Reprinted from JOURNAL OF MULTIVARIATE ANALYSIS
All Rights Reserved by Academic Press, New York and London

Vol. 42, No. 1, July 1992
Printed in Belgium

On the Calculation of Derived Variables in the Analysis of Multivariate Responses

D. R. COX

Nuffield College, Oxford, United Kingdom

AND

NANNY WERMUTH

Psychologisches Institut, Universität Mainz, Germany

Communicated by the Editors

The multivariate regression of a $p \times 1$ vector Y of random variables on a $q \times 1$ vector X of explanatory variables is considered. It is assumed that linear transformations of the components of Y can be the basis for useful interpretation whereas the components of X have strong individual identity. When $p \geqslant q$ a transformation is found to a new $q \times 1$ vector of responses Y^* such that in the multiple regression of, say, Y_1^* on X, only the coefficient of X_1 is nonzero, i.e. such that Y_1^* is conditionally independent of $X_2, \ldots, X_q$, given X_1. Some associated inferential procedures are sketched. An illustrative example is described in which the resulting transformation has aided interpretation. © 1992 Academic Press, Inc.

1. INTRODUCTION

Many of the standard methods of multivariate analysis derived from the multivariate normal distribution are essentially invariant under nonsingular linear transformations. A typical example is canonical correlation (or canonical regression) analysis. Here the relation between a $p \times 1$ vector response variable Y and another $q \times 1$ vector X, either of responses or of explanatory variables, is studied by finding linear combinations of the components of Y and of X that are maximally related, the resulting analysis being essentially invariant under separate linear transformations of Y and of X.

Received September 9, 1991; revised January 4, 1992.

AMS 1980 subject classification: 62J10.

Key words and phrases: canonical analysis, conditional independence, derived variable, graphical chain model, multivariate linear model.

0047-259X/92 $5.00

This invariance is sometimes, but by no means always, appealing on subject-matter grounds. For example, linear combinations of log height and log weight may form derived variables that are entirely satisfactory for the interpretation of the effect of the "size" of individuals on various medical outcomes and the summarization of blood pressure may be best carried out via a combination of diastolic and systolic blood pressures (or their logarithms), which are themselves somewhat arbitrarily chosen summaries of the blood pressure cyclical variation. On the other hand, variables such as anger and anxiety express distinct concepts and while a linear combination of them may well arise in a multiple regression equation as an expression of their relative importance in contributing to a third variable, the formation of a new derived response variable as an arbitrary linear combination of anger and anxiety is much less appealing. These remarks could be paralleled in many fields.

In the present paper we consider problems in which arbitrary linear combinations of a $p \times 1$ variable Y are allowable but in which it is desired to preserve the distinctive individual component structure of a $q \times 1$ variable X. It is convenient to begin by treating X as random, although in some applications conditioning on the observed values will be called for, this having virtually no effect on the following arguments.

There are a number of ways in which the problem can be formalized. The simplest and the one on which we shall concentrate is as follows. Suppose that $p \geqslant q$. We seek a linear transformation from $Y = (Y_1, ..., Y_p)^T$ to new variables $Y^* = (Y_1^*, ..., Y_q^*)^T = AY$ such that in the multiple regression of Y_s^* on X only the coefficient of X_s is nonzero ($s = 1, ..., q$); that is, Y_s^* is conditionally independent of all the X_t ($t \neq s$) given X_s. That is, in a sense Y_s^* is that derived response variable especially tied to X_s. In Section 2 we discuss the relatively simple situation in which $p = q$, extending the discussion to the case $p > q$ in Section 3. The construction is possible if and only if Σ_{yx} is of full rank. Section 4 outlines the inferential problems associated with the procedure and Section 5 described a specific application.

2. Equal Dimensionality

Suppose first for simplicity that $p = q$, i.e., that the dimensionalities of the two vectors are the same. For simplicity assume both vectors have zero mean and partition the joint covariance matrix Σ in terms of

$$\Sigma_{xx} = \operatorname{cov}(X) = E(XX^T),\ \Sigma_{yx} = \operatorname{cov}(Y, X) = E(YX^T) = \Sigma_{xy}^T,$$

$$\Sigma_{yy} = E(YY^T).$$

Now $\text{cov}(Y^*, X) = A\,\Sigma_{yx}$ and the matrix of regression coefficients of Y^* on X is thus

$$B_{y^*x} = A\Sigma_{yx}\Sigma_{xx}^{-1} = AB_{yx},$$

where B_{yx} is the matrix of regression coefficients of Y on X. We require this to be diagonal and if new variables are scaled to have unit regression coefficients on the explanatory variables B_{y^*x} must be the $q \times q$ identity matrix, so that

$$A = B_{yx}^{-1} = (\Sigma_{yx}\Sigma_{xx}^{-1})^{-1} = \Sigma_{xx}\Sigma_{yx}^{-1}.$$

Thus the new variable is given by

$$Y^* = \Sigma_{xx}\Sigma_{yx}^{-1}Y. \tag{1}$$

Note that $\text{cov}(Y^*, X) = E(Y^*X^{\mathrm{T}}) = \Sigma_{xx}$; i.e., the new variables are such that they have the same covariance matrix with X as does X with itself. Another interpretation is via the equation $Y = B_{yx}Y^*$.

