AN EDIT SCHEME BASED ON MULTIVARIATE DATA ANALYSIS

France Bilocq, J.-M. Berthelot, Statistics Canada
11th floor Coats Bldg, Ottawa, Canada K1A 0T6

Key words: Cross-editing, principal components analysis, orthoblique rotation, correlation matrix.

In order to facilitate the methodological development of the generalized data collection, capture and editing function at Statistics Canada, multivariate data analysis techniques are used to derive an editing scheme for the quantative variables of business surveys. Once the grouping of variables is identified, data editing procedures can be applied in an optimal fashion.

1. INTRODUCTION

The generalized data collection, capture and editing function is defined as the set of operations required for the acquisition and validation of survey data and their conversion into a computer readable format.

The validity and consistency of data provided by the respondent is ensured by the editing sub-function. Validity rules (numeric vs character, length of a variable, etc.) are applied to ensure that the captured data correspond to the data provided by the respondent, thus allowing for the automatic processing of the reported data in subsequent steps. Methods for detecting suspicious data are used to ensure consistency between variables. These techniques are used to identify units having a behaviour departing significantly from the rest of the population under study.

Many techniques which allow for checking the consistency between variables within a questionnaire exist. For example, it is possible to use historical checks, that is, to compare the value of a variable for the current survey cycle to the value of the same variable from a preceding cycle. It is also possible to compare two or more variables of the same cycle (cross editing). The use of cross editing generates the following problem: For a questionnaire having many financial variables, how does one identify the appropriate cross editing rules from the whole set of possible cross edits to ensure the consistency of the data.

However, even if it were possible to identify appropriate cross editing rules, this is not sufficient to ensure that the editing scheme will be optimal. To obtain an optimal scheme, an editing pattern that minimizes the number of cross editing rules must be found, while being efficient and effective. That is, the goal is not only to produce maximum results with minimum effort but also to produce good results.

The first approach considered was the use of Pearson's coefficient of correlation matrix. The Pearson coefficient of correlation measures the degree of linear association between two variables. By using this information, an attempt is made to group together variables which are highly correlated and to define cross editing rules within these groups. However analyzing a matrix with a large number of dimensions is not an easy task. Furthermore, a criterion is required to make a decision. Effectively it is very difficult to judge if the correlation between two variables is sufficiently large to consider that a linear relationship exists between two variables. For example, is a Pearson coefficient of correlation of 0.65 large enough to justify the grouping of two variables? To solve this problem, multivariate data analysis techniques were used.

2. THE GROUPING METHOD.

Factor analysis is a branch of multivariate data analysis which concentrates on the internal relationships of a set of variables. The grouping method described below is based on factor analysis theory.

The grouping method is a statistical tool used to identify natural groupings of variables allowing for the optimization of the editing process. The objective of the method is to partition the set of variables into two or more disjoint groups. An identified group contains variables that are highly correlated with each other. The result of the grouping is then used to develop an editing scheme allowing only variables within the same group to be cross referenced in order to minimize the total number of cross editing rules. Before detailing the utilization strategy of the classification method, a brief overview of the method used is presented.

The objective of the method is to partition a set of variables into two or more disjoint groups by using their correlation matrix. Groups are identified in such a way as to maximize the variance accounted for by the first principal component and to ensure that the proportion of variance explained by the second principal component is not too high.

The main steps of the grouping method can be summarized as follows:

Initially, the method considers all the variables as one group, then the following steps are repeated until all the conditions are met:

1) The first step consists of performing a principal component analysis on the correlation matrix of each group independently.

2) A given group is divided if the proportion of variance explained by the first principal component is insufficient or if the proportion of variance explained by the second principal component is too high. Otherwise, a group meeting these two criteria is not split.

3) A group is divided by using a variant of the orthoblique rotation proposed by H.H. Harman [3], and W. Harris and F. Keiser 1964 [4]. This technique is known as such since an oblique solution is obtained from a series of orthogonal rotations combined with a positive-definite matrix. An oblique solution implies that the reference axes are not perpendicular.

