THE C-TYPE DISTRIBUTION AS AN UNDERLYING MODEL FOR CATEGORICAL DATA AND ITS USE IN FACTOR ANALYSIS

by

JANDYRA MARIA GUIMARÃES FACHEL

Thesis submitted to the University of London for the PhD degree

LONDON SCHOOL OF ECONOMICS
AND POLITICAL SCIENCE

September 1986

ABSTRACT

The 'surface of constant association' was introduced by Pearson (1913) and termed 'Contingency-type distribution', or briefly 'C-type' by Mardia (1967). Members of the C-type family may be used as underlying distributions for binary data and related 'contingency-type correlation coefficients' may be derived. We shall generalize this idea to polytomous variables and derive the maximum likelihood estimator of the parameter of association of the C-type distribution for data given in an R×C table, from which contingency-type correlation coefficients will be obtained.

Latent variable models with the assumption of an underlying logistic distribution for the manifest and/or latent variables have been proposed by Bartholomew (1980). If we consider the C-type logistic distribution we have a bivariate distribution with marginal logistic distributions and with correlation coefficient in the range $[-1,1]$. We study the correlation in the C-type logistic distribution and also compare this distribution with the C-type normal and bivariate normal distributions. Other members of the C-type family are considered such as the C-type sum-of-two logistics distribution.

Assuming the C-type distribution as an underlying model for categorical data we consider a factor analysis model not restricted to the traditional assumption of normality. Contingency-type correlation matrices for binary and polytomous data are used as input to factor analysis methods. This approach leads to an alternative method of factor analysis for categorical data with the practical advantage of a great reduction in computing time, allowing the model to be applied to

large data sets. Numerical applications are presented and results compared with Bartholomew's factor analysis for categorical data models.

Finally, the improper solutions in factor analysis are considered and we study the occurrence of Heywood cases as a function of sample size, number of variables and magnitude of the parameters in the model.

CXKAAW

ACKNOWLEDGEMENTS

I wish, above all else, to express my deepest gratitude to my supervisor, Professor D.J. Bartholomew, for his constant guidance and continual encouragement throughout my work at LSE. I have profited immensely from his many suggestions not only for the completion of this thesis, but, I am sure, for my future professional life.

I am also very grateful to the Department of Statistics of the Federal University of Rio Grande do Sul for granting me the study leave. The financial support from CAPES, Ministry of Education, Brazil, is highly acknowledged for without it this work would not have been possible.

The very tedious job of typing the manuscript has been expertly performed by Mary Cahill, with assistance from Sue Kirkbride. Their unfailing sympathy with my many wishes has earned my appreciation and gratitude.

The many data sets used in Chapter 6 were kindly provid :d by S.C. Richardson, P.D. Winter, F.L. Ashwood, J. Walker, J. Parrish and D. Dawkins. To all these and to the many others who have helped me in the development of this work, I take this opportunity of expressing my sincere gratitude.

CONTENTS

CXKAAW

CHAPTER 1 : INTRODUCTION

1.1 Basic Ideas

Most manifest (observed) variables in the social sciences are categorical variables, that is, with nominal or ordinal level of measurement. The categorical variables may be either dichotomies, with only two categories of responses (yes/no, right/wrong, etc) or polytomies, with more than two categories (agree/no opinion/disagree, etc). The binary (or dichotomous) variables are generally coded '0 or 1' and the polytomous variables are generally labelled as '1,2,3,...,c' where c is the number of categories.

The recognition that categorical variables are frequent in fields such as Sociology, Psychology or Economics has led to the developement of a variety of special models and methods for the analysis of these variables. A class of these models incorporates the idea of a latent variable, which is not directly observed. This class of models is known as Latent Variable Models for Categorical Data.

One of the best known models where the presence of latent variables (the factors) is of fundamental importance is Factor Analysis. For this reason, recent literature on latent variable models includes factor analysis as a special case (see Everitt, 1984 and Bartholomew, forthcoming).

Bartholomew (1983) presents various latent variable models according to the cross classification of the manifest and latent variables, each being classified as either categorical or metrical variables. Metrical variables are measured in an interval or ratio level and may be discrete or continuous. According to Bartholomew, factor analysis is defined for metrical latent variables and metrical manifest variables. Factor analysis for categorical data and latent trait analysis are appropriate for metrical latent variables and categorical manifest variables. Latent structure analysis models may also involve latent and manifest categorical variables. Other models may be classified using this typology, but we give only examples.

The models and methods for factor analysis of categorical data and latent trait analysis are presented in Bartholomew (forthcoming) from a new point of view. He considers two main approaches to the construction of the models: the 'Response Function Approach' with origins in the theory of educational testing and the 'Underlying Variable (UV) Approach', in the factor analysis tradition, where the categorical observed variables are supposed as being produced by underlying continuous variables. Bartholomew also shows that the two approaches are equivalent for binary variables but are generally not equivalent for polytomous variables.

We shall consider the underlying variable approach in this study and in the next section we present its general formulation.

1.2 The Underlying Variable Model for Categorical Data

Consider the linear factor analysis model given by

$$x_i = \sum_{i=1}^{q} \lambda_{ij} z_j + e_i \qquad i=1,\ldots,p \tag{1.1}$$

where $x_i (i=1,2,\ldots,p)$ is the manifest (observed) variable, $z_j (j=1,..,q)$ is the latent variable (factor), λ_{ij} is the factor loading and e_i is the error term. We suppose $E(z_j)=0$, $E(e_i)=0$ and $Var(X_i)=1$.

It is assumed that the p manifest variables, $x_1,\ldots,x_p$ depend upon $q<p$ latent variables z_j and that the set of q factors should explain the whole pattern of dependence among the x's, so that

$$g(\underset{\sim}{x}|\underset{\sim}{z}) = \prod_{i=1}^{p} g_i(x_i|\underset{\sim}{z}) \tag{1.2}$$

This is the assumption of conditional (or local) independence that is basic for all latent variable models. As we can only observe the x's, any inference about the parameters of the model must be based on the joint distribution given by

$$f(\underset{\sim}{x}) = \int g(\underset{\sim}{x}|\underset{\sim}{z})\ h(\underset{\sim}{z})d\underset{\sim}{z} \tag{1.3}$$

It is, also, usually supposed that the z's are independent continuous variables with zero mean and unit variance, such that

$$h(\underset{\sim}{z}) = \prod_{j=1}^{q} h_j(z_j) \tag{1.4}$$

and that each e_i is independent of all the other e's and of all z's.

With the specifications given above, the dependence structure of the model is given by

$$\underset{\sim}{\Sigma}_x = \underset{\sim}{\Lambda}\underset{\sim}{\Lambda}' + \underset{\sim}{\psi} \tag{1.5}$$

where $\underset{\sim}{\Sigma}_x$ is the dispersion matrix of the x's, $\underset{\sim}{\Lambda}$ is the p×q matrix of factor loadings and $\underset{\sim}{\psi}$ is the dispersion matrix of the e's which is a p×p diagonal matrix with diagonal elements $\psi_i (i=1,\ldots,p)$. The diagonal elements of the matrix $\Lambda\Lambda'$, h_i^2, are known as communalities and $h_i^2=1-\psi_i$.

Since by model (1.1), x_i is assumed to be a weighted sum of continuous variables, the assumption of x_i as categorical variable would be inconsistent with the model. To avoid this we suppose that what is actually observed is not x_i but the categorical variable x_i^* with c_i categories labelled $1,2,\ldots,c_i$ (any other 'labels' could be used, as for example: $0,1,2,\ldots,c_i-1$); when $c_i=2$, $i=1,2,\ldots,p$ we have the binary variables case. We also suppose that the distribution of $\underset{\sim}{x}$ is underlying that of $\underset{\sim}{x}^*$ and that the multinomial distribution for the categorical vector $\underset{\sim}{x}^*$ can be deduced by integration over $\underset{\sim}{x}$. (This formulation is similar to that of Mutnén and Kaplan (1985), but here we are supposing that no arithmetic operation can be performed with the labels of the categories, therefore there is no meaning for quantities such as the mean and variance of categorical variables or covariances between them).

Now if the x's are each related to one or more of the z's by the factor analysis model (1.1), there will be correlations among the x's. Therefore, given a pair of observed categorical variables (x_i^*,x_j^*), we have a two-way contingency table that we suppose has been formed from the underlying bivariate continuous distribution. A measure of association for the two-way table, as for example the tetrachoric coefficient or other similar measure will be an estimate of the correlation coefficient of the bivariate continuous distribution. Thus we will say that the association structure given by $\underset{\sim}{\Sigma}_{x^*} = \underset{\sim}{\Lambda}\underset{\sim}{\Lambda}' + \underset{\sim}{\psi}$

holds for $\underset{\sim}{x}^*$.

Normal Variables

If in the factor analysis model (1.1) we suppose that $\underset{\sim}{z}$ and $\underset{\sim}{e}$ are variables with normal distributions, $\underset{\sim}{x}$ will be normal and we can fit the model if we estimate the correlation coefficients using the tetrachoric or polychoric correlation coefficients for the case of binary or polytomous variables respectively.

The tetrachoric correlation coefficient is a well known measure of association introduced by Pearson in the beginning of this century, with the purpose of estimating the correlation coefficient of the bivariate normal distribution for data given in a 2x2 contingency table. Computer routines are available in the BMDP package, for example (see also, Divgi, 1979).

The polychoric correlation coefficient is a measure of association for polytomous variables when we suppose an underlying bivariate normal distribution with parameter ρ. The polychoric coefficient was introduced by Lancaster and Hamdan (1964) as a generalization of the tetrachoric. Olsson (1979) proposes a maximum likelihood method to estimate ρ. A computer routine is available in the LISREL package.

Given the correlation matrix with tetrachoric or polychoric correlation coefficients, these can be used as input to a standard factor analysis program.

Bartholomew (forthcoming) points out that the method of factor analysis for categorical variables using tetrachoric or polychoric coefficients has the disadvantage that the correlation matrix may not be positive definite, but as Lord & Novick (1968 p.349) comment, this is an empirical difficulty rather than a theoretical objection to the use of tetrachoric (polychoric) correlations. Alternative and more efficient methods for the factor analysis of categorical variables have been suggested in the literature both for binary and polytomous variables and we shall review some of the methods in Chapter 4.

Non-normal variables

The assumption of a normal distribution for $\underset{\sim}{z}$ and $\underset{\sim}{e}$ and hence for $\underset{\sim}{x}$ in the underlying variable model given in section 1.3 is motivated by the desire to make the model consistent with the standard theory for continuous variables, as Bartholomew (forthcoming) points out. As we are observing categorical variables it may be argued whether this assumption can be expected to hold with any generality.

De Leeuw et al (1983) point out that multivariate normality is the exception rather than the rule in social science situations and that in the linear structural model context this has led to the construction of asymptotically distribution free models (see Browne, 1982, 1984).

Latent variable models for categorical data with the assumption of an underlying logistic distribution for the manifest and/or latent variables have been proposed in the literature. As examples we have Birbaum's logistic test model (Lord and Novick, 1968), the Logit model (Bartholomew, 1980), the multivariate logistic latent trait model (Bock, 1972). It is known that there is no bivariate logistic distribution with logistic margins and unconstrained correlation coefficients, as Muthén (1983) points out. Therefore the assumption of an underlying bivariate logistic distribution for the observed cross tables would be of no practical advantage.

Bartholomew (forthcoming) considers different distributional forms for the latent variables ($\underset{\sim}{z}$) and for the error term ($\underset{\sim}{e}$) in the underlying variable model (1.1). He considers the logistic distribution for $\underset{\sim}{e}$ and normal for $\underset{\sim}{z}$ or logistic for $\underset{\sim}{z}$ and $\underset{\sim}{e}$, or even one special case where $\underset{\sim}{e}$ has the Type I extreme value distribution. If $\underset{\sim}{z}$ and $\underset{\sim}{e}$ are non-normal and if the number q of factors is not too small, $\underset{\sim}{x}$ as a linear combination of independent random variables will be approximately normal by the central limit theorem. If, however, $\underset{\sim}{z}$ and $\underset{\sim}{e}$ are non-normal and q=1, we should clearly consider other distributional forms as underlying marginal distribution for $\underset{\sim}{x}$.

As we have seen in section 1.2 we should concentrate on procedures which use the marginals up to the second order, that is, which use all cross-tables. The elements of the dispersion matrix in the factor analysis model should make sense as measures of association. We know that multinormality is a sufficient condition for correlations (tetrachorics and polychorics) to make sense as association measures. Is there any other correlation coefficient or association measure that would be used in situations where the tetrachorics and polychorics are not appropriate? In the next section we will consider the problem for binary data and introduce the main topic of the thesis that is the C-type distribution as an underlying model for categorical variables.

1.3 The C-type distribution as an underlying model for categorical variables

2×2 contingency tables case

Suppose that the observed 2×2 tables for binary variables have been formed from an underlying continuous bivariate distribution and that the dichotomies are formed by cutting the marginal distributions in some point of dichotomy (threshold). Since these points are often arbitrary, we should have a bivariate distribution function with the property that when it is cut anywhere by lines parallel to the axes X_i and X_j, the probabilities in the four quadrants viewed as a 2×2 contingency table would imply a constant association. The association, in this case, is measured by the cross product ratio $\psi = p_{11}p_{22}/p_{21}p_{12}$ where the p_{ij}'s are the probabilities in the four quadrants.

Pearson and Heron (1913) showed that it is always possible to construct a surface for which the parameter of association ψ is constant for every fourfold division. This distribution was called the surface of constant association. Plackett (1965) reintroduced the same distribution as a one-parameter class of bivariate distributions from given margins. Mardia (1967, 1970) gave the distributional properties of this class and termed it the Contingency-type distribution or briefly, "C-type". For reasons that will be clear later, we choose Mardia's nomenclature, as the best way to refer to this distribution.

In Chapter 2 we shall present the C-type distributions in detail, including a historical note.

The parameter of association, ψ, of the C-type distribution is estimated, in a 2×2 contingency table, by the cross product ratio (odds ratio). For any member of the C-type distribution family, the correlation coefficient is a function of ψ only.

Chambers (1982) points out that the C-type distribution (or, as he calls it, constant odds-distribution) underlies all the correlation coefficients that are functions of ψ. By an analogy with the name of the distribution, we shall call all correlation coefficients that are functions of the cross-product ratio as "Contingency-type correlation coefficients".

Mardia (1967) presents the moment formulae for the C-type uniform and the C-type normal distributions, having uniform and normal margins respectively.

We have seen in section 1.4 that the logistic distribution plays an important role in Latent Variable Models. In Chapter 3 we compare the C-type logistic distribution with the C-type normal and bivariate normal distributions. We also study the correlation in the C-type logistic distribution. We present the distribution of the sum of two logistic random variables and use this distribution as the margins for the C-type, getting one more member of the family: the C-type sum-of-two logistics distribution. Using numerical methods we obtain the distribution of the sum of three logistic random variables and the sum of a normal plus a logistic random variables. The C-type distributions with this mixture of distributions in the marginals are compared with the bivariate normal. The results are presented in Chapter 3.

The assumption of an underlying C-type distribution for data given in 2×2 contingency tables lead us not only to estimates of the correlation coefficients that are much easier to calculate than the tetrachoric correlation coefficient but also provides a convenient approximation to the normal model.

R×C contingency tables case

Consider the general case when the data are given in an R×C contingency table, having marginal variables with R and C ordered categories respectively, that is, when the manifest variables are polytomous rather than dichotomous. How should we estimate the parameter ψ of the C-type distribution? For the binary case, we have seen in the last section that ψ is estimated by the sample cross product ratio from the 2×2 table.

In Chapter 3 we present the maximum likelihood method of estimating the parameter ψ for data given in an R×C table. The likelihood equations are derived and Fisher's scoring method is used to obtain an iterative solution. The asymptotic standard errors of the estimate are also presented. A computer program for the method is enclosed.

The maximum likelihood estimate of ψ is then used to obtain contingency-type correlation coefficients for polytomous variables. Numerical examples are given and related methods compared.

1.4 The Underlying Variable Model based on C-type distributions

Assuming the C-type distribution as an underlying model for the manifest variables x_i^* and x_j^*, introduced in Section 1.2, we obtain simple methods for estimating the parameter of association of the underlying distribution for various forms of the marginal distributions.

A factor analysis model can be fitted using the estimated contingency-type correlation coefficients for binary or polytomous variables as input for a standard factor analysis program.

This approach leads to an alternative method of factor analysis for categorical data with the practical advantage of a great reduction in computing time compared with other methods of factor analysis for categorical data. This fact allows the method to be applied to large data sets.

In Chapter 4 we give a brief account of the models and methods developed for factor analysis of categorical data.

In Chapter 5 we analyse several data sets with binary manifest variables using the correlation methods based on the C-type distribution and compare the results with Bartholomew's factor analysis for categorical data models.

In Chapter 6 numerical applications are presented for polytomous data. We compare the factor analysis results for the contingency-type correlation coefficients versus factor analysis results using as input the Pearson product moment correlation coefficients.

Before the advent of methods for polytomous data these variables were handled by transforming them into binary variables and then using one of the methods for binary data. This may lead to a loss of information that can be serious if the number of categories is relatively large. Using the contingency-type correlation coefficients, both for polytomous data and for the binary version of the same data set, we compare the results in Chapter 5.

On using the standard methods of Factor Analysis for analysing the data we notice a high frequency of Heywood cases, not only for contingency-type correlation coefficients but also for other correlations matrices. The occurrence of the improper solutions in factor analysis is not a rare event, not only in our examples but in many other factor analysis examples in the literature. In the last chapter of the thesis we review what is known about Heywood cases in the literature and we study the probability of an improper parameter estimate for different values of the sample size, number of manifest variables and number of factors. Using simulated data we try to identify the situations where the occurrence of improper solutions in factor analysis is more probable and how they can be avoided.

CHAPTER 2 : THE C-TYPE DISTRIBUTION

2.1 Definition

By developing an analogy with the cross product ratio, one of the measures of association in a 2×2 contingency table, Plackett (1965) has presented a class of bivariate distributions for given margins and just one parameter to measure the degree of association. Mardia (1967,1970) has termed this class of distributions, the Contingency-type distribution or C-type distribution. Mardia has also derived the moment-formulae appropriate to this class and the distributional properties.

The origins of the C-type distribution can be traced back to the beginning of the century in a study about theories of association by Pearson and Heron (1913) and Pearson (1913) where the distribution was called "the surface of constant association". Pearson (1913) constructed the distribution for the especial case of normal margins and constant Yulean coefficient of association. Photographs of the surface, the regression lines and the contours were also shown by Pearson (1913).

Suppose we have two random variables X and Y with distribution function F(x) and G(y) respectively, with joint distribution function H(x,y). Any bivariate distribution with d.f.H and marginal d.f's F and G can be dichotomized at an arbitrary point (x,y), giving a 2×2 contingency table.

Let p_{ij}, i,j=1,2, be the probability that an observation falls into cell (i,j) as determined by the dichotomies at point (x,y). Putting p_{11} = H, p_{12} = F-H, p_{21} = G-H and P_{22} = 1-F-G+H, we have the data summarized in a 2×2 table as shown below:

		y	
x	H	F−H	F
	G−H	1−F−G+H	1−F
	G	1−G	

The cross product ratio, ψ, for this table is given by

$$\psi = \frac{H(1-F-G+H)}{(F-H)(G-H)} \qquad (2.1)$$

and from (2.1) we have the equation:

$$(\psi-1)\ H^2 - \{1 + (F-G)(\psi-1)\}\ H + \psi FG = 0\ , \qquad \psi>0 \qquad (2.2)$$

Plackett (1965) has shown that when F and G are given, ψ is a monotonic increasing function of H, taking the value zero when $H = \max(0, F+G-1)$ and the value ∞, when $H = \min(F,G)$. Hence for given F,G and H, there corresponds a single H to satisfy

$$\max\ (0,F+G-1) \leqslant H \leqslant \min\ (F,G) \qquad (2.3)$$

Mardia (1967) considering (2.2) and (2.3) has shown that the only possible root of the quadratic equation (2.2) is given by

$$H = \begin{cases} S - \{S^2 - 4\psi(\psi-1)FG\}^{1/2}\ /\ \{2(\psi-1)\} & (\psi\neq 1) \\ FG & (\psi=1) \end{cases} \qquad (2.4)$$

where $S=1+(F+G)(\psi-1)$.

The expression given by (2.4) defines the C-type distribution when X and Y are normal variables, the distribution given by this expression is a C-type normal distribution and when $F(x) = x$ and $G(y) = y$ we have the C-type uniform distribution.

Mosteller (1968) also following Plackett (1965) has derived a similar expression for what he called the invariant distribution, because that distribution has the property that wherever the dichotomies are made, the probabilities in the four quadrants, viewed as a contingency table, have the invariance property. Consequently the cross product ratios are the same, independently of the point of dichotomy.

Using this idea we can suppose an R×C contingency table which we imagine to have been formed from a C-type distribution with parameter of association ψ. How should we estimate ψ from the data given in an R×C contingency table? In the next chapter we shall present the maximum likelihood method of estimation of the parameter ψ for R×C tables. The results are then used to obtain simple methods of estimation of the correlation coefficient for polytomous data.

In this chapter we present an account of the earlier research related to the C-type distribution (Section 2.2). In Section 2.3 we review Mardia's results about the correlation for the C-type uniform and C-type normal distributions, other correlation coefficients for data given in a 2×2 contingency table are also reviewed. A new member of the C-type family is studied and the correlation in the C-type logistic distribution is presented in Section 2.4.

The distribution of the sum of two logistic random variables is obtained and using this result we obtain the C-type sum-of-two logistic distribution in Section 2.5. Comparison of the C-type distributions with the bivariate normal distribution are presented in Sections 2.6 and 2.7.

2.2 Earlier Research

On analysing contingency tables we may distinguish two kinds of problems. First we may be interested in analysing the tables according to the pattern of association between the two sets of categories. Secondly, we may be interested in estimating the association or correlation parameter from the table.

In an R×C contingency table the association parameter could be estimated by any of the (r-1).(c-1) cross product ratios formed by dividing up the table in 2×2 subtables. If the underlying distribution is the C-type distribution we should expect approximately equal values.

Wahrendorf (1980) deriving the asymptotic distribution for (r-1)(c-1) cross product ratios, gives a statistic for testing the equality of the estimates of the parameter of association in an R×C table. This test is also useful to determine whether the hypothesis of one parameter constant association model is plausible. Wahrendorf also proposes a "weighted average estimator" of the parameter based on the (r-1)(c-1) estimators.

Goodman (1979) presents a class of association models. One of these models, namely the uniform association model was subsequently called the distribution with "constant local association" (Goodman, 1981) as to distinguish it from the constant association model introduced by Pearson (1913). Goodman (1981) shows that the constant local association model agrees closely with the bivariate normal distribution and he proposes a polychoric correlation coefficient based on his constant local association model. Goodman's models form a class of log linear models.

Ogborn (1984) compares Goodman's local association model with the cumulative odds-ratio model of uniformity of association presented by Wahrendorf (1980). In his paper, Ogborn shows that the cumulative model can be adapted to the deletion of cells, in order to improve the fit of the model in some special cases, as for example, for social mobility data. Ogborn also proposes computational algorithms for both models.

Dale (1984) also compares the local versus global or cumulative models for bivariate ordered responses with emphasis on the differences between the types of associations rather than the parameterizations. As Dale points out, local association parameters are cross ratios of 2×2 subtables of adjacent cell probabilities, while global association parameters are cross-ratios of quadrant probabilities. Dale relates, in passing, the global and local models with univariate responses models such as the generalized logistic model described by McCullagh (1980) which may be considered as an univariate specialization of the global association models. According to Dale (1984), the local association model is appropriate only if marginal categories are well defined.

As we observe, several names for the same model have been used in the literature: surface of constant association (Pearson, 1913); class of bivariate distributions with parameter of association ψ (Plackett, 1965); Contingency-type (or C-type) distribution (Mardia, 1967); invariant distribution (Mosteller, 1968); constant-odds distribution (Chambers, 1982); cumulative odds-ratio model of uniformity of association (Ogborn, 1984); global association model (Dale, 1984).

We shall prefer Mardia's nomenclature as it has the advantage of specifying completely the members of the family of distributions in this class, as for example when we use the names C-type normal distribution, C-type uniform distribution, etc.

Most of the above cited work gives emphasis to the analysis of the pattern of association in the contingency table. In this thesis we shall use the C-type distribution as an underlying model for categorical data and emphasis will be given to the estimation of correlation or association parameters from the contingency tables. Correlation coefficients as function of the association parameter of the C-type distribution, which we shall call contingency-type correlation coefficients, will then be used as input for factor analysis methods.

If the bivariate normal distribution is supposed as an underlying model for categorical data, earlier research concerning the estimation of the latent correlation may be summarized as follows.

For binary data, the tetrachoric correlation coefficient, introduced by Pearson (1901) as a measure of bivariate normal latent correlation is a well-known measure of correlation. An algorithm for its calculation has been described by Digvi (1979) and an easily accessible computer routine is available on the BMDP statistical analysis system based on the method discribed by Brown and Benedetti (1977).

A generalization of the arguments behind the tetrachoric coefficient to polytomous variables has been presented by Lancaster and Hamdam (1964). On using the theory of orthonormal functions, the correlation between latent variables under the normality assumption is estimated from a general R×C contingency table.

Olsson (1979) has used the maximum likelihood method for estimating the polychoric correlation from data given in an RxC table, also with the normality specification on the latent response variables. Olsson's method is a generalization of the method presented by Tallis (1962) for 2×2 and 3×3 contingency tables.

Olsson has derived his polychoric correlation estimates using two methods: the full maximum likelihood estimation method, when the correlation coefficient and the thresholds are estimated simultaneously, and the "two-step maximum likelihood" estimation, when the thresholds are computed from the observed marginal proportions. The comparison of the estimates, using the two methods, for generated samples, as it is presented in Olsson (1979, p.453-455), shows that the two methods are practically equivalent. He also points out that the full maximum likelihood estimate may lead to different threshold estimates for variable x when ρ_{xy} is estimated than when ρ_{xz} is estimated. The "two-step maximum likelihood" estimation method has the advantage of reducing the computational work.

Lancaster and Hamdam's method and Olsson's method are equivalent for 2×2 tables when the tetrachoric correlation coefficient is being estimated. The method proposed by Olsson (1979) is utilized by Muthén (1983) in his three stage estimation method for structural equation modelling with categorical variables. A computer routine for evaluating the polychoric correlation coefficient is available in the LISREL package.

2.3 Contingency-type correlation coefficients

For 2×2 contingency tables various measures of correlation have been suggested in the literature and most of the measures are functions of the cross product ratio $\psi = p_{11}p_{22}/p_{12}p_{21}$.

Yule's coefficient of association Q and his coefficient of colligation r_y are given by

$$Q = (\psi-1)/(\psi+1)$$

$$r_y = (\psi^{\frac{1}{2}}-1)/(\psi^{\frac{1}{2}}+1)$$

Pearson's Q_3 which is an approximation to the tetrachoric correlation coefficient is

$$Q_3 = \cos[\pi/(\psi^{\frac{1}{2}}+1)]$$

Mardia (1967) studies the correlation coefficient for the C-type uniform distribution and obtains

$$\rho_u(\psi) = (\psi^2-1-2\psi\log\psi)/(\psi-1)^2$$

The correlation in the C-type normal distribution is also considered by Mardia (1967) and he presents the values of the coefficient $\rho_N(\psi)$ for various values of ψ (see section 2.4).

Chambers (1982) also considers the correlation coefficient of the C-type uniform and C-type normal distributions and observes that several of these measures of association can be closely and conveniently approximated by a generalization of Yule's coefficients given by

$$r_\nu = (\psi^\nu -1)/(\psi^\nu +1)$$

According to Chambers for $\nu = 2/3$, we obtain an approximation of the correlation coefficient of the C-type uniform distribution to within 2 per cent,

$$\rho_u(\psi) \simeq r_{2/3}$$

and for $\nu = 0.64$, we obtain an approximation of the correlation coefficient of the C-type normal distribution

$$\rho_N(\psi) \simeq r_{0.64}$$

Chambers coefficient r_ν is very useful not only in providing reasonable estimates for the latent correlation coefficient of the underlying C-type distribution but also for the bivariate normal distribution. In this case,

$$r_{0.74} = (\psi^{0.74} -1)/(\psi^{0.74} + 1)$$

where ψ is the observed cross-product ratio for the 2x2 table, is a reasonably unbiased estimate of ρ for the bivariate normal distribution, according to Chambers (1982).

Bishop, Fienberg and Holland (1975, Chapter 11) give basic properties of the cross-product ratio and also consider measures of association based on the cross-product ratio, given by the general formula :

$$g(\psi) = \frac{f(\psi)-1}{f(\psi)+1}$$

As Chambers (1982) points out, the C-type distribution underlies all the coefficients determined by the parameter ψ and by an analogy with the name of the distribution we shall call all correlation coefficients that are function of the parameter of association ψ as "contingency-type correlation coefficient".

2.4 Correlation in the C-type logistic distribution

The logistic curve was first used as a description of population growth in 1920 and was called the "logistic function" (see Ashton, 1972). The logistic distribution function is given by

$$F(x) = \frac{1}{1+e^{-(\alpha+\beta x)}} \qquad -\infty < x < \infty \qquad (2.5)$$

Differentiation yields the form

$$f(x) = \beta F(x)(1-F(x)) \qquad -\infty < x < \infty \qquad (2.6)$$

Bivariate logistic distributions have been studied by Gumbel (1961) motivated by the fact that the logistic distribution closely resembles to the normal, both being symmetrical. Two different bivariate logistic distribution were considered by Gumbel, but for both distributions the correlation coefficient is either constant or restricted to the interval (-0.304; 0.304).

In this section we shall study the correlation in the C-type logistic distribution, which is a member of the C-type family with logistic marginal distributions.

Mardia (1970) has shown that for any continuous distribution

$$\operatorname{corr}(X,Y) = \frac{1}{\sigma_1\sigma_2} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (H-FG)\ dxdy \qquad (2.7)$$

where σ_1, σ_2 are the standard deviation, H is the joint distribution function of x and y and F and G are the marginal distribution functions. Consider the standard logistic distribution with $\alpha=0$ and $\beta=1$ in (2.5) which has the standard deviation given by $\pi/\sqrt{3}$ (see Gumbel, 1961). Then

$$F(x) = [1+e^{-x}]^{-1} \text{ and } G(y) = [1+e^{-y}]^{-1}$$

Using the fact that for the logistic distribution

$$x = \ln F - \ln(1-F) \text{ and } y = \ln G - \ln(1-G)$$

the integral (2.7) becomes

$$\mathrm{corr}(X,Y) = \frac{3}{\pi^2} \int_0^1 \int_0^1 (H-FG) \frac{dF}{F(1-F)} \frac{dG}{G(1-G)} \text{ .} \qquad (2.8)$$

where H is the C-type distribution function given by (2.4). Numerical integration of (2.8) was performed using the NAG subroutine D01FBF, which evaluates double integrals with the Gauss-Legendre formula . The program uses 24 nodes in each dimension. The numerical values of the correlation coefficient $\rho_L(\psi)$ of the C-type logistic distribution are then tabulated for various values of the parameter of association ψ and presented in table 2.1.

In order to compare the values of the correlation coefficients of the C-type logistic distribution $\rho_L(\psi)$, obtained by numerical integration of (2.8) with those of the C-type normal distribution, $\rho_N(\psi)$ and C-type uniform distribution, $\rho_U(\psi)$, obtained by Mardia (1967), we also present $\rho_N(\psi)$ and $\rho_U(\psi)$ in table 2.1.

Table 2.1 - Correlation coefficients $\rho_L(\psi)$, $\rho_N(\psi)$ and $\rho_U(\psi)$

ψ	$\rho_L(\psi)$	$\rho_N(\psi)$	$\rho_U(\psi)$
1.0	0.0000	0.0000	0.0000
1.1	.0290	.0303	.0318
1.2	.0554	.0580	.0607
1.3	.0796	.0833	.0873
1.4	.1020	.1067	.1117
1.5	.1227	.1284	.1344
1.6	.1420	.1486	.1555
1.7	.1600	.1675	.1752
1.8	.1770	.1852	.1937
1.9	.1929	.2018	.2111
2.0	.2079	.2175	.2274
2.1	.2221	.2323	.2429
2.2	.2356	.2463	.2575
2.3	.2484	.2597	.2714
2.4	.2606	.2723	.2846
2.5	.2722	.2844	.2971
2.6	.2833	.2960	.3091
2.7	.2939	.3070	.3206
2.8	.3040	.3175	.3315
2.9	.3138	.3277	.3420
3.0	.3231	.3374	.3521
3.2	.3408	.3557	.3710
3.4	.3572	.3727	.3886
3.6	.3725	.3885	.4049
3.8	.3867	.4032	.4202
4.0	.4001	.4147	.4344
4.2	.4127	.4300	.4478
4.4	.4246	.4423	.4604
4.6	.4358	.4539	.4722
4.8	.4465	.4648	.4835
5.0	.4566	.4752	.4941
5.5	.4797	.4989	.5184
6.0	.5004	.5201	.5400
6.5	.5189	.5390	.5592
7.0	.5357	.5561	.5766
7.5	.5509	.5716	.5923
8.0	.5649	.5858	.6067
8.5	.5778	.5989	.6199
9.0	.5897	.6110	.6320
9.5	.6008	.6221	.6433

10.0	.6111	.6325	.6537
11.0	.6297	.6513	.6725
12.0	.6462	.6678	.6881
14.0	.6740	.6957	.7166
16.0	.6968	.7183	.7390
18.0	.7159	.7372	.7576
20.0	.7322	.7533	.7733
25.0	.7643	.7848	.8039
30.0	.7882	.8080	.8263
35.0	.8068	.8260	.8435
40.0	.8219	.8404	.8573
50.0	.8448	.8622	.8779
75.0	.8801	.8954	.9088
100.0	.9007	.9144	.9262
150.0	.9244	.9360	.9457
200.0	.9379	.9482	.9565
300.0	.9534	.9617	.9684
400.0	.9621	.9692	.9749
600.0	.9719	.9775	.9819
1200.0	.9836	.9870	.9898
2000.0	.9894	.9914	.9934
∞	1.0000	1.0000	1.0000

Chambers (1982) has presented a generalization of Yule's coefficient of association that is a simple function of the parameter of association ψ, as we have seen in Section 2.3. Chambers' formula is given by

$$r_\nu = (\psi^\nu -1)/(\psi^\nu +1).$$

Following Chambers, we observe that the correlation coefficient for the C-type logistic distribution can be conveniently approximated by r_ν for $\nu = 0.61$, that is $\rho_L(\psi) \approx r_{0.61}$ to within 2 per cent over the whole range $-1 < \rho_L(\psi) < 1$.

A better approximation, however, is obtained by observing that $\rho_L(\psi)$, obtained by numerical integration yields values a few per cent smaller than $\rho_L(\psi)$ and that, for given ψ,

$$\rho_L(\psi) \approx \rho_N(\psi) \; [0.954 + 0.046|\rho_N(\psi)|^3] \qquad (2.9)$$

Chambers (1982) had observed that this same approximation formula was valid for relating $\rho_N(\psi)$ with $\rho_U(\psi)$, that is

$$\rho_N(\psi) \simeq \rho_U(\psi)\ [0.954 + 0.046|\rho_U(\psi)|^3] \tag{2.10}$$

On substituting (2.10) in (2.9), we obtain

$$\rho_L(\psi) \simeq 0.910|\rho_U(\psi)|+0.082|\rho_U(\psi)|^4+0.007|\rho_U(\psi)|^7+0.001|\rho_U(\psi)|^{10} \tag{2.11}$$

The approximation given by (2.11) gives values for the correlation coefficient of the C-type logistic to within ±0.008 of $\rho_L(\psi)$ for all values of ψ.

The expression (2.11) has the advantage that it is a closed form for $\rho_L(\psi)$ if we consider that

$$\rho_U(\psi) = \frac{\psi+1}{\psi-1} - \frac{2\psi\ell n\psi}{(\psi-1)^2} \tag{2.12}$$

Therefore, substituting (2.12) in (2.11) we obtain a very good approximation to $\rho_L(\psi)$ without using numerical integration routines.

In Table 2.2 we present the values of the correlation coefficient for the C-type logistic distribution obtained by the approximation formulae $r_{0.61}$ and $\rho_L(\psi)$ given by (2.11) for several values of ψ. The true values of $\rho_L(\psi)$ are also presented for comparison.

Table 2.2 - Correlation coefficients for the C-type logistic distribution obtained by approximate formulae

ψ	Exact value of $\rho_L(\psi)$	Approximate value given by expression (2.11)	Approximate value given by $r_{0.61}$
1.0	0.0000	0.0000	0.0000
1.2	.0554	.0553	.0555
1.4	.1020	.1017	.1023
1.6	.1420	.1416	.1424
1.8	.1770	.1764	.1777
2.0	.2079	.2072	.2083
2.2	.2356	.2347	.2359
2.4	.2606	.2595	.2608
2.6	.2833	.2821	.2834
2.8	.3040	.3027	.3041
3.0	.3231	.3217	.3231
4.0	.4001	.3983	.3993
5.0	.4566	.4546	.4549
6.0	.5004	.4985	.4979
7.0	.5357	.5340	.5324
8.0	.5649	.5635	.5610
9.0	.5897	.5886	.5851
10.0	.6111	.6103	.6058
12.0	.6462	.6460	.6398
14.0	.6740	.6746	.6668
16.0	.6968	.6979	.6889
18.0	.7159	.7176	.7072
20.0	.7322	.7344	.7229
30.0	.7882	.7923	.7769
40.0	.8219	.8271	.8093
50.0	.8448	.8508	.8316
100.0	.9007	.9079	.8863
200.0	.9379	.9449	.9240
300.0	.9534	.9597	.9402
400.0	.9621	.9679	.9496
600.0	.9719	.9768	.9604
1200.0	.9836	.9869	.9739
2000.0	.9894	.9915	.9808
∞	1.0000	1.0000	1.0000

2.5 The C-Type Sum-of-two Logistics Distribution

The underlying variable factor analysis model, presented in section 1.2 is given by

$$\underset{\sim}{x} = \Lambda\underset{\sim}{z} + \underset{\sim}{e} \tag{2.13}$$

where x is the continuous response variable underlying the categorical manifest variable. This model can be fitted using only the correlation coefficients (measures of association) of the bivariate distributions (x_i, x_j) estimated from the observed R×C contingency tables which we are supposing have been formed by the underlying continuous bivariate distribution.

If $\underset{\sim}{z}$ and $\underset{\sim}{e}$ are both normal, the joint distribution of (x_i,x_j) is bivariate normal, which is not the same as the C-type distribution. However, since the pioneer work about the C-type distribution, or surface of constant association, by Pearson (1913), he and many other writers (see, for example, Plackett, 1965; Mosteller, 1968; Mardia, 1967) have shown that the C-type normal and the bivariate normal are similar.

In the next sections we shall review the way the two distributions were compared, suggest a new form of considering the equivalence between the parameters of both distributions and compare the bivariate normal with some other members of the C-type family as, for example, the C-type logistic, the C-type sum-of-two logistic distribution. The reason for choosing such distributions is explained below.

We have seen, in Chapter 1, that Bartholomew's response function (RF) models, which include the logit/logit model, the logit/probit model and the probit/probit model are equivalent to the underlying variable (UV) model for the binary variables case. (Bartholomew, forthcoming).

Bartholomew's models will be reviewed in Chapter 4, but now we consider, for simplicity, the binary case and the one-factor model. The response function version of the model is given by

$$G^{-1}\{\Pi_i(y)\} = \alpha_{i0} + \alpha_{i1}\, H^{-1}(y) \qquad i=1,2,\dots,p \qquad (2.14)$$

where G^{-1} and H^{-1} are inverse distribution functions of symmetrical random variables with zero mean and unit variance. $\Pi_i(y) = \Pr\{X_i=1/y\}$ is the response function and y is uniform on $(0,1)$.

The equivalent UV model for binary data and only one factor is

$$x_i = \lambda_i z + e_i \qquad i=1,2,\dots,p \qquad (2.15)$$

where z and e_i are independent with zero mean, z has unit variance and $\text{var}(e_i) = \psi_i$. The binary manifest variable x_i^* may be defined as

$$x_i^* = 1 \text{ if } x_i \leqslant \tau_i \text{ and } x_i^* = 0 \quad \text{otherwise}$$

where τ_i is a threshold value.

The distribution of e_i in the UV model corresponds to the distribution G in the RF model and the distribution of z (the factor) corresponds to the choice of H in the RF model. (We will present Bartholomew's formal proof of the equivalence of the models in Chapter 4).

In the logit/logit model of Bartholomew (1980), G^{-1} and H^{-1} are both logit functions, so this is equivalent to taking the distribution of z and e_i in (2.15) as logistic distributions. In the logit/probit model (G^{-1} is logit and H^{-1} is probit), we have z normal and the error (e_i) logistic. If we use the logit/logit model or the logit/probit model what distribution should be supposed for the variable x_i? Given that we assume that the distribution of x_i is the distribution of the sum-of-two logistics or the distribution of the sum of a normal plus a logistic random variable, how close is the bivariate C-type sum-of-two logistic, for example, to the bivariate normal distribution? The answer to these questions is the main purpose of this chapter. Before comparing the C-type distribution we shall obtain the distribution of the sum of two logistic random variables.

<u>The distribution of the sum of two independent logistic random variables</u>

Let X_i, i=1,2 be i.i.d.r.v's with standard logistic distribution. Then

$$F(x_i) = \frac{1}{(1+e^{-x_i})} \qquad -\infty<x_i<\infty$$

$$f(x_i) = F(x_i)[1-F(x_i)] = \frac{e^{-x_i}}{(1+e^{-xi})^2}$$

$$E(X_i) = 0 \text{ and } Var(X_i) = \Pi^2/_3$$

The logistic distribution and its properties are completely described in Johnson and Kotz (1970, Chapter 22).

