Transitional ideal point models for longitudinal multinomial outcomes

Mark de Rooij

Statistical Modelling 2011 11: 115
DOI: 10.1177/1471082X1001100202

Statistical Modelling 2011; **11**(**2**): 115–135

Transitional ideal point models for longitudinal multinomial outcomes

Mark de Rooij
Methodology and Statistics Unit, Psychological Institute, Leiden University, The Netherlands

Abstract: For the analysis of longitudinal data, three families of models are generally distinguished: the marginal, the transitional and the subject-specific family. In this paper, we will propose a transitional model for the analysis of change for a nominal response variable. Such an analysis is often hampered by the dimensionality of the problem. We use multidimensional scaling techniques, more specifically the ideal point model, in order to reduce the dimensionality. The model can handle pure transitional data but also allows for explanatory variables. Two empirical examples will be discussed in order to illustrate all the virtues of the model.

Key words: biplots; categorical data; Markov models; multidimensional scaling

Received May 2008; **first revision:** December 2008; **second revision:** May 2009; **accepted** May 2009

1 Introduction

Longitudinal data arise in many fields of research. When the outcome variable is normally distributed, sufficient tools exist for the analysis of such data. For categorical variables, the last decade showed a boost of studies; surveys can be found in Diggle, *et al.* (2002) and Molenberghs and Verbeke (2005). Going through these two books, it can be noticed that for binary, count and ordered multinomial data, there have been quite some developments, mainly using generalizations of the generalized linear model. In contrast, for multinomial unordered categories, i.e., nominal variables, such generalizations of the multivariate generalized linear model are limited. Some exceptions can be found in Hedeker (1999), Hartzel *et al.* (2001) and Lipsitz *et al.* (1994). It can be argued that for nominal outcome variables, development is hampered by the dimensionality of the problem. With a discrete outcome variable having J classes, the dimensionality is $J-1$, i.e., for each explanatory variable $J-1$, regression parameters have to be estimated and interpreted in a multinomial logit model. Often the J classes do not differ on $J-1$ attributes, but the number of attributes

Address for correspondence: Mark de Rooij, Methodology and Statistics Group, Leiden University Institute for Psychological Research, PO Box 9555, 2300 RB Leiden, The Netherlands. E-mail: rooijm@fsw.leidenuniv.nl

 10.1177/1471082X1001100202

that substantially differentiate between the classes maybe far less than $J-1$. Take as an example the Dutch parliamentary election studies (to be discussed in more detail later). In the Netherlands, we have a multiparty electoral system, where citizens have to choose between a dozen parties. In 2003, e.g., 16 different parties took part in the elections. Even if only the largest eight parties are considered, building a multinomial logit model for this would require seven regression equations: Dimension reduction is required in order to understand what is going on.

Dimension reduction can be implemented using multidimensional scaling techniques, a class of models that represent data in low-dimensional Euclidean space. In the case outlined before, it is most natural to have a point in Euclidean space for each subject at each time point and a class point. The Euclidean distance between these two points determines the probability that a certain class is chosen. Explanatory variables can be included in such a graphical display, to obtain a biplot representation (Gower and Hand, 1996). The next two sections will develop this idea in more detail.

Following Diggle *et al.* (2002) and Molenberghs and Verbeke (2005), we can distinguish between three families of models for longitudinal categorical data: marginal models, transitional models and subject-specific models. In the first type, marginal models, responses are modelled marginalized over all other responses; the association structure is typically captured by a set of association parameters. In transitional models, any response in the sequence is modelled conditional upon (a subset of) past responses. In subject-specific models, the responses are assumed independent given a set of subject-specific parameters. The three types of model typically answer different questions. The marginal approach handles the question how, on average, the system of probabilities evolves over time in the population, whereas the subject-specific approach handles the question how this system of probabilities evolves over time for each individual subject. The transitional approach takes into account the response at previous time points which is qualitatively different from the two other perspectives. Whereas in the case of a normally distributed outcome variable, these types of model are naturally connected, for categorical outcomes there is no close connection (Diggle *et al.*, 2002; Molenberghs and Verbeke, 2005).

In this paper, we will develop a transitional model. De Rooij (2008)—generalizing and unifying earlier work by De Rooij and Heiser (2005) and De Rooij (2001, 2002)—developed multidimensional scaling models for transition frequency tables. These are change data without explanatory variables. In that case, it is assumed that the data come from a homogeneous group, i.e., it is assumed that all subjects follow the same change pattern. Often explanatory variables are available, and the main interest is in differential modelling of change. We will propose a modelling framework that incorporates such explanatory variables. These variables might be continuous or categorical.

In Section 2, we will outline our basic model for a multinomial outcome variable, and we develop the transitional model. In Section 3, we discuss other statistical models that reduce the dimensionality and compare them to our model. Section 4 gives two examples. We conclude this paper in Section 5 with some discussion of the modelling framework.

2 Transitional modelling of longitudinal multinomial data

2.1 An ideal point model for multinomial data

The response variable G has J classes indexed by $j = 1, \ldots, J$. For each of n subjects, we have a p-dimensional vector of explanatory variables $\mathbf{x}_i$, $i = 1, \ldots, n$. The probability that subject i chooses class j will be denoted by $P(G = j|\mathbf{x}_i) = \pi_j(\mathbf{x}_i)$, with $\sum_j \pi_j(\mathbf{x}_i) = 1$.

This probability will be modelled using a distance between two points in a Euclidean space of dimensionality M: one point, sometimes called an ideal point ($\mathbf{y}_i = [y_{i1}, \ldots, y_{iM}]^T$), represents subject i and the other ($\mathbf{z}_j = [z_{j1}, \ldots, z_{jM}]^T$) is a point for category j. The smaller the distance between the two points, the larger the probability that the subject chooses that category. The ideal point classification model (IPCM) is then

$$\pi_j(\mathbf{x}_i) = \frac{\exp\left(-d^2(\mathbf{y}_i, \mathbf{z}_j)\right)}{\sum_l \exp\left(-d^2(\mathbf{y}_i, \mathbf{z}_l)\right)}, \tag{2.1}$$

where $d^2(\cdot, \cdot)$ is the squared Euclidean distance between the two entries. The subject points are taken to be linear combinations of the predictor variables $\mathbf{x}_i$,

$$\mathbf{y}_i = \mathbf{B}^T\mathbf{x}_i, \tag{2.2}$$

where $\mathbf{B}$ is a $p \times M$ matrix with regression weights. The parameters of this model are the regression weights and the class points. It can be shown that in dimensionality $J - 1$, this model equals the multinomial logit model (De Rooij, 2009a), so that IPCM provides the possibility for dimension reduction when $M < (J - 1)$ with a simply interpreted graphical display.

