QÜESTIIÓ, Vol 17, 3 pp. 327-338, 1993

# LONGITUDINAL $K$–SETS ANALYSIS USING LAGGED VARIABLES

CATRIEN C.J.H. BIJLEVELD and EEKE VAN DER BURG

*We present an application of nonlinear Generalised Canonical Analysis (GCA) for analysing longitudinal data. The application uses lagged versions of variables to accommodate the time-dependence in the measurements. The usefulness of the proposed method is illustrated in an example from developmental psychology, in which we explore the relationship between mother and child dyadic interaction during the first six months after birth, demonstrating how child behaviour can elicit mother behaviour. We discuss the relationship between our proposed method and the most closely resembling SERIALS (Van Buuren, 1990) method for nonlinear time series analysis.*

**Key words:** Generalised canonical analysis, optimal scaling, lagged variables, developmental research.

## INTRODUCTION

Generalised Canonical Analysis or $K$–sets canonical analysis relates several sets of variables, searching for what is common between the sets (Carroll, 1968). Van der Burg, de Leeuw and Verdegaal (1988) presented an extension of this technique, in which categorical variables can be rescaled by optimal scaling; this technique was implemented in the computer program OVERALS (SPSS, 1990).

Catrien C.J.H. Bijleveld and Eeke van der Burg. Department of Psychometrics & Research Methodology. Faculty of Social and Behavioural Sciences. Leiden University. P.O.Box 9555, 2300 RB Leiden, the Netherlands.

–Article presentat al Seventh International Conference on Multivariate Analysis, setembre 1992.
–Acceptat el maig de 1993.

(Non)linear GCA or OVERALS has not been applied often in longitudinal research, although it offers several opportunities for doing so. The most obvious possibility is to treat the variables at each new time point as a new set of variables; a disadvantage of this setup is that the technique then seeks for what is common between the sets (in this case, the time points), while one is usually interested rather in what changes from time point to time point. A second method for analysing longitudinal data by nonlinear GCA, using a dummy time variable, was demonstrated in Van der Burg and Bijleveld (1993). In the present paper we will illustrate a different method, working with time-lagged versions of variables. This also enables us to model the time-dependence and to explore the causal mechanisms in the phenomena under investigation. We will not elaborate on (nonlinear) generalised canonical analysis, but refer to Van der Burg, de Leeuw and Verdegaal (1988), Gifi (1990), as well as to Van der Burg and Bijleveld (1993) for details.

## FLATTENING THE DATA BOX AND MODELLING TIME WITH LAGGED VARIABLES

The use of lagged variables is probably best introduced through an example. Suppose we have obtained daily recordings of the severity of a subject's headaches. We name the headache variable $y$, and the measurements at the respective days are indicated as $y_t$, with $t = 1, \ldots, T$. Suppose furthermore that we are interested in the effect of alcohol consumption on headache complaints, and that we have also recorded the subject's daily alcohol intake, and that we name the daily alcohol measurements $x_t$, with $t = 1, \ldots, T$. The measurements $x_t$ and $y_t$ are stacked vertically in vectors, of size $T$, that thus contain $x_1$ to $x_T$ and $y_1$ to $y_T$ respectively. A research hypothesis could be 'liberal intake of alcohol on day $t$ leads to headache on day $t + 1$'.

We can investigate this hypothesis by analysing lagged versions of the variables. If the hypothesis were true, this would mean that we would find a high positive correlation between the scores on the alcohol variable at days $t$, and the scores on the headache variable on days $t + 1$. What we now propose to do, is to alter the two variables, in the sense that we eliminate the last measurement of the alcohol variable, and we eliminate the first measurement of the alcohol variable; next, we 'shift' the two vectors (that have now both size $(T-1)$) so that they match again. Thus, using these variables, we correlate $x_t(t = 1, \ldots, T-1)$ to $y_t(t = 2, \ldots, T)$. In the literature, this lagged version of the $x-$variable is often referred to as 'the lag1 version of variable $x$'; the lagged version of y is

often referred to as 'the lag0 version of $y$'. An example of such a situation for six time points is presented in Figure 1.

