Multivariate Analysis: Future Directions
C.R. Rao, editor

Chapter 6

SOME RECENT WORK ON METHODS FOR THE ANALYSIS OF MULTIVARIATE OBSERVATIONAL DATA IN THE SOCIAL SCIENCES

D.R. COX and N. WERMUTH

A review is given of some recent work on a number of themes: the construction of derived response variables, when some of the vector variables have components where individual interpretation is to be preserved, the examination of the adequacy of covariance matrices as a means of capturing dependency and association and the study of special patterns of conditional independence.

1. Introduction

The purpose of this chapter is to outline some of our recent work connected with the analysis of multivariate data with primary although not exclusive emphasis on observational studies in the social sciences. The majority of the chapter deals with continuous variables where structure is assumed to be adequately described via a vector of means and by a covariance matrix. For further discussion of the important case of mixed discrete and continuous variables, see Lauritzen and Wermuth (1989), Wermuth and Lauritzen (1990) and Cox and Wermuth (1992a).

We concentrate on the structure of models of dependence and association and on their interpretation rather than on methods of formal inference which can typically be achieved via maximum-likelihood, augmented by examination for outliers, etc. or, when large amounts of data have to be dealt with rather automatically, via corresponding robust methods.

Among the features of 'classical' normal-theory multivariate analysis as set out, e.g., in the books by Rao (1973) and Anderson (1984) are the following:

(*i*) except for methods of internal analysis, such as principal component analysis, there is invariance under linear transformations of the components;

(*ii*) the covariance matrix as a summarizer of dependency structure al-

lows no possibility of representing nonlinear or interactive effects, except via nonlinear transformation of the vector concerned; e.g., for three components we cannot detect a dependence of the regression coefficient of Y_1 on Y_2 given $Y_3 = y_3$ on the value of y_3;

(*iii*) a $p \times p$ covariance matrix contains $\frac{1}{2}p(p-1)$ correlation coefficients so that, especially for large p, there is from various points of view a need to reduce the number of adjustable parameters.

All these features have both positive and negativeaspects. We discuss these briefly in turn. The invariance of, e.g., the canonical correlation and regression analysis of the $p \times 1$ vector Y on the $q \times 1$ vector X under nonsingular transformation of Y and of X leads to the elegant and ultimately geometric theory involved (Dempster 1969; Chapter 6).

Substantively, however, the invariance may or may not be sensible. Thus if the components of Y are log height and log weight, it is sometimes plausible that other linear combinations of log height and log weight are the appropriate basis for interpretation and the invariance of an analysis has some appeal. On the other hand if the components are, say, anger and anxiety it is more likely that an interpretation should preserve the individual identity of the components as representing distinct features or properties. If these were explanatory variables a linear combination representing their relative effects on some response could be a reasonable base for interpretation; when the components are responses the position may be different. In Section 2 we outline work in which some component variables may be subject to linear transformations while others are required to preserve their specific identity.

The points (*ii*) and (*iii*) are somewhat contradictory in that the former points to a lack of richness in specification via the covariance matrix whereas the latter aspect stresses possible overparametrisation in dealing with largish arbitrary covariance matrices. There are general implications for the desirability of introducing substantive knowledge into the analysis. Section 3 discusses the detection of effects not represented by covariance matrices and Section 4 reviews ways of examining special covariance structures. A fundamental subject-matter distinction in such discussions is between response variables, intermediate response variables and explanatory variables.

2. Derived variables

In Section 1 we distinguished between vector variables to which linear transformations could be applied and vector variables where components have an individual identity to be preserved for interpretation. Of course, in applications the distinction is bound to be to some extent provisional.

If we have a $p \times 1$ vector Y of response variables and a $q \times 1$ vector X of explanatory variables then in the absence of further subject-matter information we may proceed as follows:

(a) to preserve the individual identity of the components of Y, consider the multivariate regression of Y on X, i.e., the component by component multiple regression of Y on X and then look for substantively meaningful simplifications and interpretations of which one extreme form might be that each component Y_i has regression on a distinct subset of the components of X;

(b) if $p \geq q$ and it is required to preserve the components of X but not those of Y, Cox and Wermuth (1992b) constructed a $q \times 1$ vector Y^* such that when regressed on X, the ith component Y_i^* has nonzero regression only on the corresponding X_i, i.e., Y_i^* is conditionally independent of X_j $(j \neq i)$ given X_i. In essence a preliminary transformation to the q components of the canonical regression of Y on X is made and then, provided all canonical correlations are nonzero, a nonsingular transformation of these variables achieves the desired structure. Equivalently,

$$Y^* = \Sigma_{xx}(\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx})^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}Y, \tag{2.1}$$

where the covariance matrix Σ of $(X^T, Y^T)^T$ has been partitioned in the usual way. Wermuth and Cox (1992) discussed this further and gave examples, additional to the one of Cox and Wermuth (1992b), where the method led to simple representations of the data via the replacement of Y^* as determined by (2.1) by components with simple interpretation.

An equivalent formulation is to note that Y_1^*, say, has maximal partial correlation with X_1 given $X_2, \ldots, X_q$. It can thus be calculated as the linear discriminant function (one-dimensional canonical variable) from the regression of Y on X_1 adjusting for $X_2, \ldots, X_q$ and residuals of a full regression of Y on $X_1, \ldots, X_q$.

In (b) if there are some virtually zero canonical correlations or if $p < q$, it is either too ambitious or impossible to find a component Y_i^* to go with each component of X. For example, one might eliminate certain components of X altogether or divide X into two parts $(X^{(1)}, X^{(2)})^T$ and require Y_i to be independent of $X_j^{(1)}$ $(j \neq i)$ given $X_i^{(1)}$ and $X^{(2)}$. Except for small values of p, q, guidance from subject-matter considerations is highly desirable.