Parameter estimates can be found by maximizing the log-likelihood function

$$\sum_{i=1}^{n} \log \prod_{j=1}^{J} \pi_j(\mathbf{x}_i)^{f_{ij}}, \tag{2.3}$$

where f_{ij} equals one if subject i is in category j, zero otherwise. The model has rotational freedom and a more intricate indeterminacy, i.e., the probabilities remain the same when a constant is added for each subject (De Rooij, 2009a,b), i.e.,

$$\pi_{j|i} = \frac{\exp\left(-d_{ij}^2\right)}{\sum_l \exp\left(-d_{il}^2\right)} = \frac{\exp\left(-d_{ij}^2 + c_i\right)}{\sum_l \exp\left(-d_{il}^2 + c_i\right)} \tag{2.4}$$

Since the probabilities in our model are solely based on squared distances, we have that a model based on the $n \times J$ squared distance matrix $\mathbf{D}^*$ defined by $\mathbf{D}^* = \mathbf{D} + \mathbf{c}\mathbf{1}^T$ gives the same probabilities as a model based on squared distances $\mathbf{D}$. If we define $\mathbf{T}$

to be a $M \times M$ matrix and $\mathbf{v}$ a vector of length M, it was shown by De Rooij (2009a) that $\mathbf{B}$ and $\mathbf{Z}$ can be transformed to

$$\mathbf{Z}_* = \mathbf{1}\mathbf{v}^T + \mathbf{Z}\mathbf{T} \text{ and}$$
$$\mathbf{B}_* = \mathbf{B}(\mathbf{T}^{-1})^T,$$

under the restriction that $\text{diag}(\mathbf{Z}_*(\mathbf{Z}_*)^T) = \text{diag}(\mathbf{Z}\mathbf{Z}^T) + q\mathbf{1}$, without changing the probabilities. Furthermore, a rotation is always possible. The number of indeterminacies is thus $\max(M(M-1)/2, M(M+1) - (J-1))$, where the first entry corresponds to the rotational freedom and the second to the freedom given by the above-said transformation equations. The number of independent parameters is

$$\text{npar} = (p + J)M - \max(M(M-1)/2, M(M+1) - (J-1)). \tag{2.5}$$

In order to obtain an identified solution, we observe that row-wise centering makes solutions equal. Moreover, if we define $\Pi = \{\pi_j(\mathbf{x}_i)\}$ and $\Delta = \log \Pi$, we also have $-\Delta\mathbf{J} = \mathbf{D}\mathbf{J}$, with $\mathbf{J} = \mathbf{I}_J - \mathbf{1}_J\mathbf{1}_J^T/J$. This makes it possible to use the metric unfolding with single centering (Heiser, 1981; De Rooij, 2009a) for identification. This procedure works fine, except in the situation of maximum dimensionality, i.e., $M = J - 1$. In this case, we identify the solution by a transformation of $\mathbf{Y}$ such that $\mathbf{Y}^T\mathbf{Y} = n\mathbf{I}$ (which can be obtained using a singular-value decomposition), and solve for $\mathbf{v}$.

2.2 Repeated measurements and transitional modelling

Generalizations occur when each subject is measured several times. Let the outcome vector for subject i be $\mathbf{G}_i = (G_{i1}, G_{i2}, \ldots, G_{iT_i})^T$, with for each setting a vector of predictor values $\mathbf{x}_{it}$ $(t = 1, \ldots, T_i)$ which are gathered in a matrix $\mathbf{X}_i$. In order to build transitional models, the joint distribution of the responses given the explanatory variables for subject i can be factored using

$$f(G_{i1}, G_{i2}, \ldots, G_{iT_i}|\mathbf{X}_i) = f(G_{i1}|\mathbf{X}_i) \prod_{t=2}^{T_i} f(G_{it}|G_{i1}, \ldots, G_{i(t-1)}, \mathbf{X}_i). \tag{2.6}$$

Transitional models make use of this factorization. Bonney (1987) showed that standard software can be used in case of a binary response variable, by appropriately defining the matrix with explanatory variables. The general set-up is shown in Table 1 (upper part).

For longitudinal data, often we would like to impose a simpler structure, e.g., a structure in which only the previous choice has an influence on the new choice. The set-up of such a Markov model is shown in the second part of Table 1; in the case shown, it is a first-order Markov model. Higher order Markov models can be set up in a similar way. A further simplification is that the influence one step forward is constant, i.e., the Markov chain is said to be stationary, such a set-up for a second-order stationary Markov model is shown in the third part of Table 1. The data set-ups

Table 1 Data set-up for subject i

Response	Explanatory				
		General set-up			
G_1	$\mathbf{x}_1$	–	–	–	–
G_2	$\mathbf{x}_2$	G_1	–	–	–
G_3	$\mathbf{x}_3$	G_1	G_2	–	–
⋮	⋮	⋮	⋮	⋮	⋮
G_{T_i}	$\mathbf{x}_{T_i}$	G_1	G_2	⋯	$G_{(T_i-1)}$
		First-order Markov Model			
G_1	$\mathbf{x}_1$	–	–	–	–
G_2	$\mathbf{x}_2$	G_1	–	–	–
G_3	$\mathbf{x}_3$	–	G_2	–	–
⋮	⋮	⋮	⋮	⋮	⋮
G_{T_i}	$\mathbf{x}_{T_i}$	–	–	–	$G_{(T_i-1)}$
		Second-order Stationary Markov Model			
G_1	$\mathbf{x}_1$	–	–		
G_2	$\mathbf{x}_2$	G_1	–		
G_3	$\mathbf{x}_3$	G_2	G_1		
⋮	⋮	⋮	⋮		
G_{T_i}	$\mathbf{x}_{T_i}$	G_{T_i-1}	G_{T_i-2}		

for Markov models involve the assumption that earlier responses do not influence the current response and/or the assumption that transitions are homogeneous.

In IPCM (2.1), the ideal points $\mathbf{y}_i$ are linear combinations of the explanatory variables. For a transitional model, the ideal points should be linear combinations of the explanatory variables (if available) and the previous responses. Therefore, first define the $(J-1)$-dimensional *effect coded* dummy variable $\mathbf{h}_{i(t-r)}$ $(r = 1, \ldots, t-1)$ with the response r time points ago. Now gather these vectors in another vector $\mathbf{h}_{it}$, i.e., $\mathbf{h}_{it}$ represents the history for subject i before time point t. For the different cases shown in Table 1, this vector is set up in different, but straightforward, ways.

The probability that subject i chooses class j at time point $t = 1, \ldots, T_i$ is denoted by $\pi_{jt}(\mathbf{x}_{it}, \mathbf{h}_{it})$. The general IPCM has then the following form:

$$\pi_{jt}(\mathbf{x}_{it}, \mathbf{h}_{it}) = \frac{\exp\left(-d^2(\mathbf{y}_{it}, \mathbf{z}_j)\right)}{\sum_l \exp\left(-d^2(\mathbf{y}_{it}, \mathbf{z}_l)\right)}. \tag{2.7}$$

The ideal points are now defined by

$$\mathbf{y}_{it} = \mathbf{B}^T\mathbf{x}_{it} + \mathbf{A}^T\mathbf{h}_{it}, \tag{2.8}$$

where **B** are the usual regression weights and **A** are the regression weights for the past responses. Often the interest is in the interaction between explanatory variables and previous choices. These can be incorporated easily by variable multiplication in the linear predictor of the ideal points, i.e., the ideal points are defined by

$$\mathbf{y}_{it} = \mathbf{B}^T\mathbf{x}_{it} + \mathbf{A}^T\mathbf{h}_{it} + \mathbf{C}^T(\mathbf{x}_{it} \odot \mathbf{h}_{it}), \tag{2.9}$$

where $\odot$ represents variable-wise multiplications of $\mathbf{x}_{it}$ and $\mathbf{h}_{it}$. For longitudinal data, as we consider here, the Markov models are of special interest. For other types of multivariate outcomes, the more general set-up could be used. The contribution of the first time point will be ignored, as is often done with this kind of modelling (see (Agresti, 2002)). The model can be estimated by maximizing the likelihood which is

$$\sum_{i=1}^{n} \log \prod_{t=1}^{T_i} \prod_{j=1}^{J} \pi_{jt}(\mathbf{x}_{it}, \mathbf{h}_{it})^{f_{ijt}}, \tag{2.10}$$

where f_{ijt} is one if subject i chooses category j at time point t, zero otherwise. We optimize this likelihood function with a quasi-Newton algorithm. The methods are implemented in MATLAB (Mathworks, 2007) and can be obtained from the author upon request.

