Factor Analysis for Categorical Data

By D. J. Bartholomew

London School of Economics and Political Science

[Read before the Royal Statistical Society at a meeting organized by the Research Section on Wednesday, May 21st, 1980, Professor P. Whittle in the Chair]

Summary

The method of factor analysis is widely used as an exploratory tool to reduce the dimensionality of multivariate data. The fact that the standard model is strictly applicable only when the manifest variables are scaled is a serious limitation in social science where the variables are often categorical. In this paper we aim to provide a theoretical framework within which methods for the factor analysis of categorical data can be devised and compared. Discussion is restricted to the case of ordered categories where the latent variables are continuous. It is argued that the choice of model should be made from a restricted set which includes two existing models as special cases. A new method is proposed together with a simple approximate technique of fitting for the one-factor model. The paper concludes with an evaluation of existing methods and makes some suggestions about the direction which future research should take.

Keywords: Factor Analysis; Latent Structure Analysis; Multivariate Analysis; Categorical Data; Multi-Dimensional Contingency Tables; Data Reduction; Scaling; Ordinal Data

1. The Background

An important object of much multivariate analysis is to reduce the dimensionality of the data. This is particularly desirable in the exploratory stages of an investigation both to provide an intelligible summary and to suggest fruitful lines for model building. When the variables are continuous and measured on a common scale, principal component analysis often serves this purpose. Factor analysis achieves much the same end by setting up a model in which the observed variables are related to a smaller set of latent variables and to an “error”. Neither of these methods is directly applicable to categorical variables yet the need for data reduction techniques in such circumstances is no less pressing. This is particularly true in the social sciences where much of the data arising is categorical. The aim of this paper is to provide a framework for the development of methods for use when all the variables are measured on an ordered categorical scale. In the process we shall show how earlier approaches for dichotomous variables arise as special cases of our general formulation.

We assume that we have a simple random sample of size $N$ whose members are cross-classified on $p$ categorical ordered variables. Ordering is very common but it can always be achieved, at the loss of some information, by reducing each dimension to a dichotomy. The sample data can be set out in a multi-way contingency table whose cell frequencies have a joint multinomial distribution. There is no requirement that any marginal frequencies be fixed.

Our aim is to determine whether the $p$-variate representation of the original contingency table can be replaced, without significant loss of information, by one in a smaller number of dimensions. We shall argue that the theoretical framework for this to be done already exists in latent structure analysis, of which normal factor analysis is a special case.

Latent structure analysis has received little attention from statisticians. It appears to have originated with Lazarsfeld, a sociologist, and is expounded in Lazarsfeld and Henry (1968). A more recent discussion of some aspects from a statistical point of view is contained in Goodman.
(1978) and a useful introduction is provided by Fielding (1977). Sociologists have used latent structure analysis as a tool for investigating and measuring attributes. The raw data in such cases consist of the responses by individuals to questions designed to elicit the attitude under investigation. The interest then lies in whether the multidimensional responses are consistent with the existence of one (or more) underlying attitude scales. Similar problems are encountered by psychologists concerned with abilities rather than attitudes. A key statistical reference is Lord and Novick (1968). Some of this work shares the theoretical foundation of latent structure analysis; Lazarsfeld and Henry (1968) point out that Guttman scaling arises as a special case of their model.

Although these two fields of applications share a common mathematical structure their ways of proceeding are somewhat different. When scaling abilities one starts from the supposition that an ability (such as general intelligence) exists and is capable of being measured on a numerical scale. Response variables are therefore selected which are believed to be related to the underlying dimension of ability. Similar conditions may also apply with attitudes but in sociological enquiries it is more usual to proceed in the reverse direction. That is the responses are given and the aim is to discover whether there is evidence of one or more underlying dimensions which could account for the response pattern. This is the path usually followed in principal component analysis and exploratory factor analysis and is the one we shall adopt here.

Latent variable models have not found wide favour with statisticians. Partly this is because of the tedious and somewhat arbitrary methods used for fitting the models. This difficulty has been largely overcome by using computers to implement efficient estimation procedures. More serious, have been doubts about the apparent subjectiveness and arbitrariness of the methods used for interpreting the results. Possibly, this is because the substantive and technical aspects of the methods are so closely intertwined that only an expert in the applied field can deploy them effectively. Latent structure models have certainly arisen to meet real needs in psychology and sociology and there is a growing interest among economists (see, for example, Aigner and Goldberger, 1977). I would argue that such models are implicit in most qualitative analyses of social phenomena and that it is the business of statisticians to make their nature explicit and, as far as possible, quantitative. Attention by statisticians is therefore overdue but in the present paper our aim is more modest. It is to take the relatively neglected area of categorical data and consider how best to carry out what may reasonably be described as factor analysis.

This is not the first attempt to treat the factor analysis of categorical data. Some earlier attempts have aimed, by some means or other, to bring the problem within the framework of the standard normal theory, common factor model. Muthén (1978) is the latest in a group of papers including Bock and Lieberman (1970) and Christofferson (1975) which deal with the case where all variables are dichotomous. In essence, they do this by supposing that the $2^p$ contingency table arises from grouping each dimension of a $p$-variate multinormal distribution into two categories. The underlying variables are then assumed to have the linear structure of the factor model. This approach is very closely related to that of Lazarsfeld and Henry (1968) and Lord and Novick (1968). All of these models arise as special cases of our general approach.

McDonald (1969) proposed a method for analysing multi-category data and he also reviewed much of the earlier work. Like us he made the latent structure model his starting point. McDonald's method does not utilize the ordering of the categories; neither does that of Bock (1972), based on a multivariate logistic model.

In this paper the aim has been to approach the problem from first principles and the emphasis is on fundamentals rather than computational techniques. Much more remains to be done especially on the computational side but the results achieved so far are encouraging.

2. THE MATHEMATICAL FRAMEWORK
2.1. Terminology and Notation

We begin by setting the problem in its general context and then introduce categorical data as a special case. The variables which we observe will be called manifest variables and are denoted
by \( x = (x_1, x_2, \ldots, x_p)^T \). A latent structure model supposes these variables to be related to a set of \( q \) unobservable latent variables denoted by \( y = (y_1, y_2, \ldots, y_q)^T \). For the model to be practically useful \( q \) needs to be much smaller than \( p \). The relationship between \( x \) and \( y \) is stochastic and may be expressed by a conditional probability function \( \pi(x \mid y) \) being the distribution of \( x \) given \( y \). This will be a density or probability mass according as \( x \) is continuous or categorical. The problem is to infer something about \( y \) from the observed values of \( x \). Let \( p(y) \) denote the joint distribution of the \( y \)'s and \( f(x) \) that of the \( x \)'s then the two are related by

\[
f(x) = \int_{\mathcal{R}} \pi(x \mid y) p(y) \, dy,
\]

where \( \mathcal{R} \) is the range space of the latent variables. After \( x \) has been observed our knowledge about \( y \) is given by

\[
p(y \mid x) = \frac{p(y) \pi(x \mid y)}{f(x)}.
\]

The data reduction we are seeking is thus achieved from the fact that the distribution of \( y \) is of smaller dimension than that of \( x \). In practice we may well be content with some suitable summary measure of the conditional distribution of \( y \) such as \( E(y \mid x) \).

In the case of the multi-way contingency table, \( x \) will identify a cell of the table and \( f(x) \) will be its multinomial probability. We shall label the categories along each dimension by 0, 1, 2, \ldots, 0 being the "lowest" level, 1 the next and so on. Thus, for example, the designation (0, 2, 1, 3) refers to the cell where the first variable at level 0, the second at level 2, the third at level 1 and the fourth at level 3. A distinctive feature of our model is that the latent variables are continuous; \( p(y) \) and \( p(y \mid x) \) are thus densities. This choice is based on the fact that most latent variables which arise in social science discourse are thought of as being continuous. For example, quality of life, standard of living, political hue and aggressiveness are all regarded in this way. The case where the latent variables are better treated as categorical may be handled by latent class analysis, for which see Goodman (1978).

2.2. Assumptions

Little progress can be made without some assumptions about the various functions which we have defined. The first assumption we make is that the \( y \)'s are independent, that is that

\[
p(y) = \prod_{i=1}^{q} p(y_i).
\]

There is no completely compelling reason for this assumption but it makes the analysis easier to carry out and interpret. It therefore seems reasonable to adopt it until practical considerations dictate otherwise. The second assumption is about the form of \( p(y_i) \). We shall argue below that this distribution is essentially arbitrary and that the choice may be made to suit our convenience. For this reason we have made it uniform on (0, 1). The justification for this assertion requires us to look more closely at the nature of a latent variable. There seem to be two distinct cases as follows:

(a) The latent variable may be "real" in the sense that it could, in principle, be measured directly. An example would be some sensitive quantity like personal wealth. To avoid asking the direct question we might ask a battery of questions about possessions and life-style in the hope that they might enable us to identify and scale the underlying variable-wealth in this case. The distribution of wealth is certainly not arbitrary and it would be quite inappropriate to assume that it was uniform. Such cases seem quite rare. More commonly we have the second case.

(b) The latent variable is not "real" meaning that it could not be measured directly, even in principle. It is a mental construct used to facilitate economy of thought. Attitudes and abilities largely come into this category.
Since there is no "natural" scale in such cases we are at liberty to construct one to suit our convenience. Since ordering is the highest level of measurement available in the manifest variables it seems reasonable to ask for no more than an ordinal level of measurement of the latent variables also. Such a scale is arbitrary to the extent that any monotonic transformation of the chosen scale would serve equally well. Thus whatever the distribution of the latent variable on a chosen scale it can always be given a desired distribution, such as the uniform, by an appropriate monotonic change of scale.

The remaining function to be specified is \( \pi(y \mid x) \). We make two assumptions about this. The crucial assumption, which is fundamental to the rationale of the method, is that of conditional independence. We assume that

\[
\pi(x \mid y) = \prod_{i=1}^{p} \pi(x_i \mid y).
\]

This means that the observed dependence among the x's is wholly explained by their dependence on the y's. Eliminating variation in the latter removes the inter-dependence of the x's. In that sense the association among the x's is fully explained by their dependence on the latent variables. This gives formal expression to the hypothesis that the observed variables are describable in terms of a smaller number of latent dimensions. If (3) were not true it would imply that there was some other variable exerting a common influence on the x's.

Under the assumptions made so far (1) becomes

\[
f(x) = \int_{0}^{1} \cdots \int_{0}^{1} \prod_{i=1}^{p} \pi(x_i \mid y) \, dy = E \prod_{i=1}^{p} \pi(x_i \mid y). \tag{4}
\]

The choice of the form of the response function \( \pi_i(x_i \mid y) \), sometimes called the trace function, is the final step in the specification of the model.

3. THE CHOICE OF RESPONSE FUNCTION

For the application to contingency tables we shall suppose, initially, that each variable is dichotomous. This requirement will be relaxed in Section 7. In this case we may write

\[
\pi_i(x_i \mid y) = \{\pi_i(y)\}^{x_i} \{1 - \pi_i(y)^{1-x_i}\}, \quad (x_i = 0, 1); \tag{5}
\]

\( \pi_i(y) \) is thus the conditional probability of a response at the "upper" level on the ith manifest variable (also spoken of as a positive response).

We begin by listing some properties which the function \( \pi_i(y) \) should possess and then consider whether functions exist which meet the specification. Let \( \mathcal{F} \) denote the family of acceptable functions then we claim that \( \mathcal{F} \) should possess the following properties:

(i) \( 0 \leq \pi(y) \leq 1 \) for \( 0 \leq y_i \leq 1 \) \( (i = 1, 2, \ldots, q) \) for all \( \pi(y) \in \mathcal{F} \).

(ii) \( \pi(y) \) should be monotonic non-increasing or non-decreasing in each argument.

(iii) If \( \pi(y) \in \mathcal{F} \) then the function obtained by replacing any sub-set of the elements of y by their complements should also belong to \( \mathcal{F} \).

(iv) If \( \pi(y) \in \mathcal{F} \) then \( 1 - \pi(y) \in \mathcal{F} \).

(v) \( \pi(y) = \pi \in \mathcal{F} \).

(vi) If \( q = 1, \pi(y) = 0, 0 \leq y \leq y_0 \)

\[ = 1, \ y_0 < y \leq 1 \] \( \in \mathcal{F} \) for all \( y_0(0 < y_0 < 1) \).

Less formally we may add

(vii) \( \pi(y) \) should be capable of describing a wide variety of shapes.

(viii) The number of parameters in \( \pi(y) \) should be small.

(ix) The expectations of (4) should be easy to evaluate.

