A proposed expert system for selecting exploratory factor analysis procedures

Samia El Azzab[*]
Samy Abu Naser[**]
Ossama Sulisel[***]

ملخص البحث

وضع نموذج لعملية اختيار الاجراء المناسب في التحليل العاملي الاستكشافي هي أحد المسائل الهامة التي تحتاج الى خبرة خاصة ولكنها لم تجذب انتباه الباحثين المهتمين بتطبيقات النظم الخبيرة ويعود ذلك للطبيعة المعقدة لهذه العملية. أحد الأهداف الهامة لهذا البحث هو تطوير منهجية لنمذجة هذه العملية و بناء نظام أولي خبير لمساعدة الباحثين في استخدام التحليل العاملي بغض النظر عن مستوي خبرتهم فيه. و يهدف البحث أيضا الي تحديد مدي ملائمة استخدام النظم الخبيرة لمثل هذه العملية المعقدة. في هذا البحث تم وضع منهجية واستخدمت في بناء نظام أولي خبير وتم اختبار هذا النظام وكانت نتائج الاختبارات مشجعة جدا".

Abstract

Modeling the selection of the appropriate exploratory factor analysis (EFA) is one very important area of expertise that did not attract any attention from researchers interested in the applications of expert systems (ES) technology. This probably due to the complex and highly demanding nature of the (EFA) procedure selection problem. Developing a methodology for modeling the (EFA) procedure selection problem and building a prototype expert system to help novice and experienced researchers is one important aim of this research. Another goal of this research is to examine the appropriateness of ES as a supporting development paradigm for modeling such a complex process. A methodology was developed as well as a prototype expert system. The system was tested and the results obtained were very encouraging.

[*] Professor Samia El Azzab, Mathematics Dept. Ain Shams, Egypt.

[**] Assistant Professor Samy S. Abu Naser, Computer Science Dept. Al-Azhar University Gaza.

[***] T. Ossama S. Sulisel, Educational Computer Dept. Arafat University, Gaza.

Introduction

Since it was first commercially introduced in the beginning of the past decade, Expert systems (ES) technology has been applied to numerous problem domains including statistical analysis (Turban, 1995). The expertise of one or more expert analysts in a specific statistical analysis problem domain can be captured in a computer program to achieve several critical goals such as providing consistent and timely statistical knowledge to users.

Selecting the appropriate exploratory factor analysis (EFA) procedures is one very important area of expertise that did not attract any attention from researchers interested in the application of expert systems to advanced statistics. This is probably due to the complex and highly demanding nature of the (EFA) procedure selection problem.

Although there are many analytical models that are available to help in conducting EFA, no attempt has been made to develop a decision aid to help researchers in selecting the appropriate (EFA) procedure given a research assumption. The importance of this problem stems from the fact that some times to avoid confusion, a researcher may choose the default settings of the factor analysis package he/she is using to conduct an EFA, and since factor analysis is a multistage process and each stage contains several procedures each based on different settings and assumptions, using the default setting usually leads to invalid results and thus, invalid interpretation of the analysis results. In the following sections we will give an over view of Exploratory factor analysis, examine the different aspects of modeling the (EFA) procedure sele ction problem, and develop a methodology to be used in building a prototype expert system to help experienced and novice researchers in the selection process. We will also examine the appropriateness of expert systems as a supporting development paradigm for modeling such a choice problem.

Expert Systems And Statistical Analysis

There has been a wide spread use of expert system (ES) in the area of statistical analysis. These applications can be classified according to two different levels of abstraction: the general problem domain discussed and the particular statistical technique investigated. Under the generic category, ES have been used for diagnosis, selection, interpretation, control, prediction as well as many other domains. For example, (Kumar & Cheng, 1988) gave an illustration of a system that detects trends from historical data and selects an appropriate forecasting model. (Mellichamp, 1987) developed an expert system that interprets the results of simulation experiments. (Remus & Kotteman, 1986) describe a hypothetical statistical expert system for predicting the appropriate statistical technique for the specific problem at hand. (White, 1995) describes a prototype expert system for choosing statistical tests. The system asks the user a set of questions and builds up a picture until it is able to suggest the appropriate statistical test. (Witten & Frant, 2000) describe an expert system based on the Weka's Java algorithm to discern meaningful patterns in the data, and how to adapt them for specialized data mining applications.