There are numerous extensions of the above idea, among them the following. If Y and X are the $p \times 1$ vectors of the same variables measured at two different times (two-phase panel study) then it would sometimes be reasonable to require that if a transformation to new variables is used it should be the same for both Y and X. That is we write $Y^* = AY$,

$X^* = AX$. Then

$$\text{cov}(Y^*, X^*) = A\Sigma_{yx}A^T, \qquad \text{cov}(X^*) = A\Sigma_{xx}A^T$$

and the regression coefficients of Y^* on X^* are

$$B_{Y^*X^*} = (A\Sigma_{yx}A^T)(A\Sigma_{xx}A^T)^{-1} = A\Sigma_{yx}\Sigma_{xx}^{-1}A^{-1}$$

and a possible requirement is that this is a diagonal matrix D, say, whose elements may be of either sign. That is

$$A(\Sigma_{yx}\Sigma_{xx}^{-1})A^{-1} = D. \tag{2.2}$$

When $\Sigma_{yx}\Sigma_{xx}^{-1}$ has distinct nonzero eigenvalues, there exists a matrix A satisfying (2.2) (Rao 1973, p.43) but, in general, the eigenvalues and eigenvectors will be complex and hence unsuitable for statistical interpretation.

By exploiting the knowledge that X and Y are the same variables measured at two time points there is a route which may sometimes circumvent such problems. It is conceivable in some such situations that the matrix of concentrations Σ^{yx} is nearly a diagonal matrix since such a diagonal form implies conditional independence of components Y_i, X_j for $(j \neq i)$ given all other variables. However, if Σ^{yx} is of diagonal form then the matrix of regression coefficients is a symmetric matrix since

$$B_{YX} = -(\Sigma^{yy})^{-1}\Sigma^{yx}.$$

This suggests to take first $X^{**} = (\Sigma^{yx})^{-1}X$ and to apply (2.2) to Y and X^{**}. This leads to

$$B_{YX^{**}} = -(\Sigma^{yy})^{-1} = -\Sigma_{yy.x},$$

which as a positive definite matrix permits an orthogonal decomposition $CB_{YX^{**}}C^T = T$, where $C^T = C^{-1}$ and T is a diagonal matrix with positive diagonal elements, leading finally to

$$Y^* = CY, \qquad X^* = CX^{**} = C(\Sigma^{yx})^{-1}X.$$

With more than two time points there are connections with multiple time series and cointegration analysis (Engle and Granger 1987) in econometrics. If we have vectors Y, X, V of respectively response, intermediate response, and explanatory variables further possibilities arise when transformations of one or both of (Y, X) are allowable. In the simpler problem with just two vectors Y, X the formal inclusion of squared and cross-product terms of original variables as components of Y could be employed primarily as a device for assessing the desirability of nonlinear transformation of the components. A more formal approach could be based on maximum likelihood estimation of a parametric family of transformations of the components of Y together with a matrix analogous to that in (2.1).

3. Adequacy of representation by covariance matrix

We now turn to the second general issue mentioned in Section 1. Is it adequate to describe the associations between the component variables by a covariance matrix? To some extent this amounts to testing multivariate normality, although in many practical contexts it is not so much the distributional form that is of primary concern as the possible existence of more complex forms of dependency not revealed by the covariance matrix, including the occurrence of outliers.

Again a distinction is to be drawn between methods that are invariant under nonsingular transformations of the observed vector Y and methods that retain the original components (Cox and Small 1978) and here we concentrate on the latter.

Two broadly contrasting approaches are the plotting of residuals (Anscombe, 1973) and the calculation of test statistics. For the former there is some evidence (Wermuth and Cox 1991) that rather than plotting y_i versus y_j $(i \neq j)$, a more sensitive analysis results from plotting the corresponding complete residuals r^i versus r^j $(i \neq j)$, where r^i is the deviation of y_i from its linear least-squares regression on all other variables. This hinges partly on the result that

$$\text{corr}(r^i, r^j) = -\rho_{ij.\{k_{ij}\}}, \tag{3.1}$$

where k_{ij} is the set of all variables other than i and j and, in the usual notation, the correlation coefficient on the right-hand side of (3.1) is the partial correlation between Y_i and Y_j given all other variables.

For more formal tests with all components on an equal footing, we may calculate

(*i*) the Student t statistic Q_{ij} for regression of Y_i on Y_j^2 adjusting for Y_j, i.e., in the regression of Y_i on both Y_j^2 and Y_j. This yields $p(p-1)$ statistics in all;

(*ii*) the Student t statistic $Q_{i(jk)}$ for regression of Y_i on Y_jY_k adjusting for Y_j, Y_k. This yields $\frac{1}{2}p(p-1)(p-2)$ statistics in all.

It is clearly desirable to reduce the number of such statistics and this can be done primarily via subject-matter considerations which may indicate concentration on particular subsets of Q_{ij} and $Q_{i(jk)}$. If there is strong linear regression present of Y_i on a particular variable or set of variables Y_r, it may be advisable to eliminate linear regression on Y_r from the statistics Q_{ij}, $Q_{i(jk)}$.

There are a number of approaches to the assessment of this array of statistics depending on the number of statistics for consideration and on the extent to which an approximate calculation of significance is important. We may consider the most extreme significance level adjusted for selection

or the whole set of values may be plotted against expected normal order statistics, the effect of correlation on such plots usually being small.

It may be helpful to arrange the 'squared' terms Q_{ij} in a square array, and to calculate row and column mean squares or absolute values to detect effects associated with a particular component. A similar purpose is achieved by summing $Q_{i(jk)}$ first over j,k for fixed i and then over i,k for fixed j.

