International Workshop on
Multidimensional Data Analysis
Meeting of Dutch & Italian Schools
October 2 — 5, 1991
Anacapri, Italy

PREDICTABLE LINEAR COMBINATIONS FROM CATEGORICAL TIME SERIES

Stef van Buuren

TNO Institute of Preventive Health Care, Leiden

This paper outlines an approach to analyze categorical time series by means of optimal scaling. We study an extension of the canonical analysis proposed by Box and Tiao (1977) to categorical data. Briefly, the canonical analysis aims to transform an m–dimensional autoregressive process into m univariate components that are ordered from most to least predictable. Predictability is defined as the proportion of explained variance obtained by a p–th order autoregressive model fitted on the components. The most predictable components may serve as useful summary aggregates.

Keywords: categorical time series, canonical correlation, predictable components, exploratory data analysis

INTRODUCTION

A central problem in the study of multivariate time series is the reduction of dimensionality. For cross–sectional data, many methods exist to achieve a reduction of information, but multivariate time series are more difficult to analyze since two different types of relationships are generally of interest, namely relations *within* time and relations *between* time. Within–time relations apply to the values of different variables within the same point of time and thus ignores any dependencies among the observational units. This is a standard MVA problem and it is usually solved by constructing useful linear combinations of the data. Between–time relations are studied by means of time series models like the univariate Box–Jenkins ARMA methodology (Box and Jenkins, 1976). The specific problem of multiple time series is that relations *between series* as well as relations *between time* are of interest, so a summary of the data should adequately reflect both facets. The present paper outlines a kind of component analysis for time series which will look for linear combinations, called predictable components, that embody both between–series as well as between–time aspects.

The majority of time series techniques is based on numerical data. In practice however, data can be measured on nominal, ordinal and numerical scales, or mixtures of these. Some examples in social science research are: sequential behavioral observations in a laboratory setting, presence and absence of assumed therapeutic factors, psychophysiological measures, histories of life events and diary data. In order to be able to deal with this type of data the method allows for categorical data by means of optimal scaling (cf. Gifi, 1990). Briefly, optimal scaling replaces each category by one or more numerical values which are optimal in some sense. The present paper defines optimality in terms of contribution to the predictable components.

This paper is concerned with finding linear combinations of the observed series that capture important time dependent information. We call these aggregates predictable components. The aggregate which depends most heavily on the past, and thus is best predictable, may serve as a useful summary of the data. Like PCA, the second aggregate has maximal predictability conditional on orthogonality with the first, and so on. From a dimension reduction point of view, the information contained in the first few dimensions will be of primary importance. The technique can be used as a dimension reduction device to bring out the major time dependent characteristics of a multivariate data set.

Since lower dimensions capitalize on high autocorrelations, the first few predictable components will in general exhibit very smooth behavior, while the last few components tend to irregular white noise patterns. The transform can also be regarded as a multivariate smoother in which the most predictable component defines an optimally smoothed linear combination of the observed series. An obvious use of the technique is to identify those components that can serve as smoothed indicators of overall growth in for example the stock market.

METHOD

Let the $n \times m$ matrix $X = [x_{tj}]$ ($t = 1,\ldots,n$; $j = 1,\ldots,m$) contain m quantitative measurements sampled at n points of time. It is convenient to assume that $1'X = 0$, i.e. observations are given in deviations for their means. Subscript t is used to index time and subscript j is used to index series. The m–vector x_t represents the data at time t and the n–vector x_j denotes series j. Both x_t and x_j are column vectors. In the sequel we use of the backshift matrix B. The backshift matrix applies to finite series and it is functionally equivalent to the infinite backshift operator found in many textbooks. The backshift matrix is defined as the n x n matrix

$$B = \begin{bmatrix} 000... & 00 \\ 100 & 00 \\ 010 & . \\ 001 & . \\ .. & . \\ . & \\ . & \\ 000 & 00 \\ 000... & 10 \end{bmatrix}$$

Premultiplication of a series x_j results in the lagged series Bx_j. The t–th row of Bx_j contains the observation at time t - 1. Note how end effects are treated. The first observation of the lag becomes zero and the final observation at time t = n is lost. This way to handle end effects is consistent with the standard methods to compute autocovariances, autocorrelations and so on. Multiplying B by itself yields higher order backshift matrices. For example $B_2 = BB$ defines a second order backshift matrix. The zero order backshift B_0 is defined as the n x n identity matrix and B' is the forward shifting matrix.

