a value of zero indicates the best fit.While the guideline for determining a "close fit" using RMSEA is highly contested, most researchers concur that an RMSEA of .1 or more indicates poor fit.

Standardized Root Mean Residual (SRMR)

The SRMR is a popular absolute fit indicator. Hu and Bentler (1999) suggested .08 or smaller as a guideline for good fit. Kline (2011) suggested .1 or smaller as a guideline for good fit.

Comparative Fit Index (CFI)

In examining baseline comparisons, the CFI depends in large part on the average size of the correlations in the data. If the average correlation between variables is not high, then the CFI will not be very high. A CFI value of .95 or higher is desirable.[19]

For each measure of fit, a decision as to what represents a good-enough fit between the model and the data must reflect other contextual factors such as sample size, the ratio of indicators to factors, and the overall complexity of the model. For example, very large samples make the Chi-squared test overly sensitive and more likely to indicate a lack of model-data fit. [20])

Model modification

The model may need to be modified in order to improve the fit, thereby estimating the most likely relationships between variables. Many programs provide modification indices which may guide minor modifications. Modification indices report the change in χ^2 that result from freeing fixed parameters: usually, therefore adding a path to a model which is currently set to zero. Modifications that improve model fit may be flagged as potential changes that can be made to the model. Modifications to a model, especially the structural model, are changes to the theory claimed to be true. Modifications therefore must make sense in terms of the theory being tested, or be acknowledged as limitations of that theory. Changes to measurement model are effectively claims that the items/data are impure indicators of the latent variables specified by theory.

Models should not be led by MI, as Maccallum (1986) demonstrated: "even under favorable conditions, models arising from specification searches must be viewed with caution."

Sample size and power

While researchers agree that large sample sizes are required to provide sufficient statistical power and precise estimates using SEM, there is no general consensus on the appropriate method for determining adequate sample size. Generally, the considerations for determining sample size include the number of observations per parameter, the number of observations required for fit indexes to perform adequately, and the number of observations per degree of freedom. Researchers have proposed guidelines based on simulation studies (Chou &Bentler, 1995), professional experience (Bentler and Chou, 1987), and mathematical formulas (MacCallum, Browne, and Sugawara, 1996; Westland, 2010).

Sample size requirements to achieve a particular significance and power in SEM hypothesis testing are similar for the same model when any of the three algorithms (PLS-PA, LISREL or systems of regression equations) are used for testing.

Interpretation and communication

The set of models are then interpreted so that claims about the constructs can be made, based on the best fitting model.

Caution should always be taken when making claims of causality even when experimentation or time-ordered studies have been done. The term *causal model* must be understood to mean "a model that conveys causal assumptions", not necessarily a model that produces validated causal conclusions. Collecting data at multiple time points and using an experimental or quasi-experimental design can help rule out certain rival hypotheses but even a randomized experiment cannot rule out all such threats to causal inference. Good fit by a model consistent with one causal hypothesis invariably entails equally good fit by another model consistent with an opposing causal hypothesis. No research design, no matter how clever, can help distinguish such rival hypotheses, save for interventional experiments.[12]

As in any science, subsequent replication and perhaps modification will proceed from the initial finding.

Project implementation: 1. Extraction and analysis of latent properties of project in terms of function and impact.

2. Integration of mutually conflicting properties into a common agenda and mutually resonating performance. Thus, it eliminates more error and ensures better harmony and co-operation in project function.

13. MULTIVARIATE ANALYSIS OF VARIANCE – MANOVA

Concept: This technique examines the relationship between several categorical independent variables and two or more metric dependent variables. Whereas analysis of variance (ANOVA) assesses the differences between groups (by using T tests for two means and F tests between three or more means), MANOVA examines the dependence relationship between a set of dependent measures across

a set of groups. Typically this analysis is used in experimental design, and usually a hypothesized relationship between dependent measures is used. This technique is slightly different in that the independent variables are categorical and the dependent variable is metric. Sample size is an issue, with 15-20 observations needed per cell. However, too many observations per cell (over 30) and the technique loses its practical significance. Cell sizes should be roughly equal, with the largest cell having less than 1.5 times the observations of the smallest cell. That is because, in this technique, normality of the dependent variables is important. The model fit is determined by examining mean vector equivalents across groups. If there is a significant difference in the means, the null hypothesis can be rejected and treatment differences can be determined. A multivariate analysis of variance (MANOVA) could be used to test this hypothesis. Instead of a uni-variate *F* value, we would obtain a multivariate *F* value (Wilks' Lambda*?*) based on a comparison of the error variance/covariance matrix and the effect variance/covariance matrix. Although we only mention Wilks' Lambda.Here, there are other statistics that may be used, including Hotelling's trace and Pillai's criterion. MANOVA is useful in experimental situations where at least some of the independent variables are manipulated. It has several advantages over ANOVA.

