Longitudinal mixed-effects models for latent cognitive function

Ardo van den Hout
Department of Statistical Science, University College London, UK
E-mail: ardo.vandenhout@ucl.ac.uk

Jean-Paul Fox
Department of Research Methodology, Measurement and Data Analysis
Twente University, Netherlands

Graciela Muniz-Terrera
Medical Research Council Unit for Lifelong Health and Ageing, London, UK

Abstract

A mixed-effects regression model with a bent-cable change-point predictor is formulated to describe potential decline of cognitive function over time in the older population. For the individual trajectories, cognitive function is considered to be a latent variable measured through an item response theory model given longitudinal test data. Individual-specific parameters are defined for both cognitive function and the rate of change over time, using the change-point predictor for non-linear trends. Bayesian inference is used, where the Deviance Information Criterion and the L-criterion are investigated for model comparison. Special attention is given to the identifiability of the item response parameters. Item response theory makes it possible to use dichotomous and polytomous test items, and to take into account missing data and survey-design change during follow-up. This will be illustrated in an application where data stem from the Cambridge City over-75s Cohort Study.

Key words: bent-cable; change point; cognition; growth-curve model; item-response theory; longitudinal data analysis.

1 Introduction

Of interest is cognitive function in the older population in the years before death. Potential decline and the possibility of a one-off change in the trend of the decline (Riegel and Riegel, 1972) will be investigated using mixed-effects regression models for longitudinal data, where the response variable is latent cognitive function and the predictors are linear or non-linear.

Data stem from the Cambridge City over-75s Cohort Study (CC75C), where cognitive function is measured using a modified version of the Mini-Mental State Examination (MMSE, see Folstein *et al,* 1975). The MMSE is often used as a screening for dementia or mild cognitive impairment. Holling *et al.* (2012) show that the examination has a better diagnostic accuracy for dementia than for mild cognitive impairment. The MMSE consists of a questionnaire with dichotomous and polytomous items. The distribution of the integer sum score is skewed since most of the items are not difficult for an individual with normal cognition, and many individuals score close to the upper bound of the scale. The discreteness of the sum score and the skewness of its distribution means that a regression model with the sum score as response variable is problematic when the conditional distribution for the response is assumed to be normal.

In this paper, the response variable in the regression models is a latent continuous variable which explains how well individuals score in the examination. The link between the latent variable and the longitudinal scores on the individual items in the questionnaire is described by an Item Response Theory (IRT) model (Van der Linden and Hambleton, 1997). Hence, the latent response is interpreted as the underlying cognitive function which explains observed cognitive performance.

A one-off change in the trend of cognitive decline cannot be properly modeled using a linear predictor, which implies a constant rate of decline. For this reason, a change-point predictor will be used to describe the change of cognitive function in the older population. The most basic change-point model is the broken-stick model. The broken-stick model implies a non-linear predictor such that there are two linear parts (with different slopes) that intersect at the change point. For the models in the current work, we will use a smooth version of the broken-stick model, which is called the bent-cable model (Chiu *et al,* 2006; Van den Hout *et al,* 2013). Change-point models have been used in various applications, e.g., in medical statistics (Stasinopoulos and Rigby, 1992), in demography (Cohen 2008), and in transport (Lévy-Leduc and Roueff, 2009).

Bayesian inference will be applied using Markov chain Monte-Carlo (MCMC) techniques. The Deviance Information Criterion (DIC, Spiegelhalter *et al,* 2002) and

the L-criterion (Laud and Ibrahim, 1995; Gelfand and Gosh, 1998) are investigated with respect to model comparison.

The combination of an IRT model and a change-point regression model has not been investigated before but seems promising in scope. In modeling latent cognitive function, the change-point regression model makes it possible to evaluate within-subject change in growth rates given latent cognitive function as an outcome variable. The common assumption of constant growth rates in linear models may not be realistic with respect to cognitive function in the years before death. The IRT model is formulated for longitudinal question-specific data without relying on less informative aggregate data information such as sum scores. Hence, IRT acknowledges that different items have different characteristics in terms of difficulty and discriminatory effect. A further advantage of IRT is that missing data can be dealt with at the level of the individual questions. There are six waves in our application, and data on a selection of the dichotomous items are collapsed into polytomous data in the first three waves but are available in the last three waves. Because of this design, data are missing by design and hence *missing at random* (Rubin 1976). It will be shown that using IRT makes it relatively easy to deal with missing data due to a change of design.

Our work is building upon Bayesian inference for IRT as presented in Johnson and Albert (1999) and Fox (2010). Longitudinal IRT models which include regression models with linear predictors have been discussed in a Bayesian framework by Douglas (1999), Fox and Glas (2001), and Klein-Entink *et al.* (2011). Special attention will be given to the way the longitudinal IRT model is identified, exploring both restrictions on the item parameters, and restrictions on the scale of the latent variable.

With respect to the change-point modeling of questionnaire data, the current paper aims to extend the work in Van den Hout *et al.* (2011), where the assumption of the normal distribution for the conditional (manifest) response variable may not always be the optimal choice in practice. See also Jacqmin-Gadda *et al.* (2006), who used a change-point model with the normal distribution for a test score as response. In addition to handling non-normality of the response, the current paper extends the modeling by combining an IRT measurement model with a structural model for latent cognitive function. Instead of using a fixed measure of cognitive function, response pattern information will be used.

The time scale in the analysis of the CC75C data is rather specific. The majority of the participants in CC75C have passed away since the start of the study in 1985. By ignoring the data from the small group of survivors, it is possible to use years-to-death as the time scale. The presented methodology, however, is general and can

also be used in longitudinal models with different time scales.

In Section 2, a brief summary of the CC75C data is given. Section 3 and 4 present the models and the Bayesian inference, respectively. Section 5 discusses the data analysis after investigating choices of parameter restrictions. Section 6 is the conclusion.

2 Cambridge City over-75s Cohort Study

The Cambridge City over-75s Cohort Study (CC75C, `www.cc75c.group.cam.ac.uk`) is a UK population-based longitudinal study of aging that started in 1985 with participants aged at least 75 years old in Cambridge city. Topics in the study are, e.g., cognitive decline and dementia, patterns of cognitive change, depression and depressive symptoms, socio-demographics and social contacts, falls and functional ability, and genetics. Here we focus on the measuring of cognitive function using a modified version of the Mini Mental State Examination (MMSE). In the examination there are items on, for instance, orientation ("What day of the week is it?", "What floor of the building are we on?"), on recognizing objects ("What is this called?"), and on memory ("Can you tell me what were the objects in the colored pictures I showed you a little while ago?").

Due to the long follow-up and advanced age at baseline, almost all of the participants passed away since the start of the study. After baseline (wave 1), further interviews were conducted on average 2, 7, 9, 12, 17, and 21 years later. The sample size of wave 1 is 2165 individuals. There is a large dropout between wave 1 and 2. Because dealing with the complex mechanism behind this dropout is outside the scope of the methods in the present paper, we will only use data from the 1204 individuals in wave 2 up to 7. However, for 25 of these 1204 individuals a death time is not available. The data of these survivors will be ignored. The resulting data contains observations from 1179 individuals. This sample includes the data from 40 individuals who were observed at wave 1 and have follow-up data from wave 3 onwards only. There is also intermittent missing data in the follow-up of the individuals who are observed at wave 2. This kind of missingness is very common in longitudinal data. We assume that the mechanism for the intermittent missing data is missing at random. Specifying the growth-curve model as a random-effects model, see next section, should provide some robustness against possible violation of this missing-at-random assumption (Verbeke and Molenbergh 2000). The frequencies for the number of times individuals are observed before death are 507, 320, 195, 112, 38, and 7, for one up to six times, respectively.