In the specific technique category, ES has been used mainly to Model regression and MANOVA techniques (Pregibon & Gale, 1984; Hand, 1986; Cunningham & Holmes, 1999; Hunt & Jorgnsen, 1999).

An over view of exploratory factor Analysis

Researchers are often concerned with identifying constructs and investigating relationships among them. Constructs are theoretical concepts or abstractions that help us explain and organize our environment (Pedhazur & Schmelkin, 1991). Because constructs are theoretical abstractions, they can not be directly observed. Instead, they must be indirectly defined by their observed manifestation. For example, a researcher may be interested in studying test anxiety. Although test anxiety itself can not be observed, it might be hypothesized that indicators of it are nails biting, trembling hands, inability to concentrate on the test, or other such physical or psychological symptoms.

Factor analysis is often a useful method for investigating constructs because it provides a model that links the observations or manifestations of the process to the theories and constructs through which we interpret and understand them (Ecob & Cuttance, 1987). More specifically, factor analysis is an analytic technique used to express unobserved (latent) variables through defining and measuring their observed indicators. After these relation ships are established, it is possible to investigate relation ships among the underlying variables (called factors) Further, or to examine the relationships between the set of factors and other processes the researcher may be interested in studying.

In general, there are two basic approaches to the investigation of under lying factors: The exploratory approach and the confirmatory approach. The exploratory approach is useful when the researchers intent is to identify a set of latent factor that may be responsible for relationships (i.e., correlation or covariance) among a set of observed variables. In this manner, the structure of the data is simplified by exploring which sub sets of observed variables maybe grouped together by a common, underlying factor. In contrast, when are searcher already has a specific theoretical model in mind and can specify the relationships among the underlying factors and their observed indicators before hand, he or she can attempt to "confirm" its existence with the data by using confirmatory factor analysis. (Heck, 1998)

Exploratory factor analysis has two main stages: extraction and rotation. The extraction stage is concerned with determining the number of factor needed to explain the variance in the data and to reproduce the correlation matrix. There are several procedures that might be used in such an extraction, the most widely used are: principal components, maximum likelihood, and generalized least squares. After determining the number of factors (i.e., the ability of the selected factor structure to reproduce the observed correlations has been examined), the Rotation stage attempts an interpretation of the retained factors. There are two types of rotation: orthogonal rotation and oblique rotation. Orthogonal rotation is performed when the researcher assumes that the factors are uncorrelated, and oblique rotation is performed when the researcher assumes that the factors are intercorrelated. Within each type of rotation there are several procedures. The procedures within each of the factor analysis stages have the same goal (e.g., extraction), however, they do differ in their underlying assumptions and settings (Marcoulides & Hershberger, 1997; Heck, 1998; Loehlin, 1998)

The main interest of this research is to attempt to model the (EFA) procedure selection problem. The confirmatory factor analysis is beyond the scope of this research thus, it will be left for future research. We come now to the EFA procedure selection problem.

