Journal of Classification 30:338-369 (2013)
DOI: 10.1007/s00357-013-9138-3

Model Selection for the Trend Vector Model

Hsiu-Ting Yu

McGill University, Canada

Mark de Rooij

Leiden University, The Netherlands

Abstract: Model selection is an important component of data analysis. This study focuses on issues of model selection for the trend vector model, a model for the analysis of longitudinal multinomial outcomes. The trend vector model is a so-called marginal model, focusing on population averaged evolutions over time. A quasi-likelihood method is employed to obtain parameter estimates. Such an optimization function in theory invalidates likelihood-based statistics, such as the likelihood ratio statistic. Moreover, standard errors obtained from the Hessian are biased. In this paper, the performances of different model selection methods for the trend vector model are studied in detail. We specifically focused on two aspects of model selection: variable selection and dimensionality determination. Based on the quasi-likelihood function, selection criteria analogous to the likelihood ratio statistics, AIC and BIC, were employed. Additionally, Wald and resampling statistics were included as variable selection criteria. A series of simulations were carried out to evaluate the relative performance of these criteria. The results suggest that model selection can be best performed using either the quasi likelihood ratio statistic or the quasi-BIC. A special study on dimensionality selection found that the quasi-AIC also performs well for cases with degrees of freedom greater than 8. Another important finding is that the sandwich estimator for standard errors used in Wald statistics does not perform well. Even for larger sample sizes, the bias-correction procedure for the sandwich estimator is needed to give satisfactory results.

Keywords: Model selection; Longitudinal multinomial data; Information criterion; Resampling methods; Sandwich variance estimator.

This research was conducted while both authors were sponsored by the Netherlands Organisation for Scientific Research (NWO), Innovational Grant, no. 452-06-002. The first author would also like to thank Ming-Mei Wang for her extensive help in revising this paper.

Corresponding Author's Address: Hsiu-Ting Yu, Department of Psychology, McGill University, 1205 Dr. Penfield Avenue, Montreal, QC H3A 1B1, Canada, tel:(514) 398-6109, e-mail: hsiu-ting.yu@mcgill.ca.

Published online: 24 October 2013

1. Introduction

There was an increased interest in statistical modeling for longitudinal data in the last decades. Numerous papers have been written about modeling of normally distributed response variables. Only recently, these models were generalized to non-normally distributed response variables like dichotomous or count data. Major developments were the Generalized Estimating Equation (GEE) model and the Generalized Linear Mixed Model (GLMM), an overview can be found in Molenberghs and Verbeke (2005). These methods apply to response variables having a distribution in the exponential family. In this paper we are interested in analyzing longitudinal multinomial data. A major issue with these data relates to dimensionality. For example, the dimensionality is $G-1$ in the multinomial logistic models when the response variable has G classes. An approach to address the problem of dimensionality reduction was recently developed with application of multidimensional scaling techniques. The resulting model is referred to as the trend vector model (De Rooij 2009).

1.1 The Trend Vector Model

Recently, De Rooij (2009) proposed the trend vector model for the analysis of longitudinal multinomial data. It uses multidimensional scaling techniques to reduce dimensionality. In the trend vector model, the conditional probability $\pi_{jt}(\mathbf{x}_{it})$ of an outcome category $j = 1, \ldots, G$ at time point $t = 1, \ldots, T_i$, for subject i ($i = 1, \ldots, N$) with p-dimensional covariate vector $\mathbf{x}_{it} = (x_{it1}, \ldots, x_{itp})^T$ is modeled. This vector contains time and possibly other explanatory variables (e.g., treatment groups, gender). The conditional probability is modeled by the squared distance between two points in Euclidean space of dimensionality M ($M \leq G-1$): an ideal point $\mathbf{y}_{it}$ for the subject and a class-point $\mathbf{z}_j$ for the category. The ideal points $\mathbf{y}_{it} = (y_{it1}, \ldots, y_{itM})^T$ are assumed to be a linear combination of the predictor variables $\mathbf{x}_{it}$, i.e.,

$$\mathbf{y}_{it} = \mathbf{B}^T \mathbf{x}_{it}.$$

The conditional probability that subject i at time point t will be in class j is then equal to

$$\pi_{jt}(\mathbf{x}_{it}) = \frac{\exp(-d^2_{(it)(j)})}{\sum_k \exp(-d^2_{(it)(k)})}, \tag{1}$$

where $d^2_{(it)(j)}$ is the squared Euclidean distance between the ideal point for subject i at time point t and the class point for category j in M-dimensional space, i.e.,

$$d^2_{(it)(j)} = \sum_{m=1}^{M} (y_{itm} - z_{jm})^2.$$

The parameters of the trend vector model are the regression weights ($\mathbf{B}$) and the class points (gathered in a matrix $\mathbf{Z}$), and will be denoted by $\boldsymbol{\theta}$. The total number of free parameters is $P = [(p + G) \cdot M - \max(M(M - 1)/2, M(M + 1) - (G - 1))]$ (De Rooij 2009). The trend vector model results in a biplot type of display, where trends are represented for groups of subjects. Examples can be found in De Rooij (2009).

1.2 Estimation

For estimating the trend vector model, De Rooij (2009) maximized

$$QL(\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{t=1}^{T_i} \prod_{j=1}^{G} \pi_{jt}(\mathbf{x}_{it})^{f_{itj}}, \tag{2}$$

where f_{itj} equals 1 if subject i at time point t is in category j, zero otherwise. The function QL in Equation (2) is the likelihood function for cross-sectional data. In our case, where we have repeated measurements, QL is not a genuine likelihood since it ignores the dependencies among the responses for every subject. Liang and Zeger (1986) showed that maximizing the cross sectional likelihood (i.e., Equation 2) with repeated measurements still provides consistent estimates of the model parameters for responses in the exponential family. Lipsitz, Kim and Zhao (1994) generalized the work of Liang and Zeger (1986) to the case of multinomial responses. Standard errors obtained from the corresponding Hessian are biased, but this bias can be repaired by using sandwich-type adjustments to the Hessian (Liang and Zeger 1986, Lipsitz, Kim and Zhao 1994, also see the discussion in a later section on the Sandwich Covariance Estimator).

1.3 Application and Problem Definition

We use data published in Adachi (2000) as an example of applying the trend vector model. In this research, Japanese boys and girls were asked to report their preferred type of TV programme at 5 time points. 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 TV programme categories are Animation (A), Cinema (C), Drama (D), Music (M), Sport (S), and Variety (V). Table 1 gives a summary of the data.

Table 1. Summary of TV preference data for Japanese boys and girls.

		Time point				
Group	Status	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
Boys	A	36	26	2	1	0
	C	0	2	4	4	10
	D	1	0	7	14	4
	M	1	2	10	10	10
	S	3	4	9	7	14
	V	8	15	17	13	11
Girls	A	49	31	8	2	1
	C	0	0	1	3	11
	D	0	6	26	21	12
	M	0	1	6	13	15
	S	0	1	0	2	4
	V	2	12	10	10	8

We constructed a time variable T by using the midpoints of the ages at each of the five time points as scores (i.e., 6.5 for the first time point, 9.5 for the second, etc.) and centering around the mean (12.25). The question is whether boys and girls differ in their trends in watching behavior and what the trends look like. Figure 1 gives the solution where there is a main effect for gender and the time development follows a polynomial of degree two. Boys and girls have the same quadratic trend over time, but a different starting position. Although boys and girls both start with preferring 'animation' the trend for boys passes through 'variety', 'music', and ends in 'sports', while that for girls passes through 'drama' and ends in 'music'.

The main question that we will deal with in the rest of this manuscript is how to select the 'best' model from all candidate models. In the example shown above, a main effect for gender was selected and a polynomial of degree two for both boys and girls was selected. However, the possible effects in the example above are

(G) a dummy variable for gender;
(T) linear time;
(T^2) quadratic time;
(T^3) cubic time;
(T^4) quartic time;
(GT) gender by time interaction variable;
(GT^2) gender by quadratic time interaction variable;
(GT^3) gender by cubic time interaction variable;
(GT^4) gender by quartic time interaction variable.

In other examples, there might be other background variables to be included in the model. Therefore, there are many possible candidate models to choose

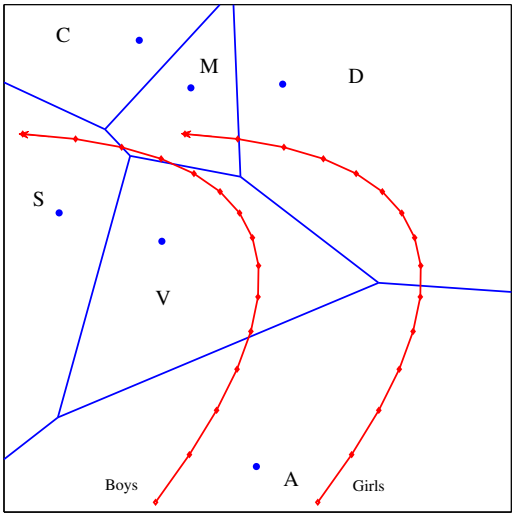


Figure 1. Solution of the trend vector model on the TV preference data

from. In addition, since the response variable has six classes, the dimensionality may be any number between one and five. Since the model is estimated using a type of quasi-likelihood, statistics based on genuine likelihoods (e.g., likelihood ratio statistics and information criteria) cannot formally be applied. Moreover, as we already discussed above, standard errors derived from the Hessian obtained from maximizing the quasi-likelihood are generally biased. Sandwich-type adjustments have been proposed to repair this bias (Liang and Zeter 1986; Lipsitz, Kim and Zhao 1994), but the appropriateness of such adjustment for the trend vector model is not yet established. In what follows, we study likelihood-based methods, the sandwich estimators, and resampling methods like the bootstrap and jackknife, as model selection tools for the trend vector model.

2. Model Selection for the Trend Vector Model

Model selection is the task of selecting a statistical model from a set of potential models given the data. It is an essential part of any statistical analysis. This paper presents a detailed study of the performance of different model selection methods for the trend vector model. We specifically focus on two aspects of model selection: variable selection and dimensionality determination. The goal for the *variable selection* is to correctly identify "important" predictor variables in the final model. In other words, whether

the irrelevant or unimportant predictor variables can be effectively excluded from the final model. *Dimensionality determination* aims to select the appropriate dimensionality that represents the class and ideal points in a Euclidean space.