Suppose that x_t can be modelled by the multiple autoregressive process

$$x_t = \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + \ldots + \Phi_p x_{t-p} + e_t, \qquad (1)$$

where the Φ_s are m x m matrices and where e_t is a realization of an m–dimensional white noise process. Box and Tiao (1977) show that this multivariate proces can be reparametrized as a collection of m uncoupled autoregressive processes on some new series u_t. The transforms works by finding linear combinations $u_j = Xa_j$ for j = 1,...,m that are contemporaneously independent, i.e. $E(u_j'u_{j'}) = 0$ for $j \neq j'$, and that are ordered according to their respective predictive powers. The predictability measure γ_j reflects how much the j–th component can predict itself by a univariate p–th order autoregressive model

$$\begin{aligned} u_{t,j} &= f_1 u_{t-1,j} + f_2 u_{t-2,j} + \ldots + f_p u_{t-p,j} + e_{t,j} \qquad (2) \\ &= \underline{u}_{t,j} + e_{t,j} \end{aligned}$$

where f_s are the autoregressive weights and where $u_{t,j}$ denotes the expectation of $u_{t,j}$ conditional on the past. Let $\underline{\sigma}_j^2 = E(\underline{u}_{tj}^2)$ and let $\sigma_j^2 = E(u_{tj}^2)$ then the predictability for u_j is equal to $\gamma_j = \underline{\sigma}_j^2 / \sigma_j^2$, i.e. the proportion of variance of u_j explained by the historical predictor u_j.

For the first predictable component, the goal is to find the weight vector a_1 such that the linear combination $u_1 = Xa_1$ has maximum predictability γ_1. Next, a second predictable component, orthogonal to the first, can be identified, and so on. To see how this problem can be solved with canonical correlation analysis let (2) be expressed for all j = 1,...,m simultaneously as

$$U = \sum_{s=1}^{p} B_s U F_s + E. \qquad (3)$$

Since U = XA we can express (4) in terms of the observed data as

$$XA = \sum_{s=1}^{p} B_s XAF_s + E = \sum_{s=1}^{p} B_s XA_s + E, \qquad (4)$$

where $A_s = AF_s$. It is now easy to see that the problem of determining maximum predictability is equivalent to finding the largest canonical correlations between a set of the observed series X and a set of lagged series [BX, B_2X, ..., B_sX]. This relationship with CCA has also been pointed out by Parzen and Newton (1980) and Velu, Reinsel and Wichern (1986). The latter authors found that γ_j is equal to the squared canonical correlation.

It is straightforward to express the Box–Tiao transform as a least squares loss function. For a given series X we must minimize

$$\sigma(Z; A, A_1, \ldots, A_s) = SSQ\,(Z - XA) + SSQ\,(Z - \sum_{s=1}^{p} B_s XA_s) \qquad (5)$$

over Z and A, $A_1,\ldots,A_s$ under constraints 1'Z = 0 and Z'Z = I. The predictable components are equal to U = XA. Since U approaches the orthogonal matrix Z the components are nearly orthogonal. It can be shown that minimizing (5) comes down to maximizing the sum of the m canonical correlations between X and $[BX, B_2X, \ldots, B_sX]$.

If the data are categorical then X is not fixed but is it parametrized as X = GY, where Y is a matrix of scaling values. To see how this works, suppose that the categorical series are coded into indicator matrices G_j (cf. Gifi, 1990). Categories may be quantified by postmultiplying G_j by a vector of initially unknown category quantifications y_j, so the product $x_j = G_j y_j$ produces a quantified series x_j. It is possible to accommodate for ordinal variables by restricting the sequence of y_j values to be weakly monotonely increasing. For interval variables the elements of y_j increment with fixed steps. Now X can be written as X = GY, where $G = [G_1,\ldots,G_m]$ and, where Y is the s × m matrix containing $y_1,\ldots,y_m$ as its diagonal blocks and where s is the total number of categories. It is convenient to scale X by 1'X = 0 and dg X'X = I. So if the data are categorical a number of additional scaling parameters in Y have to estimated.