The new variable Y^* exists and is uniquely defined provided that Σ_{yx} is nonsingular. A necessary and sufficient condition for this is that no linear combination of the components of Y be uncorrelated with all the components of X. Singularity will occur for some simple patterns, as for example when all cross-correlations are equal. In the singular case certain components of Y^* may nevertheless be determined.

There are other criteria that might be used to express the notion that each component of the transformed vector is attached to a unique component of X. For example, one might require that the sth component of the new vector have zero marginal correlation with all components of X except the sth. We shall not explore this further.

The special case $q = 2$ throws some light on the above formulae. If we normalize all variables to unit standard deviation, i.e., replace covariance matrices by correlation matrices, we need the condition $\rho_{x_1y_1}\rho_{x_2y_2} - \rho_{x_1y_2}\rho_{x_2y_1} \neq 0$ for nonsingularity. Subject to this and ignoring a constant of proportionality we can take

$$Y_1^* = (\rho_{x_2y_2} - \rho_{x_1x_2}\rho_{x_1y_2})\,Y_1 - (\rho_{x_2y_1} - \rho_{x_1x_2}\rho_{x_1y_1})\,Y_2,$$
$$Y_2^* = -(\rho_{x_1y_2} - \rho_{x_1x_2}\rho_{x_2y_2})\,Y_1 + (\rho_{x_1y_1} - \rho_{x_1x_2}\rho_{x_2y_1})\,Y_2.$$

Note that, for example, Y_1^* depends only on Y_1 if and only if the partial correlation of Y_1 with X_2 given X_1 vanishes, as is clear on general grounds. Note also that if X_1 and X_2 are uncorrelated the derived variables take a simple form, namely

$$Y_1^* = \rho_{x_2y_2}Y_1 - \rho_{x_2y_1}Y_2, \qquad Y_2^* = -\rho_{x_1y_2}Y_1 + \rho_{x_1y_1}Y_2.$$

3. Unequal Dimensionality

Suppose now that $p > q$. It can be shown that there is no unique linear combination Y_1^*, say, with the required property of dependence only on X_1. A sensible approach is to first reduce Y to the $q \times 1$ vector of canonical variables that contain the regression on X; for this we use the theory of canonical correlation or regression. Then the results of Section 2 can be applied to the new variables. The resulting procedure will choose from the multiplicity of solutions that Y^* with maximal regression on X. Indeed a nonsingular transformation can be made to the canonical variables supplemented by a set of $p-q$ random variables independent of the canonical variables and moreover independent of X. Under normal theory it follows from the factorization of the likelihood and the resulting sufficiency that inference about the dependence of Y on X should involve Y only via the canonical variables.

Now the q canonical variables formed from Y for capturing the regression on X are $c_1^T Y, \ldots, c_q^T Y$, where the c_j are eigenvectors of $\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$ (Rao [2, Sect. 8f]), the eigenvalues being the corresponding squared canonical correlations. It is not necessary to impose any particular normalization on the c_j, although it is convenient for exposition to require that $c_j^T \Sigma_{yy} c_j = 1$; it is known that $c_j^T \Sigma_{yy} c_k = 0$ $(j \neq k)$. We now transform from Y to the $q \times 1$ vector $Z = C^T Y$, where C is the $p \times q$ matrix $(c_1 \cdots c_q)$.

Then the covariance matrix of Z is the identity and

$$\operatorname{cov}(Z, X) = E(ZX^T) = C^T \Sigma_{yx}.$$

If now we apply the results of the previous section to the regression of Z on X we obtain a new variable given by

$$\begin{aligned} Y^* &= \Sigma_{xx}(C^T \Sigma_{yx})^{-1} Z \\ &= \Sigma_{xx}(C^T \Sigma_{yx})^{-1} C^T Y; \end{aligned} \tag{2}$$

it is easily verified that when $p = q$ (2) reduces to (1), then Y^* does not depend on the particular normalization of C, and that, as before,

$$\operatorname{cov}(Y^*, X) = \Sigma_{xx}.$$

We are very grateful to the referee for deriving by a different route involving the Loewner of matrices the solution

$$Y^* = \Sigma_{xx}(\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx})^{-1}\Sigma_{xy}\Sigma_{yy}^{-1} Y. \tag{3}$$