The distribution of the sum of two independent random variables can be obtained by the distribution function method, that is:

Let $Y = X_1 + X_2$

then

$$F(y) = \int_{-\infty}^{\infty} F_{X_2}(y-x_1)\, f_{X_1}(x_1)\, dx_1$$

$$= \int_{-\infty}^{\infty} \frac{1}{[1+e^{-(y-x_1)}]} \cdot \frac{e^{-x_1}}{(1+e^{-x_1})^2}\, dx_1$$

Substituting $z = \frac{1}{1+e^{-x_1}}$, we have

$$F(y) = \int_0^1 \frac{1}{[1+e^{-y}z/(1-z)]}\, dz$$

$$F(y) = \int_0^1 \frac{1-z}{1+(e^{-y}-1)z}\, dz$$

$$= \int_0^1 \frac{1}{1+(e^{-y}-1)z}\, dz - \int_0^1 \frac{z}{1+(e^{-y}-1)z}\, dz \qquad (2.16)$$

Using results of integration for rational functions (see, for example, Maxwell, 1954, Vol.II) we have

$$\int \frac{1}{1+(e^{-y}-1)z}\, dz = \frac{1}{(e^{-y}-1)} \log|1 + (e^{-y}-1)z|$$

The second integral in (2.16) may be solved considering the identity

$$\frac{z}{1+(e^{-y}-1)z} = \frac{1}{(e^{-y}-1)} - \frac{1}{(e^{-y}-1)|1+(e^{-y}-1)z|}$$

Therefore

$$F(y) = \frac{(-y)}{(e^{-y}-1)} - \frac{1}{(e^{-y}-1)} + \frac{1}{(e^{-y}-1)} \cdot \frac{(-y)}{(e^{-y}-1)}$$

or

$$F(y) = \frac{1}{(1-e^{-y})} - \frac{ye^{-y}}{(1-e^{-y})^2} \qquad -\infty<y<\infty \qquad (2.17)$$

which is the distribution function of the sum of two independent standard logistic random variables.

The corresponding density function is obtained by differentiation, then

$$f(y) = \frac{ye^{-y}(1+e^{-y})}{(1-e^{-y})^3} - \frac{2e^{-y}}{(1-e^{-y})^2} \qquad -\infty<y<\infty \qquad (2.18)$$

The characteristic function of the standard logistic distribution is given by $E(e^{itx}) = \pi t \operatorname{cosech} \pi t$. Therefore, the characteristic function of the sum of two independent logistic r.v. is given by

$$E(e^{ity}) = (\pi t \operatorname{cosech} \pi t)^2$$

The mean and variance of $Y = X_1 + X_2$ are

$$E(Y) = 0 \text{ and } Var(Y) = \frac{2\Pi^2}{3}$$

The distribution of Y is symmetrical about $Y = 0$ and all moments of order odd are zero.

The moment generating function of the sum of two logistics is given by

$$M_{Y_2}(\theta) = M_{X_1+X_2}(\theta) = (\Pi\theta \text{ cosec } \Pi\theta)^2 = [\Gamma(\theta+1)\Gamma(1-\theta)]^2$$

Using the expression

$$E(Y^r) = \left| \frac{\partial^r \log M_Y(\theta)}{\partial\theta} \right|_{\theta=0}$$

we obtain

$$E(Y^4) = \frac{\Pi^4}{4}$$

therefore the first two moment-ratios of the distribution of the sum of two logistics are

$$\sqrt{\beta_1} = \alpha_3 = 0$$

$$\beta_2 = \alpha_4 = 3 + \frac{E(X^4)}{\sigma^4} = 3.5625 .$$

Finally, on using results of hyperbolic functions it can be shown that the density function of the sum of two independent standard logistic random variables can be written as

$$f(y) = (1/2)\text{cosech}^2(y/2)[(y/2)\cosh(y/2)\text{cosech}(y/2) - 1] \quad -\infty<y<\infty$$

The distribution of the sum of three or more independent logistic random variables

Let $Y_3 = X_1 + X_2 + X_3 = Y_2 + X_3$

where $Y_2 = X_1 + X_2$

and X_i $i=1,2,3$ are standard logistic i.i.d.r.v's.

Then, using the function distribution method, we have

$$F_3(y) = \int_{-\infty}^{\infty} F_{Y_2}(y-x_3) f_{X_3}(x_3)\,dx_3$$

$$= \int_{-\infty}^{\infty} \left| \frac{1}{1-e^{-(y-x_3)}} - \frac{(y-x_3)e^{-(y-x3)}}{[1-e^{-(y-x_3)}]^2} \right| \frac{e^{-x_3}}{(1+e^{-x_3})^2}\,dx_3$$

Using a numerical integration routine we can obtain the distribution function of the sum of three logistics (We have used NAG routine; Gauss-Hermite quadrature with 46 nodes).

The mean and variance of Y_3 can be easily obtained:

$$E(Y_3) = \sum_{i=1}^{3} E(X_i) = 0$$

$$Var(Y_3) = \sum_{i=1}^{3} Var(X_i) = \pi^2$$

The characteristic function of the sum of three or more independent logistic random variables is also easily obtained.

Let $Y_n = \sum_{i=1}^{n} X_i$

then

$$E(e^{ity_n}) = (\pi t \text{ cosech } \pi t)^n$$

and the moment generating function is given by

$$E(e^{\theta y_n}) = My_n(\theta) = (\pi\theta \text{ cosec } \pi\theta)^n = [\Gamma(\theta+1)\Gamma(1-\theta)]^n.$$

Comparison of the distribution of the sum of two logistics and sum of three logistics with the univariate normal distribution

In order to compare the distribution of sum of logistic variables with the normal distribution function we need to rescale the variables such that in each case we have mean zero and variance one. We shall use Y/σ_y when comparing the distributions.

Let $F_n(y)$ be the distribution function of the sum of n independent logistic random variables. So $F_1(y)$ is the d.f. of the logistic, $F_2(y)$ is the d.f. of the sum of two logistics, $F_3(y)$ is the d.f. of the sum of three logistics.

In Table 2.3 we present the values of the distribution function $F_n(y)$, n=1,2,3 for some values of Y_n/σ_n together with the corresponding values of the normal distribution.

TABLE 2.3 : Comparison of some values of the distribution functions for normal, logistic and sum of logistics.

X/σ_X	Logistic D.F.	Sum of two Logistics D.F.	Sum of three Logistics D.F.	Normal D.F.
0.0	.50000	50000	50000	50000
0.2	.58971	58476	58300	57926
0.4	.67382	66522	66255	65542
0.6	.74806	73784	73404	72575
0.8	.81016	80033	79707	78815
1.0	.85982	85179	84912	84135
1.2	.89812	89252	89075	88493
1.4	.92685	92366	92284	91924
1.6	.94795	94675	94675	94520
1.8	.96320	96344	96404	96407
2.0	.97411	97524	97619	97725
2.2	.98184	98343	98452	98610
2.4	.98730	98902	99011	99180
2.6	.99113	99278	99378	99534
2.8	.99381	99529	99615	99745
3.0	.99569	99695	99765	99865
3.2	.99699	99804	99858	99931
3.4	.99791	99874	99915	99966
3.6	.99854	99920	99950	99984
3.8	.99899	99949	99971	99993
4.0	.99929	99968	99983	99997

The logistic distribution has a shape similar to that of the normal. According to Johnson & Kotz (1970), if the cumulative distribution functions

$$\phi(x) = \frac{1}{\sqrt{2\Pi}} \int_{-\infty}^{\infty} e^{-u^2} du$$

and

$$F(\frac{\Pi}{\sqrt{3}} x) = \frac{1}{1+exp(-\Pi x/\sqrt{3})}$$

of the standardized normal and logistic distribution respectively, are compared, we have

$$|F(\frac{\Pi}{\sqrt{3}}) - \Phi(x)| < 0.0228$$

the maximum value of the difference attained when x=0.7. Also it is shown that this maximum may be reduced to a value less than 0.01 by changing the scale of x in F, that is

$$|F[15/16(\pi/\sqrt{3})x] - \phi(x)| < 0.01$$

Comparing now the cumulative distribution function of the sum of two logistics(F_2) with the normal we observe that

$$|F_2[(\sqrt{2\pi}/\sqrt{3})x] - \phi(x)| < 0.0125$$

the maximum difference, d = 0.012404, attained also when x = 0.7. This result shows that the distribution of the sum of two logistic is more similar to the normal distribution than the logistic distribution. Actually, when compared in a graph the differences are imperceptible. We then show in Figure 2.1 the differences $F_2(x\sqrt{2\pi}/\sqrt{3})-\phi(x)$. The differences $F_2(x(\sqrt{2\pi}/\sqrt{3})15/16) - \phi(x)$ are also shown in Figure 2.1.

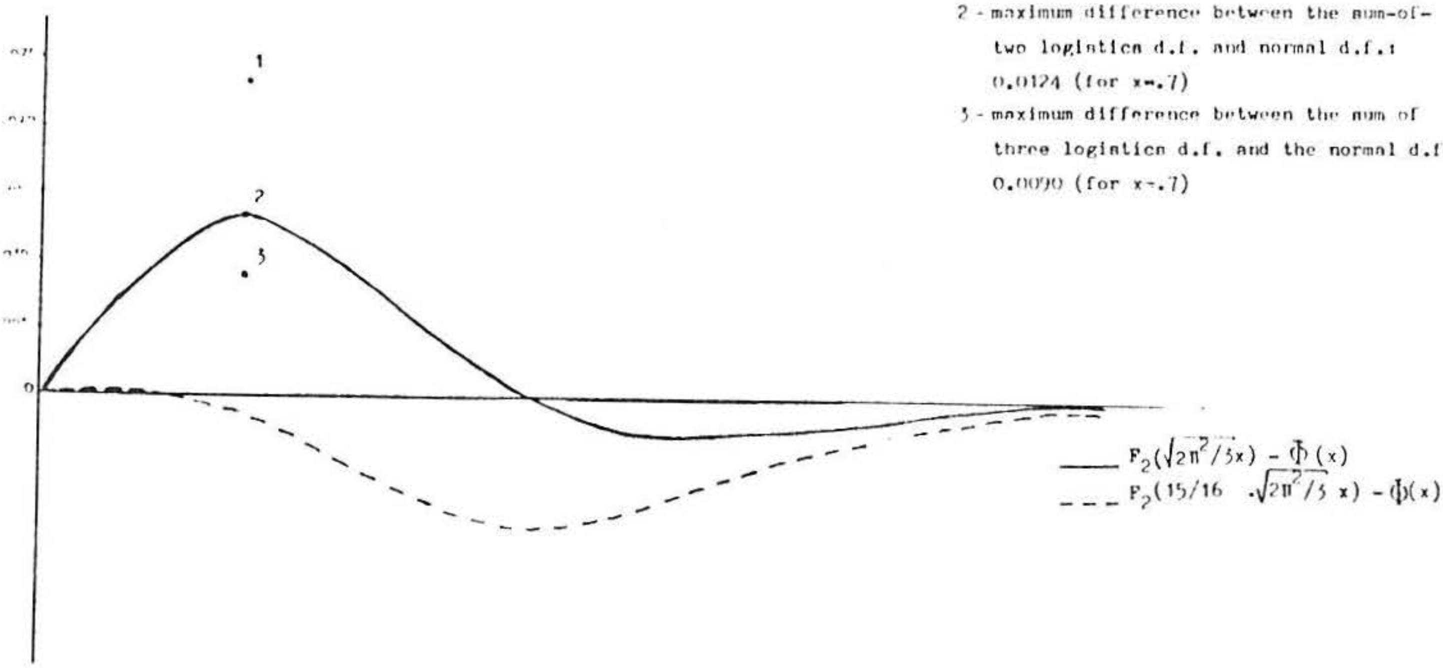
In Figure 2.2 we show the differences between the cumulative distributions of the sum of three logistics and the normal distribution function $\phi(x)$. In this case

$$|F_3(\Pi x) - \phi(x)| < 0.009034 \quad \text{for all } x$$

the maximum attained at $x = 0.7$. This shows the expected result that the sum of three logistics has the distribution more similar to the normal than the sum of two logistics or the logistic.

The differences between the normal and the sum of logistics increase when the scale of x is changed by multiplying the factor 15/16. Therefore this procedure is not recommended when we have sum of logistics (observe the increasing differences around the point $x = 1.8$ in Figures 2.1 and 2.2.

FIGURE 2.1 - COMPARICON OF THE SUM-OF-TWO LOGISTICS D.F. WITH THE NORMAL D.F.

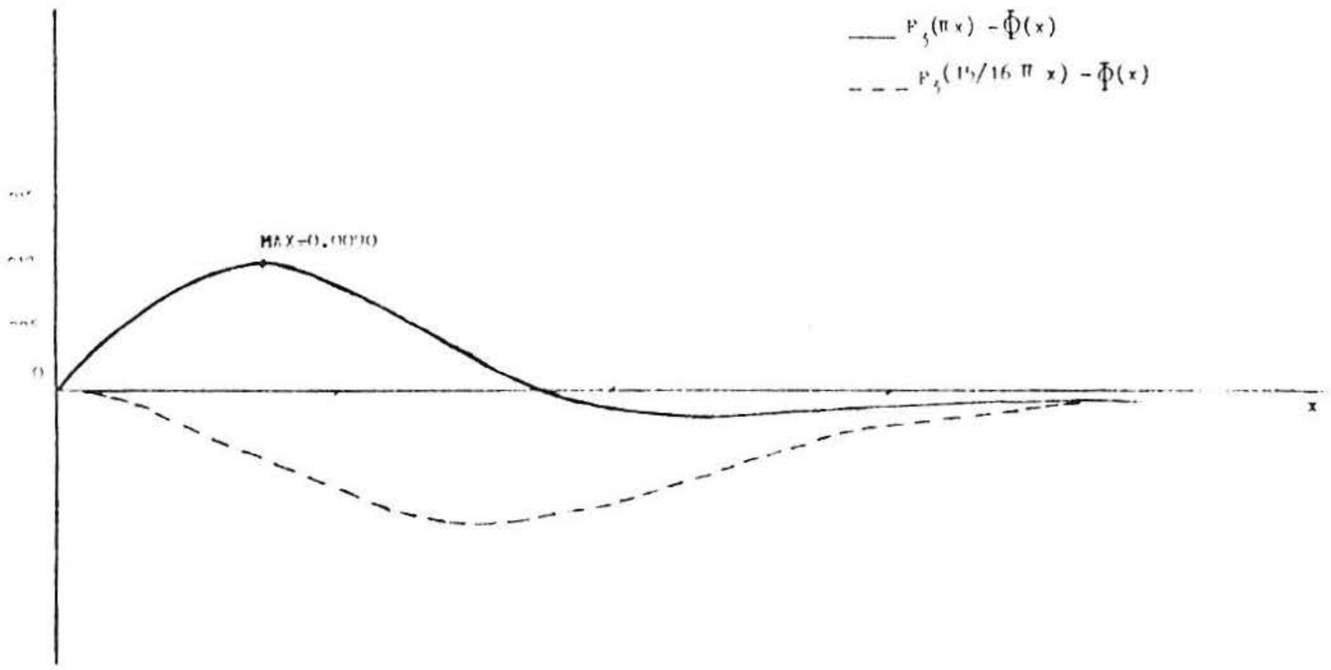


1 maximum difference between the logistic df. and the normal d.f.:0.0228(for x=.7)

2 - maximum difference between the sum-of-two logistics d.f. and normal d.f.: 0.0124 (for x=.7)

3 - maximum difference between the sum of three logistics d.f. and the normal d.f. 0.0090 (for x=.7)

FIGURE 2.2 - COMPARICON OF THE SUM OF THREE LOGISTICS D.F. WITH THE NORMAL D.F.



2.6 Comparison of the C-type normal with the bivariate normal distributions

Since the C-type distribution or surface of constant association was proposed by Pearson at the beginning of the century, several suggestions of how to relate the parameter ψ of this distribution with the correlation parameter ρ of the bivariate normal, have appeared in the literature.

Pearson (1913) compares the surface of constant association for normal margins and $Q = 0.6$, where Q is Yule's coefficient of association, $Q = (\psi-1)/(\psi+1)$, with the bivariate normal distribution with parameter $\rho = 0.5$. That means that he compares the C-type normal with $\psi = 4$ with the normal surface with $\rho = 0.5$.

Plackett (1965) shows that the C-type normal and the normal surface agree well on relating ψ and ρ by

$$\rho = \cos\{\pi/(1+\sqrt{\psi})\} \qquad (2.19)$$

which also gives $\rho = 0.5$ for $\psi = 4$ (or $\rho = -0.5$ for $\psi = 1/4$).

Mardia (1967) suggests that the most natural way of comparison between the two distributions should be relating ρ with $\rho_N(\psi)$ where $\rho_N(\psi)$ is the correlation coefficient of the C-type normal distribution, but he concludes that this method is not so good as Plackett's method. Mardia proposes, then, another expression for relating ρ and ψ which is:

$$\rho = 2 \sin\{ \frac{\pi}{6} \rho_U(\psi)\} \qquad (2.20)$$

where $\rho_U(\psi)$ is the correlation coefficient of the C-type uniform distribution and is given by

$$\rho_U(\psi) = (\psi^2-1-2\psi \log \psi)/(\psi-1)^2$$

Mardia (1967) concludes that this method is better than Plackett's except for the points in the neighbourhood of (0,0). Using (2.19) for

$\rho = 0.5$ we have $\psi = 5$ or $\rho = -0.5$ for $\psi = 0.20$.

We have already discussed in Section 2.3 that the Chambers coefficient (Chamber, 1982) given by

$$r_{.74} = \frac{\psi^{0.74} - 1}{\psi^{0.74} + 1} \qquad (2.21)$$

is a very good estimator of the correlation coefficient of the normal bivariate normal distribution. We, therefore, propose relating ψ and ρ by using the expression $\rho = r_{.74}$. For $\rho = -0.5$ we have $\psi = 0.2266$.

In Table 2.4 we present the cumulative distribution function of the normal bivariate with $\rho = -0.5$ for some values of x and y. We compare the normal probabilities with those of the C-type normal distribution function $H(x,y;\psi)$ for three different values of ψ:

a) $\psi = 0.20$ obtained by using expression (2.20) for relating ρ and ψ;

b) $\psi = 0.2266$, using expression (2.21) and

c) $\psi = 0.25$ obtained by expression (2.19) for $\rho = -0.5$.

Table 2.4 : Comparison of the bivariate normal d.f. with the C-type normal d.f. for different parameters of association.

First entry : Bivariate normal d.f. with $\rho = -0.5$
Second entry : C-type normal d.f. with $\psi = 0.20$
Third entry : C-type normal d.f. with $\psi = 0.2266$
Fourth entry : C-type normal d.f. with $\psi = 0.25$

x y	0.0	0.5	1.0	1.5
-1.5	.0092 .0120 .0132 .0143	.0200 .0225 .0242 .0256	.0344 .0371 .0389 .0403	.0485 .0517 .0528 .0537
-1.0	.0313 .0318 .0347 .0371	.0612 .0597 .0635 .0665	.0961 .0964 .0996 .1022	.1262 .1289 .1308 .1321
-0.5	.0817 .0747 .0801 .0844	.1452 .1376 .1433 .1478	.2111 .2096 .2133 .2164	.2617 .2642 .2659 .2659
0.0	.1667 .1545 .1613 .1667	.2731 .2662 .2716 .2759	.3726 .3732 .3760 .3785	.4424 .4452 .4464 .4475
0.5	.2731 .2662 .2716 .2759	.4192 .4186 .4220 .4248	.5452 .5482 .5498 .5513	.6279 .6306 .6312 .6318
1.0	.3726 .3732 .3760 .3785	.5452 .5482 .5498 .5513	.6865 .6894 .6901 .6909	.7754 .7771 .7774 .7777
1.5	.4424 .4452 .4464 .4475	.6279 .6306 .6312 .6318	.7754 .7771 .7774 .7777	.8665 .8674 .8675 .8676

Observing Table 2.4 we conclude that, regardless of the method chosen for relating the parameters, the bivariate normal and the C-type normal are similar. Our suggestion (third entry in Table 2.4) using the expression $\rho = r_{.74}$ for relating the parameters of the distributions has the advantage of giving better values (more similar to those of the normal surface) in the neighbourhood of (0,0) compared with Mardia's suggestion (second entry) and Mardia's method is better than Plackett's method (fourth entry) everywhere except for the points around (0,0).

2.7 Comparison of the C-type logistic, C-type sum-of-two logistics and C-type sum of three logistics with the bivariate normal distribution

It has been shown in the last section that the C-type normal distribution is very similar to the bivariate normal and we have suggested using $\rho = r_{.74}$ for relating the parameters. We also have shown that the marginal distributions, logistic, sum-of-two logistics, sum of three logistics are very similar to the normal distribution (see Table 2.4). We now compare the C-type distributions with the bivariate normal. We shall use the same parameter $\psi = 0.2266$ for all members of the C-type family.

Table 2.5 : Comparison of some members of the C-type distribution family (ψ = 0.2266) with the normal bivariate distribution function (ρ = -0.5)

First entry : Normal bivariate d.f.
Second entry : C-type normal d.f.
Third entry : C-type sum of 3 logistics d.f.
Fourth entry : C-type sum of 2 logistics d.f.
Fifth entry : C-type logistic d.f.

y/σ_y	0.0	0.5	1.0	1.5
-1.5	.0092 .0134 .0127 .0126	.0200 .0242 .0239 .0239	.0344 .0389 .0384 .0384	.0485 .0528 .0512 .0509
-1.0	.0313 .0348 .0328 .0321 .0301	.0612 .0635 .0613 .0607 .0586	.0961 .0997 .0964 .0953 .0918	.1262 .1308 .1249 .1227 .1164
-0.5	.0817 .0801 .0774 .0762 .0787	1452 1433 1418 1410 1390	2111 2134 2113 2098 2061	2617 2660 2604 2573 2496
0.0	1667 1613 1613 1613 1613	2731 2715 2765 2789 2852	3726 2761 3819 3839 3899	4424 4464 4484 4487 4504
0.5	2731 2716 2765 2789 2852	4192 4220 4347 4407 4571	5252 5498 5638 5694 5856	6279 6312 6409 6447 6561
1.0	3726 2761 3819 3839 3899	5452 5499 5638 5695 5856	6865 6901 7049 7100 7253	7454 7774 7875 7905 8004
1.5	4424 4464 4484 4487 4504	6279 6312 6409 6447 6561	7754 7774 7875 7905 8004	8665 8675 8725 8733 8774

CXCAAR

In Table 2.5 we show the values of the cumulative distribution functions of the C-type family and the bivariate normal d.f. for values of x/σ_x equal to 0; 0.5; 1.0; 1.5 and for values of y/σ_y from -1.5 to 1.5 (0.5). Observing the table we conclude that the C-type distributions are similar to the bivariate normal. For a better approximation of the C-type logistic we could multiply the value of the marginal random variables by 15/16 as it is done in the univariate case.

We have compared other members of the C-type family considered as marginal distributions, mixture of distributions such as the sum of the logistic plus normal random variables, the sum of two logistics plus normal and the sum of two normals plus logistic. All these marginal distributions were obtained using numerical integration routines in the same way as we obtained the distribution of the sum of three logistics. Comparing the various members of the C-type family, considering the mixture of distribution above described, with the bivariate normal, we have observed similar results.

Our conclusions in this section provide arguments for using the C-type distribution as an underling model for Factor Analysis of Categorical Data to be discussed in Chapter 4.

CHAPTER 3 - MAXIMUM LIKELIHOOD ESTIMATION OF THE PARAMETER OF ASSOCIATION OF THE C-TYPE DISTRIBUTION FOR DATA GIVEN IN AN RXC TABLE

3.1 Introduction

Suppose we observe two ordinal variables U^* and V^*, that are classified in R and C categories respectively. A cross-tabulation of U^* and V^* gives the observed frequencies as denoted in Table 3.1. We further assume that underlying U^* and V^* there are some latent continuous variables X and Y with a joint bivariate C-type distribution.

Table 3.1 - A cross tabulation of U^* and V^*: observed frequencies

			y		
U^* \ V^*	1	2	b	c	
1	n_{11}	n_{12}	n_{1b}	n_{1c}	$n1.$
2	n_{21}	n_{22}	n_{2b}	n_{2c}	$n2.$
.	.	.	.		
.	.	.	.		
.	.	.	.		
a	n_{a1}	n_{a2}	n_{ab}	n_{ac}	
x					
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
r	n_{r1}	n_{r2}	n_{rb}	n_{rc}	$nr.$
	$n.1$	$n.2$	$n.b$	$n.c$	n

Suppose that F(x) and G(y) are the distribution functions of X and Y respectively and H(x,y) is their joint distribution function. Suppose that the forms of the F(x) and G(y) are known.

Let p_{ij} be the probability that an observation falls in cell (i,j), $i = 1,\ldots,r$; $j = 1,\ldots,c$.

Let f_i, $i = 1,2,..,R$ and g_j, $j = 1,...,C$ be the marginal probabilities in the contingency table, such that

$$F_a = \sum_{i=1}^{a} f_i = P(U^* \leq a) \qquad a = 1,\ldots,r-1$$

$$G_b = \sum_{j=1}^{b} g_j = P(V^* < b) \qquad b = 1,\ldots,c-1$$

where a and b are the categories determined by the cut of the distribution by lines parallel to the axes X and Y. In other words, we are supposing a fourfold table determined by the point (x,y) or by a dichotomy of the variables U^* and V^* at categories a and b respectively. Table 3.2 presents the probabilities supposing given margins $F(x)$ and $G(y)$.

Let $H_{ab} = \sum_{i=1}^{a} \sum_{j=1}^{b} p_{ij}$ $\qquad a = 1,\ldots,r-1$
$\qquad b = 1,\ldots,c-1$

Table 3.2 - RxC contingency table: probabilities for given margins $F(x)$ and $G(y)$

U^* \ V^*	1	2	b		
1	p_{11}	p_{12} ···	p_{1b}	$f_1 - \sum_{j=1}^{b} p_{1j}$	f_1
2	p_{21}	p_{22} ···	p_{2b}	$f_2 - \sum_{j=1}^{b} p_{2j}$	f_2
	⋮	⋮	⋮	⋮	⋮
a	p_{a1}	p_{a2} ···	p_{ab}	$f_a - \sum_{j=1}^{b} p_{aj}$	f_a
	g_{1-} $- \sum_{i=1}^{a} p_{i1}$	···	g_{b-} $- \sum_{i=1}^{a} p_{ib}$	$1 - \sum_{i=1}^{a} f_i - \sum_{j=1}^{b} g_j$ $+ \sum_{i=1}^{a} \sum_{j=1}^{b} p_{ij}$	$1 - \sum_{i=1}^{a} f_i$
	g_1	···	g_b	$1 - \sum_{j=1}^{b} g_j$	1

We are supposing that X and Y have a bivariate C-type distribution or in other words, that the cross product ratios ψ_{ab} $a = 1,\dots,r-1$; $b = 1,2,\dots,c-1$ for dichotomies at the categories a and b of the variables U^* and V^* are constant and equal to ψ, thus

$$\frac{\left(\sum_{i=1}^{a}\sum_{j=1}^{b} p_{ij}\right)\left(1 - \sum_{i=1}^{a} f_i - \sum_{j=1}^{b} g_j + \sum_{i=1}^{a}\sum_{j=1}^{b} p_{ij}\right)}{\left(\sum_{i=1}^{a} f_i - \sum_{i=1}^{a}\sum_{j=1}^{b} p_{ij}\right)\left(\sum_{j=1}^{b} g_j - \sum_{i=1}^{a}\sum_{j=1}^{b} p_{ij}\right)} = \psi$$

$$a = 1,2,\dots,r-1, \quad \psi > 0$$
$$b = 1,2,\dots,c-1$$

or, using the notation given above

$$\frac{H_{ab}(x,y)(1 - (F_a(x) + G_b(y)) + H_{ab}(x,y))}{(F_a(x) - H_{ab}(x,y))(G_b(y) - H_{ab}(x,y))} = \psi$$

$$a = 1,\dots,r-1$$
$$b = 1,\dots,c-1$$
$$\psi > 0$$

and the C-type distribution is then given by

$$H_{ab}(x,y) = \begin{cases} [S_{ab} - \{S_{ab}^2 - 4\psi(\psi-1)F_a(x)G_b(y)\}^{\frac{1}{2}}]/[2(\psi-1)], & (\psi \neq 1) \\ F_a(x)G_b(y), \quad (\psi=1) & (3.1) \end{cases}$$

where $S_{ab} = 1 + (\psi-1)[F_a(x) + G_b(y)]$; $a = 1,\dots,r-1$
$b = 1,\dots,c-1$

It is our purpose to estimate the parameter of association ψ of this distribution on using the method of maximum likelihood. Before proceeding, we shall derive the expressions for the expected proportions of the cell (i,j) $i = 1,\dots,r$, $j = 1,\dots,c$, supposing a C-type population underlying the data given in a RxC contingency table.

For simplicity of notation, we shall use $H_{ab}(x,y) = H_{ab}$, $F_a(x) = F_a$, $G_b(y) = G_b$ and

$$H_{ab} = (S_{ab} - A^{\frac{1}{2}}_{ab})/\{2(\psi-1)\} \qquad \begin{array}{l} a = 1,\dots,r-1 \\ b = 1,\dots,c-1 \end{array} \qquad (3.2)$$

where

$$S_{ab} = 1 + (\psi-1)(F_a+G_b)$$

and

$$A_{ab} = S^2_{ab} - 4\psi(\psi-1)F_aG_b$$

By using the expression (3.2) we can calculate the expected proportion in the cell (1,1)

$$p_{11} = H_{11} = (S_{11}-A^{\frac{1}{2}}_{11})/[2(\psi-1)] \qquad (3.3)$$

where $S_{11} = 1 + (\psi-1)(f_1+g_1)$

$A_{11} = S_{11} - 4\psi(\psi-1)f_1g_1$

For the cells (1,2) and (2,1) we have

$$p_{12} = H_{12} - H_{11} = [(S_{12}-A^{\frac{1}{2}}_{12}) - (S_{11}-A^{\frac{1}{2}}_{11})]/[2(\psi-1)]$$

$$p_{21} = H_{21} - H_{11} = [(S_{21}-A^{\frac{1}{2}}_{21}) - (S_{11}-A^{\frac{1}{2}}_{11})]/[2(\psi-1)]$$

In general, for the cells (1,b), b = 2,...,c-1 we have

$$p_{1b} = H_{1b} - H_{1,b-1} = [(S_{1b}-A^{\frac{1}{2}}_{1b}) - (S_{1,b-1}-A^{\frac{1}{2}}_{1,b-1})]/[2(\psi-1)]$$

$$b = 2,3,\dots,c-1 \qquad (3.4)$$

and for the cells (a,1), a = 2,...,r-1 we have

$$p_{a1} = H_{a1} - H_{a-1,1} = [(S_{a1} - A_{a1}^{\frac{1}{2}}) - (S_{a-1,1} - A_{a-1,a}^{\frac{1}{2}})]/[2(\psi-1)], \qquad (3.5)$$

$$a = 2,\ldots,r-1$$

The expected proportion in the cell (a,b), $a = 2,\ldots,r-1$ and $b = 2,\ldots,c-1$ is given by

$$p_{ab} = H_{ab} - H_{a-1,b} - H_{a,b-1} + H_{a-1,b-1} \qquad (3.6)$$

$$a = 2,\ldots,r-1$$
$$b = 2,\ldots,c-1$$

where H_{ab} is given by (3.2).

The expression (3.6) can be written as:

$$p_{ab} = \sum_{i=1}^{a}\sum_{j=1}^{b} p_{ij} - \sum_{i=1}^{a-1}\sum_{j=1}^{b} p_{ij} - \sum_{i=1}^{a}\sum_{j=1}^{b-1} p_{ij} + \sum_{i=1}^{a-1}\sum_{j=1}^{b-1} p_{ij}$$

$$a = 2,\ldots,r-1$$
$$b = 2,\ldots,c-1$$

Finally, the expected proportions in the last row and column of the RxC table are given by

$$p_{r1} = g_1 - H_{r-1,1}$$

$$p_{rb} = H_{rb} - H_{r-1,b} - H_{r,b-1} + H_{r-1,b-1} = g_b - H_{r-1,b} + H_{r-1,c-1}$$

$$b = 1,\ldots,c-1$$

$$p_{1c} = f_1 - H_{1,c-1}$$

$$p_{ac} = f_a - H_{a,c-1} - H_{a-1,c-1}$$

$$a = 1,\ldots,r-1$$

$$p_{rc} = 1 - \sum_{i=1}^{a} f_i - \sum_{j=1}^{b} g_j + H_{r-1,c-1} \qquad (3.7)$$

The expected frequencies in the cells (i,j) of an RxC contingency table are given by np_{ij} where n is the sample size and p_{ij}, i = 1,...,r, j = 1,...,c are given by the expressions (3.3) to (3.7) and are function of ψ. The marginal proportions f_a, a = 1,...,r-1 and g_b, b = 1,...,c-1 are estimated by equating observed and expected marginal proportions. The observed marginal proportions are the maximum likelihood estimates of the expected marginal proportions (see Kendall and Stuart, 1979, Vol.2, pg 445). Considering the results presented by Olsson (1979) as explained in section 2.2, we shall follow here the "two-step maximum likelihood" estimation approach.

In order to estimate the parameter ψ by the maximum likelihood method we shall now consider the derivative of H_{ab} with respect to ψ. From (3.2) we have

$$H_{ab} = \frac{1+(\alpha-1)(F_a+G_b)-\{[1+(\psi-1)(F_a+G_b)]^2-4\psi(\psi-1)F_aG_b\}^{\frac{1}{2}}}{2(\psi-1)}$$

$$a = 1,\ldots,r-1$$
$$b = 1,\ldots,c-1$$

so that

$$\frac{\partial H_{ab}}{\partial\psi} = \frac{1}{2(\psi-1)^2}\left[\frac{1+(\psi-1)(F_a+G_b)-2(\psi-1)F_aG_b}{\{[1+(\psi-1)(F_a+G_b)]^2-4\psi(\psi-1)F_aG_b\}^{\frac{1}{2}}} - 1\right]$$

and using the simplified notation we have

$$H'_{ab} = \frac{1}{2(\psi-1)^2}\left[\frac{S_{ab}-2(\psi-1)F_aG_b}{A_{ab}^{\frac{1}{2}}} - 1\right] \qquad \begin{array}{l} a = 1,\ldots,r-1 \\ b = 1,\ldots,c-1 \end{array} \qquad (3.8)$$

where a prime denotes differentiation with respect to ψ,

$S_{ab} = 1+(\psi-1)(F_a+G_b)$
$A_{ab} = S_{ab}^2-4\psi(\psi-1)F_aG_b$

3.2 Derivation of the Likelihood Equation

The data are given in an RxC contingency table with observed frequencies n_{ij} as given in Table 3.1. The probability p_{ij} that an observation falls into cell (i,j) are given in section 3.1 above. Therefore the likelihood of the sample is

$$L = C \prod_{i=1}^{r} \prod_{j=1}^{c} p_{ij}^{n_{ij}}$$

where C is a constant which does not depend on the parameter ψ to be estimated and $\sum_i \sum_j p_{ij} = 1$.

Taking logarithms, we have

$$\ell = \ell n\ L = \ell n\ C + \sum_{i=1}^{R} \sum_{j=1}^{C} n_{ij}\ \ell n\ p_{ij}$$

and differentiating with respect to ψ we obtain

$$\frac{\partial \ell}{\partial \psi} = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{n_{ij}}{p_{ij}} \frac{\partial p_{ij}}{\partial \psi} \qquad (3.9)$$

Although we have the required formulae for deriving the log likelihood equation, we shall need a computationally more manageable form. In order to obtain a general expression for the likelihood equation for the RxC table case, we shall show how it is obtained from tables of dimensions 2x3, 3x3 for example. The general expression for the maximum likelihood equation for estimating ψ from data given in a RxC table is then obtained by direct generalization of the expression for tables of lower dimensions.

2x3 contingency tables

The necessary formilae for estimating ψ from 2x3 contingency tables are summarised in table 3.3.

On using (3.9) and the expressions given in table 3.3 we have

$$\frac{\partial \ell}{\partial \psi} = \frac{n_{11}}{p_{11}} H'_{11} + \frac{n_{12}}{p_{12}} (H'_{12}-H'_{11}) + \frac{n_{21}}{p_{21}} (-H'_{11}) + \frac{n_{22}}{p_{22}} (-H'_{12}+H'_{11}) + $$
$$+ \frac{n_{12}}{p_{12}} (-H'_{12}) + \frac{n_{23}}{p_{23}} H'_{12}$$

Table 3.3 - Necessary formulae for estimating ψ from 2x3 tables

Observed frequencies n_{ij}	Expected proportions p_{ij}	$\frac{\partial p_{ij}}{\partial \psi}$
n_{11}	$p_{11} = H_{11}$	H'_{11}
n_{12}	$p_{12} = H_{12}-H_{11}$	$H'_{12}-H'_{11}$
n_{21}	$p_{21} = g_1-H_{11}$	$-H'_{11}$
n_{22}	$p_{22} = g_2-H_{12}+H_{11}$	$-H'_{12}+H'_{11}$
n_{13}	$p_{13} = f_1-H_{12}$	$-H'_{12}$
n_{23}	$p_{23} = 1-f_1+g_1+g_2+H_{12}$	H'_{12}

so that

$$\frac{\partial \ell}{\partial \psi} = \left(\frac{n_{11}}{p_{11}} - \frac{n_{12}}{p_{12}} - \frac{n_{21}}{p_{21}} + \frac{n_{22}}{p_{22}} \right) H'_{11} + \left(\frac{n_{12}}{p_{12}} - \frac{n_{22}}{p_{22}} - \frac{n_{12}}{p_{12}} + \frac{n_{23}}{p_{23}} \right) H'_{12}$$

<u>3x3 contingency tables</u>

The necessary formulae for estimating ψ from 3x3 contingency table are summarised in table 3.4.

The log likelihood function for 3x3 tables is obtained using the expression (3.9) and the formulae given in table 3.4. Thus

$$\frac{\partial \ell}{\partial \psi} = \frac{n_{11}}{p_{11}} H'_{11} + \frac{n_{12}}{p_{12}} (H'_{12}-H'_{11}) + \frac{n_{21}}{p_{21}} (H'_{21}-H'_{11}) + \frac{n_{22}}{p_{22}} (H'_{22}-H'_{12}-H'_{21}+H'_{11})$$

$$+ \frac{n_{13}}{p_{13}} (-H'_{12}) + \frac{n_{31}}{p_{31}} (-H'_{21}) + \frac{n_{23}}{p_{23}} (-H'_{22}+H'_{12}) + \frac{n_{32}}{p_{32}} (-H'_{22}+H'_{21}) +$$

$$+ \frac{n_{33}}{p_{33}} H'_{22}$$

Table 3.4 - Formulae for estimating ψ from 3x3 contingency tables

Observed frequencies n_{ij}	Expected proportions p_{ij}	$\frac{\partial p_{ij}}{\partial \psi}$
n_{11}	$p_{11} = H_{11}$	H'_{11}
n_{12}	$p_{12} = H_{12}-H_{11}$	$H'_{12}-H'_{11}$
n_{21}	$p_{21} = H_{21}-H_{11}$	$H'_{21}-H'_{11}$
n_{22}	$p_{22} = H_{22}-H_{12}-H_{21}+H_{11}$	$H'_{22}-H'_{12}-H'_{21}+H'_{11}$
n_{13}	$p_{13} = f_1-H_{12}$	$-H'_{12}$
n_{31}	$p_{31} = g_1-H_{21}$	$-H'_{21}$
n_{23}	$p_{23} = f_2-H_{22}+H_{12}$	$-H'_{22}+H'_{12}$
n_{32}	$p_{32} = g_2-H_{22}+H_{21}$	$-H'_{22}+H'_{21}$
n_{33}	$p_{33} = 1- \sum_{i=1}^{2} f_i - \sum_{j=1}^{2} g_j + H_{22}$	H'_{22}

such that

$$\frac{\partial \ell}{\partial \psi} = \left(\frac{n_{11}}{p_{11}} - \frac{n_{12}}{p_{12}} - \frac{u_{21}}{p_{21}} + \frac{u_{22}}{p_{22}} \right) H'_{11} + \left(\frac{n_{12}}{p_{12}} - \frac{n_{22}}{p_{22}} - \frac{n_{13}}{p_{13}} + \frac{u_{23}}{p_{23}} \right) H'_{12}$$

$$+ \left(\frac{n_{21}}{p_{21}} - \frac{n_{22}}{p_{22}} - \frac{n_{31}}{p_{31}} + \frac{n_{32}}{p_{32}} \right) H'_{21} + \left(\frac{n_{22}}{p_{22}} - \frac{n_{23}}{p_{23}} - \frac{n_{32}}{p_{32}} + \frac{n_{33}}{p_{33}} \right) H'_{22}$$

or

$$\frac{\partial \ell}{\partial \psi} = \sum_{a=1}^{2} \sum_{b=1}^{2} \left[\left(\frac{n_{ab}}{p_{ab}} - \frac{n_{a,b+1}}{p_{a,b+1}} - \frac{n_{a+1,b}}{p_{a+1,b}} + \frac{n_{a+1,b+1}}{p_{a+1,b+1}} \right) H'_{ab} \right]$$

RxC contingency tables

In table 3.5 we present the necessary formulae for estimating ψ from RxC contingency tables.