If clear evidence is found via one or more of the statistics Q for nonlinear behaviour, subject-matter interpretation will involve first inspection of the corresponding scatter plots leading either to qualitative explanation, to transformation of variables or to the fitting of an explicit model.

It may at first sight seem odd to suggest calculating both Q_{ij} and Q_{ji}, e.g., in the case of two components X,Y to be treated on an equal footing regressing both Y on X^2 and X on Y^2. Some theoretical justification is provided below. It can be shown (Cox and Small 1978) that asymptotically

$$\operatorname{corr}(Q_{YX}, Q_{XY}) = \rho_{XY}(2 - 3\rho_{XY}^2), \tag{3.2}$$

so that the form

$$(Q_{YX}, Q_{XY}) \begin{pmatrix} 1 & r_{XY}(2 - 3r_{XY}^2) \\ . & 1 \end{pmatrix}^{-1} \begin{pmatrix} Q_{YX} \\ Q_{XY} \end{pmatrix} \tag{3.3}$$

has asymptotically a chi-squared distribution with two degrees of freedom, where r_{XY} is the sample estimate of ρ_{XY}. The same formula can be used to combine Q_{ij} and Q_{ji} $(i \neq j)$ in the general case.

Because (3.2) was given previously without proof we outline here the arguments involved. For independent pairs $(X_1,Y_1),\ldots,(X_n,Y_n)$, the numerator of the statistic Q_{YX} is

$$Q'_{YX} = \sum Y_j \left[X_j^2 - \tilde{m}_{02} - \frac{\tilde{m}_{03} - \tilde{m}_{02}\tilde{m}_{01}}{\tilde{m}_{02} - \tilde{m}_{01}^2}(X_j - \tilde{m}_{01}) \right], \tag{3.4}$$

where $\tilde{m}_{0r} = \left(\sum X_j^r\right)/n$ and the multiplier of Y_j is X_j^2 orthogonalized with respect to the vectors $\{X_j\}$ and $\{1\}$. The statistic Q_{YX} itself is (3.4) divided by a consistent estimate of its standard error. We may suppose without loss of generality that the bivariate normal distribution of (X,Y) has zero mean, unit variance and correlation ρ. We write $Y_j = \rho X_j + Z_j$; it follows from the orthogonalization that in (3.4) Y_j can be replaced by Z_j. Further as $n \longrightarrow \infty$, $\tilde{m}_{01} \longrightarrow 0$, $\tilde{m}_{02} \longrightarrow 1$, and $\tilde{m}_{03} \longrightarrow 0$ so that Q'_{YX} is approximately

$$\sum Z_j(X_j^2 - 1).$$

This and the corresponding expression for Q'_{XY} are sums of n independent

and identically distributed terms. Therefore, $\mathrm{corr}(Q'_{YX}, Q'_{XY})$ is the same as the correlation of a single term:

$$\mathrm{corr}(Q'_{YX}, Q'_{XY}) = \mathrm{corr}[(Y-\rho X)(X^2-1), (X-\rho Y)(Y^2-1)].$$

A direct calculation with the moments up to order 6 of the bivariate normal distribution gives the required result, hence verifying (3.2). A similar but more complicated calculation would yield the approximate covariance matrix of the full set of quadratic statistics.

We give here only the bivariate case although a general multivariate extension is available. The need for two statistics to assess nonlinear dependence between two variables treated on a symmetrical footing and a more formal justification of the above procedures can be obtained by taking as an alternative to the bivariate normal distribution the modification introduced by one correction term of an Edgeworth expansion. For convenience we suppose the random variables (X, Y) are scaled to zero mean and unit variances and have correlation ρ, and take the density in the asymmetric form (Barndorff-Nielsen and Cox 1979)

$$\phi_2(x, y; \rho)\{1 + \tfrac{1}{6}[\, \rho'_{30} H_3(x) + 3\rho'_{21} H_2(x) H_1(y') + 3\rho'_{12} H_1(x) H_2(y') + \rho'_{03} H_3(y')]\}, \tag{3.5}$$

where ϕ_2 is the standardized bivariate normal density of correlation ρ, $H_r(\cdot)$ are Hermite polynomials and ρ'_{rs} are standardized cumulants of the orthogonalized variables X and $Y' = (Y-\rho X)/\sqrt{(1-\rho^2)}$. This is slightly simpler for detailed calculation than the more symmetric formulation in tensorial polynomials (Barndorff-Nielsen and Cox 1989, Section 4). In terms of the standardized cumulants ρ_{rs} of X and Y, we have

$$\begin{aligned}
&\rho'_{30} = \rho_{30}, \qquad \rho'_{21} = (\rho_{21} - \rho\rho_{30})/\sqrt{(1-\rho^2)},\\
&\rho'_{12} = (\rho_{12} - 2\rho\rho_{21} + \rho^2\rho_{30})(1-\rho^2),\\
&\rho'_{03} = (\rho_{03} - 3\rho\rho_{12} + 3\rho^2\rho_{21} + \rho^3\rho_{30})/(1-\rho^2)^{\frac{3}{2}}.
\end{aligned} \tag{3.6}$$

The equivalent form of (3.5) in terms of Y and $X' = (X-\rho Y)/\sqrt{(1-\rho^2)}$ has coefficients ρ''_{30}, ρ''_{21}, ρ''_{12}, ρ''_{03} found by interchanging suffices in (3.6). Thus, $\rho''_{03} = \rho_{03}$, $\rho_{12} = (\rho_{12} - \rho\rho_{03})\sqrt{(1-\rho^2)}$, etc.