2.3 Graphical representations

The model introduced earlier results in a graphical display. In this display, we use points for the response classes. Since the model is based on Euclidean distances solely, the odds of choosing one category j above another j' is even on the line perpendicular to the line joining j and j' and intersects at their centroid. To facilitate interpretation, we include these decision lines in the graphical display.

Previous responses are also represented as points in the same Euclidean space. When explanatory variables are available, these will be plotted for every previous response using variable axes with markers. When there exists an interaction between the previous response and the explanatory variables (as in 2.9), the axes will have different length and direction per previous response. Without such an interaction, the length and direction will be the same for all previous responses. Ideal points ($\mathbf{y}_{it}$) can be obtained by completing parallelograms (a more detailed explanation will be given in Section 4.2 and Figure 5). This is, however, cumbersome when the number of explanatory variables exceeds 2. In that case, it is easier to use the vector sum method as detailed in Gower and Hand (1996, 13).

3 Other approaches to deal with the dimensionality problem

The dimensionality problem of multinomial unordered data (as discussed in Section 1) has been attacked from two points of view. The first approach finds

continuous underlying dimensions that represent the differences between the observed classes. This approach is related to multidimensional scaling and principal component analysis. The second approach clusters the observed classes in a fewer number of latent classes. In this section, these approaches and their relationship towards our ideal point model will be discussed.

3.1 Continuous dimension reduction

We discuss three different models that use dimension reduction in the continuous sense. The first is the well-known RC(M)-association model (Goodman, 1979, 1985) with reparameterizations as proposed in De Rooij and Heiser (2005) and De Rooij (2007, 2008). The second is ideal point discriminant analysis (IPDA) (Takane, 1987; Takane *et al.*, 1987) and the last correspondence analysis. In all three of these models, main effect parameters for the classes are estimated. In our ideal point model, we do not have these. As is shown in De Rooij (2009a) and Takane (1998), such parameters complicate the interpretation of the graphical display.

3.1.1 RC(M)-association model

The RC(M)-association model (Goodman, 1979, 1985) is often used to analyse cross-classified data. The RC(M)-association model is defined as

$$\log(\mu_{ij}) = \lambda + \lambda_i^R + \lambda_j^C + \sum_{m=1}^{M} \phi_m \vartheta_{im} \nu_{jm}, \tag{3.1}$$

with μ_{ij} the expected frequencies under the model given a Poisson sampling scheme, λ a general intercept parameters, λ_i^R and λ_j^C main effect parameters for the rows and columns, respectively, ϕ_m represent M intrinsic association parameters and ϑ_{im} and ν_{jm} sets of row and column scores. Identification constraints are needed on the main effects and the row and column scores to obtain a unique solution.

Graphical displays are commonly used to interpret the RC(M)-association model. A joint graphical display of the rows and the columns can show how any one category of the row variable is associated with some category of the column variable. The association can be represented in a joint plot where the categories of the row variable have coordinates $\vartheta_{im}^* = \phi_m^\tau \vartheta_{im}$ and the categories of the column variable have coordinates $\nu_{jm}^* = \phi_m^\kappa \nu_{jm}$, where $\tau + \kappa = 1$. The association should be interpreted by an inner product rule. In order to do so, one set should be drawn using vectors and the points of the other set can be projected onto these vectors to represent the association.

De Rooij and Heiser (2005) show that the joint plot of the RC(M)-association model can also be interpreted using a distance rule. In order to do so, the main effects should be adapted for the squared terms in the Euclidean distance. These new 'main effects' can be interpreted in terms of masses to obtain a mass distance law of gravity

interpretation (see De Rooij, 2008). Identification issues for the RC(M)-association model in relation to a distance interpretation are extensively discussed in De Rooij (2007).

The RC(M)-association model can also be applied to more general contingency tables, e.g., where the rows are a cross-classification of several variables. If constraints are used on the row scores, one obtains a model very similar to ours. A main difference is that in our model there is no main effect for the column variable. Such main effects complicate the interpretation of the graphical display (Takane, 1998; De Rooij, 2009a).

3.1.2 Ideal point discriminant analysis

IPDA was first proposed by Takane *et al.* (1987) and described for contingency tables in Takane (1987). The model is very similar to ours except that it includes bias parameters for the categories of the response variable. The model is given by the following equation:

$$\pi_j(\mathbf{x}_i) = \frac{\beta_j \exp\left(-d^2(\mathbf{y}_i, \mathbf{z}_j)\right)}{\sum_l \beta_l \exp\left(-d^2(\mathbf{y}_i, \mathbf{z}_l)\right)}. \tag{3.2}$$

This model seems a bit more general than our model, but as is shown in De Rooij (2009a) in maximum dimensionality the two models are equal and in reduced dimensionality the fit of the two models is very close. The bias terms, however, do complicate the interpretation of the graphical display since decision boundaries shift away from the class with smallest bias term (Takane, 1998; De Rooij, 2009a).

Takane *et al.* (1987) and Takane (1987) proposed to constrain the class points to lie in the centroid of the subjects who chose that category. Such a constrain further reduces the number of parameters to be fitted. For the model without the centroid restriction, Takane (1987) showed that it is equivalent to the RC(M)-association model.

3.1.3 Correspondence analysis

Correspondence analysis is a method for visualization of association in a contingency table. It has a long history, an overview can be found, e.g., in Greenacre (2007). Correspondence analysis is a general technique that is represented by the following equation:

$$\pi_{ij} = \pi_i \pi_j \left(1 + \sum_m \phi_m \vartheta_{im} \nu_{jm}\right).$$

The model structure is very similar to the RC(M)-association model, for differences and similarities see Goodman (1991). Often the correspondence analysis model is estimated using least squares. There exists, however, also a maximum likelihood variant. Correspondence analysis, like the methods discussed earlier, also results

in a graphical display of the row and column entities which should be interpreted using an inner product rule. In many cases, the π_i and π_j terms are neglected in the interpretation of correspondence analysis.

Van der Heijden and De Leeuw (1985) showed that correspondence analysis gives a graphical display of the residuals of the log-linear model of independence. They then generalized this procedure to represent residuals of other log-linear models for multiway data and two-way tables with special structure. Extensions of correspondence analysis with external variables has been proposed by Böckenholt and Takane (1994). As is shown in Van der Heijden *et al.* (1994), the RC(M)-association model, IPDA and correspondence analysis (both least squares and maximum likelihood) often give approximately the same results.

3.2 Discrete dimensional reduction

Another possibility to reduce the dimensionality is to further classify the categories of the multinomial response variable into latent classes. Here, we briefly discuss this approach and a generalization to longitudinal data.

