Derived variables for longitudinal studies

D. R. Cox[†‡] and Nanny Wermuth[§]

[†]Nuffield College, Oxford OX1 1NF, United Kingdom; and [§]Zentrum für Umfragen, Methoden und Analysen (ZUMA), 68072 Mannheim, Germany

Contributed by D. R. Cox, August 26, 1999

Suppose that for each individual a vector of features is measured at a number of time points. We look for a transformation of the features, the same at all time points, that will induce a simple dependency structure. In the simplest situation this requires that a certain asymmetric matrix has real nonzero eigenvalues. Extensions are considered.

In investigations in which several or many features are measured on each unit of study it is common to simplify analysis and hopefully assist interpretation by transforming the features to a set of derived variables. Often these are chosen from subject-matter knowledge or experience, a simple example being the replacement of body mass and height by Quetelet's index, mass divided by height squared. In other situations empirical analysis of the current data may guide the choice of derived variables. This may be by means of analysis of internal structure as in principal component analysis or by means of analysis of dependency, external analysis. The most common external analysis uses Hotelling's method of canonical variables (1), in which the derived variables are chosen to maximize measures of dependency between two sets of variables. Another possibility is to transform so that a simple form of dependency, for example the one in econometrics called the seemingly unrelated regression structure of Zellner (2), is achieved. Cox and Wermuth (3) developed the theory of this approach and as one of several examples showed that the dependency between the concentrations of three fatty acids and patient characteristics in a pain clinic could be captured by some simple linear combinations of log concentrations. In the present paper the argument is extended to deal with longitudinal data in which the same features are measured at more than one time point.

Suppose then that at time t a $p \times 1$ vector is measured on each study individual and is represented by the random column vector $Y_t = (Y_{t1}, \ldots, Y_{tp})^T$. We study first the dependence of Y_t on Y_{t-1}. We deal with linear dependencies described by a $p \times p$ matrix of population least squares regression coefficients $B_{t,t-1}$, where the jth row specifies the regression coefficients of responses Y_{tj} in the population least-squares multiple regression analysis of the components of Y_t on the full set of Y_{t-1}. We have, on writing C_{rs} for $\mathrm{cov}(Y_r, Y_s)$, that

$$B_{t,t-1} = C_{t,t-1}(C_{t-1,t-1})^{-1}, \qquad [1]$$

where the covariance matrices are $p \times p$. Now suppose that both vectors are transformed by the same linear transformation A_t to $Y^*_s = A_t Y_s$ for $s = t, t-1$. We now regress Y^*_t on Y^*_{t-1} to obtain the matrix of least-squares regression coefficients

$$B^*_{t,t-1} = \mathrm{cov}(Y^*_t, Y^*_{t-1})\{\mathrm{cov}(Y^*_{t-1})\}^{-1} = A_t B_{t,t-1} A_t^{-1}.$$

Choose A_t so that $B^*_{t,\,t-1}$ is a diagonal matrix, D_t say. This amounts to requiring that in a least-squares sense Y_{tj} is conditionally independent of $Y_{t-1,k}$ $(k \neq j)$ given $Y_{t-1,j}$, and this can be regarded as one time-related version of the seemingly unrelated regression property. We thus require

$$A_t B_{t,t-1} = D_t A_t,$$

so that the rows of A_t are the left eigenvectors of $B_{t,t-1}$ corresponding to the elements of D_t as eigenvalues. If we arrange the elements of D_t in order of decreasing absolute value, that matrix is an invariant of the system under linear transformation of the original vectors; the elements have a direct interpretation as correlation coefficients across time between the transformed components. The rows of A_t can be standardized in any convenient way.

In analyzing data we replace the population regression coefficients by the corresponding sample estimates $\hat{B}_{t,t-1}$ leading to estimates $\hat{D}_t$ and $\hat{A}_t$.

Even in this simplest situation with just two time points a number of issues arise.

First, the matrix $\hat{B}_{t,t-1}$ is not symmetric and some of the eigenvalues may be complex. If in some sense roots are significantly complex, then the proposed structure is incompatible with the data. It is unclear just what aspect of the dependency would lead to this conclusion. If a conjugate pair of complex eigenvalues had only small imaginary part it would be possible to set that part to zero and to proceed with the resulting pair of real values.