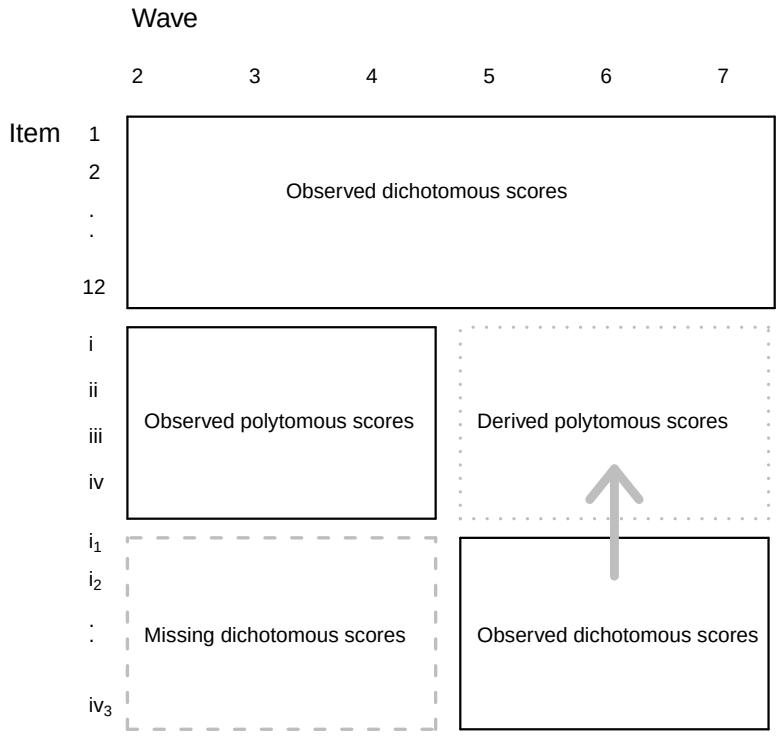

Figure 1: Availability of the item scores in CC75C.

For waves 2, 3, and 4, the data contain the answers to twelve dichotomous items and four polytomous items. The latter four items are actually the result of summarizing scores for dichotomous items. For waves 2, 3, and 4, the scores of these underlying dichotomous items are not available. However, for waves 5, 6, and 7, these scores are available. The maximum for the total sum is 23: maximum for the some of twelve dichotomous items is 12, and the maxima for the polytomous item are 2, 3, 3, and 3, respectively. The diagram in Figure 1 illustrates the availability of the item scores in the data set. As will be shown, the change in information across waves can be accounted for in the IRT modeling. More information on the individual items will be given in the section with the data analysis.

3 Models

Latent cognitive function is described by regression models where random effects are used to take into account dependencies between observations within an individual. This kind of models for longitudinal data are sometimes called *latent growth models*.

Let the latent variable be given by $\boldsymbol{\theta}_i = \theta_{i1}, ..., \theta_{in_i}$, for individual i at times $t_{i1}, ..., t_{in_i}$, where time of death is $t = 0$ and the times before death are represented by negative values. So t_{in_i} is the last time individual i was observed in the study.

Model I is the linear regression model for θ_{ij} given by

$$\begin{aligned} \theta_{ij} &= \eta_{1i} + \eta_{2i} t_{ij} + e_{ij} \qquad & e_{ij} \sim N(0, \sigma^2) \\ \eta_{1i} &= \beta_1 + b_{1i} & \\ \eta_{2i} &= \beta_2 + b_{2i} \qquad & (b_{1i}, b_{2i}) \sim MVN\left(\mathbf{0}, \boldsymbol{\Sigma}\right). \end{aligned} \tag{3.1}$$

That is, parameter vectors (b_{1i}, b_{2i}) are multivariate normally distributed with mean zero and 2×2 variance-covariance matrix $\boldsymbol{\Sigma}$. The conditional distribution of θ_{ij} is normal with unknown variance σ^2. Random intercept η_{1i} is the value of θ_{ij} at time of death $t = 0$, and random slope η_{2i} reflects the change of θ_{ij} over the time before death, i.e., for $t < 0$. If the response variable would be manifest, then (3.1) would be a standard linear mixed-effects model (Pinheiro and Bates, 2000; Molenberghs and Verbeke, 2001).

Model I implies a linear change of latent cognitive function. As an alternative, a regression model with a change-point predictor will be specified. This model will be denoted Model II and is an extension of the fixed-effects bent-cable regression model for a manifest response as introduced by Tischler and Zang (1981) and further developed and investigated by Chiu *et al.* (2006). The bent-cable regression model can be seen as a smoothed broken-stick model. The broken-stick change-point model consists of two linear splines that intersect at the change point. The basic idea in bent-cable regression is that the kink in the broken-stick change-point model is replaced by a quadratic bend.

Given latent θ_{ij} for cognitive function, Model II describes the change of θ_{ij} over time by fitting two linear parts, which are connected smoothly by a third part. Corresponding to the three parts, the formulation of the model consists of three

equations and it given by

$$\theta_{ij} = \begin{cases} \eta_{1i} + \eta_{2i} t_{ij} + e_{ij} & t_{ij} \leq \tau_i - \delta \\ \eta_{1i} + \eta_{2i} t_{ij} + \eta_{3i}(t_{ij} - \tau_i + \delta)^2/4\delta + e_{ij} & \tau_i - \delta < t_{ij} \leq \tau_i + \delta \\ \eta_{1i} + (\eta_{2i} + \eta_{3i}) t_{ij} - \eta_{3i}\tau_i + e_{ij} & \tau_i + \delta < t_{ij}, \end{cases}$$

$$\begin{aligned} \eta_{1i} &= \beta_1 + b_{1i} & e_{ij} &\sim N(0, \sigma^2) \\ \eta_{2i} &= \beta_2 + b_{2i} & & \\ \eta_{3i} &= \beta_3 + b_{3i} & & \\ \tau_i &= g(\beta_4 + b_{4i}) & (b_{1i}, b_{2i}, b_{3i}, b_{4i}) &\sim MVN\left(\mathbf{0}, \mathbf{\Sigma}\right). \end{aligned} \tag{3.2}$$

where $\delta > 0$, τ_i is the random-effect change point, and $\mathbf{\Sigma}$ is a 4×4 variance-covariance matrix. Note that the location of the change point is midway the part that connects the two linear parts. Thus the quadratic bend has half-width δ and location at τ_i. The function g is the link function between the change point τ_i and its linear predictor. This function can be used to impose a restriction on the support of the change point.

Model II specifies a bent-cable curve for each individual i. Coefficient η_{2i} is the slope of the first linear part, and $\eta_{2i} + \eta_{3i}$ is the slope of the second linear part. The intercept η_{1i} is the value of the extrapolation of the first linear part at the time of death $t = 0$.

In the application, the value of transition parameter δ is fixed. To estimate δ from data, an intensive follow-up is needed - especially around the change point. In CC75C, we do not have that kind of data. We consider δ to be a nuisance parameter: it enables to define a smooth curve but its value is of limited importance in the current setting.

A restriction used in the application, and probably of interest in general, is the restriction of the support of the change point. We will use $\mathcal{L}$ and $\mathcal{B}$ to denote the lower and upper bound of τ_i. The specification of $\mathcal{L}$ and $\mathcal{B}$ may require prior subject-matter knowledge. If the change point is a fixed effect, i.e., $\tau_i = \beta_4$, then the lower and upper bounds can be enforced by using a uniform density as the prior for β_4, and specifying g as the identity link. A more general solution to enforce the bounds is to use a logistic link, see Muggeo *et al.* (2014). For example, when τ_i is including as a random effect the bounds are enforced by

$$\tau_i = g(\beta_4 + b_{4i}|\mathcal{L}, \mathcal{B}) = \frac{\mathcal{L} + \mathcal{B}\exp(\beta_4 + b_{4i})}{1 + \exp(\beta_4 + b_{4i})}. \tag{3.3}$$

This link function allows easily for adding covariates to the modelling of τ_i. This will be illustrated in the application.