Condition (i) merely reflects the fact that \( \pi(y) \) is a probability; ideally, we would not wish to impose artificial constraints on the parameters to ensure this. Condition (ii) is not essential but
most latent variables considered in practice are such that the probability of response increases, or decreases, in step with changes in the variable. Condition (iii) reflects the arbitrariness of the direction in which the latent variable is measured. For example, we can equally well measure the political spectrum from left to right or vice versa. Condition (iv) is important and arises from the arbitrariness in the direction of the ordering of the categories. Answering "yes" to a question is the same as answering "no" to its negation. This condition ensures that the outcome of the analysis does not depend on which choice we make. Conditions (v) and (vi) ensure that two special cases are included; (v) is the case of complete independence when no reduction in dimensionality is possible; (vi) which is less important, is the case of a perfect scale (Guttman scale). All those with \( y \) above \( y_0 \) respond positively and all those below negatively.

There is no difficulty about finding functions which satisfy (i)–(vi) but most do not meet (vii)–(ix) and there is a natural conflict between (vii) and (viii). It is not, of course, possible to express \( \pi_i(y) \) as a linear function of its parameters without violating (i). Instead we shall consider the class of functions given by

\[
G\{\pi_i(y)\} = \alpha_{i0} + \sum_{j=1}^q \alpha_{ij} H(y_j), \quad (i = 1, 2, ..., p). \tag{6}
\]

The coefficients \( \{\alpha_{ij}\} \) can thus be interpreted in the usual way as factor loadings. We must select the functions \( G \) and \( H \) to meet the conditions set out above. If we choose \( H \) so that \( H(y_j) = -H(1 - y_j) \) condition (iii) is satisfied, the effect being to change the sign of the coefficient when the direction of measurement of \( y_j \) is changed. Conditions (i) and (iv) imply that \( G^{-1} \) must have the form of the distribution function of a random variable distributed symmetrically about zero. This is equivalent to requiring that \( G(v) = -G(1 - v) \) so that both \( G \) and \( H \) must be selected from the same class of functions.

In practice the choice is very limited, the commonly used functions being the logit (logit \( v = \log \frac{v}{1 - v} \)) and the probit (probit \( v = \Phi^{-1}(v) \) where \( \Phi \) is the standard normal distribution function). Another possibility is the inverse Cauchy distribution function; the simplest choice is \( H(v) = G(v) = v - \frac{1}{2} \) but this violates condition (i). The complementary log log function is ruled out by conditions (iv) and (iii). Lord and Novick (1968), who discussed the case \( q = 1 \), used the logit for \( G \) and the probit for \( H \). Bock and Lieberman (1970) used the probit in both cases. We shall argue that there are good reasons for preferring the logit function for both \( G \) and \( H \). Our choice may therefore be expressed as

\[
 \text{logit} \, \pi_i(y) = \log \{\pi_i(y)/(1 - \pi_i(y))\}
\]

\[
= \log \pi_i/(1 - \pi_i) + \sum_{j=1}^q \alpha_{ij} \log y_j/(1 - y_j) \tag{7a}
\]

or

\[
\pi_i(y) = \pi_i \prod_{j=1}^q y_j^{\alpha_{ij}} \left\{ \prod_{j=1}^q y_j^{\alpha_{ij}} + (1 - \pi_i) \prod_{j=1}^q (1 - y_j)^{\alpha_{ij}} \right\} \tag{7b}
\]

\[
(0 < \pi_i \leq 1; -\infty < \alpha_{ij} < +\infty, \ i = 1, 2, ..., p, \ j = 1, 2, ..., q).
\]

We shall refer to this as the logit model. If we transform to a new scale for the latent variables given by \( u_j = \log y_j/(1 - y_j) \) then

\[
\pi_i(u) = \pi_i \left\{ \pi_i + (1 - \pi_i) \exp \left( -\sum_{j=1}^q \alpha_{ij} u_j \right) \right\} \tag{7c}
\]

If the \( y \)'s are uniform the \( u \)'s have a logistic distribution with mean 0 and variance \( \sigma^2 = \pi^2/3 \). This transformation emphasizes the point already made about the arbitrariness in the choice of \( p(y) \) and \( \pi(y) \); the choice is, in fact, a joint one. The model is incapable of distinguishing between
having $y_i$ uniform with $\pi_i(y)$ as in (7b) and $u_j$ logistic with $\pi_i(u)$ as in (7c). Any attempt to interpret particular forms of either function in physical terms is therefore incapable of receiving empirical support. It may easily be verified that the functions given by (6b) satisfy conditions (i)–(v) but (vi) only holds for $y_0 = \frac{1}{2}$ in the limit as $\alpha \to \infty$.

The parameter $\pi_i$ has a direct and useful interpretation. It is the value of $\pi_i(y)$ when $y_j = \frac{1}{2}$ for all $j$ and hence is the probability of a positive response for an individual at the median position on each latent dimension. In that sense, therefore, it is the typical probability of response on the $i$th dimension.

4. Expected Frequencies, $y$-Scores and the Fit of the Model

In order to fit the model we have to evaluate the expectations given by (4). For the $2^p$ table this involves the determination of integrals of the form

$$\int_0^1 \cdots \int_0^1 \pi_1(y) \cdots \pi_1(y) \cdots dy,$$

where the integrand contains $p$ factors. Explicit expressions can be found when $q = 1$ for $\alpha = \frac{1}{2}, 1$ and 2. However, it is a straightforward matter to evaluate (8) numerically and this is the method we have used in the examples which follow. The computational time increases rapidly with $q$.

As in factor analysis we may wish to go further and derive the analogue of "factor scores". This may be thought of as finding a mapping of the cells of the contingency table into $q$-dimensional Euclidean space. We approach this problem via the conditional distribution of $y$ given $x$. For the $2^p$ table this has the form

$$p(y | x) = \prod_{i=1}^p \{ \pi_i(y) \}^{x_i} \{ 1 - \pi_i(y) \}^{1-x_i} / f(x).$$

This tells us how $y$ is distributed given $x$. No single value of $y$ is associated with a given $x$ but we may reasonably take some measure of location of the distribution as a typical value of $y$ for that $x$. Since $y$ is uniformly distributed, the element $y_r$ in $y$ may be interpreted as the quantile of the distribution on the $r$th latent dimension on which the individual stands. The expectation $E(y_r | x)$ is thus the expected proportion of the population "below" an individual with manifest vector $x$.

We shall define this to be the $y$-score of the individual on dimension $r$. It requires the evaluation of integrals of the form

$$\int_0^1 \cdots \int_0^1 y_r \pi_1(y) \pi_1(y) \cdots dy$$

which may be obtained numerically.

Having fitted the model it is useful to have some measure of how successful we have been. In principal component analysis we do this by computing the proportion of the total variation which is accounted for by each of the components. In a multi-way contingency table a similar measure may be based on measures of goodness of fit. This could be chi-squared but we have preferred to use

$$\Lambda = 2 \sum_i O_i \ln O_i / E_i$$

because it is a linear transform of the log-likelihood; $O_i$ and $E_i$ are the observed and expected frequencies and the summation is taken over all cells of the table. As a base-line for the comparison we take the value of $\Lambda$ when the $E$'s are calculated on the assumption of complete independence. This is a measure of the total departure from independence which we hope the model will explain; denote it by $\Lambda_0$. Let $\Lambda_q$ be the same quantity when the $E$'s are those obtained by fitting a model with $q$ latent variables then the ratio $(\Lambda_0 - \Lambda_q)/\Lambda_0$ is a measure of how much the original departure from independence is accounted for by the model fitted. If the parameters
are fitted by an efficient method $\Lambda$ will have, approximately, a $\chi^2$-distribution with degrees of freedom $(2^p - \text{number of parameters} - 1)$ and the goodness of fit may be judged by this means.

5. SOME BASIC RESULTS

Since we wish to reduce the dimensionality of the data as much as possible, a natural way to proceed is to take values of $q$ in order, beginning with $q = 1$ and stop as soon as a good enough fit is obtained. However, increasing $q$ increases the number of parameters to be fitted and there comes a point where the model is under-identified. As the literature on factor analysis testifies the question of identifiability has subtle features and the same is true of our model.

The case $q = 1$ presents the least number of problems. It is useful, therefore, to have the following theorem which helps to show whether a one-factor model has any prospect of fitting the data. Let us denote by $R_{ij}$ the cross-product ratio formed from the expected frequencies when the table is collapsed over all dimensions except $i$ and $j$. That is

$$R_{ij} = \frac{E\pi_i(y)\pi_j(y)E(1 - \pi_i(y))(1 - \pi_j(y))}{E\pi_i(y)(1 - \pi_j(y))E\pi_j(y)(1 - \pi_i(y))} \quad (12)$$

($R_{ij}$ will play a key role later on).

**Theorem 1.** If $\pi_i(y)$ and $\pi_j(y)$ are both monotonic non-increasing or non-decreasing then $R_{ij} - 1 > 0$, otherwise $R_{ij} - 1 \leq 0$ with equality only if at least one of $\pi_i(y)$ and $\pi_j(y)$ are constant.

**Proof.** $R_{ij} - 1 = \{E\pi_i(y)\pi_j(y) - E\pi_i(y)E\pi_j(y)\}/E\pi_i(y)(1 - \pi_j(y))E\pi_j(y)(1 - \pi_i(y))$ hence the sign of $R_{ij} - 1$ is the same as that of $d_{ij} = E\pi_i(y)\pi_j(y) - E\pi_i(y)E\pi_j(y)$. Now

$$d_{ij} = \int_0^1 \pi_i(y)\{\pi_j(y) - E\pi_j(y)\} dy.$$  

Suppose first that $\pi_j(y)$ is monotonic decreasing, then we can find $y = y^*$ such that $\pi_j(y) \geq E\pi_j(y)$ for $y \geq y^*$ and $\pi_j(y) < E\pi_j(y)$ for $y < y^*$ so that $d_{ij}$ may be written

$$d_{ij} = \int_0^{y^*} \pi_i(y)\{\pi_j(y) - E\pi_j(y)\} dy + \int_{y^*}^1 \pi_i(y)\{\pi_j(y) - E\pi_j(y)\} dy.$$  

If $\pi_i(y)$ is also monotonic non-decreasing

$$d_{ij} \geq \pi_i(y^*)\int_0^{y^*} \{\pi_j(y) - E\pi_j(y)\} dy + \pi_i(y^*)\int_{y^*}^1 \{\pi_j(y) - E\pi_j(y)\} dy = 0.$$  

If both functions are monotonic non-increasing a similar argument leads to the same conclusion; otherwise the inequality is reversed. Equality obviously occurs only when one or both functions are constant.

The practical relevance of this theorem is as follows. If we reverse the order of the categories on dimension $i$ the probability of a positive response will be $1 - \pi_i(y)$ instead of $\pi_i(y)$. If $\pi_i(y)$ was formerly decreasing the corresponding probability for that dimension with the categories reversed will be increasing. In the one-factor case, therefore, it must be possible to re-order the dimensions so that the response functions either all increase or all decrease. If this is done all the $(R_{ij} - 1)$'s will be positive. For a one-factor model to be appropriate it is thus necessary (but not sufficient) that an ordering of the manifest dimensions exists such that all the cross-product ratios are greater than one. Suppose, for example, that a table with $p = 4$ has $R_{12} > 1$, $R_{13}$, $R_{14}$, $R_{23}$, $R_{24} < 1$, $R_{34} > 1$. Reversing the categories on a dimension changes the sign of $R_{ij} - 1$. In this case reversing the order on dimensions 1 and 2 will produce a set of positive values.

In practice, of course, the $R_{ij}$'s have to be estimated from the sample cross-product ratios and so the signs cannot be determined with certainty. Nevertheless, if all or most of the $(R_{ij} - 1)$'s can be made positive it is worth trying a one-factor model.
The following theorem holds for all members of the chosen family of response functions. It provides the basis for the approach to estimation proposed in the next Section and it also serves to display a feature which links various methods of factor analysis together.

**Theorem 2.**

\[
E \pi_i(y) \pi_j(y) - E \pi_i(y) E \pi_j(y) = \tau^2 G^{-1}(a_{i0}) G^{-1}(a_{j0}) \sum_{k=1}^{q} \alpha_{ik} \alpha_{jk} + \text{terms of the 4th degree in } \alpha_k \text{ and } \alpha_j,
\]

where \( \tau^2 = EH^2(y) \) \( (i, j = 1, 2, ..., p; i \neq j) \).

The proof is based on a straightforward Taylor expansion of the response function and term by term integration. The left-hand side is the predicted covariance between \( x_i \) and \( x_j \) and the theorem shows that this has a simple form if the departure from complete independence, as measured by the \( \alpha \)'s, is small. The covariances in the normal theory factor model have the form \( \sum_{k=1}^{q} \alpha_{ik} \alpha_{jk} \) and this suggests we should look for sample functions which are estimates of the quantities

\[
E \frac{\pi_i(y) \pi_j(y) - E \pi_i(y) E \pi_j(y)}{\tau^2 G^{-1}(a_{i0}) G^{-1}(a_{j0})} \quad (i \neq j).
\]

(13) 

If such can be found they would have the same structure (for small \( \alpha \)'s) as in the normal case and hence known methods of estimation could be used. If we take the logit form for \( G \) and \( H \) it turns out that

\[
\frac{R_{ij} - 1}{\sigma^2} = \sum_{k=1}^{q} \alpha_{ik} \alpha_{jk} + \text{terms of 4th degree},
\]

(14)

where \( \sigma^2 = E \log^2 \{ y/(1 - y) \} = 3.289,868 \).