Next it is possible that some of the eigenvalues are very small, effectively zero. This would signal that the dependency can be captured in a reduced number of dimensions.

In applications, as typical with these kinds of multivariate analysis, it will often aid interpretation to replace the elements of $\hat{A}_t$ by simpler quantities—for example, simple integer multiples of log concentrations in the instance mentioned above—or to replace small values by zero. Further, it will be wise to look for possible nonlinearities in the dependency structure and to deal with these, for example by nonlinear transformation of the initial features.

In all these issues there are associated problems of formal statistical inference—i.e., of assessing the precision of quantities derived from $\hat{B}_{t,t-1}$. These are probably best addressed by simulation or data splitting.

We now consider a number of developments which are potentially useful in a largely exploratory sense.

First, there may be data at time points $0, 1, \ldots, m$ on the same set of individuals. We may then apply the above results with $t = 1, \ldots, m$ leading to m matrices of regression coefficients and to m estimated transformation matrices $\hat{A}_t$, for $t = 1, \ldots, m$ and m sets of eigenvalues. There are now many possible forms of interesting stability of structure that might arise. Thus, except for sampling error, the whole system might be consistent with a single matrix of regression coefficients and hence to a single A and D; note though that the means and residual covariance matrices could still vary arbitrarily across time. A weaker form of stability would involve constant A and changing D_t and a weaker still version would have only some rows of A_t stable.

Next, even with just three time points we may explore the Markov character of the dependence, in its strongest form examining whether Y_t is conditionally independent of $Y_{t-2}, \ldots$ given Y_{t-1}. If this property in its entirety fails it is possible that the non-Markovian character might be captured in a few components of the transformed variable Y^*_t. Another possibil-

[‡]To whom reprint requests should be addressed. E-mail: david.cox@nuf.ox.ac.uk.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

ity is that the transformed components are individually second-order Markov but that the conditional independencies in the new coordinate system are preserved in that Y^*_{tj} is conditionally independent of $Y^*_{t-1,k}$, $Y^*_{t-2,l}$ $(k, l \neq j)$ for all j. By the same argument as before this requires the strong condition that

$$(A_tC_{t,t-1}A_t^T, A_tC_{t,t-2}A_t^T) = (D_{t1}D_{t2}) \cdot \begin{pmatrix} A_tC_{t-1,t-1}A_t^T & A_tC_{t-1,t-2}A_t^T \\ A_tC_{t-1,t-2}A_t^T & A_tC_{t-2,t-2}A_t^T \end{pmatrix},$$

where D_{t1}, D_{t2} are diagonal matrices.

It can be shown that this requires the rows of the matrix A_t to be simultaneously the left eigenvectors of two different matrices, namely

$$(C_{t,t-1}C^{-1}_{t-1,t-2} - C_{t,t-2}C^{-1}_{t-2,t-2}) \cdot (C_{t-1,t-1}C^{-1}_{t-1,t-2} - C_{t-1,t-2}C^{-1}_{t-2,t-2})^{-1}$$

and

$$(C_{t,t-1}C^{-1}_{t-1,t-1} - C_{t,t-2}C^{-1}_{t-1,t-2}) \cdot (C_{t-1,t-2}C^{-1}_{t-1,t-1} - C_{t-2,t-2}C^{-1}_{t-1,t-2})^{-1}.$$

Often in addition to the vectors Y there will be at baseline a $q \times 1$ vector X of explanatory variables. The simplest procedure is then to use the above procedures, taking all covariance matrices residual to least-squares regression on X, and then to study how the derived variables Y^* depend on X. Andrew Roddam (personal communication) has applied this idea to a study in which X refers to maternal characteristics and with $p = 2$ the vector Y consists of log height and log body mass of infants at a number of ages up to 5 yr.

The above discussion is aimed at problems in which a considerable number of study individuals are measured at a limited number of time points. There are no assumptions of stationarity; indeed, the means of the various features may vary arbitrarily across time. If applied to a single long realization of a stationary vector time series, the technique could be used for estimating the matrix of regression coefficients by means of the lag zero and lag one matrices of cross-correlations.

1. Hotelling, H. (1936) *Biometrika* **28,** 321–377.
2. Zellner, A. (1962) *J. Am. Stat. Assoc.* **57,** 348–368.
3. Cox, D. R. & Wermuth, N. (1992) *J. Multivariate Analysis* **42,** 162–170.