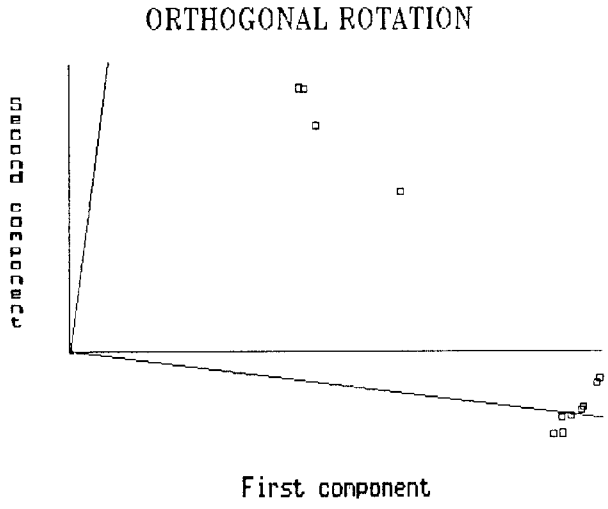
4) An iterative process re-assigns variables between groups in order to maximize the quantity of variance explained by the first principal component of each group.

5) Go back to step 1). The process is terminated when all groups meet the specified criteria for the proportion of variance explained by the first and the second principal component of a group.

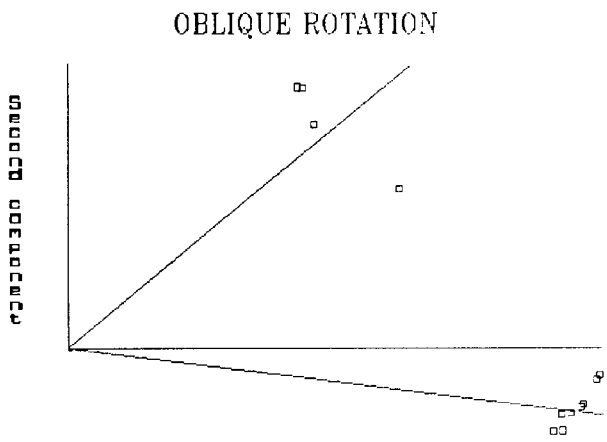
One of the basic principles of factor analysis is to always search for a simple solution. Here, simple solution means a solution for which each variable is well represented by a small number of axes (preferably only one). However, such a solution is not always found automatically by the factor analysis of a correlation matrix. This is why rotation of axes is used.

Principal component analysis provides one of a number of possible solutions for reference axes. Rotation allows pivoting these axes so as to obtain a better representation of the variables on them. This step helps to find a solution that explains the variability of a group of variables more adequately.

Why look for an orthoblique rotation rather than orthogonal? By forcing the rotation to be orthogonal, the same constraint is imposed on the axes. To get a clear picture and to ensure that each variables is well represented by one axis, the axes are often required to be oblique. For example, a correlation exists between the variables of the non-manufacturing and the variables of the manufacturing products (Canadian Census of Manufactures). However, this link is not strong enough to justify grouping these two categories of variables in such a way that they are well represented by only one axis. If an orthogonal rotation is imposed, the solution obtained will represent clearly one group at the disadvantage of the other. If an oblique solution is allowed, it is then possible to clearly represent simultaneously both groups with different axes. In order to visualize this situation, the plane of the first two principal components is presented for an oblique and an orthogonal solution in the following diagrams.



The orthogonal rotation improve only the representation of the group which is associated with the first principal component.



The oblique solution is easier to interpret. Using such a solution allows for a more detailed classification of variables.

The number of groups obtained by the grouping method is controlled by the use of parameters, specifying the proportion of variance explained by the first two principal components. In fact, if the parameter for the variance explained by the first component is increased (or the parameter for the variance explained by the second principal component is decreased), then the number of groups obtained will be higher.

To split a group in two, the correlation between each variable and the first two axes obtained by the orthoblique rotation is used. Each variable is paired with the axis with which it is most correlated. Since the orthoblique rotation provides a solution with axes which are not orthogonal, the grouping method provides the coefficient of correlation between the groups.

The algebraic details of the method are presented in the appendix.

3. UTILIZATION STRATEGY

The generalized data collection, capture and editing function is divided into two distinct parts: the designer system and the production system.

The designer system is used by survey managers, methodologists and computer analysts to define the survey requirements in terms of collection, capture and editing. This consists of the design and printing of the questionnaire, the building of an editing scheme, the computation of the parameters required for the outlier detection and others. The production system is used to run the collection, capture and editing of survey data and is under the responsibility of the operation staff. Even though there are different objectives and uses of the two systems, they are considered to be modules of an integrated system.