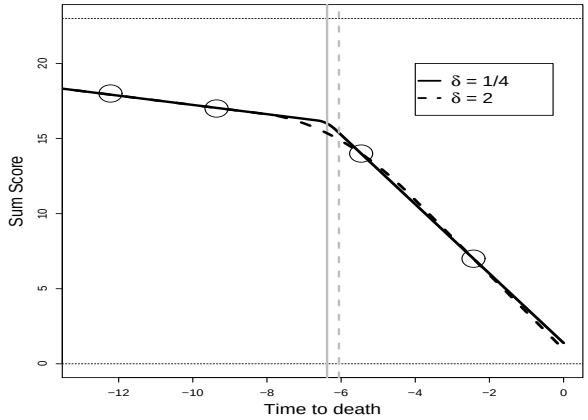

Figure 2: Example of a fixed-effects bent-cable model for CC75C data for a selected individual. Using the normal distribution for the response, and with vertical lines for the estimated location of the change point τ.

To illustrate the bent-cable regression model, Figure 2 depicts CC75C data from a selected individual and the fit of a fixed-effects model. For $\delta \to 0$ the model converges to the shape of a broken-stick change-point model (although it will stay smooth). The vertical lines are the estimated location of the change point τ for the two values of δ. The fit of the bent-cable model is by maximum likelihood estimation, where the normal distribution is used for the integer sum score with response scale $\{1, 2, ..., 23\}$. Note that using the normal distribution in a situation such as this may result in fitted values outside the range of the test score, which would cause a dependence between residuals and fitted values - a violation of model assumptions. This is one of the reasons to investigate IRT models.

The likelihoods for Models I and II conditional on the random effects are straightforward products of normal densities for values of θ_{ij} with means specified by the regression equations, and variance σ^2. The likelihoods are generically denoted by $p(\boldsymbol{\theta}|\boldsymbol{\beta}, \boldsymbol{b}, \sigma^2)$, where $\boldsymbol{\beta}$ is the vector with the fixed effects, and $\boldsymbol{b}$ is the vector with the random effects.

Cognitive function is latent since it is not directly observed but measured by a test (a questionnaire). At every observation time, the test consists of K items (questions). We formulate the normal ogive version of the graded-response model (Samejima, 1997; Fox, 2010). Let $\boldsymbol{y}_{ij} = (y_{ij1}, ..., y_{ijK})$ denote data for individual i at time t_{ij}. For item k with response categories 1 up to R (with the latter denoting the best score), the model has $R - 1$ ordered thresholds parameters $d_{k1}, .., d_{kR-1}$.

Together with the bounds $-\infty$ and ∞ and the ordering $-\infty < d_{k1} < ... < d_{kR-1} < \infty$, these thresholds define segments on the real line. The graded-response model is given by

$$p(y_{ijk} = m|\theta_{ij}, c_k, \boldsymbol{d}_k) \quad = \quad \Phi(c_k\theta_{ij} - d_{km-1}) - \Phi(c_k\theta_{ij} - d_{km}), \tag{3.4}$$

where $\boldsymbol{d}_k = (d_{k1}, ..., d_{kR-1})$. For item k, parameter c_k is the discrimination parameter, and $\boldsymbol{d}_k$ is the vector with the difficulty parameters. Given a value of θ_{ij}, these parameters define the probabilities of the answer categories.

In case $R = 2$ and answer categories 1 and 2, the graded response model (3.4) reduces to the ogive model $p(y_{ijk} = 2|\theta_{ij}, c_k, d_{k1}) = \Phi(c_k\theta_{ij} - d_{k1})$.

Mixed responses (dichotomous and polytomous items) can be formulated by making R item-dependent. A further extension is to make the item parameters c_k and $\boldsymbol{d}_k$ time-dependent, with notation c_{jk} and $\boldsymbol{d}_{jk} = (d_{jk1}, ..., d_{jkR-1})$ for item k at wave j. However, time-dependent item characteristics can lead to identification problems. If all items are time dependent, then it is not possible to distinguish change of individual latent ability over time from change of test characteristics. For that reason, at least one time-invariant item is required.

The conditional density for questionnaire data $\boldsymbol{y}$ with K items, item-dependent R, and wave-dependent item parameters is

$$p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{d}) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \prod_{k=1}^{K} \sum_{m=1}^{R_k} p(y_{ijk} = m|\theta_{ij}, c_{jk}, \boldsymbol{d}_{jk}) I(y_{ijk} = m),$$

where $I(y = m) = 1$ if $y = m$ and 0 otherwise.

The model for longitudinal questionnaire MMSE data that combines the regression model and the IRT model is not identified. This is caused by the IRT model where θ is latent and has no metric. A common choice in cross-sectional IRT models is to restrict θ_{ij} such that it is normally distributed with mean 0 and variance 1. An alternative is to impose restrictions on the scale of the item parameters. If $R = 2$ for all items, for example, restrictions $\prod_{k=1}^{K} c_k = 1$ and $\sum_{k=1}^{K} d_{k1} = 0$ can be used. For the longitudinal models, we will investigate similar restrictions in the CC75C data analysis.

4 Bayesian Inference

4.1 Markov Chain Monte Carlo

For the mixed-effects regression model with non-linear change-point predictor we formulate the basic steps in the Gibbs sampler for the model parameters given mixed responses (dichotomous and polytomous items). Define $\boldsymbol{y}^{\text{bin}}$ and $\boldsymbol{y}^{\text{pol}}$ as the dichotomous and polytomous subset of $\boldsymbol{y}$, respectively. Likewise, let $\boldsymbol{c}^{\text{bin}}$, $\boldsymbol{d}^{\text{bin}}$, $\boldsymbol{c}^{\text{pol}}$, and $\boldsymbol{d}^{\text{pol}}$ denote the corresponding subsets of the item parameters.

For dichotomous items, using an auxiliary variable allows for a straightforward implementation of sample techniques for the item parameters (Johnson and Albert 1999). Define $\boldsymbol{z}$ as a continuous representation of $\boldsymbol{y}^{\text{bin}}$ such that, corresponding to y^{bin}_{ijk}, z_{ijk} is normally distributed with mean $c_k\theta_{ij} - d_k$ and standard deviation 1. Value $y^{\text{bin}}_{ijk} = 2$ is observed when $z_{ijk} > 0$, and $y^{\text{bin}}_{ijk} = 1$ is observed, when $z_{ijk} \leq 0$.

A Gibbs sampler is a Markov chain Monte Carlo (MCMC) where each of the model parameters is sampled from a distribution which is conditional on the values of all other parameters. The conditional distributions that are used in the Gibbs sampler are given in shorthand notation by

$$\begin{aligned}
p(\boldsymbol{z}|...) &= p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{c}^{\text{bin}}, \boldsymbol{d}^{\text{bin}}, \boldsymbol{y}^{\text{bin}}) \\
p(\boldsymbol{c}^{\text{bin}}|...) &\propto p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{c}^{\text{bin}}, \boldsymbol{d}^{\text{bin}})p(\boldsymbol{c}^{\text{bin}}) \\
p(\boldsymbol{d}^{\text{bin}}|...) &\propto p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{c}^{\text{bin}}, \boldsymbol{d}^{\text{bin}})p(\boldsymbol{d}^{\text{bin}}) \\
p(\boldsymbol{c}^{\text{pol}}|...) &\propto p(\boldsymbol{y}^{\text{pol}}|\boldsymbol{\theta}, \boldsymbol{c}^{\text{pol}}, \boldsymbol{d}^{\text{pol}})p(\boldsymbol{c}^{\text{pol}}) \\
p(\boldsymbol{d}^{\text{poly}}|...) &\propto p(\boldsymbol{y}^{\text{pol}}|\boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{d}^{\text{pol}})p(\boldsymbol{d}^{\text{pol}}) \\
p(\boldsymbol{\theta}|...) &\propto p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{d})p(\boldsymbol{\theta}|\boldsymbol{\beta}, \boldsymbol{b}, \sigma, \boldsymbol{t}) \\
p(\boldsymbol{\beta}|...) &\propto p(\boldsymbol{\theta}|\boldsymbol{\beta}, \boldsymbol{b}, \sigma, \boldsymbol{t})p(\boldsymbol{\beta}) \\
p(\sigma|...) &\propto p(\boldsymbol{\theta}|\boldsymbol{\beta}, \boldsymbol{b}, \sigma, \boldsymbol{t})p(\sigma) \\
p(\boldsymbol{b}|...) &\propto p(\boldsymbol{\theta}|\boldsymbol{\beta}, \boldsymbol{b}, \sigma, \boldsymbol{t})p(\boldsymbol{b}|\boldsymbol{\Sigma}) \\
p(\boldsymbol{\Sigma}|...) &\propto p(\boldsymbol{b}|\boldsymbol{\Sigma})p(\boldsymbol{\Sigma}),
\end{aligned}$$

In the application, the specification of the random-effects distributions results in $\boldsymbol{\Sigma}$ being either a 2×2 matrix parameterized by two standard deviations and a correlation, or a 3×3 matrix parameterized by three standard deviations and one correlation. Parameter restrictions to identify the model are imposed in every run of the Gibbs sampler, this will be discussed in Section 5.1. Prior densities for parameters vectors $\boldsymbol{\beta}$, $\boldsymbol{c}$, and $\boldsymbol{d}$ assume independence between the coefficients, for example, $p(\boldsymbol{\beta}) = p(\beta_1)p(\beta_2)p(\beta_3)p(\beta_4)$.