The EFA procedure selection problem

As stated earlier, where little is known about the underlying structure of the data conducting an (EFA) may be an important first step in identifying a set of underlying relationships. The major goal of an (EFA) is to extract the minimum number of factors needed to reproduce the variation present in a set of observed variables. EFA is a multistage process involving factor extraction, rotation, and results interpretation. Within each stage, several procedures exist, each with different assumptions and settings, which gives rise to the procedure selection problem (i.e., which procedure should be used? and in what order?). It is so often that a researcher encounters a situation where he/she has to choose from a number of procedures within the same stage. Some times to a void the confusion, a researcher may choose the default settings of the factor and analysis package he/she is using. Consequently, several important assumptions may be overlooked; thus leading to invalid results. Another problem is that because of the exploratory nature of factor analysis, it is quite possible that the researcher may select a certain procedure (according to some criteria and conduct the analysis based on his selection, but the results may suggest a more appropriate procedure with different assumptions, and the researcher has to repeat the analysis guided by the results of the first analysis. For example, in the extraction stage, the researcher may choose a certain criteria for determining the number of factors needed. When the final report is the researcher usually repeats the analysis either to compare his results to other criteria's or to check that the number of factors suggested is sufficient. Another example comes from the rotation stage. The researcher may assume that the factors are correlated and hence perform an oblique rotation, in the final report, the researcher has to check the validity of his assumption by checking the factor correlation matrix in the final report. If the matrix does not show significant correlation between factors, then the researcher has to repeat the analysis using orthogonal rotation. It can be seen from these examples that EFA is usually an iterative process which adds to the complexity of the selection problem. Another problem that adds more complexity to the EFA procedure selection problem is that factor analysis is based on the normality assumption (i.e., the variables are normally distributed), so the researcher must perform a normality check on his data before conducting an EFA. If the data fail to meat the normality requirement the researcher must find a suitable transformation that will normalize the data and then analyze the transformed data.

The above discussion gives a clear view of the complex nature of modeling the EFA procedure selection problem. The selection problem can be summarized as follows:

An EFA is a multistage process conducted by a human expert in the following order: Data normality check, factor extraction, rotation and report analysis. The procedure selection problem is encountered because:

1. Eeach stage contains several procedures each with different assumptions and settings, but all have a common purpose (e.g., data normalization, factor extraction, ..., etc) which is the main source of confusion to obtain both experienced and novice researchers.
2. An EFA is usually an iterative process in which the results of the initial analysis may indicate the need to repeat the analysis with different assumptions or settings.

The above statement will be used as a basis for the development of a methodology for modeling the EFA procedure selection problem as will be shown next.

The proposed methodology:

To model the EFA Procedure selection problem we will divide the EFA into four stages: data normality check, factor extraction, rotation, and report interpretation, within each stage the procedures will be organized according to the assumptions they are based on, and a criteria will be Placed that matches a research assumption with the most suitable procedure. The proposed division helps in:

1. Organizing the knowledge needed to conduct an EFA and interpret the results.
2. Choosing a suitable knowledge representation method to be used in building the knowledge base of the prototype expert system.

3. Developing a decision making tool to aid in the selection problem which enables the system to justify its advice.

We star by examining the first stage. The main goal of the data normality check stage is to ensure that data are normally distributed which is the basic assumption of EFA. The organization of knowledge in this stage will be done by stating the basic assumption of this stage, followed by a criteria for validating this assumption, followed by a set of actions to be taken if the criteria **is not satisfied** as follows:

1. **Assumption:** the variables are normally distributed.
2. **Criteria:** under mild skweness (-1 to +1) or kurtosis (-2 to +2). Variables are considered normal.
3. **action if the criteria is not satisfied:**

Use one of the following transformations (log x, $\sqrt{X}$,or 1/x) to normalize the data and check the criteria after transforming the data.

In the second stage (factor extraction), all procedures have one goal in common which is to extract factors, but they differ in the assumptions they are based on. So the knowledge will be organized in the form of assumption, advice, and justification, as follows:

1- Assumption 1: The determinant of the correlation matrix is equal to zero (i.e. det R=0).

Advice: the choice is limited to one method of extraction "principal components".

Justification: All other extraction methods (procedure) require that det R $\neq 0$.

2-Assumption 2: The user has a predetermined number of factors in mind and would like the extraction method to test the goodness of fit of the proposed number of factors.

Advice: use the maximum likely hood method.

Justification: This method is the only method that allows a goodness of fit test.

3- Assumption 3: The researcher would like to take into account the uniqueness (error of measurement) of each variable, so that variable with high uniqueness is given less weight than those with low uniqueness.

