Biometrika (2002), **89,** 2, *pp.* 462–469

Printed in Great Britain

On some models for multivariate binary variables parallel in complexity with the multivariate Gaussian distribution

BY D. R. COX
Nuffield College, Oxford OX1 1NF, U.K.
david.cox@nuf.ox.ac.uk

AND NANNY WERMUTH
Psychological Institute, University of Mainz, D55099, Germany
nanny.wermuth@uni-mainz.de

SUMMARY

It is shown that both the simple form of the Rasch model for binary data and a generalisation are essentially equivalent to special dichotomised Gaussian models. In these the underlying Gaussian structure is of single factor form; that is, the correlations between the binary variables arise via a single underlying variable, called in psychometrics a latent trait. The implications for scoring of the binary variables are discussed, in particular regarding the scoring system as in effect estimating the latent trait. In particular, the role of the simple sum score, in effect the total number of 'successes', is examined. Relations with the principal component analysis of binary data are outlined and some connections with the quadratic exponential binary model are sketched.

Some key words: Logistic function; Median dichotomy; Multivariate Gaussian distribution; Principal components; Probit; Rasch model; Sheppard's formula.

1. INTRODUCTION

There are a number of different types of model for the joint distribution of a set of binary variables. When the number of components p is small a multinomial distribution on the 2^p possible values may be used but for larger values of p some distribution with a smaller number of adjustable parameters will often be useful, in some sense an analogue of the multivariate Gaussian distribution for continuous variables. Three important such special distributions are the dichotomised Gaussian distribution (Pearson, 1909), the Rasch model (Rasch, 1960, 1961) and the quadratic binary exponential model (Cox, 1972; Cox & Wermuth, 1994), a special case of a log-linear representation of probabilities.

Denote the vector $p \times 1$ binary random variable by A, the components taking values 1 and -1, which we occasionally call success and failure. The convention of choosing 1 and -1 rather than the more usual 1 and 0 simplifies some calculations slightly.

There are a number of distinct purposes for which a model for the joint distribution might be required. We consider here internal analysis (Bartlett, 1947), i.e. the study of the internal relation between the components, as contrasted with external analysis in which the components are studied in their relationship to a second set of variables. Even in internal analysis the emphasis changes somewhat depending on whether the components are of intrinsic interest or are what we shall call items, recorded because they form multiple indicators for some underlying variable, W. Our emphasis here is on the latter possibility, in particular on the use of sum scores, that is weighted linear combinations of the components of A, to estimate the latent trait, W. In particular, we examine

the efficiency of the simple sum score in which the components are given equal weight, this being essentially equivalent to counting the number of components with the 'upper' level of response, here conventionally called successes.

2. The dichotomised Gaussian distribution

A specification with a long history (Pearson, 1909) is obtained by supposing that the binary vector A is derived from an unobserved Gaussian vector, U. Without loss of generality suppose that U has zero mean and unit standard deviations; denote its covariance matrix by Σ_{UU} and suppose that $A_s = 1$ if and only if $U_s > -\gamma_s$. Then, for example,

$$E(A_s) = 2\Phi(\gamma_s) - 1, \quad \text{var}(A_s) = 4\Phi(\gamma_s)\Phi(-\gamma_s), \quad \text{cov}(A_s, A_t) = 4\Psi_2(\gamma_s, \gamma_t; \rho_{st}),$$

for $s \neq t$, where ρ_{st} is the correlation coefficient of (U_s, U_t) and

$$\Psi_2(x, y; \rho) = \Phi_2(x, y; \rho) - \Phi(x)\Phi(y),$$

where $\Phi(x)$ is the standardised Gaussian cumulative distribution function and $\Psi_2(x, y; \rho)$ denotes the cumulative distribution function of the standardised bivariate normal distribution of correlation ρ. Now, by a formula of Sheppard (1898), we have in the special case of median dichotomy, i.e. when $\gamma_s = \gamma_t = 0$, that

$$\Psi_2(0, 0; \rho) = 2\pi^{-1} \sin^{-1} \rho. \tag{1}$$

We shall explore the consequences of a very special dichotomised Gaussian distribution in which the underlying latent variables themselves have a single factor structure. This is based on there being two levels of latent variable. That is, the U's which generate the binary variables are themselves generated from a single underlying variable W with error in such a way that $U_1, \ldots, U_p$ are mutually independent given W. This corresponds to the simple graphical Markov model (Edwards, 2000, p. 189; Lauritzen, 1996, p. 4; Cox & Wermuth, 1996, p. 30) of Fig. 1(a). It implies the corresponding independence structure in the binary variables A derived from U; see Fig. 1(b).