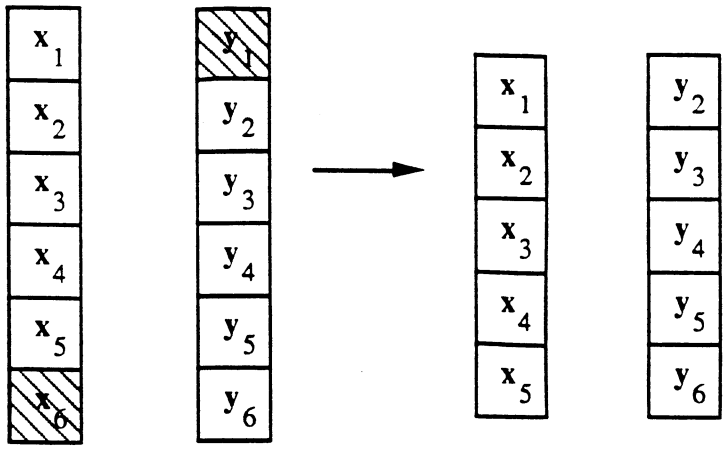


**Figure 1.**
Example of a Lag1 Variable Relating to a Lag0 Variable.

A high correlation between the lag1 version of $x$, and the lag0 version of $y$ is usually interpreted as an indication that lag1$-x$ causes lag0$-y$, that is, alcohol intake at day $t$ causes headaches on day $t+1$. Theoretically, the high correlation could also mean that headaches at day $t+1$ cause alcohol intake on day $t$, but as causes are supposed to occur before or at most simultaneously with effects, this alternative explanation need not be considered. The time order does not exclude spurious causal effects via a third variable.

Other phenomena may be modelled using lagged variables. For instance, one might be interested to know to what extent certain behaviour types, or physiological indicators depend on, or can be predicted from their past values. In that case, one would want to model the influence of a past version of a variable, on its present version. Suppose for instance that such a variable is blood pressure, and that measurements have again been obtained at $T$ time points, that have again been combined in a vector $x$ of size $T$. In order to assess to what extent blood pressure measurements can be predicted from their prior values, we would then have to 'copy' the $x$ variable, so that we have it two

times, eliminate from one the first measurement, and eliminate from the other the last measurement, and shift them, so that they match again. As an example, the squared correlation between lag1$-x$ and lag0$-x$ equals the percentage of the variance of blood pressure that we can predict from its prior versions. An illustration of such a situation with six time points is in Figure 2.

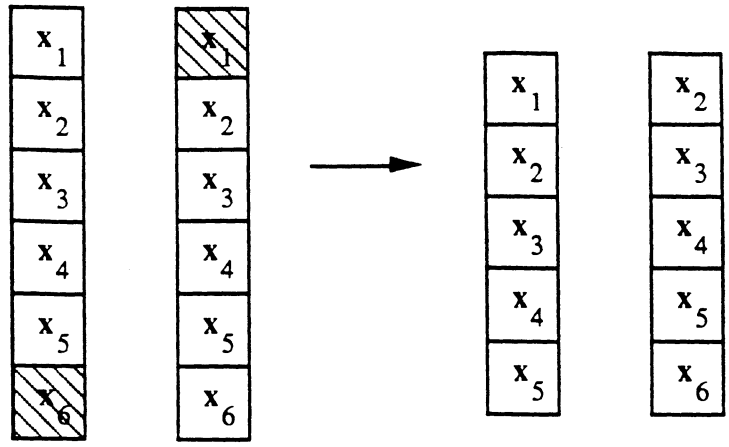


**Figure 2.**
Example of a Lag1 Variable Relating to its Lag0 Version.

Other variations on the theme are possible. For instance, higher-order lags may be specified, relating for instance in a multiple regression type of model, lag2 or lag3 versions of variables to lag0 versions of other variables (or of the same variables). One then approaches the ARMA-type models (that we will not discuss here) that were proposed by Box and Jenkins (1976). Conceptually attractive are lagged versions of variables that can model seasonal effects, such as weekly cycles, monthly (28 days for instance in menstruation research) or yearly cycles. An example is a situation with daily measurements, where we predict one set of lag0-variables from a (different) set of lag1-variables, that thus captures the immediate past, as well as from an additional set of lag7-variables, that thus models the dependence on the measurements at the same day last week. In that case, the lag0 set would thus contain the measurements from $t = 8$ until $t = T$, the lag1 set those from $t = 7$ to $t = T - 1$, and the lag7 set from $t = 1$ to $t = T - 7$. See Figure 3.

