The Mixed Effects Trend Vector Model

Mark de Rooij[1] & Martijn Schouteden[2]

[1] Leiden University, Psychological Institute

[2] Katholieke Universiteit Leuven, Department of Psycholgy

November 9, 2011

Abstract

Maximum likelihood estimation of mixed effect baseline category logit models for multinomial longitudinal data can be prohibitive due to the integral dimension of the random effects distribution. We propose to use multidimensional unfolding methodology to reduce the dimensionality of the problem. As a by product readily interpretable graphical displays representing change are obtained. The methodology can be applied to both nominal as well as ordinal response variables. Relationships to standard statistical models for multinomial data will be presented. Several empirical examples will be given to show the merits of the proposed modeling framework.

Key-Words: Categorical data; Longitudinal data; Gauss-Hermite quadrature; Multilevel model; Multidimensional Scaling; Multidimensional Unfolding.

1 Introduction

In this paper we will present a new methodology for multinomial longitudinal data. Such data are often collected in social and medical sciences. An important class of models for longitudinal data are the mixed effects models, also called multilevel models for change (Singer and Willett, 2003). For multinomial data, however, the main representative of this class has several problems, as elaborated below. With the new methodology we aim to solve these problems. We start this introduction by showing three examples of empirical data that will be used to present our methodology, followed by an extensive problem formulation on the basis of one of the examples.

1.1 Longitudinal multinomial data

Consider, as a first detailed example, data from the McKinney Homeless Research Project in San Diego as described in chapters 10 and 11 in the book by Hedeker and Gibbons (2006). The aim of this project was to evaluate the effectiveness of using an incentive as a means of providing independent housing to homeless people with severe mental illness. Housing certificates were allocated to local authorities in San Diego by the Department of Housing and Urban Development. These housing certificates were designed to make it possible for individuals with low income to choose and obtain independent housing in the community. A sample of 361 individuals took part in this longitudinal study and were randomly assigned to the experimental or control condition. Eligibility for the project was restricted to individuals diagnosed with a severe and persistent mental illness who were either homeless or at high risk of becoming homeless at the start of the study. Individuals' housing status was assessed using three categories (living on the street / living in a community center / living independently) at baseline and at 6, 12, and 24 month follow up.

In the second example (data can be found in Adachi (2000)) 49 male and 51 female Japanese undergraduates were followed from the age of 6 to 20. At five time points they were asked which of six TV program categories they liked best. The five time points are the first year of elementary school (ages 6-7), the fourth year of elementary school (ages 9-10), first year of junior high school (ages 12-13), first year of high school (ages 15-16), and as university freshmen (ages 18-20). The six TV program categories were:

Animation (A), Cinema (C), Drama (D), Music (M), Sport (S), and Variety (V). In this case the response variable had six categories and was measured at five time points. There was one background variable, gender.

Although our focus will primarily be on models for unordered (nominal) categorical response variables, they are also applicable to ordinal multinomial data. In the third example (data were described in Hedeker and Gibbons (2006)) 437 schizophrenic patients were randomly assigned to receive one of four medications: placebo, chlorpromazine, fluphenazine, or thioridazine. Patients were measured at weeks 0, 1, 3, and 6. The observed response variable had four ordered categories: (1) normal or borderline mentally ill, (2) mildly or moderately ill, (3) markedly ill, and (4) severely or among the most extremely ill. The main question was whether there is a differential change across time for the treatment group (chlorpromazine, fluphenazine, and thioridazine) relative to the placebo group. Various descriptive statistics can be found in Hedeker and Gibbons (2006).

1.2 Notation and data format

Before we discuss modeling the type of data presented above, we introduce some notation. A sample consists of n subjects and for each subject i ($i = 1, \ldots, n$) there are measurements on n_i occasions. Let $\mathcal{G}_{it}$ denote the t-th observation ($t = 1, \ldots, n_i$) for subject i, with $\mathcal{G}_{it} = c$ ($c = 1, \ldots, C$) and response probabilities $\pi_{itc} = P(\mathcal{G}_{it} = c)$. Furthermore let $\mathbf{g}_{it}$ be the corresponding vector $\mathbf{g}_{it} = [g_{it1}, \ldots, g_{itC}]^\mathsf{T}$ with $g_{itc} = 1$ if subject i at time point t belongs to/chooses/is diagnosed as category c ($c = 1, \ldots, C$) and $g_{itc} = 0$ otherwise. For every subject at every time point there are p explanatory variables x_{itj}, $j = 1, \ldots, p$. The general layout of the data as used in this paper is shown in Table 1. The data are in so-called person-time format where for each subject on a specific time point the data are given on a single row of the matrix. This data format is common for mixed effects models.

1.3 Mixed effects multinomial baseline category logit models

One important class of models for longitudinal data are the mixed effects models, also known as multilevel models, random effects models, or subject specific models. For nominal response data the Multinomial Baseline Category Model (MBCL, Agresti (2002)) is a standard regression model. The mixed effects MBCL model (Hartzel, Agresti and

Table 1: The data format

Subject	Time	Response	Explanatory variables				
1	1	$\mathcal{G}_{11}$	x_{111}	x_{112}	x_{113}	$\dots$	x_{11p}
1	2	$\mathcal{G}_{12}$	x_{121}	x_{122}	x_{123}	$\dots$	x_{12p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	n_1	$\mathcal{G}_{1n_1}$	x_{1n_11}	x_{1n_12}	x_{1n_13}	$\dots$	x_{1n_1p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	t	$\mathcal{G}_{it}$	x_{it1}	x_{it2}	x_{it3}	$\dots$	x_{itp}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	n_n	$\mathcal{G}_{nn_n}$	x_{nn_n1}	x_{nn_n2}	x_{nn_n3}	$\dots$	x_{nn_np}

Caffo, 2001; Hedeker, 2003) is the generalization of the MBCL model with random effects and is the main representative of the class of mixed effects models. For the mixed effects MBCL model the conditional distribution of $\mathbf{g}_{it}$ given a set of subject specific parameters $\mathbf{u}_i$, $f(\mathbf{g}_{it}|\mathbf{u}_i)$, is the multinomial distribution with expectation

$$E(\mathbf{g}_{it}|\mathbf{u}_i) = \boldsymbol{\pi}_{it} = [\pi_{it1}, \dots, \pi_{itC}]^\mathsf{T}.$$

The probabilities (π_{itc}) are related to a linear predictor (η_{itc}) by the vector of link functions $\mathbf{h}_l(\cdot)$, i.e.

$$\boldsymbol{\pi}_{it} = \mathbf{h}_l(\boldsymbol{\eta}_{it}),$$

and $\mathbf{h}_l(\cdot) = [h_{l1}(\cdot), \dots, h_{lC}(\cdot)]$, where $h_{lc}(\cdot)$ is

$$h_{lc}(\eta_{it1}, \dots, \eta_{itC}) = \frac{\exp(\eta_{itc})}{\sum_h \exp(\eta_{ith})}.$$

The c-th linear predictor is given by

$$\eta_{itc} = \alpha_c + \mathbf{x}_{it}^\mathsf{T}\boldsymbol{\beta}_c + \mathbf{z}_{it}^\mathsf{T}\mathbf{u}_{ic},$$

where $\mathbf{x}_{it}$ is the design vector for the fixed effects, $\mathbf{z}_{it}$ is the design vector for the random effects, and α_c, $\boldsymbol{\beta}_c$ are fixed effect parameters (see the next section for a detailed example of $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$). In order to identify the model, one set of parameters is fixed to zero, that is $\alpha_1 = 0$, $\boldsymbol{\beta}_1 = \mathbf{0}$, and $\mathbf{u}_{i1} = 0$. A multivariate normal distribution is assumed for the random effects, that is

$$\mathbf{u}_{ic} \sim N(\mathbf{0}, \mathbf{\Sigma}), c = 2, \ldots, C.$$

Conditional on the random effects it is assumed that the observations for each subject are independent. To obtain maximum likelihood estimates both Hartzel, Agresti and Caffo (2001) and Hedeker (2003) use Gauss-Hermite quadrature procedure to approximate the likelihood function (see Section 3.2) .

1.4 Application of the mixed effects MBCL model

The McKinney Homeless Research Project (MHRP) data (first example above) are analyzed here with the mixed effects MBCL model. In Hedeker and Gibbons (2006) proportions for each of the three response categories at the four time points are given for the experimental and control group. Hedeker and Gibbons (2006) show that some proportions first go up and then go down and other proportions first go down and then go up. Moreover, the proportions for the two groups are rather different. This information led us to define a model with different quadratic time trends for the two groups and random intercepts. The linear predictor for this model equals

$$\eta_{itc} = \alpha_c + \beta_{1c}G_i + \beta_{2c}T_{it} + \beta_{3c}T_{it}^2 + \beta_{4c}G_iT_{it} + \beta_{5c}G_iT_{it}^2 + u_{ic},$$

where G_i is an indicator for group membership ($G_i = 1$ for incentive) for participant i, and T_{it} represents the time variable with Month - 10. In this model the design vector for the fixed effects is $\mathbf{x}_{it} = [G_i, T_{it}, T_{it}^2, G_iT_{it}, G_iT_{it}^2]^\mathsf{T}$, and the design vector for the random effects is $\mathbf{z}_{it} = [1]$. The parameter estimates for this model are given in Table 2. It can be seen that for both contrasts (community versus street and independent versus street) the time effect is positive but the quadratic time effect is negative. These estimates give the development over time for the control group. For the experimental group, which did receive the incentive, parameter estimates have to be combined in order to find the development over time. The time effect in the contrast community/street equals $0.1599 + (-0.0862) = 0.0737$ and the squared time effect is $-0.0144 + 0.0077 = -0.0067$. In a similar way the effects for the other contrast can be found. For both contrasts there is a random intercept with standard deviations of 1.52 and 2.30, respectively. The

Table 2: Parameter estimates of the mixed effects MBCL model for the MHRP data. In the C/S column the parameter estimates for the community versus street contrast are given, whereas in the I/S column estimates for the contrast independent versus street are given. The correlation between the two random intercepts equals 0.692.