3.2.1 Latent class and Markov model

The latent class model seeks to explain observed relationships among several discrete variables with a set of latent classes. The relationships among the observed variables are explained by the class membership of a subject. That is, given that a subject is in a specific class, the responses are independent. The basic latent class model for a three-way contingency table is given by

$$\pi_{abc} = \sum_w \pi_w \pi_{a|w} \pi_{b|w} \pi_{c|w},$$

where π_{abc} is the joint probability of an observation in cell abc. The variable with classes w represents an unobserved categorical variable. The $\pi_{a|w}$ represent conditional response probabilities given the latent class w. For each class there is a unique set of conditional response probabilities. Early treatments of the latent class model are given by Green (1951) and Lazarsfeld and Henry (1968). Goodman (1974) extended the methodology to nominal variables and proposed a maximum likelihood algorithm to obtain parameter estimates. The latent class model is closely related to correspondence analysis as was shown in Van der Heijden *et al.* (1999).

For longitudinal data, the latent class model has been extended to the latent Markov model. In this model, one (or more) discrete variables are observed over time. It is assumed that a few number of latent classes underlie the responses, as in the latent class model. A Markov model is then used to model the transitions between the latent classes. Often the conditional response probabilities are assumed to be homogeneous over time (the stationary Markov chain assumption). This model was first proposed by Wiggins (1973) and later extended by Vermunt *et al.* (1999)

to include time constant and time-varying predictors on the transition probabilities as well as the response probabilities.

3.3 Comparison to our IPCM

We briefly discussed two types of model that reduce the dimensionality of a multinomial response variable. The first used multidimensional scaling, principal component type of dimension reduction. We discussed models with inner product interpretation (RC(M)-association model and correspondence analysis) and distance interpretations (IPDA and mass distance law of gravity models). In all these models, main effect parameters for the response variable are included in the model. These make interpretations of related graphical displays cumbersome as was first shown in Takane (1998). Our IPCM does not have such parameters, and therefore the graphical display can be interpreted using the pure Euclidean distance function.

Another strategy to reduce the dimensionality is to cluster the categories of the response variable into a smaller number of classes. This can be done using latent class models and extensions thereoff. The choice between continuous or discrete dimension reduction is a difficult one and depends on the application context.

4 Applications

4.1 Pure transitional data

For illustration, the model will be applied to data obtained from Upton (1978, 128) where a sample of 1651 Swedish people were asked for their votes at three consecutive elections. There are four political parties: the *Social democrats* (SD), the *Center party* (C), the *People's party* (P) and the *Conservatives* (Con). The distribution of votes is shown in Figure 1 where it can be seen that the SD lose some votes in 1970, C gains twice, P lose and regain, while the Con lose a bit in the last election. For a politician of a losing party, an important question is 'where did my votes go?'. For a politician of a winning party, it is important to know where the subjects who now voted on his (her) party came from.

These data were analysed by De Rooij (2008) and will be treated here again. Whereas De Rooij (2008) models the joint probability of the three observations, here we model the conditional observation of the 1968 given 1964 and the conditional observation of 1970 given 1968 and 1964. More similarities and differences with the results of De Rooij (2008) will be discussed shortly. For comparisons with the models in De Rooij (2008), we consider models in two-dimensional space only.

The log-likelihood of the first-order stationary Markov model equals -1726.7; adding the responses of two time points back has a significant contribution the likelihood ratio statistic equals 140.57 with 6 degrees of freedom. Allowing for an interaction between the previous vote and the vote two time points back provides a

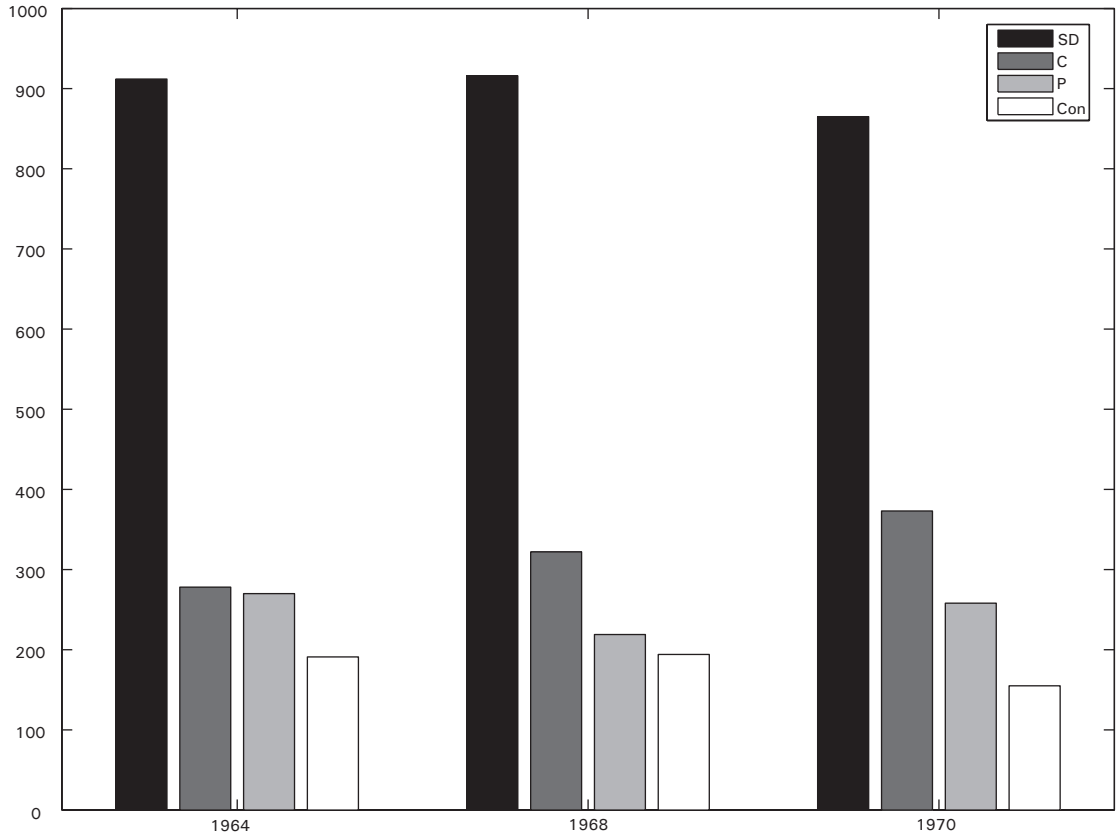

Figure 1 Histograms of Swedish vote data for the three time points

significant better solution (likelihood ratio statistics equals 69.87 with 18 degrees of freedom). The solution of this two-dimensional model is shown in Figure 2.