Methods of selecting an appropriate model can be categorized into two types of indices. The first type is based on (quasi)-likelihood, and the second type on forming confidence intervals around the estimates for Wald-type statistical testing. These two types of indices are described in detail in the following sections.

2.1 Indices Based on (Quasi)-Likelihood

A standard approach in model selection uses test statistics built on likelihood. For the trend vector model, we do not have a genuine likelihood, but we investigate whether these statistics can still be used when the likelihood is replaced by the QL-function (Equation (2)). Formally, the likelihood function is defined as $L(\boldsymbol{\theta}|\text{data}) = P(\text{data}|\boldsymbol{\theta})$, that is, the likelihood of a set of parameters ($\boldsymbol{\theta}$) given the data is equal to the probability of the observed data given the set of parameters. Therefore, "likelihood" can be viewed as "evidence" that supports one set of parameters against the other. In this paper, we consider two likelihood-based model selection methods: the likelihood ratio test and information criteria.

2.1.1 Likelihood Ratio Test

The *likelihood ratio test* (LRT) compares log-likelihood values of a full model (a model containing all predictor variables) and a reduced model (a model containing fewer variables than the full model). Let $\boldsymbol{\theta}$ denote the vector of parameters of interest, and $\hat{\boldsymbol{\theta}}$ the *unrestricted maximum likelihood* estimate. The likelihood evaluated at this value is $L(\hat{\boldsymbol{\theta}})$. Furthermore, let $\hat{\boldsymbol{\theta}}_R$ denote the *restricted* maximum likelihood estimate and the likelihood evaluated at this value, $L(\hat{\boldsymbol{\theta}}_R)$. The likelihood ratio, $L(\hat{\boldsymbol{\theta}}_R)/L(\hat{\boldsymbol{\theta}})$, is the degree to which the observations support parameters $\hat{\boldsymbol{\theta}}_R$ relative to $\hat{\boldsymbol{\theta}}$; the LRT is a transformation of the log-likelihood ratio,

$$\text{LRT} = -2 \ln \frac{L(\hat{\boldsymbol{\theta}}_R)}{L(\hat{\boldsymbol{\theta}})} = -2[\ln L(\hat{\boldsymbol{\theta}}_R) - \ln L(\hat{\boldsymbol{\theta}})]. \tag{3}$$

The LRT test statistic, $-2 \ln L(\hat{\boldsymbol{\theta}}_R)/L(\hat{\boldsymbol{\theta}})$, has an asymptotically chi-square distribution with degrees of freedom equal to the difference in dimensionality of $\boldsymbol{\theta}_R$ and $\boldsymbol{\theta}$.

The LRT is a simple and widely used method for model selection. We investigate whether the LRT can still be used for model selection for the

trend vector model. We denote this test as LRT_Q to emphasize that $QL(\boldsymbol{\theta})$ is used instead of $L(\boldsymbol{\theta})$.

2.1.2 Information Criteria

Other types of indices based on the likelihood are information criteria. Akaike's information criterion (AIC) (Akaike 1974), and Bayesian information criterion (BIC) (Schwarz 1978) are typical criteria used for comparing a range of competing models. The model(s) with the lowest AIC or BIC value are considered to be the most attractive.

Akaike's Information Criterion (AIC) (Akaike 1974) was developed to compare non-nested models adjusting for the number of estimated parameters. The AIC is defined as

$$AIC = -2 \times \ln(L(\hat{\theta})) + 2 \times P. \tag{4}$$

The AIC assigns a penalty for complexity.

A similar criterion is Schwarz's Bayesian Information Criterion (BIC) (Schwarz 1978):

$$BIC = -2 \times \ln(L(\hat{\theta})) + \ln(N) \times P. \tag{5}$$

The Bayesian information criterion (Schwarz 1978) takes sample size into account. For most sample sizes, the BIC places a larger penalty ($\ln(N)$ instead of 2) on complex models compared to the AIC, which leads to a preference for simpler models.

Pan (2001) proposed the AIC with QL instead of the true likelihood under the name QIC_u for model selection in generalized linear models. Here we study this statistic for model selection in the trend vector model. A similar adaption is made to the BIC. We use the terminology AIC_Q and BIC_Q in the remainder of this manuscript to represent (4) and (5) with $QL(\boldsymbol{\theta})$ instead of $L(\boldsymbol{\theta})$.

2.2 Indices Based on Confidence Intervals

Two categories of indices for constructing confidence intervals for model selection are considered. First, resampling methods to obtain the empirical distribution of parameters. Second, sandwich covariance estimators as proposed by Liange and Zeger (1986) and Lipstiz, Kim and Zhao (1994).

2.2.1 The Bootstrap

Bootstrapping is a general technique for statistical inference based on building a sampling distribution for a statistic by resampling from the

data at hand. Denote by $\mathbf{S}$ the sample from the population of interest. The basic idea of the bootstrap is as follows: Draw with replacement a *bootstrap sample* of size n from the elements of $\mathbf{S}$. For longitudinal data, the sampling is performed at the individual level rather at the level of a single measurement of an individual, in order to deal with the within-individual dependency (Sherman and Le Cessie 1997). This procedure is repeated a large number of times, B; the bth such bootstrap sample is denoted as $\mathbf{S}_b^*$. From these bootstrap samples we can calculate the statistic of interest, θ_b^*. There are several approaches to constructing confidence intervals based on bootstrapping.

The *bootstrap normal-theory interval* (*BSn*) assumes that the statistic θ is normally distributed, and uses the bootstrap samples to estimate the sampling variance. Let $\bar{\theta}^*$ denote the average of the bootstrapped statistics, that is, $\bar{\theta}^* = \sum_{b=1}^{B} \theta_b^*/B$; and sampling variance of θ is $\text{Var}(\theta^*) = \sum_{b=1}^{B}(\theta_b^* - \bar{\theta}^*)^2/(B-1)$. The $100(1-\alpha)$-percent confidence interval based on standard theory is

$$\bar{\theta}^* \pm z_{1-\alpha/2}\text{SE}(\theta^*), \tag{6}$$

where $\text{SE}(\theta^*) = \sqrt{\text{Var}(\theta^*)}$, and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ percentile of the standard-normal distribution (e.g., $z_{1-\alpha/2} = 1.96$ for a 95-percent confidence interval with $\alpha = .05$).

An alternative approach, called the *bootstrap percentile interval* (*BSp*), is to use the empirical distribution of θ_b^* to form a confidence interval for θ:

$$\theta^*_{(lower)} < \theta < \theta^*_{(upper)}, \tag{7}$$

where $\theta^*_{(1)}, \theta^*_{(2)}, \ldots, \theta^*_{(B)}$ are now the ordered bootstrap replicates of the statistic, *lower* $= [(B+1)\alpha/2]$, and *upper* $= [(B+1)(1-\alpha/2)]$, the square brackets indicate rounding to the nearest integer.

2.2.2 The Jackknife

The basic idea behind the jackknife estimator lies in systematically re-estimating the statistic while leaving out one observation at a time from the sample set. The delete-one estimates, $\hat{\theta}_{(-i)}$, are obtained by computing the parameter estimates for each of n subsets of the data when leaving the i-th individual out. Again, this is on the individual level instead of the level of single measurements in order to preserve the within-subject dependencies. The jackknife estimate of the standard error is given by

$$\widehat{\text{SE}}(\hat{\theta}) = \left[\frac{(n-1)}{n}\sum_{i=1}^{n}(\hat{\theta}_{(-i)} - \hat{\theta}_{(\cdot)})^2\right]^{\frac{1}{2}}, \tag{8}$$

where

$$\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \hat{\theta}_{(-i)}/n. \tag{9}$$

The jackknife confidence interval can be obtained by

$$\hat{\theta}_{(\cdot)} \pm t_{n-1}^{1-\alpha} \widehat{\mathrm{SE}}(\hat{\theta}), \tag{10}$$

where $t_{n-1}^{1-\alpha}$ is the $(1-\alpha)$th percentile of the t distribution with $n-1$ degrees of freedom (for details see Efron and Tibshirani 1993).

2.2.3 The Bias-Corrected and Accelerated Bootstrap (*BCa*)

This procedure provides an improved version of percentile intervals. It introduces a 'correction factor' (z) to deal with asymmetry among the bootstrap estimates and a quantity called *acceleration* (a) to correct for skewness. The correction factor is obtained by

$$z = \Phi^{-1}\left[\frac{\#_{b=1}^{B}(\hat{\theta}_b^* \leq \hat{\theta})}{B+1}\right], \tag{11}$$

where $\Phi(\cdot)$ is the standard-normal cumulative distribution function, and $\#_{b=1}^{B}(\hat{\theta}_b^* \leq \hat{\theta})/(B+1)$ is the adjusted proportion of bootstrap replicates at or below the original-sample estimate $\hat{\theta}$. If the bootstrap sampling distribution is symmetric, and if θ is unbiased, then this proportion will be close to .5 and the 'correction factor' z will be close to 0.

The acceleration (a) is constructed from the delete-one estimates of the jackknife procedure $(\hat{\theta}_{(-i)})$ and the mean of these n delete-one estimates as defined in Equation (9). The acceleration (a) is calculated by

$$a = \frac{\sum_{i=1}^{n}(\hat{\theta}_{(-i)} - \hat{\theta}_{(\cdot)})^3}{6\left[\sum_{i=1}^{n}(\hat{\theta}_{(-i)} - \hat{\theta}_{(\cdot)}^2)^{\frac{3}{2}}\right]}. \tag{12}$$

With the correction factor (z) and acceleration (a), the endpoints of the bias-corrected confidence interval are

$$\hat{\theta}_{BCa_{lower}} = \theta^*_{(C_a lower)} < \theta < \hat{\theta}_{BCa_{upper}} = \theta^*_{(C_a upper)},$$

where

$$\begin{aligned}
[C_a lower] &= (B+1) \times \Phi\left[z + \frac{z - z_{1-\alpha/2}}{1 - a(z - z_{1-\alpha/2})}\right] \\
[C_a upper] &= (B+1) \times \Phi\left[z + \frac{z + z_{1-\alpha/2}}{1 - a(z + z_{1-\alpha/2})}\right].
\end{aligned} \tag{13}$$

The bias-corrected and accelerated bootstrap confidence interval improves on the percentile interval, giving better coverage for the distribution of $\hat{\theta}$ that may be biased and/or skewed (Efron and Tibshirani 1993).