Effect	C/S	SE	I/S	SE
Constant	2.4358	0.3007	1.2932	0.3627
Time	0.1599	0.0187	0.2363	0.0240
Time Squared	-0.0144	0.0023	-0.0154	0.0027
Incentive	-0.9138	0.4199	2.0720	0.4854
Incentive × Time	-0.0862	0.0247	0.0216	0.0308
Incentive × Time squared	0.0077	0.0032	-0.0075	0.0037
Standard deviation	1.5190	0.1992	2.2993	0.2251

correlation between the random intercepts equals 0.692.

1.5 Problems with mixed effects MBCL models

Several problems with the mixed effects MBCL models are noted in the literature:

1. These models may become computationally very intensive when there are two or more random effects, and computationally unfeasible when there are more than five or six random effects (Hartzel, Agresti and Caffo, 2001). The reason is that the integrals appearing in the likelihood function must be solved using approximation methods, such as linearization methods, numerical integration methods, or simulation methods. For the baseline category logit random effects model with only random intercepts, the integral dimension equals the number of categories of the response variable minus one. If the number of classes is larger than six these quadrature methods are computationally unfeasible. Moreover, if in addition to random intercepts, random slopes are envisaged, the number of categories should not exceed three.

2. These models rely on the untestable assumption that random coefficients come from a multivariate normal distribution. Results may be biased when this assumption is violated (Vermunt, 2007; Aitkin, 1999).

3. It is not at all straightforward to interpret the parameters associated with the random effects (Vermunt, 2007).

4. The interpretation of regression coefficients is not simple, especially in cases with interactions and/or higher order treatment of variables. The interpretation is further complicated because the coefficients refer to contrasts of categories of the response variable with a baseline category (Fox and Andersen, 2006).

In the literature several ways have been proposed to deal with these problems, either simultaneously or separately. To deal with the first problem Bock (1972), Skrondal and Rabe-Hesketh (2003), and Takane (1996) use dimension reduction techniques, whereby the first two use an inner product parametrization and Takane uses a distance parametrization. Another approach to simultaneously deal with problems 1, 2, and 4 is to use categorical latent variables as in latent Markov models with or without explanatory variables (Langeheine and Van der Pol, 1990, 1994; Vermunt, Langeheine, and Böckenholt, 1999) or latent class growth curve models (Vermunt, 2007). To deal with problem 4, Fox and Andersen (2006) propose effect displays.

In the next Section we propose the mixed effects trend vector model to deal with problems 1, 3, and 4. This model is a generalization of the the ideal point classification model (IPC model, De Rooij (2009a)), a probabilistic multidimensional unfolding model for nominal response variables. This is a generalization of the approach of Takane (1996) in the sense that he used random intercepts only and not random slopes or fixed effect parameters.

2 The mixed effects trend vector model

The trend vector model is an extension of the IPC model (De Rooij, 2009a,b) for longitudinal data. The IPC model is a probabilistic multidimensional unfolding model where the probability of belonging to a category is a monotonically decreasing function of the relative squared Euclidean distance between the position of a subject and that of a category, compared to the squared distances towards the other categories. The position of a subject at a specific time point is given by a linear combination of the explanatory variables. In the unfolding literature, these variables are sometimes called auxiliary, supplementary, or concomitant variables. In line with the generalized linear model literature we will continue to call them explanatory variables. When time is one of the main explanatory variables the model is coined the trend vector model. In De Rooij

(2009b) the trend vector model was first proposed. This model was a marginal model modeling the population averaged trend over time and treating the dependencies among the observations as a nuisance. Yu and De Rooij (2009) discuss model selection issues for this marginal model. For a comparison of marginal and subject specific models see Diggle, Heagerty, Liang, and Zeger (2002) and Molenberghs and Verbeke (2005). In this paper we will extend the trend vector model with random effects in order to obtain a subject specific model, that is a model where each subject has its own trend over time. This is an important generalization because a random effects model provides insight into how specific individuals change across time. Moreover, the inclusion of random effects accounts for the influence of participants on their repeated observations. The random effects thus describe each person's trend over time and explain the association structure of the longitudinal data. Additionally, they indicate the degree of subject variation that exists in the population of participants (Hedeker and Gibbons, 2006).

2.1 Formal model definition

Like we did for the mixed effects MBCL model, here we also assume that the conditional distribution of $\mathbf{g}_{it}$ given the subject specific parameters $\mathbf{u}_i$, $f(\mathbf{g}_{it}|\mathbf{u}_i)$, is the multinomial with expectation

$$E(\mathbf{g}_{it}|\mathbf{u}_i) = \boldsymbol{\pi}_{it} = [\pi_{it1}, \ldots, \pi_{itC}]^{\mathsf{T}}.$$

The probabilities are related to squared Euclidean distances ($\boldsymbol{\delta}_{it}$) by the vector of link functions $\mathbf{h}_d(\cdot)$, i.e.

$$\boldsymbol{\pi}_{it} = \mathbf{h}_d(\boldsymbol{\delta}_{it}),$$

with

$$\boldsymbol{\delta}_{it} = [\delta_{it1}, \ldots, \delta_{itC}]^{\mathsf{T}},$$

and $\mathbf{h}_d(\cdot) = [h_{d1}(\cdot), \ldots, h_{dC}(\cdot)]$, where $h_{dc}(\cdot)$ is the Gaussian decay function

$$h_{dc}(\delta_{it1}, \ldots, \delta_{itC}) = \frac{\exp(-\delta_{itc})}{\sum_h \exp(-\delta_{ith})}. \tag{1}$$

The squared distance δ_{itc} is a so-called unfolding or two-mode distance in M-dimensional Euclidean space between a position (known as ideal point in the multidimensional unfolding literature) for subject i at time point t with coordinates η_{itm} $(m = 1, \ldots, M)$ and a point for category c with coordinates γ_{cm}; more specifically

$$\delta_{itc} = \sum_{m=1}^{M} (\eta_{itm} - \gamma_{cm})^2. \tag{2}$$

The coordinates for the position of subject i at time point t on the m-th dimension are given by the linear predictor

$$\eta_{itm} = \alpha_m + \mathbf{x}_{it}^{\mathsf{T}} \boldsymbol{\beta}_m + \mathbf{z}_{it}^{\mathsf{T}} \mathbf{u}_{im}. \tag{3}$$

We assume a multivariate normal distribution for the random effects, that is

$$\mathbf{u}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}).$$

2.2 Model interpretation

For interpretation of the model graphical displays are used like those in multidimensional unfolding. In such a display every individual at every time point has a position (in the unfolding literature this is often called an ideal point). The relative squared distance from this position towards the class points represents the probability for that class, as can be seen from the model formulae (Equations 1 and 2). The odds that a specific subject i at time point t chooses category a instead of b are given by $\exp(\delta_{itb} - \delta_{ita})$: the odds are in favor of the closest category. To make interpretation easier, we make prediction regions in the Euclidean space. For every class this prediction region represents that part of the space where the probability of the corresponding class is highest. The class point itself is within the prediction region. On the boundaries, the (log) odds for two classes are even.

The coordinates of the position of a subject at a given time point are a linear combination of the explanatory variables ($\alpha_m + \mathbf{x}_{it}^{\mathsf{T}} \boldsymbol{\beta}_m$) plus a random part ($\mathbf{z}_{it}^{\mathsf{T}} \mathbf{u}_{im}$). In the examples discussed in the Introduction there are multiple groups that may follow different trajectories over time. The mean trajectory of a group is given by the fixed part of the linear model, whereas the departure of individuals from their group mean is represented by the random effects.

We will draw the group mean trajectories using trend vectors. The procedure will be exemplified using the housing example of the introduction. The coordinates on dimension m are given by the following linear predictor

$$\eta_{itm} = \alpha_m + \beta_{1m} G_i + \beta_{2m} T_{it} + \beta_{3m} T_{it}^2 + \beta_{4m} G_i T_{it} + \beta_{5m} G_i T_{it}^2 + u_{im}.$$

To draw the trajectory for the control group, with $G_i = 0$, we compute the vectors $\boldsymbol{\eta}_m$ as

$$\boldsymbol{\eta}_m = \hat{\alpha}_m + \mathbf{t}\hat{\beta}_{2m} + \mathbf{t}^2\hat{\beta}_{3m}$$

where $\mathbf{t}$ is a vector with values within the observed time range, i.e. $\mathbf{t} = [0.0, 0.1, 0.2, \ldots, 6.0, \ldots, 12.0, \ldots, 24.0]^\mathsf{T}$ and $\mathbf{t}^2$ are the element wise squares of the values of $\mathbf{t}$. The points defined by the coordinates on dimension 1 ($\boldsymbol{\eta}_\mathbf{1}$) and dimension 2 ($\boldsymbol{\eta}_\mathbf{2}$) define a smooth trend vector for the control group. Similarly, for the treatment group for which $G_i = 1$

$$\boldsymbol{\eta}_m = (\hat{\alpha}_m + \hat{\beta}_{1m}) + \mathbf{t}(\hat{\beta}_{2m} + \hat{\beta}_{4m}) + \mathbf{t}^2(\hat{\beta}_{3m} + \hat{\beta}_{5m})$$

defines a trend vector for the treatment group. It is important that the vector $\mathbf{t}$ only has values within the observed time frame in order to prevent extrapolation. Every point on the trend vectors represents a time point. For the observed time points (in the MHRP data 0, 6, 12, and 24) we add markers to the trend vectors. For each marker point we can look at the distances towards the category points. The closest category point has the highest probability for an average person in the group corresponding to the trend vector at that specific time point. The farthest point has the lowest probability. Each subject has his or her own personal deviation from their group trajectory defined by the random effects.

To represent the random effects we will use ellipses. We present ellipses that include 68% of the observations, that is our ellipses are defined using major axis equal to the standard deviations of the random effects. For the models with only random intercepts the random effects represents changes in starting position, that is every trajectory has the same shape. For models including random slopes the shape of the trend vectors will be different for every subject. In the latter case it can be advantageous to represent all individual trend vectors, provided this does not clutter the graphical display.