For the 1968 votes, only one previous vote is available, so for interpretation one should look at the single-coded points in the middle of the 'stars'. These are all relatively close to the class points of the same political party, from which it can be concluded that most subjects repeat their previous vote. In general, the star centers are closer to certain borders than the class points. For previous P voters, the point is close to the class point of P but towards the decision line with C; estimated (rounded) conditional probabilities are 0.02 (SD), 0.27 (C), 0.58 (P) and 0.12 (Con). The position of previous SD voters is pulled away from the SD class point along the direction of the decision line with C into the direction of Con. Estimated conditional probabilities are 0.94 (SD), 0.04 (C), 0.01 (P) and 0.02 (Con), i.e., a very large probability of staying at the SD. For previous C voters, the position is close to the C class point, somewhat into the direction of the three other parties. The estimated

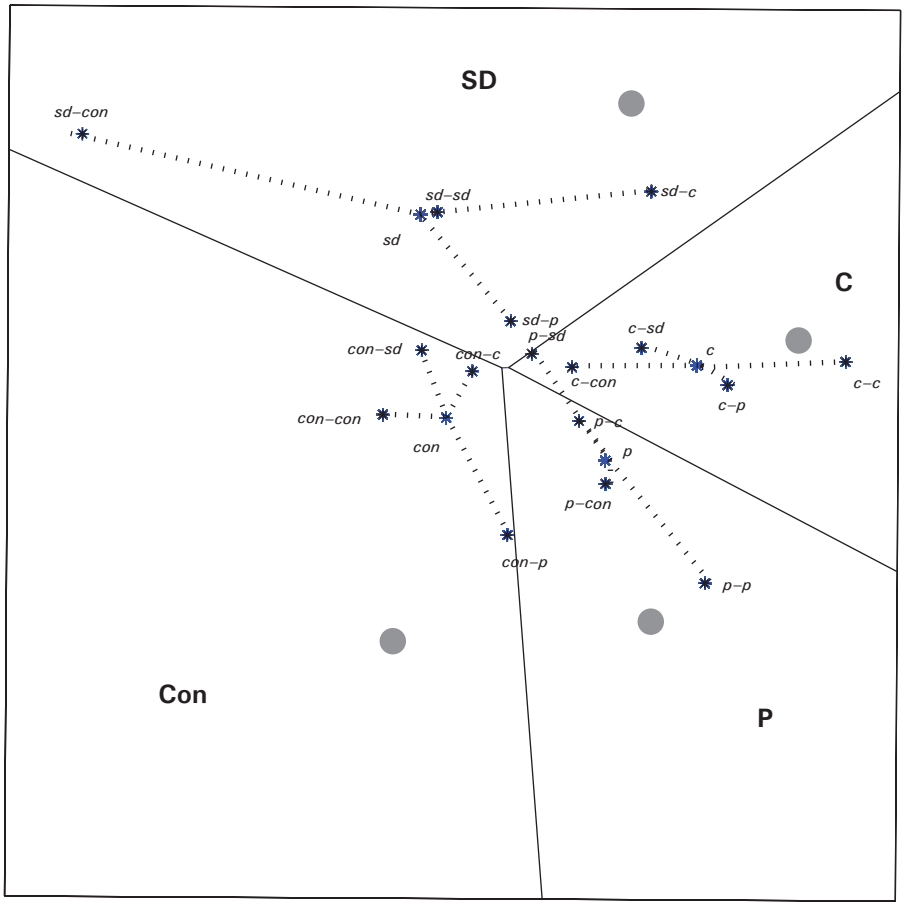

Figure 2 Two-dimensional solution of Swedish vote data. The large dots represent the response variable points, whereas the smaller dots represent the explanatory (history) variables. Double-coded points, e.g., *c-p* stands for previous vote C and the vote two time points back was P. Single-coded points *c* denote the previous choice. Decision boundaries are given which facilitate the interpretation

conditional probabilities are 0.11 (SD), 0.76 (C), 0.13 (P) and 0.01 (Con). Similarly, for the previous Con voters, the position is near the Con class point but somewhat into the direction of all other parties; estimated conditional probabilities are 0.05 (SD), 0.06 (C), 0.25 (P), 0.65 (Con).

For the 1970 votes the two previous choices are responsible for the prediction and the double-coded points are of use. The general pattern is that if the two previous choices were the same (say twice C), the point is more firmly into its own prediction region (further away from the decision boundaries with SD, P and Con), increasing the odds of the same vote again. Another general pattern is that when between 1964 and 1968 a transition was made, the probability of going back to the 1964 choice is

increased. There is only one single history, P–SD, where the probability of a change is the largest, i.e., only the point for *p-sd* is across the decision line. Note that with this history, the estimated odds seem to be in favour of C. Estimated conditional probabilities are 0.32 (SD), 0.34 (C), 0.21 (P) and 0.13 (Con). Looking at the data, there are nine people with this history, of which two vote for SD, three for C and four for P. So, although the odds are in favour of leaving P (5 to 4), all pairwise odds are in favour of staying (i.e., 2 to 4 and 3 to 4). The ideal point for subjects with history *p-sd* represents all this information.

The two-dimensional picture gives a concise summary of the data, with a very easy interpretation. The model is much easier compared to that in De Rooij (2008) where the joint probabilities were modelled. In the modelling of the joint probabilities, extra parameters were needed to deal with 'stayers', i.e., the people who voted for the same party on all three occasions. Such parameters complicate the interpretation of the model. The graphical representation in Figure 2 pertains to all data, and not only to the 'movers'. Another difference is that in the modelling framework of De Rooij (2008), masses are attached to the political parties at all three time points. These complicate the interpretation of the Euclidean space, since decision boundaries are not exactly in the middle of two points but are shifted away from the points with the larger masses, sometimes even beyond the other point. A final difference is that in this modelling approach, an interaction between the two previous time points is needed to model the 1970 votes, whereas in the model for joint probabilities only pairwise association were needed. This maybe due to the absence of masses or parameters for the stayers in our transitional model.

4.2 Change data with explanatory variables

This second example, in which data from the Dutch parliamentary election studies 2002–03 (Irwin *et al.*, 2003) will be analysed, shows an example with explanatory variables. Although many more political parties took part in the election, we will confine our analysis to the eight largest parties: the labor party (PvdA), the Christian democratic party (CDA), the conservative liberals (VVD), the progressive liberals (D66), the green left party (GL), the Christian Union (CU), List Pim Fortuyn (LPF) and the Socialists party (SP). The subjects were asked their vote intention before the election of 2002, their vote at the election of May 2002 and the election of January 2003.

The sample size is $n = 872$ with complete data on all three occasions. Histograms of the vote distribution are given in Figure 3 where considerable changes can be seen. A decline of the LPF in 2003, a rise of the CDA over time, the PvdA that gains in 2003 and a decline of the GL over time. Besides the choices at each of the three time points, we have two background variables both measured on a five-point scale: An assessment of one's own *social class* (–2, working class; –1, upper working class; 0, middle class; 1, upper middle class; 2, upper class) and *degree of urbanization* (–2, not urban [1–499 adresses/km^2]; –1, hardly urban [500–999