The sample cross-product ratio can be used to estimate \( R_{ij} \) and hence the \( \alpha \)'s via (14). Since cross-product ratios are the "natural" measures of association in \( 2^p \) contingency tables it is satisfying to find them arising in this context thus supporting the choice of the logit function.

If the probit functions are chosen then (13) is a first approximation to the tetrachoric correlation coefficients which may be taken as partly justifying the heuristic method which carries out a normal factor analysis on these coefficients.

6. Fitting the One-Factor Model Logit to the \( 2^p \) Table

6.1. Methods

Since the logit and probit functions are very similar one would expect both versions of the general model we have considered to give similar results and to involve about the same amount of calculation. The logit function is easier to compute than the probit but this advantage is likely to be fairly marginal. Bock and Lieberman (1970) developed a maximum likelihood method for the probit model and illustrated it on two examples with \( p = 5 \) and \( q = 1 \). The method involved extensive numerical integration and they suggested that it would not be feasible for \( p \) in excess of 10 or 12. Christofferson (1975) found a faster method using a least squares fit of \( E \pi_i(y) \) and \( E \pi_i(y) \pi_j(y) \) \( (i, j = 1, 2, ..., p) \) to their sample estimates. Muthén (1978) has made further improvements on this method by substantially reducing the amount of numerical integration required. It appears from this work that little information is lost by using only the first and second order margins for estimation. Programs could be provided for the logit model using the same methods and they would presumably involve similar amounts of computation.

However, the logit model has an important property which often makes it possible to obtain a simple approximate solution when \( q = 1 \). This solution also offers a good starting point for an iterative procedure by which it can be improved. The basis of the method rests on the fact that the approximation given by (14) is remarkably good even when the \( \alpha \)'s are far beyond the range
when they can be described as "small". Table 1 gives values of \((R_{ij} - 1)/\alpha_i \alpha_j \sigma^2\) for various combinations of \((\pi_i, \pi_j)\) and \((\alpha_i, \alpha_j)\). The approximation is to be judged by the closeness of the ratios to 1 (the second subscript on \(\alpha\) has been dropped).

### Table 1

Values of \(c_{ij} = (R_{ij} - 1)/\alpha_i \alpha_j \sigma^2\). The entries are unchanged if \((\pi_i, \pi_j)\) is replaced by \((1 - \pi_i, 1 - \pi_j)\) and \((\alpha_i, \alpha_j)\) by \((\alpha_j, \alpha_i)\).

<table>
<thead>
<tr>
<th>((\pi_i, \pi_j))</th>
<th>((1, 1))</th>
<th>((1, 2))</th>
<th>((1, \frac{10}{3}))</th>
<th>((\frac{10}{3}, \frac{10}{3}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>((2, 2))</td>
<td>0.942</td>
<td>1.192</td>
<td>0.984</td>
<td>1.119</td>
</tr>
<tr>
<td>((2, 1))</td>
<td>0.801</td>
<td>0.944</td>
<td>0.850</td>
<td>1.008</td>
</tr>
<tr>
<td>((2, \frac{10}{3}))</td>
<td>0.614</td>
<td>0.668</td>
<td>0.644</td>
<td>0.731</td>
</tr>
<tr>
<td>((1, 1))</td>
<td>0.912</td>
<td>1.063</td>
<td>0.988</td>
<td>1.245</td>
</tr>
<tr>
<td>((1, \frac{10}{3}))</td>
<td>0.846</td>
<td>0.934</td>
<td>0.912</td>
<td>1.125</td>
</tr>
<tr>
<td>((\frac{10}{3}, 1))</td>
<td>0.935</td>
<td>1.011</td>
<td>1.015</td>
<td>1.263</td>
</tr>
<tr>
<td>((\frac{10}{3}, \frac{10}{3}))</td>
<td>0.917</td>
<td>0.965</td>
<td>0.977</td>
<td>1.139</td>
</tr>
<tr>
<td>((\frac{10}{3}, \frac{10}{3}))</td>
<td>0.971</td>
<td>1.001</td>
<td>1.016</td>
<td>1.119</td>
</tr>
</tbody>
</table>

The worst cases occur when \(\alpha_i\) and \(\alpha_j\) are far apart and when \(\pi_i\) and \(\pi_j\) are small. Only positive values of \(\alpha_i\) and \(\alpha_j\) have been considered for reasons which will emerge below.

\(R_{ij} - 1\) is not the only function of the expectations which has \(\alpha_i \alpha_j \sigma^2\) as the first term of its expansion. The same is true, for example, of \(\ln R_{ij}\). Unfortunately the approximation is much less good for this, and other functions, which we have investigated.

The basis of our method of estimation is to find estimates \(\alpha\) and \(\pi\) such that (a) the cross-product ratios for the model are as close as possible to those observed and (b) the marginal proportions of the model and the data agree exactly. The \(\alpha\)'s are found iteratively using the result of (14) as a starting point. We proceed by the following steps.

1. Find a vector \(\alpha\) such that \(\alpha_i \alpha_j\) is as close as possible to the estimated values of \(c_{ij} = (R_{ij} - 1)/\sigma^2\) for \(i, j = 1, 2, \ldots, p; i \neq j\).
2. Find \(\pi\) by equating \(E \pi_i(y)\) (\(i = 1, 2, \ldots, p\)) to the corresponding marginal proportion using the vector \(\alpha\) obtained in step 1.
3. Improve the estimate of \(\alpha\) by a method to be described.
4. Re-estimate \(\pi\) using the improved \(\alpha\).
5. Repeat the cycle until \(\pi\) and \(\alpha\) (or \(\Lambda\)) converge.

If many cycles of the iteration are required the amount of numerical integration required will prevent the method being used on a routine basis with present-day computers if \(p\) is, say, greater than 10. However, we shall show that the first approximation is often quite adequate for practical purposes. This can be obtained rapidly on a computer with no practical limit on \(p\). The calculation of the expected frequencies from the estimated parameters does require numerical integration (with all methods) and this may take a considerable time for large \(p\).

The method requires \(p \geq 3\); if \(p > 3\) it is not possible to reproduce the \(c_{ij}\)'s exactly so we must find an \(\alpha\) such that the distance between the \(c_{ij}\)'s and the \(\alpha_i \alpha_j\)'s is as small, in some sense, as possible. Precisely this problem arises in normal theory factor analysis in which context the \(c_{ij}\)'s are covariances and the \(\alpha\)'s are factor loadings. One solution, also applicable here, is based on minimizing

\[
\sum_{i=1}^{p} \sum_{j=1}^{p} (\hat{c}_{ij} - \alpha_i \alpha_j)^2.
\]
It is an iterative procedure known as the “minres” method and will be found, for example, in Harman (1970).

An alternative method which is both intuitively appealing and easy to apply is as follows. We shall call it the row and column method.

Consider the matrix with elements \( \{ a_{ij} \} \) \((i, j = 1, 2, \ldots, p)\). This matrix has the property that
\[
(i, j)\text{th element} = (\text{Row } i \text{ total}) \times (\text{Column } j \text{ total})/\text{Grand total}.
\]
(15)

If we regard the \( \hat{c}_{ij} \)'s as estimates of the off-diagonal elements we can treat the estimation problem as one of finding diagonal elements for this matrix such that (15) holds. We can, in fact, ensure that (15) holds for every diagonal element. The \( \alpha \)'s we seek will therefore satisfy the equations
\[
\alpha_i^2 = \left\{ \hat{C}_i + \sum_{j=1}^{p} \alpha_j^2 \right\} (i = 1, 2, \ldots, p),
\]
(16)

where
\[
\hat{C}_i = \sum_{j=1, j \neq i}^{p} \hat{c}_{ij}, \quad \hat{C} = \sum_{i=1}^{p} \hat{C}_i.
\]

These equations are equivalent to
\[
\alpha_i \left( \sum_{j=1}^{p} \alpha_j - \alpha_i \right) = \hat{C}_i \quad (i = 1, 2, \ldots, p).
\]
(17)

They have considerable appeal in their own right as a means of estimation since they result from equating the off-diagonal row totals to their observed values. Their solution is simplest if the \( \alpha \)'s are all of the same sign. On this question we have the following lemma:

**Lemma.** The two \( \alpha \)-vectors which satisfy (17) have elements of the same sign if and only if \( \hat{C}_i \geq 0 \) for all \( i \).

Before proceeding with the method of solution it is therefore necessary to ensure that \( \hat{C}_i \geq 0 \) \((i = 1, 2, \ldots, p)\) by changing the order of categories, if necessary.

Writing \( A = \sum_{i=1}^{p} \alpha_i \), (17) becomes
\[
\alpha_i (A - \alpha_i) = \hat{C}_i \quad \text{or} \quad \alpha_i^2 - A \alpha_i + \hat{C}_i = 0
\]
(18)

which yields
\[
\alpha_i = \frac{1}{2} A \pm \frac{1}{2} (A^2 - 4\hat{C}_i)^{\frac{1}{2}} \quad (i = 1, 2, \ldots, p).
\]
(19)

If we sum both sides of this expression over \( i \) we shall obtain an equation for \( A \). Once this is solved (19) will provide estimates of \( \alpha_i \) \((i = 1, 2, \ldots, p)\).

There is an ambiguity of sign involved in (19) which leads to two alternative equations for \( A \). \( A \) is the real root, where it exists, of
\[
p - 2 = \sum_{i=1}^{p} (1 - 4\hat{C}_i/A^2)^{\frac{1}{2}}.
\]
(20)

Otherwise \( A \) is the real root of
\[
p - 2 = \sum_{i=1}^{p-1} (1 - 4\hat{C}_i/A^2)^{\frac{1}{2}} - (1 - 4\hat{C}_p/A^2)^{\frac{1}{2}}.
\]
(21)

It is not immediately obvious in what sense this procedure minimizes the distance between the \( \hat{c}_{ij} \)'s and the products \( \alpha_i \alpha_j \). This becomes apparent when we observe that the same estimating
equations result from maximizing

\[ \phi = \sum_{i=1}^{p} \sum_{j=1 \neq j}^{p} \hat{c}_{ij} \ln \left( \frac{\alpha_i \alpha_j}{\sum_{i=1}^{p} \sum_{j=1 \neq j}^{p} \alpha_i \alpha_j} \right) \]  

(22)

with respect to the \( \alpha \)'s. The greatest possible value of \( \phi \) is \( \sum \Sigma_{i=1}^{p} \hat{c}_{ij} \log(c_{ij}/\hat{C}) \) so by making (22) as near to this value as we can we are achieving the best fit in a certain sense. For the solution of (17) to be a maximum it is necessary for \( \hat{C}_{ij} \geq 0 \) \( (i = 1, 2, ..., p) \). This also ensures that the \( \alpha \)'s will all have the same sign and thus that the argument of the logarithm in (22) is positive.

Having obtained \( \alpha \) we next estimate the \( \pi_i \)'s from

\[ \frac{N_i}{N} = \int_0^1 \frac{\pi_i y^{\alpha_i}}{\pi_i y^{\alpha_i} + (1 - \pi_i)(1 - y)^\alpha} dy, \quad (i = 1, 2, ..., p), \]  

(23)

where \( N_i \) is the number of positive responses on dimension \( i \). These equations may be solved iteratively by the usual Newton–Raphson method using \( \pi_i = N_i/N \) as a starting value.

The method so far rests on the supposition that \( c_{ij} = \alpha_i \alpha_j \). This is only an approximation so we next write \( c_{ij} = \alpha_i \alpha_j \theta_{ij} \) where \( \theta_{ij} \) depends weakly on \( \pi_i, \pi_j, \alpha_i \), and \( \alpha_j \) but will usually be close to 1. Using the estimates of \( \alpha \) and \( \pi \) already obtained we next estimate \( \theta_{ij} \) by

\[ \hat{\theta}_{ij} = \hat{c}_{ij}(\hat{\pi}_i, \hat{\pi}_j, \hat{\alpha}_i, \hat{\alpha}_j)/\hat{\alpha}_i \hat{\alpha}_j \quad (i, j = 1, 2, ..., p; \ i \neq j), \]  

(24)

\( \hat{c}_{ij}(\hat{\pi}_i, \hat{\pi}_j, \hat{\alpha}_i, \hat{\alpha}_j) \) being the value of \( c_{ij} \) when the parameters are given the same values as the preliminary estimates. The cycle of estimation is now repeated by replacing the starting values \( \hat{c}_{ij} \) by \( \hat{c}_{ij}/\hat{\theta}_{ij} \). It can happen, as one of the examples below shows, that the estimates of \( \alpha \) do not appear to converge at all. The possibility of this is apparent from the fact that there is nothing in the model to prevent one or more \( \alpha \)'s being infinite. In such a case an iterative procedure starting from finite values may never terminate. This feature is not as serious as it might seem. When an \( \alpha \) is large, greater than 2 say, big changes in \( \alpha \) produce only small changes in the shape of \( \pi_i(y) \) and hence in the overall fit of the model. From the point of view of interpretation all that matters is that the \( \alpha \) in question is “large”. In practice, therefore, no difficulty arises if we stop the iteration as soon as no worth-while improvement in the fit is obtained.