Integration of (3.5) gives that the marginal density of X is

$$\phi(x)[1 + \tfrac{1}{6}\rho'_{30} H_3(x)],$$

where ϕ is the standardized normal density. Thus to the first order in the ρ's the conditional density of Y' given $X = x$ is

$$\phi(y')\{1 + \tfrac{1}{6}[3\rho'_{21} H_2(x) H_1(y')] + 3\rho'_{12} H_1(x) H_2(y') + \rho'_{03} H_3(y')\}$$

so that the standardized third cumulant of the conditional distribution of Y is ρ'_{03} and

$$\mathrm{E}(Y' \mid X = x) = \tfrac{1}{2}\rho'_{21}(x^2 - 1),$$
$$\mathrm{var}(Y' \mid X = x) = (1 + \rho'_{12}x),$$
$$\mathrm{E}(Y \mid X = x) = \rho x + \tfrac{1}{2}\rho'_{21}\sqrt{(1-\rho^2)}(x^2 - 1),$$
$$\mathrm{var}(Y \mid X = x) = (1 - \rho^2)(1 + \rho'_{12}x),$$

and symmetrically

$$\mathrm{E}(X \mid Y = y) = \rho y + \tfrac{1}{2}\rho''_{12}\sqrt{(1-\rho^2)}(y^2 - 1),$$
$$\mathrm{var}(X \mid Y = y) = (1 - \rho^2)(1 + \rho''_{21}y).$$

Now (3.5) represents a distribution only approximately even for very small ρ's when the region of negative values of (3.5) has very small probability. Numerical work suggests that the formula is nevertheless capable of representing reasonably a useful range of distributional shapes. There are four parameters defining nonnormality, two representing skewness and two more directly concerned with dependence. Thus a test of bivariate normality locally against the full family (3.5) would involve four statistics leading, as the statistic corresponding to (3.3), to a chi-squared with four degrees of freedom. From a substantive viewpoint, distributional shape may be of relatively minor interest and this is the justification for a reduction to two statistics. This could be achieved in various ways, but the simplest is to restrict attention to distributions with zero marginal skewness, i.e., with $\rho_{03} = \rho_{30} = \rho'_{30} = \rho''_{03} = 0$; note that this is not the same as requiring zero skewness of the conditional distributions, which would require $\rho'_{03} = \rho''_{30} = 0$. In this special case, $\rho'_{21}\sqrt{(1-\rho^2)} = \rho_{21}$ and $\rho'_{12}(1-\rho^2) = \rho_{12} - 2\rho\rho_{21}$ so that for standardized variables

$$\mathrm{E}(Y \mid X = x) = \rho x + \tfrac{1}{2}\rho_{21}(x^2 - 1),$$
$$\mathrm{var}(Y \mid X = x) = (1 - \rho^2) + (\rho_{12} - 2\rho\rho_{21})x,$$
$$\mathrm{E}(X \mid Y = y) = \rho y + \tfrac{1}{2}\rho_{12}(y^2 - 1),$$
$$\mathrm{var}(X \mid Y = y) = (1 - \rho^2) + (\rho_{21} - 2\rho\rho_{12})y.$$

Note the double interpretation of ρ_{12}, ρ_{21} as determining quadratic regressions of means and linear regressions of variances; the special case $\rho = 0$ makes this clear; see Figure 1. In particular if Y is in some sense a response to X a nonzero value of ρ'_{21} can be interpreted as suggesting a nonlinear regression in the usual sense, whereas if X is a response to Y it can be interpreted as suggesting systematic changes in conditional variance. See the example of the diabetes data below.