(a) By measuring several dependent variables in a single experiment, there is a better chance of discovering which factor is truly important.

(b) It can protect against Type I errors that might occur if multiple ANOVA's were conducted independently. Additionally, it can reveal differences not discovered by ANOVA tests.

However, there are several cautions as well.

(a) It is a substantially more complicated design than ANOVA, and therefore there can be some ambiguity about which independent variable affects each dependent variable. Thus, the observer must make many potentially subjective assumptions.

(b) Moreover, one degree of freedom is lost for each dependent variable that is added. The gain of power obtained from decreased SS error may be offset by the loss in these degrees of freedom.

(c) Finally, the dependent variables should be largely uncorrelated. If the dependent variables are highly correlated, there is little advantage in including more than one in the test given the resultant loss in degrees of freedom. Under these circumstances, use of a single ANOVA test would be preferable.

Assumptions

Normal Distribution: - The dependent variable should be normally distributed within groups. Overall, the *F* test is robust to non-normality, if the non-normality is caused by skewness rather than by outliers (outliers are values that are very low or very high as compared to the most values in the data set). Tests for outliers should be run before performing a MANOVA, and outliers should be transformed or removed.

Linearity - MANOVA assumes that there are linear relationships among all pairs of dependent variables, all pairs of covariates, and all dependent variable-covariate pairs in each cell. Therefore, when the relationship deviates from linearity, the power of the analysis will be compromised.

Homogeneity of Variances: - Homogeneity of variances assumes that the dependent variables exhibit equal levels of variance across the range of predictor variables. Remember that the error variance is computed (SS error) by adding up the sums of squares within each group. If the variances in the two groups are different from each other, then adding the two together is not appropriate, and will not yield an estimate of the common within-group variance.

Homogeneity of Variances and Co-variances: - In multivariate designs, with multiple dependent measures, the homogeneity of variances assumption described earlier also applies. However, since there are multiple dependent variables, it is

also required that their inter-correlations (co-variances) are homogeneous across the cells of the design. There are various specific tests of this assumption.

Special Cases

Two special cases arise in MANOVA, the inclusion of within-subjects independent variables and unequal sample sizes in cells. Unequal sample sizes - As in ANOVA, when cells in a factorial MANOVA have different sample sizes, the sum of squares for effect plus error does not equal the total sum of squares. This causes tests of main effects and interactions to be correlated. SPSS offers and adjustment for unequal sample sizes in MANOVA.

Within-subjects design - Problems arise if the researcher measures several different dependent variables on different occasions. This situation can be viewed as a within-subject independent variable with as many levels as occasions. Or, it can be viewed as a separate dependent variables for each occasion.

Additional Limitations

Outliers - Like ANOVA, MANOVA is extremely sensitive to outliers. Outliers may produce either a Type I or Type II error and give no indication as to which type of error is occurring in the analysis. There are several programs available to test for unilabiate and multivariate outliers. Multi-co linearity and Singularity - When there is high correlation between dependent variables, one dependent variable becomes a near-linear combination of the other dependent variables. Under such circumstances, it would become statistically redundant and suspect to include both combinations.

Project implementation:1. In many ecological or biological studies, the variables are not independent at all. Many times they have strong actual or potential interactions, inflating the error even more highly. In many cases where multiple

ANOVAs were done, MANOVA was actually the more appropriate test.