To see that (2) and (3) are in fact identical, note that by the properties

of canonical variables (Rao [2, Sect. 8f 1.2 and 1.5]) there exists a nonsingular matrix F such that

$$C^{\mathrm{T}}\Sigma_{yx}F=\Gamma_q, \qquad \Sigma_{yy}^{-1}\Sigma_{yx}F=C\Gamma_q, \qquad F^{\mathrm{T}}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}F=\Gamma_q^2,$$

where Γ_q is the diagonal matrix of canonical correlations. Thus, on using these three results in turn, we have that the right-hand side of (2) is equal to

$$\begin{aligned}\Sigma_{xx}(\Gamma_q F^{-1})^{-1}C^{\mathrm{T}} &= \Sigma_{xx}(F\Gamma_q^{-1})\,\Gamma_q^{-1}F^{\mathrm{T}}\Sigma_{xy}\Sigma_{yy}^{-1}\\ &= \Sigma_{xx}FF^{-1}(\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx})^{-1}(F^{\mathrm{T}})^{-1}F^{\mathrm{T}}\Sigma_{xy}\Sigma_{yy}^{-1}\end{aligned}$$

which reduces to the right-hand side of (3). The form (3) has the major advantage of avoiding the eigen analysis involved in (2); on the other hand, in applications, we have found it wise always to compute the canonical correlations as some check that the smallest of them is large enough to make Y^* reasonably well defined.

It can be shown by calculating the covariance matrix of Y^* that if the explanatory variables are uncorrelated, then the derived variables also are uncorrelated.

When the number of explanatory variables exceeds the number of responses, $p<q$, it will be necessary to select p or fewer variables or combinations from the q before applying the method.

4. Inference

The above results are for probability distributions. For application to data we shall assume multivariate normality, at least of Y given X, and therefore replace all population covariance matrices by the corresponding matrices of mean sums of squares and products. We regard the method as primarily a way of suggesting relatively simple derived variables and therefore to be used rather flexibly; thus elaborate discussion of formal inference procedures would be out of place. Nevertheless some simple results are available.

When $p=q$, we can obtain a confidence cone for the coefficients of, say, the first component of Y^*. Let it be hypothesized that $a^{\mathrm{T}}Y$ is conditionally independent of $X_2, \ldots, X_p$ given X_1. This can be tested in the multiple regression of $a^{\mathrm{T}}Y$ on X by a standard F test with degrees of freedom $(p-1, n-p-1)$, where n is the number of observations. A confidence cone can be formed from the subset of a not rejected in such a significance test.

The hypothesis that Y_1^*, say, does not depend on a particular set of components of Y, say the last t components $Y_{p-t+1}, \ldots, Y_p$, can be tested by

checking that the multivariate regression of $Y_1, ..., Y_{p-t}$ on X contains no contribution from $X_2, ..., X_p$ using any of the standard multivariate analysis of variance test statistics, e.g. the determinantal ratio (Rao [2, Sect. 8c.3]). If $t = 1$, i.e., only one component is hypothesized to be missing from the derived response, a standard F test is available.

The hope in using the present method will often be that one can find quite simple linear combinations of the components of the original Y that can replace the Y^* and that have a specific subject-matter interpretation. Simplicity here may mean that the coefficients defining Y have a simple interpretation or that each component Y_s^* involves only a limited number of the components of the original Y.

In some, but by no means all, cases the latter argument can be used as an alternative to the introduction of the canonical variables of Section 3. For example, suppose that $p = 3$, $q = 2$, and that on substantive grounds it is suspected that Y_1 is conditionally independent of Y_2 given (Y_3, X_1, X_2), that Y_1 is conditionally independent of X_2 given (Y_3, X_1), and that Y_2 is conditionally independent of X_1 given (Y_3, X_2). Suppose further that these relations are consistent with the data. Then an alternative to the use of canonical correlation as a method of reducing the dimension of Y from three to two is to restrict Y_1^* to be a combination of Y_1 and Y_3 and at the same time Y_2^* to be a combination of Y_2 and Y_3. It can then be verified that the appropriate combinations are, in the standard notation for conditional covariances,

$$\begin{aligned} Y_1^* &= Y_1 - Y_3 \sigma_{Y_1 X_2 \cdot X_1} / \sigma_{Y_3 X_2 \cdot X_1}, \\ Y_2^* &= Y_2 - Y_3 \sigma_{Y_2 X_1 \cdot X_2} / \sigma_{Y_3 X_1 \cdot X_2}. \end{aligned} \tag{3}$$

This relates the present analysis to the study of graphical models of conditional independency (Lauritzen and Wermuth [1]). We shall not explore this further here; in the example to be discussed in Section 5 this approach leads to essentially the same answer as reported there obtained by the method of Section 3.