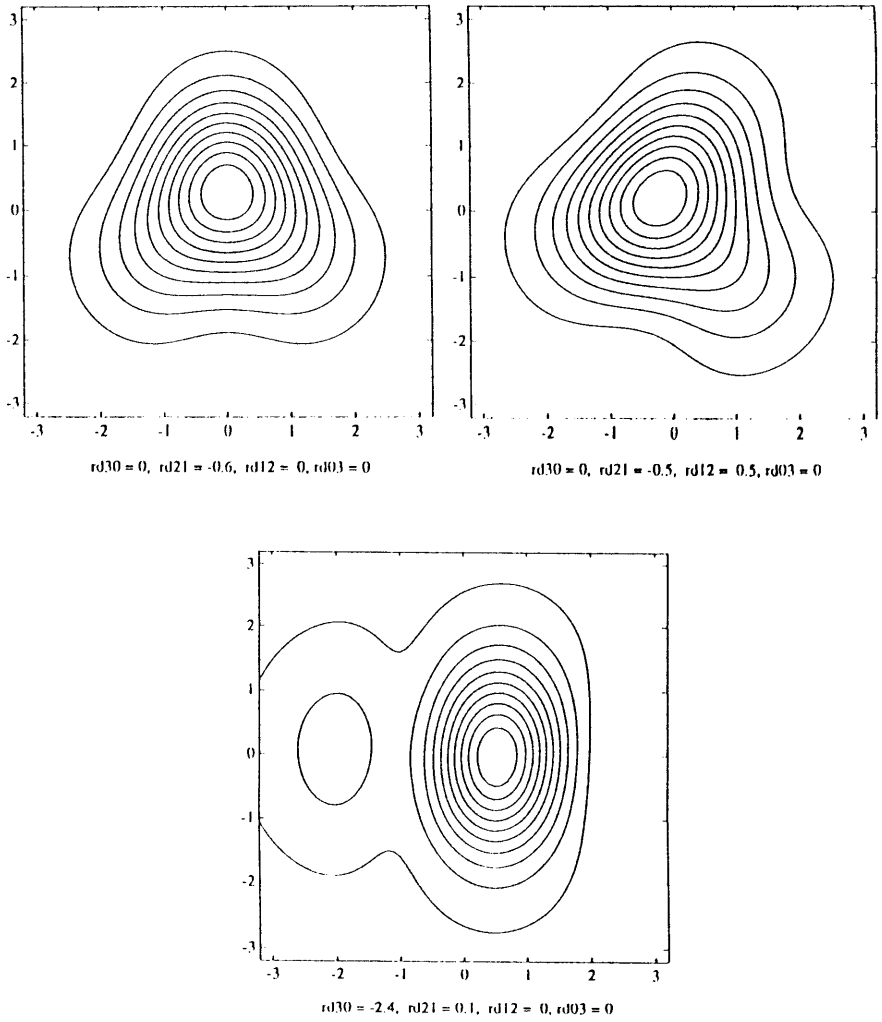

Fig. 1. Contours of density derived from Edgeworthexpansion with $\rho = 0$ and (top) $\rho_{30} = 0, \rho_{21} = -0.6, \rho_{12} = 0, \rho_{03} = 0$; (middle) $\rho_{30} = 0, \rho_{21} = -0.5, \rho_{12} = 0.5, \rho_{03} = 0$; (bottom) $\rho_{30} = -2.4, \rho_{21} = 0.1, \rho_{12} = 0, \rho_{03} = 0$

The appropriate choice of signs of ρ_{12}, ρ_{21} gives any combination of convexity and concavity in the two regressions. The likelihood associated with (3.5) in seven parameter form (i.e., with unknown mean and covariance matrix and small nonzero ρ_{12}, ρ_{21}) is best calculated locally by exponentiating the correction factor leading to the local (9,7) exponential family in which the sample mean and covariance matrix is augmented by the four third-order statistics ($\sum X_i^3$, $\sum X_i^2 Y_i$, $\sum X_i Y_i^2$, $\sum Y_i^3$). By invariance and or conditioning these can be replaced by the standardized marginal third cumulants and the quadratic regression coefficients of Y on X^2 and of X on Y^2. The marginal skewnesses are uninformative separately about ρ_{12} and ρ_{21}, although in principle there may be some information in their joint behaviour. Nevertheless, consideration of the relative variances suggest that, especially when ρ is small, the amount of information carried by third cumulants is likely to be small compared with that in the regression coefficients.

If in unstandardized units the regression coefficient of Y on X^2 adjusted for X is estimated as $\beta_{YX^2.X}$, then the corresponding estimate of ρ_{21}, assuming negligible skewness is

$$\hat{\rho}_{21} = 2\hat{\beta}_{YX^2.X}\hat{\sigma}_X^2/\hat{\sigma}_Y. \tag{3.7}$$

If the component variables in a multivariate normal distribution are dichotomized at their medians, the resulting multidimensional binary distribution is such that the frequencies in each of the resulting $2 \times 2 \times 2$ tables are equal in pairs and are given by simple functions of the marginal correlations (McFadden 1955). In particular, the three factor interaction terms in a log linear fit all vanish. This can be used to provide a simple test of whether a continuous multivariate distribution is consistent with a multivariate normal distribution with possible nonlinear transformation of the individual components.

In the diabetes data discussed below all standardized three factor interactions are small except for the one for the three variables isolated directly for nonlinear relations, where the studentized value is 2.2. Of course, this is short of the value needed to establish nonnormality on this basis alone; the form of the interaction is, however, precisely in line with the interpretation derived via the cross product Q statistics of (*ii*).

Example. From ongoing investigations of determinants of blood glucose control (Kohlmann et al. 1991, 1993) we have observations for 70 diabetic patients, all having less than 10 years of formal schooling. The variables are Y, a particular metabolic parameter, the glycosylated haemoglobin (abbreviated as GHb); X, a standardized score for particular knowledge about diabetes; W, duration of illness in months. Furthermore, three different

attitudes of the patients are measured as subscale sum scores of a questionnaire. The attitudes are to capture to whom or to what the patient attributes what is happening in relation to his illness: Z, social externality (powerful others are responsible); U, fatalistic externality (mere chance determines what occurs); V, internality (the patient sees himself as mainly responsible).

The observed correlations among these variables are given in Table 1 together with the correlations to two further constructed variables used to describe the type of nonlinearity of some of the relations. Even the largest correlations are still moderate in size, but this is not surprising since we consider variables for which large variations between persons are typical.

For these data there is no indication of deviations from linearity for the relations of variables X, Z, U, V, W from the Q statistics of (i) and (ii). However, in the regressions involving variable Y there are four statistics larger than 2: $Q_{WY} = -2.4$, $Q_{YZ} = 2.1$, $Q_{Y(ZW)} = 2.6$ and $Q_{Z(YW)} = 2.6$. Because our purpose is to seek a qualitative explanation we make no adjustments for the multiple tests involved.

The scatterplot in Figure 2 of standardized variables Y (good metabolic adjustment corresponding to low values) on W (duration of illness) shows a decreasing variability in Y over time with a concentration at a slightly better adjustment the longer the duration of the illness, i.e., the longer the patient's experience with the illness. For the contours added, a density derived from an Edgeworth expansion is assumed for Y regressed on W with $\rho = \rho_{30} = \rho_{03} = \rho'_{21} = 0$ and $\rho'_{12} = -0.243$ as estimated using (3.7).

One explanation of the nonlinear relations between Y, Z, W is that the dependence of metabolic adjustment Y on social externality Z changes with the duration of the illness, i.e.,

$$\hat{y} = \hat{\alpha} + (\hat{\beta} + \hat{\gamma}w)z + \hat{\delta}w = 14.0605 - (0.1466 - 0.0012w)z - 0.0376w.$$

The estimated values in this equation imply that adjustment is better if social externality is high in the early years of the illness, but it is worse if social externality is high for patients who have had the illness for many years. One rather tentative interpretation is that social externality measures to some extent the readiness of the patient to adhere strongly to the physicians advice: this appears to be helpful for a good metabolic adjustment in the early years of the illness but harmful the longer the illness has lasted.