Loss function (5) is a special case of the canonical function class defined by van Buuren (1990, p. 130, p. 135). An algorithm to minimize (5) both with or without optimal scaling can be found in the same reference and we will not repeat the solution here.

EXAMPLE

As an example, we analyze the diary series listed in van Buuren (1990, p. 107). The series records a number of psychological and medical factors and some daily activities of a woman in her mid–twenties for 131 consecutive days. The data are measured on nominal and ordinal scales. The variables are: emotional state (down, normal, good, active–hysteric), physical state (ill, healthy), sexual activity (nothing(1) – much(3)), indisposed (no, yes), smoking (none, some, much), food (italian, dutch, bread, snacks, other) and alcohol (none, some, much). We are interested summarizing the most salient features of the data. Assuming that the influence of an observation will wear out in about five days we include up to five terms in the autoregressive model, i.e. $p = 5$. The predictability for each of the seven components as measured by the squared canonical correlations is respectively 0.74, 0.58, 0.32, 0.26, 0.26, 0.22 and 0.16. For the first two components these values are much larger than the others so these aggregates depend most heavily on the past and will be studied in further detail below.

The first and second predictable components are plotted in Figure 1. The first component exhibits a regular periodic shape corresponding to the menstruation

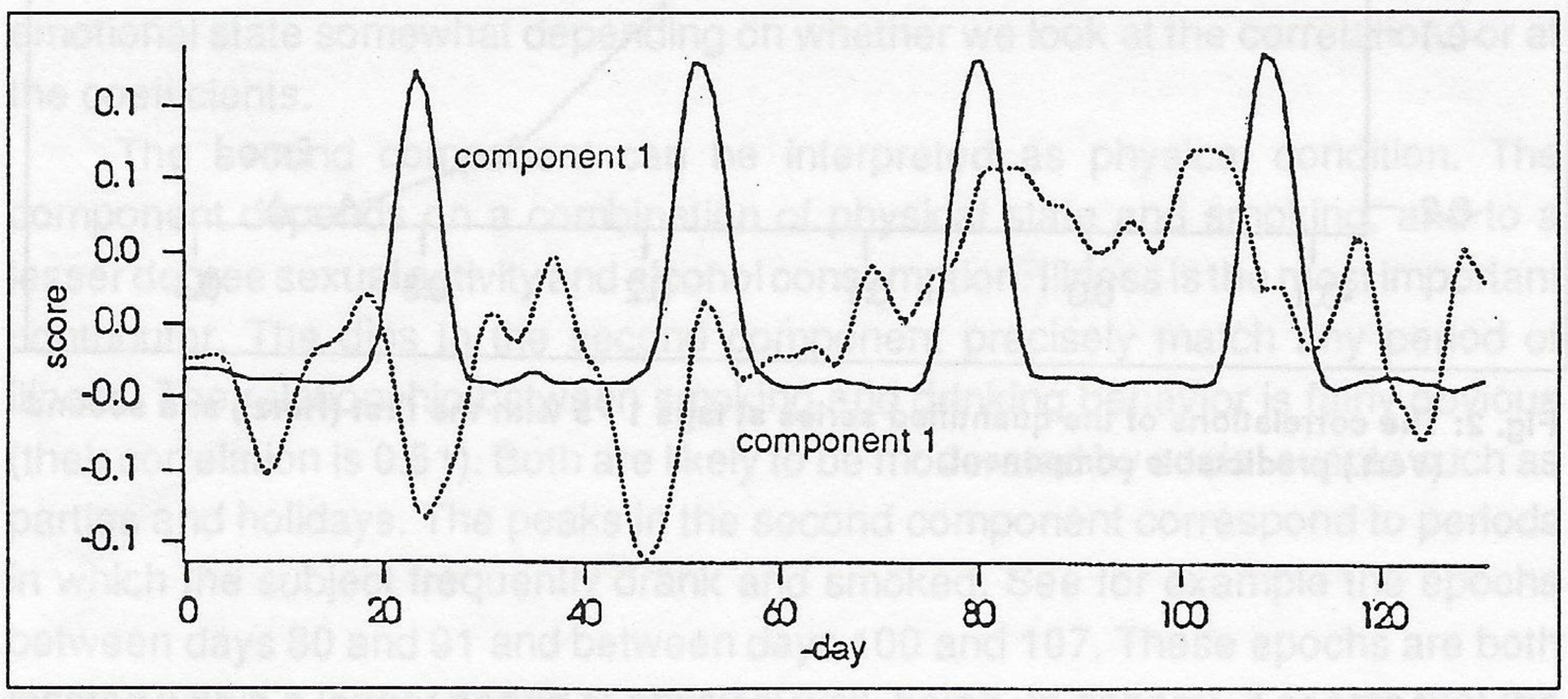

Fig. 1: The first and second predictable components of the diary data plotted against time (n=131).