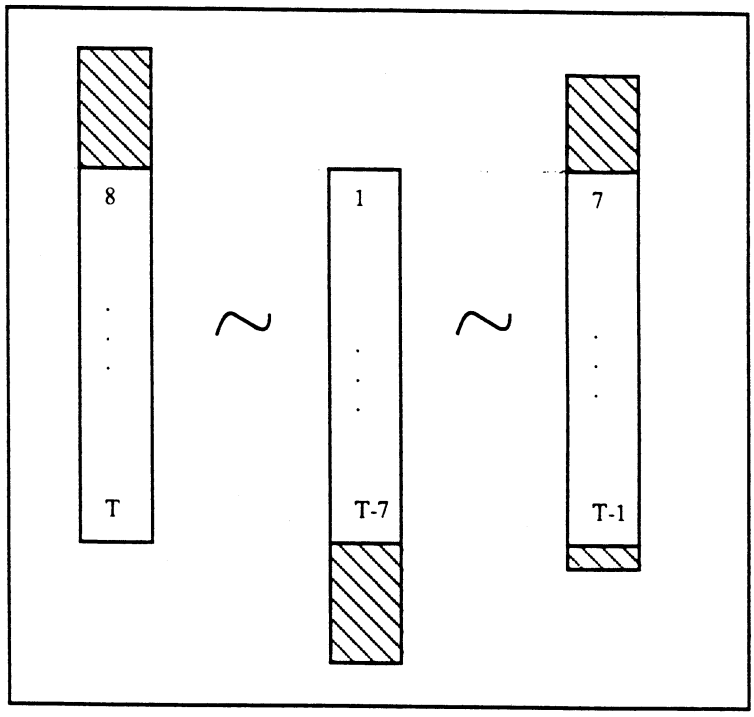


**Figure 3.**
Schematic Representation of Analysis Relating Lag0, Lag7 and Lag1 Versions of Variables.

In principle, lagged versions of variables can be used in any type of analysis. Lag1 relationships approximate the so-called Markov-type dependencies.

Solutions from an analysis that uses lagged variables should be interpreted just as solutions with only lag0-versions of variables, that is, by using for instance component loadings or category quantifications. Normally, if two variables have similar (or opposite) high component loadings, we say that these variables have a lot in common. In the case with lagged variables, if for instance the component loadings of a lag0 and lag1 variable are similar and high, one might conclude that the lag1 variable has an impact on the lag0 variable. Below, we will illustrate the use of lagged versions of variables in generalised canonical analysis. The implications can easily be generalised to other types of exploratory categorical data analysis.

Technically speaking, one 'loses' more time points the higher the order of the lags. Which means that higher-order lags can become unattractive, or impossible to model when there are not too many time points. Analyses with lagged versions are usually harder to interpret than ordinary types of analysis.

## A LONGITUDINAL EXAMPLE: DEVELOPMENT OF ATTACHMENT IN YOUNG CHILDREN

For the example, we will analyse data collected by Van den Boom (1988). Mothers and children had been investigated at 6 time points (in six consecutive months after birth); the frequencies of occurrence of 8 types of mother behaviour formed 8 variables in the mother set, the frequencies of occurrence of 6 types of child behaviour constituted 6 variables in the child set. The behaviour variables are in Table 1. Van der Burg and Bijleveld (1993) analysed the same data in a different fashion.

**Table 1**

Mother and Child Behaviour Variables

| child variables | mother variables |
|---|---|
| positive sociable behaviour | observing baby |
| observing persons and objects | effective stimulation |
| vocalising | vocalising/offering objects |
| whining/crying | physical contact |
| exploration | comforting |
| sucking | uninvolved |
| | responsiveness to crying |
| | responsiveness to positive behaviour |

After one year, the children had been classified into three attachment groups, namely *secure, avoidant* and *resistant.* One child had gone to hospital during the last two months of the study, which, with lots of crying, caused rather deviant scores, so for this child we deleted the last two months from the data set. To investigate the influence of prior mother behaviour on the present behaviour of their children, and vice versa, we constructed data sets with lagged variables. In the first set, we stacked all mother measurements from time point 1 until time point 5, this is the so-called mother lag1 set. In the second set we stacked all children measurements from time point 2 until time point 6, this is the so-called child lag0 set. (If these latter two sets are analysed together, we model the influence of mother's behaviour in the prior month on children's behaviour in

the present month.) Next, we constructed a third and a fourth set with lagged variables: in the third set, we stacked all child measurements from time point 1 until time point 5, the child lag1 set; in the fourth set we stacked all mother measurements from time point 2 until time point 6, the mother lag0 set. (If these two sets are analysed together, we model the influence of children's behaviour in the prior month on mother's behaviour in the present month.)