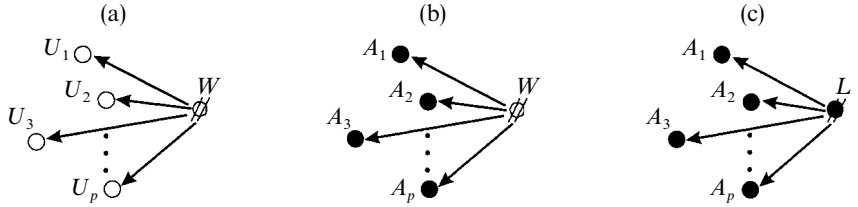

Fig. 1. (a) Normally distributed variables conditionally independent given latent variable W; correlations obey tetrad conditions. (b) Binary variables derived from (a) by dichotomy; tetrachoric correlations obey tetrad condition. (c) Binary variables in latent class model derived from latent class model with two latent classes defined by L; conditional log-odds obey tetrad condition.

Explicitly let $U_1, \ldots, U_p$ be unobserved standardised Gaussian variables such that

$$U_s = \lambda_s W + \sqrt{(1 - \lambda_s^2)} V_s,$$

where $W, V_1, \ldots, V_p$ are independent standard Gaussian variables. The parameter λ_s, which is the least squares regression coefficient of U_s on W, is called a loading in the context of a single-factor model. The joint marginal distribution of the U_s is thus a standardised multivariate Gaussian distribution with the special correlation matrix $\mathscr{T}(\lambda)$, that is having (s, t)th element $\lambda_s \lambda_t$. This satisfies the so-called tetrad condition for correlations involving four different indices (q, r, s, t), namely that for nonzero correlations

$$\rho_{st}/\rho_{rt} = \rho_{sq}/\rho_{rq} = \lambda_s/\lambda_r.$$

The tetrad condition is a consequence of the linear structure underlying the Gaussian distribution and so will not apply to correlations derived from general random variables with the independence structure of Fig. 1(a) or (b). In particular it will not apply to binary variables. For median dichotomy, however, (1) applies and, for small x, $\sin^{-1}(x)$ is approximately $x + x^3/6$; in fact $\sin^{-1}(x)$ differs from x by less than 10% provided that $x < 0{\cdot}65$. Thus the correlation between the binary variables is approximately proportional to the correlation between the corresponding Gaussian components and thus the tetrad condition will hold approximately for the binary variables, so long as the correlations are not too large. The tetrachoric correlations derived by finding via the bivariate normal integral the correlations in the underlying Gaussian distribution will, however, continue to satisfy a tetrad condition.

We denote the covariance matrix of the vector A by Σ_{AA}. Further we write Σ_{WA} for the row vector $\mathrm{cov}(W, A_s)$; it has elements

$$2\lambda_s \phi(\gamma_s),$$

where $\phi(x)$ denotes the standard normal density function.

3. The Rasch model

Let $L(x) = e^x/(1+e^x) = 1 - L(-x)$ denote the unit logistic function. Then in the Rasch model (Rasch, 1960, 1961) the probability that, for example, all the p binary variables for a given subject are successes is

$$\prod_s L(\alpha'_s + \beta'_s w), \tag{2}$$

where w is a subject effect on some standardised scale and α'_s, β'_s characterise the sth variable. The general form is

$$\mathrm{pr}\{A_s = a_s \ \ (s = 1, \ldots, p)\} = \prod_s L\{a_s(\alpha'_s + \beta'_s w)\}. \tag{3}$$

In the context of educational testing, α'_s is a measure of item general difficulty and β'_s a measure of its selectivity. In the simple form of the Rasch model the β'_s are all equal. This leads to the simple sum score, equivalent to the total number of successes, as the sufficient statistic for an individual with unknown w.