The solutions of the mixed effects trend vector models can be represented in low dimensional displays. In applications where the categories of the response variable have an intrinsic meaning or when there are many explanatory variables the dimensions can be given an interpretation. An example of the first can be found in Spinhoven *et.al* (2011) where the response variable consists of nine categories that jointly represent a two dimensional solution. Here, one dimension can be interpreted as a fear dimension while the second has a distress interpretation. In cases with many explanatory variables, these variables can give a substantive interpretation to the dimensions. In many applications, such as the ones presented in the introduction, there is not enough information to provide an interpretation of the dimensions. We will refrain from interpreting the dimensions in this paper. Nevertheless, the graphical representation can be interpreted by looking at the distances.

2.3 Nominal vs ordinal response variables

Trend vector models can be fitted in one, two, or higher dimensional spaces. In a one dimensional solution the categories of the response variable lie on a single continuum in a specific order. If the one dimensional solution is optimal in the sense that it gives the best goodness-of-fit statistics, the response variable is ordinal with respect to the explanatory variables. Nominal response variables usually require two or higher dimensional solutions.

For an ordinal response variable (like the one in the third example of the introduction), the ordinality with respect to the explanatory variables can be investigated using the trend vector models. The model does not determine the order by itself. The ordinality can be imposed on, but can also be inferred from the data. Different forms of ordinality may be assumed:

- Interval: the categories are equally spaced, i.e. $\gamma_c = \gamma_1 + (c-1) \times \tau$, with $\tau > 0$;
- Ordinal: the categories are ordered, i.e. $\gamma_c = \gamma_1 + \sum_{s=2}^{c} \tau_s$, with $\tau_s > 0$;
- Unconstrained: the categories are ordered but we do not know the ordering, i.e. unconstrained estimation of the γ_c's.

In each of these three cases the response variable is ordinal with respect to the explanatory variables. For an assumed ordinal variable the relationship with the explanatory variables does not need to take an ordinal form (just as the relationship between two continuous variables need not be linear). An advantage of the trend vector model is that the ordinality assumption can be investigated by extending the model to a two (or higher) dimensional form. In Section 5.3 we show such an investigation.

3 Fitting the mixed effects trend vector model

In this section we discuss estimation of the mixed effects trend vector model. This can be done using either maximum likelihood or Bayesian estimation. In Bayesian estimation prior distributions have to be formulated for all model parameters. With the likelihood and the priors a posterior distribution can be formulated from which samples can be drawn, for example using the Gibbs sampler. We will not use this approach here, but focus on maximum likelihood estimation in order to find estimates of the model parameters α_m, β_{jm}, γ_{cm}, and $\mathbf{\Sigma}$.

3.1 The likelihood function

Define $\mathbf{g}_i$ to be the vector of length $n_i \times C$ with the measurements for subject i, that is $\mathbf{g}_i = [\mathbf{g}_{i1}^\mathsf{T}, \ldots, \mathbf{g}_{in_i}^\mathsf{T}]^\mathsf{T}$. Let $f(\mathbf{g}_i|\mathbf{u}_i; \boldsymbol{\beta}_m, \boldsymbol{\gamma}_m)$ denote the conditional mass function of $\mathbf{g}_i$ given the random effects $\mathbf{u}_i$. Furthermore, let $f(\mathbf{u}_i; \mathbf{\Sigma})$ denote the normal density function for the random effects. The likelihood function is the probability mass function viewed as a function of α_m, $\boldsymbol{\beta}_m$, $\boldsymbol{\gamma}_m$, and $\mathbf{\Sigma}$, i.e. the likelihood function is

$$L = \prod_i \int \cdots \int f(\mathbf{g}_i|\mathbf{u}_i; \boldsymbol{\beta}_m, \boldsymbol{\gamma}_m) f(\mathbf{u}_i; \mathbf{\Sigma}) d\mathbf{u}_i. \tag{4}$$

It is assumed that, conditional on the random effects, the responses of a subject are independent multinomial distributed, so that

$$f(\mathbf{g}_i|\mathbf{u}_i; \boldsymbol{\beta}_m, \boldsymbol{\gamma}_m) = \prod_{t=1}^{n_i} \prod_{c=1}^{C} \pi_{itc}^{g_{itc}}.$$

3.2 Approximating the likelihood function

In general no analytical solutions are available for the integrals in (4), hence numerical approximations are needed. We can either approximate the integrand, the data, or the integral. For an overview of these three methods see Tuerlinckx *et. al* (2006) or Molenberghs and Verbeke (2005). The integrand may be approximated using the Laplace method, while the data may be approximated using a Penalized (PQL) or Marginal Quasi Likelihood (MQL) approach. Molenberghs and Verbeke (2005) show that these latter two perform poorly with binary response data. For multinomial data we do not expect the performance of PQL and MQL to be better.

When the integral is approximated this can be done *directly* using (adaptive) quadrature methods or Monte Carlo integration or *indirectly* using the EM-algorithm (see Tuerlinckx *et. al*, 2006), where in the E-step again quadrature or Monte Carlo integration is needed. We take the direct approach using Gauss-Hermite quadrature. In this method the integral is replaced by a weighted summation over a set of nodes. The more nodes used, the better the approximation of the likelihood, but the slower the algorithm. For details see again Tuerlinckx *et. al* (2006), Molenberghs and Verbeke (2005), or Hartzel, Agresti and Caffo (2001). The approximated likelihood is maximized using a quasi-Newton algorithm.

3.3 Indeterminacies

The trend vector model, as presented so far, is not identified. First of all, as in any distance model, there is a translation and a rotation problem. A third indeterminacy, due to the Gaussian decay function, is that a constant might be added for each subject's squared distance without changing the probabilities, that is

$$h_c(\delta_{it1}, \ldots, \delta_{itC}) = \frac{\exp(-\delta_{itc})}{\sum_h \exp(-\delta_{ith})} = \frac{\exp(-\delta_{itc} + s_{it})}{\sum_h \exp(-\delta_{ith} + s_{it})},$$

which is a type of scaling problem. These three indeterminacies can be solved by fixing class point coordinates. First, to solve the translational indeterminacy we set $\gamma_{1m} = 0$. The rotational indeterminacy is solved by setting $\gamma_{m,m+1} = 0$ and the scaling indeterminacy by setting $\gamma_{m+1,m} = 1$. For two and three dimensional solutions the matrix with class coordinates then have the following forms

$$2D: \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ \gamma_{31} & 1 \\ \gamma_{41} & \gamma_{42} \\ \vdots & \vdots \\ \gamma_{C1} & \gamma_{C2} \end{bmatrix}, \quad 3D: \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ \gamma_{31} & 1 & 0 \\ \gamma_{41} & \gamma_{42} & 1 \\ \vdots & \vdots & \vdots \\ \gamma_{C1} & \gamma_{C2} & \gamma_{C3} \end{bmatrix}.$$

When $C = 3$ and $M = 2$ we need one more constraint, i.e. $\gamma_{31} = 0$. For the case $C = 4$ and $M = 3$ we need three more constraints: $\gamma_{31} = \gamma_{41} = \gamma_{42} = 0$. Both are full dimensional solutions in which no class point coordinates have to be estimated.

There are other sets of possible identification constraints. Instead of setting a single category coordinate per dimension equal to zero we could fix the mean of the coordinates as equal to zero. Furthermore we could constrain the variance of the coordinates to equal 1, and the coordinates of any dimension to be uncorrelated with those of another dimension. Such sets of constraints are more difficult to incorporate.

3.4 Empirical Bayes estimation of random effects

Prediction of the random effects can be done using *expected a posteriori* (EAP) or empirical Bayes estimation, where the expectation of the random effect value for subject i is given by

$$E(\mathbf{u}_i|\mathbf{g}_i, \hat{\boldsymbol{\beta}}_m, \hat{\boldsymbol{\gamma}}_m, \hat{\boldsymbol{\Sigma}}) = \frac{\int \cdots \int \mathbf{u}_i f(\mathbf{g}_i|\mathbf{u}_i; \hat{\boldsymbol{\beta}}_m, \hat{\boldsymbol{\gamma}}_m) f(\mathbf{u}_i; \hat{\boldsymbol{\Sigma}}) d\mathbf{u}_i}{\int \cdots \int f(\mathbf{g}_i|\mathbf{u}_i; \hat{\boldsymbol{\beta}}_m, \hat{\boldsymbol{\gamma}}_m) f(\mathbf{u}_i; \hat{\boldsymbol{\Sigma}}) d\mathbf{u}_i}.$$

Like before, the integrals can be approximated using Gauss-Hermite quadrature methods.

3.5 Software

All models as described above can be estimated using the `NLMIXED` procedure in `SAS`. In the appendix annotated code is presented for the models shown in Section 5.

4 Comparison with other models

In this section we compare the trend vector model (or more generally the IPC model) to some well known logit models. The focus is on the IPC model, because in the comparison between models the time variable does not play an important role.

4.1 MBCL model

For this first comparison we leave out the random effects for the moment, and make a small change to the definition of the IPC model

$$h_{dc}(\delta_{it1}, \ldots, \delta_{itC}) = \frac{\exp(-\frac{1}{2}\delta_{itc})}{\sum_h \exp(-\frac{1}{2}\delta_{ith})}.$$

which does not change the model fit nor display. In maximum dimensionality, i.e. $M = C - 1$ the IPC model is equivalent to the MBCL model. Using the identification constraints as proposed in Section 3.3 we have that $\alpha^I_{0m} - \frac{1}{2} = \alpha^B_{0c}$ and $\beta^I_m = \beta^M_c$, where the superscript I identifies the parameters from the IPC model and the superscript B that of the MBCL model. The IPC model can thus be used as a visualization model for the MBCL model.

The same reasoning is true when both models are augmented with random effects. The identification constraints are placed on the class points and the random effects IPC model can be used as a visualization tool for the random effects MBCL model. Such a visualization solves problems 3 and 4 as discussed in the introduction, and is shown in Section 5.1.