ISBN: 978-93-85822-78-0

14. ARTIFICIAL NEURAL NETWORK ANALYSIS

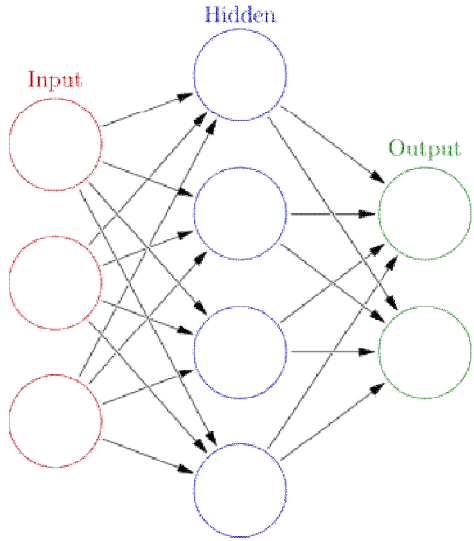
Concept: Artificial neural networks (**ANN**) or **connectionist systems** are computing systems vaguely inspired by the biological neural networks that constitute animal brains. The neural network itself isn't an algorithm, but rather a framework for many different machine learning algorithms to work together and process complex data inputs. Such systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules. For example, in image recognition, they might learn to identify images that contain cats by analyzing example images that have been manually labeled as "cat" or "no cat" and using the results to identify cats in other images. They do this without any prior knowledge about cats, e.g., that they have fur, tails, whiskers and cat-like faces. Instead, they automatically generate identifying characteristics from the learning material that they process.

An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal from one artificial neuron to another. An artificial neuron that receives a signal can process it and then signal additional artificial neurons connected to it.

In common ANN implementations, the signal at a connection between artificial neurons is a real number, and the output of each artificial neuron is computed by some non-linear function of the sum of its inputs. The connections between artificial neurons are called 'edges'. Artificial neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Artificial neurons may have a threshold such that the signal is only sent if the aggregate signal crosses that threshold. Typically, artificial neurons are aggregated into layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the

first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times.

The original goal of the ANN approach was to solve problems in the same way that a human brain would. However, over time, attention moved to performing specific tasks, leading to deviations from biology. Artificial neural networks have been used on a variety of tasks, including computer vision, speech recognition, machine translation, social network filtering, playing board and video games and medical diagnosis.



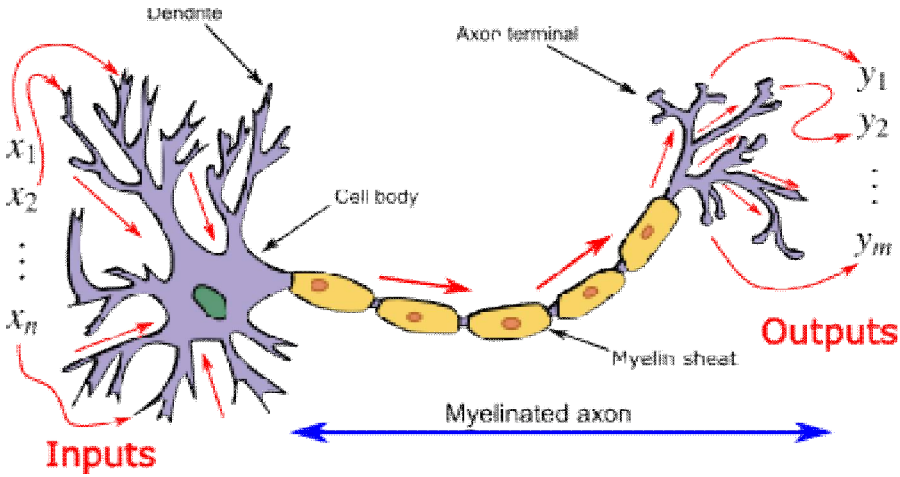
Analysis: An *artificial neural network* is a network of simple elements called *artificial neurons*, which receive input, change their internal state (*activation*) according to that input, and produce output depending on the input and activation.

An artificial neuron mimics the working of a biophysical neuron with inputs and outputs, but is not a biological neuron model.

The *network* forms by connecting the output of certain neurons to the input of other neurons forming a directed, weighted graph. The weights as well as the functions that compute the activation can be modified by a process called *learning* which is governed by a *learning rule*.

Components of an artificial neural network

Neurons



A neuron with label receiving an input from predecessor neurons consists of the following components:

an *activation* - depending on a discrete time parameter,

possibly a *threshold*- which stays fixed unless changed by a learning function,

an *activation function*- that computes the new activation at a given time from and the net input giving rise to the relation

and an *output function*- computing the output from the activation

Often the output function is simply the Identity function.