Table 3.5 - Necessary formulae for estimating ψ from RxC tables

Observed frequencies n_{ij}	Expected proportions p_{ij}	$\frac{\partial p_{ij}}{\partial \psi}$
n_{11}	$p_{11} = H_{11}$	H'_{11}
n_{12}	$p_{12} = H_{12}-H_{11}$	$H'_{12}-H'_{11}$
n_{21}	$p_{21} = H_{21}-H_{11}$	$H'_{21}-H'_{11}$
Nab $a = 2,\dots,r-1$ $b = 2,\dots,c-1$ $r,c > 2$	Pab	$\frac{\partial p_{ab}}{\partial \psi}$
n_{a1}	$p_{a1} = H_{a1}-H_{a-1,1}$	$H'_{a1}-H'_{a-1,1}$
n_{1_b}	$p_{1_b} = H_{1_b}-H_{1,b-1}$	$H'_{1_b}-H'_{1_{b-1}}$
$\vdots$	$\vdots$	$\vdots$
n_{ab}	$p_{ab} = H_{ab}-H_{a-1,b}-H_{a,b-1}+H_{a-1,b-1}$	$H'_{ab}-H'_{a-1,b}-H'_{a,b-1}+H'_{a-1,b}$
$\vdots$	$\vdots$	$\vdots$
n_{ac}	$p_{ac} = f_a-H_{a,c-1}+H_{a-1.c-1}$	$-H'_{a,c-1}+H'_{a-1,c-1}$
$\vdots$	$\vdots$	$\vdots$
n_{rb}	$p_{rb} = g_b-H_{r-1,b}+H_{r-1,c-1}$	$-H'_{r-1,b}+H'_{r-1,c-1}$
$\vdots$	$\vdots$	$\vdots$
n_{1_c}	$p_{1_c} = f_1-H_{1,c-1}$	$-H'_{1,c-1}$
n_{r1}	$p_{r1} = g_1-H_{r-1,1}$	$-H'_{r-1,1}$
n_{rc}	$p_{rc} = 1-\sum_{i=j}^{a} f_i-\sum_{j=1}^{b} g_j+H_{r-1,c-1}$	$H'_{r-1,c-1}$

CUGAAA

Since the probabilities p_{ij}, $i = 1,2,\dots,r$, $j = 1,\dots,c$ are given in table 3.5 and also $\frac{\partial p_{ij}}{\partial\psi}$, $i = 1,\dots,r$; $j = 1,\dots,c$; from (3.2) and (3.6) we have

$$H_{ab} = (S_{ab} - A_{ab}^{\frac{1}{2}})/(2(\psi-1)) \qquad \begin{matrix} a = 1,\dots,r-1 \\ b = 1,\dots,c-1 \end{matrix}$$

$$H'_{ab} = \frac{1}{2(\psi-1)^2}\left[\frac{S_{ab}-2(\psi-1)F_a G_b}{A_{ab}^{\frac{1}{2}}}\right]$$

the loglikelihood equation given by (3.9) can be determined from Table 3.5 by multiplying appropriate elements and summing over i and j. The parameter ψ can then be estimated using an iterative method for solving the likelihood equation and by equating the observed and expected marginal proportions.

However, as we have seen in examples for small contingency tables, a more convenient expression can be obtained by observing that in the expansion of the expression (3.9) the elements H'_{ab} appears only in the terms corresponding to a fourfold table formed by cells (a,b) $(a,b+1)$, $(a+1,b)$, $(a+1,b+1)$; $a=1,\dots,r-1$, $b=1,\dots,c-1$. There are $(r-1)(c-1)$ such fourfold tables in an RxC contingency table, so we have after putting in evidence the element H_{ab}, a general term given by

$$\left(\frac{n_{ab}}{p_{ab}} - \frac{n_{a,b+1}}{p_{a,b+1}} - \frac{n_{a+1,b}}{p_{a+1,b}} + \frac{n_{a+1,b+1}}{p_{a+1,b+1}}\right)H'_{ab}$$

Summing over a and b and equating to zero we have the likelihood equation

$$\frac{\partial \ell}{\partial\psi} = \sum_{a=1}^{r-1}\sum_{b=1}^{c-1}\left[\left(\frac{n_{ab}}{p_{ab}} - \frac{n_{a,b+1}}{p_{a,b+1}} - \frac{n_{a+1,b}}{p_{a+1,b}} + \frac{n_{a+1,b+1}}{p_{a+1,b+1}}\right)H'_{ab}\right] = 0 \qquad (3.1$$

3.3 Solution of the likelihood equation

The solution of the likelihood equation is obtained by the iterative method known as 'the method of scoring for parameters' (Kendall and Stuart (1979), pg.52). The iterative procedure

$$\hat{\psi} = t-\left(\frac{\partial \ell nL}{\partial\psi}\right)_t / \left[E\left(\frac{\partial^2 \ell nL}{\partial\psi^2}\right)\right]_t = t+\left(\frac{\partial \ell nL}{\partial\psi}\right)_t (\text{var}\ \hat{\psi})_t \qquad (3.11)$$

where var $\hat{\psi}$ is the asymptotic variance. The process starts from some trial value t and can be repeated until no further correction is achieved to the desired degree of accuracy.

Differentiating (3.10) with respect to ψ we have

$$\frac{\partial^2 \ell}{\partial\psi^2} = \sum_{a=1}^{r-1}\sum_{b=1}^{c-1}\left[\left(\frac{n_{ab}}{p_{ab}} - \frac{n_{a,b+1}}{p_{a,b+1}} - \frac{n_{a+1,b}}{p_{a+1,b}} + \frac{n_{a+1,b+1}}{p_{a+1,b+1}}\right)\frac{\partial^2 H_{ab}}{\partial\psi^2}\right.$$
$$+ \frac{\partial}{\partial\psi} H_{ab}\left(-\frac{n_{ab}}{p^2_{ab}}\frac{\partial p_{ab}}{\partial\psi} + \frac{n_{a,b+1}}{p^2_{a,b+1}}\frac{\partial p_{a,b+1}}{\partial\psi} +\right.$$
$$\left.\left. + \frac{n_{a+1,b}}{p^2_{a+1,b}}\frac{\partial p_{a+1,b}}{\partial\psi} - \frac{n_{a+1,b+1}}{p^2_{a+1,b+1}}\frac{\partial p_{a+1,b+1}}{\partial\psi}\right)\right] \qquad (3.12)$$

The expected value of $\frac{\partial^2 \ell}{\partial\psi^2}$ is then easily obtained observing that the first term of the expression (3.12) is reduced to zero. So

$$E\left(\frac{\partial^2 \ell}{\partial\psi^2}\right) = n\sum_{a=1}^{r-1}\sum_{b=1}^{c-1}\left[H'_{ab}\left(-\frac{1}{p_{ab}}\frac{\partial p_{ab}}{\partial\psi} + \frac{1}{p_{a,b+1}}\frac{\partial p_{a,b+1}}{\partial\psi} +\right.\right.$$
$$\left.\left. + \frac{1}{p_{a+1,b}}\frac{\partial p_{a+1,b}}{\partial\psi} - \frac{1}{p_{a+1,b+1}}\frac{\partial p_{a+1,b+1}}{\partial\psi}\right)\right] \qquad (3.13)$$

The required formulae for p_{ij} and $\frac{\partial p_{ij}}{\partial\psi}$ are set out in Table 3.5, then the expressions $\frac{\partial \ell}{\partial\psi}$ and $E\left(\frac{\partial^2 \ell}{\partial\psi^2}\right)$ can be calculated and also the iterative method given by (3.11).

The numerical process for estimating ψ must start from some trial value t. A simple estimate of ψ can be constructed supposing that the joint distribution is divided into four quadrants by lines X=x and Y=y. Let A,B,C,D be the frequencies of pairs (u_i, v_i) in the RxC contingency table in the quadrants (X<x,Y<y), (X<x,Y>y), (X>x,Y<y) and (X>x,Y>y) respectively. A natural estimate of ψ is given by t = AD/BC. The frequencies A,B,C,D depend on the point of dichotomy (x,y). Mardia (1970) points out that an optimum choice for (x,y) is that which minimizes var(t) and shows that the variance is minimized with respect to X and Y if (x,y) is the population median vector, ie., when $F(x) = G(y) = \frac{1}{2}$. In the numerical process we can start from some dichotomy point as near as possible to $\frac{1}{2}$. The particular choice of the starting value is not so important in this iterative process because the process converges rapidly in the first few iterations. (We have observed empirically, that for a starting value t = 0.5, the iterative process gives the same result as for t = AD/BC with approximately the same number of iterations.)

3.4 A Series Giving The C-Type Distribution Function

In this section we shall present an alternative expression for the C-type distribution function H = H(x,y), obtained by using a binomial expansion of the original formula. The main reason for searching for an alternative expression for H is to avoid the numerical problems arising from evaluating the maximum likelihood estimate for values of ψ in the neighbourhood of 1. (ψ=1 where X and Y are independent r.v.'s).

Let $\alpha = \psi - 1$. The C-type distribution function, H, becomes

$$H = \frac{1}{2\alpha}\left\{1 + \alpha(F+G) - \left[(1+\alpha(F+G))^2 - 4\alpha(\alpha+1)FG\right]^{\frac{1}{2}}\right\} \qquad (3.14)$$

or

$$H = \frac{1}{2}(F+G) + \frac{1}{2\alpha}\{1 - [1+2\alpha(F+G-2FG) + \alpha^2(F-G)^2]^{\frac{1}{2}}\}, \quad \alpha>-1 \qquad (3.15)$$

On using the binomial expansion for the term $(1+Z)^{\frac{1}{2}}$ where $Z = 2\alpha(F+G-2FG) + \alpha^2(F-G)^2$, we have, after some reduction the following expression, where $\bar{F} = 1-F$,

$$\bar{G} = 1 - G \quad \text{and} \quad |2\alpha(F+G-2FG) + \alpha^2(F-G)^2| < 1$$

$$\begin{aligned}
H = {} & FG + \alpha FG\bar{F}\bar{G} - \alpha^2[F^2G\bar{F}\bar{G}^2 + FG^2\bar{F}^2\bar{G}] \\
& + \alpha^3[F^3G\bar{F}\bar{G}^3 + 3F^2G^2\bar{F}^2\bar{G}^2 + FG^3\bar{F}^3\bar{G}] \\
& - \alpha^4[F^4G\bar{F}\bar{G}^4 + 6F^3G^2\bar{F}^2\bar{G}^3 + 6F^2G^3\bar{F}^3\bar{G}^2 + FG^4\bar{F}^4\bar{G}] \\
& + \alpha^5[F^5G\bar{F}\bar{G}^5 + 10F^4G^2\bar{F}^2\bar{G}^4 + 20F^3G^3\bar{F}^3\bar{G}^3 + 10F^2G^4\bar{F}^4\bar{G}^2 + FG^5\bar{F}^5\bar{G}] \\
& - \alpha^6[F^6G\bar{F}\bar{G}^6 + 15F^5G^2\bar{F}^2\bar{G}^5 + 50F^4G^3\bar{F}^3\bar{G}^4 + 50F^3G^4\bar{F}^4\bar{G}^3 + \\
& \qquad + 15F^2G^5\bar{F}^5\bar{G}^2 + FG^6\bar{F}^6\bar{G}] \\
& + \alpha^7[F^7G\bar{F}\bar{G}^7 + 21F^6G^2\bar{F}^2\bar{G}^6 + 105F^5G^3\bar{F}^3\bar{G}^5 + \\
& \qquad + 175F^4G^4\bar{F}^4\bar{G}^4 + 105F^3G^5\bar{F}^5\bar{G}^3 + 21F^2G^6\bar{F}^6\bar{G}^2 + \\
& \qquad + FG^7\bar{F}^7\bar{G}] \\
& + \ldots
\end{aligned} \qquad (3.16)$$

In order to simplify notation let

$$U = FG\bar{F}\bar{G} \quad \text{and} \quad V = F\bar{G} + \bar{F}G = F+G-2FG.$$

After expanding the above series until the term of the order α^{13} and after tedious algebra we have

$$
\begin{aligned}
H - FG \;=\; & \\
& + \alpha(UV^{o}) \\
& - \alpha^{2}(UV^{1}) \\
& + \alpha^{3}(UV^{2} + U^{2}V^{o}) \\
& - \alpha^{4}(UV^{3} + 3U^{2}V^{1}) \qquad\qquad (3.17) \\
& + \alpha^{5}(UV^{4} + 6U^{2}V^{2} + 2U^{3}V^{o}) \\
& - \alpha^{6}(UV^{5} + 10U^{2}V^{3} + 10U^{3}V^{1}) \\
& + \alpha^{7}(UV^{6} + 15U^{2}V^{4} + 30U^{3}V^{2} + 5U^{4}V^{o}) \\
& - \alpha^{8}(UV^{7} + 21U^{2}V^{5} + 70U^{3}V^{3} + 35U^{4}V^{1}) \\
& + \alpha^{9}(UV^{8} + 28U^{2}V^{6} + 140U^{3}V^{4} + 140U^{4}V^{2} + 14U^{5}V^{o}) \\
& - \alpha^{10}(UV^{9} + 36U^{2}V^{7} + 252U^{3}V^{5} + 420U^{4}V^{3} + 126U^{5}V^{1}) \\
& + \alpha^{11}(UV^{10} + 45U^{2}V^{3} + 420U^{3}V^{6} + 1050U^{4}V^{4} + 630U^{5}V^{2} + 42U^{6}V^{o}) \\
& - \alpha^{12}(UV^{11} + 55U^{2}V^{9} + 660U^{3}V^{7} + 2310U^{4}V^{5} + 2310U^{5}V^{3} + 462U^{6}V^{1}) \\
& + \alpha^{13}(UV^{12} + 66U^{2}V^{10} + 990U^{3}V^{8} + 4620U^{4}V^{6} + 6930U^{5}V^{4} + 2772U^{6}V^{2} \\
& \qquad + 132U^{7}V^{o}) - \ldots.
\end{aligned}
$$

In addition we know that

$$
\begin{aligned}
(1+\alpha V)^{-1} &= 1 - \alpha V + \alpha^{2}V^{2} - \alpha^{3}V^{3} + \alpha V^{4} \ldots \qquad , \quad -1<\alpha V<1 \\
(1+\alpha V)^{-3} &= 1 - 3\alpha V + 6\alpha^{2}V^{2} - 10\alpha^{3}V^{3} + 15\alpha^{4}V^{4} - \ldots, \quad -1<\alpha V<1 \\
(1+\alpha V)^{-5} &= 1 - 5\alpha V + 15\alpha^{2}V^{2} - 35\alpha^{3}V^{3} + \ldots \qquad -1<\alpha V<1
\end{aligned}
\qquad (3.18)
$$

Combining (3.17) and (3.18) we obtain

$$
\begin{aligned}
H = FG + \frac{\alpha U}{1+\alpha V} + \frac{\alpha^{3}U^{2}}{(1+\alpha V)^{3}} + \frac{2\alpha^{5}U^{3}}{(1+\alpha V)^{5}} + \frac{5\alpha^{7}U^{4}}{(1+\alpha V)^{7}} + \frac{14\alpha^{9}U^{5}}{(1+\alpha V)^{9}} + \\
+ \frac{42\alpha^{11}U^{6}}{(1+\alpha V)^{11}} + \frac{132\alpha^{13}U^{7}}{(1+\alpha V)^{13}} + \ldots
\end{aligned}
\qquad (3.19)
$$

This is the series giving H, where $U = FG\overline{FG}$ and $V = F\overline{G} + \overline{F}G$. We shall prove now that the series (3.19) converges to the expression (3.14).

Let $T = \dfrac{4\alpha^2U}{(1+\alpha V)^2}$, on using the binomial expansion of $(1-T)^{\frac{1}{2}}$ we have

$$\left[1- \frac{4\alpha^2U}{(1+\alpha V)^2}\right]^{\frac{1}{2}} = 1 - \frac{2\alpha^2U}{(1+\alpha V)^2} - \frac{2\alpha^4U^2}{(1+\alpha V)^4} - \frac{2.2\alpha^6U^3}{(1+\alpha V)^6} - \frac{2.5\alpha^8U^4}{(1+\alpha V)^8} - \frac{2.14\alpha^{10}U^5}{(1+\alpha V)^{10}} - \ldots$$

$$= 1 - \frac{2\alpha^2U}{(1+\alpha V)^2}\left[1+ \frac{\alpha^2U}{(1+\alpha V)^2} + \frac{2\alpha^4U^2}{(1+\alpha V)^4} + \frac{5\alpha^6U^3}{(1+\alpha V)^6} +\ldots\right]$$

Then

$$1 - \left[1- \frac{4\alpha^2U}{(1+\alpha V)^2}\right]^{\frac{1}{2}} = \frac{2\alpha^2U}{(1+\alpha V)^2}\left[1 + \frac{\alpha^2U}{(1+\alpha V)^2} + \frac{2\alpha^4U^2}{(1+\alpha V)^4} + \frac{5\alpha^6U^3}{(1+\alpha V)^6} +\ldots\right]$$

and we have

$$1 + \frac{\alpha^2U}{(1+\alpha V)^2} + \frac{2\alpha^4U^2}{(1+\alpha V)^4} + \frac{5\alpha^6U^3}{(1+\alpha V)^6} + \ldots = \frac{(1+\alpha V)^2}{2\alpha^2U}\{1 - [1 - \frac{4\alpha^2U}{(1+\alpha V)^2} \tag{3.20}$$

On the other hand, expression (3.19) may be written as

$$H = FG + \frac{\alpha U}{1+\alpha V}\left[1 + \frac{\alpha^2U}{(1+\alpha V)^2} + \frac{2\alpha^4U^2}{(1+\alpha V)^4} + \frac{5\alpha^6U^3}{(1+\alpha V)^6} + \frac{14\alpha^8U^4}{(1+\alpha V)^8} + \ldots\right]$$

or, on using (3.20)

$$H = FG + \frac{\alpha U}{1+\alpha V} \cdot \frac{(1+\alpha V)^2}{2\alpha^2U}\left\{1 - \left[1 - \frac{4\alpha^2U}{(1+\alpha V)^2}\right]^{\frac{1}{2}}\right\}$$

$$= FG + \frac{1+\alpha V}{2\alpha}\left\{1 - \left[1 - \frac{4\alpha^2U}{(1+\alpha V)^2}\right]^{\frac{1}{2}}\right\}$$

$$= FG + \left\{\frac{1+\alpha V}{2\alpha} - \frac{1}{2\alpha}\left[(1+\alpha V)^2 - 4\alpha^2U\right]^{\frac{1}{2}}\right\}$$

and finally, returning to the F and G notation, remembering that $U = FG\overline{FG}$ and

$V = F\bar{G} + \bar{F}G = F + G - 2FG$, we have

$$H = FG + \frac{1}{2\alpha}\left\{1 + \alpha(F+G-2FG) - \left[(1 + \alpha(F+G-2FG))^2 - 4\alpha^2 FG\bar{F}\bar{G}\right]^{\frac{1}{2}}\right\}$$

$$= \frac{1}{2\alpha}\left\{1 + \alpha(F+G) - \left[(1 + \alpha(F+G))^2 - 4\alpha(\alpha+1)FG\right]^{\frac{1}{2}}\right\}$$

which is the expression (3.14) given above.

Therefore, the series giving the C-type distribution function H, is given by expression (3.19), which in the F and G notation is

$$\begin{aligned} H = FG &+ \frac{\alpha FG\bar{F}\bar{G}}{1+\alpha(F\bar{G}+\bar{F}G)} + \frac{\alpha^3(FG\bar{F}\bar{G})^2}{[1+\alpha(F\bar{G}+\bar{F}G)]^3} + \frac{2\alpha^5(FG\bar{F}\bar{G})^3}{[1+\alpha(F\bar{G}+\bar{F}G)]^5} \\ &+ \frac{5\alpha^7(FG\bar{F}\bar{G})^4}{[1+\alpha(F\bar{G}+\bar{F}G)]^7} + \frac{14\alpha^9(FG\bar{F}\bar{G})^5}{[1+\alpha(F\bar{G}+\bar{F}G)]^9} + \frac{42\alpha^{11}(FG\bar{F}\bar{G})^6}{[1+\alpha(F\bar{G}+\bar{F}G)]^{11}} + \\ &+ \frac{132\alpha^{13}(FG\bar{F}\bar{G})^7}{[1+\alpha(F\bar{G}+\bar{F}G)]^{13}} + \dots \qquad |\alpha| < \frac{1}{(F\bar{G}+\bar{F}G)} \qquad (3.21) \end{aligned}$$

The expression (3.21) may be very useful in situations where we want to avoid numerical problems, as for example, using the C-type distribution function for $\alpha \equiv 0$.

Because of the numerical advantages of the expression (3.21) we have decided to include this alternative expression in the computer program for the maximum likelihood method described in section 3.5 when $0.98 \leq \psi \leq 1.02$.

Before comparing the numerical values for H given by expressions (3.14) and (3.21), we present the derivative of H with respect to α, which is also necessary in the maximum likelihood estimation method for values of ψ near 1. From (3.21) it is easily seen that

$$H' = \frac{\partial H}{\partial \alpha} = \frac{FG\bar{F}\bar{G}}{[1+\alpha(F\bar{G}+\bar{F}G)]^2} + \frac{3\alpha^2(FG\bar{F}\bar{G})^2}{[1+\alpha(F\bar{G}+\bar{F}G)]^4} + \frac{10\alpha^4(FG\bar{F}\bar{G})^3}{[1+\alpha(F\bar{G}+\bar{F}G)]^6} + \dots \qquad (3.22)$$

In Table 3.6 we present some numerical values for the C-type distribution function for some values of the parameter of association $\psi = \alpha+1$ and for some values of F and G. We observe that the approximation given by the expression (3.21) is very good indeed, even for large values

TABLE 3.6 - COMPARISON OF NUMERICAL VALUES OF THE C-TYPE DISTRIBUTION FUNCTION WITH THE APPROXIMATIVE FORMULAS GIVEN BY THE EXPRESSIONS (3.21) AND (3.16) FOR SOME VALUES OF $\psi = \alpha+1$ AND FOR SOME VALUES OF F AND G.

F	G	ψ	C-TYPE D.F. H	H GIVEN BY (3.21)	H GIVEN BY (3.16)
0.5	0.5	0.1000000	0.120127	0.120502	0.131762
0.5	0.5	0.5000000	0.207107	0.207107	0.207144
0.5	0.5	0.9800000	0.248738	0.248737	0.248737
0.5	0.5	0.9999999	0.264706	0.250000	0.250000
0.5	0.5	1.0010000	0.250089	0.250063	0.250063
0.5	0.5	1.0200000	0.251237	0.251238	0.251238
0.5	0.5	2	0.292893	0.292893	0.295853
0.5	0.5	10	0.379874	0.379498	*
0.5	0.5	100	0.454545	0.437513	*
0.5	0.5	200	0.466980	0.442429	*
0.1	0.9	0.1000000	0.058840	0.058842	0.064770
0.1	0.9	0.5000000	0.083095	0.083095	0.083113
0.1	0.9	0.9800000	0.089835	0.089835	0.089835
0.1	0.9	0.9999999	0.088235	0.090000	0.090000
0.1	0.9	1.0010000	0.090000	0.090008	0.090008
0.1	0.9	1.0200000	0.090161	0.090159	0.090159
0.1	0.9	2	0.094461	0.094461	0.095819
0.1	0.9	10	0.098782	0.098782	*
0.1	0.9	100	0.099875	0.099875	*
0.1	0.9	200	0.099937	0.099937	*
0.9	0.9	0.100000	0.801218	0.801218	0.801218
0.9	0.9	0.500000	0.805538	0.805538	0.805538
0.9	0.9	0.980000	0.809837	0.809837	0.809837
0.9	0.9	0.999999	0.823529	0.810000	0.810000
0.9	0.9	1.001000	0.810049	0.810008	0.810000
0.9	0.9	1.020000	0.810162	0.810161	0.810161
0.9	0.9	2	0.816905	0.816905	0.816905
0.9	0.9	10	0.841160	0.841158	*
0.9	0.9	100	0.874479	0.872070	*
0.9	0.9	200	0.881099	0.876235	*

* values greater than one.

of ψ (or α). The approximation given by (3.16) is valid only for small values of α. For values of ψ very near 1, the approximation (3.21) should be used instead of the expression (3.14) (see for example the case $\psi = 0.999999$ in the table). Table 3.6 is presented here for illustrative purposes only.

In Fig. 3.1 we present the derivative of the log-likelihood function $\left(\frac{\partial L}{\partial \psi} \right)$ in the neighbourhood of $\psi = 1$, before and after using the approximation given by expression (3.22) obtained from series expansion of H. In this example, the table analysed is given by the 3x3 table:

529	107	653
49	37	60
131	29	172

the row marginal proportions are 0.7295; 0.0826 and 0.1879 and the column marginal proportions are 0.4012; 0.0979 and 0.5008. The derivative of the ML function is plotted for values of ψ in the interval $0.98 \leqslant \psi \leqslant 1.02$. It is clear from Fig. 3.1 that the series giving the C-type distribution function is very useful in situations where the ML estimate of ψ is in or near this interval. Before using the approximation nonconvergence of the iteration process was observed in this example. The figure shows the numerical problems near the point of singularity of the function ($\alpha = 0$ or $\psi = 1$).

On using the maximum likelihood method designed to estimate ψ for a set of empirical data where several variables are practically independent ($\psi \equiv 1$) we observed a high proportion of cases of either a large number of iterations or no convergence. After including the approximation for H given by expression (3.19) the problem was solved, leading to an improvement of the method. The test was performed for several contingency tables. In

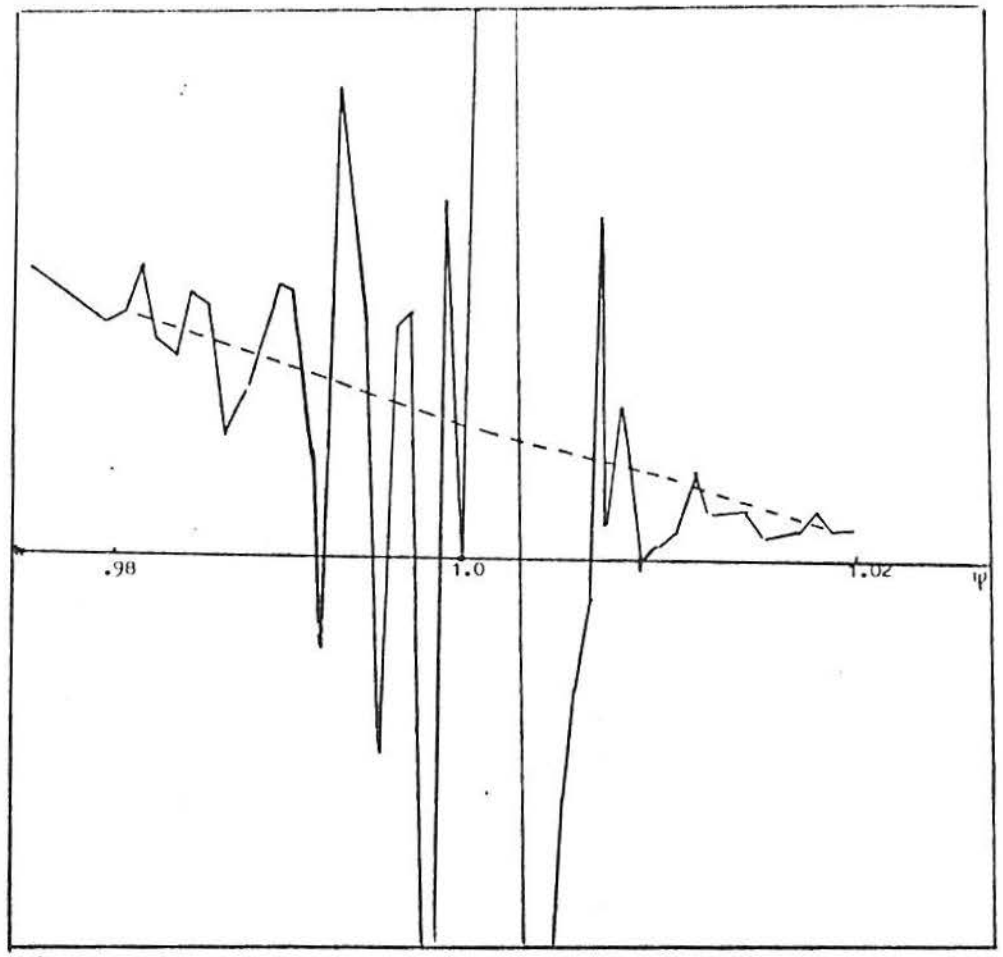


FIGURE 3.1 - DERIVATIVE OF THE MAXIMUM LIKELIHOOD FUNCTION IN THE NEIGHBOURHOOD OF $\psi=1$

____ BEFORE USING THE APPROXIMATION

_ _ _ AFTER USING APPROXIMATION

NOTE: THE APPROXIMATION WAS USED IN THE INTERVAL $0.98 \leq \psi \leq 1.02$

table 3.7 we present some of the results showing the estimate of ψ, the derivative of the loglikelihood function and the number of iterations of the iterative process for the maximum likelihood method before and after using the series given by (3.19) in the interval $0.98 < \psi < 1.02$. We point out that no numerical problems were observed with the algorithm for the method outside this interval.

Table 3.7 - Parameter estimate, derivative of the loglikelihood function, number of iterations for a set of zero correlated variables(*) before and after the inclusion of the series giving H

BEFORE			AFTER		
$\hat{\psi}$	$\frac{\partial L}{\partial \psi}$	N of Iterations	$\hat{\psi}$	$\frac{\partial L}{\partial \psi}$	N of iterations
1.00008	100281	400	0.990	-0.0028	2
0.99997	-439308	3	0.997	0.0012	3
0.99922	0	116	1.0037	0.0028	5
1.00008	306553	178	0.9874	0.0050	4
0.98838	0.0118	67	0.9919	0.0051	3
1.000347	0	140	0.9960	0.0004	2
0.99128	-0.0160	8	0.9940	0.0062	2
1.02211	-0.0098	63	1.0175	-0.0057	2
0.9737	-0.0102	18	0.9802	0.0010	7

(*) The data used in this table are from the Data Set No.4 (Greek Data) to be described in chapter 6.

Finally, we might consider that in some cases, the alternative expressions for H, may be useful for evaluating the correlation coefficient for given margins. As an example we have used the series given by (3.16) for evaluating an approximate expression for the correlation coefficient of the C-type logistic distribution.

On using the expression given by Mardia (1967):

$$\mathrm{corr}(x,y) \;=\; \frac{1}{\sigma_1}\,\frac{1}{\sigma_2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(H-FG)\,dx\,dy$$

and putting $F = 1/(1+e^{-x})$; $G = 1/(1+e^{-y})$ and H given by (3.16) we have after integration the following expression for the correlation coefficient of the C-type logistic distribution

$$\rho_L(\alpha) \;=\; \frac{3\alpha}{\pi^2}\Big[1 - \frac{\alpha}{2} + \frac{11\alpha^2}{36} - \frac{5}{24}\,\alpha^3 + \frac{137}{900}\,\alpha^4 - \frac{7}{60}\,\alpha^5 + \frac{363}{3920}\,\alpha^6 - \frac{761}{10080}\,\alpha^7 + \ldots\Big| = \qquad |\alpha| < 1 \qquad (3.23)$$

Formula (3.23) gives a very good approximation for small values of α, even if we use a small number of terms. However, for values of $\alpha > 1$, the table 2.1 presented in chapter 2 should be used.

3.5 Contingency-type correlation coefficients for polytomous data

In the preceeding sections we have presented the maximum likelihood method for estimating the parameter ψ of the C-type distribution for data given in an RxC contingency table. The parameter of association ψ is the constant cross-product ratio for the (r-1)(c-1) fourfold tables in the RxC table and may be defined as the global cross-product ratio when the underlying model is the C-type distribution. Therefore, the measures of correlation as function of the cross-product ratio presented in section 2.3 for 2x2 contingency tables may be used in the same way as estimators of the latent correlation coefficient for RxC contingency tables. We then have contingency-type correlation coefficients for polytomous data.

Altham (1970) points out that it would be convenient to have a single expression for the measure of association of the rows and columns of a table, rather than the whole set of $(r-1)(s-1)$ cross-ratios, and she proposes the definition of a metric on equivalence classes as a possible way of finding a single coefficient of association.

Assuming the C-type distribution as an underlying model for two-way contingency tables, we have a single measure of association of the rows and columns of the table, which is given by the parameter of association ψ. Contingency-type correlation coefficients are then defined according to the formulae chosen to transform the association measure into a correlation coefficient. For uniform margins we shall use Mardia's coefficient given by

$$\rho_U(\psi) = \frac{\psi+1}{\psi-1} - \frac{2\psi \ \ell n \ \psi}{(\psi-1)^2} \tag{3.24}$$

presented in section 2.3 (see also Mardia, 1967).

Chambers' coefficient given by

$$r_\nu = (\psi^\nu - 1)/(\psi^\nu + 1) \tag{3.25}$$

can also be used for estimating the latent correlation coefficient. As we have seen in chapter 2 a number of other measures of correlation can be conveniently approximated by (3.25) for some values of ν. For example, for $\nu = 2/3$ we obtain an approximation to Mardia's coefficient $\rho_U(\psi)$; for $\nu = 0.64$ we have $r_{0.64} \equiv \rho_N(\psi)$ where $\rho_N(\psi)$ is the correlation coefficient for the C-type normal distribution. We have shown in section 2.4 that for

$\nu = 0.61$, $r_{0.61} \equiv \rho_L(\psi)$ where $\rho_L(\psi)$ is the correlation coefficient for the C-type logistic distribution.

Plackett (1965) compares the bivariate normal distribution with the C-type normal and, as we have seen in section 2.6, he shows that the two distributions agree well relating ψ and ρ by

$$r_p(\psi) \quad = \quad - \cos[\pi\psi^{\frac{1}{2}}/(1 + \psi^{\frac{1}{2}})] \qquad (3.26)$$

In Plackett's formula, ψ is the cross product ratio for 2×2 tables. On taking the MLE of ψ, $\hat{\psi}$, Plackett's coefficient can also be regarded as a contingency type correlation coefficient for R×C tables. Nevertheless, we shall show in this section that the Chambers coefficient given by

$$r_{0.74} \quad = \quad (\hat{\psi}^{0.74} - 1)/(\hat{\psi}^{0.74} + 1) \qquad (3.27)$$

where $\hat{\psi}$ is the MLE of ψ for RxC contingency tables, is a better estimate of the correlation coefficient of an underlying bivariate normal distribution, for data given in RxC tables.

Asymptotic variance of the correlation coefficient estimates

One of the advantages of the maximum likelihood method of estimation presented in this chapter is that we obtain the asymptotic variance of $\hat{\psi}$, which is given by the expression

$$\text{var}(\hat{\psi}) \quad = \quad - \frac{1}{\left[E\left(\dfrac{\partial^2 \ell nL}{\partial\psi^2}\right)\right]_{\hat{\psi}}}$$

where $E\left(\dfrac{\partial^2 \ell nL}{\partial\psi^2}\right)$ is given by (3.13).

For a general function $f(\psi)$, the asymptotic variance of $f(\hat{\psi})$ is given by

$$\mathrm{var}[f(\hat{\psi})] \doteq [f'(\hat{\psi})]^2 \mathrm{var}(\hat{\psi})$$

where $f'(\hat{\psi})$ is the derivative of f with respect to ψ. (Kendall and Stuart, 1979, vol.II, pg. 53).

Therefore we can easily derive the expressions for the asymptotic variance for each of the estimators presented in the beginning of this section. Thus, on using the notation r for the estimate, we have

$$\mathrm{var}[r_U(\hat{\psi})] = [\frac{2(\hat{\psi}^2-1)\ell n\hat{\psi} - 4(\hat{\psi}-1)^2}{(\psi-1)^4}]^2 \mathrm{var}(\hat{\psi})$$

which is the asymptotic variance of the contingency-type correlation coefficient using Mardia's formula [expression (3.24)].

The asymptotic variance of the contingency-type coefficient using Chambers' formula [expression (3.25)] is given by

$$\mathrm{var}[r_\nu\ (\hat{\psi})] = [\frac{2\nu\hat{\psi}^{\nu\ -1}}{(\psi^\nu+1)^2}]^2 \ .\ \mathrm{var}(\hat{\psi})$$

The asymptotic variance of the contingency-type coefficient estimate using Plackett's formula [expression (3.26)] is

$$\mathrm{var}[r_p(\hat{\psi})] = [\pi\ \sin(\frac{\pi\hat{\psi}^{\frac{1}{2}}}{1+\psi^{\frac{1}{2}}})/2\hat{\psi}^{\frac{1}{2}}(1 + \hat{\psi}^{\frac{1}{2}})^2]^2 \mathrm{var}(\hat{\psi})\ .$$

3.6 Numerical examples

In this section we present some numerical examples of the contingency-type correlation coefficients for polytomous data and we compare the estimates of the correlation coefficients as function of the maximum likelihood estimate of ψ with other methods available in the literature.

Pearson and Heron (1913) in their study about theories of association used Pearson's Family Data and arranged 1000 cases according to the magnitude of the stature of father and son in a 7×7 contingency table. We present the data in Table 3.8, part (a), (see Pearson and Heron, 1913, Table XV, p.220).

Pearson and Heron (1913) also present two tables obtained by dividing up a bivariate normal distribution with correlation parameters $\rho = 0.5$ and $\rho = 0.3$ into the same group as the Eye-Colour grouping used for Pearson's Family Data. We present these artificial data in part (a) of table 3.9 ($\rho = 0.5$) and table 3.10 ($\rho = 0.3$). We notice that the tables were modified to give whole numbers in the cells.

We have thus three tables, one with empirical data where the assumption of an underlying bivariate distribution is plausible and two, that for practical purposes, are artificial data from normal surfaces. We shall apply our method to these three tables and present the various estimates of the correlation coefficients. The results can be compared with the product moment correlation coefficient as calculated by Pearson and Heron (1913) from the original data in the case of the empirical data or with the parameters in the case of the theoretical tables. The expected frequencies under the C-type distribution model are also presented for illustrative purposes, in part (b) of the tables above cited. We shall present the tables followed by the analysis, where we show the MLE of ψ and, where relevant, the standard error and the chi-square statistic of goodness of fit (part (c) of tables 3.8, 3.9 and 3.10). The estimates of the contingency type correlation coefficients are also presented, with the standard error for the case of empirical data (part (d)).

The chi-square (χ^2) statistic used in the examples is the usual statistic

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij}-e_{ij})^2}{e_{ij}}$$

with the degrees of freedom given by

$$DF = (r-1)(c-1) - 1$$

(one degree of freedom is lost because we are estimating the parameter ψ).

The numerical examples confirm that the coefficient $r_{0.74}$ defined by (3.26) is a very good estimate of the correlation coefficient of the underlying bivariate normal distribution estimated from an RxC contingency table. The coefficient $r_u(\psi)$ would be more appropriate when a C-type uniform distribution underlies the data. The coefficient $r_p(\psi)$ overestimates the correlation coefficient and we have always $|r_p(\psi)| > |r_{0.74}(\psi)|$. On the other hand, the coefficient $r_{0.74}(\psi)$ is a simpler alternative for the polychoric correlation coefficient.

Table 3.8 - (a) Empirical Data: Stature of Father and Son in Eye-Colour Groups for Pearson's Family Data (Pearson and Heron, 1913)

SON'S STATURE CLASS	FATHER'S STATURE CLASS							
	1	2	3	4	5+6	7	8	Total
1	4	22	7	-	1	-	-	34
2	23	154	84	26	8	6	-	301
3	8	87	75	66	22	24	2	284
4	1	29	36	37	14	14	6	137
5+6	-	18	27	26	11	18	5	105
7	-	9	26	19	7	29	8	98
8	-	3	9	6	6	10	7	41
Total	36	322	264	180	69	101	28	1000

(b) Expected frequencies for table (a) under the C-type distribution using Maximum likelihood method

SON'S STATURE CLASS	FATHER'S STATURE CLASS							Total
	1	2	3	4	5+6	7	8	
1	4.7	19.7	5.7	2.2	0.7	0.8	0.2	34
2	20.3	156.0	73.7	29.6	8.5	10.4	2.5	301
3	6.7	89.4	94.5	53.3	16.1	19.4	4.6	284
4	1.9	26.3	40.2	35.0	12.9	16.7	4.0	137
5+6	1.1	15.2	24.9	27.7	12.6	18.6	4.9	105
7	0.9	11.3	18.5	23.5	12.9	23.5	7.4	98
8	0.4	4.1	6.5	8.7	5.3	11.6	4.4	41
Total	36	322	264	180	69	101	28	1000

(c) MLE of the global cross-product ratio for table (a) and Chi-square goodness of fit statistic

MLE of ψ	Chi-square statistic	Significance
4.833 ± 0.505	$\chi^2 = 49.80$ (DF=35)	$p = 0.05$

(d) Estimate of correlation coefficients and standard errors for table (a), using the value of ψ given in (c)

Contingency type coefficient $r_U(\hat{\psi})$	0.485 ± 0.025
Contingency type coefficient $r_P(\hat{\psi})$	0.555 ± 0.027
Contingency type coefficient $r_{0.74}(\hat{\psi})$	0.524 ± 0.026
Product moment correlation	0.523(*)

(*) From Pearson and Heron (1913).