In the MCMC, we use the logistic transformation in (3.3) – also when change point τ_i is a fixed effect. Because of the transformation, we have $\tau_i \in (\mathcal{L}, \mathcal{B})$ for any value of $\beta_4 + b_{4i}$, which allows for unrestricted sampling of β and b_{4i}.

The first five steps in the Gibbs sampler above are derived from MCMC schemes that can be found in the literature for cross-sectional IRT data analysis. Note that when conditioning on $\boldsymbol{\theta}$, the conditional distributions for the item parameters have the same form as in cross-sectional IRT. The first three steps are detailed in Johnson and Albert (1999, Chapter 6), where the conditional distributions are specified as normal distributions, the fourth and the fifth are discussed in Fox (2011, Section 4.3.4). We use Metropolis steps for $\boldsymbol{c}^{\text{pol}}$ and Metropolis-Hastings steps for $\boldsymbol{d}^{\text{pol}}$ to sample from the conditional distributions. In the sampling of candidates values for $\boldsymbol{d}^{\text{pol}}$, the ordering of the threshold is maintained.

The sampling of $\boldsymbol{\theta}$ is undertaken using Metropolis steps. Parameters for the regression model can be sampled using the methods in Gelfand *et al.* (1990) when the predictor is linear, or by using methods for non-linear regression models as discussed in, e.g., Gelman *et al.* (2004). Since we use non-conjugate prior densities for the standard deviations, we also use Metropolis steps to sample values for the standard deviations.

The above Gibbs sampler results in draws of parameters values from the posterior of Model II. The expression for this posterior is

$$p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{c}, \boldsymbol{d}, \boldsymbol{\beta}, \boldsymbol{b}, \sigma, \boldsymbol{\Sigma} | \boldsymbol{y}, \boldsymbol{t}) \propto$$
$$p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{c}^{\text{bin}}, \boldsymbol{d}^{\text{bin}}, \boldsymbol{y}^{\text{bin}}) p(\boldsymbol{y}^{\text{pol}}|\boldsymbol{\theta}, \boldsymbol{c}^{\text{pol}}, \boldsymbol{d}^{\text{pol}}) p(\boldsymbol{\theta}|\boldsymbol{\beta}, \boldsymbol{b}, \sigma, \boldsymbol{t}) p(\boldsymbol{b}|\boldsymbol{\Sigma}) p(\boldsymbol{c}) p(\boldsymbol{d}) p(\boldsymbol{\beta}) p(\sigma) p(\boldsymbol{\Sigma}).$$

For missing item scores, we assume that the values are missing at random, i.e., the missingness does not depend on the missing value itself, but may depend on observed data. The MCMC can easily be extended to take this kind of missing data into account: first missing values are sampled from their conditional distributions given current parameter values, and next the MCMC steps for complete data are undertaken. In formulas, if dichotomous value y_{ijk}^{bin} is missing in wave 2, 3, or 4 (see Figure 1), then it is sampled using a Bernoulli trial with success probability $\Phi(c_k\theta_{ij} - d_{k1})$. If polytomous value y_{ijk}^{pol} is missing, then it is sampled using a multinomial distribution with probabilities given by (3.4).

Posterior inference for means, credible intervals, and other derived quantities are based upon two chains, each with a burn-in and additional updates used for inference. Convergence of the chains for the item parameters and the parameters for the growth model is be assessed by visual inspection of the chains and by diagnostics tools provided in the R-package coda (Plummer *et al.* 2006). This will be illustrated in the application.

4.2 Deviance Information Criterion and L-criterion

To compare models, we use the Deviance Information Criterion (DIC, Spiegelhalter *et al.* 2002) and the L-criterion (Laud and Ibrahim, 1995; Gelfand and Gosh, 1998). The DIC comparison is based on a trade-off between the fit of the data to the model and the complexity of the model. Models with smaller DIC are better supported by the data. The DIC in the current setting is based on the deviance, which is specified for the questionnaire data by

$$D(\boldsymbol{y}, \boldsymbol{\Omega}) = -2 \log p(\boldsymbol{y}|\boldsymbol{\Omega}), \tag{4.1}$$

where $\boldsymbol{\Omega} = (\boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{d})$. The DIC is given by

$$\text{DIC} = \widehat{D} + 2p_D, \tag{4.2}$$

where $\widehat{D} = D(\boldsymbol{y}, E(\boldsymbol{\Omega}))$ is the *plug-in deviance* and p_D denotes the effective number of parameters. The *expected deviance* is denoted $\overline{D}$ and is used to define p_D as $\overline{D} - \widehat{D}$. The expected deviance is estimated by $M^{-1} \sum_{m=1}^{M} D(\boldsymbol{y}, \boldsymbol{\Omega}^{(m)})$ with m denoting the iterations in the MCMC algorithm. The plug-in deviance is estimated by using the posterior means of the model parameters.

The plug-in deviance is not invariant to parametrization and does not take into account the precision of the estimates. The expected deviance, however, is a function of the posterior of the model parameters and does account for the precision of the estimates (Plummer 2008).

Although the DIC is widely used, it is not without problems, see, e.g., Carlin and Louis (2009) and the discussion in the seminal paper Spiegelhalter *et al.* (2002). The DIC can given inappropriate results if there are highly non-normal posterior distributions of the parameters on which prior distributions have been placed (Lunn *et al,* 2009). Given the complexity of the current model, some caution when using the DIC is therefore recommended. As alternative, we look at the L-criterion, which is not justified by relying on asymptotic results. The L-criterion is a posterior predictive check and is derived from the sum of the variance of the predictions and a distance measure with respect to predicted and observed values. The L-criterion for replicates x_l^{rep} of observed x_l is given by

$$L = \sqrt{\sum_l Var[x_l^{rep}] + (E[x_l^{rep}] - x_l)^2}. \tag{4.3}$$

As is clear from the expression itself, smaller values of L are better. In the IRT context, this criterion can be formulated with respect to the individual item scores y_{ijk}, in which case the summation in (4.3) is over all the combinations of i, j, and k.

We will only discuss and compare the criterions in light of the application, a more theoretical and wider discussion is outside the scope of the present paper.

5 Data Analysis

Section 5.1 investigates parameter restrictions using a subsample from the CC75C data with dichotomous items only. The results provide insight in the effect of restrictions and will also function as a preliminary step for the change-point data analysis in Section 5.2

5.1 Parameter restrictions

Parameter restrictions to identify the model are imposed in every run of the Gibbs sampler. Restrictions are needed because the latent cognition parameterized by θ_{ij} does not have a metric. For cross-sectional IRT models parameter restrictions are discussed in Fox (2010, Section 4.4.2). In the following similar restrictions are used for longitudinal models. Using CC75C data, the effects of the restrictions are investigated with respect to model comparison and convergence of the MCMC sampling.

Since running the MCMC is computationally intensive we use a random subsample of the data with sample size $N = 400$. For this sample size, there is enough information in the data for parameter inference whilst the running of the MCMC is not too time consuming. Frequencies of the number of times the MMSE is observed per individual in this subsample are 168, 113, 61, 49, 6, and 3, for one up to six times, respectively. Hence there are 821 observations in total.