2.2.4 Sandwich Covariance Estimator

Let $\mathbf{f}_{it}$ denote the dummy coded vector of the multinomial response for the ith individual at time t (e.g., a response of 2 from a 3-category outcome is coded as $[0, 1, 0]^T$), and $\mathbf{F}_i$ is the vertical concatenation of $\mathbf{f}_{it}^T$ with rows corresponding to the different time points. Without loss of generality, the last column of $\mathbf{F}_i$ can be dropped, since probabilities add up to one. Therefore, $\mathbf{F}_i$ is a $T \times (G-1)$ matrix. Define $\mathbf{f}_i$ to be a vector of length $T \cdot (G-1)$, which is the vectorized form of $\mathbf{F}_i^T$.

The sandwich covariance estimator (Liang and Zeger 1986) for model parameters has the form

$$\left(\sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i\right)^{-1} \left(\sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} \text{cov}(\mathbf{f}_i) \mathbf{V}_i^{-1} \mathbf{D}_i\right) \left(\sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i\right)^{-1}, \tag{14}$$

where $\mathbf{D}_i = \partial \boldsymbol{\pi}_i / \partial \boldsymbol{\theta}$, and $\boldsymbol{\pi}_i$ is the vector with probabilities corresponding to $\mathbf{f}_i$. The derivation of $\mathbf{D}_i$ is given in Appendix A. For our set-up $\mathbf{V}_i$ is a block-diagonal matrix with diagonal blocks $\mathbf{V}_{it}$ defined by

$$\mathbf{V}_{it} = \text{Diag}[\boldsymbol{\pi}_{it}] - \boldsymbol{\pi}_{it} \boldsymbol{\pi}_{it}^T, \tag{15}$$

where $\text{Diag}[\boldsymbol{\pi}_{it}]$ denotes a diagonal matrix with the elements of $\boldsymbol{\pi}_{it}$ on the main diagonal. The $\text{cov}(\mathbf{f}_i)$ in Equation (14) is typically estimated by $(\mathbf{f}_i - \hat{\boldsymbol{\pi}}_i)(\mathbf{f}_i - \hat{\boldsymbol{\pi}}_i)^T$.

It has been noted in the literature that for small samples (e.g., fewer than 40 individuals) the sandwich estimator tends to underestimate the variance of the regression weights (e.g., Fay and Graubard 2001; Mancl and DeRouen 2001; Kauermann and Carroll 2001; Pan and Wall 2002). One possible reason is that the $(\mathbf{f}_i - \hat{\boldsymbol{\pi}}_i)(\mathbf{f}_i - \hat{\boldsymbol{\pi}}_i)^T$ used to estimate $\text{cov}(\mathbf{f}_i)$ in Equation (14) is a biased estimator. Mancl and DeRouen (2001) proposed a method to adjust for such underestimation. They suggested replacing the $(\mathbf{f}_i - \hat{\boldsymbol{\pi}}_i)(\mathbf{f}_i - \hat{\boldsymbol{\pi}}_i)^T$ to estimate $\text{cov}(\mathbf{f}_i)$ by

$$(\mathbf{I} - \mathbf{H}_i)^{-1} (\mathbf{f}_i - \hat{\boldsymbol{\pi}}_i)(\mathbf{f}_i - \hat{\boldsymbol{\pi}}_i)^T (\mathbf{I} - \mathbf{H}_i^T)^{-1}, \tag{16}$$

where $\mathbf{I}$ is an $T \cdot (G-1) \times T \cdot (G-1)$ identity matrix and $\mathbf{H}_i$ is

$$\mathbf{H}_i = \mathbf{D}_i^T \left(\sum_{i=1}^{n} \mathbf{D}_i \mathbf{V}_i^{-1} \mathbf{D}_i^T\right)^{-1} \mathbf{D}_i \mathbf{V}_i^{-1} \tag{17}$$

(Preisser and Qaquish 1996; Mancl and DeRouen 2001). The estimates obtained by the uncorrected sandwich estimator and the bias-corrected sandwich estimator are denoted as $Sand$ and $Sand_{BC}$, respectively.

3. Simulation Design

To compare the performances of different indices, we carried out three simulation studies. The first one is designed to investigate dimensionality determination, the second one to investigate variable selection, and the third one to study the effects of degrees of freedom in model fitting on dimensionality determination. The designs and specifications of the first two studies follow the same basic approach. Data are generated with the following prescriptions. Data are generated for a longitudinal design. There are five explanatory variables $(X_1, \ldots, X_5)$ of which $X_1 - X_3$ are important and $X_4 - X_5$ are noise variables. In all cases, the true model is a two-dimensional model. We varied the following factors:

- Number of time points, $T = 3$ and $T = 5$;
- Three sample sizes are used, $N = 40$, $N = 80$, and $N = 160$;
- Response (outcome) categories are well separated on both dimensions or not;
- Explanatory variables $X_1 - X_3$ are well separated on both dimensions or not;
- The strength of the relationship between a relevant explanatory variable (X_3) and a noise variable (X_4) is varied.

Data are generated according to the trend vector model. Since a positive association among the responses is often found in longitudinal studies, we used a mixed model set-up as described in De Rooij and Schouteden (2012) to obtain correlated responses; that is, ideal points are defined by

$$\mathbf{y}_{it} = \mathbf{B}^T \mathbf{x}_{it} + \mathbf{u}_i, \tag{18}$$

where $\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{I})$, with $\mathbf{I}$ the identity matrix of order M. Together with class points these specify Euclidean distances and thus probabilities. Multinomial distributions are used to generate the observations (response categories) from the specified probabilities.

3.1 Simulation Design for Dimensionality Determination

A four-class problem $(G = 4)$ is generated to study dimensionality determination. The dimensionality is either one, two, or three for a four-class model. We systematically varied the pattern of relationships among the explanatory variables $(\mathbf{X})$, the parameters of the trend vector model: the

Table 2. Specified regression weights ($\mathbf{B}$).

	Pattern 1		Pattern 2	
	Dim-1	Dim-2	Dim-1	Dim-2
X_1	1	0	1	0
X_2	$\frac{1}{2}$	$\frac{\sqrt{3}}{2}$	$\sqrt{3}$	$\frac{1}{2}$
X_3	$\frac{\sqrt{3}}{2}$	$-\frac{1}{2}$	$\sqrt{3}$	$-\frac{1}{2}$
X_4	0	0	0	0
X_5	0	0	0	0

regression weights ($\mathbf{B}$), class locations ($\mathbf{Z}$), number of time points ($\mathbf{T}$), and sample sizes ($\mathbf{N}$). The details of the specifications are described in the following sections.

3.1.1 Explanatory Variables

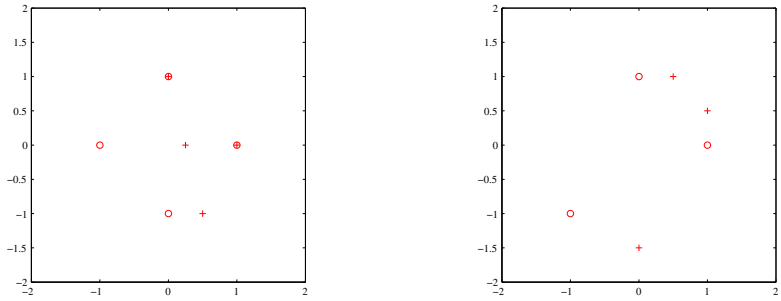
Explanatory variable X_1 is designed to be the time indicator variable, and is coded as 0, 1, 2 in the case $T = 3$ and 0, 1, 2, 3, 4 in the case $T = 5$. Explanatory variables X_1 to X_3 are meaningful explanatory variables, while X_4 and X_5 are noise variables. We systematically controlled the correlation between explanatory variables X_3 and X_4 to have four different values: 0, 0.2, 0.5, and 0.8; other variables are independent of each other. The explanatory variables (X_2 to X_5) are drawn from a multivariate normal distribution with mean $\mathbf{0}$ and variance $\boldsymbol{\sigma}^2$, where the correlation between X_3 and X_4 is specified as described above.

3.1.2 Regression Weights

We designed two patterns of regression weights ($\mathbf{B}$), the values of the weights on each dimension are listed in Table 2. Since the regression weights of X_4 and X_5 on both dimensions in both patterns are zeros, the predictor variables X_4 and X_5 are noise or irrelevant explanatory variables. In the first pattern of regression weights, the predictor variables have similar weights on each of the two dimensions. In the second pattern, the first dimension is weighted more than the second dimension. Note that the regression weight of the first predictor variable on the second dimension (b_{12}) was fixed to be zero to deal with rotation indeterminacy (for details see DeRooij 2009).

3.1.3 Class Locations

A four-class model is generated for studying dimensionality determination. The two patterns of class locations are presented in Figure 2(a). Classes of the first pattern are marked as circles, the four classes are well



(a) Four-class (b) Three-class

Figure 2. Specified class locations ($\mathbf{Z}$): Circles represent the class points for the case that the classes are well separated; pluses for the case of less separation.

separated in this case. In the second pattern, represented by plus signs, two classes are close to each other on the vertical dimension and separated from the other two classes.

3.1.4 Number of Time Points

There are two levels on this condition. The shorter sequence has three time points, and longer one has five time points.

3.1.5 Sample Sizes

Three different sample sizes ($N = 40, 80, 160$) are chosen to represent typical small, median, and large sample sizes.

3.2 Simulation Design for Variable Selection

The simulation design for the variable selection of the trend vector model focuses on a three-class model ($G = 3$) in two dimensions. The specifications of regression weights ($\mathbf{B}$), correlation patterns among predictor variables ($\mathbf{X}$), number of time points ($\mathbf{T}$), sample sizes ($\mathbf{N}$), and the within-individual dependency between time points are identical to the simulation design for dimensional determination in the previous section.