4.2 Proportional odds model

An important model for ordinal data is the proportional odds logit model (McCullagh, 1980). In the proportional odds logit model (POM) the probabilities are related to a linear predictor by the vector of link functions $\mathbf{h}_p(\cdot)$, i.e.

$$\boldsymbol{\pi}_{it} = \mathbf{h}_p(\boldsymbol{\eta}_{it}),$$

and $\mathbf{h}_p(\cdot) = [h_{p1}(\cdot), \ldots, h_{lp}(\cdot)]$, where $h_{pc}(\cdot)$ is

$$\begin{aligned} h_{p1}(\eta_{it1}, \ldots, \eta_{itC}) &= \frac{\exp(\eta_{it1})}{1+\exp(\eta_{it1})} \\ h_{pc}(\eta_{it1}, \ldots, \eta_{itC}) &= \frac{\exp(\eta_{itc})}{1+\exp(\eta_{itc})} - \frac{\exp(\eta_{it(c-1)})}{1+\exp(\eta_{it(c-1)})}, c = 2, \ldots, C. \end{aligned}$$

In this model the linear predictor is defined as

$$\eta_{itc} = \alpha_c + \mathbf{x}_{it}^{\mathsf{T}}\boldsymbol{\beta} + \mathbf{z}_{it}^{\mathsf{T}}\mathbf{u}_i.$$

An important feature of the fixed effects (i.e. without random effects) proportional odds logistic model is the simple form of the cumulative odds ratio. That is, the odds of making response $\leq c$ at $\mathbf{x}_{it} = \mathbf{x}_1$ are $\exp((\mathbf{x}_1 - \mathbf{x}_2)^{\mathsf{T}}\boldsymbol{\beta})$ times the odds at $\mathbf{x}_{it} = \mathbf{x}_2$, or in other words, proportional to the signed distance between $\mathbf{x}_1$ and $\mathbf{x}_2$ and independent of c.

For our fixed effects IPC model the odds (*not* the cumulative odds) for choosing category a over b at $\mathbf{x}_{it} = \mathbf{x}_1$ are

$$\exp\left\{2(\mathbf{x}_1 - \mathbf{x}_2)^{\mathsf{T}}\boldsymbol{\beta}(\gamma_a - \gamma_b)\right\}$$

times the odds at $\mathbf{x}_{it} = \mathbf{x}_2$, i.e. the odds ratio depends on the signed distance between $\mathbf{x}_1$ and $\mathbf{x}_2$ and the signed distance between γ_a and γ_b. If the signed distances between consecutive categories are all equal (the interval case in Section 2.3), the odds ratio only depends on the distance between $\mathbf{x}_1$ and $\mathbf{x}_2$. If the response categories are not equally spaced the distance between two category points co-determine this odds ratio.

The ideal point parametrization has some interpretational advantages over the proportional odds model. For instance, the location at which the odds changes from being in favor of one response category to another is exactly in the middle of the two category locations. Each category has at least a single position where it is most probable, unless two (or more) category coordinates are equal. Furthermore, the ideal point model is a model for the probabilities of each outcome, whereas the proportional odds model is a model for cumulative probabilities; the former is easier to comprehend.

The proportional odds model is sometimes thought to be too restrictive, i.e. the proportional odds assumption might not be valid for some variables. To deal with such a situation the partial proportional odds model was proposed by Peterson and Harrell (1990). A serious disadvantage of this model is that it can lead to negative values of the response probabilities (Hedeker and Gibbons, 2006, p.192). As shown above the ideal point model is not as strict as the proportional odds logistic model, but still the ordinality assumption might be too coarse for some explanatory variables. Our model can also be fitted in two (or more) dimensions to investigate disordinalty.

4.3 Stereotype model

Another model for ordinal data which can be generalized to nominal data is proposed by Anderson (1984) under the name stereotype model. Johnson (2007) recently extended the stereotype model to a random effects model. In the stereotype model the probabilities are related to a linear predictor by the same vector of link functions as used in the baseline category logit model ($\mathbf{h}_l(\cdot)$), i.e.

$$\boldsymbol{\pi}_{it} = \mathbf{h}_l(\boldsymbol{\eta}_{it}),$$

and $\mathbf{h}_l(\cdot) = [h_{l1}(\cdot), \ldots, h_{lC}(\cdot)]$, where $h_{lc}(\cdot)$ is

$$h_{lc}(\eta_{it1}, \ldots, \eta_{itC}) = \frac{\exp(\eta_{itc})}{\sum_h \exp(\eta_{ith})}.$$

The c-th linear predictor is given by (see Johnson, 2007)

$$\begin{aligned}
\eta_{itc} &= \alpha_c - \mathbf{x}_{it}^{\mathsf{T}}\boldsymbol{\beta}\phi_c + \mathbf{z}_{it}^{\mathsf{T}}\mathbf{u}_i, \text{ or} \\
\eta_{itc} &= \alpha_c - \mathbf{x}_{it}^{\mathsf{T}}\boldsymbol{\beta}\phi_c + \mathbf{z}_{it}^{\mathsf{T}}\mathbf{u}_{ic}.
\end{aligned}$$

Under this model the regression weights $\boldsymbol{\beta}_c = -\boldsymbol{\beta}\phi_c$ are parallel. To achieve an ordinal model the ϕ_c should be ordered, i.e. $1 = \phi_1 > \phi_2 > \ldots > \phi_C$.

Our one dimensional trend vector model can be rewritten as follows

$$\begin{aligned}
\pi_{itc} &= \frac{\exp\left(-(\eta_{it}-\gamma_c)^2\right)}{\sum_h \exp\left(-(\eta_{it}-\gamma_h)^2\right)} \\
&= \frac{\exp\left(-\eta_{it}^2-\gamma_c^2+2\eta_{it}\gamma_c\right)}{\sum_h \exp\left(-\eta_{it}^2-\gamma_h^2+2\eta_{it}\gamma_h\right)} \\
&= \frac{\exp\left(-\gamma_c^2+2\eta_{it}\gamma_c\right)}{\sum_h \exp\left(-\gamma_h^2+2\eta_{it}\gamma_h\right)} \\
&= \frac{\exp\left(-\gamma_c^2+2(\alpha+\mathbf{x}_{it}^{\mathsf{T}}\boldsymbol{\beta}+\mathbf{z}_{it}\mathbf{u}_i)\gamma_c\right)}{\sum_h \exp\left(-\gamma_h^2+2(\alpha+\mathbf{x}_{it}^{\mathsf{T}}\boldsymbol{\beta}+\mathbf{z}_{it}\mathbf{u}_i)\gamma_h\right)} \\
&= \frac{\exp\left(-\gamma_c^2+2\alpha\gamma_c+2\mathbf{x}_{it}^{\mathsf{T}}\boldsymbol{\beta}\gamma_c+2\mathbf{z}_{it}\mathbf{u}_i\gamma_c\right)}{\sum_h \exp\left(-\gamma_h^2+2\alpha\gamma_h+2\mathbf{x}_{it}^{\mathsf{T}}\boldsymbol{\beta}\gamma_h+2\mathbf{z}_{it}\mathbf{u}_i\gamma_h\right)} \\
&= \frac{\exp\left(\alpha_c^*+2\mathbf{x}_{it}^{\mathsf{T}}\boldsymbol{\beta}\gamma_c+2\mathbf{z}_{it}\mathbf{u}_i\gamma_c\right)}{\sum_h \exp\left(\alpha_h^*+2\mathbf{x}_{it}^{\mathsf{T}}\boldsymbol{\beta}\gamma_h+2\mathbf{z}_{it}\mathbf{u}_i\gamma_h\right)},
\end{aligned}$$

where $\alpha_c^* = -\gamma_c^2 + 2\alpha\gamma_c$. The last equation shows that the trend vector model is closely related to the stereotype model.

5 Empirical Examples

We analyzed the three data sets described in the introduction. The first (MHRP data) will be used as an example of the trend vector model as a visualization tool for the mixed effects MBCL model. The second example, on the TV watching preferences, shows that for this data we can fit mixed effects trend vector models. This would be impossible with the mixed effects MBCL model due to the number of random effects. The final example shows that we can test the ordinality assumption for ordinal response variables.

To compare models several fit indices will be given. First the deviance is presented (-2 times the Log likelihood) together with the number of fitted parameters (npar). Furthermore information criteria are used to compare models. The BIC statistic defined by

$$\text{BIC} = -2 \times LL + \log(n) \times \text{npar},$$

where n is the number of subjects and LL the log likelihood. The AIC statistic

$$\text{AIC} = -2 \times LL + 2 \times \text{npar},$$

and the AICC which is the AIC with a second order correction for small sample sizes

$$\text{AICC} = \text{AIC} + \frac{2k(k+1)}{n-k-1},$$

where $k = \text{npar} + 1$. Note that, because we set up the data in ungrouped format (see Agresti, 2002, pp. 174-175) we cannot use the deviance as a model test. Moreover, for such a set-up there is no sensible saturated model.

5.1 Housing example: graphical representation of MBCL model

Figure 1 gives the solution corresponding to the analysis shown in the introduction. The trend vector for both the incentive and the control group is presented. Small circles represent the time points 0, 6, and 12 months and the vector ends at 24 months. These vectors represent the fixed effects of the model and can be interpreted as follows. The starting point of the two groups and the quadratic time trend both differ. The starting point for the incentive group is already somewhat favorable. Furthermore, we see that the incentive group quickly moves into the Independent housing region, whereas the control group moves into the community house region and only at the end of the study gets closer to the independent housing region. The joint influence of time and time squared for both groups can be easily understood from the graph, whereas it is quite difficult to obtain a thorough understanding from the parameter estimates themselves (see Table 2).

The ellipse represents a 68% region for the random intercepts. The very large ellipse means that there are large differences in individual starting points. The trends are, however, homogeneous for all participants (no random effect of time). These differences in starting point dominate the solution in the sense that the differences in starting point are much larger than the trends over time. More specifically, for someone who has his/her starting point at the lower right of the space the trend vector follows a trajectory within the community housing region, while for someone in the upper left corner of the space the trajectory will be within the independent housing region. This large variation in the intercepts also means that there is a large positive association among the responses of every individual, i.e. participants tend to be in the same category at consecutive time points.