Table 3.9 - (a) Artificial Data I: Bivariate Normal Distribution for $\rho=0.5$ adjusted to give whole units in cells (from Pearson and Heron, 1913, p.220)

	1	2	3	4	5+6	7	8	TOTAL
1	7	20	5	2	-	-	-	34
2	21	145	79	36	10	9	1	301
3	6	94	85	54	19	22	4	284
4	2	32	39	31	12	17	4	137
5+6	-	18	28	25	11	18	5	105
7	-	11	22	24	12	22	7	98
8	-	2	6	8	5	13	7	41
Total	36	322	264	180	69	101	28	1000

(b) Expected Frequencies for table (a) under the C-type distribution model

	1	2	3	4	5+6	7	8	TOTAL
1	4.5	19.4	5.9	2.4	0.7	0.9	0.2	34
2	20.0	153.3	74.4	30.7	8.9	11.0	2.7	301
3	7.0	89.9	92.7	53.2	16.4	20.0	4.8	284
4	2.0	27.1	39.9	34.3	12.8	16.8	4.1	137
5+6	1.2	15.9	25.2	27.2	12.3	18.3	4.9	105
7	0.9	12.0	19.1	23.5	12.6	22.8	7.1	98
8	0.4	4.4	6.8	8.7	5.3	11.2	4.2	41
Total	36	322	264	180	69	101	28	1000

(c) Global cross-product ratio for table (a) using maximum likelihood method

$\psi = 4.486$

(d) MLE of the correlation coefficient for table (a) supposing a underlying C-type diistribution

Contingency type Coefficient $r_U(\psi)$	0.465
Contingency type Coefficient $r^U(\psi)$	0.534
Contingency type Coefficient $r^U_{0.74}(\psi)$	0.504
Correlation coefficient parameter	0.500

Table 3.10 - (a) Artificial Data II: Bivariate Normal Distribution with $\rho=0.3$ adjusted to give whole units in cells (from Pearson and Heron, 1913)

	1	2	3	4	5+6	7	8	TOTALS
1	4	17	8	3	1	1	0	34
2	17	123	80	45	15	18	3	301
3	9	93	78	52	19	26	7	284
4	3	38	37	27	11	16	4	137
5+6	2	25	28	22	9	15	4	105
7	1	20	24	21	10	16	6	98
8	0	6	9	9	4	9	4	41
TOTALS	36	322	264	180	69	101	28	1000

(b) Expected Frequencies under the C-type distribution for table (a)

	1	2	3	4	5+6	7	8	TOTALS
1	2.6	16.4	7.9	3.8	1.2	1.6	0.4	34
2	16.6	128.9	78.3	41.1	13.6	17.9	4.6	301
3	9.2	92.0	80.8	51.8	18.4	25.2	6.6	284
4	3.1	34.9	37.5	29.0	11.4	16.6	4.5	137
5+6	2.1	23.2	27.0	23.3	9.9	15.3	4.4	105
7	1.7	19.2	23.3	21.9	10.1	16.7	5.0	98
8	0.7	7.4	9.2	9.1	4.4	7.8	2.5	41
TOTALS	36	322	264	180	69	101	28	1000

(c) Global cross-product ratio (ψ) for table (a)

$$\psi = 2.353$$

(d) MLE of the correlation coefficient for table (a) supposing an underlying C-type distribution

Contingency type Coefficient $r_U(\psi)$	0.278
Contingency type Coefficient $r^U(\psi)$	0.325
Contingency type Coefficient $r^P_{0.74}(\psi)$	0.306
Correlation Coefficient parameter	0.300

From the analysis of tables 3.8, 3.9 and 3.10 we observe that the data sets fit the C-type distribution well, and we notice that the fit for the artificial data from the bivariate normal model is extremely good, showing that the normal surface can be well approximated by a C-type normal distribution.

The example which follows compares the contingency-type correlation $r_{0.74}(\hat{\psi})$ with other polychoric correlation coefficients proposed in the literature. We take the data analysed by Lancaster and Hamdam (1964) and by Goodman (1981). We shall compare our method for RxC tables with Lancaster-Hamdam's coefficient obtained by polychoric series method and with two Goodman's coefficients: ρ_S for the local uniform association model and ρ_S' for the Goodman's model I.

The stature of fathers and daughters data were originally analysed by Pearson and as in our previous examples the assumption of bivariate normal distribution is plausible; it was actually tested by Lancaster (see Lancaster and Hamdam, 1964). We shall consider here only the 3×3 table based on a natural grouping of rows and columns (neighbouring classes being pooled) from the original 18×18 contingency table. In table 3.11(a) we present the observed frequencies for the 3×3 table, in part (b) the expected frequencies under the C-type distribution model, obtained by the maximum likelihood method are presented. Part (c) of the table shows the maximum likelihood estimate of the parameter ψ with standard error and the Chi-square statistic of goodness-of-fit. Finally part (d) shows the estimates of the correlation coefficient obtained by different methods.

TABLE 3.11 - (a) Stature of Fathers and Daughters - 3×3 TABLE
(From Lancaster and Hamdam, 1964)

DAUGHTER'S HEIGHT CLASS	FATHER'S HEIGHT CLASS 1	 2	 3	 TOTALS
1	70	99	15	184
2	128	432	183	743
3	20	177	252	449
TOTALS	218	708	450	1376

(b) Expected Frequencies under C-type distribution maximum likelihood method for Table (a)

DAUGHTER'S HEIGHT CLASS	FATHER'S HEIGHT CLASS 1	 2	 3	 TOTALS
1	72.7	91.6	19.7	184
2	121.4	443.4	178.2	743
3	23.9	173.0	252.0	449
TOTALS	218	708	450	1376

(c) MLE of ψ and Chi-square statistic for table (a)

MLE of ψ	Chi-square statistic
4.713 ± 0.474	$\chi^2 = 3.333$ (DF=3)

(d) Estimates of correlation coefficient for table (a)

Contingency type coefficient $r_U(\hat{\psi})$	0.479 ± 0.026
Contingency type coefficient $^{U}r(\hat{\psi})$	0.548 ± 0.028
Contingency type coefficient $r^{D}_{U\,.74}(\hat{\psi})$	0.518 ± 0.027
Pearson contingency method	0.381 (*)
Lancaster and Hamdan method	0.497 (*)
Goodman's coefficient (unif.assoc.)	0.494 (*)
Goodman's coefficient (MODEL I)	0.492 (*)
Product Moment Correlation Coefficient	0.517 (*)

(*) Taken from Goodman (1981).

The analysis of the table 3.11 confirms our previous conclusion, the contingency-type coefficient $r_{0.74}(\psi)$ is an extremely good approximation for the product moment correlation coefficient. Considering that the product moment correlation is calculated from the original data and thus, the total of information from the sample is not the same as in the table 3×3, the advantages of the coefficient $r_{0.74}$ seem to be considerable. Again, in this example the fit to the C-type model is good. Comparing the estimate $r_{0.74}(\hat{\psi})$ with the other methods, in this example, we can see that the method of maximum likelihood produces better results.

The next two examples have the purpose of comparing our method with the maximum likelihood estimator of the correlation coefficient (the polychoric coefficient), proposed by Tallis (1962) and generalized by Olsson (1979), supposing a normal bivariate surface as the parent distribution. The first of these two examples consists of the data, presented by Tallis and analysed also by Plackett (1965) where he calculates Plackett's coefficient as function of the cross-product ratio of a 2×2 table. In table 3.12 we present the results.

TABLE 3.12 - (a) Empirical data taken from Tallis (1962)

		1952		
1953	No lambs	1 lamb	2 lambs	Total
No lambs	58	52	1	111
1 lamb	26	58	3	87
2 lambs	8	12	9	29
TOTAL	92	122	13	227

(b) Expected Frequencies under the C-type distribution model

	1952			
1953	No lambs	1 lamb	2 lambs	Total
No lambs	61.6	46.5	2.9	111
1 lamb	25.1	55.9	6.0	87
2 lambs	5.3	19.6	4.1	29
TOTAL	92	122	13	227

(c) MLE of ψ

$$\psi = 3.521 \pm 0.907$$

(d) Estimates of Correlation coefficient

Tallis' MLE supposing Normal Bivariate distribution		0.420 ± 0.076(*)
Plackett method (2×2 table)		0.365 ± 0.096(*)
MLE method supposing C-type distribution:		
Contingency-type coefficient	$r_U(\hat{\psi})$	0.398 ± 0.073
Contingency-type coefficient	$r_p(\hat{\psi})$	0.460 ± 0.081
Contingency-type coefficient	$r_{0.74}(\hat{\psi})$	0.434 ± 0.077

(*) From Plackett (1965).

From table 3.12 we can conclude that the contingency-type coefficient $r_p(\psi)$ as a function of the MLE of ψ is better than Plackett's method of estimating ψ by the cross product ratio of the 2×2 table. Secondly, the estimates obtained by Tallis' method and by the contingency-type coefficient $r_{0.74}(\psi)$ are similar, the difference being 0.014.

Olsson (1979) analyses artificial data using a multinomial routine for generating samples of size 500 for various sets of parameters. One of these generated cross-tables is presented in Olsson (1979, pg.456), where the sample estimates using Olsson's maximum likelihood method are also shown. We analyse the table using contingency-type correlation coefficients, computed as functions of MLE of ψ. The data and comparisons are presented in table 3.13.

TABLE 3.13 - (a) Artificial data taken from Olsson (1979))

x/y	1	2	3	Total
1	13	6	0	19
2	69	113	22	204
3	41	132	104	277
Total	123	251	126	500

(b) Expected frequencies under the C-type distribution model

x/y	1	2	3	Total
1	10.83	6.76	1.41	19
2	76.52	103.20	24.28	204
3	35.65	141.04	100.31	277
Total	1123	251	126	500

(c) MLE of ψ

$$\psi = 4.361 \pm 0.783$$

(d) Estimates of correlation coefficient

Olsson's polychoric coefficient	0.49 ± 0.048 (*)
MLE method supposing C-type distribution:	
Contingency-type coefficient $r_u(\psi)$	0.458 ± 0.048
Contingency-type coefficient $r_{0.74}(\psi)$	0.497 ± 0.050
Value of the correlation parameter	0.50 (*)

(*) from Olsson (1979, pg.456).

The comparison of the results in table 3.13 shows that the coefficient $r_{0.74}(\psi)$, as function of the MLE of ψ is a very good estimate of the true correlation parameter of the table and it is similar to Olsson's polychoric coefficient.

The numerical examples presented in this section show that: first, the method of maximum likelihood for estimating the parameter ψ of the C-type distribution can provide us with a new correlation coefficient for polytomous data, given by $r_{0.74}(\hat{\psi})$ which is a good estimator of the correlation coefficient of an underlying bivariate normal distribution. Secondly, the C-type distribution agrees very well with the normal bivariate distribution in the examples presented in this section.

Computer program for the maximum likelihood method of estimating the parameter ψ

A computer program for the maximum likelihood estimator of ψ using the method presented in this chapter is available in FORTRAN. The program has as input the dimensions R and C of the contingency table and the observed frequencies of the table in free format.

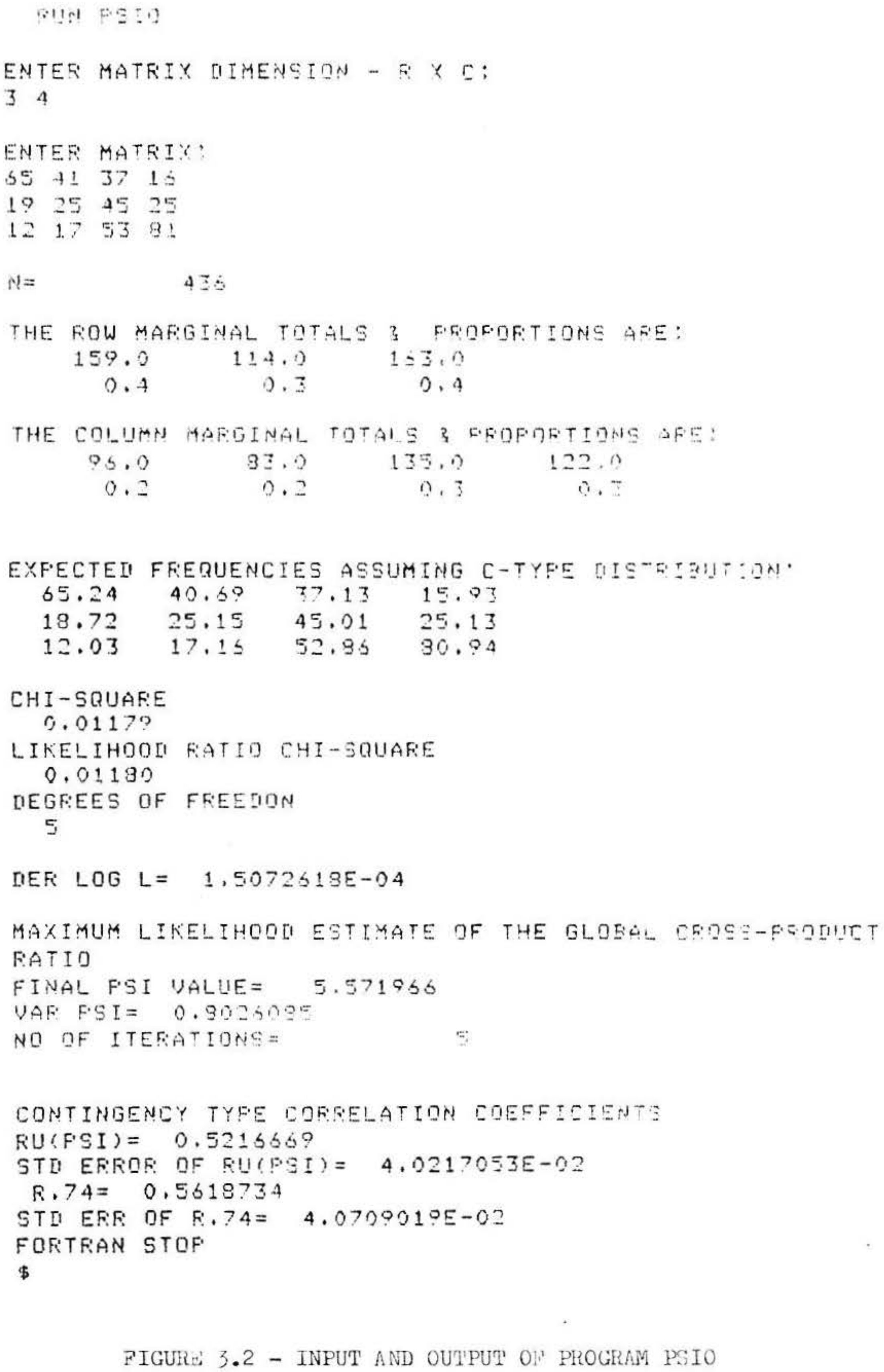
Given the input, the starting value for ψ is computed and the scoring method for parameters is used as the iterative method for obtaining the maximum likelihood estimate of ψ. The output of the program consists of final estimate, variance of the estimate and the function value at the maximum. The number of iterations and the expected proportions of the cells at convergence are also printed in the output.

The program also computes the contingency-type correlation coefficients presented in section 3.5 with the respective standard errors. As additional information the output contains the row and column marginal proportions for the RxC table and also the sample size N.

CUGAA_{B}

A listing of the program is presented in Appendix I and in Figure 3.2 in this section we present an example of the input and output of the program.

CUGAAB



```
  RUN PSIO

ENTER MATRIX DIMENSION - R X C:
3 4

ENTER MATRIX:
65 41 37 16
19 25 45 25
12 17 53 81

N=            436

THE ROW MARGINAL TOTALS &  PROPORTIONS ARE:
    159.0        114.0        163.0
      0.4          0.3          0.4

THE COLUMN MARGINAL TOTALS & PROPORTIONS ARE:
     96.0        83.0        135.0        122.0
      0.2         0.2          0.3          0.3


EXPECTED FREQUENCIES ASSUMING C-TYPE DISTRIBUTION:
  65.24   40.69   37.13   15.93
  18.72   25.15   45.01   25.13
  12.03   17.16   52.86   80.94

CHI-SQUARE
  0.01179
LIKELIHOOD RATIO CHI-SQUARE
  0.01180
DEGREES OF FREEDOM
  5

DER LOG L=  1.5072618E-04

MAXIMUM LIKELIHOOD ESTIMATE OF THE GLOBAL CROSS-PRODUCT
RATIO
FINAL PSI VALUE=   5.571966
VAR PSI=  0.8026095
NO OF ITERATIONS=              5


CONTINGENCY TYPE CORRELATION COEFFICIENTS
RU(PSI)=  0.5216669
STD ERROR OF RU(PSI)=  4.0217053E-02
 R.74=  0.5618734
STD ERR OF R.74=  4.0709019E-02
FORTRAN STOP
$
```

FIGURE 3.2 - INPUT AND OUTPUT OF PROGRAM PSIO

CHAPTER 4 : FACTOR ANALYSIS METHODS

4.1 Factor analysis for categorical data : a historical note

Since the early stages of factor analysis in the first half of this century, the need for methods of factor analysis for qualitative data has been evident. However, it seems that it was not before 1950 that the first paper dealing explicitly with the subject appeared in the literature. The paper, entitled "The factorial analysis of qualitative data" by Burt (1950) deals not only with factor analysis for dichotomous data but also with variables with "manifold classification" (polytomous variables). Burt suggested that for dichotomous variables, the following correlation matrices could be factored: Phi-coefficient; Yule's colligation coefficient; Yule' association coefficient; Pearson's coefficient of contingency, corrected for degrees of freedom by Tchuprow's formula ; Pearson's tetrachoric coefficient and the correlation ratio.

For binary variables, the tetrachoric coefficient seems to have been used earlier and also methods not based on tetrachoric were suggested, as for example, by Slater (1947) in the paper entitled 'The factor analysis of a matrix of 2×2 tables'.

For polytomous variables, Burt (1950) suggested factor analysing a symmetrical positive-definite matrix of relative frequencies. This suggestion is similar, in principle, to latent structure analysis, where the matrix of the joint occurrence proportions for pairs of items is factored.

The relationship between latent trait theory and factor analysis was explicitly formulated by Green (1952) and by Lord and Novick (1968, Ch.24). The former compares factor analysis with latent structure analysis which was formulated and developed by Lazarsfeld (1950) for binary items and with applications in Sociology. Lord and Novick show the similarities between factor analysis and latent trait theory (also for binary manifest variables) in the test theory context.

The general multiple-factor model for categorical variables has been studied by Christoffersson (1975) and Muthén (1978), for binary variables; McDonald (1969) and Bartholomew (1980) for polytomous variables. Several other contributions have been made using the "response function approach" as for example, Samejima (1969), Bock and Lieberman (1970).

Mulaik (1986) emphasizes the tendency in the direction of generalizing the ideas of common factor analysis to the analysis of covariance structures and he gives a complete account of the recent developments in the analysis of covariance structures, unified with linear structural equations modeling (see also Bentler, 1986). According to Mulaik the contemporary developments of the subject are leading to a "synthesis of linear structural equation models with latent trait models" and then he makes reference to the work by Muthén (1984) who proposes a structural equation model with a generalized measurement part, allowing for dichotomous and ordered categorical variables in addition to continuous ones. Muthén's work generalizes the Jöreskog-Sörbom (1984) LISREL methodology for structural equation models, to deal with categorical variables. In particular, the paper also extends the Muthén-Christoffersson methodolgy for factor analysis of dichotomous variables (Muthén, 1978; Muthén and Christoffersson, 1981) to handle ordered categorical and continuous data. According to

Muthén (1984) his method is computationally heavy and it is limited for practical use to small sized problems (15-20 variables), but it gives a possibility for a detailed analysis.

Everitt (1984) gives a brief account of the recent work about factor analysis for binary data. Bartholomew (forthcoming) reviews the developments in latent variable models and also give an account of the recent methodological contributions for factor analysis for categorical data.

In the next section we shall review, briefly, Bartholomew's models and methods. In Chapter 5 and 6 comparisons of numerical results using the underlying variable model based on the C-type distribution will be made with Bartholomew's models. In Section 4.3 we review the traditional factor analysis methods and in Section 4.4, we present the computer program for using the underlying variable model based on the C-type distribution as a factor analysis for categorical data method.

4.2 Bartholomew's factor analysis for categorical data models

As we pointed out in Chapter 1, Bartholomew (forthcoming) presents the models and methods of factor analysis for categorical data using two approaches to the construction of models: the Response Function (RF) approach and the Underlying Variable (UV) model approach.

Bartholomew (1980) introduced his model for factor analysis for categorical data using the response function approach. The extension of the model from the binary case to the polytomous variables case as well as the generalization for two or more latent variables (factors) was also proposed in Bartholomew (1980). In a sequence of papers, Bartholomew (1981, 1983, 1984a, 1984b, 1985) presents different aspects of the model, its generalizations and its relationship with other latent variable models.

In this section we present Bartholomew's models for binary variables according to Bartholomew (forthcoming) where also methods for fitting the model are considered.

Suppose we have a set of p dichotomous variables (items) $\underset{\sim}{x} = (x_1, x_2, \ldots x_p)$. Each row of this matrix is called a response pattern. For p binary variables, we have 2^p different response patterns. Let $f(\underset{\sim}{x})$ denote the associated probability function of $\underset{\sim}{x}$ and $\pi_i(y) = \Pr\{x_i = 1/y\}$ denote the response function giving the probability of a positive response for an individual with latent position $\underset{\sim}{y}$. Suppose $y_j (j=1,2,\ldots q)$ independent and uniformly distributed on $(0,1)$.

A general class of linear models is then defined by

$$G^{-1}\{\Pi_i(y)\} = \alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} H^{-1}(y_j) \;, \quad i=1,2,\ldots,p \qquad (4.1)$$

where the functions G^{-1} and H^{-1} are arbitrary but such that their inverses G and H are distribution functions of random variables symmetrically distributed about zero; α_{i0} and α_{ij} $(i=1,\ldots,p)$ are item parameters. From the general model (4.1) three particular cases are considered.

(a) The logit/logit model - when we take G and H as the logistic distribution function. In this case we have

$$G^{-1}(u) = H^{-1}(u) = \text{logit}(u) = \log[u/(1-u)]$$

and taking

$$\alpha_{i0} = \text{logit}\,\Pi_i = \log\,[\Pi_i/(1-\Pi_i)] \qquad i=1,2,\ldots p$$

and

$$z_j = \text{logit}(y_j) \qquad j=1,\ldots q$$

we have

$$\Pi_i(z) = \frac{\Pi_i}{\Pi_i + (1-\Pi_i)\exp\left(-\sum_{j=1}^{q}\alpha_{ij}z_j\right)} \qquad i=1,2,\ldots,p$$

which is the response function for the logit/logit model (or simply logit model). The parameter Π_i has the following interpretation: Suppose that $y_j = \frac{1}{2}$ for all j, thus $z_j = 0$ for all j and $\Pi_i = \Pi_i(\underset{\sim}{z})$. Therefore π_i is the probability of a positive response for an individual at the median position on each latent dimension.

(b) The logit/probit model - when G is the logistic distribution function and H is the normal distribution function. In this case we have

$$G^{-1}(u) = \text{logit}(u) \text{ and } H^{-1}(u) = \Phi^{-1}(u)$$

where Φ is the cumulative distribution function of a standard normal variable.

If we make the transformation $z_j = \Phi^{-1}(y_j)$ and $\alpha_{i0} = \text{logit } \Pi_i$, the response function for the logit/probit model is given by

$$\Pi_i(\underset{\sim}{z}) = \frac{\Pi_i}{\Pi_i + (1-\Pi_i) \exp (- \exp \sum_{j=1}^{q} \alpha_{ij} z_j)} \qquad i=1,2,\ldots p$$

(c) The probit/probit model - when G and H are taken as the normal distribution function. In this case $G^{-1}(u) = H^{-1}(u) = \phi^{-1}(u)$. If we make the following transformation

$$z_j = \Phi^{-1}(y_j) \qquad j=1,\ldots q$$

we have

$$\Pi_i(\underset{\sim}{z}) = \Phi(\alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} z_j)$$

It is shown by Bartholomew (forthcoming) that for binary observed variables, the response function approach and the underlying variable model approach are equivalent because they lead to the same joint probability distributions of $\underset{\sim}{x}$ under the conditions stated below.

The underlying variable model (UV) is given by

$$\xi_i = \mu_i + \sum_{j=1}^{q} \Lambda_{ij} z_j + e_i \qquad (4.2)$$

where
$$x_i = \begin{cases} 1 & \text{if } \xi_i > \tau_i \\ 0 & \text{if } \xi_i < \tau_i . \end{cases}$$

The equivalence between the RF and the UV approach exists if the distribution function of $e_i/\psi_i^{\frac{1}{2}}$ (where ψ_i is the variance of e_i) is the same as G and if

$$\alpha_{i0} = (\mu_i - \tau_i)/\psi_i^{\frac{1}{2}} \text{ and } \alpha_{ij} = \lambda_{ij}/\psi_i^{\frac{1}{2}}$$

Supposing that $\mathrm{var}(\varepsilon_i) = 1$, we have from the UV model that

$$\psi_i = 1 - \sum_{j=1}^{q} \lambda_{ij}^2$$

therefore

$$\alpha_{ij} = \frac{\lambda_{ij}}{(1-\sum_{j=1}^{q} \lambda_{ij}^2)^{\frac{1}{2}}} \quad \text{or} \quad \lambda_{ij} = \frac{\alpha_{ij}}{(1+\sum_{j=1}^{q} \alpha_{ij}^2)^{\frac{1}{2}}}$$

The λ_{ij} are the factor loadings in the factor analysis model given by (4.2), therefore the paramaters α_{ij} can also be interpreted as "factor loadings" or weights. In the test theory context usually the latent variable space is one-dimensional (q=1) and the parameters of the model have special interpretations. The parameter $\alpha_{i1} = \alpha_i$, $(i=1,\dots p)$ is the "discriminating power" of the item, because the bigger the absolute value of α_i, the greater the difference in their probabilities of giving a positive response, therefore the easier to discriminate between them (see Bartholomew, forthcoming). The parameter α_{i0} (or Π_i) are related to the "difficulty level" of item i. In test theory, the curve $\Pi_i(y)$ is referred to as the "item characteristic curve (ICC)".

We have seen that the function G in (4.1) plays the same role as the distribution of $\underset{\sim}{e}$ in (4.2). Therefore the probit/probit model, for binary variables, is precisely equivalent to the "underlying normal" factor analysis model given by (4.2) with $\underset{\sim}{z}$ and $\underset{\sim}{e}$ both standard normal variables. Hence, fitting the model using tetrachoric correlation coefficients as input to some standard factor analysis program will provide estimates of α_{ij} and α_{i0} in (4.1). The threshold value estimate τ_i is obtained by the expression

$$p_i = \Pr(x_i = 1) = \Pr(\xi_i > \tau_i) = \Phi(\tau_i)$$

$$\text{or } \tau_i = \Phi^{-1}(p_i)$$

where p_i is the observed proportion of positive responses for item i in the sample.

Bartholomew (1980) has proposed an approximate method for estimating the parameters of the model (4.1) when G^{-1} is the logit function and the items are dichotomous. The method was motivated by the fact that one natural measure of association for a 2×2 table is the cross product ratio, which should contain most of the information about the associations (or covariances) between the manifest variables. Observing that

$$\psi_{ij} = 1 + \sigma^2 \sum_{k=1}^{q} \alpha_{ik}\alpha_{jk} + O(\alpha^4) \qquad i \neq 1$$

where ψ_{ij} is the cross product ratio and σ^2 is $E(H^{-1}(y_j))^2 = E(z_j{}^2)$ for all j, Bartholomew proposes an iterative method using the first few terms of the above series as an approximation to the association between the variables, if the α's are small. The approximation is good if q = 1. The program for the method, for one-factor logit model is called MODFAC.

More efficient methods of fitting the general model given by (4.1) has been described by Bartholomew (forthcoming). One recent approach is based on the E-M algorithm, and it is similar to the method presented by Bock and Aitkin (1981) for the probit model. A program for the method, using the E-M algorithm for the maximum likelihood estimation method and fitting any of the following models: logit/logit model; logit/probit model an probit/probit model at the user's option, is called FACONE (see Shea, 1984). The program is designed for the one-factor model and for binary variables.

The factor score (y-score) of an individual for a given score pattern $\underset{\sim}{x}_k$ from the 2^p possible score patterns is computed by the FACONE program, as the posterior mean $E(y/\underset{\sim}{x}_k)$. The "components" given by

$$X_k = \sum_{j=1}^{p} \alpha_{j1} \; x_{kj}$$

are also provided by the program. For a complete description of the "components" see Bartholomew (1984b).

The FACONE program can be used for 50-60 variables but it is computationally heavy, even for powerful mainframe computers (see Shea, 1984).

Goodness-of-fit

A test of goodness of fit of the models presented in this section can be carried out using the likelihood ratio statistic given by

$$\Lambda = 2 \sum_{k=1}^{s} O_k \ln O_k/E_k$$

where O_k and E_k are the observed and expected frequencies of the score patterns, in the usual way. Λ has, approximately, a χ^2 distribution with (2^p-2p-1) degrees of freedom. This test is appropriate if the number of the manifest variables (p) is not very large. If p is large, 2^p will be large and the expected frequencies of the score pattern will be very small, making it necessary to pool adjacent score patterns; 's' in the above expression is then the number of score patterns (e.g. $s = 2^p$ if the pooling of the score patterns is not necessary).

4.3 Traditional factor analysis methods

For continuous manifest variables x_i $(i=1,\ldots p)$, the factor analysis model is given by

$$\underset{\sim}{x} = \underset{\sim}{\Lambda}\underset{\sim}{z} + \underset{\sim}{e} \tag{4.3}$$

where $\underset{\sim}{x}' = [x_1,\ldots,x_p]$, $\underset{\sim}{\Lambda} = [\lambda_{ij}]$ is the p×q matrix of factor loadings; $\underset{\sim}{z}' = [z_1,z_2,\ldots z_q]$ represent the vector of factors and $\underset{\sim}{e}' = [e_1,e_2,\ldots e_p]$ is the vector of residual terms, or error terms. With the usual specifications of the model, as given in Section 1.2 and assuming orthogonal factors ($\Phi = I$ in the usual notation) we have the dependence structure of the model given by

$$\underset{\sim}{\Sigma} = \underset{\sim}{\Lambda}\underset{\sim}{\Lambda}' + \underset{\sim}{\psi} \tag{4.4}$$

where $\underset{\sim}{\psi}$ is the dispersion matrix of $\underset{\sim}{e}$.

Several methods for estimating the parameters of the traditional factor analysis model have been presented and discussed in the literature. At the present stage of computing development, almost all methods are available in the most popular computer packages, such as SPSS-X (Statistical Package for Social Sciences; X series), BMDP (Biomedical Computer Programs; P-series) and SAS (Statistical Analysis System). A brief description of the methods available in the programs is given in this section. We also consider two of the main practical problems faced by the user of factor analysis programs: the number of factors problem and the criteria for judging the adequacy of a factor analysis solution.

Factor Analysis methods

The maximum likelihood factor analysis method (ML or MLFA) is the best method for estimating the parameters of the model from the point of view of statistical properties, when the underlying distribution of the variables is the multivariate normal distribution. It was introduced by Lawley (1940) (see also Lawley and Maxwell, 1971) and it produces estimates with the properties of asymptotic efficiency and invariance under changes of scale of the variables.

The generalized least-squares (GLS) method minimizes the sum of the squares of the differences between the observed and reproduced dispersion matrices (ignoring the diagonals). In this method the covariances are weighted inversely by the error variance (uniqueness) of the variables, that is, correlations involving variables with high error variance are given less weight than correlations involving variables with low error variance.

The unweighted least squares (ULS) method also minimizes the sum of the squared residuals, but the dispersion matrix is unweighted.

One of the computational procedures for the MLFA, GLS and ULS methods are described by Joreskog and van Tillo (1971). In general, the methods attempt to fit a dispersion matrix $\mathbf{\Sigma}$ given by the model (4.4) to the observed dispersion matrix $\mathbf{S}$ by the minimization of a function $F(\mathbf{S},\mathbf{\Sigma}(\mathbf{\Lambda},\boldsymbol{\psi}))$. This function is different for each of the three methods. The conditional minimum of F for given $\boldsymbol{\psi}$ is found, giving a function $f(\boldsymbol{\psi})$. Using the Newton-Raphson procedure, $f(\boldsymbol{\psi})$ is then minimized numerically. Function values and derivatives of first and second order are given in terms of the eigenvalues and eigenvectors of a matrix $\mathbf{A}$, say. For the MLFA and GLS methods, $\mathbf{A} = \boldsymbol{\psi}^{\frac{1}{2}}\mathbf{S}^{-1}\boldsymbol{\psi}^{\frac{1}{2}}$, for ULS, $\mathbf{S} = \mathbf{S}-\boldsymbol{\psi}$. GLS and MLFA yield estimates with the same asymptotic properties, when multivariate normality is assumed. GLS is also scale free.

The maximum likelihood factor analysis method has an associated Chi-square statistic for testing the number of factors of the model. Given the ML estimates of Λ and ψ it is possible to test the hypothesis that the q-factor model accounts satisfactorily for the covariances of the observed variables. The likelihood ratio test statistic is given by

$$T = n'\min F(\underset{\sim}{S},\underset{\sim}{\Sigma}(\underset{\sim}{\Lambda},\underset{\sim}{\psi})) \qquad (4.5)$$

where $f(S,\Sigma,(\Lambda,\psi))$ is the likelihood criterion minimized and $n' = n-1-(2p+5)/6-2q/3$. The statistic T given in (4.5) is tested as a chi-square variable with $\frac{1}{2}[(p-q)^2-(p+q)]$ degrees of freedom. Usually the test is used as a sequential procedure for determining q. For the multivariate asymptotic distribution of sequential chi-square test statistics see Steiger, Shapiro and Browne (1985).

Rao's canonical factor analysis (RAO method in SPSS) was introduced by Rao (1955) and provides one of the possible solutions to the maximum likelihood equations of lawley (1940). Rao's factors are derived from the latent vector of the matrix $\underset{\sim}{\Lambda} = \underset{\sim}{\psi}^{-\frac{1}{2}}\underset{\sim}{S}\underset{\sim}{\psi}^{-\frac{1}{2}}$. A program for the method is available in earlier versions of the SPSS package and it is no longer available in SPSS-X.

Principal factor analysis (PFA) or principal axis factoring (PAF), one of the popular methods of factor analysis, employs an iterative procedure for improving the estimates of the communalities. The initial trial for the communalities is, usually, the well-known Guttman's greatest lower bound for the communality given by the square multiple correlation of each variable on the remaining p-1. The main reference for the method is Harman (1976). PFA determines the factor matrix (matrix of factor loadings) from the eigenvectors of the matrix $\underset{\sim}{S}-\underset{\sim}{\psi}$, the reduced correlation matrix.

Alpha factor analysis (ALPHA) was suggested by Kaiser and Caffrey (1965) and it is based upon the psychometric concept of generalizability, the measure of which is known as the Kuder-Richardson reliability coefficient (or Crombach's alpha). The factors determined by this method have maximum "generalizability" and only factors with positive generalizability are retained. It is equivalent to saying that the factors associated with the eigenvalues greater than one of the matrix $\underset{\sim}{A} = [\text{diag}(\underset{\sim}{S}-\underset{\sim}{\psi})]^{-\frac{1}{2}}\, \underset{\sim}{S}[\text{diag}(\underset{\sim}{S}-\underset{\sim}{\psi})]^{-\frac{1}{2}}$ are retained. This method has the property of invariance under change of scale of the variables and operates in the metric of the communalities. In ALPHA factor analysis, the variables included in the analysis are considered a sample from the universe of variables. This involves a psychometric inference, not a statistical inference in the usual sense.

Image factor analysis (IMAGE) and Little Jiffy (LJIFFY) methods are based in the image theory introduced by Guttman. The methods were developed by Kaiser (1963, 1970). The Little Jiffy method is essentially Image analysis with some modifications on the rule of decision about the number of factors. In Image analysis, the number of factors are determined by the eigenvalues greater than one of the Image covariance matrix, given by $[\text{diag } \underset{\sim}{S}^{-1}]^{\frac{1}{2}}\, \underset{\sim}{S}\, [\text{diag } \underset{\sim}{S}^{-1}]^{\frac{1}{2}}$. For the Little Jiffy method, the number of factors is determined from the unaltered correlation matrix. This number is usually smaller than for Image analysis. The estimates of the factor loadings for the factor that are retained are the same in both methods, but the communalities are different.

Alpha factor analysis results are different, in general, from the Image or Little Jiffy results. These three methods yield considerably different results from MLFA, GLS, ULS or PFA and are not recommended for general use unless the user has specific reasons for doing so.

Principal Component Analysis (PCA) is the well known method developed by Hotelling, where the eigenvalues and eigenvectors of the unaltered correlation matrix are obtained. It is not properly a factor analysis method, although it can be used with the same purposes as factor analysis, in which case, only components (factors) associated with eigenvalues greater than one are recommended to be retained. The method is available in all three packages in the factor analysis chapter.

In Table 4.1 we compare the methods with relation to the different matrices, from which the eigenvalues (eigenvectors) are calculated.

Table 4.1 - Matrices from which the eigenvalues are calculated for each different factor analysis method.

Method	Matrix
MLFA	$\psi^{\frac{1}{2}} S^{-1} \psi^{\frac{1}{2}}$
GLS	$\psi^{\frac{1}{2}} S^{-1} \psi^{\frac{1}{2}}$
RAO	$\psi^{-\frac{1}{2}} S \psi^{-\frac{1}{2}}$
ULS	$S-\psi$
PFA/PA2	$S-\psi$
ALPHA	$[\mathrm{diag}(S-\psi)]^{-\frac{1}{2}}(S-\psi)\mathrm{diag}(S-\psi)]^{-\frac{1}{2}}$
IMAGE/LJIFFY	$[\mathrm{diag}\ S^{-1}]^{\frac{1}{2}} S\ [\mathrm{diag}\ S^{-1}]^{\frac{1}{2}}$
PCA	S

Note: S is the observed dispersion matrix.

The number of factors problem

One important issue for a successful factor analysis solution is the decision rule about the number of factors. Several criteria have been suggested in the literature and the problem is not yet completely solved. The SAS (1985) User's guide points out, properly, that "no computer program is capable of reliably determining the optimal number of factors since the decision is ultimately subjective. You should not accept blindly the number of factors obtained by default. Use your own judgment to make an intelligent decision". This citation from the SAS manual shows very well how the emphasis on the 'number of factor' problem has changed in the last few years in the user guides instructions. A few years ago few options other than the default criterion of the packages - usually retaining the factors associated with the eigenvalues (of the correlation matrix) greater than one - were available. Now, if not completely solved the problem, the user is advised to try different decision rules, what is becoming easier with each new version of the factor analysis programs. (The SPSS-X program allows more than one kind of factor analysis solution each time we use the program).

The criterion of retaining the factors that have eigenvalues greater than one is known as the "Kaiser criterion" (Kaiser, 1960). This criterion has an intuitive appeal in the sense that only factors, that account for at least as much variance as does a single variable, are retained.

Cattell (1966) proposed a test based on the fact that the magnitude of the eigenvalues would cease to change very much after the nontrivial common variance had been removed from a correlation matrix.

It is called the "scree test" and consists in observing the plot of the eigenvalues (scree plot) looking for "breaks" in the curve. The name "scree" comes from the resemblance of such a plot to the rock slope of a mountain with a mass of rubble called the scree or talus at the bottom. Experimental evidence indicates that the scree begins at the q-th factor, where q would be the true number of factors. Cattell suggests that the scree test can be objectified by taking the first differential of the curve and finding at what point it departs significantly from zero (Cattell and Jaspers, 1967, p.41). Although the scree test may be very useful in many cases, it is still a subjective decision rule as we have no adequate definition of what a "break" is. As we shall see in Chapter 7, the occurrence of improper solutions in factor analysis is strongly related with an inappropriate decision about the number of factors. Therefore calling attention to the number of factors problem as the SAS manual does and including several options in a factor analysis program will certainly lead to more successful analysis by the increasing number of users of the statistical analysis packages.

The number of factors problem is discussed in some detail by several authors as, for example, Thorndike (1978), Gorsuch (1983), Cureton and D'Agostino (1983) among others. Thorndike also reviews several criteria to define the adequacy of a factor analysis solution. The criteria for factor solutions are related to the number of factors problem and it is another of the practical problems faced by the user of factor analysis programs.

Criteria for judging the adequacy of a factor analysis solution

One of the most used criteria for testing the adequacy of a factor analysis solution is the Goodness-of-fit test associated with the maximum likelihood factor analysis method (see expression 4.5). The test was derived under restricted conditions, that is, for continuous variables, multinormally distributed and analysis based on the sample covariance matrix as opposed to the sample correlation matrix. These conditions are seldom met in practice. Some authors (e.g. Gorsuch, 1983) have pointed out that because the test is dependent upon sample size, for large samples, a model that is trivially false is likely to be rejected.

According to Thorndike (1978), a potentially useful way to judge whether an additional factor adds enough information to the previous solution is to compare the matrix of residual correlations obtained with the extra factor to the matrix obtained without it. The residual correlation matrix is the difference between the reproduced correlation matrix by the model and the observed correlation matrix. This matrix is now available as an option in all three factor analysis packages. The SAS program prints also the root mean square residual (RMS) given by

$$\mathrm{RMS} = [\sum_{i<j} (r_{ij}-\hat{r}_{ij})^2/k]^{\frac{1}{2}}$$

where k is the number of off-diagonal elements of the residual matrix, which is $p(p-1)/2$; r_{ij} are the observed correlation coefficients and $\hat{r}_{ij}$ are the correlation coefficients reproduced by the model. This measure is not available in the other programs, but SPSS-X prints the proportion of elements of the residual matrix that are greater, in absolute value, than 0.05. This clearly is not an appropriate criterion because the magnitude of the residuals depend on sample size. Thurstone (1947) has used the criterion of insignificant residuals as an index of the adequacy of a factor analysis solution. McNemar (1942) has discussed the theory of residual distribution in an effort to provide a statistical test of when the residuals are sufficiently small (cf. Thorndike, 1978). Under McNemar's proposal, factors should be extracted until the distribution of the residuals had a standard deviation no greater than the standard error of a zero-order correlation. We agree with Thorndike when he says that "although this approach would seem to merit consideration, it has not been widely used".

The root-mean-square residual (RMS) is appropriate for comparing different solutions (different number of factors) from the same correlation matrix, although the criteria may lead to the inclusion of too many factors because the residuals approximate to zero, as the number of factors increase.

Finally, we shall quote here a comment made by Johnson and Wichern (1982) - unfortunately, they say, "the criterion for judging the quality of any factor analysis has not been well quantified. Rather it seems to depend on a 'WOW criterion'. If while scrutinizing the factor analysis, the investigator can shout: 'Wow! I understand these factors - the application is deemed successful'". This shows that the adequacy of a model must be judged by multiple criteria.

To avoid the subjective element two relatively new criteria have been introduced in the factor analysis program of the SAS package. The criteria are associated with the maximum likelihood factor analysis method:

Akaike's Information Criterion (AIC) for the MLFA mmethod (Akaike, 1973, 1983) is a general criterion for estimating the best number of parameters to include in a model if MLFA is used. The number of factors that yields the smallest value of AIC is considered best. According to the SAS (1985) user's guide, the Akaike's criterion like the chi-square test, "tends to include factors that are statistically significant but inconsequential for practical purposes".