The locations of the three classes are illustrated in Figure 2(b). The first pattern (circles) has all three classes well separated in space, the second pattern (pluses) has two classes close to each other in one dimension and separated from the third class.

3.3 Simulation Design for a Special Study on Dimensionality

It can be seen from Equations (3) and (4) that the model selection criteria LRT_Q and AIC_Q for comparing model i with a larger model j (i.e.,

a model with more free parameters or a model with higher dimensionality in the current context) are related as:

$$\text{LRT}_Q = AIC_Q(i) - AIC_Q(j) + 2k. \tag{19}$$

Note that k is the difference between the number of free parameters (P_i) in model i and (P_j) in model j. That is, $k = P_j - P_i > 0$ is the number of additional free parameters in the larger model.

The common practice is that when the likelihood ratio statistic approaches a χ^2 distribution, the larger model (i.e., model j) is favored if $\text{LRT}_Q(i, j)$ exceeds $\chi^2_{k,.05}$, where k is the degrees of freedom and .05 is the significance level of the χ^2 distribution. Otherwise, the decision is to keep the smaller model. Likewise, the decision based on AIC_Q is to favor model j if $AIC_Q(j) < AIC_Q(i)$, or equivalently, if $\text{LRT}_Q(i, j) > 2k$. It follows that there might be a reverse in decisions (i.e., arriving at opposite decisions) between the two different selection criteria. Specifically, decisions based on AIC_Q may be slightly more liberal in favor of a larger model when k is less than 8 and thus $\chi^2_{k,.05}$ is greater than $2k$. On the other hand, decisions based on AIC_Q may be slightly more conservative in choosing a larger model when k is equal to or greater than 8 and thus $\chi^2_{k,.05}$ is less than $2k$.

To study empirically how the relative performances of LRT_Q and AIC_Q on dimensionality selection may be affected by the degrees of freedom k, we have designed a series of simulations involving different sizes of k by varying the number of classes (G) and the number of explanatory variables (p). In total, there are 16 sets of simulations: $G = 3$, 4, 5, and 6 by $p = 4$, 5, 6, and 7. The number of time points and sample size are the same across sets of simulations: $T = 5$ and $N = 80$. In addition, the class points and regression weights maintain a common pattern across simulation sets. Data are generated with a two dimensional model in the same manner as described earlier using multinomial distributions. For each simulation condition, 100 replications are obtained. The simulation plan is presented in Table 3. As can be seen in the table, k ranges from 6 to 12, including the pivotal value $k = 8$, for which $\chi^2_{k,.05} = 15.51$ and $2k = 16$.

For each replication, the data were fitted with three trend-vector models: one-dimensional, two-dimensional, and three-dimensional. In order to cover a sufficient range of k, we consider models from the perspective of item response processes. Thus the models are subject to rotational indeterminacy[1]. Table 3 also shows the number of parameters estimated for

1. It should be noted that the more general trend vector model as interpreted from the metric unfolding perspective may be subject to additional indeterminacies if certain restrictions are put on class points (for a detailed discussion on identification of the model, see De Rooij 2009). However, the additional indeterminacies would in effect reduce the degrees of freedom k, resulting in an inadequate coverage of the range of interest in the current study.

Table 3. Simulation plan for the special dimensionality study.

		# par(P)			# par diff(k)				# par(P)			# par diff(k)	
G	p	M1	M2	M3	M2-M1	M3-M2	G	p	M1	M2	M3	M2-M1	M3-M2
	4	7	13	na	6	na		4	9	17	24	**8**	7
3	5	8	15	na	7	na	5	5	10	19	27	9	**8**
	6	9	17	na	**8**	na		6	11	21	30	10	9
	7	10	19	na	9	na		7	12	23	33	11	10
	4	8	15	21	7	6		4	10	19	27	9	**8**
4	5	9	17	24	**8**	7	6	5	11	21	30	10	9
	6	10	19	27	9	**8**		6	12	23	33	11	10
	7	11	21	30	10	9		7	13	25	36	12	11

Note. For G=3, the maximum dimension fitted is 2 and model selection between dimensionality of 2 and 3 is not applicable (na).

each model, the difference between one-dimensional and two-dimensional models, and the difference between two-dimensional and three-dimensional models.

3.4 Analysis Procedures

For the first two simulation studies, we systematically control for five factors: The pattern of class locations (well-separated vs. not well-separated), the degree of correlation of a noise variable with one valid explanatory variable ($r_{34} = 0, 0.2, 0.5, 0.8$), the regression weights of predictor variables attributed to each dimension (similar weights on two dimensions vs. emphasis on one dimension), the sample size ($N = 40, 80, 160$), and the number of time points ($T = 3, 5$). We simulated 50 data sets for each combination of these five control factors. A trend vector model was fitted to each data set, and we studied the performance of different indices for dimensionality determination and variable selection.

The performance of each index for dimensionality determination and variable selection is investigated by studying the number of replications out of the 50 that a particular method can correctly identify the true dimensionality or discriminate between relevant and irrelevant predictor variables. ANOVA tests are performed to investigate whether the performances of different indices (within factor) are affected by the control factors (between factors). The measure of effect size (η^2) is also included for each significant effect.

For the special simulation study, the relative performances of LRT_Q and AIC_Q on dimensionality selection are evaluated in terms of the numbers of replications (out of 100) for selecting a particular dimension as the best model. Of special interest is how the patterns of selection results differ between the two criteria as k varies.

4. Results

In this section, we describe the results of our simulation studies. An important prerequisite before we can evaluate the performance for various model selection indices is whether parameters are recovered well by the trend vector model. We investigated parameter recovery, the results are described in Appendix B. Overall, the parameters are recovered well; therefore, there is no impediment to deal with the model selection issues.

4.1 Simulation Results for Dimensionality Determination

For each data set, 1-, 2-, and 3-dimensional trend vector models were fitted. Since methods based on confidence intervals are difficult to use for dimensionality determination, we focus on the performance of the quasi-likelihood based methods (i.e., LRT_Q, AIC_Q, and BIC_Q). The LRT_Q compares the difference between two nested models. The number of parameters for the 1-, 2-, and 3-dimensional trend vector models are 9, 15, and 18, respectively. The difference between the 1- and 2-dimensional model is 6 parameters, and the difference between the 2- and 3-dimensional model is 3. Based on the asymptotic Chi-square distribution, the critical values for the LRT_Q for choosing between the 1-dimensional versus 2-dimensional and between the 2-dimensional versus 3-dimensional models are 12.59 and 7.81, respectively. These values are used to determine the best dimensionality by using the LRT_Q method between pairs of models. The best dimensionality using AIC_Q and BIC_Q criteria is chosen with the lowest value among the 1-, 2-, and 3-dimensional models for each simulated sample.

Table 4 summarizes the ANOVA results for each criterion in each column with F-values and effect sizes for factors that have reached significance. The first column of Table 4 shows that LRT_Q performs better when classes are well-separated, but the performance was not affected by other factors. The performance of BIC_Q showed differences in four factors: BIC_Q performs better with larger sample sizes, symmetric regression weights, shorter time sequences, and well-separated classes. The performance of AIC_Q in identifying the correct dimensionality is not affected by any of the five design factors. By examining the pattern on the number of times each dimensionality was chosen with AIC_Q, we found that AIC_Q tends to choose models that are more complex than necessary.

Next, we compare across the three methods (LRT_Q, AIC_Q, and BIC_Q) for identifying correct dimensionality. An ANOVA is conducted with the three methods as the within factor, and the five design factors (i.e., class location, regression weight, correlation between predictors, number of time points, sample size) as the between factors. The F-statistic and effect size

Table 4. F-statistic and effect size of the ANOVA results for each LRT_Q, AIC_Q, and BIC_Q on dimensionality determination.

	LRT_Q	AIC_Q	BIC_Q
$\mathbf{T}$ (2)	-	-	4.63 ($\eta^2 = .03$)
$\mathbf{N}$ (3)	-	-	5.50 ($\eta^2 = .08$)
$\mathbf{B}$ (2)	-	-	4.95 ($\eta^2 = .04$)
$\mathbf{Z}$ (2)	11.87 ($\eta^2 = .10$)	-	25.79 ($\eta^2 = .19$)
$\mathbf{r}_{34}$ (4)	-	-	-

Table 5. ANOVA results for dimensionality determination. F-values, p-values, and effect size measures for each of the factors. Statistic is a within factor with levels AIC_Q, BIC_Q, and LRT_Q.

Factor	F-value	$p < .05$	η_p^2
Statistic	26.91	*	.21
$\mathbf{T}$ (2)	2.60		
$\mathbf{N}$ (3)	4.12	*	.07
$\mathbf{B}$ (2)	8.53	*	.07
$\mathbf{Z}$ (2)	16.01	*	.13
$\mathbf{r}_{34}$ (4)	0.23		
Statistic $\times\mathbf{T}$	2.79		
Statistic $\times\mathbf{N}$	1.18		
Statistic $\times\mathbf{B}$	0.03		
Statistic $\times\mathbf{Z}$	9.12	*	.07
Statistic $\times\mathbf{r}_{34}$	0.05		

Note. The symbol "*" indicates the criterion of $p < .05$ is met.

of each significant effect of this analysis are summarized in Table 5. The test shows that the performances of the three methods on dimensionality determination differ significantly. By examining the mean patterns of the three methods, it clearly shows that LRT_Q and BIC_Q consistently outperform AIC_Q. This result can be explained primarily by the more lenient criteria for choosing a model of higher dimensionity when the degrees of freedom is less than 8 (k=6 between 1-dimensional and 2-dimensional comparisons and k=3 between 2-dimensional and 3-dimensional comparisons).

For the between factors, the results also show that all three criteria perform better when (a) class locations are well-separated, (b) regression weights are more evenly distributed over the dimensions, and (c) for larger sample sizes. Moreover, there is an interaction between methods and class locations. The interaction shows that LRT_Q and BIC_Q perform better when class locations are well separated; however, the AIC_Q criterion is not affected by this factor and performs worst among the three methods.

Table 6. Number of times the LRT_Q turns out non-significant out of the 50 for given sets of omitted predictor variables (N=160 and T=3).