In a random subject version, w is the value of a random variable W so that the probability (3) is replaced by its expectation over the distribution of W. In particular we shall suppose for the present paper that W has the standard normal distribution. From this viewpoint the model has in W a single latent variable. This is often also called a construct and sometimes one aims to estimate its value for an individual via multiple observed indicators, A.

4. An approximate relationship between formulations

It is known that to a close approximation

$$L(x) = \Phi(cx) \tag{4}$$

for a suitable constant c. If we combine this with the assumption that the random variable W has a standard Gaussian distribution we have that (2) becomes approximately

$$\int_{-\infty}^{\infty} dw \phi(w) \prod_s \Phi(\alpha_s + \beta_s w), \tag{5}$$

where $\alpha_s = c\alpha'_s$ and $\beta_s = c\beta'_s$. Note incidentally that use of a different scaling constant c for each s is allowable and that if the effective range of the probability of success is very different for the different items this would lead to an improved approximation (4). The approximation will work

worst when there are some component variables with much steeper slopes, that is β_s, than the majority. In that case for some values of w the arguments of the distribution functions are likely to be at the extremes of the range and the Gaussian distribution then assigns more extreme probabilities than does the logistic.

It can now be shown that the integral (5) is the p-dimensional standard Gaussian distribution function

$$\Phi_p\{\alpha_s/\sqrt{(1+\beta_s^2)}\ (s=1,\ldots,p);\ \mathscr{T}(\beta_s/\sqrt{(1+\beta_s^2)})\}. \tag{6}$$

Note that for the simple Rasch model, that is the model with all slopes, i.e. loadings, equal, $\mathscr{T}$ becomes an intraclass correlation matrix, i.e. a matrix with all correlations the same, namely $\beta/\sqrt{(1+\beta^2)}$.

The result (6) can be proved analytically from the integral I in (5) by forming

$$\partial^p I/\partial\alpha_1 \ldots \partial\alpha_p,$$

simplifying and then integrating the resulting exponentiated quadratic form. A direct probabilistic proof is as follows. Let $V_1,\ldots,V_p$ be independent standard Gaussian random variables. Then conditionally on $W=w$ the required probability has the form

$$\mathrm{pr}\{V_s<\alpha_s+\beta_s w\ \ (s=1,\ldots,p)\},$$

and when we integrate over the distribution of W this becomes

$$\mathrm{pr}\{(V_s-\beta_s W)/\sqrt{(1+\beta_s^2)}<\alpha_s/\sqrt{(1+\beta_s^2)}\ \ (s=1,\ldots,p)\}. \tag{7}$$

Note that the random variables in (7) are all standardised to zero mean and unit variance and that their correlation matrix is the tetrad matrix $\mathscr{T}$ defined above, thus proving the approximate equivalence of (5) and (6).

5. Linear functions of a binary variable A

When the A_s are regarded primarily as multiple indicators for some unobserved feature, i.e. a latent trait or a construct, which we shall take to be W, we are then interested in summary scores based on A, in particular in the simple unweighted sum score $S=1_p^{\mathrm{T}}A$, where 1_p is a column vector of ones, equivalent to the total number of successes. We shall consider only linear combinations of the individual item responses, although for binary responses the restriction to linear sum scores, however convenient in practice, has no obvious formal statistical justification. There are, however, clear practical arguments in favour of such a restriction. The arguments set out below can be extended to include polynomial, and hence entirely general, functions of A.

The linear least squares regression of W on A gives the optimal linear combination $\Sigma_{WA}\Sigma_{AA}^{-1}A$ for estimating the unobserved W and, noting that the marginal variance of W is one, we may define the efficiency of the corresponding estimator of W by the squared multiple correlation of W on the vector A, namely

$$\mathscr{E}_A=\Sigma_{WA}\Sigma_{AA}^{-1}\Sigma_{WA}^{\mathrm{T}}.$$

By contrast the efficiency of the simple sum score, S, is

$$\mathscr{E}_S=\{\mathrm{cov}(S,W)\}^2/\mathrm{var}(S)=(\Sigma_{WA}1_p)^2/(1_p^{\mathrm{T}}\ \Sigma_{AA}1_p).$$

Under the simple Rasch model, corresponding to equal slopes, S is, among functions of A, fully efficient for estimating W.