Advice: use the generalized (weighted) least square method.

Justification: In this method, correlations are weighted by the inverse of their uniqueness, so that variables with high uniqueness are given less weight

Note that in this stage the number of factors to be extracted depends on kaizer Eigen value criteria (except for the maximum likely hood method). Other criteria's are only possible in the report stage.

4- Assumption 4: det $R \neq 0$ and assumption 2 and 3 do not hold.

Advice: use principal components

Justification 4: This method can be used to examine the nature of the factor solution of the first analysis the, results of the analysis may suggest new assumptions.

The third stage (rotation) is concerned with simplifying the factor solution obtained from the extraction stage to improve interpretability of the factors. The procedures of this stage are divided into tow categories depending on whether or not the factors are assumed to be correlated. Within each category the procedures have different assumption. So the knowledge about this stage will be organized as the previous stage.

1- Assumption 1: The factors are not correlated

Advice: use orthogonal rotation.

Justification: The procedures in this type of rotation produce factors that are Uncorrelated with one another.

1.a. assumption 1.a: The researcher is interested in a factor solution containing a general factor.

Advice: Use the Quartimax procedure

Justification: This procedure encourages the appearance of a general factor.

Assumption 1.a: The researcher is interested in a voiding solutions containing a general factor.

Advice: Use the varimax procedure.

Justification: This procedure discourages the appearance of a general factor.

Assumption 1.b: In doubt.

Advice: Try both procedures.

Justification: If both procedures arrive at the same solution, well and good. If not, the researcher knows that there is more than one plausible interpretation of the data.

2-Assumption 2: The factors are correlated.

Advice: Use oblique rotation.

Justification: The procedure in this type of rotation produce factors that are correlated.

Assumption 2.a: The researcher is interested in controlling the degree of correlation between factors.

Advice: Use the direct oblimin procedure.

Justification: This procedure allows the researcher to control the degree to which correlation among the factors is encouraged.

Assumption 2.b: The researcher is not interested in controlling the degree of correlation between factors.

Advice: Use the Promax procedure then compare the solution obtained to the solution obtained by the direct oblimin procedure.

Justification: If both procedures arrive at the solution, well and good, otherwise the researcher has to relay on his research assumptions to choose the best fitting solution.

The final stage is 'report interpretation'. This stage is mainly concerned with the determination of the number of factors required to reproduce the variation present in the data. The knowledge needed for this stage involves criteria's for selecting the appropriate number of factors, testing for the goodness of fit of a predetermined number of factors, and interpreting the factor loadings (i.e., determining to which factor a specified variable belongs).

The organization of knowledge in this stage is based on the set of possible procedures selected by the researcher in the previous stages and the different criteria's of factor extraction. The organization will be in the form of advice and Justification as follows:

1-Advice: check the commonalty table and repeat the analysis removing the variables with commonalties less than 0.3.

Justification: The extraction communality of each variable represent the variance of the variable accounted for by the common factors and is always less than 1. The rest is error in measurement. So it's advisable to drop variables with the commonalties less than 0.3 (the error in measurement is too high).

2-Advice: check the "Total variance explained" table. The potential factors are those with initial eigen values >= 1.

Justification: This is the kaizer criterion.

3-Advice: If the user wants to inspect using another criteria try the following: with each variable (or factor) with eigen value >= 1 the table displays the % of variance accounted for by that factor, and the cumulative % of variance accounted for by that factor and all factors proceeding it in the table. Choose the factor that account for the % of variance you consider of practical value to your study (usually % 50 to % 80).

Justification: It is logical to choose the factor that account for as much of the % of variance the researcher considers of practical value.

4-Advice: To further confirm your findings check the values of the cumulative % of the initial eigen values and the cumulative % of extraction sumf squared loadings. If the tow values are not equal or nearly equal, try another extraction method until the tow values are equal or nearly so (if possible).