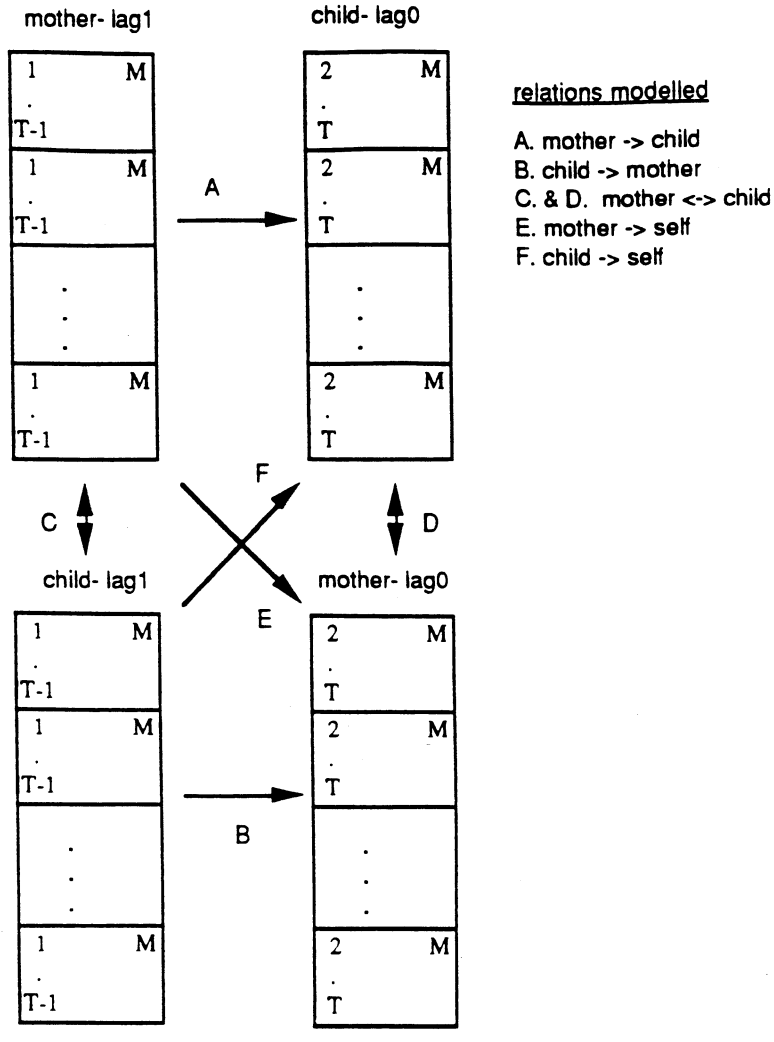


**Figure 4.**
Schematic Representation of Lagged Analysis.

Instead of doing separate analyses of the influence of children upon mothers and vice versa, we analysed the four sets together in one analysis, which had the added advantage that a number of cross- and auto-regressive influences can be investigated simultaneously. The influence of mothers' past behaviour on their present behaviour, is given by the combination of the mother lag0 and mother lag1 set; the influence of children's past behaviour on their present behaviour, is given by the combination of the child lag0 and child lag1 set. Instantaneous relations between mothers and children are modelled by the combination of mother and child sets of same lag. For a schematic representation see Figure 4.

As we have more than two sets, we performed multi-set analysis for this set up, using the computer program OVERALS. All variables were treated as ordinal variables.

## ANALYSIS RESULTS

The eigenvalues represent a fit measure of an OVERALS analysis; they always vary between zero and one. The eigenvalues resulting from the mother and child data analysis were .801 for the first dimension, and .730 for the second dimension, which is quite nice. The component loadings, that correspond to the correlations of the rescaled variables with the dimensions, are in Figure 5. Mother variables are typed plain face, child variables italic face; the points of the component loadings are not connected to the origin, but instead the component loadings of lag1-versions are connected to those of the lag0-versions of variables, an arrow indicating the lag0-version. The first and second dimensions have been switched in this picture, to make it more comparable with the Figure 2 of Van der Burg and Bijleveld (1993), although it is still tilted slightly counter-clockwise with respect to this figure.