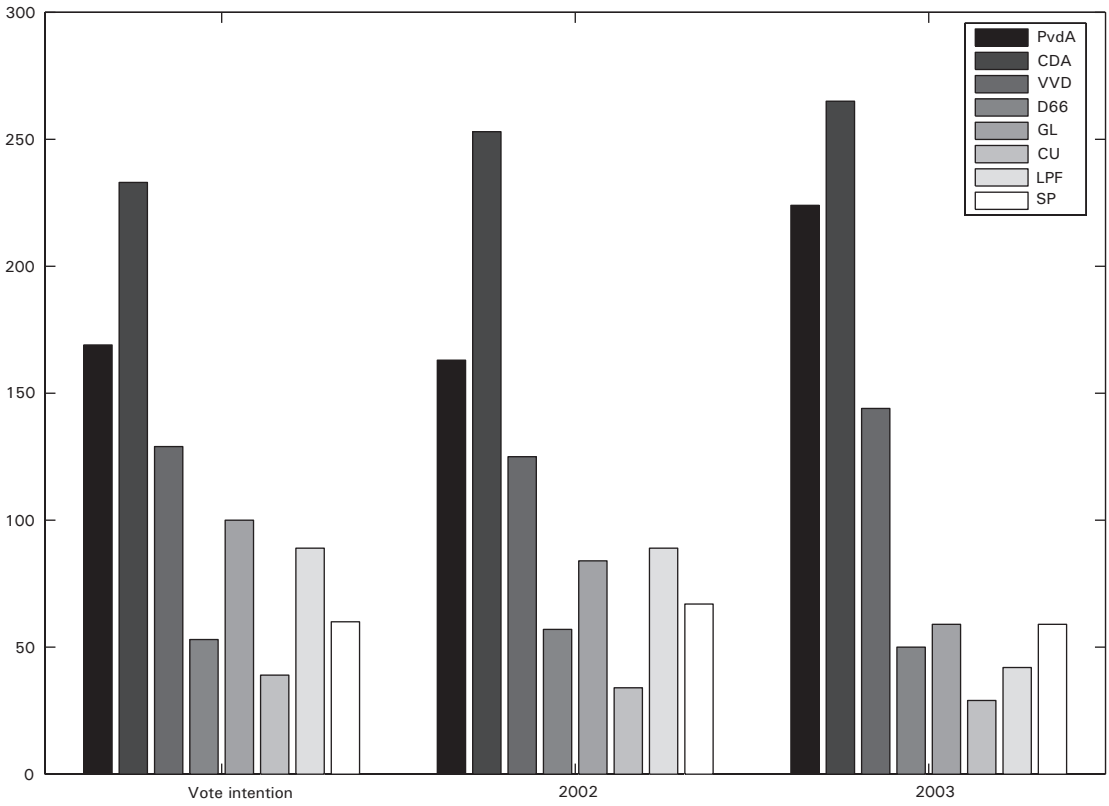

Figure 3 Histograms of Dutch parliamentary election studies for the three time points

adresses/km^2]; 0, mildly urban [1000–1499 adresses/km^2]; 1, strongly urban [1500–2499 adresses/km^2]; 2, very strongly urban [>2500 adresses/km^2]). These variables were treated as continuous as is often done with variables measured on a five-point scale. As in the previous section, for losing politicians, it is important to know where votes went, and for winning politicians, where did the votes come from. In this case, extra information is available, such that the question can be raised: what type of people went to another party (i.e., subjects with a high social economic profile from the large city left our party) or what type of people did we gain?

It is often assumed that the Dutch political system is two dimensional: a left–right continuum and a progressive–conservative one (Pennings and Keman, 2003; Van Holsteyn and Irwin, 2003). Therefore, we will do all analyses in two dimensions. Compared to the multinomial logit model, this gives a substantial reduction in the number of parameters to be estimated. We started with a stationary first-order Markov model (log-likelihood equals –1480.9) and added the two explanatory variables. They have a significant contribution, the likelihood ratio statistic being

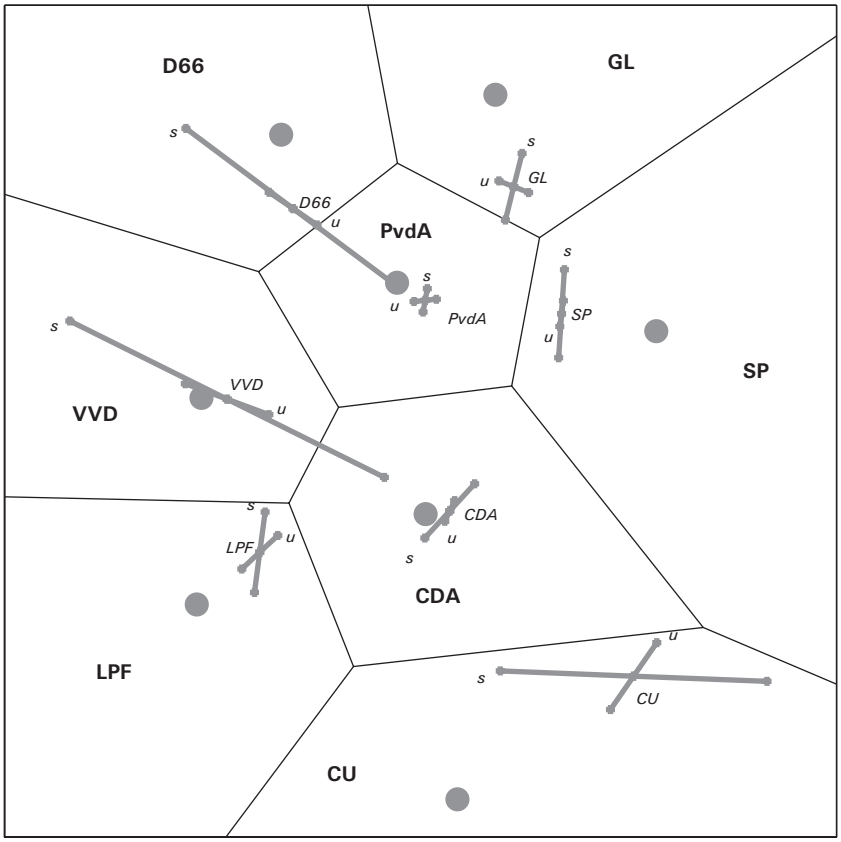

Figure 4 Two-dimensional solution of Dutch parliamentary election studies. The large dots represent the positions of the categories of the response variables. The name of the parties is within the region. For the labels of the political parties see text. Labels of the variables are at the positive ends of the variables: u is for the variable degree of urbanization and s for social class

14.39 with 4 degrees of freedom. Then it was seen whether there was an interaction between the two explanatory variables and the previous vote, the likelihood ratio statistic equals 77.91 with 28 degrees of freedom. The solution of the last analysis is shown in Figure 4.

The configuration of response categories is as expected (see Pennings and Keman, 2003, Figure 3). The two explanatory variables, social class and degree of urbanization, have a different direction and magnitude for each of the previous votes (as it should be when an interaction exists). When the previous vote was CDA or PvdA, the two explanatory variables do not matter much (their magnitude is small), contrary to a previous VVD, D66 or CU vote where the variables play a major role.