Justification: In a good factor solution the tow values are equal or nearly so.

5-Advice: If the user has decided on the number of factors that he considers reasonable. it is advisable to repeat the analysis using the maximum likely hood extraction method, and setting the number of factors option to the number he/she wants to test. The report will now contain a chi-sequar test. If the ratio (chi-sequar / degrees of freedam) is less than 2 then the number factors chosen fits other wise the user should increase the number of factors until this criteria is satisfied.

Justification: this criterion allows the user to test the goodness of fit of his obtained factor solution and used by EFA experts in such a situation.

6-Advice: after determining the final number of factor, repeat the analysis setting the number of factors option to that number

Justification: This is necessary to determine the factor loadings (assignment of variables to factors) neglecting the rest the factor the researcher now considers negligible.

7-Advice: If the user did not perform rotation and whishes to determine to which factors the variables belong; the user should inspect the component matrix. This matrix contains the loadings (correlation) of each variable on each factor. The variable belongs to the factor on which it has highest positive loading.

Justification: a factor loading represent the degree to which a certain variable is close to a certain factor. The higher the loading is, the closer the variable is to the factor.

8-Advice: If the user has requested rotation. The factor loadings are reported in a structure matrix (In the case of oblique rotation) or in a rotated components matrix (in the case of orthogonal rotation). Variable belongs to the factor on which it has the highest positive loading.

Justification: a factor loading represent the degree to which a certain variable is close to a certain factor. The higher the loading is, the closer the variable is to the factor.

9-Advice: check the final table in the report. This Table gives the correlations between factors. If the user requested oblique rotation and there was no significant correlation between factors then the analysis should be repeated using orthogonal rotation.

Justification: checking the correlation between factors determines whether or not the rotation method selected is valid. This is very important because, based on invalid assumption the factor solution obtained (factors and their loading) will be misleading.

The above methodology serves as a decision tool that matches research assumptions to the most appropriate procedures based on practical and statistical criteria. It also provides a frame work that organizes the knowledge about EFA scattered across the statistical literature in the form of a set of Justifiable advice in an attempt to model the procedure selection process conducted by a human expert.

In this framework, we feel that the best strategy is to examine as many criteria as possible. If the results are similar, determining the number of factors and their associated loading (i.e., the factor solution) is easy. If the results are different, the researcher should look for some consensus among the criteria.

Never the less, one should remember that the final judgment should be based on the reasonableness of the number of factors and their potential interpretation. That is, a good solution "makes sense" in that it appears to fit what is known about the phenomenon being studied. That is why it is important to remember that theory should be a guide when performing EFA.

Details of the system

A prototype Expert system was developed using VPX (a micro computer shell) to model the selection process as discussed in the previous section. The expert system operates in a dialogue mode. The process starts by asking the end user to answer a set of questions that compose the pre-constructed knowledge base. Based on the replies of the user, a selection is made for a procedure in each of the factor analysis stages.

The system was developed using IF-THEN rules. The rules are based on the methodology discussed in the previous section.

The rules are divided according to the stage they represent knowledge about (i.e., normality check, extraction, rotation, and report interpretation), and associated with each rule is a justification for the advice (or action). The justifications of the system are exactly those given in the previous section. In fact, the rules may concerned as a direct translation of the proposed methodology. Fore example, there is a rule that states the condition for selecting the principal components procedure in the extraction stage as follow.

IF:

det R = 0 (R is the determinant of the correlation matrix)

THEN:

Use the principal components procedure.

Justification:

All other extraction methods require that det R ? 0.

The various procedures of factor analysis are the goals of system, and system uses backward – chaining infrencing. Back ward chaining is a goal driven mechanism, which starts with the goal and works backward searching for arguments that satisfy the specified objective. To illustrate, in order to reach "advice 2" which could be "oblique rotation", the system will search the then part of the rules of the rotation stage to find the first "value" (goal). Once the system locates the first rule that satisfies this advice, the system will try to prove the rules condition and fire the rule and its associated justification.