The most striking feature of the picture is that present and prior variables are almost always located closest to one another. This is most apparent in the periphery, for instance prior and present exploration and crying by the baby are situated very closely, past and present watching by the mother are very close, as are past and present versions of uninvolvedness and physical contact. This implies, that present behaviour of mothers as well as present behaviour of children is always most strongly related to their own past behaviour. Thus, in preference over past or present behaviour of the other, the best predictor of the behaviour of mothers and children, are their respective past behaviours. One could translate this into saying that inter-individual differences are larger than intra-individual differences, with mothers as well as children fluctuating at more

or less individually particular levels of behavioural activity, that do not change dramatically in the course of the first six months after birth.

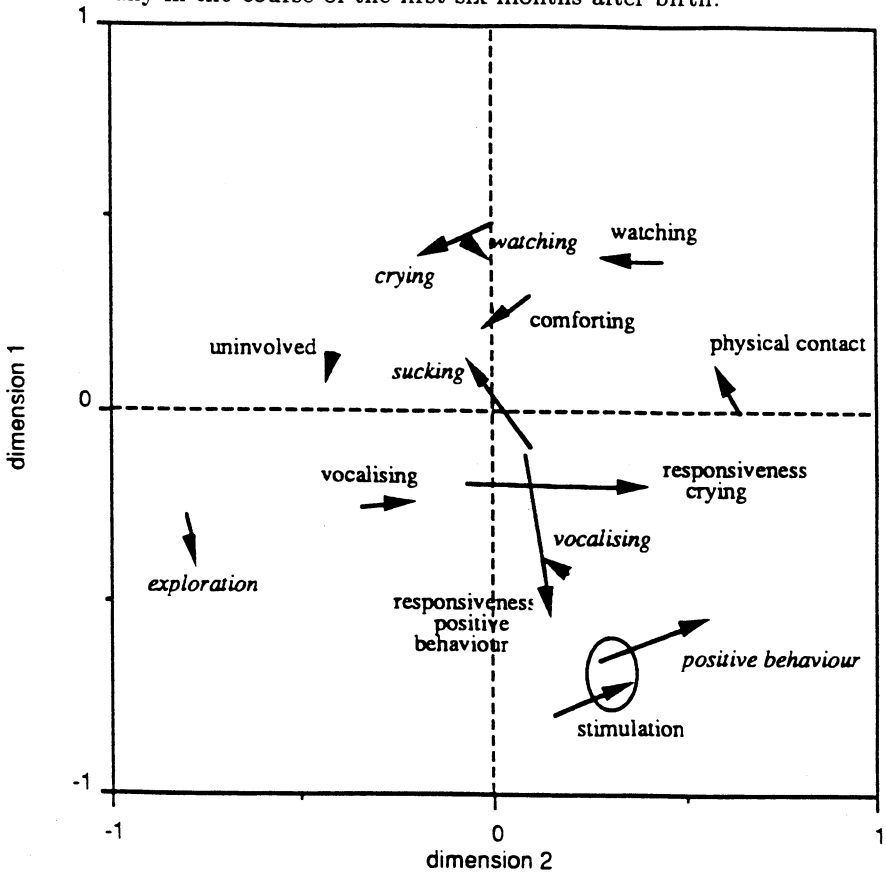


**Figure 5.**
Component Loadings of Mother (plain face) and Child Variables (italic face) in Lagged Analysis.

An exception to this pattern is the past positive behaviour of the child with present effective stimulation by the mother (circled in the plot). These are located very close to one another, indicating that positive behaviour of children in one month in general goes together with effective stimulation by the mother in the next month. Maybe, once children start exhibiting this type of positive behaviour, this triggers the mother's stimulative behaviour; this result might be explained by supposing that mothers only start stimulating their children in this way, once the children give the signal that they are ready for it.