			Deleted predictor										
Z	**B**	r_{34}	X_1	X_2	X_3	X_4	X_5	(X_4, X_5)	(X_3, X_4)	(X_3, X_5)	(X_1, X_5)	(X_1, X_4)	(X_1, X_3)
1	1	0.0	0	0	0	50	50	48	0	0	0	0	0
1	1	0.2	0	0	0	49	48	50	0	0	0	0	0
1	1	0.5	0	0	0	49	49	49	0	0	0	0	0
1	1	0.8	0	0	0	48	49	47	0	0	0	0	0
1	2	0.0	0	0	0	48	49	47	0	0	0	0	0
1	2	0.2	0	0	0	46	47	49	0	0	0	0	0
1	2	0.5	0	0	0	46	48	50	0	0	0	0	0
1	2	0.8	0	0	0	49	48	48	0	0	0	0	0
2	1	0.0	0	0	0	48	47	47	1	1	1	1	1
2	1	0.2	0	0	0	48	46	46	0	0	0	0	0
2	1	0.5	0	0	0	48	49	48	0	0	0	0	0
2	1	0.8	0	0	8	47	49	47	0	10	0	0	0
2	2	0.0	0	0	0	48	49	49	0	0	1	1	0
2	2	0.2	0	0	0	47	48	47	0	0	1	2	0
2	2	0.5	0	0	0	50	50	50	0	0	0	0	0
2	2	0.8	0	0	1	49	48	49	0	0	2	3	0

Note. All other "deleted-two predictor" combinations not listed in the table demonstrate significance in the LRT_Q (i.e., the full model was retained).

4.2 Simulation Results for Variable Selection

4.2.1 Likelihood-Based Indices

The LRT_Q compares the quasi-likelihood between two nested models. For the purpose of variable selection, each predictor or a pair of predictors are left out from the model, and the quasi-likelihood is obtained by fitting a 2-dimensional trend vector model. We use the condition of $N = 160$ and three time points as an exemplary case to illustrate the findings. Table 6 lists the number of times that LRT_Q indicated a non-significant loss out of the 50 analyses when omitting certain predictor variables. Although there are 10 possible combinations of deleting two predictors from the five predictors, only patterns demonstrating non-significance (i.e., accepting the reduced model) of the LRT_Q are shown. As indicated in Table 6, the LRT_Q is quite successful in detecting that X_4 or X_5 are not needed in the model when testing is done by deleting a single variable at a time. The LRT_Q is also able to detect that both X_4 and X_5 can be dropped from the model simultaneously.

Next we turn to variable selections using information criteria. The AIC_Q and BIC_Q are calculated for every combination of deleting none,

Table 7. Number of times out of 50 for which the model without explanatory variables (X_4, X_5) is selected to be the best model using AIC_Q and BIC_Q.

				AIC_Q			BIC_Q		
T	**Z**	**B**	r_{34}	N=40	N=80	N=160	N=40	N=80	N=160
3	1	1	0.0	24	22	26	49	50	50
3	1	1	0.2	17	20	26	47	50	50
3	1	1	0.5	24	19	20	49	50	50
3	1	1	0.8	20	23	25	41	45	49
3	1	2	0.0	16	8	21	49	44	49
3	1	2	0.2	14	12	25	49	40	50
3	1	2	0.5	20	11	19	48	40	50
3	1	2	0.8	21	10	26	49	41	50
3	2	1	0.0	19	15	21	38	48	50
3	2	1	0.2	20	23	24	40	50	49
3	2	1	0.5	19	20	25	43	50	50
3	2	1	0.8	21	16	26	38	45	47
3	2	2	0.0	15	21	26	35	43	50
3	2	2	0.2	16	21	17	48	45	48
3	2	2	0.5	29	26	17	50	47	50
3	2	2	0.8	24	12	19	48	44	48
5	1	1	0.0	19	26	30	25	32	38
5	1	1	0.2	20	28	27	24	33	39
5	1	1	0.5	15	29	32	16	32	43
5	1	1	0.8	22	23	31	22	28	41
5	1	2	0.0	27	32	34	41	50	50
5	1	2	0.2	24	37	29	39	50	50
5	1	2	0.5	27	33	33	39	49	50
5	1	2	0.8	26	29	37	38	45	48
5	2	1	0.0	18	28	34	30	41	48
5	2	1	0.2	19	22	25	26	40	49
5	2	1	0.5	21	34	26	26	44	47
5	2	1	0.8	14	29	29	21	38	43
5	2	2	0.0	33	39	31	42	50	50
5	2	2	0.2	33	35	34	47	50	50
5	2	2	0.5	29	30	34	38	49	50
5	2	2	0.8	28	26	33	37	43	50

one, or two predictors from the model. We rank all 16 possible patterns of combination for different sample sizes based on the AIC_Q and BIC_Q. Table 7 presents the number of times that the model with both X_4 and X_5 dropped (i.e., having only X_1, X_2, and X_3 in the model) is selected to be the best model for different sample sizes. As shown in Table 7, BIC_Q performs much better than AIC_Q to select the true model and both criteria tend to perform better with larger sample sizes.

If we treated indices AIC_Q and BIC_Q as the within factor, and the patterns of regression weights, class locations, correlation patterns among

Table 8. The number of times the joint confidence interval for X_4/X_5 includes the origin for six different methods (N=160 and T=5).

Z	**B**	r_{34}	BSn	BSp	$Jackknife$	BCa	$Sand$	$Sand_{BC}$
1	1	0.0	45/41	46/42	46/43	47/46	38/31	46/44
1	1	0.2	47/41	47/40	47/42	47/41	34/33	45/40
1	1	0.5	48/44	47/45	48/46	49/45	39/27	46/42
1	1	0.8	46/42	45/44	48/43	46/45	34/35	46/45
1	2	0.0	45/45	45/45	45/46	45/46	31/27	43/46
1	2	0.2	44/43	46/43	45/41	45/43	28/27	44/40
1	2	0.5	42/45	43/44	43/45	42/44	24/32	42/39
1	2	0.8	46/40	46/40	46/40	46/40	31/25	44/38
2	1	0.0	46/44	47/44	46/45	47/44	36/34	46/45
2	1	0.2	46/42	45/43	46/43	45/43	38/32	47/43
2	1	0.5	44/46	43/44	44/48	44/44	36/31	44/45
2	1	0.8	45/45	44/45	44/45	45/46	38/38	46/46
2	2	0.0	45/43	45/42	45/44	45/45	28/30	45/42
2	2	0.2	48/47	46/45	48/47	47/45	31/39	45/44
2	2	0.5	48/44	47/44	48/45	49/45	35/38	45/45
2	2	0.8	45/46	46/46	46/47	46/46	30/33	42/45

variables, number of time points, and sample size as between factors, the ANOVA test indicated that AIC_Q and BIC_Q significantly differ in terms of the ability to identify the correct model ($F(1, 87) = 1687, p < .05, \eta^2 = 0.83$). The regression weight ($F(1, 87) = 11.40, p < .05, \eta^2 = 0.07$) and sample sizes ($F(2, 87) = 27.12, p < .05, \eta^2 = 0.35$) have significant effects, that is, indices perform better when regression weights are more evenly distributed and when sample sizes are larger. The two indices do not differ in the control factors since there is no interaction present in the analysis.

4.2.2 Confidence Intervals

We compare six different methods that are based on forming confidence intervals around the parameter of interest. Since the model is a two-dimensional model, we counted the number of times out of 50 that the joint confidence intervals of the regression weights for a specific variable includes the origin. The results for X_4 and X_5 with $N = 160$ and five time points are listed in Table 8 to show a general impression. It is clear that indices based on bootstrapping performed much better than those based on the regular sandwich covariance estimator. This pattern holds consistently for both explanatory variables under all conditions. As can be seen from the last two columns of Table 8, the bias-corrected sandwich covariance estimator shows a clear improvement across all conditions.

Table 9. ANOVA results for variable selection of X_4, F-values, p-values, and effect size measures for each of the factors. Statistic is a within factor with levels BS_n, BS_p, $Jackknife$, BCa, $Sand$, $Sand_{BC}$.

Factor	F-value	$p < .05$	η_p^2
Statistic	892.62	*	.88
T (2)	0.30		
N (3)	4.18	*	.08
B (2)	1.31		
Z (2)	0.18		
$\mathbf{r}_{34}$ (4)	0.91		
Statistic $\times$**T**	3.86	*	.00
Statistic $\times$**N**	10.22	*	.02
Statistic $\times$**B**	0.64		
Statistic $\times$**Z**	3.10	*	.00
Statistic $\times\mathbf{r}_{34}$	0.89		

The results of using the six indices as the within factor and the five design factors as the between factors on X_4 are summarized in Table 9. The six indices differ significantly, it is mainly due to the fact that the sandwich covariance estimator performs considerably worse than the other methods. Larger sample sizes also improve the performance of these indices. These indices also show differences in the levels of these factors: number of time points, sample sizes, and the class locations; however, the effect sizes indicate that these differences are quite small. The mean patterns indicated that the method of the bias-corrected sandwich estimator performed worse than other methods when sample size is small (N=40). In other words, the improvement of performance through the bias-corrected procedure on the sandwich estimator is relatively less for small sample sizes.

To compare the predictors X_4 and X_5, we ran an ANOVA using the predictors (X_4 and X_5) and the six indices as the within factor, and the five design factors as the between factor. The test showed that the six indices differ significantly ($F(5, 435) = 1463.48, p < .05, \eta^2 = 0.95$), but there is no significant difference between the two predictor variables X_4 and X_5.

4.3 Simulation Results for the Special Study on Dimensionality

The special study concerns how the performances of AIC_Q and LRT_Q dimensionality selection may be affected by the associated degrees of freedom. In the preceding study of dimensionality determination, the degrees of freedom (k) are 6 and 3, respectively, when comparing a 1-dimensional with a 2-dimensional model and when comparing a 2-dimensional and a 3-dimensional model. In this special study, simulation design covers a range of

Table 10. Results for a special study to compare AIC_Q and LRT_Q in dimensionality selection.