cycle. The interpretation of the second component is less obvious. It is informative to inspect the correlations between the predictable components and the series from which they were constructed. There is a total of 5 (lags) × 7 (variables) × 2 (components) = 70 correlations to look for. Figure 2 is a graphic representation of a subset of these correlations. The correlations with the first and second component serve as the coordinates on respectively the first and second axis. The points are marked by variable name and lag number. The correlations for 'sexual activity' and 'food' are close to zero and are not plotted.

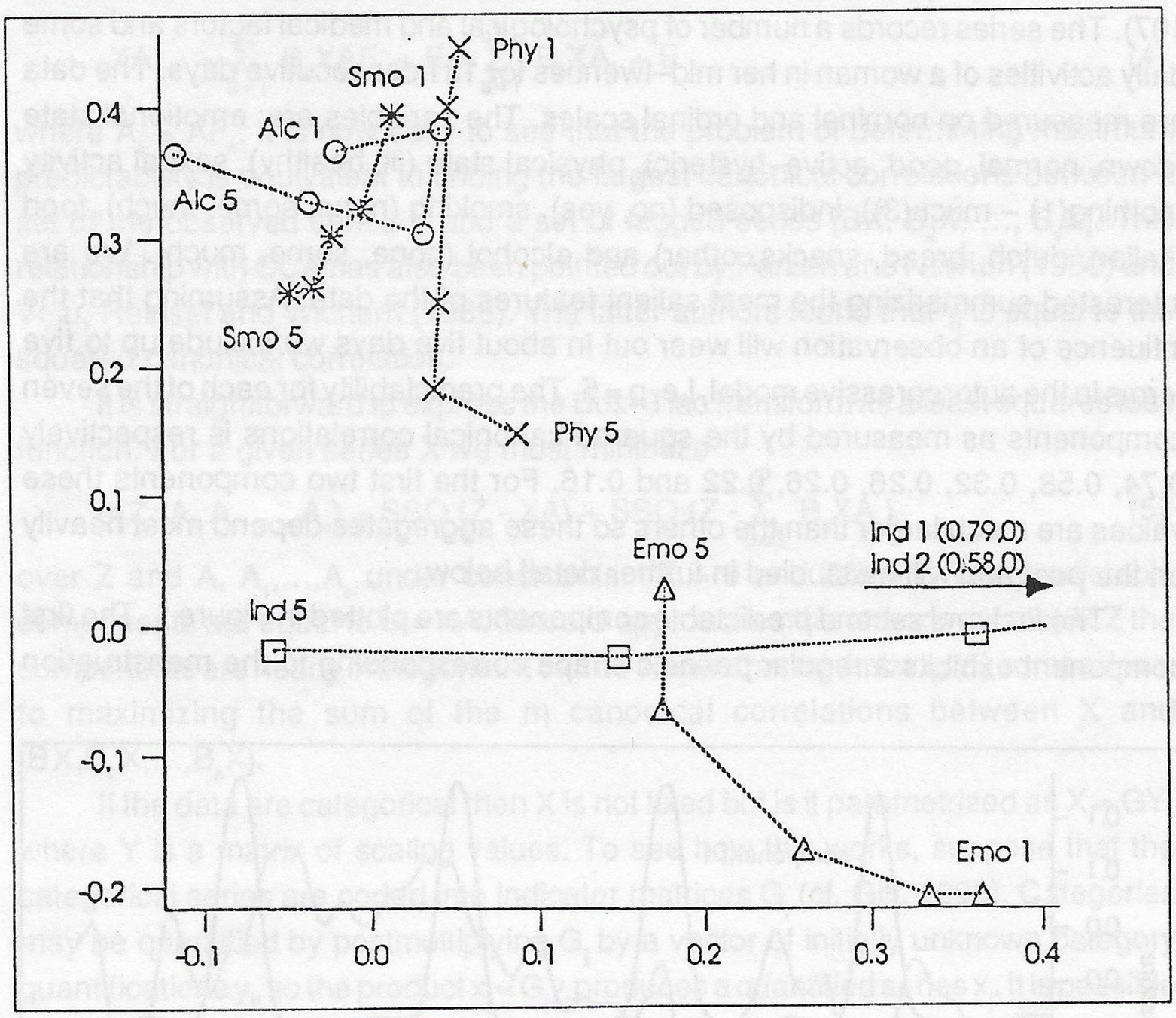

Fig. 2: The correlations of the quantified series at lags 1 - 5 with the first (Horz.) and second (Vert.) predictable component.

Tab. I: Weights YA × 1000 for components 1 and 2. Adding the seven weights for a given data profile and dividing the result by 1000 yields the predictable component scores.

EMOTIONAL STATE			SMOKING		
down	−2	−10	none, missing	−2	−33
normal	−1	−2	some	2	40
good	2	10	much	3	75
active–hysteric	−1	−6	FOOD		
PHYSICAL STATE			Italian	−1	−1
ill	0	−111	Dutch	−4	−5
healthy	0	25	Bread	12	18
SEXUAL ACTIVITY			Snacks	2	3
nothing	0	−1	Other	−5	−8
moderate	0	−1	ALCOHOL		
much	3	37	none, missing	4	−22
INDISPOSED			some	0	0
no	−39	−1	much	−5	23
yes	181	6			

Table 1 contains quantifying weights YA that were used to construct the predictable components. For example, the first component score for the data profile {good, healthy, nothing, no, some, Dutch, some} would be (2 + 0 + 0 - 39 + 2 - 4 + 0) / 1000 = -0.039. The elements of Table 1 thus measures how much a specific category contributes to the components.

As expected, the first component is highly correlated with the menstruation cycle: the correlations between the lagged data and the component is 0.79. Emotional state is also correlated. However, we may infer from Table 1 that its contribution is minimal. Also, other series do not affect to the first component so the menstruation cycle is relatively independent of other factors, possibly except for emotional state somewhat depending on whether we look at the correlations or at the coefficients.