An *input neuron* has no predecessor but serves as input interface for the whole network. Similarly an *output neuron* has no successor and thus serves as output interface of the whole network.

Connections, weights and biases

The *network* consists of connections, each connection transferring the output of a neuron to the input of a neuron. In this sense is the predecessor of and is the successor of . Each connection is assigned a weight. Sometimes a bias term

ISBN: 978-93-85822-78-0

added to total weighted sum of inputs to serve as threshold to shift the activation function.

Propagation function

The *propagation function* computes the *input* to the neuron from the outputs of predecessor neurons and typically has the form.

When a bias value added with the function, the above form changes to following where is a bias.

Learning rule

The *learning rule* is a rule or an algorithm which modifies the parameters of the neural network, in order for a given input to the network to produce a favoured output. This *learning* process typically amounts to modifying the weights and thresholds of the variables within the network.

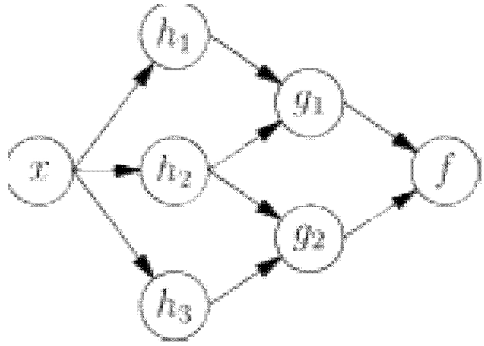
Neural networks as functions

Neural network models can be viewed as simple mathematical models defining a function or a distribution over or both and. Sometimes models are intimately associated with a particular learning rule. A common use of the phrase "ANN model" is really the definition of a *class* of such functions (where members of the class are obtained by varying parameters, connection weights, or specifics of the architecture such as the number of neurons or their connectivity).

Mathematically, a neuron's network function is defined as a composition of other functions that can further be decomposed into other functions. This can be conveniently represented as a network structure, with arrows depicting the dependencies between functions. A widely used type of composition is the *nonlinear weighted sum*, where , where (commonly referred to as the activation function is some predefined function, such as the hyperbolic tangent or sigmoid function or softmax function or rectifier function. The important characteristic of the activation function is that it provides a smooth transition as input values

ISBN: 978-93-85822-78-0

change, i.e. a small change in input produces a small change in output. The following refers to a collection of functions as a vector .



ANN dependency graph

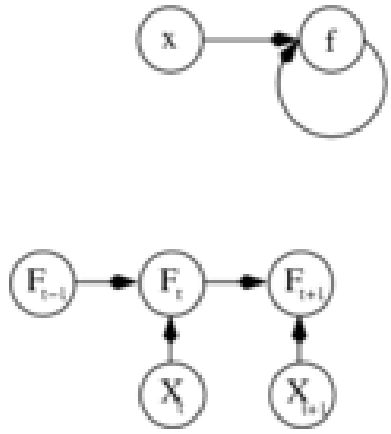
This figure depicts such a decomposition of with dependencies between variables indicated by arrows. These can be interpreted in two ways.

The first view is the functional view: the input is transformed into a 3-dimensional vector , which is then transformed into a 2-dimensional vector , which is finally transformed into . This view is most commonly encountered in the context of optimization.

The second view is the probabilistic view: the random variable depends upon the random variable which depends upon which depends upon the random variable . This view is most commonly encountered in the context of graphical models.

The two views are largely equivalent. In either case, for this particular architecture, the components of individual layers are independent of each other (e.g., the components of are independent of each other given their input). This naturally enables a degree of parallelism in the implementation.

ISBN: 978-93-85822-78-0



Two separate depictions of the recurrent ANN dependency graph

Networks such as the previous one are commonly called feed-forward, because their graph is a directed acyclic graph. Networks with cycles are commonly called recurrent. Such networks are commonly depicted in the manner shown at the top of the figure, where is shown as being dependent upon itself. However, an implied temporal dependence is not shown.

Learning

The possibility of learning has attracted the most interest in neural networks. Given a specific *task* to solve, and a class of functions learning means using a set of observations to find which solves the task in some optimal sense.