Schwarz's Bayesian criterion (SBC) for the MLFA method (see Gweke and Singleton, 1980) is also a criterion for determining the best number of parameters of a model if MLFA is used. The number of factors that yields the smallest value of SBC is considered best. According to SAS instructions, SBC seems to be less inclined to include trivial factors than either AIC or the chi-square test.

Methods and options available in factor analysis routines

A comparative study of the factor analysis programs in three packages: SPSS, BMPD and SAS was presented by MacCallum (1984), therefore we only summarize the comparisons in terms of options available in the most current available versions of the packages, as for example the new SPSS-X, which was not included in MacCallum's study. The versions considered here are as follows:

BMDP, April 1985 version	[Dixon et al, 1983]
SAS, Version 5, 1985	[SAS Institute Inc, 1985]
SPSS-X Release 2.1, 1986	[SPSS^X Inc, 1986; Norusis, 1985]

SPSS-X seems superior to the other two packages with relation to the limitation problem. In SPSS-X there is no limitations to the number of variables, the number of analysis, the number of extractions or the number of rotations. It is the only program at the moment where more than one extraction method for a given 'ANALYSIS' subcommand can be specified. The user can also specify more than one rotation method for a given extraction or even more than one analysis for each problem. SPSS-X now accepts a correlation matrix in lower triangular form as BMDP does, but SAS still only accepts a correlation matrix in square symmetric form. SPSS-X does not accept a covariance matrix as input and this is a disadvantage of the package. The flexibility in BMDP with respect to input is then one point in favour of BMDP, but the limitations in the number of extraction methods in one analysis makes BMDP and SAS inferior to SPSS-X at the moment. Other comparisons can be made from the options available in each package. Table 4.2 presents the methods and options available in each package.

Table 4.2 - Options available in the current versions of the BMDP, SAS and SPSS-X factor analysis programs

	BMDP	SAS	SPSS-X
Maximum number of variables in the analysis	100 variables (60 with MLFA)	250 variables	No limit
Forms of input:			
Raw data	Y	Y	Y
Correlation Matrix	Y	Y	Y
Covariance Matrix	Y	Y	N
Prior solution	Y	Y	Y
Factoring methods:	Y	Y	Y
Principal factor analysis	Y	Y	Y
Maximum likelihood	N	N	Y
Generalized least squares	N	Y	Y
Unweighted least squares	N	Y	Y
Alpha factor analysis	N	Y	Y
Image factor analysis	N	Y	Y
Kaiser's Little Jiffy	Y	N	N
Harris component analysis	N	Y	N
Principal component analysis	Y	Y	Y
Communality estimates:			
Square multiple correlation (SMC)	Y	Y	Y
Adjusted SMC	N	Y	N
Maximum absolute correlation	Y	Y	N
A list of values (input)	Y	Y	Y
Random no. (uniform distr.)	N	Y	N
Number of factors criteria			
Maximum number (input)	Y	Y	Y
minimum eigenvalue	Y	Y	Y
proportion of variance	N	Y	N
Rotation methods			
Varimax	Y	Y	Y
Quartimax	Y	Y	Y
Equimax	Y	Y	Y
Orthomax	Y	Y	N
Promax	N	Y	N
Orthoblique	Y	Y	N
Direct oblimin	Y	N	Y
Direct Quartimin	Y	N	N
Procrustes	N	Y	N

Adequacy of FA criteria			
goodness-of-fit (χ^2) (ML)	N	Y	Y
Likelihood criterion (ML)	Y	N	N
Residual correlation	Y	Y	Y
Root mean square residual	N	Y	N
Akaike's Information criterion(ML)	N	Y	N
Schwarz's Baysian criterion (ML)	N	Y	N
Measure of sampling adequacy	N	Y	Y
Plots			
Scree plot (eigenvalues)	N	Y	Y
Factor loadings	Y	Y	Y
Rotated factor loadings	Y	Y	Y
Factor scores	Y	N	N
Factor scores method			
Regression method	Y	Y	Y
Bartlett method	N	N	Y
Anderson-Rubin method	N	N	Y

Note: Y = Yes ; N = No.

4.4 Factor analysis for categorical data: the underlying variable model based on C-type distributions

In this section we present the factor analysis for categorical data method, introduced in Chapter 1, which we call the underlying variable factor analysis method (UVFA) based on the C-type distribution. The theoretical framework for the method is given in Section 1.2, for any underlying distribution of the manifest variable. We then suppose that the C-type distribution is the underlying distribution and contingency type correlation coefficients (functions of the parameter of the association of the C-type distribution) are then obtained. A maximum likelihood method for estimating the parameter of the association of the C-type distribution (the global cross product ratio from R×C contingency tables) is presented in Chapter 3. We then suggest using the contingency-type correlation coefficients as input to the factor extraction methods available in the computer routines. A computer program that yields the contingency-type correlation matrix for using as input in the statistical packages is now described.

A FORTRAN program, call CROSSPSI, was designed for the maximum likelihood method of estimating the parameter of association (ψ) of the C-type distribution and contingency-type correlation coefficients. The program reads the raw data and cross-tabulates the p variables. The parameter ψ for each of the $p(p-1)/2$ crosstables is estimated by the ML method using the iterative method of scoring. The correlation coefficients using Mardia's formula , $r_u(\psi)$, and Chambers' formula , $r_{0.74}(\psi)$ (see Section 2.3) are then calculated. The output of the CROSSPSI program consists of the contingency-type correlation matrices in lower triangular form, in appropriate format for subsequent use in factor analysis programs. The input system of the CROSSPSI program is described and the output is illustrated with an example.

The input for CROSSPSI consists of the data file and some simple information about the data file:

1) Title of the problem
2) Number of cases (sample size)
3) Number of variables
4) Number of categories of each variable
5) Input format of the raw data.

Input format for the CROSSPSI program

1) The title of the problem should be input using the following format:

```
(1H0, 'TITLE')
```

where "TITLE" is the title of the problem to be analysed and should contain at most 20 characters.

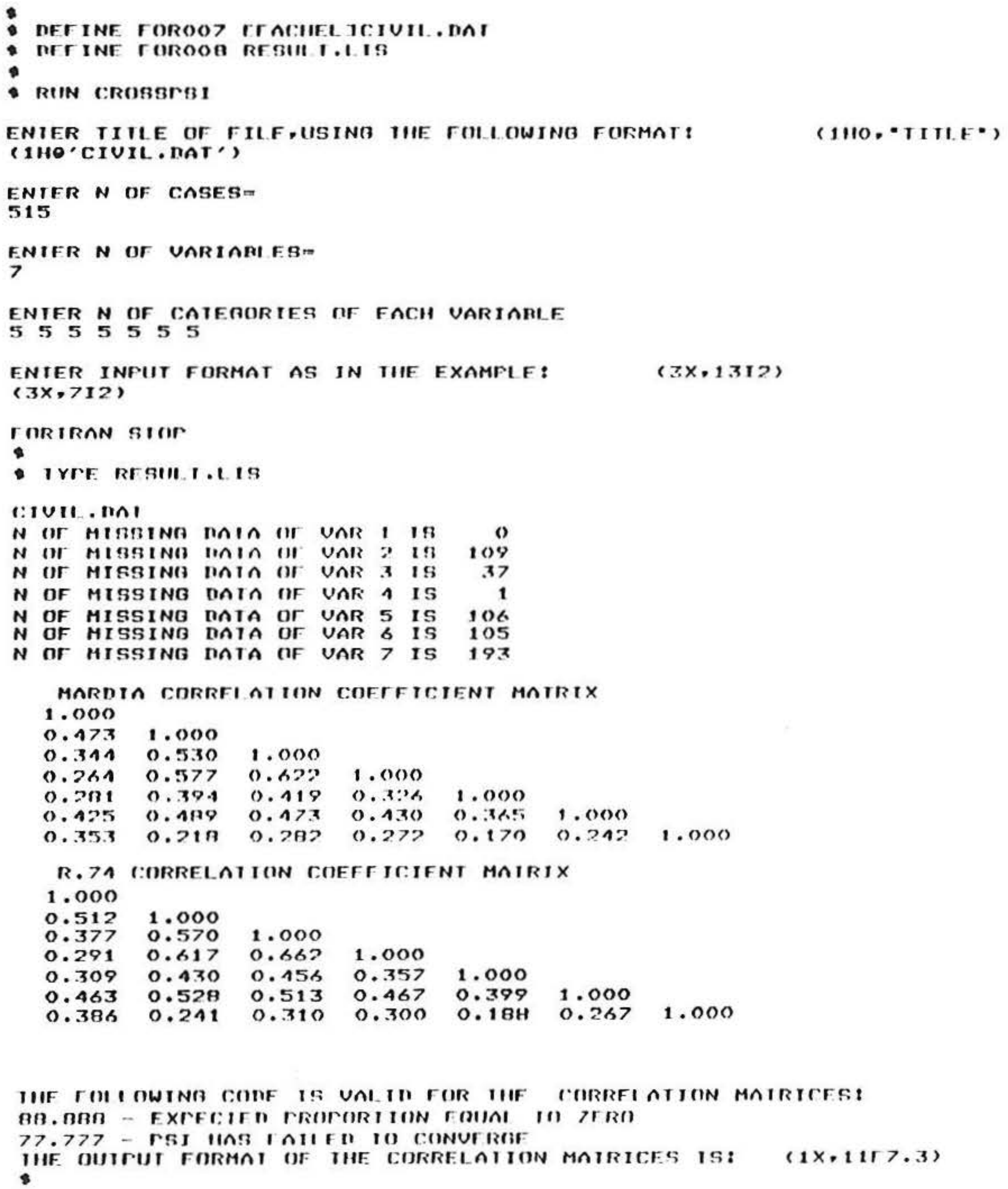
2) Number of cases - free format

3) Number of variables - free format

4) Number of categories of each variable: free format, but one integer number (c_i) should be input for each of the p variables.

5) The input format of the data to be read in should conform with a FORTRAN format as for example (kX, pIw) where k means the number of spaces (if any) at the beginning of each data file line; p is the number of variables and Iw implies the variable to be read in is of type INTEGER and occupies a field of width w in the current data line. (The format specification Fw.0 may be used instead of Iw).

The program is written in FORTRAN 77. The input channel number is assumed to be 5 for the input items 1 to 5 given above and the input channel number 7 should be used for the data file (raw data). The



```
$
$ DEFINE FOR007 [FACHEL]CIVIL.DAT
$ DEFINE FOR008 RESULT.LIS
$
$ RUN CROSSPSI

ENTER TITLE OF FILE,USING THE FOLLOWING FORMAT!          (1H0,"TITLE")
(1H0'CIVIL.DAT')

ENTER N OF CASES=
515

ENTER N OF VARIABLES=
7

ENTER N OF CATEGORIES OF EACH VARIABLE
5 5 5 5 5 5 5

ENTER INPUT FORMAT AS IN THE EXAMPLE!        (3X,13I2)
(3X,7I2)

FORTRAN STOP
$
$ TYPE RESULT.LIS

CIVIL.DAT
N OF MISSING DATA OF VAR 1 IS     0
N OF MISSING DATA OF VAR 2 IS   109
N OF MISSING DATA OF VAR 3 IS    37
N OF MISSING DATA OF VAR 4 IS     1
N OF MISSING DATA OF VAR 5 IS   106
N OF MISSING DATA OF VAR 6 IS   105
N OF MISSING DATA OF VAR 7 IS   193

   MARDIA CORRELATION COEFFICIENT MATRIX
  1.000
  0.473  1.000
  0.344  0.530  1.000
  0.264  0.577  0.622  1.000
  0.281  0.394  0.419  0.326  1.000
  0.425  0.489  0.473  0.430  0.365  1.000
  0.353  0.218  0.282  0.272  0.170  0.242  1.000

   R.74 CORRELATION COEFFICIENT MATRIX
  1.000
  0.512  1.000
  0.377  0.570  1.000
  0.291  0.617  0.662  1.000
  0.309  0.430  0.456  0.357  1.000
  0.463  0.528  0.513  0.467  0.399  1.000
  0.386  0.241  0.310  0.300  0.188  0.267  1.000


THE FOLLOWING CODE IS VALID FOR THE  CORRELATION MATRICES!
88.888 - EXPECTED PROPORTION EQUAL  TO ZERO
77.777 - PSI HAS FAILED TO CONVERGE
THE OUTPUT FORMAT OF THE CORRELATION MATRICES IS:   (1X,11F7.3)
$
```

FIGURE 4.1 - INPUT AND OUTPUT OF PROGRAM CROSSPSI

output is printed in a new file using the output channel number 8.

Observation: Using the program on the VAX computer at LSE, the commands for defining the input and outputs files before running the CROSSPSI program should be:

```
DEFINE FOR007  DATA FILE NAME
DEFINE FOR008  OUTPUT FILE NAME
```

Example

The Civil Service I data (see Chapter 6, Section 6.2) are used to illustrate the input and output of the program CROSSPSI. In Figure 4.1 we show the sequence of commands to be used.

In the example we use only the seven first variables of the Civil Service I data for illustrative purposes. The raw data file called CIVIL.DAT consists of seven categorical variables, each one with five categories and the data file format is (3X, 7I2). The input items are answered iteratively in a computer terminal session as illustrated in Figure 4.1. The output file, which in the example is called RESULT.LIS is printed and it contains:

1) Number of missing data for each variable. CROSSPSI uses pairwise deletions of missing data; that is, if the value of any variable in a case is missing or out of range, the case is omitted from the computation of the correlation coefficients of this variable with any other variable.

2) $R_u(\psi)$ correlation matrix, or Mardia correlation matrix which is formed by the contingency type correlation coefficients using Mardia's formula $r_u(\psi)$ given in Section 2.3.

3) $R_{0.74}$ correlation matrix which is formed by the contingency type correlation coefficients using Chambers' formula $r_{0.74}(\psi)$ given in Section 2.3.

The output format of the correlation matrices is in the lower triangular form and the standard FORTRAN format is used (1X,11F7.3). The format is appropriate for using as input for the factor analysis program of the BMDP and SPSS-X. (The current version of the SAS factor program requires the square correlation matrix and a special SAS format for this program, therefore the whole matrix should be input).

Portability of the CROSSPSI program

No auxiliary routine is used in the CROSSPSI program. As a FORTRAN program, it can be used in any machine with a FORTRAN compiler, including the microcomputers. The limitations of the program depends on the available memory space of the machine. A listing of the program is included in Appendix I. The dimension parameters of the program are defined as follows:

MAX1 - the maximum value of the sample size to be used for any particular user

MAX2 - the maximum number of variables

MAX3 - the maximum number of categories for the variables

MAX4 - the maximum number of elements of the lower triangular correlation matrix (P(P-1)/2). It should be calculated as MAX4 = MAX2 (MAX2-1)/2

For the particular listing of the program presented in Appendix I, these limits are: MAX1 = 1800; MAX2 = 50; MAX3 = 10; MAX4 = 1225.

If the machine to be used by the user allows an increase in the capacity of the program, this is easily done by changing the parameters on the first line of the main program and in the first line of the subroutine PSIO, where MAX = MAX3.

We have described how to use the program (CROSSPSI) for obtaining contingency type correlation matrices, which may thus be used as input to the traditional factor extraction methods. This method allows us to obtain factor analysis results for categorical variables without restrictions on the number of variables. The limitations on the number of variables for the underlying variable factor analysis method based on the C-type distribution, as described in this section are, therefore, the same limitations as for the traditional factor analysis methods (see Section 4.3). Examples of the use of the UVFA method for categorical data will be presented in the next two chapters, for binary and for polytomous variables. Comparisons with Bartholomew's model will be made.

CHAPTER 5 : NUMERICAL APPLICATIONS FOR BINARY DATA AND COMPARISONS WITH BARTHOLOMEW'S MODELS

5.1 Introduction

In this chapter we shall make a comparative study of different factor analysis methods, including the Bartholomew's factor analysis for categorical data models, the underlying variable model based on the C-type distribution and traditional factor analysis methods. We shall apply the methods to six data sets from empirical experiments that have been used in the literature related with the subject. For all cases the manifest variables are dichotomous. The data sets are summarized as follows:

Set Numbers	Name	No of variables	Sample size
1	Weinreich Data	5	802
2	Abortion Data	6	1386
3	Andersen Data	5	600
4	Lombard Data	4	1729
5	McHugh Data	4	137
6	Goodman Data	4	1000

We shall describe and analyse each data set separately. For each data set, four different correlation matrices are used as input to the traditional factor analysis methods. The four cases are described below.

CASE I - INPUT : TETRACHORIC COEFFICIENTS. In this case we shall assume that the underlying distribution of the variables are really continuous and normal. The Pearson tetrachoric correlation coefficient is calculated using a computer routine from BMPD. In this program the bivariate normal integral is approximated by an infinite series and the coefficient is found implicitly by iteration. If the series does not converge within 100 terms, Gaussian quadrature is used to evaluate the integral (Brown and Benedetti, 1977).

CASE II - INPUT : CHAMBERS CORRELATION COEFFICIENTS. This is a contingency type coefficient, that is, a function of the parameter of association of the C-type distribution. The coefficient was introduced by Chambers as an estimate of the latent coefficient of the bivariate normal.

The underlying variable model approach, with normal marginal distributions, is assumed in this case. The coefficient is given by

$$r_{0.74} = \frac{\psi^{.74} - 1}{\psi^{.74} + 1}$$

CASE III - INPUT : MARDIA CORRELATION COEFFICIENTS. This coefficient was introduced by Mardia (1967) studying the moments of the C-type uniform distribution. In this case we assume that the underlying marginal distribution functions correspond to the uniform distribution on (0,1). It is given by

$$r_U(\psi) = \frac{\psi+1}{\psi-1} - \frac{2\psi \ln \psi}{(\psi-1)^2}$$

where ψ is the cross product ratio estimated from the 2×2 observed tables. (See Chapter 2, Section 2.3).

CASE IV - INPUT : PHI COEFFICIENTS. As a comparative study we shall also use the familiar Phi-coefficient, an acceptable measure of correlation when the variables are purely two-valued qualitative attributes (see Chambers, 1982). The Phi-coefficient corresponds to the product-moment correlation coefficient evaluated from the binary data.

We have already pointed out that on using Tetrachoric, Chambers or Mardia coefficients as input to the factor analysis methods, the correlation matrices may not be Gramian, but in all examples to be shown in this chapter, the correlation matrices are positive definite.

The estimated parameters, comparisons and further considerations about goodness of fit of the models will be presented for each data set. The comparisons wil be made directly on basis of the factor loadings. We have reparameterized the parameter estimates of Bartholomew's models in order to compare with the heuristic estimates (see Section 4.2).

The determination of the number of common factors is a difficult and often controversial problem of factor analysis. Again, for comparing with Bartholomew's models we have decided to use the one-factor model for all methods, but we shall consider and comment on the solutions in which the number of factors of the model is extracted by the "default criterion" of each factor analysis program. This number generally corresponds to the number of eigenvalues greater than one from appropriate matrices.

For a description of the Bartholomew models see Section 4.2. In this chapter we shall use two different estimation methods and two models: for the logit model, we use the approximation method using as input the cross product ratio (Bartholomew, 1980) which program is called MODFAC. For the probit/probit model we shall use the maximum likelihood method and the program FACONE (Shea, 1984).

With respect to the traditional factor analysis methods, we shall analyse the data sets using the methods available in the BMDP and SPSS packages. As was pointed out in Section 4.3, the RAO method (SPSS) yields the same results as MLFA (BMDP) provided we use the RAO program with an additional input card allowing a greater number of iterations than the default criterion (25 iterations). We have used a "N of ITERATION = 1000" card as input, and for all data sets analysed in this chapter we obtained identical solutions with RAO and MLFA methods. The common solution will be presented as the maximum likelihood factor analysis method (MLFA).

PFA (BMPD) and PA2 (SPSS) are also equivalent methods with small differences between them. However, for all data sets of this chapter the two methods yielded the same results and only PFA will be presented in the tables of results. We also include the results for two other factor analysis methods : ALPHA (SPSS) and LJIFFY (BMDP). We observe that IMAGE (SPSS) method is equivalent to LJIFFY. For illustrative purposes we also include the PCA (Principal Component Analysis) method.

Although the SAS package has some advantages compared with the BMDP and SPSS factor analysis programs as pointed out by MacCallum (1983), we shall not use SAS because it uses a more complex input system which is inconvenient for dealing with too many different data sets and various correlation matrices as in this study.

5.2 Weinreich data

The Weinreich data consists of five items concerning allergic reactions and it is taken from Weinreich (1982). The result for each item was "no reaction" or "positive reaction" from an allergy test of 802 patients with five sorts of grasses; (1) Onion Couch; (2) Fescue grass; (3) Couch grass; (4) Cock's foot grass and (5) Rye grass.

The cross product ratios for this set are presented in Table 5.1 and the correlation matrices in Table 5.2.

Table 5.1 - Cross-Product Ratios; Weinreich Data

Items	1	2	3	4
2	34.94			
3	29.54	32.20		
4	46.63	29.81	28.66	
5	17.92	31.52	22.68	16.25

Table 5.2 - Correlation Matrices for Weinreich Data (*)

	Tetrachoric Coefficients				Chambers Coefficients			
Items	1	2	3	4	1	2	3	4
2	.90				.87			
3	.88	.89			.85	.86		
4	.91	.87	.88		.89	.85	.85	
5	.81	.87	.85	.81	.79	.85	.82	.77

	Mardia Coefficients				Phi-Coefficients			
Items	1	2	3	4	1	2	3	4
2	.84				.71			
3	.82	.83			.67	.68		
4	.87	.83	.82		.70	.65	.68	
5	.76	.83	.79	.74	.58	.64	.64	.60

(*) The diagonal is omitted

Table 5.3 - Weinreich data: Factor loadings obtained by different factor analysis methods for various correlation matrices and the reparameterised factor loadings obtained by Bartholomew's methods. One-factor model. (*)

CASE I - INPUT: TETRACHORIC CORRELATION COEFFICIENTS						Bartholomew's methods	
ITEM	MLFA	PFA	ALPHA	LJIFFY	PCA	MODFAC	FACONE
1	95	94	94	93	95	96	94
2	95	95	95	93	96	96	95
3	94	94	94	92	94	94	94
4	94	93	93	92	95	95	94
5	89	89	89	88	90	90	89

CASE II - INPUT: CHAMBER COEFFICIENTS ($r_{.74}$)							
ITEM	MLFA	PFA	ALPHA	LJIFFY	PCA	MODFAC	FACONE
1	93	93	93	91	94	96	94
2	94	94	94	91	95	96	95
3	92	92	92	90	94	94	94
4	92	92	92	90	93	95	94
5	87	87	87	86	91	90	89

CASE III - INPUT: MARDIA COEFFICIENTS							
ITEM	MLFA	PFA	ALPHA	LJIFFY	PCA	MODFAC	FACONE
1	92	92	92	89	93	96	94
2	93	93	93	90	94	96	95
3	90	91	91	88	93	94	94
4	91	90	90	88	92	95	94
5	85	85	85	84	89	90	89

CASE IV - INPUT: PHI-COEFFICIENTS							
ITEM	MLFA	PFA	ALPHA	LJIFFY	PCA	MODFAC	FACONE
1	83	83	83	77	86	96	94
2	83	83	83	77	87	96	95
3	83	83	83	77	87	94	94
4	82	82	82	76	86	95	94
5	74	74	74	70	81	90	89

(*) Decimal point is omitted

CXCAAT

In Table 5.3 we present the factor loadings obtained by different factor analysis methods and for different correlation matrices used as input to the methods, according to the description in the last section.

One factor was extracted by all factor analysis methods by the default criterion. The eigenvalues for the different correlation matrices show a large value associated with the first factor and small values for the remainder. Table 5.4 shows the eigenvalues for the unaltered correlation matrices in each case under consideration.

Table 5.4 - Eigenvalues of the unaltered correlation matrices for Weinrich data

Correlation Matrix	Eigenvalues				
Tetrachoric	4.47	0.22	0.12	0.12	0.08
Chambers	4.36	0.26	0.15	0.12	0.11
Mardia	4.26	0.30	0.18	0.14	0.12
Phi	3.62	0.45	0.34	0.31	0.26

Bock and Lieberman (1970) suggest that when there is a well-defined "jump" between the value of the first eigenvalue and that of the remaining, the result would be taken to justify the unidimensionality of the latent space. Although this rule is very subjective, it seems to be useful to help in the determination of the number of factors of the model. Also the homogeneity of the items together with high values of the correlation coefficients as it is shown in Table 5.2, lead to the assumption of unidimensionality of the latent space in this example.

We see in Table 5.3 that the agreement between the factor loadings obtained by the factor analysis methods: MLFA, PFA ALPHA and

Bartholomew's model is very close for the tetrachoric coefficients and for Chambers coefficient cases. The agreement is reasonable for Mardia coefficients and it is poor for the Phi-coefficient case. We also observe that for each of the four cases considered the solution for MLFA, PFA and ALPHA are the same for the Weinreich data. LJIFFY estimates are lower than the MLFA estimates and PCA estimates are higher for all cases considered. We note, however, that the same relative order between the items is observed for almost all methods in the four cases. The items show approximately the same magnitude concerning the factor loadings, although the fifth item presents systematically the smallest. In other words, we can say that the items have approximately the same discriminating power for the Weinreich data.

Weinreich (1982), analysing the same data set, concludes that the data are well described by the Rasch model which assumes equal discriminating power for all items.

We shall now present some additional considerations about the goodness of fit and test of significance of the number of factors provided by the traditional factor analysis methods and also the goodness of fit test for the Bartholomew's models. As was pointed out in Section 4.3, two factor analysis methods, MLFA and RAO provide tests of significance for testing the number of factors.

We should note, however, that the test of the number of factors provided by the factor analysis methods is designed for the case of continuous normal manifest variables and covariance matrices and is not meaningful when the assumptions are not satisfied, as in the case of the heuristic approach. Therefore, we shall not consider in this chapter the results of the chi-square tests and our analysis of the fit of the factor analysis models will be based on the residual correlation matrix.

The goodness-of-fit test for Bartholomew's logit model (MODFAC), given by the log-likelihood ratio is Λ=60.94 with 21 d.f. ($p<0.001$). For the probit/probit model we obtained Λ=47.28 with 14 d.f. ($p<0.001$). See Section 4.2 for the description of the goodness-of-fit tests for Bartholomew's model. Both models are rejected for the Weinreich Data. This result is probably caused by the high observed frequencies in the low and high scoregroups (see Weinreich, 1982).

Next, we shall compare the methods regarding the residual correlation matrix. This matrix is supplied by the methods available in the BMDP package. From the residual correlation matrix we obtain the root mean square (RMS) of the off-diagonal elements of this matrix. (This criterion is available in the SAS factor analysis methods but only the residual matrix is presented as output in the BMDP package). In Table 5.5 we present the RMS values for each correlation matrix case and for each factor analysis methods. MLFA and PFA shows the best fit comparing the methods and the tetrachoric correlation matrix is the most perfectly reproduced by the factor analysis model.

Table 5.5 - The root mean square (RMS) of the residual correlation matrices for the different factor analysis methods.

Weinreich data

Correlation Matrix	MLFA	PFA	LJIFFY	PCA
Tetrachoric	0.018	0.018	0.036	0.034
Chambers coefficient	0.021	0.021	0.044	0.039
Mardia coefficient	0.024	0.024	0.050	0.045
Phi-coefficient	0.021	0.021	0.087	0.073

CXCAAT

From the discussion above, we can conclude that the heuristic approach is certainly justified when we use the tetrachoric correlation coefficients and the method MLFA for the Weinreich data. Although the agreement between the factor loadings estimates for Chambers' coefficient and Mardia's coefficients is not so good, the results are only slightly poorer. Finally, in case of the Phi-coefficients, the agreement between the methods is not good, the results being systematically lower.

5.3 Abortion data

The Abortion data set is taken from Clogg and Sawyer (1981) and it is related with attitudes towards legal abortion. It contains six items concerning different attitudes, each with 'yes' or 'no' responses for 1286 individuals. Respondents were asked if abortion should be legally available:

(1) If the woman's own health is seriously endangered by the pregnancy;

(2) If there is a strong chance of serious defect in the baby;

(3) If she becomes pregnant as a result of rape;

(4) If the family has a very low income and cannot afford more children;

(5) If she is not married and does not want to marry the man;

(6) If she is married and does not want more children.

The cross-product ratios and the correlation matrices used as input to factor analysis methods are presented in Table 5.6 and 5.7 respectively

Table 5.6 - Abortion Data: Cross-product ratios

Item	1	2	3	4	5
2	59.9				
3	24.2	23.0			
4	19.7	15.4	42.4		
5	18.0	17.5	104.9	43.9	
6	14.8	15.2	40.2	60.0	47.5

Table 5.7 - Abortion Data: Correlation Matrices

Tetrachoric coefficients						Chambers coefficient					
Item	1	2	3	4	5	Item	1	2	3	4	5
2	.89					2	.91				
3	.79	.82				3	.83	.82			
4	.70	.72	.83			4	.80	.77	.88		
5	.68	.72	.86	.91		5	.79	.78	.94	.88	
6	.65	.70	.80	.93	.92	6	.76	.76	.88	.91	.89

Mardia coefficients						Phi-coefficients					
Item	1	2	3	4	5	Item	1	2	3	4	5
2	.89					2	.61				
3	.80	.79				3	.50	.57			
4	.77	.73	.86			4	.32	.40	.46		
5	.76	.75	.93	.87		5	.29	.38	.43	.73	
6	.73	.73	.86	.89	.87	6	.27	.36	.41	.75	.74

(the diagonal is omitted)

In Table 5.8 we present the factor loadings obtained by various factor analysis methods and the reparameterized estimates from MODFAC program (logit model). The four cases according to the correlation matrix used are presented. One-factor model is used for all methods.

Table 5.8 - Abortion Data : factor loadings obtained by different factor analysis methods (one-factor model) and Bartholomew's MODFAC method.

CASE 1 - INPUT: TETRACHORIC CORRELATION MATRIX						Bartholomew's method
Item	MLFA	PFA	ALPHA	LJIFFY	PCA	MODFAC
1	75	81	82	79	86	95
2	78	85	86	83	89	93
3	89	93	93	90	93	84
4	96	93	92	93	93	96
5	96	93	92	93	94	96
6	95	90	89	92	92	95

CASE II - INPUT: CHAMBERS CORRELATION MATRIX						Bartholomew's method
Item	MLFA	PFA	ALPHA	LJIFFY	PCA	MODFAC
1	86	88	89	87	91	95
2	85	87	87	87	90	93
3	96	96	96	94	96	84
4	93	93	93	92	94	96
5	96	94	94	94	95	96
6	93	92	91	91	93	95

CASE III - INPUT : MARDIA CORRELATION MATRIX						
Item	MLFA	PFA	ALPHA	LJIFFY	PCA	MODFAC
1	83	86	87	85	89	95
2	83	85	86	84	88	93
3	96	95	95	93	95	84
4	92	92	92	90	93	96
5	95	93	93	93	94	96
6	91	90	90	90	92	95

CASE IV - INPUT: TETRACHORIC CORRELATION MATRIX						
Item	MLFA	PFA	ALPHA	LJIFFY	PCA	MODFAC
1	50	52	60	46	62	95
2	61	62	71	55	71	93
3	61	65	67	57	73	84
4	84	81	78	77	83	96
5	82	79	76	76	81	96
6	84	78	76	76	80	95

(decimal point omitted)

The results obtained by MODFAC method are repeated in each case to facilitate comparisons. For this data set the MODFAC method does not converge and actually two solutions were obtained after 21 iterations. We considered the best solution with respect to the log-likelihood ratio. The FACONE program could not be applied to this data set because the program uses as input the score pattern, which was not available in the cited references.

In Table 5.9 we present the eigenvalues of the unaltered correlation matrix for each case.

Table 5.9 - Abortion data: Eigenvalues of the unaltered correlation matrix

TETRACHORIC MATRIX	4.98	0.60	0.17	0.11	0.07	0.06
CHAMBERS MATRIX	5.20	0.40	0.15	0.11	0.07	0.06
MARDIA MATRIX	5.08	0.46	0.18	0.13	0.08	0.06
PHI-COEFFICIENT MATRIX	3.43	1.20	0.48	0.37	0.27	0.24

In this example it is not clear whether a single factor explains the data sufficiently well or more factors are required. The eigenvalue rule depends on which case we are regarding. For the tetrachoric, Chambers and Mardia coefficient cases, the one-factor model clearly emerges, but using phi-coefficients two factors seem necessary.

Observing Table 5.8 we can see that the loadings, comparing the methods are more heterogeneous than in the previous example. There is no close agreement between the first three loadings for the logit model and those obtained by traditional factor analysis methods. Within each case there is not much difference between the various traditional methods.

The goodness of fit test for the Bartholomew's logit model is 19.33 with 7 degrees of freedom ($P < 0.005$). Next we present the comparison between the methods of traditional factor analysis regarding the residual correlation matrix. The root mean square of the off-diagonal elements of the residual correlation matrix are presented in Table 5.10 for each case.

Table 5.10- Abortion Data: Root mean square off-diagonal residual correlation for the one-factor model

Correlation Matrix	MLFA	PFA	LJIFFY	PCA
Tetrachoric	0.097	0.076	0.082	0.087
Chambers	0.051	0.048	0.050	0.055
Mardia	0.058	0.053	0.057	0.063
Phi	0.153	0.127	0.150	0.162

It is seen in Table 5.10 that the best fit between the observed correlation matrix and that reproduced by the factor analysis methods is obtained using Chambers' coefficient.

The abortion data has been studied in the literature by different authors. Clogg and Sawyer (1981) analyse the data with models which assume that the latent variable is a discrete variable. In this study, the models used are response error models for assessing the scalability of a set of dichotomous items. They conclude that the abortion attitudes depart from the Guttman model in important ways, although this model provides an acceptable summary of the attitude in question. Some other models such as Lazarsfeld's latent distance model and others that are generalizations of the Guttman model were used. None of the response error models have fitted the data to an acceptable degree. The authors conclude that the assumption of a unique ordering of items for the entire population needs to be carefully examined.

A different class of models is considered under the factor analytic approach, which assumes that the latent variable(s) is continuous. Bartholomew's model used here starts from this assumption as well as the underlying variable model approach.

Muthén (1981) analyses the same data using his model (Muthén, 1978). In this model normality of the latent variables is assumed and the estimation is carried out using generalized least squares and information from the first and second order proportion. In this factor analytic formulation the two-factor model was considered. Tests of the number of factors for the one-factor model and for the two-factor model are provided. The results show that a single factor is clearly insufficient for this data set and that the two-factor is more suitable, giving a good overall fit. Muthén's two-factors solution is therefore rotated by the Promax method as in traditional factor analysis. The two factors were interpreted. The first factor, according to Muthén's solution is measured by items where the arguments in favour of abortion are for 'medical' reasons, (items 1,2) and the second factor is measured by items of social nature (items 4,5 and 6). The item 3 is related to both factors (see Muthén, 1981, p.206). Muthén also analyses the data on abortion attitudes studying the development of the factors over time (from the General Social Survey 1972-78). Our analysis is restricted to the 1975 survey.

Returning to our analysis, we had obtained two factors (by the default criterion of eigenvalues greater than one) when using the Phi-coefficients. The rotated factors were interpreted, the first factor showing high loadings for the three last items (abortion for social reasons) and the second factor high loadings for the first three items (abortion for medical reasons plus 'rape').

It is interesting to note that if the object of the analysis is to determine the dimensionality of the latent space and to interpret these dimensions, then the heuristic approach, using Phi-coefficients seems to produce more realistic results according to Muthén's analysis, although the tests of significance of the number of factors for the other correlation matrix cases had showed an evident rejection of the hypothesis of the one-factor model.

5.4 Andersen Data

We shall now analyse the Andersen data which consists of 5 items from a study of consumer complaining behaviour and it is taken from Andersen (1982). The original data are responses by 600 individuals on typical consumer situations to six items. The individuals were asked to state whether they, under the given circumstances, would complain or not. We shall use only 5 variables, the 5 last items from the original data set. We have decided to exclude the first item because it presents a very high percentage of positive responses (96%).

The cross-product ratio and the correlation coefficients for Andersen data are presented in Tables 5.11 and 5.12 respectively.

Table 5.11 - Andersen Data: Cross-product ratios

Item	1	2	3	4
2	2.91			
3	1.22	1.24		
4	2.56	2.70	2.58	
5	2.03	3.95	2.08	4.76

Table 5.12 - Andersen Data: Correlation coefficients

Tetrachoric coefficients					Chambers coefficients				
Item	1	2	3	4	Item	1	2	3	4
2	.34				2	.38			
3	.07	.08			3	.07	.08		
4	.32	.34	.34		4	.33	.35	.34	
5	.23	.43	.27	.54	5	.26	.47	.26	.52

Mardia coefficients					Phi-coefficients				
Item	1	2	3	4	Item	1	2	3	4
2	.34				2	.18			
3	.07	.07			3	.04	.04		
4	.30	.32	.30		4	.17	.19	.22	
5	.23	.43	.24	.48	5	.12	.22	.17	.34

CXCAAV

In Table 5.13 we consider the factor loading estimates according to the various factor analysis methods for the four correlation matrices.The reparameterized estimates for the logit model (MODFAC) and probit/probit model (FACONE) are presented and repeated in each case to facilitate comparisons.

Table 5.13 - Andersen Data: Factor loading estimates for different factor analysis methods (one-factor model)

CASE I - INPUT : TETRACHORIC CORRELATION						Bartholomew's methods	
Item	MLFA	PFA	ALPHA	LJIFFY	PCA	MODFAC	FACONE
1	40	41	39	36	56	41	40
2	53	53	53	45	67	54	52
3	36	34	27	30	46	33	37
4	74	74	85	57	80	75	76
5	73	72	75	57	79	74	71
CASE II - INPUT : CHAMBERS COEFFICIENTS							
Item	MLFA	PFA	ALPHA	LJIFFY	PCA	MODFAC	FACONE
1	45	45	42	39	59	41	40
2	58	58	56	49	70	54	52
3	34	32	26	29	45	33	37
4	70	72	84	56	78	75	76
5	74	73	76	57	79	74	71
CASE III - INPUT : MARDIA COEFFICIENTS							
Item	MLFA	PFA	ALPHA	LJIFFY	PCA	MODFAC	FACONE
1	41	43	40	36	57	41	40
2	56	56	53	46	69	54	52
3	32	30	25	26	43	33	37
4	67	68	80	52	76	75	76
5	71	70	73	53	77	74	71

CASE IV - INPUT : PHI-COEFFICIENTS

Item	MLFA	PFA	ALPHA	LJIFFY	PCA	MODFAC	FACONE
1	27	(*)	28	20	45	41	40
2	35		37	26	55	54	52
3	30		22	21	45	33	37
4	62		68	37	72	75	76
5	56		57	35	70	74	71

(*) No solution: the eigenvalues of the reduced correlation matrix are all too small.

We observe in Table 5.13 that there is a close agreement between the loadings of MLFA, PFA, MODFAC and FACONE, when we use the tetrachoric matrix. ALPHA, LJIFFY and PCA yielded different results. For the Chambers and Mardia coefficients the agreement between MLFA, PFA, MODFAC and FACONE is also good although slightly poorer than for the tetrachoric case. For the Phi-coefficient case, the estimates are considerably lower than Bartholomew's models estimates. The profile of the loadings is approximately the same for all cases. LJIFFY method underestimates the loadings compared with the other factor analysis methods for all correlation matrix cases.

The eigenvalues of the unaltered correlation matrix (with ones in the diagonal) are presented in Table 5.14 for Andersen Data.

Table 5.14 - Eigenvalues of the unaltered correlation matrix

Tetrachoric matrix	2.24	1.02	0.74	0.58	0.42
Chambers matrix	2.28	1.02	0.72	0.56	0.42
Mardia matrix	2.17	1.02	0.74	0.60	0.47
Phi-coefficients	1.71	1.03	0.85	0.76	0.64

Using the eigenvalue rule, two factors were extracted for all cases by the default criterion. With the two-factor solution we obtained improper solutions: a Heywood case (a loading equal to one for item 2) was observed for the tetrachoric matrix, Chambers matrix and Mardia matrix cases using the MLFA method. Using the Phi-coefficients we observed a very high loading for item 2 with the MLFA method ($\hat{\lambda}_{12}=0.93$) and no solution was obtained with the PFA method for the two-factors solution because the eigenvalues of the reduced matrix were all too small. The same occurred with the one-factor solution as shown in Table 5.13.

For the one-factor solution we now analyse the residual correlation matrices. The root mean square for each case is presented in Table 5.15.

Table 5.15 - Andersen Data: root mean square of the off-diagonal elements of the residual correlation matrices for each factor analysis method

Correlation Matrix	MLFA	PFA	LJIFFY	PCA
Tetrachoric	0.070	0.070	0.131	0.149
Chambers	0.074	0.073	0.133	0.134
Mardia	0.066	0.053	0.131	0.149
Phi	0.041	-	0.113	0.162

The fit between the observed correlation matrix and that reproduced by the factor analysis methods is reasonable for MLFA and PFA methods, showing that the one-factor model fits the data very well.

For the same data, the values of goodness-of-fit test for the MODFAC method was $\Lambda = 30.56$ for 21 d.f. ($p<0.100$) and for the FACONE method, $\Lambda = 19.42$ for 10 d.f. ($p<0.025$).

Andersen (1982) analysing this data set, applies three different models: the Rasch model, the Rasch model with a normal latent density and the latent class model. The Rasch model consists of the logistic latent trait model where the parameters α_i (the item discriminating power) are assumed to be equal for all i, and the individual location in the latent space is treated as a parameter. The second model is one of the latent variable models with the latent density normal with parameters μ and σ^2. The parameters α_i are again considered constant. Finally, the third model applied by Andersen is the latent class model where a discrete distribution is assumed for the latent variable. For a complete description of these models see Andersen (1982). Andersen concludes that the Rasch model as well as the latent class model with three latent classes give a relatively good fit to the observed data. The goodness-of-fit test for the three-latent class model, however, provides the best results. The model with a latent normal density was clearly rejected.