The system will repeat this process until there are no more goals. The knowledge base of the system contains 40 rules.

The prototype system was tested using pre-solved problems (these problem can be found in the references of this research) and using the statistical package spss 9.0.

The problems used to test the performance of the system cover areas where EFA is most widely used such as politics, education, and psychology (i.e., Lab testing of the systems components). The system arrived at the required result in each of the problems given to it (over 20 problems). To further test the system, it was used to help graduate students in their research projects in the faculty of education in Gaza (i.e., pilot testing). The research projects are concerned with building scales (an area heavily dependent on EFA). The system proved to be a valuable tools to the researchers, saving time and effort and giving accurate advice as reported by the researchers and a set of observing experts in EFA who were asked to monitor the system's performance and the validity of advice given by the system in each EFA stage (i.e., Expert inspection). The results of the tests were encouraging to go a head in building the final version of the system. The exploratory methods used to evaluate the system are based on the methodology proposed by (Iqbal et al., 1999) (i.e., expert inspection, Lab testing of the system components, pilot testing).

Conclusion

There are two goals in this research. The first is that the proposed modeling methodology and prototype expert system seeks to provide consistent and up-to-date EFA procedure selection knowledge both for novice and experienced researchers. The designed system, however, is not proposed to take the place of good intuitive judgement of these users or theory guidance about the problem studied by the researchers. The test results indicate that the system is recommended to be taken as a decision – aiding tool for researchers. The second goal is to examine whether the ES development paradigm can be used for the EFA procedure selection problem.

The results obtained show that it is quite adequate to use expert systems for such an important decision process.

References:

1. Cunningham, S. & Aolmes, G. (1999). Developing innovative applications of machine learning. Pro Southeast Asia Regional Computer confederation conference, Singapore.

2. Ecob, R., & Cuttance, P. (1987) Structural Modeling by example. Cambridge university press.

3. Hand, D. Patterns in Statistical Strategy. In Gale, W. (ed), Artificial intelligence in statistics. Addison-Wesley, Reading, MA, 1986, pp. 355-387.

4. Heck, R. (1998) Factor Analysis: Exploratory and Confirmatory Approaches. In G. Marcoulides (Eds), modern methods for Business Research. NJ: Lawrence Erlbaum associates.

5. Hunt, L. & Jorgensen, M. (1999). Mixture model clustering using the multimix program. Australian and New Zealand Journal of statistics, 41:2, pp. 153-171.

6. Iqbal, A., Oppermann, R. & Kinshuk (1999). A classification of Evaluation methods for intelligent Tutoring System. Software Ergonomie 'aa. Leipzig, pp. 169-181.

7. Kumar, S. and H.cheng. "An expert system framework for forecasting method selection". Proceedings of the Hawali international confrence on system science, 1988, pp.86-95.

8. Loehlin, J. (1998) Lafent variable models, Third edition. NJ: Lawrence Erlbaum associates.

9. Marcoulides, G. & Hersh berger,S. (1997) Multivariate statistical methods, Afirst course. NJ:Lawrence Erlbaum associates.

10. Mellichamp, J. An Expert System For FMS Design. Simulation, 48:5, 1987, pp. 201-209.

11. Pedhazur, E. & Schmelkin, L. (1991) Measurement, design, and analysis: An integrated a proach. Hillsdale, NJ: Lawrence Erlbaum Associates.

12. Pregibon, D. & Gale, W. REX: An Expert System for regression Analysis. CompSTAT, 1984, pp. 242-248.

13. Remus,W. & Kotteman, J. Towards Intelligent Decision Support System: an artificially Intelligent statistician. MIS Quarterly, 10:4, 1986, pp 403-418.

14. White, A. An Expert System for choosing statistical tests. New Review of Applied Expert Systems.

15. Witten, H. &Frank, E. (2000). Data mining: practical machine learning tools and techniques with Java implementations. Morgan Kaufmann, San Francisco.