A descriptive summary of the changes in the relations of Y and Z with time is given in Table 2.

Table 1.
Observed marginal correlations, means and standard deviations for 70 diabetic patients with less than 10 years of formal schooling.

Var.	Y	X	Z	U	V	W	$Z \times W$	Y^2
Y: GHb	1							
X: knowledge	−0.21	1						
Z: social externality	0.04	−0.33	1					
U: fatalistic externality	0.01	−0.49	0.38	1				
V: internality	0.21	0.00	−0.18	−0.37	1			
W: duration	−0.20	0.07	−0.01	0.04	0.06	1		
$Z \times W$:$(z-\bar{z})(w-\bar{w})$	0.25	−0.02	0.12	0.10	0.01	0.15	1	
Y^2:$(y-\bar{y})^2$	0.37	−0.19	0.00	0.04	0.19	−0.33	−0.01	1
Mean	9.32	0.00	26.9	20.9	40.3	127.3		
Standard deviation	2.23	1.00	7.04	6.47	5.16	86.38		

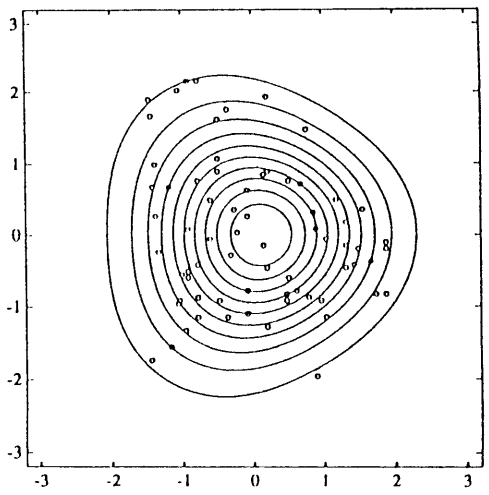

Fig. 2. Scatterplot of standardized values for metabolic adjustment Y and duration of illness W together with contours of a density derived from Edgeworth expansion

Table 2.
Changes in metabolic adjustment, GIIb (Y), with the duration of illness (W) and changes in its dependence on social externality (Z).

		Duration of illness in months		
	Overall	0–96	97–192	193–288
mean: GIIb	9.32	9.77	9.41	8.50
stand. dev.: GIIb	2.21	2.65	1.99	1.34
correlation of Y and Z	0.04	−0.32	0.37	0.53
number of obs.	70	31	20	19

4. Simplification of covariance structures

We now turn to issues in a sense complementary to those of Section 3. Suppose that dependency is adequately described by a $p \times p$ covariance matrix; is it helpful to have a more parsimonious representation?

Historically this has been approached from a number of somewhat interrelated points of view involving simplifications in terms of

(i) zero correlations or blocks of zero correlations, chosen possibly in the light of inspection of the sample correlation matrix;

(ii) simple block structure in the correlation matrix with, e.g., equal correlation between variables in a block and equal and different correlation between variables in different blocks;

(iii) zero elements in the concentration matrix;

(iv) linear relations involving latent (unobserved or hidden) variables;

(v) sets of conditional independencies among component variables, preferably expressing substantive research hypotheses (Wermuth and Lauritzen 1990).

Here (i) and (ii) represent special cases of linear covariance structures (Anderson 1973), (iii) leads to the covariance selection models of Dempster (1972), (iv) to factor analysis and linear structural models (Jöreskog 1973) and (v) is related to the substantial literature in econometrics and other fields, stemming in a sense from Wright's path analysis (Wright 1921, 1923). In the more formal framework of exponential family models, (i) and (ii) impose structure on the mean or moment parameters whereas (iii) is set out in terms of the canonical parameters and this leads to some simplifications in formal inference and in computing maximum-likelihood estimates.

Here we concentrate on (v), especially because this appears to be a fruitful way of introducing into the analysis important information both on the nature of the component variables and on the relations between them. The notion of representing the relations by graphs has been used to advantage in connection with expert systems exploiting the properties of Markov random fields. We sketch here related methods using, however, two different kinds of edge in the graphs and often dividing variables into blocks on the basis of a priori substantive knowledge (Cox and Wermuth 1993).

The conventions for constructing such graphs as illustrated in Figures 3 to 5 [where we use the notation for independence as introduced by Dawid (1979)] are in outline as follows:

(a) where possible, variables are classified as responses, intermediate responses, possibly at various levels, and explanatory variables; the nodes representing the variables of the above types arranged in boxes from left to right (in line with the notation for conditional probability);

(b) variables in the same box are to be regarded symmetrically, e.g., both

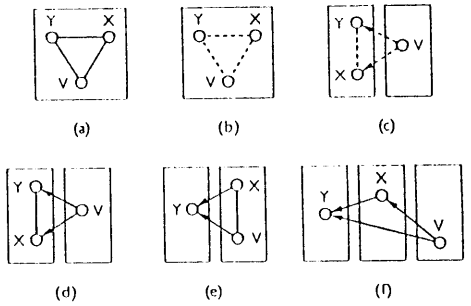

Fig. 3. Six distributionally equivalent ways of specifying a saturated model for three variables: (a) joint distribution of Y, X, V with three substantial concentrations; (b) joint distribution of Y, X, V with three substantial covariances; (c) multivariate regression chain model with regressions of Y on V and of X on V and with correlated errors; (d) block regression chain model with regressions of Y on X, V and of X on Y, V; (e) univariate regression of Y on X, V and joint distribution of X, V; (f) univariate recursive regression system with Y as response to X, V; X as intermediate response to V. For instance, graph (e) with double lines round the right-hand box represents the standard linear model for regression of Y on fixed explanatory variables X, V.