More generally we might use weighted linear scores derived by some different approach, or even more than one score simultaneously. If L is a $r\times p$ matrix defining a set of scores LA, the corresponding measure of efficiency is

$$\mathscr{E}_{LA}=\Sigma_{WA}L^{\mathrm{T}}(L\Sigma_{AA}L^{\mathrm{T}})^{-1}L\Sigma_{WA}^{\mathrm{T}}.$$

Finally, for comparison it is helpful to have the efficiency in this sense that would be achievable were the continuous variables U to be observed. This is

$$\mathscr{E}_U = \Sigma_{WU}\Sigma_{UU}^{-1}\Sigma_{WU}^{\mathrm{T}},$$

where $\Sigma_{WU} = \{\lambda_1, \ldots, \lambda_p\}$ and $\Sigma_{UU} = \mathscr{T}(\lambda_1, \ldots, \lambda_p)$ is the corresponding tetrad matrix. Since the inverse of a tetrad matrix is also of tetrad form (Bartlett, 1951) it follows that

$$\mathscr{E}_U = \frac{\Sigma\lambda_j^2/(1-\lambda_j^2)}{1+\Sigma\lambda_j^2/(1-\lambda_j^2)}.$$

If, in particular, all the λ_j are equal, say to λ, then

$$\mathscr{E}_U = \frac{p\lambda^2}{1+(p-1)\lambda^2}.$$

6. Remarks on principal component analysis

Before illustrating the above results numerically and discussing their interpretation we comment briefly on the principal component analysis of binary data.

One version focuses on the individuals and aims to map points in p dimensions into a reduced number of dimensions preserving so far as possible the Euclidean distance between pairs of points. From this point of view analysis of the covariance matrix of binary features poses no special problem; the Euclidean distance in the originating binary variables is essentially the number of discrepant components. Note that principal component analysis of the correlation matrix would not have this interpretation.

In the dual interpretation, emphasis is on the variables. Here the relevance for binary variables is much less clear, essentially because orthogonal transformation of binary variables, standardised or not, has no clear meaning. Nevertheless there is the possibility that the first principal component of the correlation matrix of binary variables may be close to the optimal linear combination of the binary responses as defined above for estimating a hypothesised underlying Gaussian single-factor latent variable model.

We shall see in § 7 that this is indeed sometimes the case. In some extreme situations this is immediately clear. For example, if there were a mixture of exchangeable binary variables following the Gaussian latent factor model and some mutually independent variables independent also of the first set, then principal component analysis would give the latter variables zero weight. The simple sum score of the first set would be recovered. Furthermore, if some of the variables are scored in the wrong direction, but otherwise the variables are exchangeable, again the optimal combination would be recovered.

7. Some consequences

We now consider some detailed results for a binary distribution derived from an underlying single factor latent Gaussian model. Note that the dichotomised Gaussian model corresponding to the simple Rasch model has to a first approximation all loadings equal.

In the simplest exchangeable case where all the variables are median dichotomised and have equal loadings λ, so that S is the optimal linear sum score, the limiting behaviour of $\mathscr{E}_S$ is slightly delicate. As λ tends to zero for fixed p, $\mathscr{E}_S$ tends to zero, as would be expected. As p tends to infinity for fixed λ, $\mathscr{E}_S$ tends to a limit which is a decreasing function of λ, the limit being close to one for small λ. This is because, when the internal correlation of the items is very high, there is little gain from replication. When λ is small, however, the limit for large p is approached slowly. Table 1 shows some typical values of $\mathscr{E}_S$. The relative loss by observing only the dichotomised variable A rather than the continuous variable U is shown by the ratio $\mathscr{E}_S/\mathscr{E}_U$ and is typically about 0·7 for $p = 4$ and slightly larger for $p = 8$.