Another striking point of Figure 5 is that the lag0 and lag1 responsiveness variables differ more from one another than do the other lag0 and lag1 combinations. Mother's lag0 responsiveness to positive behaviour is in fact closer to children's lag1 positive behaviour than it is to its own lag1 version. Probably mothers can only become responsive to their children's positive behaviour, once the children have expressed such behaviour. For lag0 responsiveness to crying, and lag0 crying, as well as for the two lag1 versions of these variables, the situation is slightly more complicated, as they are in opposite directions: responsiveness to crying is in an opposite direction to crying. This indicates that the less the child cries, the more responsive the mother is, when crying does occur; these relations should not be taken as absolute however, as the component loadings are not very high. The lag0 and lag1 versions of responsiveness to crying, are quite wide apart. Lag1 responsiveness to crying is closest to the children's vocalising at lag0, indicating that an increased responsiveness to crying might stimulate children to vocalise.

Figure 5 gave the scores for the rescaled variables. It is also possible to depict the subjects (the mother and child pairs) in this plot. Then we could track the average developments of the three attachment groups in the plot, by averaging the scores of the mother-child pairs per time point per attachment group. This would have produced similar results to those of Van der Burg and Bijleveld (1993), with the resistantly and avoidantly attached travelling in the upper part of the plot (from right to left) characterised by a lot of crying, and the securely attached travelling in the bottom of the plot (also from right to left), starting out with a lot more responsiveness to crying, positive behaviour and stimulation. All three attachment groups develop more or less from dependent to independent behaviour, that is, from lots of physical contact towards exploration.

## DISCUSSION

Lagged variables are a nice tool to use in longitudinal data analysis. Especially when working with non-numerical longitudinal data, for which many analysis techniques are unsuited, they may provide a help, at least in exploring the structure of the data or changes over time. In our example, the use of lagged variables gave some valuable insights into the relationships in time between the mother and child behaviour variables. We could demonstrate how certain types of child behaviour elicited mother's behaviour, pointing away from the customarily supposed central force of the mother (maybe a comfort for some mother readers?). Furthermore, we could depict in an attractive way mother and child behaviour development.

We could have done a similar analysis using Van Buuren's SERIALS method (van Buuren, 1990). Van Buuren's SERIALS model combines elements of the Box-Tiao transform with state space analysis, providing an extension for scaling of any non-numerical variables. In this technique, present variables are combined into one set, and past versions of the data set in the other. The two sets are then analysed using an OVERALS type model, that tries to find directions that explain most of what is common between the two sets. In principle, van Buuren's method offers the possibility to model a time dependence of the object scores on themselves as well. The main difference between our method and the SERIALS method, is that in our method the optimal quantifications of the categories of lagged and unlagged versions of variables may differ. This could constitute a problem, as lagged and unlagged variables then become qualitatively different variables, that cannot be related as if they were measuring the same phenomenon. In our example, the category quantifications of the lagged and unlagged versions of variables differed only minimally, but especially when working with nominal variables, differences may turn out substantial. In such cases, one had perhaps resort to the SERIALS method, that has the major advantage that it constructs identical quantifications of the categories of lagged and unlagged versions of variables. An important advantage of the OVERALS method, on the other hand, is that it provides the opportunity to differentiate between variables for mother and child.

## REFERENCES


[1] **Box, G.E.P. & Jenkins, G.M.** (1976). *Time series analysis, forecasting and control.* San Francisco: Holden-Day.

[2] **Carroll, J.D.** (1968). "Generalization of canonical correlation analysis to three or more sets of variables". *Proceedings of the 76th Annual Convention of the American Psychological Association*, 227–228.

[3] **Gifi, A.** (1990). *Nonlinear multivariate analysis.* Chichester: Wiley.

[4] **SPSS** (1990). *SPSS Categories.* Chicago: SPSS Inc.

[5] **Van Buuren, S.** (1990). *Optimal scaling of time series.* Leiden: DSWO-Press.

[6] **Van den Boom, D.C.** (1988). *Neonatal irritability and the development of attachment: Observation and intervention.* Unpublished PhD thesis. Leiden: Leiden University.

[7] **Van der Burg, E. & Bijleveld, C.C.J.H.** (1993). "Longitudinal $K-$sets analysis using a dummy time variable". *Qüestiió*, **17.3**, 333–345.

[8] **Van der Burg, de Leeuw & Verdegaal, R.** (1988). "Homogeneity analysis with $K$ sets of variables. An alternating least squares method with optimal scaling features". *Psychometrika*, **53**, 177–197.