To make comparisons possible, we shall now present the goodness-of-fit test result for Bartholomew's one-factor logit model considering the original 6 variables. In this case a value $\Lambda = 238.46$ for 51 d.f. ($p<0.001$) was obtained, giving a poor fit when 6 items are considered. Excluding the first item, we obtain a much better fit. Comparing the probability levels, the fit of the logit model is still better than those found by Andersen (1982).

We conclude for the Andersen data that the heuristic approach is justified, giving equivalent results to those obtained by the logit model for practical purposes.

5.5 Lombard and Doering Data

The Lombard and Doering data relates to general knowledge of cancer and consists of 4 items. It is taken from Lombard and Doering (1947) and a sample of 1729 individuals was studied. This set was also presented in Bartholomew (1980). The 4 items, concerning sources of general knowledge of cancer, have two categories: (1) radio/no radio,; (2) newspaper/no newspaper; (3) solid reading/no solid reading; (4) lectures/no lectures.

In Table 5.16 and 5.17 we present the cross-product ratios and the correlation coefficients for Lombard data.

Table 5.16 - Lombard Data: Cross-product ratios

Item	1	2	3
2	2.81		
3	1.68	5.05	
4	2.30	2.46	2.45

Table 5.17 - Lombard Data: Correlation coefficients

Tetrachoric coefficients				Chambers coefficients			
Item	1	2	3	Item	1	2	3
2	.36			2	.36		
3	.19	.56		3	.19	.54	
4	.25	.26	.27	4	.30	.32	.32

Mardia coefficients				Phi-coefficients			
Item	1	2	3	Item	1	2	3
2	.33			2	.20		
3	.17	.50		3	.11	.38	
4	.27	.29	.29	4	.11	.11	.12

The eigenvalues of the unaltered correlation matrices are shown in Table 5.18.

Table 5.18 - Lombard Data: Eigenvalues of the unaltered correlation matrices

Tetrachoric matrix	1.97	0.86	0.77	0.44
Chambers matrix	2.03	0.84	0.71	0.42
Mardia matrix	1.94	0.85	0.74	0.47
Phi matrix	1.55	0.95	0.89	0.61

One factor was extracted by all factor analysis methods with the various correlation matrices used as input. The factor loadings for the one-factor model are presented in Table 5.19.

Table 5.19 - Lombard Data: Factor loadings obtained by different factor analysis methods and by Bartholomew's methods. One-factor model.

CASE I - INPUT: TETRACHORIC COEFFICIENTS						Bartholomew's methods	
Item	MLFA	PFA	ALPHA	LJIFFY	PCA	MODFAC	FACONE
1	40	41	44	34	60	41	39
2	85	83	79	57	82	84	88
3	65	64	58	52	76	65	63
4	35	39	45	33	59	41	36

CASE II - INPUT: CHAMBERS COEFFICIENTS							
Item	MLFA	PFA	ALPHA	LJIFFY	PCA	MODFAC	FACONE
1	43	44	45	37	62	41	39
2	81	79	78	55	81	84	88
3	65	63	56	51	74	65	63
4	44	48	55	39	66	41	36

CASE III - INPUT: MARDIA COEFFICIENTS

Item	MLFA	PFA	ALPHA	LJIFFY	PCA	MODFAC	FACONE
1	41	42	43	34	60	41	39
2	78	77	75	51	80	84	88
3	62	60	54	47	73	65	63
4	42	46	52	36	64	41	36

CASE IV - INPUT: PHI-COEFFICIENTS

Item	MLFA	PFA	ALPHA	LJIFFY	PCA	MODFAC	FACONE
1	27	(*)	33	20	51	41	39
2	72		61	36	77	84	88
3	52		47	33	72	65	63
4	18		25	15	40	41	36

(*) no solution: eigenvalues of the reduced matrix, all too small

Comparing the results in Table 5.19 we observe again close agreement between MLFA, PFA, MODFAC and FACONE for Case I and Case II. For all cases LJIFFY yields underestimates and PCA yields systematically higher loadings. ALPHA method yields different solutions from the MLFA and PFA method in this example, at least for the two last items. In the Mardia coefficients case, we obtained lower estimates for the second item, compared with Bartholomew's model estimates. Finally, the solution using Phi-coefficients underestimates the loadings comparing with MODFAC and FACONE or comparing with Case I, II and III.

The goodness-of-fit test for the logit model obtained by the MODFAC method, is $\Lambda = 19.30$ on 7 degrees of freedom ($p<0.005$) and for the probit/probit model (FACONE) we obtained $\Lambda = 11.34$ on 6 degrees of freedom ($p<0.075$). As we can see the fit of the probit/probit model is acceptable.

Table 5.20 - Lombard Data: Root mean square of the off-diagonal elements of the residual correlation matrices

Correlation matrix	MLFA	PFA	LJIFFY	PCA
Tetrachoric	0.059	0.055	0.149	0.178
Chambers	0.061	0.059	0.153	0.175
Mardia	0.063	0.056	0.153	0.179
Phi	0.031	---	0.214	0.221

In Table 5.20 we observe that the fit between the reproduced correlation matrix from the parameter estimates and the observed correlation matrix is reasonable for the MLFA and PFA method and it is poor for LJIFFY and PCA.

5.6 McHugh Data

The McHugh data set is taken from Everitt (1984, p.81) and consists of 4 items, which are four machine-design subtests given to 137 engineers. The items are dichotomized into positive (above the subtest mean) and negative (below the subtest mean). The data were also analysed by McHugh (1956) using a latent class model with two classes. Everitt applied the EM algorithm to the data using the same model. In this section we shall analyse the data using Bartholomew's models and traditional factor analysis methods applied to various correlation matrices appropriate to binary data.

The cross-product ratios and the correlation coefficients for the data are presented in Tables 5.21 and 5.22.

Table 5.21 - McHugh data: Cross-product ratios

Item	1	2	3
2	8.06		
3	3.57	2.90	
4	2.32	2.37	11.79

Table 5.22 - McHugh data: Correlation coefficients

Tetrachoric coefficients				Chambers coefficients			
Item	1	2	3	Item	1	2	3
2	0.68			2	0.65		
3	0.46	0.39		3	0.44	0.38	
4	0.32	0.33	0.76	4	0.30	0.31	0.72

Mardia coefficients				Phi-coefficients			
Item	1	2	3	Item	1	2	3
2	0.61			2	0.48		
3	0.40	0.34		3	0.31	0.26	
4	0.27	0.28	0.69	4	0.21	0.21	0.55

Table 5.23 - McHugh data: eigenvalues of the unaltered correlation matrix

Tetrachoric matrix	2.48	0.97	0.33	0.22
Chambers matrix	2.40	0.98	0.37	0.25
Mardia matrix	2.30	1.01	0.40	0.29
Phi-matrix	2.00	1.02	0.53	0.44

The eigenvalues of the unaltered correlation matrix for each case are presented in Table 5.23. Using the "default criterion" of number of eigenvalues greater than one, for deciding the number of factors, we have one factor for the tetrachoric and Chambers matrix cases and two factors for the Mardia and Phi-coefficient cases. In order to compare with Bartholomew's one factor logit model and probit/probit model we present the one-factor solution for all methods. The factor loadings estimates are presented and repeated in each correlation matrix input case for comparisons.

Table 5.24 - McHugh Data: Factor loadings estimates for various factor analysis methods. One-factor model.

CASE I - INPUT: TETRACHORIC COEFFICIENTS						Bartholomew's methods	
Item	MLFA	PFA	ALPHA	LJIFFY	PCA	MODFAC	FACONE
1	51	67	70	60	78	70	58
2	46	63	66	57	76	67	55
3	93	82	82	73	84	85	91
4	80	69	62	68	77	79	78
CASE II - INPUT: CHAMBERS COEFFICIENTS							
Item	MLFA	PFA	ALPHA	LJIFFY	PCA	MODFAC	FACONE
1	50	65	68	57	77	70	58
2	45	61	64	54	74	67	55
3	91	80	81	70	83	85	91
4	78	67	61	64	76	79	78

CASE III - INPUT: MARDIA COEFFICIENTS

Item	MLFA	PFA	ALPHA	LJIFFY	PCA	MODFAC	FACONE
1	47	62	65	53	74	70	58
2	43	58	62	51	72	67	55
3	89	78	78	66	82	85	91
4	76	65	58	61	74	79	78

CASE IV - INPUT: PHI-COEFFICIENTS

Item	MLFA	PFA	ALPHA	LJIFFY	PCA	MODFAC	FACONE
1	43	53	58	43	69	70	58
2	40	50	54	41	67	67	55
3	77	70	69	52	76	85	91
4	66	59	52	47	70	79	78

All methods yield a high factor loading for item 3, followed by the loading for item 4 and lower loadings for item 1 and 2. This pattern or profile of items is observed for all methods, but there is no close agreement between the absolute values of the loadings. The solutions for Case I, tetrachoric coefficients and Case II, Chambers coefficients are similar.

We now present the root mean square of the off-diagonal elements of the residual correlation matrix for each factor analysis method and input correlation matrix case in Table 5.25.

Table 5.25 - McHugh Data: RMS of the residual correlation matrices for factor analysis methods

Correlation matrix	MLFA	PFA	LJIFFY	PCA
Tetrachoric	0.188	0.158	0.182	0.207
Chambers	0.177	0.155	0.181	0.206
Mardia	0.173	0.152	0.185	0.210
Phi	0.133	0.124	0.178	0.208

The results in Table 5.25 show a poor fit of the reproduced correlation matrix with the observed correlation matrix for all methods. This would indicate that one more factor is needed in the model. The goodness of fit statistic obtained for the probit/probit model (FACONE program) was $\Lambda = 19.10$, 4 degrees of freedom and $p<0.001$, indicating that the fit to the one-factor model is poor.

5.7 Goodman Data

The Goodman data set is taken from Goodman (1978) and consists of four items of the Lazarsfeld-Stouffer questionnaire for noncommissioned officers on attitude toward the Army. The sample size is 1000, and the items are dichotomies: favorable (1) or unfavorable (0).

The cross-product ratios and the correlation coefficients for the Goodman data set are presented in Tables 5.26 and 5.27.

Table 5.26 - Goodman data: Cross-product ratios

Item	1	2	3
2	3.02		
3	3.52	2.76	
4	4.45	2.75	3.55

Table 5.27 - Goodman data: Correlation coefficients

Tetrachoric coefficients				Chambers coefficients			
Item	1	2	3	Item	1	2	3
2	.392			2	.388		
3	.424	.368		3	.434	.359	
4	.475	.360	.439	4	.503	.357	.437

Mardia coefficients				Phi-coefficients			
Item	1	2	3	Item	1	2	3
2	.354			2	.238		
3	.398	.327		3	.244	.229	
4	.464	.326	.401	4	.259	.218	.270

Table 5.28 - Goodman data: eigenvalues of the unaltered correlation matrix

Tetrachoric matrix	2.23	0.66	0.58	0.52
Chambers matrix	2.24	0.67	0.59	0.49
Mardia matrix	2.14	0.71	0.62	0.53
Phi-matrix	1.73	0.79	0.75	0.72

In Table 5.28 we present the eigenvalues of the unaltered correlation matrix for each matrix used as input to the MLFA method. One factor emerges from the analysis of the eigenvalues for all cases. The factor-loadings for the one-factor model are presented in Table 5.29. In order to compare the factor loadings obtained by the MLFA method with the reparameterized factor loadings for Bartholomew's probit/probit model (FACONE program), we present and repeat the FACONE loadings in each part of the Table 5.29.

Table 5.29 - Goodman data: factor loadings estimates for various factor analysis methods and the reparameterized factor loadings from FACONE method.

CASE I - INPUT : TETRACHORIC COEFFICIENTS					BARTHOLOMEW'S MODEL
Item	MLFA	PFA	LJIFFY	PCA	FACONE (probit model)
1	69	69	54	77	68
2	56	56	46	69	56
3	64	64	51	75	64
4	68	68	53	77	68
CASE II - INPUT : CHAMBERS COEFFICIENTS					
Item	MLFA	PFA	LJIFFY	PCA	FACONE
1	71	71	55	79	68
2	54	54	45	68	56
3	63	63	51	74	64
4	69	69	54	78	68

CASE III - INPUT : MARDIA COEFFICIENTS

Item	MLFA	PFA	LJIFFY	PCA	FACONE
1	68	68	51	77	68
2	51	52	42	66	56
3	60	60	47	73	64
4	67	66	51	76	68

CASE IV - INPUT : PHI-COEFFICIENTS

Item	MLFA	PFA	LJIFFY	PCA	FACONE
1	50	(*)	33	66	68
2	45		30	63	56
3	51		33	67	64
4	51		33	67	68

(*) eigenvalues of the reduced correlation matrix are all too small (no solution)

We observe, in Table 5.29 that for the Goodman data, similar factor loadings are obtained with the FACONE method (probit/probit model) and the MLFA method for the tetrachoric matrix input case. Similar results are also observed between FACONE and MLFA for Case II and III. The profile of the loadings using MLFA is similar for all cases. No solution was obtained with the PFA method for the phi correlation matrix case.

In Table 5.30 we present the root mean square of the residual correlation matrices for the factor analysis methods.

Table 5.30 - Goodman data: Root-mean-square of the residual correlation matrices for factor analysis methods.

Correlation Matrix	MLFA	PFA	LJIFFY	PCA
Tetrachoric	0.012	0.012	0.154	0.148
Chambers	0.012	0.012	0.153	0.147
Mardia	0.013	0.012	0.154	0.155
Phi	0.009	-	0.139	0.189

CXCAAX

For the MLFA method, we observe a very good fit for all correlation matrices, and also for PFA, but a bad fit for LJIFFY and PCA.

The goodness-of-fit statistic for the probit/probit model (FACONE) is $\Lambda = 6.46$ based on 7 degrees of freedom, indicating a very good fit of Bartholomew's model.

CXCAAX

5.8 Final Comments

We have presented in this chapter several examples of empirical comparisons between factor analysis models for binary data using the logit model (Bartholomew, 1980), the probit/probit model (Bartholomew, forthcoming) and the underlying variable model approach assuming:

a) first order marginals with normal distribution (tetrachoric coefficient and Chambers' correlation coefficient cases);

b) first order marginals with underlying continuous uniform distribution (Mardia's coefficient case) ;

c) the heuristic solution using the Phi-coefficients as input to traditional factor analysis methods. We have compared, for each correlation matrix case, various factor analysis methods available in BMDP and SPSS.

Concerning the factor analysis programs available in the statistical packages, ALPHA factoring method (SPSS) should be used only if the items included in the analysis are considered as a sample from a population of items. Otherwise, the MLFA or even the PFA methods give more consistent results. The method LJIFFY should not be used, as it produces systematically lower estimates for the factor loadings compared with the other methods and no consistent results considering the profile of the loadings. We have also included in our comparative study, the principal components analysis method, for illustrative purposes only. As expected it yields different results and, in general, higher component coefficients than the loadings for the same number of factors.

With the above considerations in mind, we shall concentrate our attention for comparisons between the factor loadings obtained by MLFA and PFA factor analysis methods versus the reparameterized parameter estimates for Bartholomew's models.

Before proceeding with our comparative conclusions between the methods, we shall point out some considerations about the correlation matrices used as input to the factor analysis methods. We have observed, for all six data sets used as examples in this chapter, that the phi-coefficients matrix yields factor loadings considerably lower than the correspondent estimates for the other correlation matrices cases. Theoretically, we had already concluded that the heuristic method using phi-coefficients, should not be applied, because there is no plausible bivariate distribution underlying this coefficient.

For all data sets, we observed close agreement between the MLFA results using tetrachoric coefficients and Bartholomew's probit/probit model (FACONE). The only exception was with Abortion Data, in which case there were some dissimilarities between the loadings. It should be considered, however, that according to other studies of the same data set, the two-factors model is more appropriate for describing the correlation structure.

We also observed, in general, similar results between the Chambers coefficients case and the tetrachoric coefficients case for all factor analysis methods. The results for the Mardia matrix input case are a few percent lower than for the Chambers matrix case following the tendency observed between the correlation coefficients, but the profile or pattern structure of the loadings are the same for Chambers and Mardia coefficients input cases.

Finally, we notice that there is no appropriate goodness-of-fit test for the traditional factor analysis methods applied to binary data. For this reason, we include in the analysis the root mean square criterion for the residual correlation matrix. This criterion, if not completely satisfactory either, at least identifies if there are considerable differences between the reproduced correlation matrix by the factor analysis model and the observed matrix. When no comparative study is necessary, a simple observation of the residual correlation matrix should give a first idea about the goodness-of-fit of the factor model. The chi-square test, on the other hand, is not appropriate for testing the dimensionality of the model when the assumptions of the normality and interval scale for the variables are not satisfied, as in the case of factor analysis for binary data. It should be noted that the test of the dimensionality of the latent space is a controversial area in factor analysis also for continuous variables. Seber (1984) points out that a satisfactory method for determining the number of factors does not seem to be available, and fictitious factors are all too readily generated. Seber's conclusion is based on a simulation study by Francis (1974), which shows that the goodness-of-fit test for the number of factors, even if the assumptions of the model are met, may not lead to the correct number of factors in the model.

In this chapter all data sets contain binary items and, for the three first input cases, we assume that there is some latent variable underlying each dichotomy, as explained before. Nevertheless, from the observed marginals, it is impossible to infer the form of the latent continuous distribution. Consequently, we may suppose different underlying distributions for these manifest responses, and the analysis of the data using the underlying variable model for binary data based on the C-type distribution is one of the possibilities.

CHAPTER 6 : NUMERICAL APPLICATIONS FOR POLYTOMOUS DATA

6.1 Introduction

In this chapter we shall present some numerical applications of factor analysis methods for polytomous data, using as input, correlation coefficients obtained as functions of the parameter of association of the C-type distribution, which is estimated by the maximum likelihood method presented in Chapter 3. Two contingency type correlation coefficients will be used: the coefficient $r_u(\psi)$ using Mardia's formula [expression (3.24)] and the coefficient $r_{0.74}(\psi)$ using Chambers' formula [expression (3.27)].

Four data sets with different numbers of variables (p), different sample sizes (n) and different number of categories of the variables are used. The data sets can be summarized as follows:

Data set	Name	No of variables	Sample size
1	Civil Service I Data	13	515
2	Boots Data	9	1181
3	Civil Service II Data	14	548
4	Greek Data	50	1784

The description and analysis of each data set will be presented separately. A more complete analysis will be presented for the first data set, in which case the data are analysed, first as binary data and then as polytomous data. Comparisons between factor analysis methods with Bartholomew's logit model (MODFAC) will be presented for the binary version of the Civil Service I data. For all data sets, factor analysis results using the underlying variable approach based on the C-type distribution, will be compared with traditional factor analysis methods using as input the product moment correlation coefficient. All data sets in this chapter contain polytomous ordered variables with values $X_i = 1,2,\ldots,C_i$, where C_i is the number of categories of the ith variable and the product moment correlation is evaluated from the raw data. The factor analysis method used in this chapter is the maximum likelihood factor analysis (MLFA) method from BMDP.

In Appendix II we present all correlation matrices used as input for each data set analysed in this chapter.

For all examples we shall use the following notation:

Let $R_u(\psi)$ denote the correlation matrix formed by the contingency-type coefficients using Mardia's formula $r_u(\psi)$ as function of the MLE of ψ, assuming an underlying C-type distribution, where

$$r_u(\psi) = \frac{\psi-1}{\psi-1} - \frac{2\psi \ln\psi}{(\psi-1)^2}$$

Let $R_{0.74}(\psi)$ denote the correlation matrix formed by the contingency-type coefficients using Chambers' formula $r_{0.74}(\psi)$ as function of the MLE of ψ, assuming an underlying C-type distribution, where

$$r_{0.74} = (\psi^{0.74} - 1)/(\psi^{0.74} + 1)$$

And finally, let R denote the product moment correlation matrix, as usual.

6.2 Civil Service I Data

The Civil Service I data set consists of 13 variables (items) related to aspects of performance obtained on 515 individuals. The variables are categorical, with ordered categories 1 to 6, where the rating 1 is related to outstanding performance and the rating 6 means unsatisfactory performance. Briefly, the variables are:

1. Foresight (anticipates problems and develops soltions in advance)
2. Penetration (gets straight to the roots of a problem)
3. Judgment (proposals or decisions are consistently sound)
4. Ability to produce constructive ideas
5. Expression on paper
6. Oral expression
7. Numerical ability
8. Relations with colleagues
9. Relations with others
10. Acceptance of responsibility
11. Management of staff
12. Reliability under pressure
13. Drive and determination

Using the computer program CROSSPSI described in Chapter 4, which evaluates the maximum likelihood estimate of the parameter of association ψ of the C-type distribution for each pair of variables of the data set, from the observed two-way cross tables, we obtained the two contingency-type correlation matrices: $R_u(\psi)$ and $R_{0.74}(\psi)$. We then analysed the two matrices using the MLFA (BMDP) program. We also obtained the factor analysis results for Civil Service I data using the Pearson product moment correlation coefficient. In Table 6.1

we summarize some information about the three correlation matrices used as input to the MLFA method.

Table 6.1 - Civil Service I Data: some numerical aspects of the correlation matrices

	Correlation matrix		
	$R_u(\psi)$	$R_{0.74}(\psi)$	R
$\underset{i,j}{\mathrm{Max}}\lvert r_{ij}\rvert$	0.908(V8×V9)	0.920(V8×V9)	0.755(V8×V9)
$\underset{i,j}{\mathrm{Min}}\lvert r_{ij}\rvert$	0.167(V7×V11)	0.185(V7×V11)	0.178(V5×V7)
determinant	0.7×10^{-3}	0.2×10^{-3}	7.8×10^{-3}
eigenvalues > 1	5.95 1.31 1.04	6.36 1.28 1.03	5.35 1.30 1.01

For the Civil Service I data we have three eigenvalues greater than one, as we see in Table 6.1, and the first eigenvalue of the correlation matrices is very high compared with the others. Using the "scree test" for the number of factors, the one-factor model should be chosen. Even though we analyse the data with three factors, two factors and finally with one factor. Some features of the results are summarized in Table 6.2. As in the last chapter, the chi-square test for the MLFA method will not be considered here because the test is designed for continuous normal variables.

Table 6.2 - Civil Service I Data : Comparison of some factor analysis features for different correlation matrices

(a) Three-factors solution; MLFA method

Correlation matrix		$R_u(\psi)$	$R_{0.74}(\psi)$	R
% var explained		63.9	66.7	59.0
Communality:	Min	0.167(V7)	0.192(V7)	0.137(V7)
	Max	1.000(V1;V8) (*)	1.000(V1;V8) (*)	0.866(V8)

(*) Improper solution: Heywood cases for variables 1 and 8.

(b) Two-factors solution; MLFA method

Correlation matrix		$R_u(\psi)$	$R_{0.74}(\psi)$	R
% var explained		55.8	58.8	51.3
Communality:	Min	0.150(V7)	0.169(V7)	0.137(V7)
	Max	1.000(V8) (*)	1.000(V8) (*)	0.856(V8)

(*) Improper solution: Heywood cases for variable 8.

(c) One-factor solution; MLFA method

Correlation matrix		$R_u(\psi)$	$R_{0.74}(\psi)$	R
% var explained		45.7	48.9	41.2
Communality:	Min	0.146(V7)	0.164(V7)	0.180(V7)
	Max	0.510(V11)	0.549(V11)	0.461(11)

CXCABA

We observe in Table 6.2 that improper solutions were obtained for the three and two factors solutions using the contingency type correlation matrices as input to the MLFA method. No improper solution was observed for the Pearson correlation matrix case. The occurrence of an improper solution for three and two factors suggest that these models are not appropriate and that too many factors are included in the solution. The one-factor model is clearly the correct factor analysis model for the Civil Service I data. The factor loadings for the one-factor solution are presented in Table 6.3 for each correlation matrix.

Table 6.3 - Civil Service I Data: factor loadings for the one-factor solution for the three correlation matrices used as input - MLFA method

VAR	$R_u(\psi)$	$R_{0.74}$	R
1	0.65	0.68	0.67
2	0.67	0.70	0.63
3	0.70	0.73	0.66
4	0.65	0.68	0.62
5	0.43	0.46	0.44
6	0.68	0.70	0.62
7	0.38	0.40	0.42
8	0.65	0.67	0.55
9	0.68	0.70	0.57
10	0.66	0.69	0.62
11	0.71	0.74	0.66
12	0.71	0.74	0.68
13	0.70	0.73	0.63

For a better comparison of the loadings, we present in Figure 6.1 the profile of the factor loadings for each correlation matrix case are presented.

The factor loadings for the input case using $R_{0.74}(\psi)$ are a few per cent higher than the loadings for the input case using the correlation matrix $R_u(\psi)$, but the profile is the same. A slightly different profile is obtained with the product moment correlation matrix case. As explained before, the one-factor model is chosen and we conclude that one dimension is enough for measuring the performance of the individuals analysed, and that the most important aspects of the performance are Judgement (Var3), Management of staff (Var11) and Reliability under pressure (Var12). At the other extreme, with lower factor loadings, appear the items: Expression on paper (Var5) and Numerical ability (Var7). All other items have factor loadings approximately equal to the three highest factor loadings, showing that all items are equally important, except items 5 and 7. The profile resulting from the product moment correlation matrix input case shows lower loadings for item 8 and 9 comparing with the two contingency type correlation matrix input cases.

As a comparative study we have also used other correlation matrices as input to factor analysis and although we are not showing the details of the solution here, we shall comment briefly on the results for Civil Service 1 data. We have used Goodman and Kruskal's gamma coefficient, Kendall's tau b and tau c. The results of factor analysis using the gamma coefficients are similar (although a few per cent higher) than those using the contingency type correlation matrices. We also obtained an improper solution (Heywood case) for the two and three factor models using the gamma coefficients as input. On the other hand factor analysis results using tau b and tau c coefficients as input yield profiles similar to the product moment matrix input case. The factor loadings for the input case using the tau c coefficients are considerably lower than the factor loadings for the product moment matrix. No improper solution was obtained for factor analysis using tau b and tau c coefficients. It is worthwhile to point out that had we analysed the data using only the product moment correlation coefficient, a three-factor solution (by the criterion of eigenvalues greater than one) would appear appropriate and attempts would then be made to interpret two more factors erroneously. On using contingency type coefficients, we obtained correctly the one-factor solution for the items measuring performance for the Civil Service I data.

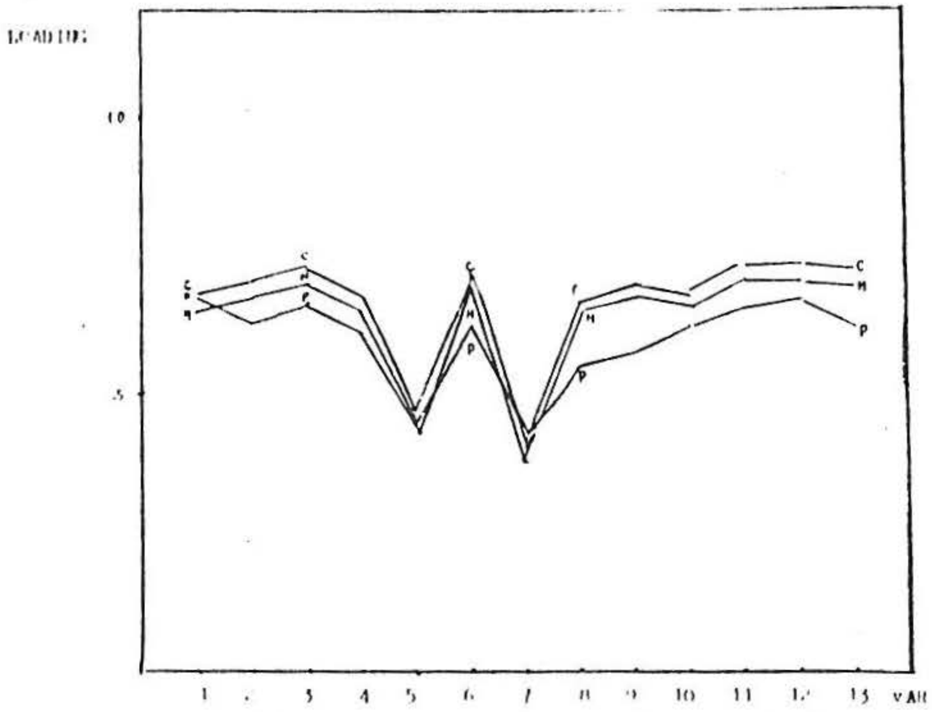


FIGURE 6.1 - PROFILE OF THE FACTOR LOADINGS. CIVIL SERVICE I DATA

CORRELATION MATRIX USED AS INPUT:
M - $R_u(\Psi)$
C - $R_{.74}(\Psi)$
P - R

Polytomous versus binary version of the Civil Service I data

Most of the available methods of factor analysis for categorical variables are designed for dealing with binary variables. One possible approach to treating categorical data is to dichotomize the variables and apply one of the methods available for binary data. In this section we dichotomize the Civil Service I data and apply Bartholomew's MODFAC method. We also evaluate the correlation coefficients as functions of the cross product ratios for the 2×2 tables formed by dichotomization of the variables and use Mardia's correlation coefficients, $r_u(\psi)$ as input to factor analysis methods. The factor analysis results for the binary version of the Civil Service I data are then compared with the factor analysis results for the polytomous version of the same data. The results for the polytomous version of the data were already shown in the last section. We repeat them here to facilitate comparisons. It is known that when we dichotomize the data we lose information. The question is "Are factor analysis results for the binary version of the data considerably different from the factor analysis results for the original polytomous data? The answer to this question cannot be conclusive with only one example. Our purpose in this section is to show the differences between both approaches for the Civil Service I data.

In Table 6.4 we present some information about the correlation matrix $R_u(\psi)$ for the dichotomous and polytomous version of the Civil Service I data.

Table 6.4 - Comparison between some aspects of the correlation matrices $R_u(\psi)$ for the dichotomous and polytomous version of the Civil Service I data.

Feature	Dichotomous data	Polytomous data
Correlation coefficients		
min $\|r_{ij}\|$	0.102 (V5×V9)	0.167 (V7×V11)
max $\|r_{ij}\|$	0.857 (V8×V9)	0.908 (V8×V9)
Determinant	1.56×10^{-3}	0.71×10^{-3}
Eigenvalues greater than one	5.66; 1.40	5.95; 1.31; 1.04

We first analysed the data using the "default criterion" of number of factors associated to eigenvalues greate r than one (Kaiser criterion). Therefore two factors were extracted for the binary version of the data and three factors for the polytomous version. Some aspects of the factor analysis results using four different factor analysis methods are presented in Table 6.5 using the Kaiser criterion for the number of factors.

Table 6.5 - Comparison between some aspects of the Factor Analysis methods using as input the correlation matrix $R_u(\psi)$ for the dichotomous and polytomous version of the Civil Service I data.

	Dichotomous data				Polytomous			
FA method	MLFA	RAO	PA2	PFA	MLFA	RAO	PA2	PFA
No of factors by the default criterion	2	2	2	2	3	3	3	3
% of variance explained by the factors	48.7	54.3	54.3	-	57.1	63.9	63.9	-
Communality								
min	.138(V7)	.136(V7)	.138(V7)	-	.167(V7)	.164(V7)	.157(V7)	-
max	.975(V8)	.890(V8)	.982(V8)	-	(V1) 1.000(V8)	.904(V8)	.999(V8)	-
No of iterations	3	>25	4	>25	8	>25	16	>25
Comment		(*)	(**)	(***)	(****)	(*)	(**)	(***)

(*) More than 25 iterations is required by RAO method (no final solution)
(**) PA2 method terminates when communality of one or more variable exceed one
(***) Communalities failed to converge. Program stops
(****) Communality of one or more variables exceeded one.

Table 6.5 is useful for illustrating the different approaches of the various factor analysis methods from BMPD (MLFA and PFA) and SPSS (RAO and PA2) for dealing with Heywood cases. As we can see in Table 6.5, improper solutions were obtained for either version of the data with the PA2 and PFA methods. The RAO factoring method stops at iteration 25 and, if no other instruction is given to the program allowing a greater number of iterations a misinterpretation of the results is easily obtained. The MLFA factoring method shows a Heywood case for variables 1 and 8, as we have already seen in the last section for the polytomous version of data, showing that th e three factors-solution is not appropriate for this data set. For the dichotomous version of the data, we obtained apparently a proper solution, although a very high communality is observed for variable 8.

When too many factors have been included in a solution, this may cause improper solutions (see Chapter 7). Occurring improper solutions we should examine more carefully the eigenvalues, their differences often provide an excellent evidence of the correct number of factors in the model, as was pointed out in Section 4.3. From Table 6.4 we observe that there is a substantial difference between the first and second eigenvalues for both version of the data. This fact suggests that the number of factor to be included in the analysis is only one. We therefore reanalyse the data with a one-factor model. In Table 6.6 we present some aspects of the factor analysis results for the one-factor model.

Table 6.6 – Comparison between some aspects of the Factor Analysis results using as input the correlation matrix $R_u(\psi)$ for the dichotomous and polytomous version of the Civil Service I data. One-factor model.

	Dichotomous data				Polytomous data			
FA method	MLFA	RAO	PA2	PFA	MLFA	RAO	PA2	PFA
% of var explained	39.2	43.6	43.6	78.4	41.48	45.7	45.7	78.5
Communality:								
min	.115(V5)	.115(V5)	.120(V5)	.120(V5)	.146(V7)	.146(V7)	.147(V7)	.147(V
max	.582(V11)	.582(V11)	.570(V11)	.570(V11)	.510(V11)	.510(V11)	.512(V12)	.513(
No of iterations	6	15	5	5	5	10	5	5

We now present the factor loadings of the one-factor model for both versions of the data. In Table 6.7 we also present the reparameterized factor loadings of the one-factor logit model (MODFAC) for binary data.

Table 6.7 - Factor loadings obtained by different factor analysis methods using as input the correlation matrix $R_u(\psi)$ for the dichotomous and polytomous version of Civil Service I data and factor loadings obtained by Bartholomew's MODFAC method for binary data.

Item	Dichotomous data					Polytomous data			
	MODFAC	MLFA	RAO	PA2	PFA	MLFA	RAO	PA2	PFA
1	70	68	68	67	67	65	65	65	65
2	67	63	63	65	65	67	67	67	67
3	70	70	70	71	70	70	70	70	70
4	59	57	57	58	58	65	65	65	65
5	35	34	34	35	35	43	43	43	43
6	68	65	65	66	66	68	68	68	68
7	36	36	36	36	37	38	38	38	38
8	85	66	66	65	65	66	65	65	65
9	82	64	64	63	63	68	68	67	67
10	64	63	63	63	63	66	66	65	65
11	78	76	76	76	76	71	71	71	71
12	69	67	67	68	68	71	71	72	72
13	69	68	68	68	68	70	70	70	70

(decimal point is omitted)

The analysis of the factor loadings of Table 4.1 shows that the agreement between the methods is very close. The greatest difference between the MODFAC method and the others appears for the loadings of the items 8 and 9. Comparing the dichotomous and polytomous version of data we find, in general, greater factor loadings for polytomous data with the exception of items 1 and 11. The loadings for items 3 and 8 are approximately the same. For a visual comparison of the loadings for dichotomous and polytomous data, we present in Figure 6.2, the profile of the factor loadings for the MLFA method and for the MODFAC method. The other methods are not presented because of the similarity of the results with the MLFA method in this case.

CXKAAK

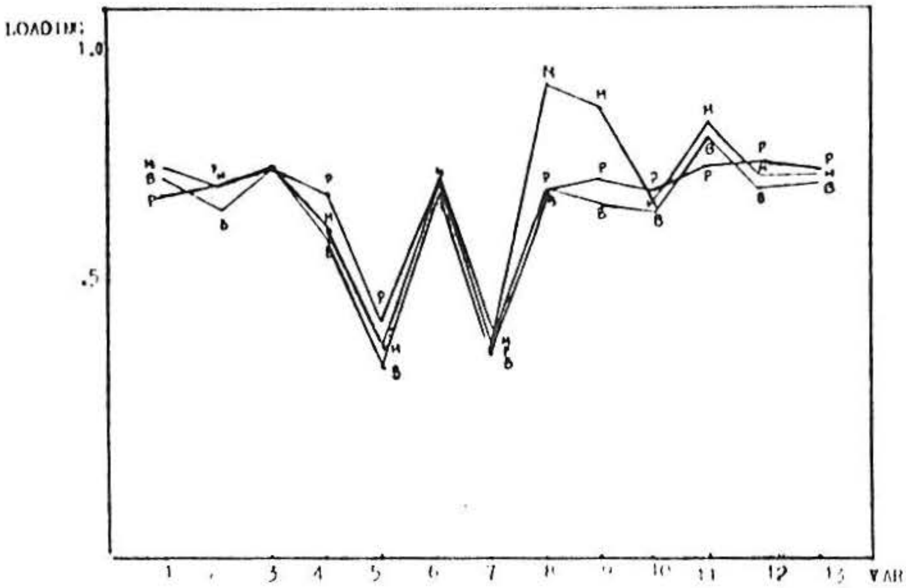


FIGURE 6.2 - PROFILE OF THE FACTOR LOADINGS FOR MODFAC AND MLFA FACTOR ANALYSIS SOLUTIONS USING BINARY AND POLYTOMOUS VERSIONS OF CIVIL SERVICE I DATA

M - MODFAC METHOD, BINARY DATA

B - MLFA METHOD, BINARY DATA

P - MLFA METHOD, POLYTOMOUS DATA

Comparing, finally, the factor analysis results for the dichotomous and polytomous version of the Civil Service 1 data it is relevant to note that the greatest difference between Bartholomew's method (MODFAC) and the traditional factor analysis is observed for items with high factor loadings (items 8 and 9). The comparison between the results for dichotomous and polytomous version of data for the one-factor model shows only small differences, but when using the Kaiser criterion for the choice of number of factors, different number of factors resulted for each version.

6.3 Boots Data

The Boots data set consists of nine items (selected from a total of 50 items) from a biographical background survey and it is part of a graduate selection improvement project carried out by the Boots Company. The sample size is 1181 individuals and the items have varying numbers of categories, as summarized in Table 6.8.

Table 6.8 - Number of categories of each item

Boots Data

Item	V1	V2	V3	V4	V5	V6	V7	V8	V9
No of categories	6	5	5	5	2	2	5	7	6

A brief description of variables is as follows.

V1 - Extra curricular activities I (Summer school, professional conference, etc.)
V2 - Number of part-time jobs during academic term
V3 - Extra curricular activities II (sport, debates, choral, etc)
V4 - Participation in clubs or societies
V5 - Elected office in a club or society
V6 - Prize for a competition
V7 - Organization of activities (sports competition, charity campaign, etc.)
V8 - Kind of holiday taken
V9 - Number of countries visited

Contingency-type correlation coefficients were used for each pair of variables from the Boots data set. The global cross product ratio for R×C contingency tables was estimated using and the program CROSSPSI. Two contingency type correlation matrices: $R_u(\psi)$ and $R_{0.74}(\psi)$ were used as before. These matrices were factor analysed using the maximum likelihood factor analysis method (MLFA-BMDP). Comparisons of the factor analysis results using the new method will be made with the factor analysis results using the Pearson moment product correlation matrix (R) as input.

In Table 6.9 we present some relevant aspects of the correlation matrices used as input to the factor analysis method.

Table 6.9 - Boots data: Extreme values of the correlation coefficients and eigenvalues greater than one for each correlation matrix.

	Correlation matrix		
	$R_u(\psi)$	$R_{0.74}(\psi)$	R
$\max_{i,j}\|r_{ij}\|$	0.312(V3×V7)	0.343(V3×V7)	0.254(V3×V7)
$\min_{i,j}\|r_{ij}\|$	0.044(V6×V8)	0.049(V6×V8)	0.056(V5×V8)
eigenvalues > 1	2.275 1.144	2.407 1.157	2.040 1.120

In this case the numbers of factors in the factor analysis model is two, using either the Kaiser criterion (number of eigenvalues greater then one) or the scree test. In Table 6.10 we present some features of factor analysis outputs for the various correlation coefficients.

Table 6.10 - Boots data: Comparison of some aspects of factor analysis outputs for the various correlation matrices used as input to the MLFA method.