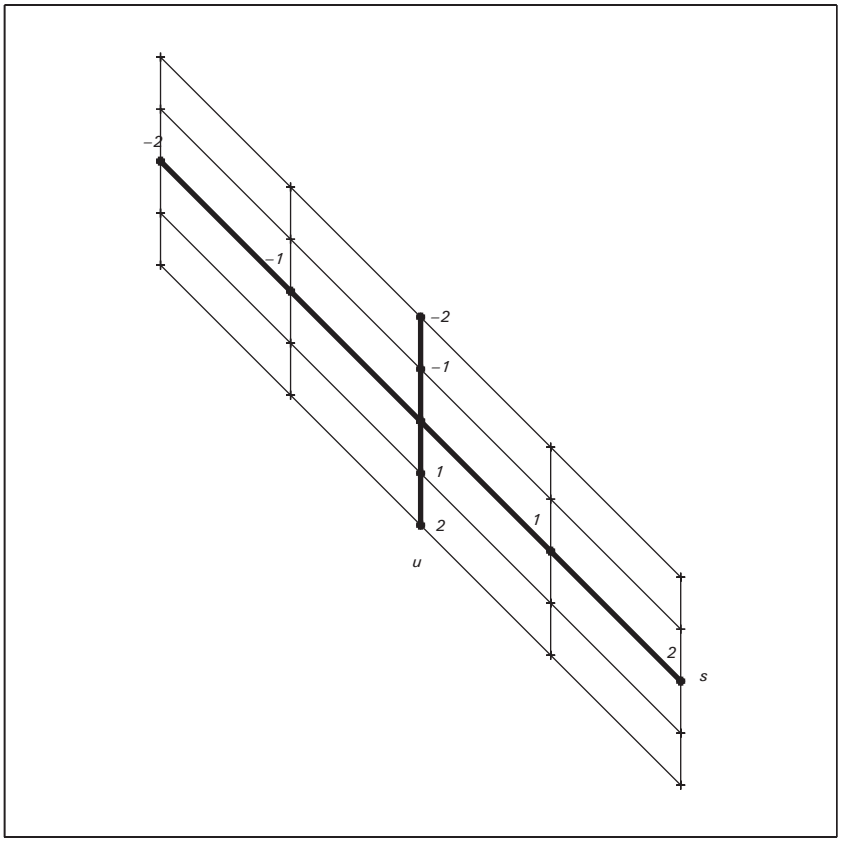

Figure 5 Obtaining the ideal points using a grid of the two explanatory variables. Labels of the variables are at the positive ends of the variables: u is for the variable degree of urbanization and s for social class

Ideal points for a subject with a given previous vote and given values on the two explanatory variables can be obtained using a grid (this is the completing parallelograms method mentioned earlier). This is illustrated in Figure 5. A grid is formed by the two explanatory variables and their marker values, i.e., the possible values for each of the variables. Each intersection of the grid lines gives a position of an ideal point. For example, a subject with values –2 on both variables has his (her) ideal point at the upper left point of the grid, while a subject with +2 for urbanization and –2 for social class is in the lower left point of the grid. Alternatively, the vector sum method can be used in which case one takes the centroid of the markers on the variable axes and multiplies the length of the origin of the grid till this centroid by 2 (the number of explanatory variables) to obtain the ideal point. This latter method can be used with any number of explanatory variables. Using this information about the ideal points, we can conclude the following.

Table 2 Cross-classification of 332 subjects whose previous vote was PvdA of social class (S) with new vote and urbanization (U) with new vote. Bold numbers indicate highlighted elements mentioned in text

Variable	Score	PvdA	CDA	VVD	D66	GL	CU	LPF	SP
S	−2	52	**3**	1	0	1	0	0	1
	−1	23	**2**	0	0	1	0	0	1
	0	142	5	1	4	6	0	1	3
	1	60	1	0	**4**	1	0	0	3
	2	14	1	0	0	1	0	0	0
U	−2	41	3	0	2	2	0	0	1
	−1	59	3	0	0	3	0	0	1
	0	58	2	0	0	1	0	0	4
	1	71	2	**1**	**4**	3	0	0	1
	2	62	2	**1**	**2**	1	0	**1**	1

1. First we look at previous PvdA voters. Low social class subjects seem to have increased probabilities for CDA while high social class subjects have increased probabilities for GL and D66. For subjects with a high degree of urbanization the probabilities for VVD, LPF, and D66 increase. Table 2 gives the cross-classification of social class/urbanization with the new vote for the people who previously voted PvdA. Looking at these data, we see that most people vote PvdA again, which explains the very small variable axes. There are minor asymmetries which explain the directions of the variables (bold numbers), i.e., from the subjects who are going to vote CDA, five have low social class while only two have high social class; from the subjects who are going to vote D66, four have high social class while none has low social class. This concurs with the direction of the social class variable for previous PvdA voters. Looking at degree of urbanization, we see that the direction of this variable is due to nine people: two of whom go to VVD, six to D66 and 1 to the LPF.
2. Previous VVD voters with a low social class have a high probability of voting CDA. Looking back into the data, we find six subjects who voted VVD and have a score –2 or –1 on the social class variable. Of these six subjects, three voted on the next occasion for CDA, one for PvdA, one for SP and one for the VVD.
3. Previous D66 voters with low social class seem to leave D66 with increased probabilities for SP and PvdA. Looking at the data, there are nine of these subjects, three of whom vote PvdA, one SP, one VVD and four D66 again.
4. Previous CU voters with a high degree of urbanization seem to have increased probability for voting CDA or SP. In the data, there are 13 subjects scoring +2 or +1 on urbanization of whom five voted for CDA, one for SP and seven for CU again.

5 Discussion

We proposed a transitional model for multinomial longitudinal data. Such data are often collected in the social sciences (as is shown in the two examples) but also in consumer studies (which brand of yoghurt is bought) and medical and health studies (does the participant live on the street, in a community house or independently Hedeker and Gibbons, 2006). It can be argued that the problem with multinomial data is its dimensionality: when the response variable has many categories, many regression equations have to be built in a multinomial regression model. In this paper, we proposed a model with two advantages. The first is dimension reduction, the number of parameters may be greatly reduced. In the example on the Dutch parliamentary election studies, e.g., we reduced the number of parameters from 168 in the multinomial logit model to 61 in our two-dimensional model. The second virtue is visualization as shown in the applications. A new type of biplot display was created to deal with previous choices and explanatory variables in a satisfactory way. The combined effects of the two previous choices (Section 4.1) or the main and interaction effects (Section 4.2) can be easily represented in a graphical display, whereas getting the same information from the regression weights of a multinomial logit model would be awkward. The graphical display in Figure 4 provides a very detailed description of the voting mechanics. The display can be used for direct interpretation but also highlights aspects of the data that merit closer attention. It provides a much clearer view on voting transitions than can ever be obtained through the multinomial regression model.

Another way to reduce the dimensionality of the multinomial logit model is through latent class or latent Markov models. For the examples shown, it would mean that the political parties are grouped into clusters. For example, a left wing cluster with PvdA, GL and SP; a Christian cluster with CDA and CU and a right wing liberal cluster with VVD, D66 and LPF. By doing so, the research questions stated in Sections 4.1 and 4.2 which are of central interest to politicians cannot be answered since in that case transitions are between clusters and not between parties. In that case, clusters lose votes to other clusters, but that does not give insight into the question which type of voters left a specific party and where these voters went.

Compared to the work of De Rooij (2008), the current models can be more easily extended to situations with more than three time points, whereas the models presented in De Rooij (2008) become very restricted for such cases and their interpretation becomes challenging. Furthermore, the current models allow for explanatory variables, which might be discrete or continuous. Such information is often available and is of central interest.

In the transitional models, as proposed, the responses are independent given the model structure. For example, in the example on the Dutch parliamentary elections, it is assumed that the association among the responses is accounted for by including the previous vote and the two background variables as explanatory variable. This assumption could be tested by including the vote of two time points back or more background information, if such information is available. However, it is not very

likely that all the dependencies among the responses are captured by including the previous vote as explanatory variable. From the theory of generalized estimating equations (Liang and Zeger, 1986), however, we know that optimizing function 2.10 still gives unbiased estimates of the model parameters. For ideal point models, under such an ill-conditioned likelihood function, Yu and De Rooij (forthcoming) show that likelihood ratio statistics and the Bayesian information criterion (BIC) are appropriate model selection tools.