This section discusses Model I and Model II for binary IRT data. The models for latent cognitive function consist of an IRT measurement model and a latent growth model. In this section, Model I and Model II have the same measurement model, but differ in the model for the latent growth.

Using $K = 12$ binary items with score 1 for an incorrect answer, and 2 for a correct one, the IRT measurement model is given by

$$p(y_{ijk} = 2|\theta_{ij}, c_k, d_k) = \Phi(c_k\theta_{ij} - d_k),$$

where i denotes the individual, j indexes the repeated observations, and k indexes the question, see (3.4).

The latent growth in Model I is specified by (3.1) using a linear predictor for the latent θ_{ij}. For the Bayesian inference, vague prior densities are used for the residual

variance and the variance components, i.e., $\sigma, \sigma_1, \sigma_2 \sim U(0,5)$. For ρ the prior is $U(-1,1)$. For the remaining model parameters, the priors are improper and equal to 1.

The latent growth in Model II is a restricted version of (3.2) and is given by

$$\begin{aligned} \eta_{1i} &= \beta_1 + b_{1i} \qquad \eta_{2i} = \beta_2 \qquad \eta_{3i} = \beta_3 + b_{3i} \\ \tau_i &= \tau = \frac{\mathcal{L} + \mathcal{B}\exp(\beta_4)}{1+\exp(\beta_4)} \qquad (b_{1i}, b_{3i}) \sim MVN\left(\mathbf{0}, \mathbf{\Sigma}\right), \end{aligned}$$

where $\mathbf{\Sigma}$ is parameterized by standard deviations σ_1 and σ_3, and correlation ρ. For the fixed-effects β_1, β_2, β_3 and β_4, the priors are improper and equal to 1. For the current time scale, $\mathcal{B} = 0$ represents the time of death, $\mathcal{L} = -12$ represents twelve year before the time of death. The choice for $\mathcal{L}$ reflects our current interest: going back more than 12 years means losing the assumed link between cognitive decline and the proximity of death. The transition parameter δ is fixed to $1/2$, representing half a year. For all other parameters the priors are as in Model I.

The scale of the 821×1 vector with latent values of θ_{ij}, denoted by $\boldsymbol{\theta}$, can be restricted in each MCMC iteration by transforming a sampled $\boldsymbol{\theta}$ such that the resulting $\boldsymbol{\theta}$ has mean 0 and variance 1. This fixing of the metric for θ_{ij} does not have to take into account the hierarchical structure of the repeated observations within individuals. The transformation is linear so the relative distances between values of θ_{ij} are maintained. Note also that fixing the metric for θ_{ij} does not fix σ, which is the conditional variance $Var(\theta_{ij}|\eta_{1i}, \eta_{2i}, t_{ij})$.

Alternatively, if the restriction concerns dichotomous item parameters $\boldsymbol{c}$ and $\boldsymbol{d}$, then re-scaling is undertaken such that $\prod_{k=1}^{K} c_k = 1$ and $\sum_{k=1}^{K} d_k = 0$. This is a common choice in cross-sectional IRT models and can easily be implemented in a longitudinal setting. If some or all items are polytomous, restrictions can be imposed by restricting the product of their discrimination parameters to be one, and by restricting, for each item, the sum of the difficulty parameters.

For each of the models, two chains are used each consisting of 40000 iterations. For Model I, half of the iterations are discarded afterwards as the burn in. For Model II, we discard the first 5000 of each chain as the burn-in. Model II is more complex and convergence diagnostics improved when using more sampled values. The jump distributions in the Metropolis steps are adjusted during the burn-in such that the acceptance rates are between 30% and 50%. Table 1 presents DIC and L-criterion statistics, and convergence diagnostics using the univariate or multivariate potential scale reduction factor (Gelman and Rubin, 1992).

In IRT, the ability parameters θ_{ij} are much more involved in the determination of the effective number of parameters p_D than the item parameters. Increased shrinkage

Table 1: For a subset of the data ($N = 400$), DIC, L-criterion, and potential scale reduction factors (psrf) for a selection of the parameters.

	DIC	$\overline{D}$	$\widehat{D}$	p_D	minimum D	L-criterion
Linear model for latent cognitive function						
Restricted $\boldsymbol{\theta}$	5750	5400	5051	350	5225	29.0
	psrf(σ) = 1.02		psrf($\boldsymbol{c}$) = 1.01			
	psrf($\boldsymbol{\beta}$) = 1.00		psrf($\boldsymbol{\rho}$) = 1.02		psrf(σ_1, σ_2) = 1.01	
Restricted $\boldsymbol{c}$ and $\boldsymbol{d}$	5662	5235	4808	427	5058	28.6
	psrf(σ) = 1.07		psrf($\boldsymbol{c}$) = 1.05			
	psrf($\boldsymbol{\beta}$) = 1.01		psrf($\boldsymbol{\rho}$) = 1.02		psrf(σ_1, σ_2) = 1.02	
Restricted $\boldsymbol{c}$ and $\boldsymbol{\theta}$	5673	5246	4818	428	5071	28.6
	psrf(σ) = 1.04		psrf($\boldsymbol{c}$) = 1.03			
	psrf($\boldsymbol{\beta}$) = 1.00		psrf($\boldsymbol{\rho}$) = 1.05		psrf(σ_1, σ_2) = 1.01	
CP Model for latent cognitive function						
Restricted on $\boldsymbol{c}$ and $\boldsymbol{d}$	5643	5236	4828	408	5052	28.6
	psrf(σ) = 1.16		psrf($\boldsymbol{c}$) = 1.05			
	psrf($\boldsymbol{\beta}$) = 1.08		psrf(ρ) = 1.06		psrf(σ_1, σ_3) = 1.07	

on θ_{ij} induces a lower value of p_D because with more shrinkage there is less variability in θ_{ij}. As a random effect, the shrinkage on θ_{ij} is determined by the variance: if the variance of θ_{ij} is higher, then there is less shrinkage and p_D is higher.

Denote the overall variance of θ_{ij} by σ^*. Restricting the discrimination parameters $\boldsymbol{c}$ by equalling their product to one will lead to a value of σ^*, which defines the amount of shrinkage of θ_{ij} and hence the value of p_D. Restricting θ_{ij} by $\sigma^* = 1$ is a second way to define the amount of shrinkage. The two choices of parameter restrictions do not lead to the same amount of shrinkage on θ_{ij}. In the application, Table 1 shows that for the chosen restrictions on $\boldsymbol{c}$ and $\boldsymbol{d}$ imply less shrinkage.

Note that all this is not an issue of reparametrization, it is about restrictions on model parameters. For specified restrictions, DIC can be used to compare models. From Table 1, e.g., we can conclude that for the current restrictions, the model with

restricted $\boldsymbol{c}$ and $\boldsymbol{d}$ performs better than the model with restricted θ_{ij}. But this does not imply that choosing to restrict item parameters is the best choice in general. The idea of the un-identified model is that we should be able to find a restriction on θ_{ij} such that the resulting DIC is equal to the DIC obtained with a given restriction on $\boldsymbol{c}$ and $\boldsymbol{d}$.

From the univariate and multivariate potential scale reduction factor in Table 1, we conclude that, for Model I, the choice of restrictions do not have an influence on the convergence of the MCMC. For Model II, longer chains were needed to attain convergence statistics similar to Model I.

Both the DIC and the L-criterion favor the chosen restrictions on $\boldsymbol{c}$ and $\boldsymbol{d}$ over the chosen restrictions on θ_{ij}. Using restrictions on $\boldsymbol{c}$ and θ_{ij} for the model with the linear predictor also works well, but will not be pursued in what follows. According to the DIC, the model with the change-point predictor is an improvement upon the model with the linear predictor. The L-criterion does not signal a difference between these two models.