In most cases we have investigated, where the \( \alpha \)'s turn out to be small and of the same order of magnitude, convergence of the parameter estimates is rapid. This is especially true of \( \pi \).

### 6.2. Examples

To illustrate the use of the one-factor logit model and to compare it with the probit method we shall give the results of fitting the model to seven sets of data. Two of these were used by Bock and Lieberman (1970), Christoffersson (1975) and Muthén (1978). They relate to 1000 cases on each of Sections VI and VII of the Law School Admission Test (LSAT). Background details and the original data are in Bock and Lieberman (1970). The results of fitting the logit and probit models are given in Table 2. For the logit we give the first approximation obtained from (14) and the final estimates after iteration. For the probit model we give Bock and Lieberman’s maximum likelihood estimates, Muthén’s generalized least squares (GLS) estimates and Muthén’s unweighted least squares estimators. The latter are obtained by doing a standard factor analysis on the tetrachoric correlations obtained from the table.

We have re-parameterized the probit model to conform with (6) as explained later. For Section VI the fit by all methods is excellent with \( \Lambda \) almost equal to its expectation. The differences between the various parameter estimates are negligible and would have no effect on the interpretation of the factor. On these grounds there is therefore nothing to choose between
the logit and probit models. In the case of Section VII the fit is less good and there is greater variation in the estimates but, again, these are not sufficient to affect the interpretation of the analysis.

**Table 2**

Comparison of parameter estimates and goodness of fit for the LSAT data using the probit and logit models

<table>
<thead>
<tr>
<th></th>
<th>Logit</th>
<th>Probit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First approximation</td>
<td>Final estimate</td>
</tr>
<tr>
<td>Section VI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>0.460</td>
<td>0.410</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>0.431</td>
<td>0.424</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>0.516</td>
<td>0.538</td>
</tr>
<tr>
<td>$\alpha_4$</td>
<td>0.401</td>
<td>0.391</td>
</tr>
<tr>
<td>$\alpha_5$</td>
<td>0.373</td>
<td>0.351</td>
</tr>
<tr>
<td>$\pi_1$</td>
<td>0.941</td>
<td>0.938</td>
</tr>
<tr>
<td>$\pi_2$</td>
<td>0.731</td>
<td>0.730</td>
</tr>
<tr>
<td>$\pi_3$</td>
<td>0.562</td>
<td>0.563</td>
</tr>
<tr>
<td>$\pi_4$</td>
<td>0.785</td>
<td>0.784</td>
</tr>
<tr>
<td>$\pi_5$</td>
<td>0.887</td>
<td>0.885</td>
</tr>
<tr>
<td>$\Lambda$</td>
<td>21.24</td>
<td>21.17</td>
</tr>
<tr>
<td>Section VII</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>0.663</td>
<td>0.604</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>0.574</td>
<td>0.581</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>0.898</td>
<td>0.907</td>
</tr>
<tr>
<td>$\alpha_4$</td>
<td>0.455</td>
<td>0.465</td>
</tr>
<tr>
<td>$\alpha_5$</td>
<td>0.444</td>
<td>0.420</td>
</tr>
<tr>
<td>$\pi_1$</td>
<td>0.876</td>
<td>0.870</td>
</tr>
<tr>
<td>$\pi_2$</td>
<td>0.687</td>
<td>0.688</td>
</tr>
<tr>
<td>$\pi_3$</td>
<td>0.848</td>
<td>0.849</td>
</tr>
<tr>
<td>$\pi_4$</td>
<td>0.620</td>
<td>0.621</td>
</tr>
<tr>
<td>$\pi_5$</td>
<td>0.868</td>
<td>0.865</td>
</tr>
<tr>
<td>$\Lambda$</td>
<td>33.1</td>
<td>32.21</td>
</tr>
</tbody>
</table>

From the computational point of view the “logit first approximation” is the simplest and can easily be carried out with a pocket calculator for problems of this size. In that case the integration in the solution of (23) can be avoided by using a normal approximation given by

$$\hat{\pi}_i = \Phi\left( (1 + \beta^2_i) \frac{1}{N} \Phi^{-1}(N_i/N) \right). \quad (25)$$

The computing effort required for the final estimates depends on how many cycles of the iteration are necessary and this, in turn, depends on the accuracy required. No exact comparisons have yet been made with the various probit methods but it seems likely to be faster than the maximum likelihood method but slower than Muthén’s GLS method.

The computation of the expected frequencies and the y-scores involves calculations of the same order for both models though here the logit has the slight advantage of being easier to calculate than the probit.

In social, as opposed to psychometric and educational applications, $p$ is often quite small and what is wanted is a simple method of extracting one or two factors and a way of providing a scale of measurement for the latent variables. We therefore give five further examples of this kind
in which our main aim will be to see how good the first approximation is and to illustrate problems which may arise in the fitting process. The sets of data used are as follows.

Set I is taken from Lombard and Doering (1947) and it relates to knowledge about cancer. A sample of 1729 individuals were classified on four dimensions concerning sources of general knowledge each having two categories as follows: (1) Radio/no radio; (2) Newspapers/no newspapers; (3) Solid reading/no solid reading; (4) Lectures/no lectures. A fifth variable was whether or not the respondent had a good knowledge of cancer. Here we shall look only at the first four variables to see whether there is evidence of a single latent variable which, we might anticipate, would have to do with how well informed people are in general.

Sets II and III are from Solomon (1961) and concern attitudes to science expressed by 2982 young people. They were divided into two equal groups on the basis of their IQ (High = II, Low = III). Attitudes were elicited in the form of positive or negative responses to four questions. The data and the questions are reproduced in Plackett (1974) which also contains Set I.

Set IV is a $2^4$ table from Upton (1978) where the questions from a survey on entry to the EEC were such as might be expected to relate to a latent political left/right variable.

Set V is taken from a study of mobility of the elderly and it is included here as an example which gives rise to problems in fitting.

### Table 3

**Parameter estimates and goodness of fit of the logit model for the five cases described in the text**

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th></th>
<th>II</th>
<th></th>
<th>III†</th>
<th></th>
<th>IV</th>
<th></th>
<th>V</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First</td>
<td>Final</td>
<td>First</td>
<td>Final</td>
<td>First</td>
<td>Final</td>
<td>First</td>
<td>Final</td>
<td>First</td>
<td>Final</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>0.444</td>
<td>0.445</td>
<td>0.169</td>
<td>0.195</td>
<td>0.168</td>
<td>0.164</td>
<td>0.962</td>
<td>0.986</td>
<td>2.757</td>
<td>1.695</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>1.228</td>
<td>1.550</td>
<td>0.448</td>
<td>0.400</td>
<td>0.097</td>
<td>0.161</td>
<td>0.351</td>
<td>0.397</td>
<td>2.682</td>
<td>2.382</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.864</td>
<td>0.860</td>
<td>0.818</td>
<td>1.068</td>
<td>1.143</td>
<td>15.225</td>
<td>0.546</td>
<td>0.571</td>
<td>0.275</td>
<td>0.457</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.506</td>
<td>0.456</td>
<td>0.217</td>
<td>0.223</td>
<td>0.242</td>
<td>0.168</td>
<td>0.998</td>
<td>1.074</td>
<td>0.386</td>
<td>0.594</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>0.213</td>
<td>0.212</td>
<td>0.818</td>
<td>0.819</td>
<td>0.839</td>
<td>0.839</td>
<td>0.704</td>
<td>0.707</td>
<td>0.008</td>
<td>0.037</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>0.604</td>
<td>0.620</td>
<td>0.174</td>
<td>0.178</td>
<td>0.169</td>
<td>0.167</td>
<td>0.454</td>
<td>0.453</td>
<td>0.526</td>
<td>0.524</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>0.461</td>
<td>0.464</td>
<td>0.646</td>
<td>0.664</td>
<td>0.526</td>
<td>0.732</td>
<td>0.469</td>
<td>0.467</td>
<td>0.237</td>
<td>0.221</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>0.057</td>
<td>0.061</td>
<td>0.543</td>
<td>0.543</td>
<td>0.446</td>
<td>0.448</td>
<td>0.703</td>
<td>0.709</td>
<td>0.411</td>
<td>0.402</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>0.389</td>
<td>0.387</td>
<td>0.237</td>
<td>0.221</td>
<td>0.469</td>
<td>0.467</td>
<td>0.237</td>
<td>0.221</td>
<td>0.411</td>
<td>0.402</td>
</tr>
<tr>
<td>$\Lambda$</td>
<td>23.71</td>
<td>19.15</td>
<td>11.80</td>
<td>11.10</td>
<td>17.03</td>
<td>12.92</td>
<td>89.83</td>
<td>90.40</td>
<td>33.31</td>
<td>21.62</td>
</tr>
<tr>
<td>Degrees of freedom</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>21</td>
<td>7</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

† After 15 cycles.

In all cases, except II, the value of $\Lambda$ suggests that a further factor might be involved but the reduction in $\Lambda$ from the case of complete independence is always very substantial. The first approximation is good on the whole though there are several marked discrepancies to which we turn in a moment.

In case III the estimates for $\alpha$ were not converging and the iteration was stopped after 15 cycles. At this point the value of $\Lambda$ had virtually converged. In other words, the fit at this stage was hardly affected by changing the parameters. The extreme case, $\alpha = \infty$, is equivalent to what is called a Heywood case in factor analysis where there is no uncertainty on the response once
the value of the latent variable is fixed. It is this kind of response function which is assumed in Guttman scaling and there seems to be no reason for regarding it as anomalous in our model. In this case our first approximation is less good but it still shows the same basic pattern.

Case IV is interesting in that although the fit of the one-factor model is poor the first approximation actually provides a slightly better fit than the full iteration.

Case V arises from a $2^4$ table where one cross-product ratio was very large (27.3) whereas the others were in the range 2–4. Here it took 26 iterations to converge and there was a marked oscillation from one cycle to the next in the early stages. Even here, the first approximation gives the broad outline of the solution and the estimates of $\pi$ are particularly good. We have other examples with very large cross-product ratios, including one for $p = 7$, where the iteration appears to diverge with many elements of $\pi$ tending to zero. This appears to be because some of the $\pi_i(y)$'s have the Guttman form with $y_0$ (see condition (vi) of Section 3) near 0 or 1. Neither the logit nor the probit model can adequately cope with that situation.

A full analysis requires the calculation of expected frequencies and $y$-scores. These are given for Case I in Table 4.

### Table 4

Fit of the one-factor model to Lombard and Doering’s data on cancer knowledge

<table>
<thead>
<tr>
<th>Cell</th>
<th>Observed frequency</th>
<th>Independence frequency†</th>
<th>Fitted frequency</th>
<th>$y$-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000</td>
<td>477</td>
<td>279.1</td>
<td>466.5</td>
<td>0.212</td>
</tr>
<tr>
<td>0001</td>
<td>12</td>
<td>23.6</td>
<td>16.1</td>
<td>0.304</td>
</tr>
<tr>
<td>0010</td>
<td>150</td>
<td>251.8</td>
<td>156.4</td>
<td>0.384</td>
</tr>
<tr>
<td>0011</td>
<td>11</td>
<td>21.3</td>
<td>8.6</td>
<td>0.475</td>
</tr>
<tr>
<td>0100</td>
<td>231</td>
<td>359.2</td>
<td>250.1</td>
<td>0.522</td>
</tr>
<tr>
<td>0101</td>
<td>13</td>
<td>30.4</td>
<td>18.9</td>
<td>0.615</td>
</tr>
<tr>
<td>0110</td>
<td>378</td>
<td>324.1</td>
<td>355.5</td>
<td>0.700</td>
</tr>
<tr>
<td>0111</td>
<td>45</td>
<td>27.4</td>
<td>44.0</td>
<td>0.797</td>
</tr>
<tr>
<td>1000</td>
<td>63</td>
<td>87.3</td>
<td>67.1</td>
<td>0.304</td>
</tr>
<tr>
<td>1001</td>
<td>7</td>
<td>7.4</td>
<td>3.0</td>
<td>0.394</td>
</tr>
<tr>
<td>1010</td>
<td>32</td>
<td>78.8</td>
<td>35.8</td>
<td>0.475</td>
</tr>
<tr>
<td>1011</td>
<td>4</td>
<td>6.7</td>
<td>2.4</td>
<td>0.567</td>
</tr>
<tr>
<td>1100</td>
<td>94</td>
<td>112.4</td>
<td>78.8</td>
<td>0.615</td>
</tr>
<tr>
<td>1101</td>
<td>12</td>
<td>9.5</td>
<td>7.5</td>
<td>0.711</td>
</tr>
<tr>
<td>1110</td>
<td>169</td>
<td>101.4</td>
<td>182.9</td>
<td>0.796</td>
</tr>
<tr>
<td>1111</td>
<td>31</td>
<td>8.6</td>
<td>34.4</td>
<td>0.889</td>
</tr>
<tr>
<td>Total</td>
<td>1729</td>
<td>1729.0</td>
<td>1729.0</td>
<td>—</td>
</tr>
<tr>
<td>$\Lambda$</td>
<td>—</td>
<td>380.57</td>
<td>19.15</td>
<td>—</td>
</tr>
</tbody>
</table>

† These are the expected frequencies on the assumption of complete independence between the manifest variables.