	Correlation Matrices		
	$R_u(\psi)$	$R_{0.74}(\psi)$	R
% Var explained	22.3%	24.6%	17.4%
Communality: min	0.064	0.071	0.081
max	0.507	0.565	0.278
No of iterations	7	8	5

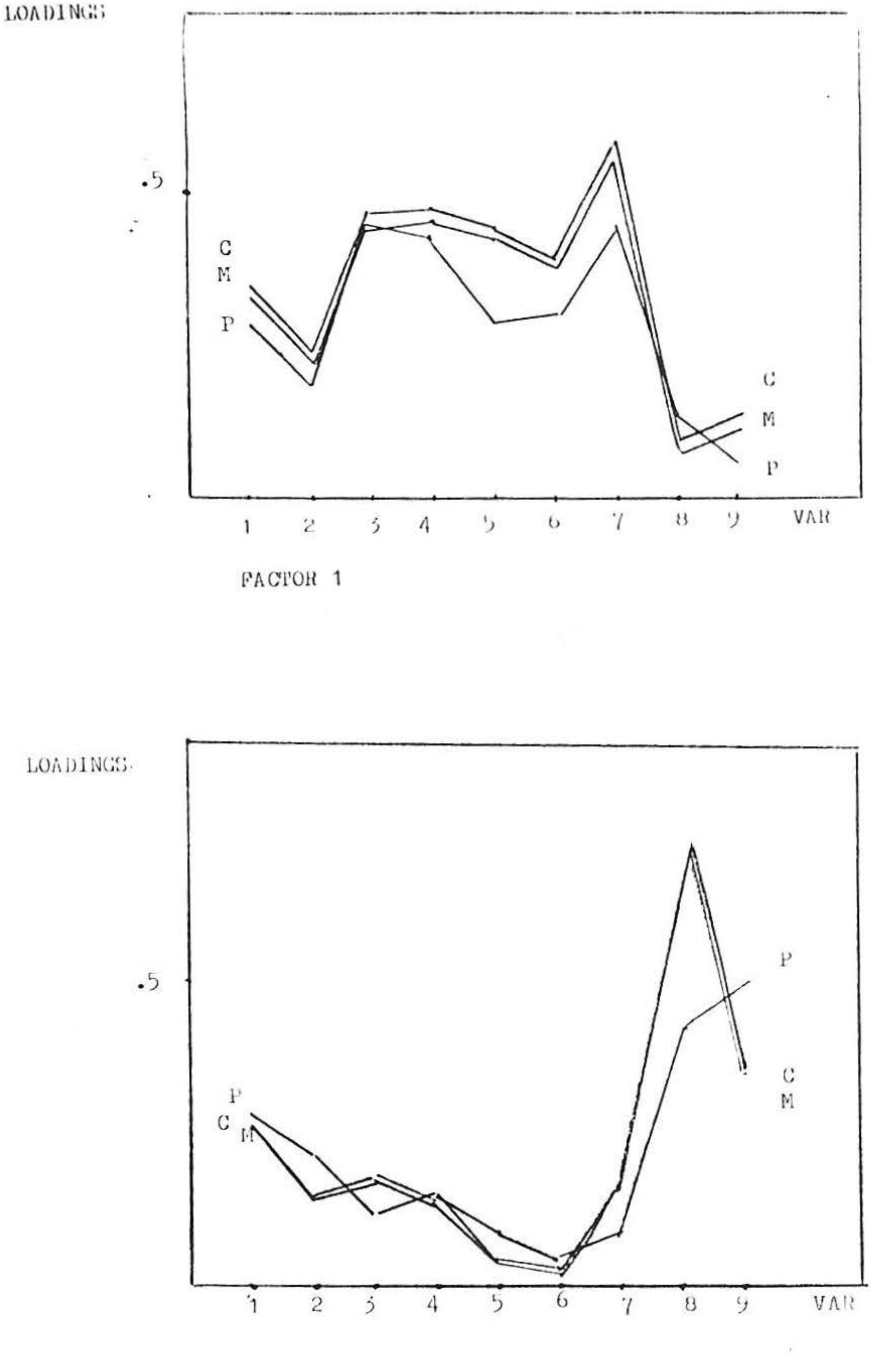
The Varimax rotated factor loadings for the two factors for each correlation matrix used as input are presented in Table 6.11. The Varimax rotation method provides, in this case, a factor pattern easily interpretable.

Table 6.11 - Boots Data: Varimax rotated factor loadings; two-factor model, MLFA method.

	Correlation matrix used as input					
	$R_u(\psi)$		$R_{0.74}(\psi)$		R	
Item	F_1	F_2	F_1	F_2	F_1	F_2
1	.33	.25	.35	.26	.28	.28
2	.21	.14	.22	.15	.18	.22
3	.44	.18	.46	.19	.46	.12
4	.45	.15	.47	.15	.43	.15
5	.43	.05	.45	.05	.29	.10
6	.37	.03	.39	.03	.30	.06
7	.56	.17	.58	.18	.44	.11
8	.08	.71	.09	.75	.14	.45
9	.12	.37	.12	.39	.06	.52

From Table 6.11 we see that there are differences between the factor loadings estimates for each correlation matrix used as input, but all three cases seem to lead to the same interpretation of factors. The factor loadings for the matrix $R_u(\psi)$ are similar to the loadings for the matrix $R_{0.74}(\psi)$, although a few per cent lower. In figure 6.3 we present the profile of the rotated factor loadings for each factor separately. Although the loadings are not equal in magnitude, if the oscillatory movement of the "curve" is similar, we can say that there is an equivalence in the results, in the sense that, the order of the loadings is the same. That is true for the results obtained for the two contingency type matrix cases, but a different movement is observed for the loadings of the Pearson correlation matrix case.

Table 6.12 shows the items with the highest factor loadings for each factor. The order of the loadings for the contingency type matrix cases are the same and are presented together. A slightly different order was observed for the Pearson case. The interpretation of the factors is the same for the three cases. The first factor could be interpreted as a factor of "Extra curricular activities" and the second factor is related to "Travel".



FACTOR 1

FACTOR 2

FIGURE 6.3 - PROFILE OF THE ROTATED FACTOR LOADINGS.

BOOTS DATA

$M - R_u(\Psi)$ $\qquad$ $P - R$

$C - R_{.74}(\Psi)$

Table 6.12 - Boots data: Items with the highest factor loadings for each factor and each input case. MLFA method - Varimax rotation.

Contingency type matrices		Product-moment matrix	
Item	Factor I	Item	Factor 1
7	Organization of activities	3	Extra curricular activities
4	Participation in clubs or societies	7	Organization of activities
3	Extra curricular activities II	4	Participation in clubs or societies
5	Elected office in club or society	6	Prize for a competition
6	Prize for a competition	5	Elected office in club or society
1	Extra curricular activities I	1	Extra curricular activities I
	Factor 2		Factor 2
8	Kind of holiday	9	Number of countries visited
9	Number of countries visited	8	Kind of holiday

As a comparative and complementary analysis we have compared the factor analysis results for the Boots data using the three correlation matrices: $R_u(\psi)$, $R_{0.74}(\psi)$ and R with other factor analysis results using different measures of association as input. We have analysed the data using gamma, tau b and tau c coefficients. Although the results are not presented in detail, a comment on the results is now presented.

The profile of the rotated factor loadings are very similar for all coefficients, although the loadings for the gamma coefficient case are systematically higher than the others and the loadings for tau b, systematically lower for all items. An improper solution was observed for the two-factors model when using the tau c coefficients as input to the MLFA method. For all correlation matrices used as input we obtained two eigenvalues greater than one and the same interpretation for the rotated factors, with the exception of the tau c case (improper solution) and of the Pearson product moment correlation matrix, in which case a slightly different profile was observed as shown in Figure 6.3. For the Pearson case we observe a considerable difference in the loadings for items 5 and 6 in Factor 1. Analysing the contents of these items, it seems to us that the profile of the first factor, for the two contingency type matrix cases is more "reasonable".

6.4 Civil Service II Data

The third data set to be analysed in this chapter is also related to aspects of performance. The items are similar to the Civil Service I data, but now we have 19 items and a different sample with 548 individuals. The variables are categorical with categories 1 to 6, where rating 1 represents outstanding performance and 6 stands for unsatisfactory performance. The items are described as follows:

1 - Foresight
2 - Penetration
3 - Judgment
4 - Constructive ideas
5 - Expression on paper
6 - Oral expression
7 - Numerical ability
8 - Relations with colleagues
9 - Relations with public
10 - Relations with official and other bodies
11 - Responsibility
12 - Use of staff and other resources
13 - Management of staff
14 - Reliability
15 - Drive
16 - Ability to organise own work
17 - Knowledge of work on which engaged
18 - Knowledge of own unit generally
19 - Knowledge of Ministry/Board

The same method of analysis is used for the Civil Service II data. We shall compare the factor analysis results using the contingency type correlation matrices: $R_u(\psi)$ and $R_{0.74}(\psi)$ as input with the traditional factor analysis results using the product moment correlation coefficients. For obtaining the contingency type correlation matrices for polytomous data we have used the program CROSSPSI. The factor analysis method used in this section is the maximum likelihood method (MLFA).

Table 6.13 shows some relevant information about the three correlation matrices for the Civil Service II data.

CXKAAL

Table 6.13 - Extreme values of the correlation coefficients and eigenvalues greater than one for three correlation matrices. Civil Service II data.

	Correlation Matrix		
	$R_u(\phi)$	$R_{0.75}(\phi)$	R
max $\|r_{ij}\|$	0.817 (V9×V10)	0.842 (V9×V10)	0.626 (V9×V10)
min $\|r_{ij}\|$	0.054 (V7×V19)	0.060 (V7×V19)	0.046 (V7×V19)
eigenvalues > 1	8.05 1.55 1.31 1.11 1.05	8.65 1.57 1.30 1.11 1.03	6.75 1.43 1.26 1.07 1.03

Using the Kaiser criterion, five factors were extracted in the first analysis. Improper solutions - three Heywood cases for the contingency type matrix cases and one Heywood case for Pearson matrix case - were observed. Looking at the magnitude of the eigenvalues, the scree test shows clearly that the one-factor model should be chosen for the Civil Service II data. The first eigenvalue, as shown in Table 6.13, is considerably greater than the others. Therefore we reanalyse the data using a one-factor model. The main aspects of the MLFA outputs for the three matrices used as input are shown in Table 6.14.

Table 6.14 - Civil Service II data: Main features of the factor analysis results for three correlation matrices used as input to MLFA.

	Correlation Matrix		
	$R_u(\psi)$	$R_{0.74}(\psi)$	R
No of factors	1	1	1
% Var. explained	39.35	42.65	32.16
Communality			
max	0.594 (V1)	0.636 (V1)	0.500 (V1)
min	0.111 (V19)	0.124 (V19)	0.077 (V19)
Iterations	6	6	4

The factor loadings for the one-factor model are presented in Table 6.15 for each input case. There is in this case a great similarity between the factor analysis results for the three correlation matrices. As in the previous examples, factor analysis results using the contingency type matrices are equivalent, although the loadings for the $R_{0.74}(\psi)$ case are a few per cent higher than for the $R_u(\psi)$ case. In this example, the Pearson case results are also similar to the contingency type correlation matrix input cases, but lower factor loadings are observed. Also a small difference in the order of the loadings is observed for the Pearson matrix case.

Table 6.15 - Civil Service II data: Factor loadings for one-factor model for the three correlation matrices, MLFA method.

Item	Correlation Matrix Item		
	$R_u(\psi)$	$R_{0.74}(\psi)$	R
1	.77	.80	.71
2	.70	.73	.65
3	.74	.77	.66
4	.61	.63	.55
5	.44	.47	.42
6	.57	.60	.51
7	.43	.45	.38
8	.58	.61	.51
9	.64	.66	.53
10	.60	.62	.50
11	.69	.72	.66
12	.62	.65	.56
13	.69	.72	.64
14	.75	.77	.69
15	.72	.74	.67
16	.73	.76	.68
17	.63	.66	.56
18	.42	.44	.38
19	.33	.35	.28

The interpretation of the results leads us to conclude that we have one general factor of performance, with high loads for items: Foresight, Reliability, Judgment, Ability to organize own work, Drive and determination, Penetration, Responsibility and Management of staff. At the other extreme we have the following items with low weight: Numerical ability, Expression on paper, Knowledge of own unit generally, Knowledge of Ministry/Board. It is interesting to note that there are some similarities between the analysis for the Civil Service II data and the Civil Service I data, although a more complete set of items was analysed for Civil Service II.

In Figure 6.4 we present the profile of the loadings for the Civil Service II data.

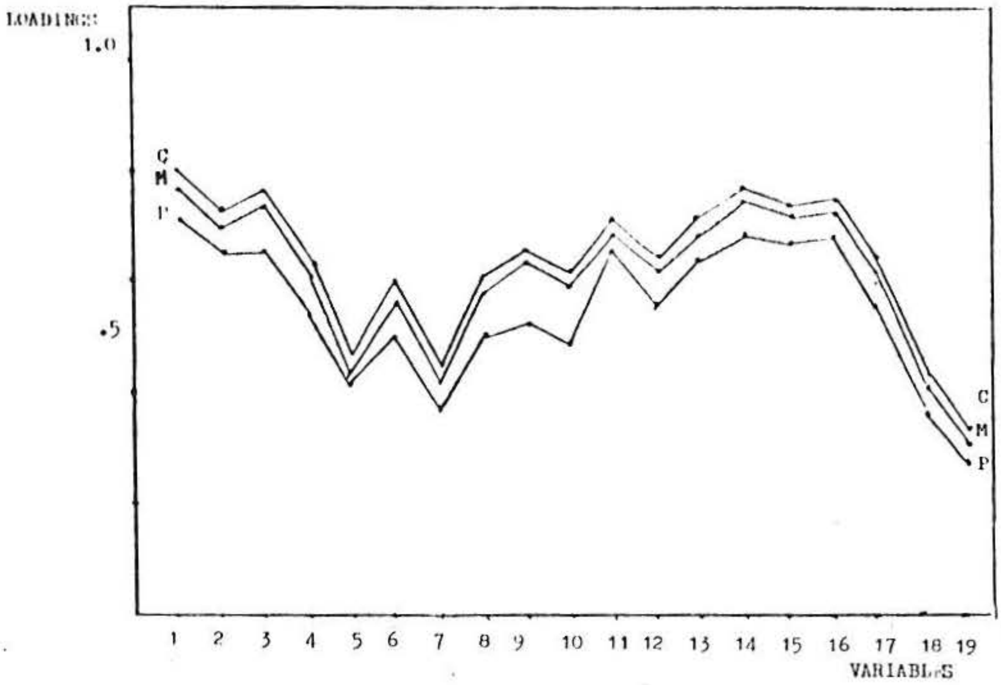


FIGURE 6.4 - PROFILE OF THE FACTOR LOADINGS. ONE FACTOR SOLUTION

CIVIL SERVICE II DATA
CORRELATION MATRIX:

$M - R_u(\Psi')$

$C - R_{.74}(\Psi')$

$P - R$

6.5 Greek Data

The last data set to be analysed in this chapter consists of the Greek translation of a children's version of the "C-Scale", a scale for measuring social attitudes. The scale consists of 50 items, each requiring no/neutral/yes responses according to agreement with (or approval of, or belief in, as appropriate) the item. The Conservative Scale (C-Scale) for adults was developed by Wilson and Petterson (1968) and the children's version was constructed by Insel and Wilson (1971).

The sample size for the Greek data is 1784. The sample consists of 14 years-old boys and girls from Greek high schools. The 50 items of the Children's Scale of Social Attitudes administered to Greek children are reproduced in Table 6.16. The C-Scale for adults has been reported in the literature as a robust scale. According to Joe (1984), the factor structure that emerged in the original studies in England is highly similar to that reported by studies employing samples from Australia, South Africa, Korea, New Zealand and United States (see also Wilson, 1973).

Table 6.16 - The Children's Scale of Social Attitudes administered to Greek children

Which of the following do you prefer (like, agree with or believe in)?
Circle Yes or No. If you can't decide, circle (?).

1. Death to thieves
2. Space travel
3. School uniforms
4. Bikinis
5. Sunday School
6. Men with beard
7. Honouring the flag
8. Modern Art
9. Obedience
10. Comics
11. Miracles
12. Dancing
13. Military service
14. Mixed schools
15. Ten commandments
16. Russians
17. White supremacy
18. Kissing
19. Beating children
20. Blasphemy
21. Servants
22. Short skirts
23. Saving money
24. Playing pranks
25. Police
26. Computers
27. Prayers
28. Going barefoot
29. Royalty
30. Female doctors
31. Science
32. Beer
33. Atomic bombs
34. Nude bathing
35. Church
36. Chinese food
37. Politeness
38. Telling fibs
39. Corporal punishment for criminals
40. Germans
41. Strict rules
42. Rock-and-roll or the Beatles
43. Death to our enemies
44. Laughing in class
45. Hunting
46. Divorce
47. Confession
48. Blacks
49. Bible reading (Religious books)
50. Playing doctors

Using the CROSSPSI program we obtained the contingency type correlation matrices - $R_u(\psi)$ and $R_{0.74}(\psi)$. The factor analysis results using these two matrices as input will be compared with the results for the Pearson product moment correlation matrix. Some aspects of the matrices are summarized in Table 6.17.

Table 6.17 - Extreme values of the correlation matrices and eigenvalues greater than one. Greek data - 50 variables

	Correlation Matrix		
	$R_u(\psi)$	$R_{0.74}(\psi)$	R
Comment	(*)	(*)	-
min $\|r_{ij}\|$	0.000	0.000	0.001
max $\|r_{ij}\|$	0.807(V35×V27)	0.834(V35×V27)	0.626(V35×V27)
eigenvalues > 1	8.27	8.91	5.31
	4.03	4.32	2.82
	3.32	3.54	2.69
	2.17	2.28	1.68
	2.06	2.16	1.61
	1.65	1.70	1.46
	1.38	1.41	1.26
	1.34	1.37	1.24
	1.16	1.17	1.16
	1.15	1.16	1.12
	1.13	1.14	1.09
	1.07	1.08	1.05
	1.05	1.05	1.02
	-	-	1.01

(*) Matrix not positive semi-definite

For the Greek data, the correlation matrices using $r_u(\psi)$ and $R_{0.74}(\psi)$ matrices are not positive semi-definite. The MLFA program prints a message and no iterative solution is carried out. The solution printed in the output corresponds to the solution without iteration. The main features of the factor analysis solutions for the three correlation matrices used as input to the MLFA program are summarized in Table 6.18.

Table 6.18 - Greek data: comparison of some aspects of factor analysis results for different correlation matrices. MLFA method.

	Correlation Matrix		
	$R_u(\psi)$	$R_{0.74}(\psi)$	R
No. of factors	13	13	14
% Var explained	59.5	62.6	29.2
Communality: min	0.394 (V32)	0.421 (V32)	0.103 (V50)
max	0.916 (V37)	0.966 (V37	0.675 (V35)
No. of iterations	0	0	6

The number of factors in Table 6.18 corresponds to the number of eigenvalues greater than one. The scree plot of the eigenvalues shows that three factors should be chosen (see Table 6.17). However, as all analysis of the C-scale reported in the literature are carried out with the number of factors equal to the number of latent roots greater than one, we have decided to maintain this criterion in our analysis, for eventual comparisons.

Comparing the results in Table 6.18 it is interesting to note the great difference between the cumulative percentage of variance in the data space associated with the respective number of factors for the Pearson case. In this case 14 factors account for only 29.2 per cent of the total variance. For the MLFA method, total variance is defined as the sum of the positive eigenvalues of the matrix. This fact - a lowest percentage of the variance explained by the factors when using the product moment correlation matrix - was also observed in the other examples of this chapter. We also observe that the magnitude of the product moment correlation for polytomous data, is in general, lower than for the contingency type correlation matrices.

In Table 6.19 we present the sorted unrotated factor loadings (and respective items) for the three correlation matrices.

Table 6.19 - Sorted unrotated factor loadings for the first factor - MLFA method. Greek data.

Correlation Matrix		$R_u(\psi)$	$R_{0.74}(\psi)$		R
Rank	Item	Loading	Loading	Item	Loading
1	Church	.82	.84	Church	.75
2	Prayers	.79	.81	Prayers	.70
3	Bible reading	.76	.78	Bible reading	.70
4	Politeness	.71	.74	Sunday School	.62
5	Ten commandments	.71	.73	Ten commandments	.61
6	Sunday school	.69	.71	Confession	.58
7	Saving money	.66	.68	Short skirts	-.38
8	Confession	.64	.66	Laughing in class	-.37
9	Blasphemy	-.62	-.64	Obedience	.36
10	Laughing in class	-.54	-.57	Saving money	.36
11	Saluting the flag	.54	.56	Playing pranks	-.36
12	Obedience	.54	.55	School uniforms	.35
13	Short skirts	-.50	-.53	Nude bathing	-.35
14	Playing pranks	-.50	-.52	Politeness	.32
15	Nude bathing	-.50	-.52	Blasphemy	-.30

In Table 6.19 only the fifteen highest factor loadings are presented. There is no difference in the order of the items for the contingency type matrix cases. Although the order of the factor loadings for the Pearson matrix case is not the same, almost all the fifteen items are the same from the set of 50 items, with exception of the item "School uniforms" that do not appear among the first items for the contingency type matrix cases. To provide a better interpretation of the factors, the Varimax rotated factor loadings for the first three factors are presented in Table 6.20. We present only the highest loadings in decreasing order (sorted rotated factors). The fourth factor for the three cases consisted of relatively low loadings, and could not be interpreted.

Comparing the three first rotated factors, for the different correlation matrices we conclude that the three methods lead to the same interpretation of the factors, although different factor loadings were observed - the loadings for the Pearson case are systematically lower. There is also some difference in the order of the items. The interpretation using the contingency type matrix columns in Table 6.20 is rather clearer than the Pearson case. Factor 1 has a definite religious theme. Factor 2 has a predominantly sexual theme and Factor 3 is related to punitiveness.

The three rotated factors and their themes are similar to those reported in previous studies of the Children' Scale of Social Attitudes (e.g. Nias, 1973) and also are similar to those reported for the Adults' C scale (e.g. Joe, 1984). In the study presented by Nias the scale was administered to 217 boys and 224 girls at an English comprehensive school. The analysis was carried out separately for boys and girls and the method applied was Principal Component Analysis followed by a Promax rotation. The first four factors were identified as relating to religion, ethnocentrism, punitiveness and sex. For the Greek data the factor related to "ethnocentrism" could not be identified among the first four factors, but the factors "religion", "sex" and "punitiveness" are basically the same though the order is not the same. The correlation coefficiennt used in other studies was the Pearson coefficient. Furthermore, looking at the items with the highest loadings for the first unrotated factor, presented in

Table 6.20 - Greek Data: the highest loadings for the first three rotated factors (decreasing order) for the three correlation matrices used as input to the MLFA method.

		Correlation Coefficient			
$R_u(\psi)$		$R_{0.74}(\psi)$		R	
Item	Loading	Item	Loading	Item	Loading
Church	.88	Church	.89	Church	.77
Bible reading	.82	Bible reading	.83	Bible reading	.72
Prayers	.82	Prayers	.83	Prayers	.66
Ten commandments	.76	Ten commandments	.78	Ten commandments	.59
Confession	.76	Confession	.78	Confession	.57
Sunday school	.73	Sunday school	.75	Sunday school	.54
Politeness	.61	Politeness	.67	Miracles	.27
Honouring the flag	.61	Honouring the flag	.66	Saving money	.22
Saving money	.49	Saving money	.54	Honouring the flag	.22
Obedience	.41	Obedience	.46	Obedience	.21
Miracles	.36	Blasphemy	-.40	Politenesss	.19
Blasphemy	-.36	School uniforms	.39	Police	.19
School uniforms	.35	Miracles	.38	Blasphemy	-.17
Short skirts	.78	Short skirts	.79	Short skirts	.60
Kissing	.76	Kissing	.79	Kissing	.55
Bikinis	.60	Bikinis	.62	Mixed schools	.42
Nude bathing	.59	Nude bathing	.60	Bikinis	.42
Mixed schools	.56	Mixed schools	.59	Nude bathing	.37
Beer	.43	Beer	.45	Beer	.30
Blasphemy	.39	Blasphemy	.40	Laughing in class	.28
Rock-and-Roll	.37	Rock-and-roll	.39	Blasphemy	.28
Corporal punishment	.80	Corporal punishment	.82	Corporal punishment	.69
Death to thieves	.80	Death to thieves	.82	Death to thieves	.66
Death to enemies	.70	Death to enemies	.71	Death to enemies	.55
Beating children	.67	Beating children	.70	Beating children	.48

Table 6.19, and comparing with the English data reported by Nias (1973), we notice that eleven out of the fifteen items are the same, although in a different order.

6.6 Final Comments

In this chapter we have applied factor analysis methods using as input contingency type correlation coefficients which are functions of the estimated parameter of association of the C-type distribution: the global cross-product ratio for R×C contingency tables. This approach is justified using the underlying variable model based on C-type distributions as explained in Chapter 4. The main advantage of the method compared with other methods of factor analysis for categorical data is the great reduction in computing time, allowing the method to be applied to large data sets, as in some examples of this chapter. Before the advent of specific factor analysis methods for categorical data, the great majority of the analysis were carried out using the product moment correlation coefficient as input to traditional factor analysis methods. For this reason, we have compared the results using the new approach with the traditional factor analysis results. In the first example of this chapter - Civil Service I data, using the Pearson matrix as input, the three-factor model would have been chosen by the Kaiser criterion (which is widely used by less experienced investigators). Using the contingency type correlation matrices, the three-factor model was rejected because it yielded improper solutions. A more careful analysis of the magnitude of the eigenvalues and a more consistent interpretation of the results showed that the one-factor model is more appropriate for the Civil Service I data. For Boots data, the second example in this chapter, we observed a different profile of the factor loadings for the Pearson matrix case compared with the contingency type correlation matrix cases. The third example, Civil Service II data, resulted in a remarkably stable solution compared with the Civil Service I data, for the same items. In this example the profile of the factor loadings showed similar results for the three correlation matrices. Finally, for the last data set, similar results were also obtained, but, again, a more satisfactory definition of the number of factors (using the scree test) is obtained for the contingency type correlation matrices.

CHAPTER 7 : HEYWOOD CASES IN UNRESTRICTED FACTOR ANALYSIS

7.1 Heywood cases and improper solutions: an introduction and related research

A Heywood solution is known in the literature of factor analysis as the occurrence of a negative or zero estimate of the error variance for one or more variables in any factor analysis solution. Occurrences of Heywood cases have been reported in the literature since the first observation of this kind of particular solution by Heywood (1931). Heywood cases may occur in any factor analysis method, they also occur in confirmatory factor analysis and there is some evidence in the literature that the maximum likelihood factor analysis method is particularly prone to the occurrence of Heywood cases. The causes for such occurrences are still not clearly understood and some few studies have tried to show, through empirical evidence, in which situations the occurrence of Heywood cases are more frequent.

We shall distinguish, in this chapter, Heywood solutions and improper solutions in factor analysis. The improper solutions in factor analysis that occur frequently are Heywood solutions, but not all Heywood solutions are improper solutions, and not every improper solution is a Heywood solution. Suppose we have a one-factor model with one or more of the factor loading parameters very high or, conversely, suppose one or more of the error variance parameters in the factor analysis model are positive but very near zero. A solution that reproduces this pattern, that is, a factor analysis solution that yields an exact (and no negative) zero error variance estimate, when the corresponding parameter is approximately zero, cannot be considered an improper solution. In this particular case, the only cause for the zero variance is the sampling variation and any small difference between the estimate and the parameter is only to be expected. From the practical point of view we can have situations in which the one-factor model fits the data and one of the variables is perfectly correlated with the single factor, meaning that this variable itself could be a good indicator of the factor. Suppose a situation where the corresponding factor loading parameter for that variable is 0.98, say. A solution that yields a factor loading estimate as 1.00 is not an "improper solution". It will be a Heywood case because the variance estimate of the error term for that variable is zero. But this is a proper solution, given the model.

There are also improper solutions in factor analysis that are not Heywood cases. If the true number of factors is known, any factor analysis solution, that has not the same number of factors as is assumed in the model, is an "improper solution". Unless we know the model for a particular factor analysis solution (as is the case in simulation studies), we cannot distinguish, in practical work, an improper solution from a Heywood solution, but, very frequently, when the number of factors is not that of the hypothesized model, a Heywood case will indicate an improper solution, as we shall see in a simulation study to be presented in this chapter.

Although we shall consider only unrestricted factor analysis in this study, a review of the earlier research about Heywood cases will be made, considering also confirmatory factor analysis.

Martin and McDonald (1975) distinguish two types of Heywood solution: an exact Heywood solution when at least one unique variance is zero but none are negative and an ultra-Heywood solution where at least one unique variance is negative. Ultra-Heywood cases are, obviously, improper solutions, because we cannot have negative variances. But an exact Heywood solution may not be an improper solution as we explained before.

Most of the factor analysis programs available in the statistical analysis packages, do not allow the communalities of the variables to exceed one. That is the case for the BMDP and SPSS packages. Some of the factor analysis programs in the SAS package have the option for ultra-Heywood cases, that is, they allow communalities to exceed one. Therefore on using either BMDP, SPSS (or SPSS-X) an ultra-Heywood case will not be observed, although SPSS will print "the communality is greater than one" and will stop the iteration process.

In the next section, we shall present a simulation study, using the BMDP program for maximum likelihood factor analysis, where we will identify some of the possible causes for Heywood cases.

In past research, there are some simulation studies relevant to the present study, although some of them are concerned with the confirmatory factor analysis model. We now review these studies.

Tumura and Fukutomi (1970) have presented some numerical experiments to investigate the occurrence of Heywood cases in six different cases, where the uniqueness of the solution is considered and also where the given number of factors (m) for the solution is different from the true number of factors of the model. Jöreskog's unrestricted maximum likelihood factor analysis method was considered in the study, which is limited in the sense that only one or two experiments per case was analysed. Nevertheless, the authors conclude that for the case where Λ is unique and m=k, Heywood cases "occur occasionally if Λ contains some row vector with their length equal to nearly one" (see also Tumura, Fukutomi and Asoo, 1968).

A Monte Carlo study is presented by Boomsma (1985) to assess the problems of nonconvergence, improper solutions and starting values in LISREL maximum likelihood estimation for confirmatory factor analysis. Results on the likelihood ratio chi-square statistic for goodness-of-fit are also presented. Twelve factor analysis models were studied, all having two factors (correlated and not correlated factors). The factor pattern Λ (p×2), where p is the number of observed variables, was chosen such that half of the observed variables had a non zero loading on the first factor and a zero loading on the second one, and the reverse for the other half (p=6 or 8). The size of the factor loadings were chosen as small (0.4; 0.6); medium (0.6; 0.8) and large (0.8; 0.9). The sample sizes were 25, 50, 100, 200 and 400 (with 300 replications of each). In this study, Boomsma considers only the ultra-Heywood cases (negative estimates of the error variance). She concludes that "there is a real danger of improper solutions" with small sample size. In the simulation results, the occurrence of improper solutions increased as 1) sample size decreased; 2) the number of variables in the model was six rather than eight and 3) the population values of the error variance were close to zero.

Anderson and Gerbing (1984) also present a Monte Carlo study for the LISREL confirmatory factor analysis method. They analyse 54 models, with 2, 3 or 4 factors, for sample sizes of 50, 75, 100, 150 and 300 (with 100 replications of each). The proportion of nonconvergent and improper solutions that occurred in obtaining 100 good solutions per cell is presented. They conclude that a sample size of 150 for models with three or more indicators per factor (6 or more variables in the model) will usually be sufficient for a convergent and proper solution. In this study the solutions are defined as improper when one or more of the unique variances is less than a positive, arbitarily small, prescribed number such as 0.005. Anderson and Gerbing also observe that the occurrence of improper solutions increased as 1) sample size decreased; 2) the number of indicators per factor (and consequently the number of variables in the model) decreased; 3) correlation between factors were 0.3 rather than 0.5. For the models analysed, they also observe that with two indicators per factor (small number of variables), loadings of 0.9 give the largest proportion of improper solutions, whereas for larger numbers of variables no improper solutions occurred for models with loadings 0.9. Results on goodness-of-fit indices are also presented in this Monte Carlo study.

Seber (1984) reports some results from a simulation study by Francis (1973, 1974). Francis' analysis is based on exploratory or unrestricted factor analysis models with two or three factors. The sample size is 50. Twelve models were generated with different factor patterns. Again in this case the solution is said to be improper if the error variances are less than an arbitrary small positive number (e.g., 0.005). Several cases of improper solutions were observed when the number of factors for a particular solution was greater than the true number of factors of the model.

Other researchers have proposed methods to avoid the occurrence of Heywood cases or for detecting the causes of Heywood cases. We now review briefly these methods.

Jöreskog (1967) proposes a procedure to deal with improper solutions for the maximum likelihood factor analysis method. He defines the problem of improper (Heywood) solution as follows: "Since the diagonal elements of ψ are variances the function $f_k(\psi)$ is defined in the region where all the diagonal elements of ψ are positive" (k is the number of factors). "We have no guarantee, however, that all partial derivatives of f_k vanish at a point where all the diagonal elements of ψ are positive. This suggests that we shall define $f_k(\psi)$ in the region R_ε , where $\psi_{ii} > \varepsilon$ for all $i = 1,2,\ldots,p$ and where ε is a positive, arbitrarily small, prescribed number. The problem, then, is to find the minimum of $f_k(\psi)$ in the region R_ε . Since R_ε is a closed region, the minimum is found either in the interior of R_ε or on the boundary. If the minimum is found in the interior of R_ε , we shall say that the minimum is a proper solution. If on the other hand, the minimum is found on the boundary of R_ε , the solution is improper".

We have transcribed Jöreskog's text because it seems to be the origin of the term "improper solution", which has been used frequently. Jöreskog (1967, p.443) also says that "such improper (Heywood) solutions occur more often than is usually expected". The procedure that he proposes to avoid such improper solutions is to eliminate partially the variables with unique variances equal to ε and the analysis continues from the conditional dispersion matrix. The solution finally accepted in this process is combined with the principal components of the eliminated variables, to give a complete solution for all the original variables.

Martin and McDonald (1975) propose a Bayesian procedure for estimation in unrestricted factor analysis. The procedure has as one of its objectives to avoid inadmissable estimates of unique variances. A choice of the form of the prior distribution is justified and empirical examples are shown.

Finally, we will review the paper by Van Driel (1978) which has been cited in almost all studies about Heywood cases. Van Driel has identified some of the causes of Heywood solutions, by dropping the constraints of positive definiteness of the matrices containing the parameters of the factor analysis model. He proposes a method, which he calls "the nonclassical approach" and analyses some artificial data drawn from 5 populations, corresponding to five factor analysis models. The models are called: "Close to zero" (one of the unique variances is close to zero and the others are all equal to 0.5); "Close to one" (one of the unique variances is close to one and the others are equal to 0.5); "Dwarf" (all unique variances are equal to 0.5 and the second factor has loadings very small comparing with the first factor); "Heywood" (the classical one-factor model example where one of the unique variances is supposed negative) and "Anderson and Rubin" (a three-factor model with unique variances equal to 0.5; the factor matrix for this population is in accordance with the Anderson and Rubin identification condition). In this study five samples are drawn from each population, each with sample size 800, and each sample is analysed with the classical and non-classical approach for every appropriate number of factors.

Van Driel (1978) referring to Jöreskog's paper calls attention to the "subtle" difference between the terms "improper solution" and

"Heywood cases", but he uses the term improper as meaning Heywood solutions (that is, at least one unique variance negative or zero - small values of the variance, such as 0.004 are considered proper by Van Driel, as for example in the "close to zero" example). Van Driel identifies three causes for Heywood cases:

1) sampling fluctuations combined with true values of ψ close to zero;
2) there does not exist any factor analysis model that fits the data;
3) indefiniteness of the model (e.g. too many true factor loadings are zero).

Starting from the results of the previous studies two main questions arise:

1) How "close to zero" should be the unique variance parameters in the factor analysis model to cause Heywood cases?

2) How often do Heywood cases occur as a consequence of chosing a given number of factors different from the true number of factors of the model?

The first question is approached by Boomsma (1985) when she generates models with "large", "medium" and "small" factor loadings leading to different magnitudes of the unique variances of the model. Boomsma's results are, however, for confirmatory factor analysis using the LISREL program. We shall present some results for unrestricted factor analysis.

Concerning the second question, suppose <u>m</u> is the given number of factors for a particular factor analysis solution and q is the true number of factors of the model. We observed that Tumura and Fukutomi (1970) did not obtain Heywood cases when m>q in their numerical experiments, but on the other hand several cases of Heywood solutions are reported by Seber with reference to Francis' results when m>q (see Seber, 1984, p.232, Table 5.20). We also observed that several numerical examples presented by Jöreskog show the occurrence of Heywood solutions when increasing the number of factors for a particular example (see Jöreskog, 1967, p.474, Table 8). We then suppose that another possible cause of Heywood cases is the inappropriateness of the solution for a given number of factors.

To assess the effect of sampling variation and model characteristics on the occurrence of Heywood cases for unrestricted factor analysis using the maximum likelihood method, a Monte Carlo study was designed. As a by-product of the study some results about the goodness-of-fit test of the model are also obtained. This simulation study is described in the next section.

7.2 The effect of sampling variation and model characteristics on the occurrence of Heywood cases for maximum likelihood factor analysis: a simulation study.

Our first objective is to study how the normal theory estimators for maximum likelihood unrestricted factor analysis perform regarding the occurrence of Heywood cases for models with specified characteristics. Estimates of the MLFA model are provided by the BMDP factor analysis program using an algorithm developed by Jenrich and Sampson (see Dixon et al, 1983).

The normal random variates are created using the Random Number Generator of the BMDP package.

Simulation design

For this Monte Carlo study three one-factor models were chosen for different magnitudes of the first factor loading. The one-factor model for variables with mean zero is given by

$$x_i = \lambda_i z + e_i \qquad i=1,2,\ldots,p$$

such that $\text{var}(x_i) = 1$, $\text{var}(e_i) = \psi_i$, z and e_i are the normal generated variables and

$$h_i^2 + \psi_i = 1$$

where h_i^2 is the communality of the i-th observed variable, and ψ_i unique variance or error variance. The first factor pattern Λ $(p\times 1)$, where p is the number of observed variables was chosen such that the first observed variable had a "close to zero" unique variance or a very high loading ($\lambda_1 = 0.98$) and all other loadings equal to 0.5 ($\lambda_j = 0.5$, $j \neq 1$). This model will be called Model I.

The other models are similar, but the idea was to vary the first loading in such a way that we had three different degrees of "close to zero" variances. The last model having far from zero but not "close to one" unique variance. The three models are:

<u>Model I</u> $$\Lambda_{I} = \{ \lambda_1 = 0.98;\ \lambda_i = 0.5\ (j \neq 1)\} \quad \text{or}$$

$$\psi_{I} = \{ \psi_1 = 0.0396;\ \psi_i = 0.75\ (j \neq 1)\}$$

<u>Model II</u> $$\Lambda_{II} = \{ \lambda_1 = 0.90;\ \lambda_i = 0.5\ (i \neq 1)\} \quad \text{or}$$

$$\psi_{II} = \{ \psi_1 = 0.19;\ \psi_i = 0.75\ (j \neq 1)\}$$

<u>Model III</u> $$\Lambda_{III} = \{ \lambda_1 = 0.70;\ \lambda_i = 0.447\ (i \neq 1)\} \quad \text{or}$$

$$\psi_{III} = \{ \psi_1 = 0.51;\ \psi_i = 0.8\ (j \neq 1)\}$$

For each model three different numbers p of observed variables were analysed so as to represent a range of values typically encountered in practice (p=5; p=10 and p=20). Sample sizes were chosen according with the criterion: small (N=50); medium (N=100) and large (N=500). For each cell of this design, 20 replications were generated.

Finally, to assess the effect of having m>q on Heywood cases, where m is the given number of factors in one solution and q is the true number of factors of the model (q=1 in this case), we chose to analyse the correlation matrices generated by Model III (where the occurrence of Heywood cases is assumed to be very small or zero) with a two-factor solution.

Although the above design produced 720 separate analyses, admittedly, this is a very limited Monte Carlo study, with respect to different models studied, different sample sizes and number of replications. Nevertheless, the study should give a good deal of important information related to the occurrence of Heywood cases, standard errors of the MLFA estimators, results on the likelihood criterion given by the MLFA (BMDP) program and results about the empirical frequency distribution of the eigenvalues. Our results are, however, limited to the cases here studied, no generalizations beyond these models will be made.

Results

The MLFA/BMDP program produces factor loadings estimates and unique variances within the parameter space or on the boundary. No ultra Heywood cases can be observed, because of the constraints in the program. In obtaining the 720 convergent analyses, we observed only 2 nonconvergent cases, for Model III and when forcing a misspecification of the model with 2 factors and 20 variables. The results to be presented in this section are related to the proportion of exact Heywood cases for each model.

In Table 7.1 we present the percentage of Heywood solutions in each cell of the simulation design for Models I and II. For each cell we observed 20 replications.

Table 7.1 - Proportion of Heywood solutions for Models I and II in the Monte Carlo study (20 replications per cell).

		Sample Size		
Model	No of var	50	100	500
	5	.80	.40	.10
MODEL I	10	.30	.10	.00
[λ_1= 0.98]	20	.50	.25	.00
	5	.20	.25	.00
MODEL II	10	.00	.00	.00
[λ_1= 0.90]	20	.00	.00	.00

No more than one zero variance estimate was observed for Models I and II although within a single replication more than one variance can be zero. The proportion of Heywood solutions decreases as the sample increases, in general. A greater proportion of Heywood solutions is observed for a small number of variables in the model. For Model I and small sample sizes a greater proportion of Heywood solutions is observed when the number of variables is 20 rather than 10. A greater number of replications per cell would be necessary to confirm this tendency. The results in Table 7.1 are in accordance with the findings of Van Driel (1978), that is the "close to zero" population is one of the causes for Heywood cases combined with sampling variation. At this point, we call attention to the fact that for small sample size, N=50, say, the proportion of Heywood solutions is very high (80%). For Model I the Heywood cases were observed always for the first variable ("close to zero" case), therefore these solutions are very similar to the true model (we observe a factor loading $\hat{\lambda}_1 = 1.0$ where the parameter is $\lambda_1 = 0.98$. Due to sampling variation, solutions with the first loading equal to one (and consequently unique variance equal to zero) are expected to occur and such a solution cannot be called "improper". They are exact Heywood cases, but the solution is proper.

In Table 7.2 we present the proportion of Heywood solutions out of 20 replications for Model III for different sample sizes. We also show the proportion of Heywood solutions that occcur as a result of a simulated misspecification of the model, that is, we knew that the model had one factor, but we asked the program to produce the two-factor solution. We then observed a very high proportion of Heywood solutions for two factors and even more than one variable with zero variance. The Heywood cases were observed for any variable, not always for the first as in the case of Models I and II. In this case we have improper solutions. We then conclude that another cause for Heywood cases is the inclusion of too many factors in the solution. We believe that many of the Heywood solutions observed in the literature are due to the fact that they are over-factored (e.g. too many factors). When analysing empirical data, it is impossible to know the true number of factors of the model. In the simulation studies we know the model, but this is an artificial situation. We suggest that, in empirical situations, when the researcher is using factor analysis and obtains a Heywood solution, he should reanalyse the data decreasing the number of factors by one. If the goodness-of-fit indices are good, that should be the best solution for factor analysis.

Table 7.2 - Proportion of Heywood solutions for Model III using one-factor solution and two-factor solution (20 replications per cell).