We also investigated the sensitivity with respect to the choice of fixing the value of δ to 1/2 in Model II. Results across various fixed values are very similar. For $\delta = 1/4$, posterior mean for the fixed-effects change point parameter τ is -6.33, and 95% credible interval is (-7.89, -5.08). The DIC is 5640. For $\delta = 1/2$, the inference is -6.42 (-7.65, -5.35), with DIC in Table 1 equal to 5643. For $\delta = 1$, the inference is -6.48 (-8.32, -5.30), with DIC = 5644. Convergence diagnostics are similar for all three settings and do not indicate problems.

There is not a lot of variation in the DICs in this sensitivity analysis, and the L-criterion is the same across the three different values, i.e., 28.6. With increasing values for δ there is a slight shift in the posterior mean further away with the time of death, but the shift is less than half a year with all the 95% credible intervals wider than 2 years.

We are not sure whether the DIC is suitable in the current setting for model comparison across models with a different random-effects specifications. The L-criterion has the advantage this it is not justified by relying on asymptotic results, but may be to too crude to compared models that differ in minor aspects only.

5.2 Random change-point model for CC75C

This section investigates the IRT growth model with random change-points for the CC75C data specified in Section 2, with sample size $N = 1179$. For the CC75C waves 2 up to 7, the twelve dichotomous items in the data are `Qweekday`, `Qdateday`, `Qmonth`, `Qyear`, `Qseason`, `Qcounty`, `Qtown`, `Qstreet`, `Qplace`, `Qifs`, `Qwrite`, and

`Qread`. The first nine are with respect to orientation, `Qifs` is about repetition of the expression "No ifs, ands or buts.", `Qwrite`, and `Qread` are tests for writing and reading, respectively.

For waves 2, 3 and 4, the four polytomous items are `Qobject`, `Qregist`, `Qrregist`, and `Qpaper`. The first is about recognizing two objects. `Qregist` is about registering three objects, and `Qrregist` is about remembering these three objects later on in the interview. `Qpaper` is about following instructions regarding handling a piece of paper. For the remaining waves 5, 6, and 7, the scores for the underlying dichotomous items are available for these four polytomous items. For example, `Qobject` is split up in two items scoring the correct recognition of a pencil (`Qpencil`) and a watch (`Qwatch`). The other three polytomous items are split up in (`Qrapple`, `Qrtable`, `Qrpenny`), (`Qrrapple`, `Qrrtable`, `Qrrpenny`), and (`Qhand`, `Qfold`, `Qlap`), respectively.

For the data analysis, the polytomous scores for waves 5, 6, and 7 are derived from the observed dichotomous scores, see Figure 1. The dichotomous scores underlying the four polytomous items in waves 2, 3 and 4 are missing by design and are therefore not taken into account in the analysis. We could have imputed these missing dichotomous scores within the MCMC but that would add unnecessary uncertainty to the analysis.

The change-point Model II with random change point that will be used is specified by the graded-response measurement model (3.4) and the latent growth model

$$\begin{aligned}
\eta_{1i} &= \beta_1 + b_{1i} \\
\eta_{2i} &= \beta_2 \\
\eta_{3i} &= \beta_3 + b_{3i} \\
\tau_i &= \frac{\mathcal{L} + \mathcal{B}\exp(\beta_4 + b_{4i} + \gamma_1 D_i + \gamma_2 S_i)}{1 + \exp(\beta_4 + b_{4i} + \gamma_1 D_i + \gamma_2 S_i)} \qquad (b_{1i}, b_{3i}, b_{i4}) \sim MVN\left(\mathbf{0}, \mathbf{\Sigma}\right),
\end{aligned}$$

where D_i is age at death for individual i, and S_i is a 0/1 dummy for women/men. The 3×3 matrix $\mathbf{\Sigma}$ is parameterized by standard deviations σ_1 and σ_3, and correlation ρ for random effects b_{1i} and b_{3i}, and standard deviation σ_4 for random effect b_{4i}. Model II has thus six fixed-effects parameters $(\beta_1, \beta_2, \beta_3, \beta_4, \gamma_1, \gamma_2)$, one residual variance parameter (standard deviation σ), three parameters for the standard deviations for the random effects $(\sigma_1, \sigma_3, \sigma_4)$, and one correlation parameter for random intercept and random second slope (ρ). There are as many ability parameters θ_{ij} as there are interviews, there are 27 discrimination parameters c (twelve for the twelve dichotomous items in all waves, eleven for the dichotomous items in the last three waves, and four for the four polytomous items). There are 23 difficulty parame-

Table 2: Posterior mean (and 95% credible interval) for the parameters in the growth-curve submodel in Model II. Number of MCMC iterations is 60000, where first 10000 are ignored as burn-in.

Coefficients			Variance components					
β_1	1.36	(1.25, 1.47)	σ_1	0.46	(0.41, 0.51)			
β_2	-0.03	(-0.04, -0.02)						
β_3	-0.22	(-0.31, -0.14)	σ_3	0.11	(0.06, 0.17)	ρ	-0.37	(-0.82, 0.12)
β_4	1.37	(0.54, 2.27)	σ_4	1.43	(0.91, 2.23)			
γ_1	-0.08	(-0.11, -0.05)	σ	0.33	(0.27, 0.38)			
γ_2	0.85	(0.37, 1.57)						

ters for the dichotomous items, two difficulty parameters for the first polytomous question and $3 \times 3 = 9$ difficulty parameters for the three remaining polytomous questions.

As in Section 5.1, the priors for the fixed-effects parameters are equal to 1, the vague prior density for σ_1 and σ_3 is $U(0, 5)$, and for σ_4 the prior is $U(0, 3)$. All other priors are improper and equal to 1. The transition parameter δ is fixed to 1/2. MCMC for the item parameters is robust and convergence is quickly attained. For the parameters in the growth-curve submodel, more iterations are needed for proper convergence. The total number of iterations is 60000, where the first 15000 are ignored as burn-in.

Table 2 presents the posterior inference for the parameters for the growth-curve modeling.

An absolute value of θ_{ij} as a measure of cognitive ability should not be interpreted on its own. Only comparisons between values are meaningful. The fixed-effects slope parameters in Table 2 show that conditional on the mean zero for the random effects, there is a slight overall decline of ability before the change point (β_2) followed by a sharper decline after the change point ($\beta_2 + \beta_3$). The correlation ρ between the random effects for the intercept and second slope (b_1 and b_3) has a negative posterior mean, but a wide 95% credible interval, which includes zero.

Posterior inference for the effect of age at death, γ_1, shows that people dying at a high age have a change point more years before death then people who die younger. The posterior mean $\gamma_2 > 0$ shows that men tend to have change points closer to death than women.

Figure 3 depicts the posterior distributions for the item parameters. The item

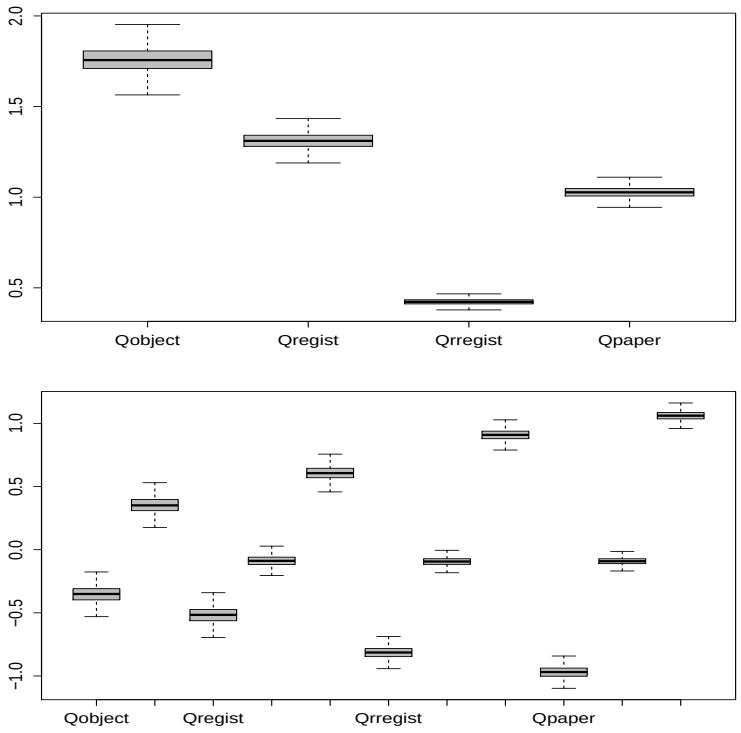

Figure 3: Posterior inference for the parameters for the four polytomous items. Discrimination parameters $\boldsymbol{c}^{\text{pol}}$ at the top and corresponding difficulty parameters $\boldsymbol{d}^{\text{pol}}$ at the bottom.