		AIC_Q			BIC_Q†			# df diff		LRT_Q		
G	p	M1	M2	M3	M1	M2	M3	M2-M1	M3-M2	M1	M2	M3
	4	17	83	na	68	32	na	6	na	17	83	na
3	5	0	100	na	11	89	na	7	na	0	100	na
	6	0	100	na	0	100	na	**8**	na	0	100	na
	7	0	100	na	0	100	na	9	na	0	100	na
	4	2	97	1	28	72	0	7	6	2	97	1
4	5	0	99	1	1	99	0	**8**	7	0	99	1
	6	0	99	1	0	100	0	9	**8**	0	99	1
	7	0	98	2	0	100	0	10	9	0	98	2
	4	0	98	2	27	73	0	**8**	7	0	98	2
5	5	0	98	2	3	97	0	9	**8**	0	98	2
	6	0	98	2	0	100	0	10	9	0	96	4
	7	0	97	3	0	100	0	11	10	0	97	3
	4	0	100	0	34	66	0	9	**8**	0	100	0
6	5	0	96	4	6	94	0	10	9	0	96	4
	6	0	97	3	0	100	0	11	10	0	92	8
	7	0	95	5	0	100	0	12	11	0	92	8

Note. Conditions are in boldface when the degrees of freedom between two models equals 8. For G=3, the maximum dimension fitted is 2 and model selection between dimensionality of 2 and 3 is not applicable (na).
† Although this simulation study concerns the performance reversal for AIC_Q and LRT_Q, the results of BIC_Q are also included in the table for references.

degrees of freedom (from 6 to 12). As explained earlier (see the description of study design), the two criteria differ in that AIC_Q is more liberal in favor of a higher dimensional model for k less than 8, while more conservative for k equal to or greater than 8.

Table 10 presents the results of this special study, showing the number of times (out of 100) a model of specific dimensionality is selected as the best fit. Data in Table 10 suggest that the AIC_Q and LRT_Q are quite comparable in successfully identifying the true dimension as the best fit. Except for the case with 3 classes and 4 predictors, both criteria attain a successful rate of over 92 percent. Small differences between them are found for a few cases: $G = 5$, $p = 6$; $G = 6$, $p = 6$; and $G = 6$, $p = 7$. In these cases, LRT_Q selects the 3-dimensional model as best slightly more often than AIC_Q (respectively, 4 versus 2, 8 versus 3, and 8 versus 5 for the three cases noted above). The results for these three cases represent a reversal of that in the first study, but are consistent with the change in direction of the differences between the two criteria when k exceeds 8. Indeed, the small differences in correct identification of the true dimensionality are expected in light of the small probability that the asymptotic Chi-square distribution

covers the interval between the two criteria. For instance, the probability of a Chi-square distribution with 9 degrees of freedom between 16.92 ($= \chi^2_{9,.05}$) and 18 ($= 2k = \chi^2_{9,.035}$) is approximately .015, implying that about 1 or 2 out of 100 cases may be expected to have a reversal in deciding between a 2-dimensional and a 3-dimensional model when k =9 (as for the case with G=5, p=6). As shown in the table, the empirical result indicates a reversal for 2 more cases where a 3-dimensional model is selected as best by LRT$_Q$.

For the interest of readers, Table 10 also presents the results for BIC_Q. Data in Table 10 show that BIC_Q is the most conservative in selecting models with lower dimensionality as best. Because BIC_Q employs a factor of $\ln(N)$ in penalizing models with larger numbers of parameters, it tends to favor smaller models when N increases. Note that $N = 80$ in the current simulations, and $\ln(80) = 4.38$. As expected, Table 10 suggests that BIC_Q is less successful in identifying the true dimension ($M = 2$) by choosing dimensionality of one as the best model in many more cases as compared to AIC_Q and LRT$_Q$.

5. Discussion and Conclusion

In this paper, we report the results of three simulation studies to evaluate the performances of model selection criteria based on the quasi-likelihood function for fitting the trend vector model to longitudinal multinomial data. Three criteria (LRT$_Q$, AIC_Q, and BIC_Q) are employed for dimensionality determination and six additional criteria based on confidence intervals (BS_n, BS_p, $Jackknife$, BCa, $Sand$, $Sand_{BC}$) are used for variable selection. We compared the performances of these criteria on identifying the true dimensionality and selecting the important explanatory variables. To support effective evaluations of the model selection criteria, we also examined parameter recovery to ensure the accuracies of estimates when fitting the generating model to the simulated data.

It is generally recommended to first determine the dimensionality and then further look into the issue of variable selection. For determination of the dimensionality, we compared the LRT$_Q$, AIC_Q, and BIC_Q. Both the LRT$_Q$ and BIC_Q gave satisfactory results. For the simulations in the first study, the AIC_Q criterion has a moderate tendency to choose more complex models. In other words, the penalty for model complexity in AIC_Q is not sufficient when models with small degrees of freedom are involved. However, the special study focused on dimensionality selection across models involving a broad range of degrees of freedom showed that AIC_Q performed well, largely comparable to LRT$_Q$ (see Table 10).

The performance of the LRT$_Q$ and BIC_Q in determining the dimensionality is affected by the factor of class locations in that both statistics

performed better when classes are well-separated. Comparing to BIC_Q, the performance of LRT_Q appears to be similar across variations of the other four controlled factors. In contrast, Table 4 shows that four of the control factors (except the correlation between an important explanatory variable and a noise variable) have significant effects on the performance of BIC_Q.

In summary, LRT_Q and BIC_Q can be used successfully to determine the appropriate dimensionality for the trend vector model even when the model is relatively simple (e.g., less than 4 classes and up to 3 relevant explanatory variables). These two criteria are based on quasi-likelihood functions which ignore dependencies between observations. Thus they are not strictly likelihood-based criteria and will approach the asymptotic Chi-square distribution much slower. Nevertheless, the current study finds them effective in recovering the true dimensionality of the model underlying the simulated data. Additional simulations suggest that AIC_Q (also based on quasi-likelihood functions) works very well in dimensionality determination for models involving larger degrees of freedom (or more complex models).

For variable selection, the performances of nine criteria are studied with empirical simulations. The criteria include 3 based on quasi-likelihood functions, 4 based on joint confidence intervals as obtained using resampling methods, and 2 based on sandwich estimators. Overall, we found that most of the methods perform well, with the exceptions of AIC_Q and sandwich estimators. The three likelihood-based methods gave satisfactory performances for variable selection. The LRT_Q statistic works remarkably well in identifying the important predictor variables. It should be noted that we did not investigate whether the LRT_Q statistic for the trend vector model is asymptotically distributed as a Chi-square distribution. We believe that such asymptotic results, although theoretically interesting, may have little value in practice.

Pan and Le (2001) showed that the Bayesian Information Criterion (BIC) performed well for correlated data under a similar estimation scheme. The results of the simulation in this study confirm the finding of Pan and Le (2001), and establish the usage of the BIC index for variable selection in trend vector models. The simulations in this study offered some evidence of using likelihood-like methods even for quasi-likelihood methods. Overall, when the regression weights are distributed more evenly on both dimensions, the methods based on information criteria (e.g., AIC_Q and BIC_Q) perform better.

The indices based on bootstrapping generally gave satisfactory results. These indices in general are able to identify the noise variables in the trend vector model. The bootstrap indices, however, can be quite time consuming. For example, for the simulations performed in this study, the bootstrap took about 3-5 days of computing time with 50 replications across 16

conditions even for sample size of 40. Nevertheless, the bootstrap methods perform well in the simulation studies. Therefore, in addition to likelihood-based methods, the bootstrapping methods for variable selection can be recommended for trend vector models.

Our simulation study showed that using the sandwich estimator to obtain the confidence intervals of the parameters for variable selection is not appropriate. In this study, the methods based on sandwich covariance estimators consistently performed very poorly across different conditions. The literature has warned us about the risk of underestimation (i.e., the estimated variance of the parameters is too small) when the sample size is smaller than 40. However, in our simulation, even with median sample sizes ($N = 80$ and 160) the performance of the sandwich estimator is not satisfactory.

The confidence intervals based on bias-corrected sandwich covariance estimator demonstrates significant improvement. With the bias-correction, the sandwich covariance estimator can reach satisfactory levels for variable selection in trend vector models. There have been many papers dealing with this methodology after the seminal paper by Liang et Zeger (1986) where they propose the so-called GEE (generalized estimating equations) methodology. The straightforward application of the sandwich estimators for the standard errors needs to proceed with caution. The simulations in this paper show that even with median to large sample size, bias-correction of the sandwich estimator is necessary.

To conclude, we recommend the use of LRT_Q or BIC_Q for dimensionality determination and variable selection in trend vector models. These simple methods give reliable results compared to AIC_Q and uncorrected sandwich estimators. Bootstrap methods can also be used, however, they require much more computational effort. An important message of this paper is that one should be aware of the underestimation of sandwich covariance estimators, the bias-correction procedure is needed even for large sample sizes.

Appendix A
Derivation of the Partial Gradient Matrix

We will illustrate the calculation of $\mathbf{D}_i$ using a 2-dimensional trend vector model with 5 predictor variables and a multinomial response with 3 levels recorded at 3 different time points.