The method for the grouping of the variables is part of the designer system. It will be used as an analytical tool in the establishment of the editing pattern. It is important to remember that the goal of the grouping method is to group variables which are linked together for the purpose of editing. Once the groups are formed, the cross editing can be restricted to the relationships within groups while ensuring a high degree of consistency

between the variables of a record. Cross editing rules using variables from different groups will be done only if they are required to ensure a higher degree of consistency for a record. Since the grouping method is the result of an orthoblique rotation of the axes, the coefficient of correlation between groups is also available. The coefficient of correlation between groups can be used to help in the selection of inter-group cross editing rules, if required.

An empirical study using real survey data from the Census Of Manufactures was done using the SAS procedure VARCLUS [6]. This procedure uses a technique similar to the one presented in this article. Since the procedure already exists in SAS, it is therefore easy to implement. The multivariate data analysis is done using the correlation matrix of the variables, implying that a small amount of time and resources are required since the work is done only on the correlation matrix and not on the original data set.

The results of the empirical study conducted on real survey data are conclusive. The grouping method extracts the correlation pattern easily from a set of variables. Furthermore, the variable groups obtained are representative of the specific characteristics of the observed data and agree with the knowledge of the subject matter officers [2]. This empirical study has shown that the grouping method is easy to use and easy to interpret.

4. CONCLUSION

The grouping method presented in this paper can be used as a tool which allows combining statistical theory with the experience and knowledge of the subject matter officers. One of the objectives of the statistical theory is to provide a scientific law or a mathematical model to explain the behaviour of the data. A prior knowledge of the behaviour of the variables under study can help to verify the coherence of the groups identified by the method. Furthermore, certain unknown links between variables may be revealed by applying this method. However, if the behaviour of the variables under study is unknown, or if experimental results are desired, then this method groups the variables empirically and objectively. This method thus enables us to provide a theoretical framework for the editing process without forgetting its intuitive nature.

5. BIBLIOGRAPHY

[1] Berthelot, J.-M. (1989), Approche générale pour la sous-fonction de vérification et de correction des données, Statistique Canada, Division des méthodes d'enquêtes entreprises, Document de travail.

[2] Bilocq. F. (1989), Analyses sur la classification des variables et sur la détection des données suspectes, Statistique Canada, Division des méthode d'enquêtes-entreprises, Document de travail.

[3] Harman, H.H. (1976), MODERN FACTOR ANALYSIS, 3rd ed, University of Chicago Press, Ill. (487 p.)

[4] Harris, W. et Keiser, F. (1964), Oblique factor analytic solutions by orthogonal transformations, PSYCHOMETRIKA, vol 29, no 24, p. 347-362.

[5] Holzinger, K.J. et Harman, H.H. (1942), FACTOR ANALYSIS, University of Chicago Press, Ill. Chicago, Ill., (417 p.).

[6] SAS institute, (1985), SAS USER'S GUIDE : STATISTICS, Version 5, North Carolina.

[7] Thorndike, A.M., (1978), CORRELATION PROCEDURES FOR RESEARCH, Gardner Press, New York.

6. APPENDIX

Theoretical description of the grouping method

This appendix presents an algebraic description of the principal components analysis and the orthoblique rotation of the grouping method.

6.1. Notation

I : Identity matrix

R : Matrix of correlations between variables

Q : Matrix of eigenvectors of the matrix R

L : Diagonal matrix containing eigenvalues of matrix R

A : Matrix of coefficients α_{ij} of the principal components analysis (factor loadings)

C_j : j^{th} principal component

T : Orthogonal rotation matrix

S : Matrix of correlations between variables and axes

F : Matrix of correlations between axes

P : Matrix of coefficients (factor loadings) obtained after transformation of the matrix A

D : Positive definite matrix

v : Number of variables

6.2. Principal component analysis (PCA)

PCA of the matrix of correlations of each group separately.