Both the nature of development over time as well as the influence of the random

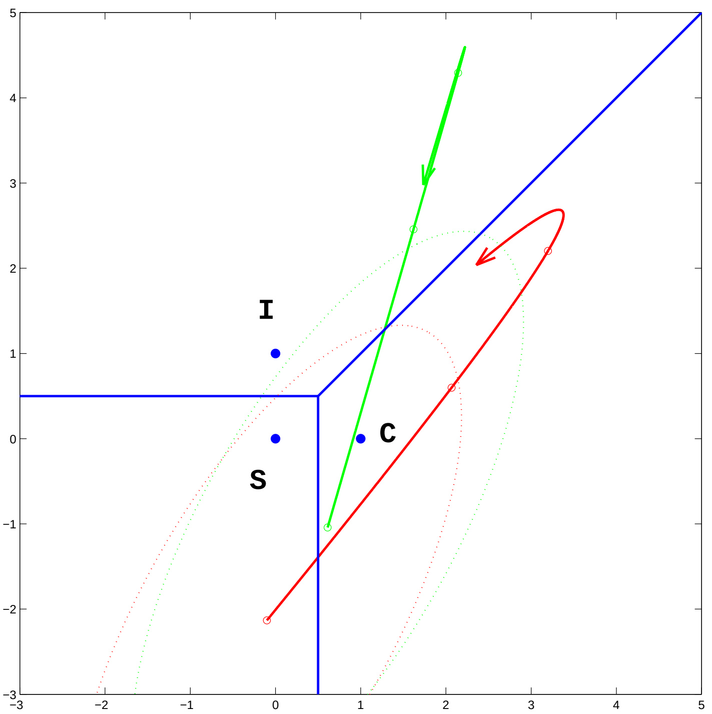

Figure 1: *Visualization of the random intercept quadratic trend MBCL model for the McKinney Homeless Research Project. The two trend vectors represent the trajectories for the incentive and control group. The dots on the trend vectors represent* $T = 0, 6, 12$*, and the trend vector ends at* $T = 24$*. The ellipses represents approximate 68% regions of the random intercepts, i.e. variation in starting points. I stands for Independent, C for Community housing, and S for Street.*

intercepts are clear from Figure 1 but cannot easily be grasped form the parameter estimates in Table 2.

5.2 TV watching example: Individual differences in change patterns

Here the results of the analyses of the TV preference data from Adachi (2000) are given. We constructed a time variable T by using the midpoints of the ages at the specific time points as scores (i.e. 6.5 for the first time point, 9.5 for the second, etc) and centered around the mean (12.25). The question is whether boys and girls (G) differ in their trends in watching behavior and what the trends look like.

For the random intercept and slope model we will distinguish between three structures for the covariance matrix of the random effects: The full structure (f), the dimension wise structure (d) and the intermediate (i) structure. In the dimension wise structure the random effects of the first dimension are assumed to be uncorrelated with those of the second dimension. For the intermediate structure we assume zero correlation between the random intercept of dimension 1 (2) and the random slope of dimension 2 (1). The structures are defined as

$$\mathbf{\Sigma}_f = \begin{bmatrix} \sigma_1^2 & & & \\ \sigma_{21} & \sigma_2^2 & & \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}, \quad \mathbf{\Sigma}_d = \begin{bmatrix} \sigma_1^2 & & & \\ \sigma_{21} & \sigma_2^2 & & \\ 0 & 0 & \sigma_3^2 & \\ 0 & 0 & \sigma_{43} & \sigma_4^2 \end{bmatrix}, \quad \mathbf{\Sigma}_i = \begin{bmatrix} \sigma_1^2 & & & \\ \sigma_{21} & \sigma_2^2 & & \\ \sigma_{31} & 0 & \sigma_3^2 & \\ 0 & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix},$$

where σ_1^2 represents the random intercept of the first dimension, σ_2^2 the random slope of the first dimension, σ_3^2 the random intercept of the second dimension, and σ_4^2 the random slope of the second dimension. In Table 3 fit statistics are given for a series of models in two dimensions.

The various model fit indices indicate different best models. The BIC points towards the model with linear predictor

$$\eta_{itm} = \alpha_m + \beta_{1m}G + \beta_{2m}T_{it} + u_{0i.m} + u_{1i.m}T_{it},$$

i.e. a random intercepts and slopes model with fixed gender and homogeneous linear time effect (i.e. no interaction). The AIC selects a more complex model with the linear predictor given by

$$\eta_{itm} = \alpha_m + \beta_{1m}G + \beta_{2m}T_{it} + \beta_{3m}T_{it}^2 + \beta_{4m}GT_{it} + \beta_{5m}GT_{it}^2 + u_{0i.m} + u_{1i.m}T_{it},$$

a model with different quadratic trends for boys and girls plus random intercepts and slopes.

To compare these two models we computed residuals to verify the fit of the two models. Therefore, we first define the Pearson residuals

$$\chi_{it}^2 = \frac{\left(1 - \pi_{it(1)}(\mathbf{x}_{it})\right)}{\pi_{it(1)}(\mathbf{x}_{it})}$$

where $\pi_{it(1)}(\mathbf{x}_{it})$ denotes the probability for the actual chosen category, and the deviance

Table 3: *Fit statistics for Adachi TV data. I represents random intercepts model, while I+S represents the random intercepts and slopes model. G is the gender variable and T the time variable. T|G represents an interaction between the time and gender variable plus their main effects. The number of fitted parameters is given under npar, -2LL is the deviance, BIC the Bayesian Information Criterium, AIC is Akaike's Information Criterion, and AICC is the small sample bias corrected version of AIC.*

Random	Structure	Fixed	npar	-2LL	BIC	AIC	AICC
I	Independence	$G+T$	15	1244.0	1313.1	1274.0	1275.4
	Correlated	$G+T$	16	1244.0	1317.7	1276.0	1277.5
	Independence	$G+T+T^2$	17	1242.5	1320.8	1276.5	1278.2
	Correlated	$G+T+T^2$	18	1229.6	1312.5	1265.1	1267.5
	Independence	$G+T\|G$	17	1241.8	1320.1	1275.8	1277.5
	Correlated	$G+T\|G$	18	1240.4	1323.3	1276.4	1278.3
	Independence	$G+T\|G+T^2\|G$	21	1233.4	1330.1	1275.4	1278.0
	Correlated	$G+T\|G+T^2\|G$	22	1220.4	1321.7	1264.4	1276.2
I+S	Dim. wise	$G+T$	19	1224.1	1311.6	1262.1	1264.2
	Intermediate	$G+T$	21	1212.7	**1309.4**	1254.7	1257.2
	Full†	$G+T$	23	1206.1	1312.0	1252.1	1255.2
	Dim. wise	$G+T+T^2$	21	1215.4	1312.1	1257.4	1260.0
	Intermediate	$G+T+T^2$	23	1207.7	1313.7	1253.7	1256.8
	Full	$G+T+T^2$	25	1205.8	1320.9	1255.8	1259.5
	Dim. wise	$G+T\|G$	21	1222.2	1318.9	1264.2	1266.8
	Intermediate	$G+T\|G$	23	1212.5	1318.4	1258.5	1261.6
	Full	$G+T\|G$	25	1204.8	1319.9	1254.8	1258.5
	Dim. wise	$G+T\|G+T^2\|G$	25	1201.6	1316.8	**1251.6**	1255.3
	Intermediate	$G+T\|G+T^2\|G$	27	1201.5	1325.9	1255.5	1259.8
	Full	$G+T\|G+T^2\|G$	29	1199.5	1333.1	1257.5	1262.5

† SAS gives the following warning: The final hessian matrix is full rank but has at least one negative eigenvalue. Second order optimality condition violated.

residuals $d_{it}^2 = 2\log(1+\chi_{it}^2)$ (Lesaffre and Albert, 1989). In Figure 2 we show the normal probability plots for the two models. On the left hand plot the deviance residuals are shown for the model with homogeneous linear trends; in the middle plot they are shown for the model with heterogeneous quadratic trends. On the right hand side a quantile-quantile plot is shown with the deviance residuals of the first against the second model.

Both models do not seem to fit well, moreover, the more complex model does not have a better fit. There are probably important missing covariates that can explain TV watching in more detail. The high residuals are for those participants with changes forth and back between two categories. For example, a participant with observed pattern D,M,S,V,S has a very high residual for the fourth observation (V), and another partici-

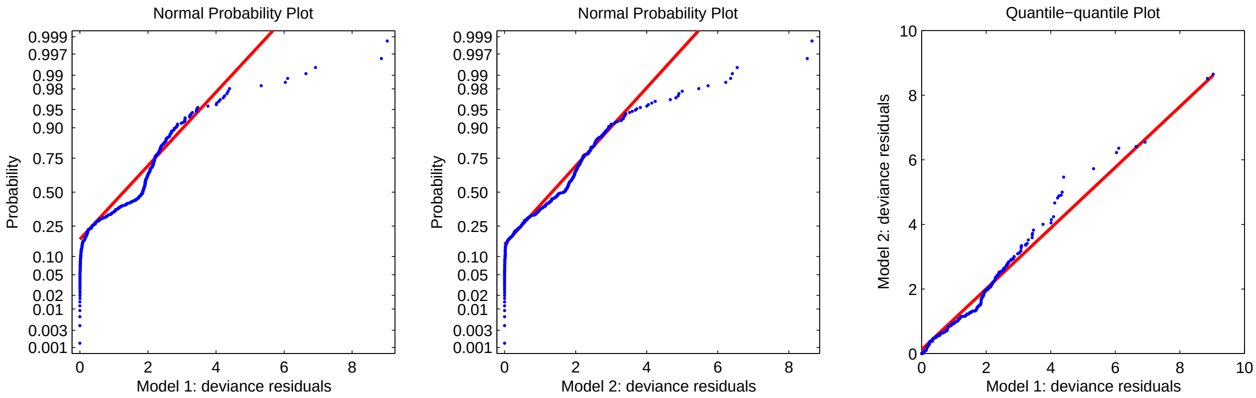

Figure 2: *Normal probability plots with deviance residuals. On the left hand plot the deviance residuals are shown for the model with homogeneous linear trends; in the middle plot they are shown for the model with heterogeneous quadratic trends. On the right hand side a quantile-quantile plot is shown with the deviance residuals of the first (left) against the second model (middle).*

pant with pattern S,C,S,V,S also has a very high residual for the fourth observation (V). Participants with these type of changes are very hard to model, not only using the trend vector model but also in MBCL models; they require un-smooth or irregular trends.