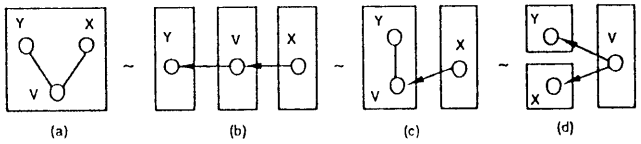

Fig. 4. Four distributionally equivalent ways of specifying $Y \perp\!\!\!\perp X \mid V$: (a) covariance selection model for Y, X, V having parameters $\rho_{yv.x} \neq 0$, $\rho_{xv.y} \neq 0$, and $\rho_{yx.v} = 0$; (b) univariate recursive regression model with $\beta_{yv.x} \neq 0$, $\beta_{yx.v} = 0$, $\beta_{vx} \neq 0$; (c) block regression chain model with Y, V as joint responses to X and with independent parameters $\rho_{yv.x} \neq 0$, $\beta_{yx.v} = 0$, $\beta_{vx.y} \neq 0$; (d) two independent regressions of Y on V and of X on V with $\beta_{yv} \neq 0$, $\beta_{xv} \neq 0$, $\rho_{yx.v} = 0$.

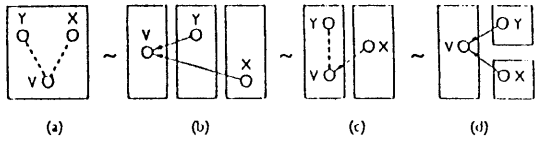

Fig. 5. Four distributionally equivalent ways of specifying $Y \perp\!\!\!\perp X$: (a) linear structure incovariances with $\rho_{yv} \neq 0$, $\rho_{xv} \neq 0$, $\rho_{yx} = 0$; (b) univariate recursive regression model with $\beta_{vx.y} \neq 0$, $\beta_{vy.x} \neq 0$, $\beta_{yx} = 0$; (c) multivariate regression chain model with $\rho_{yv.x} \neq 0$, $\beta_{vx} \neq 0$, $\beta_{yx} = 0$; (d) multiple regression of V on two independent regressors Y, X, with $\beta_{vy.x} \neq 0, \beta_{vx.y} \neq 0, \rho_{yx} = 0$.

as response variables, and joined, if at all, by an undirected edge, whereas variables in different boxes are joined, if at all, by directed edges, the arrow pointing from explanatory variable to response;

(c) there is at most one connecting edge between any pair of nodes;

(d) variables in one box are considered always conditionally on all variables in boxes to the right;

(e) if full lines are used as edges, variables are considered also conditionally on all other variables in the same box, whereas if dashed lines are used a response is considered marginalised over the responses in the same box;

(f) if there is no edge connecting two variables, the two variables are conditionally independent, the conditioning variables being as specified in (d) and (e);

(g) graphs are drawn with boxes to represent substantive research hypotheses, i.e., when the presence of an edge implies an association large enough to be of substantive interest rather than merely the absence of a zero constraint; such hypotheses thus representing models which are in some sense the simplest appropriate;

(h) if a right-hand box has two lines around it, the corresponding variables are regarded as fixed at their observed values, i.e., their distribution is not modelled;

(i) a row of unstacked boxes represents an ordered sequence of responses, intermediate responses and explanatory variables. If no order is implied the boxes are stacked.

When these ideas are applied to the diabetes data of Table 1, we start with a first classification of the variables into a sequence of dependencies and associations which is derived from substantive knowledge and from hypotheses about the variables and which is expressed here as the dependence chain in Figure 6 containing four elements, i.e., four boxes.

The strength and direction of the dependencies and associations as well as possible independencies are to be deduced from the estimates in the

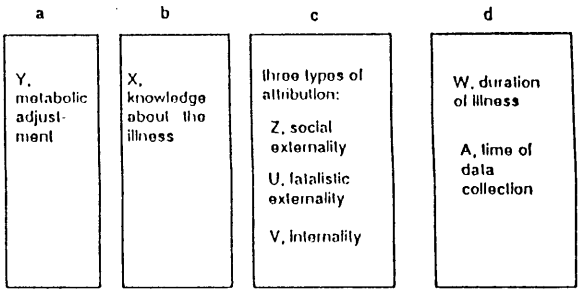

Fig. 6. A first classification of the variables into a sequence of dependencies and associations with Y as response of primary interest having all others as potential explanatory variables; with X as an intermediate variable considered conditionally given Z, V, U, W, A; with Z, V, U as joint intermediate variables on equal footing given W, A and with W, A as a background characteristics.

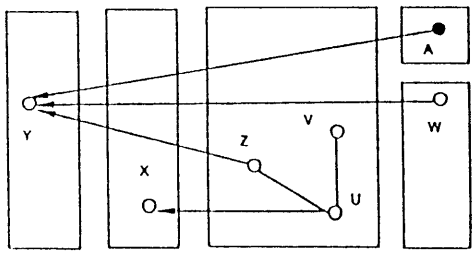

Fig. 7. Chain graph of dependencies for the diabetes data.

conditional analyses specified in this way. The independencies are displayed with the chain graph of Figure 7.

It shows in particular that of the explanatory variables considered the important direct influences for metabolic adjustment (Y) are social externality (Z), duration of illness (W) and time of data collection (A); that knowledge (X) depends directly only on fatalistic externality (U); that the attribution scores (Z, V, U) are jointly independent of the backgroundcharacteristics and that duration of illness (W) is independent of the time of data collection (A, 1990 and 1991). With A being a dichotomous variable its correlation and regression coefficients reflect differences among group means.