This entails defining a cost function such that, for the optimal solution , i.e., no solution has a cost less than the cost of the optimal solution .

The cost function is an important concept in learning, as it is a measure of how far away a particular solution is from an optimal solution to the problem to be solved. Learning algorithms search through the solution space to find a function that has the smallest possible cost.

For applications where the solution is data dependent, the cost must necessarily be a function of the observations, otherwise the model would not relate to the data. It is frequently defined as a statistic to which only approximations can

ISBN: 978-93-85822-78-0

be made. As a simple example, consider the problem of finding the model , which minimizes , for data pairs drawn from some distribution . In practical situations we would only have samples from and thus, for the above example, we would only minimize . Thus, the cost is minimized over a sample of the data rather than the entire distribution.

When some form of online machine learning must be used, where the cost is reduced as each new example is seen. While online machine learning is often used when is fixed, it is most useful in the case where the distribution changes slowly over time. In neural network methods, some form of online machine learning is frequently used for finite datasets.

Choosing a cost function

While it is possible to define an ad hoc cost function, frequently a particular cost (function) is used, either because it has desirable properties (such as convexity) or because it arises naturally from a particular formulation of the problem (e.g., in a probabilistic formulation the posterior probability of the model can be used as an inverse cost). Ultimately, the cost function depends on the task.

Back propagation

A DNN can be discriminatively trained with the standard back propagation algorithm. Back propagation is a method to calculate the gradient of the loss function (produces the cost associated with a given state) with respect to the weights in an ANN.

The basics of continuous back propagation were derived in the context of control theory by Kelley in 1960 and by Bryson in 1961, using principles of dynamic programming. In 1962, Dreyfus published a simpler derivation based only on the chain rule. Bryson and Ho described it as a multi-stage dynamic system optimization method in 1969. In 1970, Linnainmaa finally published the general method for automatic differentiation (AD) of discrete connected networks

of nested differentiable functions. This corresponds to the modern version of back propagation which is efficient even when the networks are sparse. In 1973, Dreyfus used back propagation to adapt parameters of controllers in proportion to error gradients. In 1974, Werbos mentioned the possibility of applying this principle to Artificial neural networks, and in 1982, he applied Linnainmaa's AD method to neural networks in the way that is widely used today. In 1986, Rumelhart, Hinton and Williams noted that this method can generate useful internal representations of incoming data in hidden layers of neural networks. In 1993, Wan was the first to win an international pattern recognition contest through back propagation.

The weight updates of back propagation can be done via stochastic gradient descent using the following equation: where, is the learning rate, is the cost (loss) function and a stochastic term. The choice of the cost function depends on factors such as the learning type (supervised, unsupervised, reinforcement, etc.) and the activation function. For example, when performing supervised learning on a multiclass classification problem, common choices for the activation function and cost function are the softmax function and cross entropy function, respectively. The softmax function is defined as where represents the class probability (output of the unit) and and represent the total input to units and of the same level respectively. Cross entropy is defined as where represents the target probability for output unit and is the probability output for after applying the activation function.

These can be used to output object bounding boxes in the form of a binary mask. They are also used for multi-scale regression to increase localization precision. DNN-based regression can learn features that capture geometric information in addition to serving as a good classifier. They remove the requirement to explicitly model parts and their relations. This helps to broaden the variety of objects that can be learned. The model consists of multiple layers, each

of which has a rectified linear unit as its activation function for non-linear transformation. Some layers are convolutional, while others are fully connected. Every convolutional layer has an additional max pooling. The network is trained to minimize L^2 error for predicting the mask ranging over the entire training set containing bounding boxes represented as masks.

Alternatives to back propagation include Extreme Learning Machines, "No-prop" networks, training without backtracking, "weightless" networks, and non-connectionist neural networks.

Learning paradigms

The three major learning paradigms each correspond to a particular learning task. These are supervised learning, unsupervised learning and reinforcement learning.

Supervised learning

Supervised learning uses a set of example pairs and the aim is to find a function

in the allowed class of functions that matches the examples. In other words, we wish to infer the mapping implied by the data; the cost function is related to the mismatch between our mapping and the data and it implicitly contains prior knowledge about the problem domain.