Although the one-factor model is barely adequate it is a great improvement over the “independence” fit. The $y$-scores provide a useful ranking of the cells according to the degree of knowledge which the cell members exhibit.

7. Manifest Variables with more than Two Categories

7.1. Specification of the Response Function

When there are more than two categories on any dimension we need a response function to specify the probability of falling into each category. We do this by defining the cumulative
probability

\[
\pi_{is}(y) = Pr \{\text{an individual with latent position } y \text{ falls in category } s \text{ or higher}\},
\]
\[
\pi_{i,0}(y) = 1, \quad \pi_{i,c}(y) = 0,
\]

(26)

where \(c\) is the number of categories. The probability that an individual falls into category \(s\) on the \(i\)th dimension is thus

\[
\pi_{i,s}(y) - \pi_{i,s+1}(y).
\]

(27)

We suppose that the cumulative function has the same logit form as we used for the dichotomy, with parameters \(\pi_{is}\) and \(\alpha_i\). The \(\alpha\)'s do not depend on \(s\). They measure the departure from independence and, as such, should be independent of the number of categories. This fact also ensures that the differences (27) are non-negative. The same approach could be used with the probit or any other response function of the form given by (6).

7.2. Fitting the Model

The approximate method already given for the \(2^p\) table can easily be extended to the general case. At the cost of neglecting some information, the \(\alpha\)'s can be estimated using the method of Section 6 by reducing each dimension to a dichotomy. This should be done so as to make the marginal frequencies as nearly equal as possible. More efficient methods would, of course, be desirable. Once the \(\alpha\)'s have been found the full set of \(\pi_{is}\)'s can be estimated by equating each marginal cumulative frequency to its expectation.

If the number of manifest variables and categories is at all large the computation of the expected frequencies and \(y\)-scores is a substantial undertaking, even for the \(2^p\) table. It is thus desirable to have a systematic approach to the calculations and this is provided by the result of Theorem 3. The statement of the theorem requires some further definitions and terminology.

Suppose that dimension \(i\) has \(r_i+1\) categories which, in accordance with our earlier convention, are denoted by the levels \((0, 1, 2, ..., r_i)\). Each cell of the table is identified by a sequence of levels which can be generated symbolically by forming a Kronecker (direct) product of the vectors of levels. Thus, for example, if we had two dimensions with levels \((0, 1, 2)\) and \((0, 1, 2, 3)\) the listing of the cells is

\[
\begin{bmatrix}
0 \\
1 \\
2
\end{bmatrix}
\times
\begin{bmatrix}
0 \\
1 \\
2 \\
3
\end{bmatrix}
= 
\begin{bmatrix}
00 \\
01 \\
02 \\
03 \\
10 \\
11 \\
12 \\
13 \\
20 \\
21 \\
22 \\
23
\end{bmatrix}
\]

(28)

The order of the cells depends on the order in which we take the vectors. This is arbitrary but once selected it must be adhered to throughout. It is convenient to rank the vectors in increasing order of \(r_i\), as in the above example, and to multiply out from the right-hand end.
In general, for a \( p \)-dimensional table, the listing of the cells is given by forming the product

\[
\begin{bmatrix}
0 \\
1 \\
2 \\
\vdots \\
r_1
\end{bmatrix} \times \begin{bmatrix}
0 \\
1 \\
2 \\
\vdots \\
r_2
\end{bmatrix} \times \cdots \times \begin{bmatrix}
0 \\
1 \\
2 \\
\vdots \\
r_p
\end{bmatrix}
\]

An \( E \) in front of such an expression means that the expectation is to be taken of all cell frequencies designated by the product.

Finally, we define the \((r+1) \times (r+1)\) matrix \( A_r \) as follows:

\[
A_r = \begin{bmatrix}
1 & -1 & 0 & 0 & \cdots & 0 \\
0 & 1 & -1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \cdots & \cdots & -1 \\
0 & 0 & \cdots & \cdots & \cdots & 1
\end{bmatrix}
\]

This matrix forms the first differences of the elements of any column vector which it pre-multiplies. In particular, if it pre-multiplies a vector of cumulative probabilities such as given by (26) it yields the category probabilities of (27).

The theorem is as follows:

**Theorem 3.** The expected cell frequencies can be computed from the formula

\[
E \begin{bmatrix}
0 \\
1 \\
\vdots \\
r_1
\end{bmatrix} \times \begin{bmatrix}
0 \\
1 \\
\vdots \\
r_2
\end{bmatrix} \times \cdots \times \begin{bmatrix}
0 \\
1 \\
\vdots \\
r_p
\end{bmatrix} = N[A_{r_1} \times A_{r_2} \times \cdots \times A_{r_p}]
\]

\[
E \begin{bmatrix}
\pi_{11}(y) \\
\pi_{12}(y) \\
\vdots \\
\pi_{1r_1}(y)
\end{bmatrix} \times \begin{bmatrix}
\pi_{21}(y) \\
\pi_{22}(y) \\
\vdots \\
\pi_{2r_2}(y)
\end{bmatrix} \times \cdots \times \begin{bmatrix}
\pi_{p1}(y) \\
\pi_{p2}(y) \\
\vdots \\
\pi_{pr_p}(y)
\end{bmatrix}
\]

(Note that the absence of the "\( \times \)" sign before \( E \) on the right-hand side implies matrix multiplication of the standard kind.)
Proof. The marginal probabilities that an individual falls into the various categories on the $i$th dimension, given $y$, are given by

$$A_i = \begin{bmatrix} 1 \\ \pi_{11}(y) \\ \pi_{12}(y) \\ \vdots \\ \pi_{1r_i}(y) \end{bmatrix}$$

(29)

Thus the probability of falling into any cell of the table is obtained by multiplying together the relevant marginal probabilities for that cell (since the events are independent, given $y$). This is achieved for all cells together by forming the Kronecker product of the vectors (29) taken over $i$ in the same order as in the statement of the theorem. Using the standard result that

$$(A \times B \times \ldots) = (A \times B \times \ldots)(x \times y \times \ldots)$$

the result then follows on taking expectations with respect to $y$.

The theorem identifies all the expectations which have to be evaluated. In all there are $(r_1 + 1)(r_2 + 1) \ldots (r_p + 1) - 1$ integrals to be calculated. The vector of expectations is then converted into one of expected frequencies by pre-multiplying by the Kronecker product of the differencing matrices. The theorem includes, as a special case, the $2^p$ table obtained by setting $r_i = 1$ for all $i$.

The $y$-score on dimension $s$ for cell $x$ is given by

$$E(y_s | x) = \int_0^1 \ldots \int_0^1 y_s p(x | y) dy = N \int_0^1 \ldots \int_0^1 y_s p(x | y) dy / N \int_0^1 \ldots \int_0^1 p(x | y) dy.$$  

(30)

Theorem 3 provides a formula for the denominator of this expression. The numerator can be found by multiplying the Kronecker product after the $E$ by $y_s$ and then evaluating the expression as before.

As an illustration consider the $2 \times 3$ case used as an example at the beginning of the section:

$$E \begin{bmatrix} 1 \\ \pi_{11}(y) \end{bmatrix} \times \begin{bmatrix} 1 \\ \pi_{21}(y) \\ \pi_{22}(y) \end{bmatrix} = E \begin{bmatrix} 1 \\ \pi_{21}(y) \\ \pi_{22}(y) \\ \pi_{11}(y) \\ \pi_{11}(y) \pi_{21}(y) \\ \pi_{11}(y) \pi_{22}(y) \end{bmatrix}$$

and each element of this must be calculated by numerical integration. The resulting vector is now pre-multiplied by

$$A_1 \times A_2 = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 1 \end{bmatrix} \times \begin{bmatrix} 1 \\ -1 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 0 \\ -1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ -1 \\ 1 \\ 0 \\ 0 \\ 1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

to give the expected cell frequencies.
8. EXTENSIONS AND EVALUATION

8.1. More than one Latent Variable

The methods of fitting the one-factor logit model extend in a natural way to the case of several factors. Theorem 2 holds for all \( q \) and hence any method for fitting the normal factor model to the off-diagonal elements of a covariance matrix can be used to provide approximate estimates for the \( \alpha \)'s. The new method given in Section 6.1 can also be used in the following way. First we fit the parameters \( \alpha_{0.1} = (\alpha_{2.1}, \ldots, \alpha_{p.1}) \) as already described. Next we construct the residuals \( \{\hat{c}_{ij} - \hat{d}_{ij}, \hat{d}_{ij}\} \). These will include negative quantities so signs must be changed to render the row totals positive. A second set of parameters \( \alpha_{0.2} \) is then fitted to the residuals and \( \pi \) is re-estimated. For example, in the case of Upton's (1978) data (Case IV, Table 3) the resulting estimates are

\[
\hat{\alpha}_{0.1} = (0.962, 0.351, 0.546, 0.998, 0.493), \\
\hat{\alpha}_{0.2} = (-0.518, 0.241, -0.518, 0.520, 0.158), \\
\hat{\pi} = (0.704, 0.454, 0.469, 0.703, 0.389).
\]

The estimate of \( \pi \) is very close to that for the one-factor model; \( \Lambda \) has been reduced from 89.83 to 61.47 which is still a poor fit. An iterative procedure could now be used to improve the estimates but the feasibility of this remains to be explored. Calculations of \( c_{ij} \) for two-factor models similar to those given in Table 1 suggest that the approximation given by Theorem 2 is more restricted in its usefulness than when \( q = 1 \). Further investigation is required.

With more than two factors the probit model offers computational advantages. This is because the expectations \( \{E\pi_i(y)\pi_j(y)\} \) can always be reduced to bivariate integrals however many latent variables there are. In fact it may easily be shown that

\[
E\pi_i(y) = \Phi(\lambda_{i.0}), \\
E\pi_i(y)\pi_j(y) = \int_{-\infty}^{\lambda_0} \int_{-\infty}^{\lambda_0} \phi(z_1, z_2; \rho_{ij}) dz_1 dz_2 \quad (i \neq j),
\]

where \( \phi \) is the standard bivariate normal density with correlation coefficient

\[
\rho_{ij} = \sum_{k=1}^{q} \lambda_{ik} \lambda_{jk},
\]

where \( \lambda_{ik} = \alpha_{ik}/(1 + \sum_{h=1}^{q} \alpha_{ih})^{1/2} \) (\( k = 0, 1, 2, \ldots, p \)). The various estimation methods proposed by Bock and Lieberman (1970), Christofferson (1975) and Muthén (1978) for this model are based essentially on (31) and (32). Their estimates of \( \{\lambda_{ik}\} \) can easily be converted into estimates for \( \{\alpha_{ik}\} \) which, in turn, would be good approximations to the corresponding parameters of the logit model.

From the point of view of interpretation and for reasons given below we prefer to use the logit parameterization for which

\[
\pi_i = G^{-1}(\alpha_{i.0}) = G^{-1}\left(\lambda_{i.0}/\left(1 - \sum_{h=1}^{q} \lambda_{ih}^2\right)^{1/2}\right), \\
\alpha_{ih} = \lambda_{ih}/\left(1 - \sum_{h=1}^{q} \lambda_{ih}^2\right)^{1/2}.
\]

8.2. Comparisons with other Methods

It is not unusual to find factor analyses carried out on covariance matrices regardless of the form of the distribution of the manifest variables. For the \( 2^p \) table it is perfectly possible to treat the indicator variables \( \{x_i\} \) just like any others and to perform a factor analysis on the estimated
covariance matrix. Such an analysis can be roughly interpreted in terms of our model, when the \( \alpha_k \)'s \((k = 1, 2, ..., q)\) are small, as follows. From Theorem 2, a factor model fitted to the off-diagonal elements will estimate the quantities \( \{ \tau G^{-1}(\alpha_o) \alpha_k \} \) \((k = 1, 2, ..., q)\). The variance of \( x_i \) is \( G^{-1}(\alpha_o) \{ 1 - G^{-1}(\alpha_o) \} + O(x^2) \). For the logit function \( G(v) = G(v) \{ 1 - G(v) \} \) and hence the sample variances estimate \( G^{-1}(\alpha_o) \) and therefore the \( \alpha \)'s are determined. It is doubtful whether such a procedure has any practical value.