`Qobject`, recognizing two objects discriminates best, whereas item `Qrregist` about remembering these objects later does not seem to help in discriminating between individual abilities. Figure 3 also shows that the item parameters are well identified.

Posterior inference for random change points is best undertaken by assessing the posterior for τ_i for $i = 1, ..., N$. Additional sampling using 1000 iterations was undertaking for posterior inference using the random effects. For the $N = 1179$ posterior means of the individual-specific random change points, the mean is -2.52 years before death. The distribution of the N posterior means is skewed: the quantiles are -10.58, -4.02, -2.52, -1.55, and -0.40, for the 0, 1/4, 1/2, 3/4, 1 quantiles, respectively.

The model assumes that all individuals have a change point. However, the change point assessment is only of interest regarding those individuals who experience change during the follow-up. For the stable trends, change points are fitted

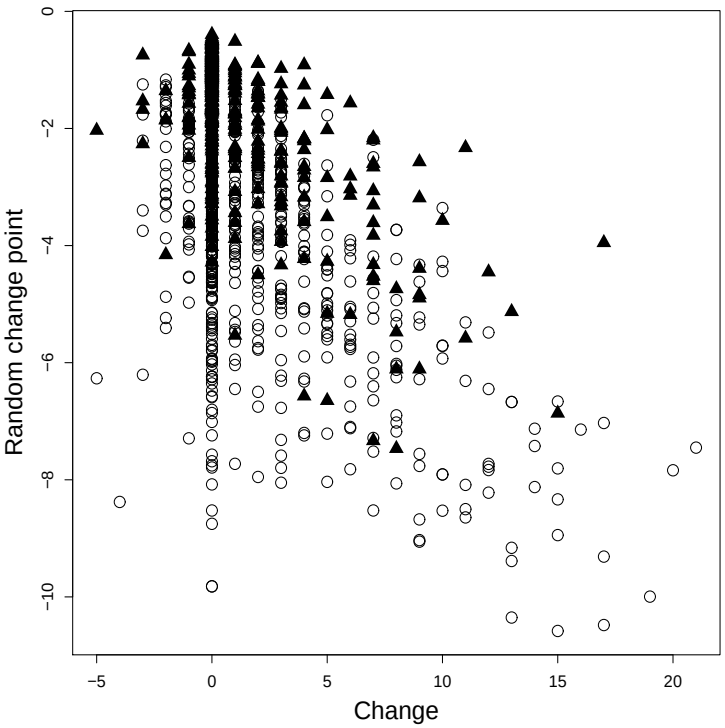

Figure 4: Posterior means for $N = 1179$ random change points against change in test score during follow-up (first observed score minus last one). Circles for women, triangles for men.

very close to the time of death and cannot be considered to be true change points.

Define change in test score as first observed score minus last observed one. This definition ignores the inherent variability of the measuring, but will help to assess the distribution of the individual change points. Figure 4 depicts posterior means for the $N = 1179$ random change points against change in test score during follow-up. The graph illustrates the skewness of the posterior means, the difference between men and women, and that more change is associated with an earlier change point on the scale years before to death. For example, for the individuals with a score change of 5 or more, the mean of the individual posterior means is -5.66 years before death (and the median is -5.48). For a score change of 10 or more, the mean is -7.07 (median -7.42).

It is possible to predict the item scores given the posterior distribution of the population parameters and the random effects. For a random sample from those

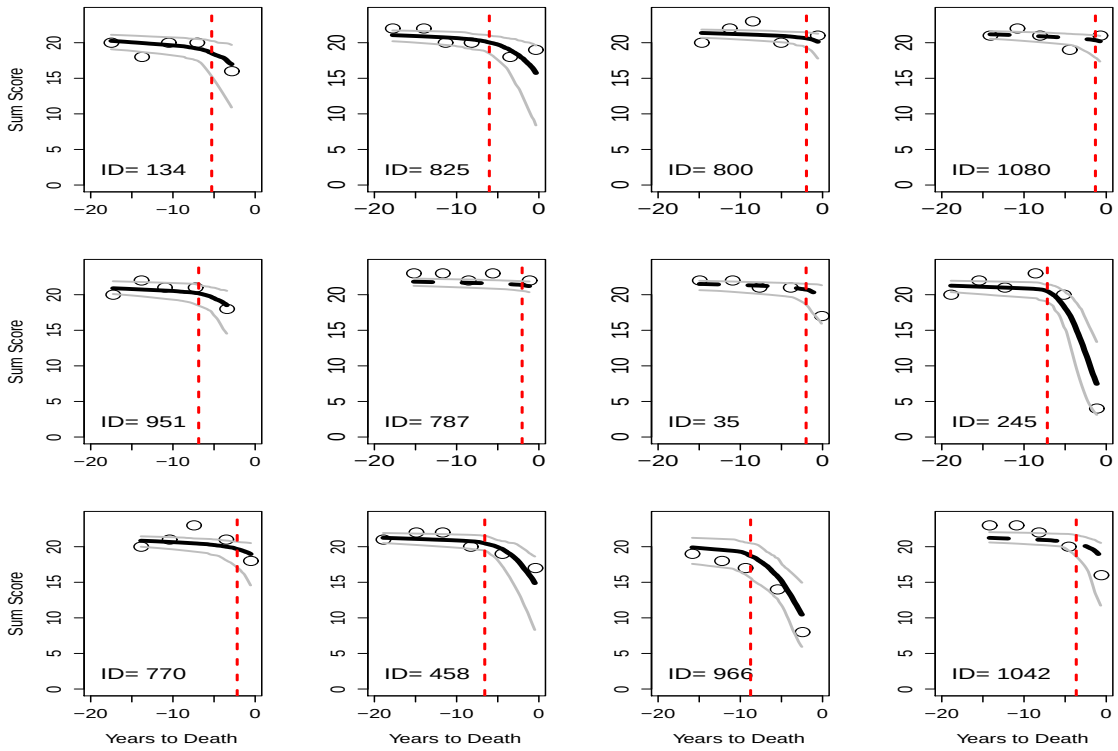

Figure 5: For a random sample from individuals who were seen five times or more, observed sum scores and fitted curves (with 95% credible band). Vertical line for location posterior mean individual change point. Dashed lines for men, solid for women.

individuals who were seen five times or more, Figure 5 shows fitted sum scores and observed scores. It is nice to see that the model captures to observed sum-score trend reasonable well for these individuals. Figure 5 also shows that the model is capable of fitting stable individual trends, see the graph for individuals ID = 1080 and ID = 787. This is an important feature of the modeling, which acknowledges that not all individuals in the sample experience cognitive decline in the years prior to death.

6 Conclusion

As stated in Klein Entink *et al.* (2011), the use of questionnaires is widespread, not only in social sciences but also the biostatistics (e.g., the measuring of depression or quality of life). Because of this, item response theory (IRT) becomes more important as it enables the measuring of an underlying variable while recognizing the psychometric properties of the questionnaire that is used. By accounting for differential item effects, IRT implies a more realistic data analysis compared to models which use the sum score and assume that each item contributes in an equal way.