A commonly used cost is the mean-squared error, which tries to minimize the average squared error between the network's output, , and the target value over all the example pairs. Minimizing this cost using gradient descent for the class of neural networks called multilayer perceptrons (MLP), produces the back propagation algorithm for training neural networks.

Tasks that fall within the paradigm of supervised learning are pattern recognition (also known as classification) and regression (also known as function approximation). The supervised learning paradigm is also applicable to sequential data (e.g., for hand writing, speech and gesture recognition). This can be thought of

as learning with a "teacher", in the form of a function that provides continuous feedback on the quality of solutions obtained thus far.

Unsupervised learning

In unsupervised learning, some data is given and the cost function to be

minimized, that can be any function of the data and the network's output, .

The cost function is dependent on the task (the model domain) and any *a priori* assumptions (the implicit properties of the model, its parameters and the observed variables).

As a trivial example, consider the model where is a constant and the cost . Minimizing this cost produces a value of that is equal to the mean of the data. The cost function can be much more complicated. Its form depends on the application: for example, in compression it could be related to the mutual information between and , whereas in statistical modelling, it could be related to the posterior probability of the model given the data (note that in both of those examples those quantities would be maximized rather than minimized).

Tasks that fall within the paradigm of unsupervised learning are in general estimation problems; the applications include clustering, the estimation of statistical distributions, compression and filtering.

Reinforcement learning

See also: Stochastic control

In reinforcement learning, data are usually not given, but generated by an agent's interactions with the environment. At each point in time, the agent performs an action and the environment generates an observation and an instantaneous cost , according to some (usually unknown) dynamics. The aim is to discover a policy for selecting actions that minimizes some measure of a long-term cost, e.g., the expected cumulative cost. The environment's dynamics and the long-term cost for each policy are usually unknown, but can be estimated.

ISBN: 978-93-85822-78-0

More formally the environment is modelled as a Markov decision process (MDP) with states and actions with the following probability distributions: the instantaneous cost distribution , the observation distribution and the transition , while a policy is defined as the conditional distribution over actions given the observations. Taken together, the two then define a Markov chain (MC). The aim is to discover the policy (i.e., the MC) that minimizes the cost.

Artificial neural networks are frequently used in reinforcement learning as part of the overall algorithm. Dynamic programming was coupled with Artificial neural networks (giving neuro-dynamic programming) by Bertsekas and Tsitsiklis and applied to multi-dimensional nonlinear problems such as those involved in vehicle routing, natural resources management or medicine because of the ability of Artificial neural networks to mitigate losses of accuracy even when reducing the discretization grid density for numerically approximating the solution of the original control problems.

Tasks that fall within the paradigm of reinforcement learning are control problems, games and other sequential decision making tasks.

Learning algorithms

Training a neural network model essentially means selecting one model from the set of allowed models (or, in a Bayesian framework, determining a distribution over the set of allowed models) that minimizes the cost. Numerous algorithms are available for training neural network models; most of them can be viewed as a straightforward application of optimization theory and statistical estimation.

Most employ some form of gradient descent, using back propagation to compute the actual gradients. This is done by simply taking the derivative of the cost function with respect to the network parameters and then changing those parameters in a gradient-related direction. Back propagation training algorithms fall into three categories: steepest descent (with variable learning rate and

momentum, resilient back propagation); quasi-Newton (Broyden-Fletcher-Goldfarb-Shanno, one step secant); Levenberg-Marquardt and conjugate gradient (Fletcher-Reeves update, Polak-Ribiére update, Powell-Beale restart, scaled conjugate gradient) Evolutionary methods, gene expression programming, simulated annealing, expectation-maximization, non-parametric methods and particle swarm optimization are other methods for training neural networks.

Convergent recursive learning algorithm

This is a learning method specially designed for cerebellar model articulation controller (CMAC) neural networks. In 2004, a recursive least squares algorithm was introduced to train CMAC neural network online. This algorithm can converge in one step and update all weights in one step with any new input data. Initially, this algorithm had computational complexity of $O(N^3)$. Based on QR decomposition, this recursive learning algorithm was simplified to be $O(N)$.

Project implementation: It helps in estimation of the possibility of success or probability of failure in accomplishment of project goal.

Because of their ability to reproduce and model nonlinear processes, Artificial neural networks have found many applications in a wide range of disciplines.