We have already noted that the logit and probit models are likely to give similar numerical results. At the conceptual level there are considerable advantages in developing both models by means of the latent structure arguments used here. The traditional “factor analysis” approach assumes that there are two tiers of latent variables. First there is supposed to be a latent variable underlying each dichotomy; a positive response is then observed if that variable exceeds a threshold value. Secondly, these variables are related to the second tier of latent variables by the usual common factor model. This may be plausible in some applications but with dichotomies based on house ownership, trade union membership and such like, the notion of an underlying latent variable and its associated threshold is somewhat artificial. When we add to this the argument that the forms of the distributions of the latent variables are essentially arbitrary the usual model appears as no more than a convenient fiction. It is for these reasons that we prefer the parameterization in (34) which has a more robust interpretation.

Similar considerations apply to the hybrid model of Lord and Novick (1968) in which \( G \) is a logit and \( H \) a probit function. Samathanan and Blumenthal (1978) have given a maximum likelihood method for estimating its parameters similar to the EM algorithm. There is clearly room for further study of the numerical aspects of all models in the light of current, and the likely future, state of computer technology.

It is unfortunate that no other suitable response function has come to light for which the various integrals have simple explicit forms. If we are prepared to abandon the symmetry conditions we could consider such functions as

\[
\pi_i(y) = y^a \quad \text{or} \quad \pi_i(y) = 1 - (1 - y)^a
\]

for \( q = 1 \). These models can be fitted very easily but, with only one parameter, they are not sufficiently flexible for most purposes. Introducing further parameters destroys their simplicity. The approximate method of fitting our logit model seems to come nearest to combining simplicity and flexibility. Whether or not it is good enough to be generally useful for \( q > 1 \) requires further investigation. In the meantime the various methods for the probit model are available.

ACKNOWLEDGEMENTS

The approach on which this paper is based was first outlined in a paper read at the Society's conference at Oxford in March 1979. I am grateful to several participants for suggestions and especially to Dr J. A. Anderson whose remarks led to a major change of direction. The suggestions of referees and other readers of an earlier version have also led to many improvements. The method of fitting the logit model in Section 6 has been programmed in FORTRAN by J. Tomenson to whom I owe special debt.

REFERENCES


Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29–51.


DISCUSSION OF PROFESSOR BARTHOLOMEW’S PAPER

Professor MURRAY AITKIN (University of Lancaster): I am pleased to propose the vote of thanks for David Bartholomew’s paper. The subject of latent variable models is of rapidly increasing practical importance and unifies a number of apparently unconnected statistical areas. Professor Bartholomew notes in Section 1 that latent variable models have not found wide favour with statisticians, partly because of the difficulty of fitting the models, and that attention to such models by statisticians is overdue. His paper tonight takes an important step towards developing proper latent variable models for categorical data.

The basis of the models is set out in Section 2. An early distinction is made between continuous and discrete latent variables, though the conditional independence model has been used with both types of latent variable. Given the properties (i)–(vi) in section 3, the choice of response function comes down essentially to a probit/logit choice for both the manifest variables and the latent variables. Bartholomew chooses the logit/logit model, for reasons which are not entirely clear. Sections 4 and 5 discuss in detail the fitting of the model. Here the cross-ratios play an important role. The LSAT example discussed shows that the computing method for the logit model gives estimates consistent with the ML estimates for the probit model of Bock and Lieberman. The method is practicable, though it is not clear that it gives efficient estimates.

Latent variable models are natural candidates for ML estimation by the EM algorithm (Dempster, Laird and Rubin 1977). The logit model is unsuitable for EM, but the probit model is very suitable, as the sufficient statistics in the “complete data” model are just the usual regression sums of squares and cross-products. DLR pointed out that ML estimation in the normal factor model could be achieved by EM using simple back-and-forward least squares computations, and John Hinde has developed at Lancaster a GENSTAT macro for exploratory factor analysis using EM. Hasselblad, Stead and Creason, in a note to appear in Biometrics, point out that the standard probit analysis model for a dose-response curve can be fitted by EM. The combination of these two approaches allows the estimation of the parameters in the probit/probit latent variable model by ML using an EM algorithm. I am currently completing joint work with Darrell Bock on this procedure.

Professor Bartholomew notes that categorical latent variables may be treated by latent class analysis. While continuous latent variables are more natural choices for abilities and attitudes, it is of some interest that the simple latent class model also provides a scaling of the cell entries on a scale which is essentially continuous. This is of practical value because the computations for fitting the latent class model are the same as those for a general mixture model, and can easily be done in GLIM using an EM algorithm. The cancer knowledge example (Set 1) provides a simple illustration.

The latent class model used is a two-component multinomial mixture. There are two classes of people: well-informed, and badly informed. In Professor Bartholomew’s notation, y in (1) is a Bernoulli variable, with \( P(y = 1) = \lambda, \) \( P(y = 0) = 1 - \lambda. \) The response function of (5) is then just

\[
\pi(y | x) = \pi y^x (1 - \pi) 1 - x,
\]
the $y$ suffix indicating that there are two sets of $\pi_i$ in the two latent classes. The conditional probability function of $y$ given $x$ in (2) is then

$$P(y = 1 \mid x) = \frac{\lambda \pi(x \mid y = 1)}{\lambda \pi(x \mid y = 1) + (1 - \lambda) \pi(x \mid y = 0)},$$

a monotone function of the likelihood ratio for the two components, and similarly for $P(y = 0 \mid x)$.

The EM algorithm begins with starting values for the probabilities of latent class membership, most simply by assigning each cell to one of the two classes. Parameter estimates are then obtained in the M-step from the conditional independence model. These are substituted into the likelihood ratio to give new probabilities of class membership in the E-step. The sequence of steps continues till convergence to the ML estimates of the $\pi_{iy}$. This is very simply accomplished in GLIM with a small macro. At convergence, the probabilities of latent class membership have also converged, and these also provide a ranking of cells from "most well informed" to "least well informed." In addition, for the conditional independence model, the log of the ratio of the probabilities of class membership is a linear function of the $x_i$—a linear discriminant function, whose coefficients are the log-odds ratios for the $i$th item. This discriminant function can also be used to scale the individual cells. A small or zero coefficient indicates that the corresponding item does not discriminate between well and poorly informed classes, and can be dropped from the scale.

In the cancer data, the two-class model gives a goodness-of-fit value $\Lambda$ of 15·4, using one extra parameter, so it fits as well as the one-factor model. The discriminant function is

$$1·43x_1 + 3·62x_2 + 2·35x_3 + 1·61x_4$$

assigning most weight to newspapers, next to solid reading, and least to radio and lectures. These coefficients are very similar in relative magnitude to the factor loadings in Table 3, column 2. The discriminant score, and the estimated probability of belonging to the well-informed group, are shown in Table D1, together with the cell code and the estimated factor score from Table 4.

<table>
<thead>
<tr>
<th>Cell</th>
<th>Factor (y) score</th>
<th>Discriminant score</th>
<th>Probability of being well-informed</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000</td>
<td>0·212</td>
<td>0·00</td>
<td>0·029</td>
</tr>
<tr>
<td>0001</td>
<td>0·304</td>
<td>1·61</td>
<td>0·131</td>
</tr>
<tr>
<td>0010</td>
<td>0·384</td>
<td>2·35</td>
<td>0·241</td>
</tr>
<tr>
<td>0011</td>
<td>0·475</td>
<td>3·96</td>
<td>0·613</td>
</tr>
<tr>
<td>0100</td>
<td>0·522</td>
<td>3·62</td>
<td>0·531</td>
</tr>
<tr>
<td>0101</td>
<td>0·615</td>
<td>5·23</td>
<td>0·850</td>
</tr>
<tr>
<td>0110</td>
<td>0·700</td>
<td>5·97</td>
<td>0·922</td>
</tr>
<tr>
<td>0111</td>
<td>0·797</td>
<td>7·58</td>
<td>0·983</td>
</tr>
<tr>
<td>1000</td>
<td>0·304</td>
<td>1·43</td>
<td>0·112</td>
</tr>
<tr>
<td>1001</td>
<td>0·394</td>
<td>3·04</td>
<td>0·386</td>
</tr>
<tr>
<td>1010</td>
<td>0·475</td>
<td>3·78</td>
<td>0·569</td>
</tr>
<tr>
<td>1011</td>
<td>0·567</td>
<td>5·39</td>
<td>0·868</td>
</tr>
<tr>
<td>1100</td>
<td>0·615</td>
<td>5·05</td>
<td>0·825</td>
</tr>
<tr>
<td>1101</td>
<td>0·711</td>
<td>6·66</td>
<td>0·959</td>
</tr>
<tr>
<td>1110</td>
<td>0·796</td>
<td>7·40</td>
<td>0·980</td>
</tr>
<tr>
<td>1111</td>
<td>0·889</td>
<td>9·01</td>
<td>0·996</td>
</tr>
</tbody>
</table>

A plot of the factor score against the discriminant score shows a very nearly linear relationship.

Similar results are obtained for the LSAT data of Bock and Lieberman. For Section 6, the value of $\Lambda$ for the mixture model after 12 iterations is 23·9, with a discriminant function of

$$1·66x_1 + 1·48x_2 + 1·91x_3 + 1·32x_4 + 1·26x_5.$$  

For Section 7, $\Lambda$ is 35·5 after 12 iterations with discriminant function

$$1·76x_1 + 2·02x_2 + 2·67x_3 + 1·48x_4 + 1·37x_5.$$  

The goodness-of-fit of both models can be improved by further iterations, without essentially changing the discriminant function.
I referred at the beginning of these comments to the value of latent variable models in unifying apparently unconnected areas. I should like to conclude with an example: regression on principal components. It is common practice, or at least a commonly advocated practice, to extract principal components from a highly correlated set of predictors, and then regress the response on a suitable subset of the principal variables. A latent variable model—the MIMIC model of Jöreskog and Goldberger (1975)—makes clear what principal component regression is trying to achieve. Underlying the set of predictors \( x \) is a set of latent variables \( z \). Given \( z \), the response \( y \) is independent of \( x \) and depends on \( z \) through a regression model. The \( x \) are conditionally independent given \( z \):

\[
\begin{align*}
y_i | z_i &\sim N_1(\mu + \beta' z_i, \sigma^2) \quad \text{independently} \\
x_i | z_i &\sim N_p(\theta + \Lambda z_i, \Psi) \quad \text{independently of } y_i
\end{align*}
\]

with \( \Psi \) a diagonal matrix, while marginally \( z_i \sim N(O, I) \).

The model can again be fitted by ML using an EM algorithm, and a GENSTAT program for this is under development by John Hinde.

In conclusion, I have much pleasure in proposing the vote of thanks for this stimulating and important paper.

Dr A. M. Skene (University of Nottingham): Professor Bartholomew has described an approach to the analysis of ordinal data which is a very welcome addition to the rather sparse body of theory which exists at present in this area. The full impact must of course await future developments which release some of the computational constraints imposed by the present estimation procedure. The restriction to one or perhaps two latent variables must appear to be a very severe restriction to those accustomed to using normal theory factor analysis. However, accepting this limitation and aware of the arbitrary nature of both \( p(y) \) and \( \pi(x | y) \) I looked to possible uses of this logit latent structure model.

There are two areas of application; modelling and data reduction.

A single continuous latent variable may provide a very good explanation for the pattern observed in a multidimensional contingency table, particularly if there are extra-statistical arguments to support the existence of such a variable. On the other hand, latent variables tend to be mental constructs and thus equally valid arguments can usually be made for a discrete formulation. The flexibility of latent class analysis as described by Goodman (1978) suggests that the discrete formulation should be our starting point, with the logit model being adopted when the latent class analysis reveals classes having a clear ordering.

The logit model effects data reduction by replacing \( x \) by \( E(y | x) \), the factor scores. It follows from the conditional independence of the manifest variables given \( y \), that it is a trivial matter to calculate \( p(y | x^{(1)}) \) where \( x^{(1)} \) is any subvector of \( x \). This, coupled with the fact that parameter estimation only requires knowledge of the one and two way margins of the manifest variables, leads to the observation that the logit model's applicability is unaffected by missing data.