Because both models fit about equally well, and because the sample size is small it is wise to have a parsimonious model. In Figure 3 we show the first model, the one favored by the BIC. In this figure the mean trend for girls starts at Animation and then goes into the Variety area, while the mean trend for boys also starts in Animation but then passes Drama, Music, and ends at the border of Music and Sports.

In addition to the mean trends for boys and girls the individual trends are shown. These can be obtained by computing the EAP estimates of the random effects for every subject. The individual trends show a large variation although they all start in or near Animation. For this solution the correlation between the random intercepts equals 0.86 and the correlation between the slopes 0.66. The intercept slope correlation on the first axis equals -0.43 and for the second axis 0.30.

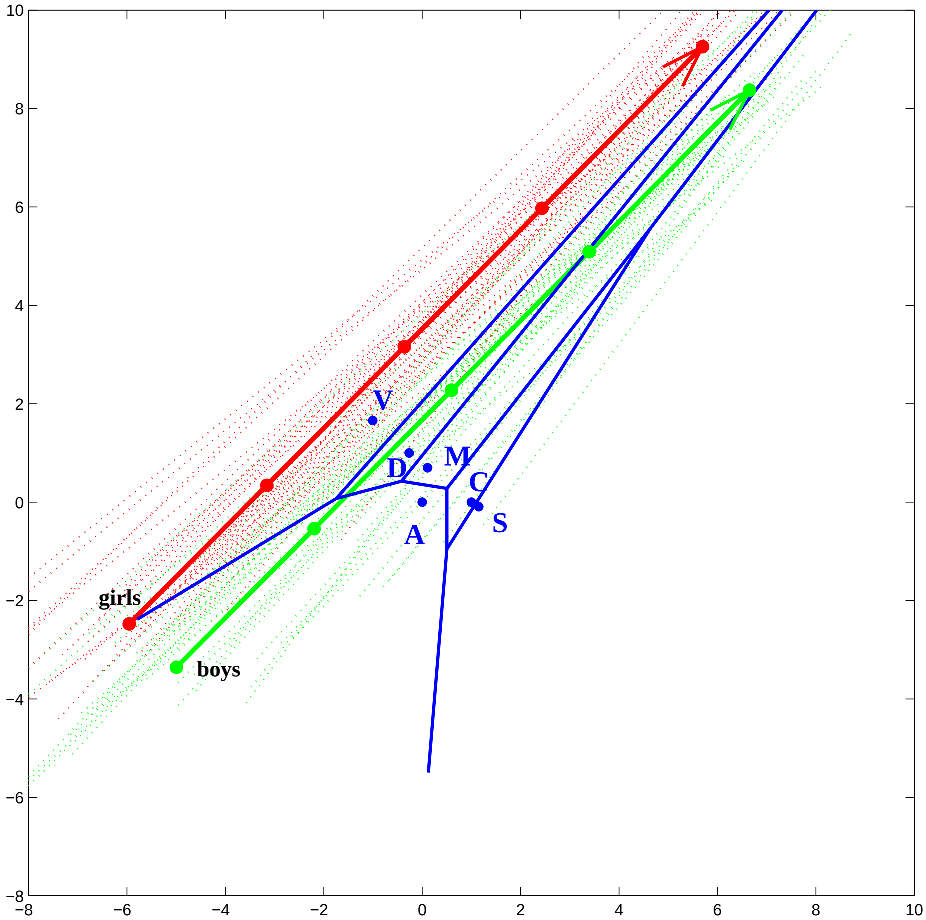

Figure 3: *The solution for the Adachi TV preference data. Mean trends for boys (green) and girls (red) are presented. The dots on the trend vectors represent the predictions at the five time points. Also shown are individual trends. Regions and category points are indicated by letters: A = Animation, C = Cinema, D = Drama, M = Music, S = Sport, and V = Variety.*

5.3 Schizophrenia example: investigating ordinality

The last example, where we re-analyze data from the National Institute of Mental Health Schizophrenia Collaborative Study, has an ordinal response variable. Following Hedeker and Gibbons (2006), we combine the three anti-psychotic drug groups into a single treatment group and the square root of the week number is used as explanatory time variable (T_{it}). The main question is whether there is a differential change across time for the treatment group relative to the placebo group. Various descriptive statistics can be found in Hedeker and Gibbons (2006).

The linear predictor for both the proportional odds model as well as the unidimensional trend vector model is

$$\eta_{it} = \alpha + \beta_1 G_i + \beta_2 T_{it} + \beta_3 G_i T_{it} + u_{0i},$$

where G_i represents the treatment ($G_i = 1$) or control group ($G_i = 0$) and T_{it} the time point. In the case of the random intercepts and slopes model the linear predictor has the following form

$$\eta_{it} = \alpha + \beta_1 G_i + \beta_2 T_{it} + \beta_3 G_i T_{it} + u_{0i} + u_{1i} T_{it}.$$

For the two dimensional model the linear predictors are

$$\eta_{itm} = \alpha_m + \beta_{1m} G_i + \beta_{2m} T_{it} + \beta_{3m} G_i T_{it} + u_{0i.m}\ (+u_{1i.m} T_{it}),$$

the parentheses at the end indicate the difference between a random intercepts model and a random intercepts and slopes model.

The statistics do not agree about the preferred model, see Table 4. BIC points towards the proportional odds model with random intercept and slope, while the AIC is smallest for the two dimensional trend vector model. A detailed description of the solution of the proportional odds model can be found in (Hedeker and Gibbons, 2006, pages 207-212). Here we will show the two dimensional trend vector solution, to discover violations of the ordinality assumption.

The solution of the two dimensional model is presented in Figure 4. The trend of the treatment group follows the direction of the ordinal dimension. The trend for the

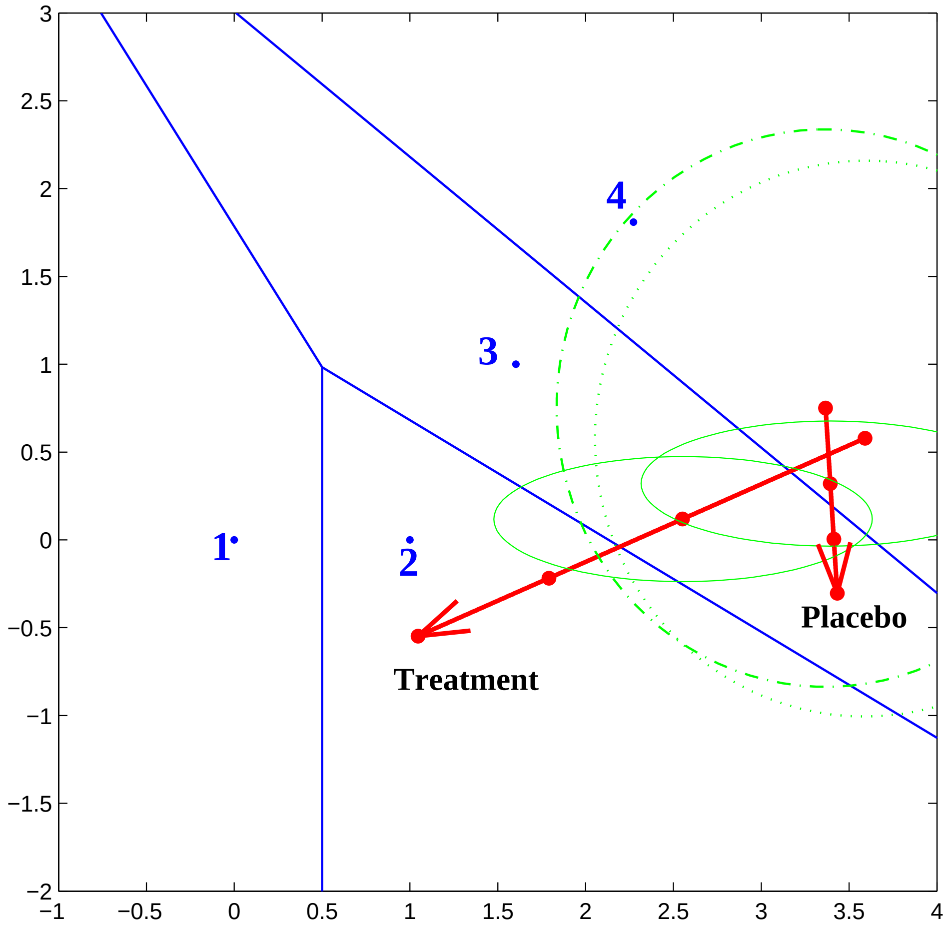

Figure 4: *The solution for two dimensional trend vector model with random intercepts and slopes on the schizophrenia data set. Approximate 68% regions for the random intercepts are given by the dashed ellipses. For the random slopes approximate 68% intervals are indicated by the solid ellipses with center equal to the second dot on the trend vectors. The digits represent the four categories: 1 normal or borderline mentally ill; 2 mildly or moderately ill; 3 markedly ill; and 4 severely or among the most extremely ill.*

Table 4: *Fit statistics for the schizophrenia data set. POM is the proportional Odds Model. Interval refers to the trend vector model assuming equally spaced category points, ordinal to the ordered category points case, and two-dimensional refers to the two dimensional model. Under Random the random effects vector is given where I represents the random intercepts model and I + S the random intercepts and slopes model. The number of fitted parameters is given under npar, -2LL is the deviance, BIC the Bayesian Information Criterium, AIC is Akaike's Information Criterion, and AICC is the small sample bias corrected version of AIC.*

Model	Random	npar	-2LL	BIC	AIC	AICC
POM	I	7	3404.2	3446.7	3418.2	3418.2
	I + S	9	3326.5	**3381.2**	3344.5	3344.6
interval	No	5	3792.6	3829.5	3802.6	3802.7
	I	6	3431.6	3468.1	3443.6	3443.7
	I + S	8	3367.3	3416.0	3383.3	3383.4
ordinal	No	7	3767.9	3819.5	3781.9	3781.9
	I	8	3417.7	3466.4	3433.7	3433.8
	I + S	10	3350.3	3411.1	3370.3	3370.4
two-dimensional	I	14	3347.0	3432.1	3375.0	3375.2
	I + S†	17	3308.8	3412.1	**3342.8**	3343.2

† The dimensionwise structure is reported. The intermediate and full structure did not fit better.

placebo group is quite different and seems to stay away from the healthy category (1). The disordinality seems to be present in the interaction effect of group by time. While the treatment group obtains a growing probability of belonging to category normal or borderline ill (1) over time, the placebo group does not develop into that direction at all. The probability for category 1 for the placebo group remains zero, see also Table 5. For the treatment group there is a definite progression, whereas for the placebo group there is still a large probability of belonging to category four, severely ill, at the last time point.