5. An Example

Table I summarizes key aspects of observations obtained on 40 patients who have not received a preoperative treatment. There are three variables measured directly before an operation, the log concentrations of the three fatty acids palmitic acid, Y_1, linoleic acid, Y_2, and oleic acid, Y_3, and for the present purpose these form the response variable, Y. There are two explanatory variables forming the vector X, blood sugar measured the

TABLE I

Observed Marginal Correlations, Means, and Standard Deviations for 40 Patients

Variable	Y_1	Y_2	Y_3	X_1	X_2
Log palmitic acid (Y_1)	1				
Log linoleic acid (Y_2)	0.90	1			
Log oleic acid (Y_3)	0.95	0.92	1		
Blood sugar (X_1)	−0.25	−0.27	−0.32	1	
Sex (X_2)	0.28	0.43	0.23	−0.03	1
Mean	4.91	4.26	4.88	80.93	0.05
Standard deviation	0.3726	0.4745	4.4073	9.1661	1.02

morning before the operation, X_1, and sex, X_2, the latter coded as 1 for females and -1 for males. Log concentrations are used partly because the concentrations themselves are positive variables with large coefficients of variation around 50% and hence with very skew distributions and partly because linear concentrations of logs with simple numerical coefficients may be hoped to have a simple interpretation.

We apply the results of Section 3 with $p=3$, $q=2$. The two nonzero canonical correlations between Y and X are 0.60 and 0.38, neither being near zero. This points to an appreciable relation between the derived responses Y_1^* and Y_2^* and the corresponding explanatory variables X_1 and X_2.

The transformation matrix obtained from (2) is

$$\begin{pmatrix} 110.3 & 17.5 & -163.5 \\ -3.0 & 8.1 & -9.7 \end{pmatrix}$$

and suggests taking as simple forms of derived variable $Y_1^* = Y_1 - Y_3$ and $Y_2^* = Y_2 - Y_3$. This implies that the ratio of palmitic to linoleic acid is primarily connected to blood sugar as is the ratio of linoleic to oileic acid to sex.

Table II gives the correlation matrix of Y_1^*, Y_2^*, X_1, X_2. In this particular example the derived variables turn out to be nearly uncorrelated and this, together with the negligible correlation between X_1 and X_2, yields a very simple structure in which (Y_1^*, X_1) are completely independent of (Y_2^*, X_2). A likelihood ratio test of consistency with this structure yields chi-squared of 0.59 with 4 degrees of freedom.

Thus the 9 nonnegligible correlations of the original variables have been reduced to a simple structure with just two appreciable correlations. This

TABLE II

Observed Marginal Correlations for the Derived Responses and the Explanatory Variables of Table I

Variable	Y_1^*	Y_2^*	X_1	X_2
$Y_1^* = Y_1 - Y_3$	1			
$Y_2^* = Y_2 - Y_3$	0.09	1		
X_1	0.32	0.01	1	
X_2	0.09	0.57	−0.03	1

is a simplification special to this problem consequent on the essentially independent explanatory variables.

In general our procedure with $p = 3$, $q = 2$ imposes two conditional independencies, namely that Y_1^* is conditionally independent of X_2 given X_1 and that Y_2^* is conditionally independent of X_1 given X_2, leading usually to a nondecomposable independency structure (Wermuth and Cox, [3]) in the multivariate regression of Y^* on X requiring iterative fitting for maximum likelihood estimation.

The analysis was repeated on a different set of 40 patients for whom the same variables but quite different correlations were observed. The method based on (2) yielded essentially the same derived variables.

The computations were done using MATLAB.

6. Discussion

While the method has been reasonably successful on the above example, it may often prove ineffective, even when the broad formulation in terms of linear combinations of Y that preserve the individual structure of X is quite appealing. The method will work best when the q canonical correlations are all reasonably large and there is no strong collinearity between the columns of X. In other cases only some of the components of the transformed vector may be reasonably well defined. For such reasons, it essential, as indeed with other relatively advanced methods, to have checks that the method is in some sense reasonably effective.

It should be verified that the derived variables do have appreciable regression on their target x-components, and if one or more of the canonical correlations is small, say less than 0.1–0.2, it is unlikely that all the components of the derived response will be effective.

Acknowledgments

We are grateful to Dr. A. Theisen, University of Mainz, for the data of Section 5, to the referee for the result referred to in Section 3, and to the Anglo German Academic Research Collaboration Programme for supporting our joint work.

References

[1] 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.

[2] Rao, C. R. (1973). *Linear Statistical Inference and Its Applications.* 2nd ed. New York: Wiley.

[3] Wermuth, N., and Cox, D. R. (1991). Explanations for nondecomposable hypotheses in terms of univariate recursive regressions including latent variables. In *Berichte zur Stochastic und verwandten Gebieten 91–2, Universität Mainz.*

Printed by Catherine Press, Ltd., Tempelhof 41, B-8000 Brugge, Belgium