In our model, the latent variable representing cognitive function is assumed to be continuous. For applications where the assumption of a discrete latent variable is suitable, other methods are available. Bartolucci *et al.* (2009) discuss a model where latent states of a first-order discrete-time Markov model explain observed longitudinal binary item scores. For cognitive function, it seems reasonable to assume that the latent scale is continuous given that we assume that change is gradual. But it is of course possible to approximate a continuous latent scale by introducing a series of latent discrete states. How many latent states to define is not clear at the outset, and this is a disadvantage of the latent-states approach. But there is a similar problem with assuming beforehand that the latent scale for cognitive function is univariate. Exploring a bivariate latent scale in our application would be an interesting extension of our current model. An additional advantage of using the mixed-effects growth model in our application is that it does not require a regular spacing of consecutive interviews. Observation times are allowed to vary between and within individuals. With latent-variable modelling, there is always the problem of interpretability of the parameter for the latent trait. In our model, the slope parameters for the change in latent cognitive function do not provide information other than their signs. However, the interpretation of the posterior distribution of the change point is clear and direct, as illustrated by Figures 4 and 5.

When models are compared, the minimum value of the Deviance Information Criterion (DIC) is intended to identify the model that performs best with respect to short-term predictions (Lunn *et al,* 2012). The criterion is particularly useful when comparing random-effects models, although as a general method for model comparison it is also subject to criticism, see Section 4.2. In the current setting, the effective number of parameters is high due to the latent variable approach in addition to the random-effects specification, and we are not sure whether the DIC is suitable for comparing models with different random-effects structures. The L-criterion is used as an alternative. The latter criterion is not justified by relying on asymptotic results. We also investigated the use of the pseudo-marginal likelihood

(Geisser and Eddy, 1979; Gelfand and Dey, 1994) as an alternative to the DIC. But the estimation of the conditional predictive ordinates was quite unstable for the models and the data at hand, and this approach was not pursued.

Although there are publications on IRT models for longitudinal questionnaire data, there is still scope for further work and improved data analysis. Douglas (1999) and Fox and Glas (2001) discuss linear regression models for time-dependent latent ability as measured by IRT. Recent work by Wang *et al.* (2013) presents Bayesian inference for IRT models where the change in latent ability over time is modeled using dynamic models. The random change point model is capable of fitting stable individual trends as shown in Figure 5. A possible extension would be to explicitly model stable trends versus change using a two-component mixture. A similar latent-class mixture approach is used in Van den Hout *et al.* (2013) who analyze a manifest outcome variable with a change-point predictor.

For the Cambridge City over-75s Cohort Study sample in the analysis, all the $N = 1179$ death times are obtained from population registers. For the intermittent missing data (missing an interview) the random-effects model should provide some robustness against violation of the MAR assumption. We used the CC75C data from wave 2 onwards only. The dropout between wave 1 and wave 2 in CC75C has not been taken into account in the current analysis, and whether this has an impact on the conclusions is still to be investigated.

The methods in the current paper show that it is possible to estimate regression models with non-linear predictors for the underlying variable and that it is worthwhile to investigate different ways of identifying the model. The application shows how this approach can be used to investigate potential decline in cognitive function taking into account the possibility of a one-off change in the trend of the decline.

Acknowledgements

The authors would like to thank the CC75C group for providing the data, see `www.cc75c.group.cam.ac.uk` for a list of contributors and funding organizations. Feedback from two anonymous referees was used to revise the manuscript.

References

Bartolucci F, Lupparelli M and Montanari GE (2009) Latent Markov model for longitudinal binary data: An application to the performance evaluation of nursing homes. *Annals of Applied Statistics*, **3**, 491–879.

Carlin BP and Louis TA (2009) *Bayesian Methods for Data Analysis, Third Edition*. Boca Raton, Florida: Chapman and Hall/CRC.

Chiu G, Lockhart R and Routledge R (2006) Bent-cable regression theory and applications. *Journal of the American Statistical Association*, **101**, 542–53.

Cohen P (Ed.) (2008) *Applied Data Analytic Techniques for Turning Points Research*. Routledge, New York.

Douglas JA (1999) Item response models for longitudinal quality of life data in clinical trials. *Statistics in Medicine*, **18**, 2917–31.

Folstein MF, Folstein SE and McHugh PR (1975) Mini-mental state. A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, **12**, 189–98.

Fox J-P (2010) *Bayesian Item Response Modeling*. New York: Springer.

Fox J-P and Glas CAW (2001) Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, **66**, 271–88.

Geisser S and Eddy WF (1979) A predictive approach to model selection. *Journal of the American Statistical Association*, **74**, 153-60.

Gelfand AE and Dey D (1994) Bayesian model choice: asymptotic and exact calculations, *Journal Royal Statistical Society B*, **56**, 501-14.

Gelfand AE and Ghosh SK (1998) Model choice: A minimum posterior predictive loss approach. *Biometrika*, **85**, 1–11.

Gelfand AE, Hills SE, Racine-Poon A and Smith AFM (1990) Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, **85**, 972–85.

Gelman A and Rubin DB (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–511.

Gelman A, Carlin JB, Stern HS and Rubin DB (2004) *Bayesian Data Analysis*. London: Chapman and Hall.

Holling H, Böhning W and Böhning D (2012) Meta-analysis of diagnostic studies based upon SROC-curves: a mixed model approach using the Lehmann family. *Statistical Modelling*, **12**, 347–375.

Jacqmin-Gadda H, Commenges D and Dartigues J-F (2006) Random changepoint model for joint modeling of cognitive decline and dementia. *Biometrics*, **62**, 254–260.

Johnson VE and Albert JH (1999) *Ordinal Data Modeling*. New York: Springer.

Klein Entink RH, Fox J-P, and Van den Hout A (2011) A mixture model for the joint analysis of latent developmental trajectories and survival. *Statistics in Medicine*, **30**, 2310–25.

Laud W and Ibrahim JG (1995) Predictive Model Selection. *Journal of the Royal Statistical Society. Series B*, **57**, 247–62.

Lévy-Leduc C, Roueff F (2009) Detection and localization of change-points in high-dimensional network traffic data. *Annals of Applied Statistics* **3**, 637–662.

Lunn D, Jackson C, Best N, Thomas A and Spiegelhalter D (2012) *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Boca Raton: Chapman and Hall/CRC.

Lunn D, Spiegelhalter D, Thomas T and Best N (2009) The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, **28**, 3049–67.

Molenberghs G and Verbeke G (2001) A review on linear mixed models for longitudinal data, possibly subject to dropout. *Statistical Modelling*, **1**, 235–269.

Muggeo VMR, Atkins DC, Gallop RJ and Sona Dimidjian (2014). Segmented mixed models with random changepoints: a maximum likelihood approach with application to treatment for depression study. *Statistical Modelling*, to appear. DOI: 10.1177/1471082X13504721

Pinheiro JC and Bates D (2000) *Mixed-Effects Models in S and S-Plus*. Springer: New York, 2000.

Plummer M (2008) Penalized loss functions for Bayesian model comparison. *Biostatistics*, **9**, 523–39.

Plummer M, Best N, Cowles K and Vines K (2006) CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, **6**, 7–11.

Riegel KF and Riegel RM (1972) Development, drop, and death. *Developmental Psychology*, **6**, 306–19.

Rubin DB (1976) Inference and missing data. *Biometrika*, **63**, 581–92.

Samejima F (1997) The graded response model. In *Handbook of modern item response theory* (eds. W.J. van der Linden and R.K. Hambleton). New York: Springer.

Spiegelhalter DJ, Best NG, Carlin BP and Van der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). *Journal Royal Statistical Society B*, **4**, 583–640.

Stasinopoulos DM, Rigby RA (1992) Detecting break points in generalised linear models. *Computational Statistics & Data Analysis* **13**, 461–471.

Tishler A and Zang I (1981) A new maximum likelihood algorithm for piecewise regression. *Journal of the American Statistical Association*, **76**, 980–87.

Van der Linden WJ and Hambelton RK (1997) *Handbook of Modern Item Response Theory*. New York: Springer.

Van den Hout A, Muniz Terrera G and Matthews FE (2011) Smooth random change point models. *Statistics in Medicine*, **30**, 599-610.

Van den Hout A, Muniz-Terrera G, Matthews FE (2013) Change point models for cognitive tests using semi-parametric likelihood. *Computational Statistics and Data Analysis* **57**, 684–698.

Verbeke G, Molenbergh G (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Wang X, Berger JO, Burdick BS (2013) Bayesian analysis of dynamic item response models in educational testing. *Annals of Applied Statistics* **7**, 126–153