In this case the linear predictor contains both random intercepts and slopes. Each individual has a different starting point and a different linear trajectory. As before, the random effects are represented by ellipses. The dashed ellipse represents the random intercept for the placebo group, while the dotted ellipse represents the random intercept for the treatment group. The solid ellipses represent random slopes. The center of these ellipses is at the second dot (representing $T = 1$) on each of the trend vectors. The random intercepts are displacements of the starting point of the trend vectors, while the random slopes change the length and direction of the linear trend vector.

Table 5: *Conditional probabilities for the four categories over time for the placebo and treatment group, both given* $\mathbf{u}_i = \mathbf{0}$

		Category			
group	T	1	2	3	4
placebo	0	0.00	0.01	0.29	0.69
	1	0.00	0.05	0.43	0.52
	3	0.00	0.11	0.51	0.38
	6	0.00	0.21	0.54	0.25
treatment	0	0.00	0.02	0.29	0.70
	1	0.00	0.27	0.57	0.16
	3	0.05	0.65	0.28	0.02
	6	0.24	0.70	0.06	0.00

6 Discussion

We presented a mixed effects trend vector model for longitudinal multinomial data. The multinomial distribution can be considered a multivariate distribution (Fahrmeier and Tutz, 2001) and hence we have a multivariate problem. In the case that the response variable has many categories standard methodology, i.e. the baseline category model, becomes unfeasible due to the dimension of the random effects. In such cases one could adopt quasi-likelihood or penalized likelihood methods but these methods are simpler and approximations are often poorer (Hartzel, Agresti and Caffo, 2001; Molenberghs and Verbeke, 2005). Following Takane (1996) we tackled the problem by using multidimensional scaling techniques. The integral dimension then is no longer dependent on the number of categories in the response variable but depends on the dimensionality of the Euclidean space. In the TV watching example data the response variable has 6 response categories. If only random intercepts are used in the mixed effects MBCL model 5 random effects would be needed. However, often we would also like random slopes, which would be computationally unfeasible. In a two dimensional trend vector model random intercepts and slopes could be fitted for the TV watching data. Dimension reduction not only makes estimation of these models possible but also increases the stability of the final model due to the smaller number of parameters.

Another way to deal with the computational complexities of the random effects models is to assume that the latent variables (random effects) are categorical as in latent class growth curve models (Vermunt, 2007) or latent Markov models (Langeheine and

Van der Pol, 1990, 1994; Vermunt, Langeheine, and Böckenholt, 1999). Latent class growth models, however, assume a discrete latent variable in contrast to the normally distributed latent variables in the trend vector model and generalized linear mixed models. Each latent class then has its own developmental trajectory over time. Assuming such a categorical latent variable solves the computational difficulties with respect to the numerical integration. However, the optimization function of latent growth curve models has many local optima. In latent class growth curve models, as in mixed effects trend vector models, the temporal dynamics are assumed to be a gradual process (Rijmen, Vansteelandt, and De Boeck, 2008). Latent Markov models, on the other hand, are capable of modeling both abrupt and smooth temporal patterns. In this respect it could be beneficial to model the TV watching data using such an approach since since we found large residuals for some cases with non smooth transitions. There is a danger, however, of obtaining very small latent classes.

Another way to deal with the numerical problems of the random effects is to assume a special structure of the random effects as in the bi-factor model (Gibbons and Hedeker, 1992; Rijmen, 2010). For a multinomial variable such an approach could only be incorporated if there is advance knowledge of which categories belong to which latent dimension. Most often such knowledge is not available, unless the variable is a cross-classification of several simpler categorical variables.

A second virtue of the presented models is that the solution can be neatly depicted in a graphical display. The graphical display immediately shows trends over time for the group(s) under study. We gave several examples throughout this paper. Understanding polynomial trends and interactions in baseline category models is usually very difficult, whereas in our graphical display this understanding is very easy.

In our mixed effects trend vector models we assumed constancy of the categories over time, while the participants change across time. In other examples the category points could change over time, for example in an election study where the political parties change their election program. When both categories and participants are allowed to change over time additional restrictions need to be imposed. Such additional constraints are necessary to prevent categories and participants to move together in the same direction in Euclidean space. Therefore, a single category or participant should be constrained to be constant over time. The time effects then need to be interpreted in terms of relative change compared to this constant category/participant.

Similar graphical displays could be obtained by using multidimensional unfolding techniques like PREFSCAL (Busing, *et. al.*, 2005) where time and group information are used as external variables. Multidimensional unfolding does not make any distributional assumptions and parameters are estimated by least squares. In our procedure we assume a multinomial distribution of the response variable given the random effects. An advantage of the trend vector model is that it is possible to test effects, for example whether there is a difference between boys and girls or treatment and control. Multidimensional unfolding does not give such tests. In multidimensional unfolding the random effects are treated as fixed effects. Therefore, if more subjects are available more parameters have to be estimated. Under maximum likelihood this would bias the parameter estimates (see Neyman and Scott, 1948).

In Section 5.2 we showed some residual plots to look at the fit of the model. For generalized linear mixed models (GLMMs) little research has been conducted on residuals and diagnostics for influential cases. Such statistics have been to be found really useful for generalized linear models and linear models. Further research is needed to study residuals and diagnostics and their properties for GLMMs and mixed effects models for multinomial data.

Recently, in the IRT framework the classical models like the Rasch model and the two parameter logistic model were extended to explanatory IRT models (Rijmen, Tuerlinckx, De Boeck, and Kuppens, 2003; De Boeck and Wilson, 2004). Our mixed effects trend vector model can also be applied to other types of hierarchical or multilevel data, for example questionnaire data. In this case a mixed effects IPC model would result. Takane (1996) proposed such a model without fixed effects or random slopes. Our mixed effects IPC model can be conceived as an explanatory variant of the model proposed by Takane.

The trend vector model has to be contrasted with the squared logistic model (Andrich, 1988) and similar unfolding IRT models. In the squared logistic model for binary agree-disagree data the probability of agreeing is modeled using a unimodal curve. An importance difference between this model and our IPC model is that the disagree category is not homogeneous in unfolding IRT models. That is, people choosing the disagree category can do so for two reasons, often denoted by disagree from below and disagree from above. This means that this category has two positions on the unidimensional scale, which is implicit in the squared logistic model but more explicit in, for example, the generalized graded unfolding model (Roberts, Donoghue, and Laughlin, 2000).

As discussed above, the multinomial distribution for a response variable with C categories can be considered as a multivariate binomial distribution, with dimensionality $C - 1$. When C is large, the problem is high dimensional. Recently, there has been increased interest in models for longitudinal multivariate data. Fieuws and Verbeke (2009) provide an overview of joint models for high dimensional data. They distinguish between four types of models, depending on two facets: whether or not latent variables are used for reducing the dimensionality of the response space and whether or not latent variables are used for the time dimension. In the first approach the variables are assumed to measure one (or more) latent factors or components. In the second approach it is assumed that the observed measurements reflect a latent evolution for each of the outcomes. By crossing both facets Fieuws and Verbeke distinguish four types of models: Models for the evolution in observed variables (Type I), models for the latent evolution in observed variables (Type II), models for the evolution in latent variables (Type III), and models for the latent evolution in latent variables (Type IV).

We considered models with a latent variable for the time dimension, hence we proposed models of type II (models in maximum dimensionality) and IV (models in reduced dimensionality). For models of type IV we used multidimensional scaling ideas for reducing the dimensionality. That is, we assumed an underlying latent space in which relationships between subject points and response category points are specified by distances. In contrast, in factor analysis or principal component models, these relationships are specified by inner-products or projections. Often inner product and distance methods result in equivalent solutions, see Tsai and Böckenholt (2001) and De Rooij and Heiser (2005). Equivalence of the inner product and distance methods occurs when there are so called main effect parameters for both the 'rows' (subjects) and the 'columns' (categories) in the model. If such a term is present in the model it absorbs the squared terms of the Euclidean distance formulation and thus the two representations are equivalent. For our model, however, there is no equivalent inner product representation, because there is no main effect term for the categories of the response variable.

Appendix: Annotated SAS code for examples

In this Appendix `SAS NLMIXED` code is shown for the examples in Sections 5.2 and 5.3.