Liang and Zeger (1986) proposed a sandwich estimator to estimate the covariance matrix of the regression weights in generalized linear models. Using a similar derivation for the trend vector model the sandwich estimator is

$$\left(\sum_{i=1}^{n}\mathbf{D}_i^T\mathbf{V}_i^{-1}\mathbf{D}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{D}_i^T\mathbf{V}_i^{-1}\text{cov}(\mathbf{y}_i)\mathbf{V}_i^{-1}\mathbf{D}_i\right)\left(\sum_{i=1}^{n}\mathbf{D}_i^T\mathbf{V}_i^{-1}\mathbf{D}_i\right)^{-1},$$

where

$$\mathbf{D}_i = \frac{\partial[\pi_{11}(\mathbf{x}_{i1}) \;\; \pi_{21}(\mathbf{x}_{i1}) \;\; \pi_{12}(\mathbf{x}_{i2}) \;\; \pi_{22}(\mathbf{x}_{i2}) \;\; \pi_{13}(\mathbf{x}_{i3}) \;\; \pi_{23}(\mathbf{x}_{i3})]}{\partial \boldsymbol{\theta}}.$$

Due to the constraint that $\sum_{j=1}^{G} \pi_{jt}(\mathbf{x}_{it}) = 1$, we only need $G-1$ classes in the calculation. The trend vector model is defined as

$$\pi_{jt}(\mathbf{x}_{it}) = \frac{\exp\left(-\sum_{m=1}^{M}(y_{itm} - z_{jm})^2\right)}{\sum_k \exp\left(-\sum_{m=1}^{M}(y_{itm} - z_{km})^2\right)} = \frac{u}{v},$$

with

$$\begin{aligned} u &= \exp\left(-\sum_{m=1}^{M}(y_{itm} - z_{jm})^2\right), \\ v &= \exp\left(-\sum_{m=1}^{M}(y_{itm} - z_{1m})^2\right) + \exp\left(-\sum_{m=1}^{M}(y_{itm} - z_{2m})^2\right) \\ &\quad + \exp\left(-\sum_{m=1}^{M}(y_{itm} - z_{3m})^2\right). \end{aligned}$$

We split up the calculation for $\mathbf{b}$ (a vectorized form of $\mathbf{B}$) and $\mathbf{z}$ (a vectorized form of $\mathbf{Z}$). The derivatives of u and v with respect to b_{qa}, the regression weight for variable q on dimension a, are

$$\begin{aligned} u' &= \exp\left(-\sum_{m=1}^{M}(y_{itm} - z_{jm})^2\right) \times (-2(y_{ita} - z_{ja})x_{itq}) \\ v' &= \exp\left(-\sum_{m=1}^{M}(y_{itm} - z_{1m})^2\right) \times (-2(y_{ita} - z_{1a})x_{itq}) \\ &\quad + \exp\left(-\sum_{m=1}^{M}(y_{itm} - z_{2m})^2\right) \times (-2(y_{ita} - z_{2a})x_{itq}) \\ &\quad + \exp\left(-\sum_{m=1}^{M}(y_{itm} - z_{3m})^2\right) \times (-2(y_{ita} - z_{3a})x_{itq}). \end{aligned}$$

Using

$$\frac{\partial \pi_{jt}(\mathbf{x}_{it})}{\partial b_{qa}} = \frac{u' \times v - u \times v'}{v^2}$$

and filling in gives the first part of $\mathbf{D}_i$.

The derivatives of u and v with respect to z_{ha}, the coordinate for class h on dimension a, are

$$u' = \exp\left(-\sum_{m=1}^{M}(y_{itm}-z_{hm})^2\right)\times(2(y_{ita}-z_{ha})).$$
$$v' = \exp\left(-\sum_{m=1}^{M}(y_{itm}-z_{hm})^2\right)\times(2(y_{ita}-z_{ha})).$$

Again using

$$\frac{\partial\pi_{jt}(\mathbf{x}_{it})}{\partial z_{ha}} = \frac{u'\times v - u\times v'}{v^2},$$

gives the second part of $\mathbf{D}_i$. Vertically concatenating the two parts gives the matrix $\mathbf{D}_i$.

Appendix B
Parameter Recovery

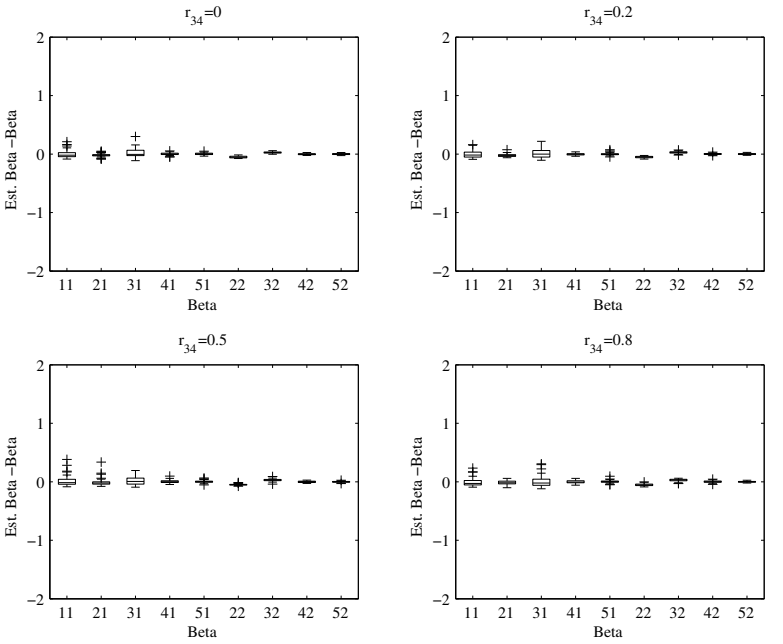
Parameter recovery is investigated by evaluating the difference between the estimated and true values, and the estimation stability is evaluated using the standard error of the differences.

The trend vector model is a marginal model or population averaged model (PA). However, the data were generated from a random-effect model (RE) in order to create within subject dependence. In order to compare the recovery of parameters, we used the relationship of the parameters based on these two approaches showed by Neuhaus (1993). He showed that

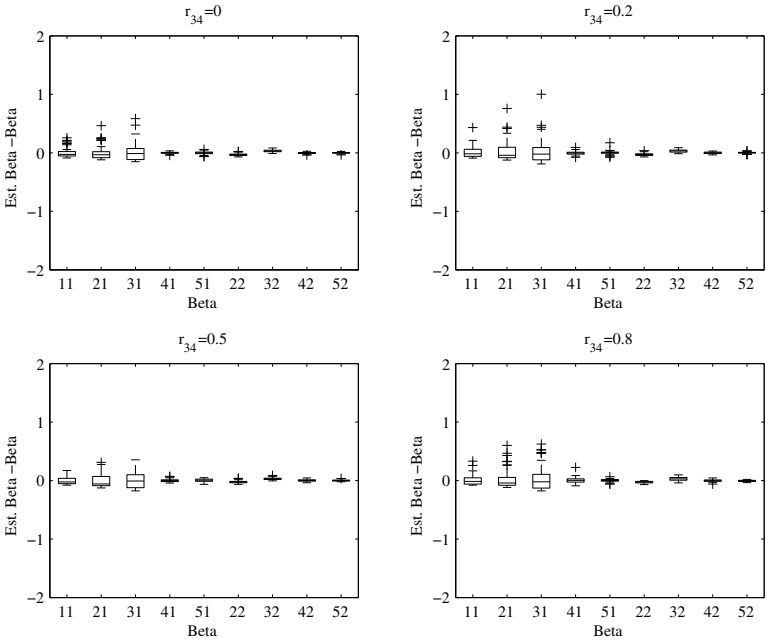
$$\theta_{PA} \approx \theta_{RE}[1-\rho_f], \tag{20}$$

where $\rho_f = \text{corr}(f_{it}, f_{it'})$ is the intra-subject correlation. For multinomial data, the variance is $\pi^2/3$, and the random effects in our data generation are specified to have variance 1, so that the intra-subject correlation equals $\rho_f = 1/[1+(\pi^2/3)] = 0.233$ (Hedeker and Gibbons 2006). Therefore, the estimates are expected to be smaller than the specified value by a factor of 0.767.

Figure B1 plots the difference between the estimated regression weights ($\mathbf{B}$) and the adjusted true values for sample size of $N = 160$ and Figure B2 presents the boxplots for the performance of the parameter recovery for class location ($\mathbf{Z}$). In general, the parameters are recovered well.