The second component can be interpreted as physical condition. The component depends on a combination of physical state and smoking, and to a lesser degree sexual activity and alcohol consumption. Illness is the most important contributor. The dips in the second component precisely match any period of illness. The relationship between smoking and drinking behavior is fairly obvious (their correlation is 0.51). Both are likely to be moderated by social events such as parties and holidays. The peaks in the second component correspond to periods in which the subject frequently drank and smoked. See for example the epochs between days 80 and 91 and between days 100 and 107. These epochs are both located within a longer period of physical well–being. In general, it seems that the

dips are caused by illness and that the peaks are caused by many cigarettes and much wine.

CONCLUSION

Predictable components can be helpful to uncover interesting relationships in multivariate time series. The first few predictable components tend to single out the slowest varying series and can be seen as low–pass multivariate smoothers. The method is meant to be applied on series that consist of at least 50 time points, preferably more. The solution will become unstable for low n.

In the conventional Box–Tiao transform, where all variables are numerical, the solution will be nested. This means that components themselve are invariant under the number of derived components. The optimal scaling version is not nested: with a different number of components, different solutions may emerge. Fortunately, in practice differences are usually so small that they can safely be ignored.

REFERENCES

Box, G.E.P. and Jenkins, G.M. : *Time series analysis, forecasting and control (revised edition)*. Holden–Day, San Francisco, 1976.

Box, G.E.P. and Tiao, G.C.: *A canonical analysis of multiple time series*. Biometrika, *64*, 355–365; 1977.

Gifi, A.: Nonlinear multivariate analysis. Wiley, New York, 1990.

Parzen, E. and Newton, H.J.: *Multiple time series modelling II*. In P.R. Krishnaiah (Ed.), Multivariate analysis V, 181–197. North–Holland, Amsterdam, 1980.

Van Buuren, S.: *Optimal scaling of time series*. Dissertation, University of Utrecht. DSWO Press, Leiden, 1990.

Velu, R.P., Reinsel, G.C. and Wichern, D.W.: *Reduced rank regression models for multiple time series*. Biometrika, *73*, 105–118; 1986.