Table 1. *Some efficiencies of the simple sum score for equal loadings and median dichotomy*

λ^2	$\mathscr{E}_S$			$\mathscr{E}_U$	
	$p=4$	$p=8$	$p\to\infty$	$p=4$	$p=8$
0·1	0·214	0·352	0·998	0·308	0·471
0·2	0·368	0·537	0·993	0·500	0·667
0·4	0·571	0·719	0·973	0·727	0·842
0·6	0·685	0·790	0·932	0·857	0·923
0·8	0·735	0·794	0·863	0·941	0·970
0·9	0·730	0·765	0·804	0·973	0·986

A reasonably detailed further study has been made for $p=4$. It is difficult to summarise the conclusions concisely but they are essentially the following. If we consider the totally exchangeable case, i.e. where all components have the same loading, λ, and cut-off point, γ, the simple sum score, the optimal linear score and the first principal component coincide, by symmetry. For fixed γ the efficiency of the linear score has a maximum as a function of λ at approximately $\lambda=0{\cdot}9$, the position depending slightly on the value of γ.

The closest parallel to the assumptions of the simple Rasch model requires all components to have the same λ but allows different marginal distributions, i.e. different λ. In these cases indeed the simple sum score has an efficiency within less than 1% of the most efficient linear combination. The first principal component of the correlation matrix of the binary variables is only very slightly less efficient; see, for example, the first row of Table 2.

Table 2. *Some efficiencies with different marginal distribution and loadings*; $p=4$. *Efficiencies are of optimal linear combination of binary responses, simple sum score, first principal component of binary items and of underlying continuous variables*

(γ, L)	$\mathscr{E}_E$	$\mathscr{E}_S$	$\mathscr{E}_{\mathrm{PCA}}$	$\mathscr{E}_U$
$(-1, 0{\cdot}98)$, $(-0{\cdot}5, 0{\cdot}98)$, $(0{\cdot}5, 0{\cdot}98)$, $(1, 0{\cdot}98)$	0·867	0·867	0·865	0·990
$(0, 0{\cdot}4)$, $(0, 0{\cdot}4)$, $(0, 0{\cdot}98)$, $(0, 0{\cdot}98)$	0·692	0·614	0·678	0·980
$(-1, 0{\cdot}4)$, $(-1, 0{\cdot}4)$, $(-1, 0{\cdot}98)$, $(-1, 0{\cdot}98)$	0·496	0·451	0·494	0·980
$(0, 0{\cdot}2)$, $(0, 0{\cdot}6)$, $(0, 0{\cdot}95)$, $(0, 0{\cdot}99)$	0·699	0·614	0·686	0·983
$(-1, 0{\cdot}2)$, $(-1, 0{\cdot}6)$, $(-1, 0{\cdot}95)$, $(-1, 0{\cdot}99)$	0·501	0·448	0·498	0·983
$(-1, 0{\cdot}98)$, $(-1, 0{\cdot}4)$, $(1, 0{\cdot}4)$, $(1, 0{\cdot}98)$	0·721	0·568	0·604	0·980

If all components have the same γ, in particular if they are median dichotomised, $\gamma=0$, but have different loadings, λ, the first principal component derived from the correlation matrix of the binary items is appreciably better than the simple sum score, but is capable of some small further improvement; see the second to fifth rows of Table 2. When, however, there are large variations in both marginal distribution and in loadings the first principal component, while still an improvement on the simple sum score, is appreciably less predictive than the optimal linear combination. See the last row of Table 2.

Note again that all this discussion is concerned with the properties of probability distributions. If the underlying single factor model is indeed assumed as the basis of interpretation, there are at least two routes to statistical estimation of a linear function intended to improve on the simple sum score. One is by principal component analysis. Another is to estimate as simply as possible the parameters λ_s and α_s and via them the coefficients of the optimal linear estimating function for W.

8. Quadratic binary exponential model

We now turn to a third distributional form. The quadratic binary distribution has

$$\mathrm{pr}\{A_s=a_s\ (s=1,\ldots,p)\}=\exp\{\Sigma\delta_s a_s+\Sigma_{s>t}\delta_{st}a_s a_t-k(\delta)\}, \tag{8}$$

where $k(\delta)$ is a normalising constant depending on the full parameter vector δ; see, for example, Cox (1972) and Cox & Wermuth (1994). The interpretation of $\delta_{st}=0$ is that $A_s \perp\!\!\!\perp A_t \mid A_u$ $(u \neq s, t)$. In this sense the symmetric matrix formed from δ_{st} is analogous to the inverse covariance or concentration matrix in the multivariate normal distribution. The simplification corresponding to the vanishing of particular quadratic coefficients is thus incisively represented via an undirected independence graph in which conditional independencies given all remaining variables are shown by missing edges; this is sometimes called a concentration graph (Cox & Wermuth, 1996). The properties of such graphs are known in generality (Lauritzen, 1996).