		Sample Size		
	No of Var	50	100	500
One-factor solution	5	.10	.05	.00
	10	.00	.00	.00
	20	.00	.00	.00
Two-factor solution	5	.90	.75	.50
	10	.55	.55	.35
	20	.30	.45	.30

As can be seen in Table 7.2, the proportion of Heywood solutions indicating an improper solution for two factor solutions is very high even for large sample size. The proportion seems to decrease as the number of variables increases. Model III is a one-factor model with the following error variances [$\psi_1 = 0.51$ and $\psi_i = 0.80$, $i \neq 1$]. It is interesting to observe that for the one-factor solution and for sample size 50, we observe cases with communality very near zero, or variances very near one, producing negative estimates of loadings, which could be considered as another kind of improper solution; the proportion of these cases was very small (1 case for p=5 and p=20 and two cases for p=10, all for N=50).

As a by-product of this simulation study we shall now present results about the Chi-square test which can be obtained from the likelihood criterion (LC) to be minimized. (The BMDP/MLFA program only prints the likelihood criterion). The Chi-square statistic can then be obtained by $\chi^2 = n$ LC, where n' is given by

$$n' = N-1-(2p+5)/6-2q/3$$

The χ^2 statistic for the unrestricted factor analysis model is tested as a chi-square variable with degrees of freedom given by

$$df = \tfrac{1}{2}[(p-q)^2 - (p+q)].$$

In Table 7.3 we present the proportion of significant chi-square values for $\alpha = 0.05$, for 20 replications in each cell for the three models analysed. We also include in Table 7.3 the results for the two-factor solutions for Model III.

Table 7.3 - Proportion of significant chi-square statistics for α = 0.05, for the three models (20 replications per cell).

		Sample Size		
Model	No of var	50	100	500
Model I	5	.00	.05	.10
	10	.05	.00	.15
	20	.15	.10	.05
Model II	5	.00	.00	.05
	10	.00	.05	.05
	20	.30	.10	.00
Model III (one-factor solution)	5	.05	.00	.00
	10	.00	.00	.05
	20	.25	.05	.00
Model III (two-factor solution)	5	.00	.00	.00
	10	.00	.00	.00
	20	.00	.05	.00

We observe in Table 7.3 that for small samples the observed proportion of significant chi-square statistics is higher than the expected proportion of 0.05 for the one-factor model. When analysing the two-factor solution the test accepts the model with two factors, which should not be accepted. But this is a known fact of this goodness-of-fit test, because it depends on the residual correlations, if we include more factors in the model the residuals become smaller, and consequently, the chi-square statistic. The factor analysis user should, for this reason, use more than one goodness-of-fit indice, including in the analysis other criteria such as Akaike's Information Criterion and Schwarz's Bayesian criterion (see Section 4.3).

Another interesting result from this simulation study is the empirical frequency distribution of the number of eigenvalues greater than one (γ>1) of the correlation matrices. In Tables 7.4 to 7.6 we present these empirical frequencies, for each of the one-factor models studied, according to sample size and number of variables in the model.

Table 7.4 - Empirical distribution of the numbers of eigenvalues greater than one in 20 replications of the simulation study; Model 1.

	N = 50		N = 100		N = 500	
	No of γ>1	%	No of γ>1	%	No of γ>1	%
P = 5	1	75	1	90	1	100
	2	25	2	10		
P = 10	1	--	1	20	1	100
	2	30	2	75		
	3	60	3	05		
	4	10				
P = 20	1	--	1	--	1	40
	2	--	2	--	2	45
	3	--	3	05	3	15
	4	--	4	20		
	5	35	5	35		
	6	50	6	30		
	7	10	7	10		
	8	05				

Table 7.5 - Empirical distribution of the number of eigenvalues greater than one for different sample sizes and number of variables of the model - Model II (20 replications per cell).

	N = 50		N = 100		N = 500	
	No of γ>1	%	No of γ>1	%	No of γ>1	%
P = 5	1	35	1	55	1	100
	2	60	2	40		
	3	05	3	05		
P = 10	1	--	1	05	1	100
	2	--	2	35	2	15
	3	55	3	50		
	4	45	4	10		
P = 20	1	--	1	--	1	--
	2	--	2	--	2	20
	3	--	3	--	3	55
	4	--	4	--	4	25
	5	--	5	05		
	6	25	6	45		
	7	65	7	50		
	8	10				

Table 7.6 - Empirical distribution of the number of eigenvalues greater than one. Model III (20 replications per cell)

	N = 50		N = 100		N = 500	
	No of γ>1	%	No of γ>1	%	No of γ>1	%
P = 5	1	70	1	90	1	100
	2	30	2	10		
P = 10	1	--	1	10	1	100
	2	35	2	60		
	3	55	3	30		
	4	10				
P = 20	1	--	1	--	1	30
	2	--	2	--	2	55
	3	--	3	--	3	10
	4	--	4	15	4	05
	5	15	5	55		
	6	75	6	20		
	7	10	7	10		

The Tables 7.4 to 7.6 show that with small sample size and for a number of variables in the model such as 20, several eigenvalues of the correlation are greater than one for the one-factor models I, II and III. If the factor analysis user chooses the number of factors by this criterion, as is still very common, with small samples and large number of variables, the inclusion of too many factors in the solution would occur. Even for a moderate sample size such as 100, that would be the case. On the other hand, for all models and cases, the scree test would be more appropriate since the magnitude pattern of the eigenvalues always shows a very high first eigenvalue compared with the others.

Finally, as another by-product of the simulation study we now present the results related to the parameter estimates of the models. In Table 7.7 we present the mean and standard deviation of the first factor loading for the three models, for each cell of the simulation design, based on 20 replications. We have included the Heywood solutions in all calculations. In Tables 7.8 and 7.9 we present the mean and standard deviation of the second and third factor loading for each model, respectively. The small number of replications in this Monte Carlo study should be kept in mind when considering the results for the cases studied.

Table 7.7 - Mean and standard deviation of the first factor loading estimates in each cell of the simulation design (20 replications per cell).

Model	Number of variables	Sample Size 50	100	200
	5	0.986 (0.033)	0.965 (0.042)	0.978 (0.016)
Model 1 (true value = 0.98)	10	0.970 (0.031)	0.967 (0.020)	0.969 (0.009)
	20	0.993 (0.010)	0.989 (0.010)	0.991 (0.005)
	5	0.918 (0.065)	0.938 (0.058)	0.905 (0.028)
Model II (true value = 0.90)	10	0.894 (0.039)	0.895 (0.026)	0.882 (0.017)
	20	0.913 (0.036)	0.920 (0.017)	0.914 (0.010)
	5	0.740 (0.198)	0.777 (0.116)	0.725 (0.051)
Model III (true value = 0.70)	10	0.691 (0.129)	0.724 (0.062)	0.710 (0.032)
	20	0.645 (0.094)	0.681 (0.045)	0.714 (0.028)

Note: Estimates are based in averaging the estimates in each cell. In parenthesis is the empirical standard deviation.

Table 7.8 - Mean and standard deviation of the second factor loading estimates in each cell of the simulation design (20 replications per cell).

Model	Number of variables	Sample Size 50	 100	 200
	5	0.498 (0.062)	0.493 (0.121)	0.479 (0.022)
Model I (true value = 0.50)	10	0.569 (0.089)	0.547 (0.058)	0.519 (0.029)
	20	0.509 (0.089)	0.470 (0.055)	0.496 (0.029)
	5	0.482 (0.083)	0.431 (0.111)	0.480 (0.036)
Model II (true value = 0.50)	10	0.567 (0.096)	0.506 (0.070)	0.534 (0.028)
	20	0.486 (0.090)	0.479 (0.070)	0.499 (0.033)
	5	0.397 (0.162)	0.359 (0.094)	0.431 (0.042)
Model III (true value = 0.447)	10	0.507 (0.166)	0.492 (0.065)	0.441 (0.032)
	20	0.325 (0.150)	0.354 (0.099)	0.470 (0.044)

Table 7.9 – Mean and standard deviation of the third factor loading estimates in each cell of the simulation design (20 replications per cell).

Model	Number of variables	Sample Size 50	 100	 200
	5	0.538 (0.109)	0.499 (0.082)	0.511 (0.028)
Model I (true value = 0.5)	10	0.443 (0.086)	0.492 (0.056)	0.505 (0.028)
	20	0.527 (0.111)	0.496 (0.066)	0.504 (0.035)
	5	0.574 (0.117)	0.535 (0.076)	0.511 (0.033)
Model II (true value = 0.5)	10	0.450 (0.111)	0.463 (0.068)	0.504 (0.028)
	20	0.584 (0.081)	0.536 (0.055)	0.500 (0.028)
	5	0.348 (0.154)	0.392 (0.074)	0.422 (0.031)
Model III (true value = 0.447)	10	0.270 (0.161)	0.384 (0.088)	0.431 (0.046)
	20	0.522 (0.099)	0.436 (0.073)	0.437 (0.034)

Discussion

In this simulation study the effect of sampling variation and model characteristics on the occurrence of Heywood cases was analysed. Two main causes of Heywood solutions in factor analysis were observed for the models analysed:

1) sampling variation combined with unique variance parameters close to zero, which is in accordance with Van Driel (1978);

2) misspecification of the model - too many factors are included in one particular solution causing improper solutions.

The occurrence of Heywood cases is much more frequent for small sample sizes. Factor analysis based on fifty or less observations should certainly be avoided, not only because of a higher possibility of Heywood cases, but because the sampling fluctuations may lead to solutions that differ substantially from the true model.

It was observed, generally, that the occurrence of Heywood cases increases as the number of the variables of the model decreases.

Our results, for unrestricted factor analysis, are in accordance with the findings of Boomsma (1985) and Anderson and Gerbing (1984) for confirmatory factor analysis, concerning the occurrence of Heywood cases.

For normal theory, the chi-square test has been shown to behave well, although a higher proportion than the expected rejects the model for small sample sizes and moderate number of variables ($p=20$).

The results from the simulation study also show that the Kaiser criterion for choosing the number of factors should not be used as it may lead to the occurrence of Heywood cases, caused by the inclusion of too many factors in a solution, mainly if the sample size is small and the number of variables large. With large sample sizes and small number of variables, the criterion may be useful if used together with other criteria.

As a final recommendation, we strongly advise that sample sizes of 100 or more are needed for reasonable factor analysis results.

7.3 The probability of occurrence of Heywood cases

A simulation study with the purpose of assessing the main causes of Heywood cases was presented in the last section. Heywood cases occurred even when a factor model with large residual variances generated the data. In this section we consider the case of one-factor models and we show how the probability of occurrence may be calculated. We also show that this probability for the one-factor model depends on sample size, parameter vectors and number of variables in the model. We shall consider the Minres method of fitting for factor analysis, that is, we shall suppose that the factor loadings are estimated under the condition that the sum of squares of the off-diagonal residuals is minimized.

The minres method was introduced by Harman and Jones (1966) although, as they point out, the idea of getting a factor solution by minimizing off-diagonal residual correlations was first posed by Thurstone (1954). They do not consider the minimization of the total residual matrix (including diagonal terms), which would lead to the principal-factor solution. We do not consider the numerical procedures for obtaining the minres solution. It is on the principle of the method that we shall base our method for obtaining the probabilities of occurrence of Heywood cases.

Consider the one-factor model given by

$$x_i = \lambda_i f + e_i \qquad i=1,2,\ldots,p \tag{7.1}$$

with $E(f) = E(e_i) = 0$; $E(fe_i) = 0$. Suppose $Var(f) = 1$, $E(e_i e_j) = 0$, $i \neq j$ and $Var(e_i) = \psi_i > 0$, $i=1,\ldots,p$. Suppose the x_i's are standardized so that $var(x_i) = 1$. Hence,

$$\lambda_i^2 + \psi_i = 1, \qquad i=1,2,\ldots,p \tag{7.2}$$

Then, we have

$$\lambda_i^2 < 1, \quad i=1,2,\ldots,p \tag{7.3}$$

For this model

$$\mathrm{corr}(x_i,x_j) = \rho_{ij} = \lambda_i\lambda_j . \tag{7.4}$$

Suppose, therefore, we fit the model by minimizing

$$S = \sum_{i=1}^{p} \sum_{j=i+1}^{p} (r_{ij} - \lambda_i\lambda_j)^2 \tag{7.5}$$

where r_{ij} is the observed correlation between variables x_i and x_j. (Actually, as x_i and x_j are standardized variables, r_{ij} and ρ_{ij} may be considered as covariances).

From a purely mathematical point of view, the problem of finding a minimum for the nonlinear function S is well defined. S is a function of the p(p-1)/2 off-diagonal residual correlations, which are dependent upon the p elements of the factor matrix $\Lambda(p\times 1)$. The minimum value for S occurs at the point where its partial derivatives with respect to the p elements of Λ are zero and its matrix of second derivatives is positive definite. Thus,

$$S'(\lambda_i) = \frac{1}{2}\frac{\partial S}{\partial \lambda_i} = \sum_{\substack{j=1\\ j\neq i}}^{p} \lambda_j r_{ij} + \lambda_i \sum_{\substack{j=1\\ j\neq i}}^{p} \lambda_j^2 \quad i=1,2,\ldots,p. \tag{7.6}$$

we also have,

$$\frac{1}{2}\ \frac{\partial^2 S}{\partial \lambda_i^2} = \sum_{\substack{j=1\\ j\neq i}}^{p} \lambda_j \geqslant 0$$

$$\frac{1}{2}\ \frac{\partial^2 S}{\partial\lambda_i\partial\lambda_j} = -r_{ij} + 2\lambda_i\lambda_j \tag{7.7}$$

If the system of equations

$$\frac{\partial S}{\partial \lambda_i} = 0 \quad i=1,2,\ldots,p$$

has a solution with $h_i^2 < 1$, then it is clearly a minimum by (7.7). However, there may not be such a solution. For fixed $\lambda_1,\ldots\lambda_{i-1},\lambda_{i+1},\ldots\lambda_p$, $S'(\lambda_i)$ is a linear increasing function of λ_i. We are interested in the behaviour of the function at the point $\lambda_i = 1$ and consequently $\psi_i = 0$. If the minimum of the function occurs at $\lambda_i = 1$, an exact Heywood case occurs. Let us consider the following cases:

1) Suppose that the derivative of S with respect to λ_i at the point $\lambda_i = 1$ is negative, that is $S'(1) < 0$.

At $\lambda_i = 1$, from (7.6) we have

$$\left[\frac{\partial S}{\partial \lambda_i} \right]_{\lambda_i=1} = -\sum_{\substack{j=1\\ j\neq i}}^{p} \lambda_j r_{ij} + \sum_{\substack{j=1\\ j\neq i}}^{p} \lambda_j^2$$

If $S'(1)<0$, this would imply that $S'(0)<0$ and that there will be no intermediate value of λ_i for which it is zero. The minimum of S will thus occur at $\lambda_i=1$, because S' is a linear increasing function of λ_i. A proof of this is given below:

Suppose $S'(1)<0$ then

$$-\sum_{\substack{j=1\\ j\neq i}}^{p} \lambda_j r_{ij} + \sum_{\substack{j=1\\ j\neq i}}^{p} \lambda_j^2 < 0$$

thus

$$\sum_{j\neq i}^{p} \lambda_j r_{ij} > \sum_{j\neq i}^{p} \lambda_j^2 > 0$$

Therefore

$$S'(0) = \left[\frac{\partial S}{\partial \lambda_i} \right]_{\lambda_i=0} = -\Sigma\, \lambda_j r_{ij} < 0$$

We recall that we are assuming $0 \leqslant \lambda_i \leqslant 1$. It is known that for the one-factor model, Λ reduces to a column vector of p elements that is unique apart from a possible change of sign of all its elements, which corresponds merely to changing the sign of the factor. According to Lawley and Maxwell (1971), such changes are merely trivial.

2) Suppose now that S'(1) is positive. Therefore S'(0) may be negative or positive and the minimum of S will not occur at the point $\lambda_i=1$.

3) Finally, if S'(1)=0, the minimum of S will occur at $\lambda_i=1$ and a Heywood case occur.

We are interested in evaluating the probability of an occurrence of a Heywood case for variable x_i. Therefore we need consider only cases 1) and 3) above. That is

$$\Pr\{S'(1)<0\} = \Pr\{ \sum_{\substack{j=1\\ j\neq i}}^{p} \lambda_j r_{ij} > \sum_{\substack{j=1\\ j\neq i}}^{p} \lambda_j^2 \} \quad i=1,2,\ldots,p \qquad (7.8)$$

For evaluating this probability we need to know the distribution of $\sum_{j\neq 1}^{p} \lambda_j r_{ij}$. An approximation may be obtained as follows.

Since the x's have zero means and unit variances

$$r_{ij} = \frac{1}{n} \sum_{h=1}^{n} x_{ih} x_{jh}$$

where h indexes the sample members. (The terms of this sum are independent for h=1,2,...,n; for sufficiently large n, this will be approximately normal by the central limit theorem).

Consider

$$z_i = \sum_{\substack{j=1\\ j\neq i}}^{p} \lambda_j r_{ij} = \frac{1}{n} \sum_{h=1}^{n} x_{ih} \sum_{j\neq i} x_{jh} \lambda_j$$

Let us denote $\sum_{j\neq i} x_{jh}\lambda_j$ by y_{ih} say. Therefore

$$z_i = \frac{1}{n} \sum_{h=1}^{n} x_{ih} y_{ih} \; .$$

We shall now find the moments of x_{ih} and y_{ih}. By definition we have

$$E(x_{ih}) = 0 \; , \; \text{var} \; (x_{ih}) = 1$$

Thus

$$E(y_{ih}) = 0 \text{ and}$$

$$\text{var}(y_{ih}) = \text{var}(\sum_{j \neq i} x_{jh} \lambda_j) = \sum_{\substack{j=1 \\ j \neq i}}^{p} \lambda_j^2 + \sum_{\substack{j=1 \\ j \neq i \\ j \neq k}}^{p} \sum_{\substack{k=1 \\ k \neq i}}^{p} \lambda_j \lambda_k \; \text{cov}(x_{jh}, x_{kh}) \tag{7.9}$$

Now

$$\text{cov}(x_{jh}, x_{kh}) = \lambda_j \lambda_k \tag{7.10}$$

thus

$$\text{cov}(y_{ih}) = \sum_{\substack{j=1 \\ j \neq i}}^{p} \lambda_j^2 + \sum_{\substack{j=1 \\ j \neq i \\ j \neq k}}^{p} \sum_{\substack{k=1 \\ k \neq i}}^{p} \lambda_j^2 \lambda_k^2 \tag{7.11}$$

$$\text{cov}(x_{ih}, y_{ih}) = \sum_{\substack{j=1 \\ j \neq i}}^{p} \lambda_j \; \text{cov}(x_{ih}, x_{jh}) = \lambda_i \sum_{\substack{j=1 \\ j \neq i}}^{p} \lambda_j^2 = E(x_{ih}, y_{ih}) \tag{7.12}$$

Let us suppose that the observed variables x_i, $i=1,\ldots p$ have a multivariate normal distribution. Thus, $y_{ik} = \sum_{\substack{j=1 \\ j \neq i}}^{p} x_{jh} \lambda_j$, which is the sum of p-1 dependent normal variables, is also normal with mean and variance given by (7.9). Hence x_{ih} and y_{ih} have a normal bivariate distribution. From Kendall and Stuart (1977, Vol.I, p.85) it is known that if (x_{ih}, y_{ih}) are normal bivariate we have, using the fact that $\sigma^2_{x_{ih}} = 1$

$$E(x^2_{ih} \cdot y^2_{ih}) = (1 + 2 \; \rho^2_{x_{ih} y_{ih}}) \; \sigma^2_{x_{ih}} \sigma^2_{y_{ih}} = \sigma^2_{y_{ih}} + 2 \; [E(x_{ih}, y_{ih})]^2$$

Thus

$$\begin{aligned} \text{Var}(x_{ih} y_{ih}) &= E(x^2_{ih} y^2_{ih}) - [E(x_{ih} y_{ih})]^2 = \\ &= \sigma^2_{y_{ih}} + 2[E(x_{ih} y_{ih})]^2 - [E(x_{ih} y_{ih})]^2 \\ &= \sigma^2_{y_{ih}} + [E(x_{ih} y_{ih})]^2 \end{aligned} \tag{7.13}$$

Now

$$z_i = \frac{1}{n} \sum_{h=1}^{n} x_{ih} y_{ih}$$

where $z_{ih} = x_{ih} y_{ih}$ are independent for $h=1,\ldots,n$. Therefore, using (7.12)

$$E(z_i) = \frac{n}{n} \; E(x_{ih} y_{ih}) = \lambda_i \sum_{\substack{j=1 \\ j \neq i}}^{p} \lambda_j^2 \qquad (7.14)$$

and using (7.11) and (7.13) we have

$$\mathrm{Var}(z_i) = \frac{1}{n} \left[\sum_{\substack{j=1 \\ j \neq i}}^{p} \lambda_j^2 + \sum_{\substack{j=1 \\ j \neq i \\ j \neq k}}^{p} \sum_{\substack{k=1 \\ k \neq i}}^{p} \lambda_j^2 \lambda_k^2 + \lambda_i^2 \left(\sum_{\substack{j=1 \\ j \neq i}}^{p} \lambda_j^2 \right)^2 \right] \qquad (7.15)$$

For n sufficiently large

$$z_i = \frac{1}{n} \sum_{h=1}^{n} x_{ih} y_{ih} = \sum_{\substack{j=1 \\ j \neq i}}^{p} \lambda_j r_{ij}$$

is approximately normal, by the central limit theorem, with mean and variance given by (7.14) and (7.15) respectively. Therefore (7.8) may be calculated approximately as

$$1 - \Phi \left[\frac{\sum_{j \neq i} \lambda_j^2 - \lambda_i \sum_{j \neq i} \lambda_j^2}{\frac{1}{\sqrt{n}} \sum_{j \neq i} \lambda_j^2 + \lambda_i^2 \left[\sum_{j \neq i} \lambda_j^2 \right] + \sum_{\substack{j \neq i \\ j \neq k}} \sum_{k \neq i} \lambda_j^2 \lambda_k^2)} \right] \quad i=1,2,\ldots p \qquad (7.16)$$

where $\Phi(z)$ is the normal distribution function.

The above expression gives the asymptotic probability of occurrence of a Heywood case for variable x_i in function of the sample size, magnitude of the factor loading parameters and the number p of variables in the model. As can be seen, if $n \to \infty$, the expression (7.16) converges to zero. If $\lambda_i \to 1$, the probability of Heywood case converges to 0.5 for any values of $\lambda_j (j \neq i)$.

We now present tables of the probability of occurrence of a Heywood case for several values of sample size, different values of p, considering

1) $\lambda_i = 0.98$ and $\lambda_j = 0.5$ $(j \neq i)$ $\quad i,j=1,2,\ldots p$
2) $\lambda_i = 0.90$ and $\lambda_j = 0.5$ $(j \neq i)$
3) $\lambda_i = 0.80$ and $\lambda_j = 0.5$ $(j \neq i)$

4) $\lambda_i = \lambda_j = 0.50$ $\quad i,j=1,2,\ldots p$
5) $\lambda_i = \lambda_j = 0.90$ $\quad i,j=1,2,\ldots p$

Table 7.10 - Asymptotic probability of occurrence of Heywood cases for various sample sizes (n), different number of variables (p) and different magnitude of the factor loading parameters (cases 1 to 5)

1) $\lambda_i=0.98 \quad \lambda_j=0.5 \ (j \neq i) \quad i,j=1,\ldots p$

n p	50	100	200	400	500	1000	2000	5000
5	.4658	.4516	.4318	.4040	.3929	.3504	.2935	.1952
10	.4628	.4474	.4258	.3958	.3838	.3380	.2772	.1750
15	.4618	.4461	.4240	.3932	.3809	.3341	.2722	.1689
20	.4613	.4453	.4230	.3917	.3793	.3320	.2695	.1656
30	.4608	.4447	.4220	.3904	.3779	.3300	.2670	.1627
40	.4605	.4443	.4215	.3897	.3770	.3289	.2655	.1610
50	.4604	.4441	.4211	.3892	.3766	.3283	.2647	.1600
100	.4601	.4438	.4207	.3887	.3759	.3274	.2635	.1586

2) $\lambda_i=0.90 \quad \lambda_j=0.5 \ (j \neq i) \quad i,j=1,\ldots p$

n p	50	100	200	400	500	1000	2000	5000
5	.3293	.2660	.1884	.1057	.0812	.0241	.0026	.0000
10	.3146	.2473	.1671	.0860	.0634	.0154	.0012	.0000
15	.3097	.2411	.1602	.0800	.0580	.0131	.0009	.0000
20	.3070	.2380	.1567	.0770	.0555	.0121	.0008	.0000
30	.3047	.2349	.1534	.0740	.0531	.0112	.0007	.0000
40	.3034	.2334	.1517	.0727	.0518	.0107	.0006	.0000
50	.3026	.2324	.1508	.0719	.0511	.0104	.0005	.0000
100	.3012	.2306	.1487	.0703	.0497	.0099	.0005	.0000

3) $\lambda_i = 0.80$ $\lambda_j = 0.5$ $(j \neq i)$ $i,j=1,2,\ldots p$

n p	50	100	200	400	500	1000	2000	5000
5	.1802	.0980	.0339	.0048	.0019	.0000	.0000	.0000
10	.1572	.0773	.0220	.0022	.0007	.0000	.0000	.0000
15	.1495	.0710	.0189	.0016	.0005	.0000	.0000	.0000
20	.1457	.0679	.0174	.0014	.0004	.0000	.0000	.0000
30	.1421	.0650	.0161	.0013	.0003	.0000	.0000	.0000
40	.1405	.0635	.0155	.0011	.0003	.0000	.0000	.0000
50	.1392	.0626	.0151	.0011	.0003	.0000	.0000	.0000
100	.1370	.0609	.0143	.0010	.0003	.0000	.0000	.0000

4) $\lambda_i = \lambda_j = 0.50$ $i,j=1,2,\ldots p$

n p	50	100	200	400	500	1000	2000	5000
5	.0062	.0002	.0000	.0000	.0000	.0000	.0000	.0000
10	.0025	.0000	.0000	.0000	.0000	.0000	.0000	.0000
15	.0017	.0000	.0000	.0000	.0000	.0000	.0000	.0000
20	.0015	.0000	.0000	.0000	.0000	.0000	.0000	.0000
30	.0012	.0000	.0000	.0000	.0000	.0000	.0000	.0000
40	.0011	.0000	.0000	.0000	.0000	.0000	.0000	.0000
50	.0010	.0000	.0000	.0000	.0000	.0000	.0000	.0000
100	.0009	.0000	.0000	.0000	.0000	.0000	.0000	.0000

5) $\lambda_i = \lambda_j = 0.90$ $i,j=1,2,\ldots,p$

n p	50	100	200	400	500	1000	2000	5000
5	.3025	.2322	.1507	.0717	.0510	.0103	.0005	.0000
10	.3009	.2303	.1483	.0700	.0494	.0098	.0005	.0000
15	.3004	.2297	.1477	.0694	.0490	.0097	.0005	.0000
20	.3002	.2294	.1474	.0692	.0488	.0096	.0005	.0000
30	.3000	.2292	.1471	.0690	.0486	.0095	.0005	.0000
40	.2999	.2290	.1457	.0689	.0485	.0095	.0005	.0000
50	.2998	.2290	.1457	.0688	.0484	.0094	.0005	.0000
100	.2997	.2288	.1457	.0687	.0484	.0094	.0005	.0000

Final comments

Comparing the proportion of occurrence of Heywood cases in the simulation study of Section 7.2 with the probability of Heywood cases as obtained in Table 7.10, we must note that the simulation study was based on the maximum likelihood factor analysis method using the algorithm of the BMDP package. The tables of probability of Heywood case are obtained from considerations about the principle of the minres method, that is, minimizing the sum of squares of the off-diagonal residual correlations. We also believe that some of the constraints in the algorithm for the MLFA method, as for example stopping the iteration for any particular variable which $\psi_i=0$, could lead to a different proportion of Heywood cases in some situations, because of a possible bias in the method.

For large sample sizes, the probability of Heywood cases is very small. We observe in Table 7.10 that the probability of Heywood cases decreases as 1) the sample size increases; 2) the number of variables in the model increases but a small variation is observed from p=5 to p=100 for all tables. Related to the magnitude of the factor loading parameters, we observe that when only one factor loading increases (the others being equal to 0.5), the probability of Heywood cases also increases. Finally, it is interesting to note, comparing cases 2) and 5) in Table 7.10, that when only one loading is equal to 0.90, the probabilities of Heywood cases are slightly higher than when all loadings are equal to 0.90.

The approximate probability of the occurrence of Heywood cases for different parameter vectors for the one-factor model may be easily evaluated using expression (7.16). In Table 7.10 we have included only some cases.

REFERENCES

Altham, P.M.E. (1970). The measurement of association of rows and columns for an rxs contingency table. Journal of the Royal Statistical Society, B, 32, 63-73.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B.N. and Csaki, F. (Eds), Second International Symposium on Information Theory. Budapest: Akademiai Kiado.

Akaike, H. (1983). Information measures and model selection. 44th Session of the International Statistical Institute, Madrid, 277-290.

Andersen, E.B. (1982). Latent trait models and ability parameter estimation. University of Copenhagen, Institute of Statistics, Research Report No.79.

Anderson, J.C. and Gerbing, D.W. (1984). The effect of sampling error on convergence, improper solutions, and goodness of fit indices for maximum likelihood confirmatory factor analysis. Psychometrika, 49, 155-173.

Ashton, W.D. (1972). The logit transformation with special reference to its uses in Bioassay. Griffin's Statistical Monographs and Courses, 32. London: Griffin

Bartholomew, D.J. (1980). Factor analysis for categorical data. Journal of the Royal Statistical Society, B, 42, 293-321.

Bartholomew, D.J. (1981). Posterior analysis of the factor model. British Journal of Mathematical and Statistical Psychology, 34, 93-99.

Bartholomew, D.J. (1983). Latent variable models for ordered categorical data. Journal of Econometrics, 22, 229-243.

Bartholomew, D.J. (1984a). The foundations of factor analysis. Biometrika, 71, 221-232.

Bartholomew, D.J. (1984b). Scaling binary data using a factor model. Journal of the Royal Statistical Society, B, 46, 120-123.

Bartholomew, D.J. (1985). Foundations of factor analysis: some practical implications. British Journal of Mathematical and Statistical Psychology, 38, 1-10.

Bartholomew, D.J. (forthcoming). Latent variable models and factor analysis. London: Charles Griffin & Co.

Bentler, P.J.M. (1986). Structural modeling and Pscyhometrika: an historical perspective on growth and achievements. *Pscyhometrika*, *51*, 35-51.

Bishop, Y.M; Fienberg, S.E. and Holland, P.W. (1975). *Discrete multivariate analysis. Theory and practice*. Cambridge, Mass.: the MIT Press.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Bock, R.D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Pscyhometrika*, *46*, 443-459.

Bock R.D. and Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*, 179-197.

Boomsma, A. (1985). Nonconvergence, improper solutions and starting values in LISREL maximum likelihood estimation. *Psychometrika*, *50*, 229-242.

Brown, M.B. and Benedetti, J.K. (1977). On the mean and variance of the tetrachoric correlation coefficient. *Psychometrika*, *42*, 347-355.

Browne, M.N. (1982). Covariance structures. In Hawkins, D.M. (ed) *Topics in Applied multivariate analysis*. London: Cambridge University Press.

Browne, M.N. (1984). Asymptotically distribution-free methods for the analysis of covariances structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62-83.

Burt, C. (1950). The factor analysis of qualitative data. *The British Journal of Psychology, Statistical Section*, *3*, 166-185.

Cattel, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245-276.

Cattel, R.B. and Jaspers, J. (1967). A general plasmode (No.30-10-5-2) for factor analytic exercises and research. *Multivariate Behavioral Research Monographs*, No.67-3.

Chambers, R.G. (1982). Correlation coefficients from 2x2 tables and from biserial data. *British Journal of Mathematical and Statistical Psychology*, *35*, 216-227.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*, 5-31.

Clogg, C.C. and Sawyer, D.O. (1981). A comparison of alternative models for analysing the scalability of response patterns. In Leinhardt, S. (ed), Sociological Methodology. San Francisco: Jossey-Bass.

Cureton, E.E. and D'Agostino, R.B. (1983). Factor analysis: An applied approach. New Jersey: Lawrence Erlbaum Associates.

Dale, J.R. (1984). Local versus global association for bivariate ordered responses. Biometrika, 71, 507-14.

De Leeuw, J.; Keller, W. and Wansbeek, T. (1983). Interfaces between Econometrics and Psychometrics. Editors' Introduction. Journal of Econometrics, 22, 1-12.

Divgi, D.R. (1979). Calculation of the tetrachoric correlation coefficient. Psychometrika, 44, 169-172.

Dixon, W.J. and Brown, M.B. (eds) (1983). BMDP Statistical Software (1983 Printing with additions). Berkeley: University of California Press.

Everitt, B.S. (1984). An introduction to latent variable models. London: Chapman and Hall

Francis, I. (1974). Factor analysis: its purpose, practice and packaged programs. Invited paper, American Statistical Association. NY, December 1973.

Francis, I. (1974). Factor analysis: fact or fabrication. Mathematical Chronicle, 3, 9-44.

Goodman, L.A. (1978). Analyzing Qualitative/Categorical Data, (Ed. J. Magidson) Cambridge, Mass: Abt Books.

Goodman, L.A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. Journal of the American Statistical Association, 74, 537-552.

Goodman, L.A. (1981). Association models and the bivariate normal for contingency tables with ordered categories. Biometrika, 68, 347-355.

Gorsuch, R.L. (1983). Factor analysis (2nd ed.) New Jersey: Lawrence Erlbaum Associates.

Green, B.F. (1952). Latent structure analysis and its relation to factor analysis. Journal of the American Statistical Association, 47, 71-76.

Gumbel, E.J. (1961). Bivariate logistic distributions. Journal of the American Statistical Association, 56, 335-349.

Gweke, J.F. and Singleton, K.J. (1980). Interpreting the likelihood ratio statistic in factor models when sample size is small. Journal of the American Statistical Association, 75, 133-137.

Harman, H.H. (1976). Modern factor analysis (3rd ed). Chicago: The University of Chicago Press.

Harman, H.H. and Fukuda, Y. (1966). Resolution of the Heywood case in the Minres solution. Psychometrika, 31, 563-571.

Harman, H.H. and Jones, W.H. (1966). Factor analysis by minimizing residuals (MINRES). Psychometrika, 31, 351-368.

Heywood, H.B. (1931). On finite sequences of real numbers. Proceedings of the Royal Society of London, 134, 486-501.

Insel, P.M. and Wilson, G.D. (1971). Measuring social attitudes in children. British Journal of Social Clinical Psychology, 10, 187-200.

Joe, V.C. (1984). Factor analysis of the conservatism scale. The Journal of Social Psychology, 124, 175-178.

Johnson, N.L. and Kotz, S. (1972). Distributions in Statistics: continuous multivariate distributions. New York: John Wiley & Sons.

Johnson, R.A. and Wichern, D.W. (1982). Applied multivariate statistical analysis. New Jersey: Prentice Hall.

Jöreskog, K.G. (1967). Some contributions to maximum likelihood factor analysis. Psychometrika, 32, 443-482.

Jöreskog, K.G. and Sörbom, D. (1984). LISREL VI. Analysis of linear structural relationships by maximum likelihood, instrumental variables and least squares methods. Mooresville, IN: Scientific Software.

Jöreskog, K.G. and Van Tillo, M. (1971). New rapid algorithms for factor analysis by unweighted least squares, generalized least squares and maximum likelihood. Research memorandum, 71-5. Princeton: Educational Testing Service.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20, 141-151.

Kaiser, H.F. (1963). Image analysis. In Harris, C.W. (ed): Problems in measuring change. Madison: The University of Wisconsin Press.

Kaiser, H.F. (1970). A second generation Little Jiffy. Psychometrika, 35, 410-415.

Kaiser, H.F. and Caffrey, J. (1965). ALPHA factor analysis. Psychometrika, 30, 1-14.

Kendall, M. and Stuart, A. (1977). The advanced theory of statistics (Vol.1, 4th ed). London: Charles Griffin & Co.

Kendall, M. and Stuart, A. (1979). The advanced theory of statistics (Vol.2, 4th ed). London: Charles Griffin & Co.

Lancaster, H.O. and Hamdan, M.A. (1964). Estimation of the correlation coefficient in contingency tables with possibly nonmetrical characters. Psychometrika, 29, 383-391.

Lawley, D.N. (1940). The estimation of factor analysis by the method of maximum likelihood. Proceedings of the Royal Society of Edinburgh, 60, 64-82.

Lawley, D.N. and Maxwell, M.A. (1971). Factor analysis as a statistical method (2nd ed). London: Butterworths.

Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. In Stouffer, SA et al (eds). Measurement and Prediction. New York: John Wiley & Sons. (Chapter 10).

Lombard, H.L. and Doering, C.R. (1947). Treatment of the fourfold tables by partial association and partial correlation as it relates to public health problems. Biometrics, 3, 123-128.;

Lord, F.M. and Novick, H, (1968). Statistical theories of mental test scores. Reading, Mass: Addison-Wesley. Publishing Co.

MacCallum, R. (1983). A comparison of factor analysis programs in SPSS, BMDP and SAS. Psychometrika, 48, 223-231.

McCullagh, P.J. (1980). Regression models for ordinal data. Journal of the Royal Statistical Society, B, 42, 109-142.

McDonald, R.P. (1969). The common factor analysis of multicategory data. British Journal of Mathematical and Statistical Psychology, 22, 165-175.

McHugh, R.B. (1956). Efficient estimation and local identification in latent class analysis. Psychometrika, 21, 331-347.

McNemar, Q. (1942). On the number of factors. Psychometrika, 7, 9-18.

Mardia, K.V. (1967). Some contributions to contingency-type bivariate distributions. Biometrika, 54, 235-249.

Mardia, K.V. (1970). Families of bivariate distributions. London: Charles Griffin & Co.

Martin, J.V. and McDonald, R.P. (1975). Bayesian estimation in unrestricted factor analysis: A treatment for Heywood cases. Psychometrika, 40, 505-517.

Maxwell, E.A. (1954). An analytical calculus for school and university. Vol.II Cambridge, England: Cambridge University Press.

Mosteller, F. (1968). Association and estimation in contingency tables. Journal of the American Statistical Association, 63, 1-28.

Mulaik, S.A. (1986). Factor analysis and Psychometrika: Major developments. Psychometrika, 51, 23-33.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. Psychometrika, 43, 551-560.

Muthén, B. (1981). Factor analysis of dichotomous variables: American attitudes toward abortion. In Jackson, D.J. and Borgatta, E.F. (eds): Factor analysis and measurement in sociological research. London: Sage Publications.

Muthén, B. (1983). Latent variable structural equation modeling with categorical data. Journal of Econometrics, 22, 43-65..

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. Psychometrika, 49, 115-132.

Muthén, B. and Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. Psychometrika, 46, 407-419.

Muthén, B. and Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of nonnormal Likert variables. British Journal of Mathematical and Statistical Psychology, 38, 171-189.

Nias, D.K.B. (1973). Measurement and structure of children's attitudes. In Wilson, 6.D (ed). The Psychology of conservatism. London: Academic Press.

Norušis, M.J. (1985). SPSS-X: Advanced statistics guide. New York: McGraw Hill.

Ogborn, J. (1984). Cumulative and local odds-ratio measures of uniformity of association in two-way ordered contingency tables. British Journal of Mathematical and Statistical Psychology, 37, 89-99.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika, 44, 443-459.

Pearson, K. (1901). I - Mathematical contributions to the theory of evolution. VII - On the correlation of characters not quantitatively measurable. Philosophical Transactions of the Royal Society, 195, 1-47.

Pearson, K. (1913). Note on the surface of constant association. Biometrika, 9, 534-537.

Pearson, K. and Heron, D. (1913). On theories of association. Biometrika, 9, 159-315.

Plackett, R.L. (1965). A class of bivariate distributions, Journal of the American Statistical Association, 60, 516-522.

Rao, C.R. (1955). Estimation and tests of significance in factor analysis. Psychometrika, 20, 93-111.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement. Psychometrika, 34, 71-100.

SAS Institute (1985). SAS User's guide: Statistics, Version 5 edition. Cary, NC: SAS Institute Inc.

Seber, G.A.F. (1984). Multivariate observations. New York: John Wiley & Sons.

Shea, B.L. (1984). FACONE User guide: A computer program for fitting the logit latent variable model by maximum likelihood. London: : LSE, Department of Statistics.

Slater, P. (1947). The factor analysis of a matrix of 2×2 tables. Journal of the Royal Statistical Society, IX (Sup), 114-127.

SPSS Inc. (1986). SPSS-X User's guide (2nd ed). New York: McGraw Hill.

Steiger, J.H.; Shapiro, A and Browne, M.W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. Psychometrika, 50, 253-264.

Tallis, G.M. (1962). The maximum likelihood estimation of correlation from contingency tables. Biometrics, 18, 342-353.

Thorndike, R.M. (1978). Correlational procedures for research. New York: Gardner Press, Inc.

Thurstone, L.L. (1947). Multiple factor analysis. Chicago: University of Chicago Press.

Thurstone, L.L. (1955). A method of factoring without communalities. 1954 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service, 59-62, 64-66.

Tumura, Y. and Fukutomi, K. (1970). On the improper solutions in factor analysis. TRU Mathematics, 6, 63-71.

Tumura, Y.; Fukutomi, K. and Asoo, Y. (1968). On the unique convergence of iterative procedures in factor analysis. TRU Mathematics, 4, 52-59.

Van Driel, O.P. (1978). On various causes of improper solutions in maximum likelihood factor analysis. Psychometrika, 43, 225-243

Wahrendorf, J. (1980). Inference in contingency tables with ordered categories using Plackett's coefficient of association for bivariate distributions. Biometrika, 67, 15-21.

Weinreich, M. (1982). An introduction to the Rasch model and the SAS procedure RACHIT. Kopenhagen: Universitets Statistiske Institut. Computer Programs No.1.

Wilson, G.D. (ed) (1973). The Psychology of conservatism. London: Academic Press.

Wilson, G.D. and Patterson, J.R. (1968). A new measure of conservatism. British Journal of Social Clinical Psychology, 7, 264-269.