Application areas include system identification and control (vehicle control, trajectory prediction, process control, natural resource management), quantum chemistry, game-playing and decision making (backgammon, chess, poker), pattern recognition (radar systems, face identification, signal classification, [208] object recognition and more), sequence recognition (gesture, speech, handwritten and printed text recognition), medical diagnosis, finance(e.g. automated trading systems), data mining, visualization, machine translation, social network filtering and e-mail spam filtering.

Artificial neural networks have been used to diagnose cancers, including lung cancer, prostate cancer, colorectal cancer and to distinguish highly invasive cancer cell lines from less invasive lines using only cell shape information.

Artificial neural networks have been used to accelerate reliability analysis of infrastructures subject to natural disasters.

Artificial neural networks have also been used for building black-box models in geo-science: hydrology, ocean modelling and coastal engineering, and geomorphology

15. SIMULATION AND MODELLING

Concept: Simulation modelling is the process of creating and analyzing a digital prototype of a physical model to predict its performance in the real world. Simulation modelling is used to help designers and engineers understand whether, under what conditions, and in which ways a part could fail and what loads it can withstand. Simulation modelling can also help to predict fluid flow and heat transfer patterns. It analyses the approximate working conditions by applying the simulation software.

Analysis: Simulation modelling follows a process much like this:

Use a 2D or 3D CAD tool to develop a virtual model, also known as a digital prototype, to represent a design.

Generate a 2D or 3D mesh for analysis calculations. Automatic algorithms can create finite element meshes, or users can create structured meshes to maintain control over element quality.

Define finite element analysis data (loads, constraints, or materials) based on analysis type (thermal, structural, or fluid). Apply boundary conditions to the model to represent how the part will be restrained during use.

Perform finite element analysis, review results, and make engineering judgments based on results.

Project implementation: Simulation modelling allows designers and engineers to avoid repeated building of multiple physical prototypes to analyze designs for new

or existing parts. Before creating the physical prototype, users can investigate many digital prototypes. Using the technique, they can:

Optimize geometry for weight and strength

Select materials that meet weight, strength, and budget requirements

Simulate part failure and identify the loading conditions that cause them

Assess extreme environmental conditions or loads not easily tested on physical prototypes, such as earthquake shock load

Verify hand calculations

Validate the likely safety and survival of a physical prototype before

Simulation modelling software programs

Any Logic, Abaqus, ANSYS, Autodesk Simulation Mechanical, Autodesk Simulation CFD, Autodesk Inventor Professional, Auto Form, COMSOL, CONSELF CFD on Cloud, FEA Tool Multiphysics, Gold Sim Pro, Insightmaker, LAMMPS, Matlab, Nastran, Nogrid points and Nogrid CAD Compass, Patran, Siemens NX CAE, SimScale, Solidworks Simulation, Simio.

Software and tools

There are an enormous number of software packages and other tools for multivariate analysis, including:

JMP (statistical software), Mini Tab, Calc, PSPP, R, SAS (software), SciPy for Python, SPSS, Stata, STATISTICA, The Unscrambler, Warp PLS, Smart PLS, MATLAB, Eviews.

References:

1. *Multivariate analysis of variance*. Available from: https://www.researchgate.net/publication/237227650_MULTIVARIATE_ANALYSIS_OF_VARIANCE [accessed Nov 23 2018].
2. Pal P, Acharya S.K., Biswas A, Research methodology design, tools and techniques

3. Extracted from https://en.wikipedia.org/wiki/Stepwise_regression
4. Extracted from https://en.wikipedia.org/wiki/Multivariate_statistics
5. Extracted from https://en.wikipedia.org/wiki/Factor-analysis
6. Extracted from https://en.wikipedia.org/wiki/Correspondence_analysis
7. Extracted from https://en.wikipedia.org/wiki/multidimensional_scaling
8. Extracted from https://en.wikipedia.org/wiki/Simulation_modeling
9. Extracted from https://en.wikipedia.org/wiki/conjoint-analysis
10. Extracted from https://en.wikipedia.org/wiki/Structural-equation_modeling
11. Extracted from https://en.wikipedia.org/wiki/Artificial_neural_network
12. Extracted from https://project-management knowledge.com/definitions/r/regression-analysis