There is a interactive effect of social externality and duration on

metabolic adjustment as described before with Table 2 and (10). There is a further main effect of time of data collection: for the first 38 patients there is a much worse average metabolic adjustment (GIIb=9.9) than for the 32 patients (GIIb=8.6) observed one year later. It is conceivable that this is a consequence of the feedback from the first study which showed pronounced differences in metabolic adjustment for patients of lower and higher educational background. This could also explain the larger variability in the metabolic adjustment in the early years of the illness shown in Figure 2: there were almost no patients with less than 8 years of illness and good metabolic adjustment observed in 1990 but a substantial number of them one year later. In addition, only in the first study the typical metabolic adjustments are considerably better for long than for short durations of illness.

As expected, knowledge is better the lower the fatalistic attribution; fatalistic externality is positively correlated with social externality but negatively with internality; there is a weak negative correlation between social externality and internality, possibly implied by the other two relations. A check of the stability of these results will be possible when data for more patients are available in about a year.

5. Discussion

We have in this chapter concentrated on different methods which we judged to be useful in the analysis of multivariate observational data:

(1) Concentration on constructing new variables such that they have special relations of conditional independence with a set of explanatory variables. Thereby the explanatory variables either remain untransformed since their identity is to be preserved or they are transformed in the same way as the responses if they coincide with the responses but are measured at an earlier time;

(2) Concentration on checking and modelling different forms of nonlinear dependencies or outliers. In particular it is explained in which sense a nonlinear dependence in a regression of Y on X can correspond to a linear regression of X on Y having a systematic change with y of the variation of X given $Y = y$;

(3) Concentration on simplifying covariance structures with chain graphs implying conditional independencies. This permits one to integrate substantive knowledge about the roles of variables as responses, intermediate and explanatory variables and to distinguish between two different types of multivariate dependencies, between block regression and multivariate regression equations.

The discussion throughout the chapter is expository or given as results

needed to interpret the features of a specific set of data. Some extensions are most desirable, in particular to incorporate derived variables for several components of a dependence chain; to derive implications of conditional nonlinear relations after marginalizing; to obtain model formulations corresponding to multivariate regressions with discrete or possibly mixtures of discrete and continuous responses.

Acknowledgement

We are grateful to the British German Academic Research Collaboration Programme for supporting our joint work. We thank G.K. Reeves and G. Hommel for helpful comments.

References

Anderson, T.W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist.* **1**, 135–141.

Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd edition. Wiley, New York.

Anscombe, F.J. (1973). Graphs in statistical analysis. *Am. Statist.* **27**, 17–21.

Barndorff-Nielsen, O.E. and Cox, D.R. (1979). Edgeworth and saddle-point approximations with statistical applications (with discussion). *J. R. Statist. Soc. Ser. B* **41**, 279–312.

Barndorff-Nielsen, O.E. and Cox, D.R. (1989). *Asymptotic Techniques for Use in Statistics*. Chapman and Hall, London.

Cox, D.R. and Small, N.H.J. (1978). Testing multivariate normality. *Biometrika* **65**, 263–272.

Cox, D.R. and Wermuth, N. (1992a). Response models for mixed binary and quantitative variables. *Biometrika* **79**, 441–461.

Cox, D.R. and Wermuth, N. (1992b). On the calculation of derived variables in the analysis of multivariate responses. *J. Multivar. Anal.* **42**, 162–170.

Cox, D.R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs. To appear in *Statist. Sci.*

Dawid, A.P. (1979). Conditional independence instatistical theory (with discussion). *J. Statist. Soc. Ser. B* **41**, 1–31.

Dempster, A.P. (1969). *Elements of Continuous Multivariate Analysis*. Addison-Wesley, Reading, MA.

Dempster, A.P. (1972). Covariance selection. *Biometrics* **28**, 157–175.

Engle, R.F. and Granger, C.W.J. (1987). Cointegration and error correction; representation, estimation and testing. *Econometrica* **55**, 251–276.

Jöreskog, K.G. (1973). A general method for estimating a linear structural equation system. In: *Structural Equation Models in the Social Sciences*, A.S. Goldberger and O.D. Duncan (Eds.) 85–112. Seminar Press, New York.

Kohlmann, C.W., Krohne, H.W., Küstner, E., Schrezenmeir, J., Walther, U. and Beyer, J. (1991). Der IPC-Diabetes-Fragebogen: ein Instrument zur Erfassung krankheitsspezifischer Kontrollüberzeugungen bei Typ-I-Diabetikern. *Diagnostica* **37**, 252–270.

Kohlmann, C.W., Küstner, E. and Beyer, J. (1993). Kontrollüberzeugungen und Diabeteseinstellung in Anhängigkeit von der Krankheitsdauer. *Z. Gesundheitspsychologie* 1, 32–48.

Lauritzen, S.L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.* 17, 31–57.

McFadden, J.A. (1955) Urn models of correlation. *Ann. Math. Statist.* 26, 478–489.

Rao, C.R. (1973) *Linear Statistical Inference and its Appplications, 2nd edition.* Wiley, New York.

Wermuth, N. and Cox, D.R. (1991). On scatterplots for partial correlations. Berichte zur Stochastik und verw. Gebieten. 91-4, Universität Mainz.

Wermuth, N. and Cox, D.R. (1992). Derived variables calculated from similar responses: some characteristics and examples. To appear in: *Comp. Stat. Data Anal.*

Wermuth, N. and Lauritzen, S.L. (1990). On substantive research hypotheses, conditional independencegraphs and graphical chain models (with discussion). *J. R. Statist. Soc. Ser. B* 52, 21–72.

Wright, S. (1921). Correlation and causation. *J. Agric. Res. (Washington, D.C.)*, 20, 557–585.

Wright, S. (1923). The theory of path coefficients: a reply to Niles' criticism. *Genetics* 8, 239–255.