A topic we did not touch is dimension selection. In both analyses shown, we had theoretical reasons to use a two-dimensional space. If there are no such reasons, selection of the appropriate dimensionality can be performed using information criterion like the Akaike's information criteria or BIC (Yu and De Rooij, forthcoming). There are indications that likelihood ratio statistics when used for dimension selection are not chi-squared distributed (Takane *et al.*, 2003). However, Yu and De Rooij (forthcoming) show that the likelihood ratio statistic performs quite well for determination of the dimensionality.

The transitional approach is criticized in Diggle *et al.* (2002, 142–44). We think that in the examples shown, a transitional analysis is the most natural approach. For a losing politician, it answers the question: Where did our votes go? For a political party that gained votes it answers the question: Where did our voters come from? This is essential information that cannot be obtained from a marginal or a subject-specific data analysis approach.

Acknowledgements

The author thanks Willem Heiser for his useful comments and remarks on an earlier version of this manuscript. This research was conducted while the author was sponsored by the Netherlands Organization for Scientific Research, Innovational Grant No. 452-06-002.

References

Agresti A (2002) *Categorical data analysis*, 2nd edition. New York: John Wiley and Sons.

Böckenholt U and Takane Y (1994) Linear constraints in correspondence analysis. In Greenacre MJ and Blasius J (eds.). *Correspondence analysis in the social sciences: recent developments and applications*. New York: Academic Press, 112–27.

Bonney GE (1987) Logistic regression for dependent binary observations. *Biometrics*, **43**, 951–73.

De Rooij M (2001) Distance association models for the analysis of repeated transition frequency tables. *Statistica Neerlandica*, **55**, 157–81.

De Rooij M (2002) Distance models for three-way tables and three-way association. *Journal of Classification*, **19**, 161–78.

De Rooij M (2007) The distance perspective of generalized biadditive models: scalings and tranformations. *Journal of Computatoinal and Graphical Statistics*, **16**, 210–27.

De Rooij M (2008) The analysis of change, Newton's law of gravity and association models. *Journal of the Royal Statistical Society, Series A*, **171**, 137–57.

De Rooij M (2009a) Ideal point discriminant analysis revisited with an emphasis on visualisation. *Psychometrika*, **74**, 317–30.

De Rooij M (2009b) Trend vector models for the analysis of change in continuous time for multiple groups. *Computational Statistics and Data Analysis*, **53**, 3209–16.

De Rooij M and Heiser WJ (2005) Graphical representations and odds ratios in a distance-association model for the analysis of cross-classified data. *Psychometrika*, **70**, 99–123.

Diggle PJ, Heagerty P, Liang K-Y and Zeger SL (2002) *Analysis of longitudinal data*. Oxford: Oxford University Press.

Goodman LA (1974) Exploratory latent structure analysis usingboth identifiable and unidentifiable models. *Biometrika*, **61**, 215–31.

Goodman LA (1979) Simple models for the analysis of association in cross classifications having ordered categories. *Journal of the American Statistical Association*, **74**, 537–52.

Goodman LA (1985) The analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models, and asymmetric models for contingency tables with or without missing entries. *The Annals of Statistics*, **13**, 10–69.

Goodman LA (1991) Measures, models, and graphical displays in the analysis of cross-classified data. *Journal of the American Statistical Association*, **86**, 1085–111.

Gower JC and Hand DJ (1996) *Biplots*. London: Chapman & Hall.

Green BF (1951) A general solution of the latent class model of latent structure analysis and latent profile analysis. *Psychometrika*, **16**, 151–66.

Greencacre MJ (2007) *Correspondence analysis in practice*. London: Chapman & Hall.

Hartzel J, Agresti A and Caffo B (2001) Multinomial logit random effects models. *Statistical Modeling*, **1**, 81–102.

Hedeker D (1999) MIXNO: A computer program for mixed-effects nominal logistic regression. *Journal of Statistical Software*, **4**, 1–92.

Hedeker D and Gibbons RD (2006) *Longitudinal data analysis*. Hoboken, NJ: John Wiley and Sons.

Heiser WJ (1981) *Unfolding analysis of proximity data*. Phd thesis, Leiden University.

Irwin GA and Van Holsteyn JJM and Den Ridder JM (2003) *Dutch parliamentary election study 2002–2003: an enterprise of the foundation for electoral research in the netherlands (SKON)*, computerfile Amsterdam: Steinmetz Archives, P1628.

Lazarsfeld PF and Henry NW (1968) *Latent structure analysis*. Boston: Houghton Mifflin.

Liang K-Y and Zeger SL (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Lipsitz SR, Kim K and Zhao L (1994) Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine*, **13**, 1149–63.

Mathworks (2007) *MATLAB: The language of technical computing*. Natick: Mathworks.

Molenberghs G and Verbeke G (2005) *Models for discrete longitudinal data*. New York: Springer.

Pennings P and Keman H (2003) The Dutch parliamentary election studies in 2002 and 2003: the rise and decline of the Fortuyn movement. *Acta Politica*, **38**, 51–68.

Takane Y (1987) Analysis of contingency tables by ideal point discriminant analysis. *Psychometrika*, **52**, 493–513.

Takane Y (1998) Visualization in ideal point discriminant analysis. In Blasius J and

Greenacre MJ (eds.). *Visualization of categorical data*. New York: Academic Press, 441–59.

Takane Y, Bozdogan H and Shibayama T (1987) Ideal point discriminant analysis. *Psychometrika*, **52**, 371–92.

Takane Y, Van der Heijden PGM and Browne MW (2003) On likelihood ratio tests for dimensionality selection. In Higuchi T, Iba Y and Ishiguro M (eds.). *Proceedings of science of modeling: the 30th anniversary meeting of the information criterion (AIC)*. Tokyo: The Institute of Statistical Mathematics, 348–49.

Upton GJG (1978) *The analysis of cross-tabulated data*. Chichester: John Wiley and Sons.

Van der Heijden PGM and De Leeuw J (1985) Correspondence analysis used complementary to loglinear analysis. *Psychometrika*, **50**, 429–47.

Van der Heijden PGM, Gilula Z and Van der Ark LA (1999) An extended study into the relationships between correspondence analysis and latent class analysis. In Sobel M and Becker M (eds.). *Sociological methodology*. Cambridge: Blackwell, 147–86.

Van der Heijden PGM, Mooijaart A and Takane Y (1994) Correspondence analysis and contingency models. In Greenacre MJ and Blasius J (eds.). *Correspondence analysis in the social sciences*. New York: Academic Press, 79–111.

Van Holsteyn JJM and Irwin GA (2003) Never a dull moment: Pim Fortuyn and the Dutch parliamentary election of 2002. *West European Politics*, **26**, 41–66.

Vermunt JK, Langeheine R and Böckenholt U (1999) Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, **24**, 179–207.

Wiggins L (1973) *Panel analysis: latent probability models for attitude and behavior processes*. New York: Elsevier Scientific Publishing Company.

Yu H-T and De Rooij M (forthcoming). Model selection for the trend vector model.