Once \( p(y | x) \) has been obtained however, its mean may be totally inappropriate as a summary measure. Fig. D1 displays two instances of \( p(y | x) \) for the Lombard and Doering data. The problem is that of finding

![Fig. D1. Conditional distributions for Lombard and Doering data. (i) \( p(y | x' = (0,0,0,0)) \). (ii) \( p(y | x' = (0,1,1,0)) \).](attachment:image.png)
suitable summary measures for heavily skewed distributions. The choice of summary statistic is even more complicated when two latent variables are fitted and it is certainly dangerous to make much here of the analogy with normal factor analysis.

Effective data reduction is striking a balance between reducing dimension and retaining that information which is relevant to a specific objective. The real value of factor scores or conditional distributions cannot be judged in the abstract and it is pointless debating the meaning of these quantities. Their ultimate value must be judged by, for example, the accuracy of the final predictive equation or the insights gained into the subject of the analysis.

This point is relevant to all latent structure models and can be illustrated by the following model used for medical diagnosis.

Given diseases \( D_i, i = 1, \ldots, I \) and symptom vector \( S \), one possible formulation of \( p(S \mid D_i) \) is the latent class model

\[
p(S \mid D_i) = \sum_{j=1}^{n} \prod_{k=1}^{K} p_{k}(S_{ik} \mid C_j) p(C_j \mid D_i). \tag{1}
\]

Conditional upon latent class, \( C_j \), we assume that the symptoms are mutually independent and independent of \( D_i \). The parameters of this model, viz. the parameters of \( p_{k}(S_{ik} \mid .) \) and the probabilities \( p(C_j \mid D_i) j = 1, \ldots, n; i = 1, \ldots, I \) can be estimated using training data, Skene (1978), and, given a particular realisation of \( S \), say \( T \), disease probabilities \( p(D_i \mid T) \propto p(T \mid D_i) p(D_i) \) can be computed.

Equation 1 makes absolutely no claim to be a representation of the truth. In any particular application it stands or falls by its ability to correctly diagnose patients.

There is a second way of writing this model. Given \( T \), we may first calculate

\[
p(C_j \mid T) \propto \prod_{k} p_{k}(T_{ik} \mid C_j) p(C_j),
\]

where

\[
p(C_j) = \sum_{i} p(C_j \mid D_i) p(D_i)
\]

and then calculate

\[
p(D_i \mid T) = \sum_{j} p(D_i \mid C_j) p(C_j \mid T).
\]

The probabilities \( p(C_j \mid T) j = 1, \ldots, n \) define a probability distribution over the latent classes and this alone is used in calculating the disease probabilities.

This particular formulation makes the two steps of the classification much clearer. The first step of data reduction is followed by the using of the transformed data. However this formulation is also very seductive as it exposes the latent classes and raises the possibility that they might have real meaning. Such emphasis is, in the main, unwarranted.

Professor Bartholomew, in effect, has described a rather different way of doing this first step of data reduction. The ultimate test of this particular model is whether effective predictions or good understanding of particular data sets result.

I have much pleasure in seconding the vote of thanks.

The vote of thanks was passed by acclamation.

Mr C. J. Skinner (University of Southampton): I should also like to thank Professor Bartholomew for a very interesting paper. I particularly enjoyed the discussions of the response function in Section 3 and I note that certain latent class models may also be included in the general formulation of equation (6), if \( H \) becomes a discrete valued function.

My main comments concern the suggestion in this paper that the logit model is preferable to the numerically similar probit model, and I should like to offer a few words in defence of the probit model. One reason given for preferring the logit model is that Section 6 provides a simple approximate solution (at least when \( q = 1 \)). However, as an iterated analogue of the simple heuristic solution for the probit model, where the estimated \( c_{ij} \)'s correspond to the tetrachoric correlations. In fact, if one attempts to iterate the heuristic method in a corresponding manner, one finds that successive iterations give an identical solution, because, under the probit parameterisation, \( E \pi_j(y) \) does not depend on the factor loadings and the corresponding \( \theta_{ij} \)'s are all unity. One advantage of such a non-iterated two-stage
procedure is that conventional factor analysis packages or more sophisticated correlation structure packages such as LISREL (Jöreskog and Sörbom, 1978) may be used with categorical variables or with combinations of categorical and continuous variables.

Available empirical evidence, as in Table 2, suggests that point estimates obtained by the heuristic procedure are very close to the full maximum likelihood estimates. A supposed problem with the heuristic procedure, as for example stated by Muthén (1978), is that of obtaining a statistical test of model fit. However, a chi-squared goodness of fit test may be obtained by direct analogy to Section 4. of this paper, where the computation of the test statistic requires the evaluation of a number of multivariate normal integrals. Perhaps a more difficult problem with the heuristic method is that of obtaining standard errors. As Professor Aitken has noted, this problem is not referred to in this paper, and it would be interesting to know if the method in Section 6 can be simply adapted to give standard errors.

Finally, I find in teaching this subject that the probit model provides an easily understood modification of normal theory factor analysis, and I find it valuable to demonstrate simple links between contingency table analysis and continuous variable analysis.

Dr G. J. G. Upton (University of Essex): Professor Bartholomew has set up an elegant mathematical superstructure which fills me, at least, with awe and wonder. I will therefore confine my comments to a discussion of the results that he has obtained.

Table 4 includes the $y$-scores for the Lombard and Doering data, which have apparently been obtained by numerical integration of (10). These scores are, however, simply connected to the estimates of the $\alpha$ parameters. Using subscripts $i, j, k$ and $l$, each taking values 0 or 1 corresponding to the cell definition, I note that an excellent fit to the $y$-scores is given by

$$4.895 y_{ijkl} = 1.024 + \alpha_1 i + \alpha_2 j + \alpha_3 k + \alpha_4 l.$$ 

I have been unable to find an explanation for this equation, but the fit is too good to be accidental.

I am unhappy about the scant attention paid by Professor Bartholomew to the interpretation of his results. In particular, I cannot believe that my book has been so widely read that it is unnecessary to define the manifest dimensions for data set IV. These are referendum votes (for or against entry into the Common Market), political allegiance in 1975, amount of schooling, union membership and social class. The order of categories of union membership is the reverse of the order given in my book. The positive $\alpha$-values place the minimally-schooled working-class anti-Common Market Labour union member at the left hand end of the political axis, which is reasonable. However, it is distinctly surprising that the strongest contrasts are those related to referendum vote and union membership rather than manifest political allegiance.

In section 8 Professor Bartholomew fits a second latent factor to the referendum data. For this factor it is the minimally-schooled non-union middle-class anti-Common Market Conservative who is at one end of an axis. Could this be an age dimension? However, I am unhappy about the assumption that one of the two latent dimensions in the two-factor model will of necessity be the dimension found in the one-factor model. I would have felt that, in a case where there were really two dimensions, the single dimension found by fitting the one-factor model is more likely to be an over-worked hybrid lying between the two. My hypothesis could be tested by creating a data set which was derived, without random variation, from two latent variables, and then fitting the single latent variable model.

Mr G. J. A. Stern (I.C.L.): Many branches of science seem to follow the course of cosmological theories suggested in the lines, which consist of a couplet by Pope and a modern sequel:

Nature, and Nature's laws lay hid in night,  
God said, "Let Newton be," and all was light.  
It did not last, the Devil shouting "Ho!"  
Let Einstein be!" restored the status quo.

Darkness is followed by a clarifying theory, which is followed by worse darkness in the shape of sophistication leading to incomprehensibility and possibly meaninglessness, so that clarification follows an upside-down U-curve.

It seems to me that Spearman's $G$ meant something: many measures of intelligence were highly correlated to a single factor. Likewise, I would suggest that linear combinations of the original variates, where the combination has a meaning, as is often the case with principal components, means something. It is harder to see what the full factor model means, with all sorts of non-orthogonality and the impossibility of expressing the factors in terms of the original variates (except by estimation).
If a factor analysis type of theory were to be applied at all, it should be, I suggest, to precise data where there are many readings so that precise estimates of parameters can be made. I would suggest that the cancer data, for example, is far from that. Can people really recall accurately from which combination of radio, papers etc they got their cancer knowledge from?

I suggest that another sample would yield different answers, which would greatly alter the estimates of the parameters. Moreover, even the one-factor model is fitting eight parameters and a variate to what are really sixteen points, and a two factor model would be worse still in leading to estimates with (as I believe) a huge variance.

I don’t know what the answer is, even after playing with the data of Table 4, but would venture these suggestions:

(1) In fact even the independence fit is not all that bad having regard to the imprecision of the data. At most, I suggest, a slight modification of this assumption is needed.

(2) More than one factor should not be considered for the above reasons.

(3) Possibly the model would be more convincing if the \( y \) had a physical meaning, perhaps related to the correlation between the answers to the questions.

(4) If parameters as well as the \( y \) are needed, I would suggest that the \( x \)'s should not be used but only the \( \pi \)'s.

In conclusion, I suggest that with social multivariate data we are often trying to explain, comment on, look at, rather imprecise figures in a way which adds to people’s understanding of what the data is saying. Has this been achieved here? I think that quite a few assumptions have been built in to the theory whose implications the user of the theory will often not fully comprehend, and so it will be hard for the user to know what has been achieved, in many cases. Certainly I think this is a core on which a clearer and simpler theory could be built, and I would hope that this will be done.

The foregoing expresses my own view.

Dr P. M. E. ALTHAM (Cambridge University): I would like to thank Professor Bartholomew for a useful and stimulating paper, and make two brief points.

(i) Although I should perhaps think further about Dr. Upton’s comments before I speak, my impression is that I would find it not too hard to interpret the latent structure models to a social scientist, and certainly not harder than interpreting a loglinear model with complicated high order interactions.

(ii) Latent structure models possess the following feature which I find attractive. The essential feature of the model is that for the observable variables \( x_1 \ldots x_p \), which are generally discrete, we postulate the existence of the latent variables \( y \), such that given \( y, x_1 \ldots x_p \) are independent. Thus the joint distribution of any subset of \( x_1 \ldots x_p \) has the same structure as that of \( x_1 \ldots x_p \). This seems a desirable property in applications where the number of observable \( x \)'s may not be very clearly defined; the social scientist would probably want to include or exclude extra \( x \)'s or “questions” without drastically altering his model. This “invariance” feature is not shared by loglinear analysis, although of course it must be recognised that loglinear and latent structure analyses are addressing rather different problems, but for the same type of data.

A consequence of this property, as pointed out already by Dr Skene, is that Professor Bartholomew’s logit model is “unaffected” by missing data; I only wish to put the positive advantages of the property more strongly.

The following contributions were received in writing, after the meeting.

Professor E. B. ANDERSEN (University of Copenhagen): It has been very stimulating to read Professor Bartholomew’s new unifying approach to latent structure analysis. The key issue is, of course, how to model in latent space. Although Professor Bartholomew argues very forcefully for always having a uniform distribution of the latent variable, it is important to note, that many arguments demand that we consider latent distributions with parameters. We may thus be interested in comparing several latent distributions or we may be interested in changes in a latent variable over time. In such cases a statistical analysis will usually take the form of a comparison of the parameters of different latent distributions. One of the models mentioned by Professor Bartholomew—with \( G \) a logit and \( H \) a probit—was considered in a paper by Andersen and Madsen (1977) and it has recently been extended to cover the type of comparisons I mentioned above (Andersen (1980)). If one compares the approach of Professor Bartholomew with the results just mentioned, it appears that the \( x \)-parameters of model (6) play different roles. Some of them are parameters connected with the manifest variables and some of them relate more to the latent variables. For an interpretation of the results of an analysis it may be worth the effort to make such a distinction.
As an example, if we consider the G-logit, H-probit model with one latent variable, $\alpha_{10}$ will combine the item parameters of the simple logistic item characteristic curve model (or Rasch model) and the mean of the latent variable, while $\alpha_{11}$ is a constant and equal to the standard deviation of the latent variable.

Dr J. A. Anderson (University of Newcastle upon Tyne): Professor Bartholomew's paper is doubly welcome because it combines two important topics, multivariate categorical data analysis and factor analysis. His paper provides a very helpful summary of necessary properties for factor models and contains the interesting results on the expectation of $R_{ij}$. I agree entirely that these factor models are useful and important but I still retain some preference for the probit model.

The method of estimation suggested here is to fit the one-way margins exactly and to optimize a measure of goodness of fit of the two-way margins. A similar approach has been established for the probit model by Mackenzie (1976), with the advantages that (i) the equation (25) is exact and (ii) maximum likelihood estimation for the $(\alpha_{ij})$ conditional on the estimates $(\beta_k)$ is feasible for many more dimensions than $p = 10$; these can be shown to be asymptotically efficient and standard errors can be derived. Bock and Lieberman's (1970) limitation, $p \leq 12$, refers to the simultaneous estimation of $(\pi_i)$ and $(\alpha_{ij})$ and may in any case be superceded by better methods of optimization.