Such a model is a special case of a whole family of log-linear models including when the full number of parameters is allowed the saturated multinomial model on 2^p points. The special interest of (8) lies in the reduced number of parameters obtained by excluding cubic and higher terms in the expansion.

An underlying continuous variable is not easily manipulated within this model in general but a binary latent class variable, L, can be included by assuming that it and A jointly have a quadratic exponential distribution in which, analogously to the situation in Fig. 1(b), $A_1, \ldots, A_p$ are mutually independent given L, see Fig. 1(c); that is, we start from the form (8) for the $p+1$ variables $L, A_1, \ldots, A_p$ in which the only nonzero quadratic coefficients are those involving L. If we now marginalise over L the quadratic exponential form is retained only as an approximation, but in that approximation $\delta_{(L)rs}$, the coefficient of $a_r a_s$ marginalised over L, is approximately of the tetrad form

$$\delta_{(L)rs} = \delta_{Lr}\delta_{Ls}\,\mathrm{sech}^2\,\delta_L;$$

that is, the underlying structure yields a tetrad condition not on the correlations or on the tetrachoric correlations but on the conditional log-odds ratios.

9. Some broad comparisons

In the general p-dimensional multivariate Gaussian distribution the standard parameterisation is in terms of p means, p variances, or equivalently conditional precisions, and $p(p-1)/2$ correlations. Alternatively one may use the means and the elements of the concentration matrix, these being equivalent to conditional variances and partial correlations. In the simple Rasch model and its Gaussian equivalent there are p location parameters and a single parameter defining interrelations. In the general Rasch model and its Gaussian equivalent there are p location parameters and p parameters defining interrelations.

In some ways the most important consequence of the present paper is that the Rasch model and the corresponding dichotomised Gaussian model are likely to lead to essentially the same conclusions. In any case, empirical discrimination between Rasch and multivariate Gaussian models with the same number of parameters would depend on aspects of the difference between logistic and integrated normal distribution functions and hence is very unlikely to be either feasible or interesting in practical cases. Discrimination between Rasch and quadratic binary exponential distributions in their general form is also likely to be difficult; it will not be discussed further here.

Their relative usefulness as tools for detailed analysis will depend on how relevant it is for interpretation to establish a connection with a hypothesised latent variable.

Acknowledgement

We are very grateful to Professor David Firth for comments on the paper and for some of the numerical work, to the referees for constructive comments and to Professor J. K. Ghosh for asking a question leading to the study of the role of principal component analysis.

References

Bartlett, M. S. (1947). Multivariate analysis (with Discussion). *Suppl. J. Statist. Soc.* **9**, 176–97.

Bartlett, M. S. (1951). An inverse matrix adjustment arising in discriminant analysis. *Ann. Math. Statist.* **22**, 107–11.

Cox, D. R. (1972). The analysis of multivariate binary data. *Appl. Statist.* **21**, 113–20.

Cox, D. R. & Wermuth, N. (1994). A note on the quadratic exponential binary model. *Biometrika* **81**, 403–8.

Cox, D. R. & Wermuth, N. (1996). *Multivariate Dependencies.* London: Chapman and Hall.

Edwards, D. (2000). *Introduction to Graphical Modelling*, 2nd ed. New York: Springer.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford University Press.

Pearson, K. (1909). On a new method of determining correlations between a measured character *A*, and a character *B*, of which only the percentage of cases wherein *B* exceeds (or falls short of) a given intensity is recorded for each grade of *A*. *Biometrika* **7**, 96–105.

Rasch, G. (1960). *Probabilistic Models for some Intelligence or Attainment Tests.* Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proc. 4th Berkeley Symp. Math. Statist. Prob.* **4**, Ed. J. Neyman, pp. 321–34. Berkeley: University of California Press.

Sheppard, W. F. (1898). On the geometric treatment of the 'normal curve' of statistics with particular reference to correlation and the theory of errors. *Proc. R. Soc. Lond.* **62**, 170–3.

[*Received February* 2001. *Revised September* 2001]