In Figure 5 we present the code for the final model on the TV preference data. On lines `002-005` the parameters are defined and starting values are given. On line `007` the identification restrictions of the model are given. On lines `009-010` the linear predictors are defined and these are used in lines `012-014` to compute quadratic Euclidean distances between the positions of the participants and the category points. From line `016` till `020` the probabilities are defined: First the denominator is computed and then the probabilities for each class are defined. With these probabilities the likelihood is defined in line `023-024` and the distribution of the random effects is specified on line `025`. Finally, the Expected A Posteriori estimates of the random intercepts and slopes are asked to be writen to files on lines `028-031`.

```
proc nlmixed data=adachi noad qpoints = 5;
PARMS
b01 0.2297
...
r12 r13 r24 r34 0;
/*identification restriction*/
z11=0;z12=0;z21=1;z22=0;z32=1;
/*Code linear predictors- age centered */
eta1 = b01 + b21*GENDER + b11*(age-12.25) + u1 + (age-12.25) * u2;
eta2 = b02 + b22*GENDER + b12*(age-12.25) + u3 + (age-12.25) * u4;
/*Code squared distances*/
dist1= (eta1-z11)*(eta1-z11) + (eta2-z12)*(eta2-z12);
...
dist6= (eta1-z61)*(eta1-z61) + (eta2-z62)*(eta2-z62);
/*compute probabilities*/
denom = exp(-(dist1)) + exp(-(dist2)) + exp(-(dist3)) + exp(-(dist4)) + exp(-(dist5)) + exp(-(dist6));
if (choice = 1) then p = exp(-(dist1)) / denom;
else if (choice = 2) then p = exp(-(dist2)) / denom;
...
else if (choice = 6) then p = exp(-(dist6)) / denom;
/*Define likelihood*/
ll = log(p);
model choice ~ general(ll);
/*Specify random effect distribution*/
random u1 u2 u3 u4   normal([0,0,0,0], [s1*s1, r12*s1*s2 ,s2*s2, r13*s1*s3 ,0,s3*s3,0,r24*s2*s4,
r34*s3*s4,s4*s4]) subject = pident;
replicate freq;
/* EAP estimates */
predict u1 out= Int1;
predict u2 out= Slp1;
predict u3 out= Int2;
predict u4 out= Slp2;
run;
```

Figure 5: SAS NLMIXED code for selected model for TV preference data.

In Figure 6 we present the code for the final model of the Schizophrenia study. On lines `002-004` the parameters are defined and starting values are given. On line `006` the identification restrictions of the model are given. On lines `008-009` the linear predictors are defined and these are used in lines `011-014` to compute quadratic Euclidean distances between the positions of the participants and the category points. From line `016` till `020` the probabilities are defined: First the denominator is computed and then the probabilities for each class are defined. With these probabilities the likelihood is defined in line `022-023` and the distribution of the random effects is specified on line `024`.

```
proc nlmixed noad qpoints=5;
parms
b1 -b8 0 z31= 0 z41=0 z42= 0
sd1= 1 sd2 = 1 sd3= 1 sd4 = 1 cov12 cov34 0;
/*restrictions*/
z11= 0; z12 = 0; z21=1; z22= 0; z32=1;
/* linear predictor */
eta1 = b1 + b2*Time + b3* G + b4*G*Time + u1 + u2*Time;
eta2 = b5 + b6*Time + b7* G + b4*G*Time + u3 + u4*Time;
/* specify squared distances */
dist1 = (eta1-z11)*(eta1-z11) + (eta2-z12)*(eta2-z12);
dist2 = (eta1-z21)*(eta1-z21) + (eta2-z22)*(eta2-z22);
dist3 = (eta1-z31)*(eta1-z31) + (eta2-z32)*(eta2-z32);
dist4 = (eta1-z41)*(eta1-z41) + (eta2-z42)*(eta2-z42);
/* specify probabilities */
denom = exp(-dist1)+exp(-dist2)+exp(-dist3)+exp(-dist4);
if (RESP = 1) then p = exp(-dist1)/denom;
else if (RESP = 2) then p = exp(-dist2)/denom;
else if (RESP = 3) then p = exp(-dist3)/denom;
else if (RESP = 4) then p = exp(-dist4)/denom;
/* specify likelihood */
ll = log(p);
model RESP ~ general(ll);
random u1 u2 u3 u4 normal([0,0,0,0],[sd1*sd1,cov12,sd2*sd2,0,0, sd3*sd3,0,0,cov34,sd4*sd4]) subject = ID;
run;
```

Figure 6: SAS NLMIXED code for selected model for the Schizophrenia study

Acknowledgements

The authors thank Frank Busing and Anne Miers for detailed comments on an earlier version of this manuscript. This research was conducted while the first author was sponsored by the Netherlands Organisation for Scientific Research (NWO), Innovational Grant, no. 452-06-002.

References

Adachi, K. (2000). Scaling of a longitudinal variable with time-varying representation of individuals. *British Journal of Mathematical and Statistical Psychology*, **53**, 233–253.

Agresti, A. (2002). *Categorical Data analysis*, (2nd Edition). New-York: John Wiley and Sons.

Anderson, J.A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society, series B*, **46**, 1–30.

Andrich, D. (1988). The application of an unfolding model of the PIRT type for the measurement of attitude. *Applied psychological measurement*, **12**, 33–51.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models *Biometrics*, **55**, 218–234.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, **37**, 29–51.

Busing, F.M.T.A., Heiser, W.J., Neufeglise, P., and Meulman, J.J. (2005). *PREFSCAL: Program for metric and nonmetric multidimensional unfolding, including individual differences and fixed coordinates.* SPSS Inc., Chicago, IL.

De Boeck, P. and Wilson, M. (2004). *Explanatory Item Response Models.* New-York: Springer-Verlag.

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**, 137–157.

De Rooij, M. (2009a). Ideal point discriminant analysis revisited with an emphasis on visualization. *Psychometrika*, **74**, 317-330.

De Rooij, M. (2009b). Trend vector models for the analysis of change in continuous time for multiple groups. *Computational Statistics and Data Analysis*, **53**, 3209-3216.

De Rooij, M. (2011). Transitional ideal point models for longitudinal multinomial outcomes. *Statistical Modelling*, **11**, 115-135.

De Rooij, M. and Heiser, W.J. (2003). A distance representation of the quasi-symmetry model and related distance models. In: Yanai, H., Okada, A., Shigemasu, K., Kano, Y., and Meulman, J.J. (Eds.) *New developments in psychometrics*, pp. 487–494. Tokyo: Springer-Verlag

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

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

Fahrmeier, L. and Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models*, (2nd Edition). New-York: Springer.

Fieuws, S. and Verbeke, G. (2009). Joint models for high-dimensional longitudinal data. In: Fitzmaurice, G, Davidian, M., Verbeke, G., and Molenberghs, G. (Eds.) *Longitudinal data analysis*, pp. 367–391. Boca Raton: Chapman & Hall/CRC.

Fox, J. and Andersen, R. (2006). Effect displays for multinomial and proportional-odds logit models. *Sociological Methodology*, **36**, 225–255.

Gibbons, R.D. and Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika*, **57**, 423-436.

Gower, J. C. and Hand, D. J. (1996). *Biplots*. London: Chapman and Hall.

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

Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. *Statistics in Medicine*, **22**, 1433–1446.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal data analysis*, Hoboken, New Jersey: John Wiley and Sons, Inc..

Johnson, T.R. (2007). Discrete choice models for ordinal response variables: A generalization of the stereotype model. *Psychometrika*, **72**, 489-504.

Langeheine, R. and Van der Pol, F. (1990). A unifying framework for Markov modeling in discrete space and discrete time. *Sociological Methods and Research*, **18**, 416–441.

Langeheine, R. and Van der Pol, F. (1994). Discrete time mixed Markov latent class models. In: Dale, A. and Davies, R.B. (Eds.) *Analyzing social and political change. A casebook of methods*, pp. 170–197. London: Sage.

Lesaffre, E. and Albert, A. (1989). Multiple-group logistic regression diagnostics. *Applied Statistics*, **38**, 425–440.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, series B*, **42**, 109–142.

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

Neyman, J. and Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, **16**, 1–32.

Peterson, B. and Harrell, F.E. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics*, **39**, 205–217.

Rijmen, F. (2010). Formal relations and an empirical comparison between the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, **47**, 361–372.

Rijmen, F., De Boeck, P., and Van der Maas, H.L.J. (2005). An IRT model with a parameter driven process for change. *Psychometrika*, **70**, 651–669.

Rijmen, F., Ip, E.H., Rapp, S., and Shaw, E.G. (2008). Qualitative longitudinal analysis of symptoms in patients with primary and metastatic brain tumors. *Journal of the Royal Statistical Society, series A*, **171**, 739–753.

Rijmen, F., Tuerlinckx, F., De Boeck, P., and Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, **8**, 185-205.

Rijmen, F., Vansteelandt, K., and De Boeck, P. (2008). Latent class models for diary method data: process for change. *Psychometrika*, **73**, 167–182.

Roberts, J.S., Donoghue, J.R., and Laughlin, J.E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, **24**, 3-32.

Singer, J.D. and Willett, J.B. (2003). *Applied longitudinal data analysis: Modeling change and event occurence*. New York: Oxford University Press.

Skrondal, A. and Rabe-Hesketh, S. (2003). Multilevel logistic regression for polytomous data and rankings. *Psychometrika*, **68**, 267–287.

Spinhoven, P., De Rooij, M., Heiser, W.J., Smit, J., and Penninx, B.W.J.H. (2011). A new picture of the prognostic value of personality dimensions for changes in (co-) morbidity of depressive and anxiety disorders. *Submitted paper*.

Takane, Y. (1996). An item response model for multidimensional analysis of multiple-choice data. *Behaviormetrika*, **23**, 153–167.

Takane, Y., Bozdogan, H. and Shibayama, T. (1987). Ideal point discriminant analysis. *Psychometrika*, **52**, 371–392.

Tsai, R.-C. and Böckenholt,U. (2001). Maximum likelihood estimation of factor and ideal point models for paired comaprison data. *Journal of Mathematical Psychology* **45** 795–811.

Tuerlinckx, F., Rijmen, F., Verbeke, G., and De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, **59**, 225–255.

Vermunt, J.K. (2007). Growth models for categorical response variables: Standard, latent class, and hybrid approaches. In: Van Montfort, K., Oud, J., and Satorra, A. (Eds.) *Longitudinal models in the behavioral and related sciences*, pp. 139–158. Mahwah, New Jersey: Lawrence Erlbaum Associates, publishers.

Vermunt, J.K., 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**, 178–205.

Yu, H.-T. and De Rooij, M. (2009). Model selection for the trend vector model. *Submitted paper.*

Zhang, Q., Snow Jones, A., Rijmen, F., and Ip, E. (2010). Multivariate discrete hidden Markov models for domain based measurements and assessment of risk factors in child development. *Journal of Computational and Graphical Statistics*, **19**, 746–765.