A more fundamental concern about the logit model, when there is more than one factor, relates to rotatability. In continuous factor analysis models with, say, $k$ factors, it is possible only to estimate a $k$-dimensional factor space (Lawley & Maxwell, 1971). The choice of factors within this space is determined subjectively or by external criteria. The probit model for categorical factor analysis has exactly the same property. Rotation of the factor space corresponds to rotation of the factor loadings. However, the logit model appears not to be rotatable as unlike the normal case, independent, homoscedastic logistic, variates are not invariant under rotation. Since the probit and logit models are so close in other respects, I am concerned that there is an approximate rotatability in the logit model which would lead to numerical instability in the estimation procedure unless recognized and dealt with as in the continuous factor case.

This field has been neglected in the statistical literature and Professor Bartholomew is to be congratulated both on his results and on stimulating our interest.

Mr C. L. F. Attfield (University of Bristol): I found the paper particularly stimulating as the topic is one with which I am not altogether familiar and so the paper served to introduce me to the previous attempts to solve the problem, which Professor Bartholomew shows, can be viewed as special cases of his more general approach. I was impressed by the accuracy of the parameter estimates obtained by the "logit first approximation" method for the one factor case which, Professor Bartholomew states, can be obtained on a pocket calculator. The approximation should prove an invaluable tool in the preliminary analysis of categorical data.

It would be interesting to see the result of relaxing the assumption of independence of the latent variables. In working with unobservable economic variables I find independence very difficult to justify. Would it be possible to work through the analysis without imposing the independence condition and then construct a test for independence?

I would argue with Professor Bartholomew's remark in Section 2 that latent variables which are "real", i.e. can in principle be measured directly such as "personal wealth", are quite rare. On the contrary most latent variables in economic theory are of exactly this form, e.g. permanent income, the expected rate of inflation, anticipated investment. It is true that in the majority of these cases there is no problem because the latent variables can be associated with manifest variables which variables (or a transformation of them) can be assumed to be continuous and distributed as multivariate normal. The models can then be estimated using the GLS procedure due to Brown (1974) or the maximum likelihood method due to Jöreskog outlined in Jöreskog and Sörbom (1974).

Professor H. Goldstein (University of London): I found Professor Bartholomew's paper interesting and useful, but I am a little puzzled by the importance he attaches to assumption (iv) in Section 3. This seems to me to lead to an unnecessarily restrictive set of models and I fail to see why, in general, $1 - \pi(y)$ should belong to the same family of functions as $\pi(y)$. For example, for an examination question with a correct/incorrect response one would normally expect an incorrect response, say to a multiple choice question, to be obtained by way of different mental processes to a correct response, and I would not therefore expect (iv) to hold. I have argued elsewhere (Goldstein, 1980) that the complementary log log function (which does not satisfy (iv) but does, incidentally, satisfy (ii)) may in some circumstances be more appropriate than the logit or probit for exam type data. Whilst I would accept Professor
Bartholomew's justification of (iv) for many kinds of data, it would seem unfortunate if it were to be used where more realistic functions are available.

On the topic of conditional independence (referred to as "local independence" in much of the psychometric literature), Professor Bartholomew asserts in Section 2.2 that if conditional independence were not true, then this implies that some other latent variable was exerting an influence on the manifest variables. I am not sure I agree. Suppose we have determined the latent space, and choose a set of individuals at a single point in this space, that is, all having the same set of values on the latent variables. If we consider a \(2^p\) table of responses then conditional independence means that the response probabilities in this table depend only on the margins of the table and through these on the latent variables. This, however, seems to be rather a strong assumption, and even if conditional independence did not hold, we might still be able to relate the appropriate "interaction" probabilities in this table, via additional parameters (loadings), to the same set of latent variables. In practice we could presumably attempt this only if we had independent replicate observations on individuals, although this is a difficult thing to achieve in the social sciences. The dimensionality of the latent space is concerned with the between-individual variation, whereas the within-individual dependencies will determine how many parameters are associated with each latent variable.

I would like to endorse strongly what Professor Bartholomew says about the care needed in interpreting results from latent variable models. The history of factor analysis seems to be full of what are essentially mathematically convenient devices being confused with substantive reality. Of course, one way of inculcating a proper caution is if one can show that reasonable but different models, including some which do not satisfy (iv), can lead to quite different interpretations of a common data set. I very much hope that Professor Bartholomew will give some further thought to this issue.

The Author replied later, in writing, as follows.

The discussion has raised many fundamental questions which merit a more extended reply than the present limit on time and space allows. I am most grateful to those who submitted contributions and the following incomplete remarks are intended as first contribution to what will I hope be a continuing discussion.

Several speakers have drawn attention to the possibility of using a latent class model in which the latent variable is categorical. As Mr Skinner pointed out, such a model with two latent classes arises as a special case of our general model. Thus suppose we choose

\[
\pi_i(y) = \begin{cases} 
\pi_{i1}, & 0 \leq y < y_o \\
\pi_{i2}, & y_o < y \leq 1
\end{cases}
\]  

(A)

then \(y_o\) is the proportion of the population in latent class 1. Models with three or more ordered latent classes arise similarly. The function in (1) satisfies all the conditions of Section 3 though it is not expressible in the form (6) and hence the parameters are not interpretable in the same way. In-so-far as (A) can approximate the logit (or probit) functions, or vice versa, we would expect the latent class model to give fits similar to ours. It is re-assuring to find from Professor Aitkin's calculations that this is indeed the case and his examples will repay detailed study. It is worth noting that the coefficients of his discriminant functions exhibit the same pattern as the \(x\)'s in our model. In particular they indicate that, in the example on cancer knowledge, written sources of information carry more weight than what is heard through radio or lectures.

The choice between a latent class and a latent variable model may be made on grounds of realism or computational convenience. My own experience suggests that continuous latent variables are almost always more realistic but in any case, as we have just shown, both kinds of model can be accommodated within the one framework. Professor Aitkin's case is thus essentially one for using step functions like (A). He claims the computational advantages which stem from the existence of computer programmes which can easily be adapted for fitting the model. There may well be short term advantages here but the main thrust of the paper was to look at a wide class of models within a framework which allows their relative merits to be assessed. I hope that this would caution against the premature adoption of second best models until the full range of options has been thoroughly explored and evaluated.

From a practical point of view there seems to be little at stake in the argument over whether a probit, logit or a hybrid version of the model should be used. My own decision to investigate the logit model was largely stimulated by the discovery of its close link with the cross-product ratios. The latter seem more natural measures of association than tetrachoric correlations and the remarkably good approximation provided by (14) may be indicative of some deeper connection yet to be uncovered. At the very least it justifies further study. In the long run I see no obstacle to making the general version of the logit model as
Discussion of the Paper by Professor Bartholomew

[No. 3,]

easily applicable as the probit is now. The simplicity of the approximate method of fitting the one factor logit model already promises well.

A number of questions were raised about the choice of \( p(y) \) and \( \pi(x \mid y) \). However, the point made in the paper about the essential arbitrariness involved in this choice does not seem to have been fully taken. Another way of making the point is to observe that any change of variable in the integral of (1) leaves \( f(x) \) unchanged. Since \( f(x) \) is all that can be estimated there is no empirical means of distinguishing between any of the combinations of functions which lead to a given \( f(x) \). It is thus pointless to argue about the most realistic form of \( p(y) \); the question is purely one of convenience. By the same token, analyses which depend in an essential way on any particular form have little meaning. For this reason I am doubtful about the consequences of Dr E. B. Andersen's proposal to make the parameters of \( p(y) \) functions of time.

Several speakers mentioned directly or by implication the lack of standard errors and the disregard of questions of efficiency. This omission was not entirely due to lack of space but arises in part from misgivings about the relevance of such concepts in an analysis which is exploratory and descriptive. The overall goodness of fit test is certainly useful. How much standard errors contribute to the interpretation of the analysis over and above that seems more questionable. The root of the matter has to do with how far it is sensible to consider repeated sampling from a population which, in a certain sense, has no real existence.

Dr Skene, Dr J. A. Anderson and Dr Upton all raise matters concerning the case \( q > 1 \). There was no space to develop this in the paper and their remarks have been noted for future use. For small \( q \) the normal theory carries over by Theorem 2 but this is scarcely adequate for general use. The probit model has real advantages here and the closedness of the logit and probit models means that the estimates of the parameters of one can easily be transformed to give estimates of the other.

No one commented on Section 7 which shows how any model of the family can be fitted to tables with several ordered manifest categories. This brings a much wider range of data within the scope of the analysis and further work is in progress to implement the method.

Apart from these general issues a number of specific points were raised and these are dealt with, in order, below.

Professor Aitkin's example involving the use of principal components in regression is a good instance of the benefits of setting latent variable problems in a general framework. His example is used to illustrate the same point in Bartholomew (1981).

Dr Skene is right to point out that the mean may not be a good measure of location for the posterior distribution but the remark has to be interpreted in the light of the arbitrariness of \( p(y) \). By choosing \( y \) to be uniform we ensured that \( E(y \mid x) \) had a "distribution-free" interpretation as the expected quantile of an individual chosen at random from those with a given \( x \).

Mr Skinner's remarks about the relationship between the probit and logit estimation methods are illuminating. I would part company with him in seeing it as a virtue of the probit model that it can be easily linked with the normal model of factor analysis. For teaching purposes I think that this obscures rather than reveals the common underlying structure shared by all latent variable models.

Dr Upton's simple formula for the \( y \)-scores is intriguing but not altogether surprising. It is another indication of the common structure which emerges, for example, in theorem 2. For sufficiently small \( \alpha \)'s the \( y \)-scores will satisfy an equation of the form

\[
Ay = B + \sum_{i=1}^{p} \alpha_i x_i
\]

It is surprising that this form is so good when the \( \alpha \)'s are not small. This example encourages the hope of finding a general linear approximation which would avoid the need to evaluate the integrals required for the \( y \)-scores. I share his regret at the scant attention paid in the paper to the interpretation of the results which, again, was solely due to lack of space. I am grateful to Dr Upton and other contributors for helping to remedy the deficiency.

As always, Mr Stern puts us on our guard against undue sophistication. In this case however, I would dispute all of his conclusions. The study of cancer knowledge was not concerned with asking people about the source of their knowledge of cancer. The point of trying to find latent variables which could be identified with "Knowledge about things in general" was to see how this related to knowledge of cancer.

Drs Altham and Skene drew attention to an important property of the model which is extremely useful in social applications not least in the regression problem referred to by Professor Aitkin. There are usually a great many possible manifest variables from which those included in the study are often chosen in an
arbitrary fashion. As these contributors imply, the manner of this choice does not invalidate the model. It is for this reason that the "naive" interpretation of linear combinations of manifest variables as recommended by Mr. Stern may be less simple than appears at first sight.

I was interested to hear from Dr. J. A. Anderson about Mackenzie's work and would commend it to Professors Aitkin and Bock since it would be interesting to see how it compares with their method.

Mr. Attfield raised two matters which deserve more attention than is possible here. Among economists, I suspect, latent variables such as those he mentions are firmly established in the language of the theory. They will often be correlated and it is natural to want to express the analysis in terms which will relate to economic theory. This can be done, in principle, for the models given here but I would prefer to work with orthogonal latent variables at the first stage and then transform them subsequently to dimensions which are economically meaningful. On the question of "real" latent variables I do not think that there is any substantial difference between us. I defined a real latent variable as one which could, in principle, be measured. I think Mr. Attfield regards a variable as real if economists believe it has meaning in economic discourse. The two definitions are not equivalent.

Professor Goldstein's points are, likewise, fundamental. Whatever our differences I fully endorse his penultimate sentence. It is certainly possible to conceive of a situation in which \( p(y) \) does not satisfy assumption (iv) of Section 3. The question at issue may be put as follows: Does the labelling of the categories convey any information which is relevant to the analysis? If not then condition (iv) follows. Even if this argument is not accepted there are substantial empirical grounds for adopting (iv). Suppose we take some simple function, not satisfying (iv), such as \( p(y) = y^2 \). Then the function \( p(y) = 1 - (1 - y)^a \) might appear to serve equally well. Which form should we choose? With \( p \) dimensions to the table there are \( 2^p \) possible combinations. It would be a formidable task to investigate all of them and choose the best. Invoking condition (iv) reduces the options to one. It we decide to use a function not satisfying (iv) we have to find some extraneous grounds for preferring a particular form. In view of all that we have said about the arbitrariness involved this seems to be a formidable task.

The crucial assumption of conditional independence is more in the nature of a definition than an assumption. It is a formal statement of what we mean when we say that the variation among the \( x \)'s is completely explained by their dependence on the \( y \)'s. If the latent variables are constructs, the assumption could never be tested empirically and so it does not seem appropriate to speak of it in terms of being true or false. There is clearly much more to be said on this, and many other of the issues raised, and I hope that the debate will continue.

REFERENCES IN THE DISCUSSION