By matrix theory, a matrix A can be found such that R = AA'

By PCA theory, $R = Q \Lambda Q'$ (correlation matrix broken down into eigenvalues and eigenvectors)

If we define : $A = Q \Lambda^{\frac{1}{2}}$

We find that : $AA' = Q \Lambda^{\frac{1}{2}} \Lambda^{\frac{1}{2}} Q'$

$= Q \Lambda Q'$

$= R$

The matrix A is called the factor loading matrix of the PCA.

Calculation of the proportion of variance accounted for by the first principal component C_j

Variance accounted for by $C_j = \sum_i \alpha_{ij}^2 = \lambda_j$ (where λ_j = j^{th} eigenvalue)

The proportion of variance explained $P_j = \lambda_j / \sum_i \lambda_i$

If $P_j < k_1$ and/or $\lambda_2 > k_2$, then the group must be separated. (k_1 and k_2 are predetermined parameters.)

6.3. Orthoblique rotation

Before explaining the theory underlying this method, a brief overview of the orthogonal and oblique methods is presented.

In the orthogonal case, matrix A (factor loadings) is transformed by means of an orthogonal rotation matrix T (where T'T=TT'=I). Let B=AT, where B becomes the new orthogonal solution.

Algebraically, the correlation between a variable (zi) and an axis (Fj) is expressed as follows :

$$S = A\Phi \quad (s_{ij} = \rho(z_i, F_j))$$

$$\rho(z_i, F_j) = \sum_{k=1}^{v} a_{ik}\rho(F_k, F_j)$$

Where v is the number of variables.

Since the axes are perpendicular, the correlation between two axes is defined as follow :

$$\sigma(F_i, F_j) = 1 \quad if \quad i = j$$

$$= 0 \quad otherwise$$

$$\Rightarrow \quad \Phi = I$$

$$\Rightarrow \quad S = A \quad , \quad s_{ij} = a_{ij}$$

In the oblique case, the matrix A is transformed by means of a rotation matrix constructed so as to obtain a solution in which the axes are not perpendicular. The matrix of the new coefficients obtained is called P to distinguish them from the old coefficients A. With an oblique rotation, the axes are correlated with each other ($F \neq I$). Then : S = PΦ

6.4. Orthoblique case

The first step is to apply the quartimax rotation method [3] to the first two eigenvectors so as to obtain the orthogonal rotation matrix T (TT'=T'T=I). The quartimax method maximizes a function of the 4th power of the eigenvectors in order to determine the angle of rotation .

Matrix T is then defined as follows :

$$T = t_{11} = \cos(\theta)$$

$$t_{12} = \sin(\theta)$$

$$t_{21} = -\sin(\theta)$$

$t_{22} = \cos(\theta)$ (where θ is derived using the quartimax method)

$$t_{ij} = 1 \quad for \quad i = j \quad i, j = 3, \ldots, v$$

$$t_{ij} = 0 \quad Otherwise$$

To complete the rotation, a diagonal matrix that will standardize the expression T 'Λ T so as to obtain the correlation matrix with 1's on the diagonal must be found. The matrix used to perform this oblique rotation is D^{-1}.

Calculation of D^{-1}

Recall that T' Λ T = [σ^2] and

that $\Phi = D^{-1}$ T' Λ T D^{-1} (Harman 1976)

We want $\Phi_{ij} = 1$ if i=j

Let define D^2 = Diagonal [T' Λ T] i.e. the σ^2 on the diagonal

Then :

$$D_i^2 = \sigma_i^2$$

$$D_i = \sigma_i$$

$$\frac{1}{D_i} = \frac{1}{\sigma_i}$$

Implying D^{-1} = [diagonal (T' Λ T)]$^{-1/2}$

$$\Rightarrow \quad D^{-1}T'\Lambda T D^{-1} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

$$if \quad i = j \quad \Rightarrow \quad \frac{\sigma_i^2}{\sigma_i \sigma_i} = 1$$

Once T and D have been calculated, the complete oblique solution is defined as follows :

P = Q T D (new coefficients)

$\Phi = D^{-1}$ T' Λ T D^{-1} (matrix of correlation between the axes)

S = Q Λ T D^{-1} (matrix of correlation between the variables and the axes)

The Sij are used to associate each of the variables on either the first or the second axis. The results of this step gives two unconnected groups of variables. The entire procedure is applied again to the groups until they all meet the criteria set for P_1 and λ_2.

The SAS VARCLUS procedure uses a similar technique to carry out the orthoblique rotation.