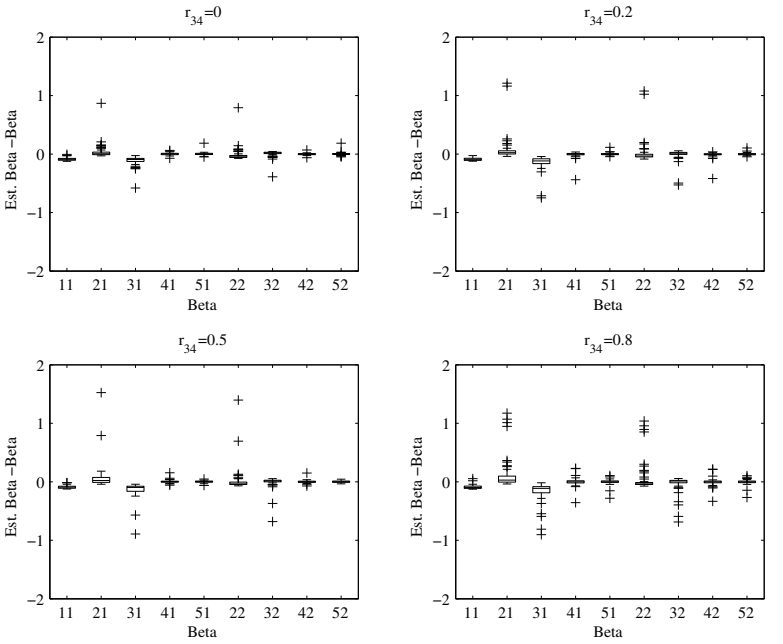


(a) Z=1,B=1

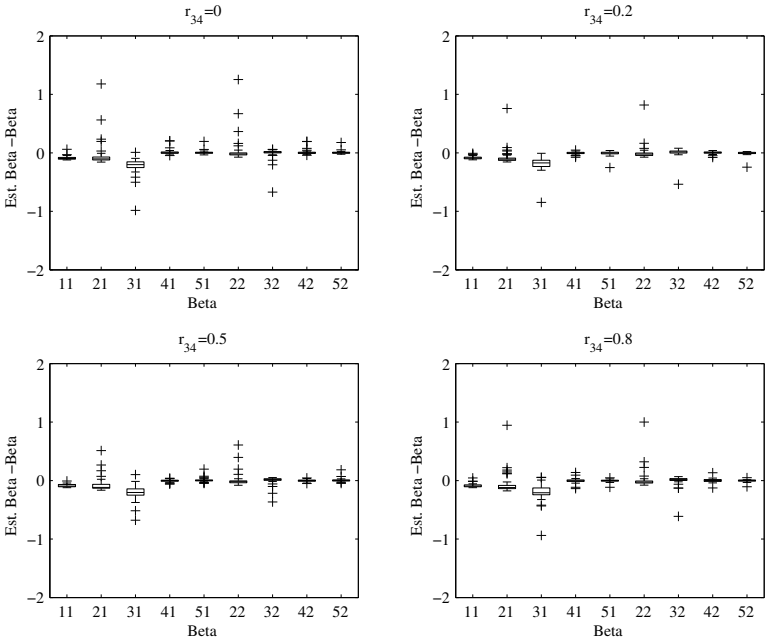


(b) Z=1,B=2

Figure B1. Part 1. Parameter recovery (β) N=160

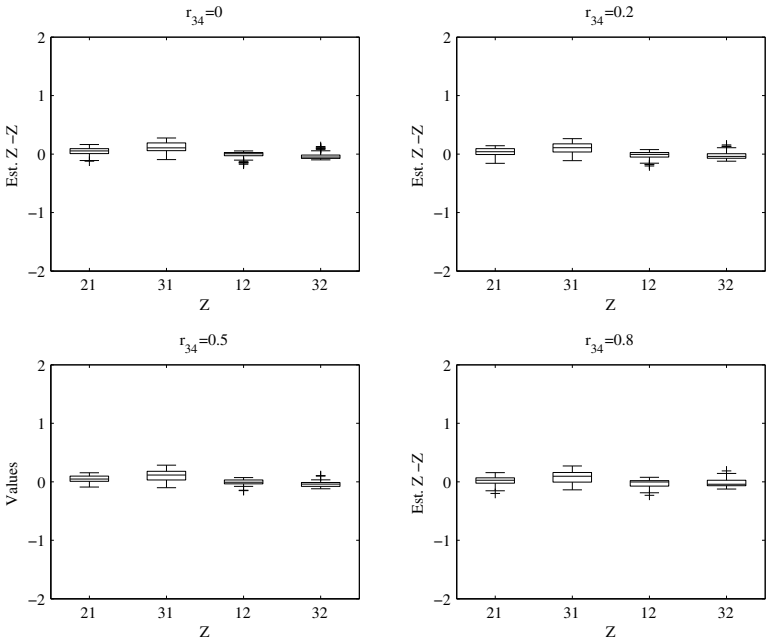


(c) Z=2,B=1

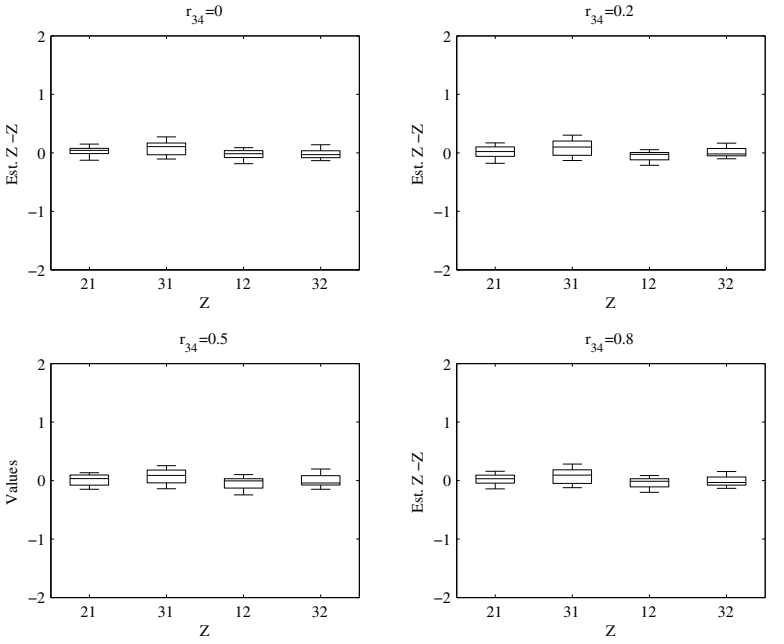


(d) Z=2,B=2

Figure B1. Part 2. Parameter recovery (β) N=160

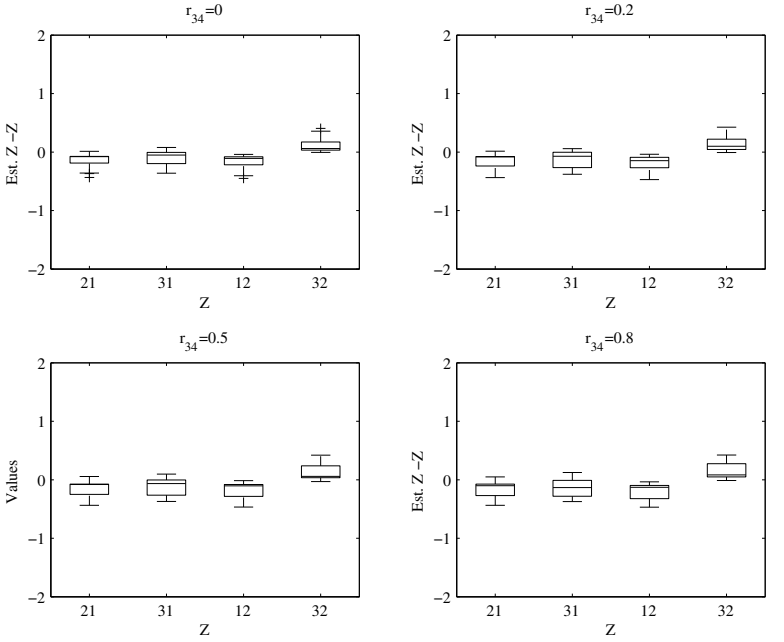


(a) Z=1,B=1

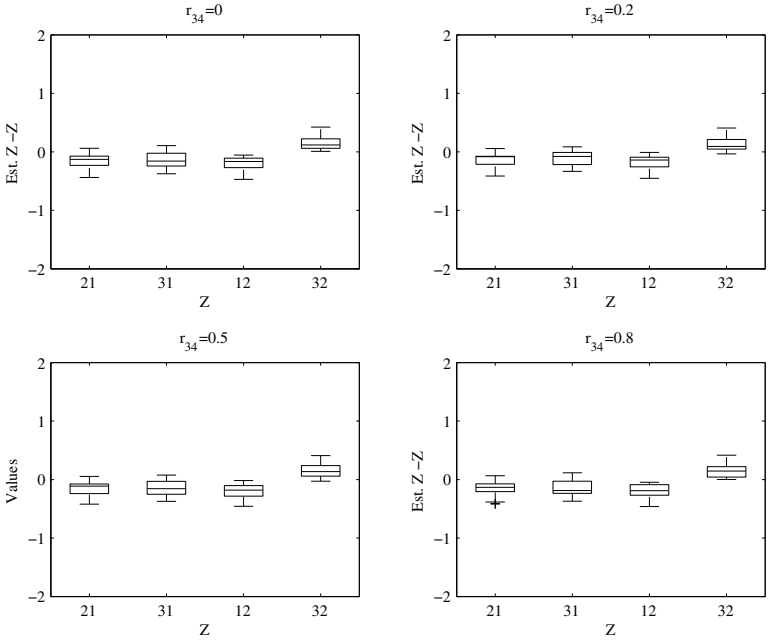


(b) Z=1,B=2

Figure B2. Part 1. Parameter recovery (**Z**) N=160



(c) Z=2,B=1



(d) Z=2,B=2

Figure B2. Part 2. Parameter recovery (**Z**) N=160

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.

AKAIKE, H. (1974), "A New Look at the Statistical Model Identification", *IEEE Transactions on Automatic Control, 19(6)*, 716–723.

DE ROOIJ, M. (2009), "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., and SCHOUTEDEN, M. (2012), "The Mixed Effects Trend Vector Model", *Multivariate Behavioral Research, 47,* 635–664.

EFRON, B., and TIBSHIRANI, R.J. (1993), *An Introduction to the Bootstrap*, New York: Chapman and Hall.

FAY, M.P., and GRAUBARD, B.I. (2001), "Small-Sample Adjustments for Wald-Type Tests Using Sandwich Estimators", *Biometrics, 57,* 1198–1206.

HEDEKER, D., and GIBBONS, R.D. (2006), *Longitudinal Data Analysis*, New York: John Wiley & Sons.

KAUERMANN, G., CARROLL, R.J. (2001), "A Note on the Efficiency of Sandwich Covariance Matrix Estimation", *Journal of the American Statistical Association, 96 (456)*, 1387–1398.

LIANG, K.Y., and ZEGER, S.L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models", *Biometrika, 73,* 13–22.

LIPSITZ, S.R., KIM, K., and ZHAO, L. (1994), "Analysis of Repeated Categorical Data Using Generalized Estimating Equations", *Statistics in Medicine, 13,* 1149–1163.

MANCL, L.A., and DEROUEN, T.A. (2001), "A Covariance Estimator for GEE with Improved Small-Sample Properties", *Biometrics, 57,* 126–134.

MOLENBERGHS, G., and VERBEKE, G. (2005), *Models for Discrete Longitudinal Data,* New York: Springer.

NEUHAUS, J.M. (1993), "Estimation Efficiency and Tests of Covariate Effects with Clustered Binary Data", *Biometrics, 49,* 989–996.

PAN, W. (2001), "Akaike's Information Criterion in Generalized Estimating Equations", *Biometrics, 57,* 120–125.

PAN, W., and LE, C.T. (2001), "Bootstrap Model Selection in Generalized Linear Models", *Journal of Agricultural, Biological, and Environmental Statistics, 6,* 49–61.

PAN, W., and WALL, M.W. (2002), "Small-Sample Adjustments in Using the Sandwich Variance Estimator in Generalized Estimating Equations", *Statistics in Medicine, 21,* 1429–1441.

PREISSER, J.S., and QAQISH, B.F. (1996), "Deletion Diagnostics for Generalized Estimating Equations, *Biometrika, 83,* 551–562.

SCHWARZ, G. (1978), "Estimating the Dimensions of a Model", *Annals of Statistics, 6,* 461–464.

SHERMAN, M., and LE CESSIE, S. (1997), "A Comparison Between Bootstrap Methods and Generalized Estimating Equations for Correlated Outcomes in Generalized Linear Models", *Communications in Statistics-Simulation and Computation, 26,* 901–925.