A varying coefficient graded response model

Bakker, T.

A Varying Coefficient Graded Response Model

T. Bakker

Thesis advisor: Prof. Dr. H. Kelderman

Second reader: Prof. Dr. W.J. Heiser

MASTER THESIS

Defended on April 11, 2017

STATISTICAL SCIENCE
FOR THE LIFE AND BEHAVIOURAL SCIENCES

Abstract

We propose a varying coefficient IRT model, in order to study the effect of a metric variable on model and population parameters estimated by IRT models. Kernel smoothing was used to capture the variation, and cross-validation to determine optimal parameters. The model was applied to a variety of simulated data sets in order to test its properties, and on a real-world personality data set. The tests on simulated data showed the ability to recover and visualize the variation of coefficients and their confidence bands over time with some success. The real-world tests showed some, but limited variation, depending on the trait studied.

Contents

1 Introduction **5**

2 Model and estimation **7**

2.1 Definitions and model 7

2.1.1 Terminology 9

2.1.2 Scoring polytomous items 10

2.1.3 Model estimation 10

2.2 Varying coefficients 12

2.2.1 Model estimation 13

2.2.2 Bandwidth selection 17

2.2.3 Confidence bands 18

3 Simulated results **19**

3.1 Data generation 19

3.2 Experiments 20

4 Application **37**

4.1 Data-set 37

4.2 Adjustments to data-set 38

4.3 Results 38

4.3.1 Bandwidth determination 39

4.3.2 Parameter estimates 39

4.3.3 Exploratory conclusions 44

4.3.4 The effect of sample size on optimal bandwidth 44

5 Discussion **46**

References **48**

A Results for the other 4 factors **51**

B Proof for 2.1.3 **56**

1 Introduction

The modern field of psychometrics, as arguably introduced by Spearman with his paper on the measurement on human intelligence [Spearman, 1904], is concerned with the quantitative measurement and description of psychological factors. Spearman's single factor model was further explored [Lawley and Maxwell, 1962] and extended to multiple factors [Thurstone, 1929] in the following century. Where factor analysis deals with continuous responses, many experiments have discrete outcomes, calling for a method specifically tailored to such results.

Item Response Theory (IRT) [Rasch, 1960, Birnbaum, 1968] is a method for the analysis of latent variables based on discrete outcomes of questionnaires or other forms of tests. Given only the answers to a questionnaire or a test, IRT allows for the estimation of model parameters (e.g., question difficulties) and scores for the skill or miscellaneous latent factor measured by a test.

The basic IRT model assumes the parameters of the response model to be invariant: it does not include any data other than test or survey answers. Sometimes we do not believe this invariance to hold, and are interested in the relationship between the parameters and some external variable. For example, one might study the differences between a latent factor in various demographic groups [Hambleton, 1991, p126-142]. In this thesis, we are interested in treating the parameters of the IRT model as functions of a metric variable (such as time), as opposed to scalars in the normal IRT model.

Hastie described varying coefficient models [Hastie and Tibshirani, 1993], and in this thesis we will attempt to create varying coefficient IRT models, in some ways similar to how Zhang [Zhang and Wang, 2014] described varying coefficient additive models. Our method would improve over other approaches: for example, it allows the use of more data for the estimates for varying coefficients than a multi-group IRT model would.

Using our method, we aim to be able to determine and visualize a relationship between a metric variable, such as time, and latent variables or parameters in item response model.

In Chapter 2, we formally introduce IRT, and then extend it to allow for varying

coefficients. In Chapter 3, we create simulated data sets with varying coefficients, and attempt to use our model to find the simulation parameters. In Chapter 4, we apply the method to a real data set, using data from a large-scale personality test published by the University of Amsterdam [Smits et al., 2013].

2 Model and estimation

In this chapter we will introduce the theoretical foundations of item response theory, and extend them into varying coefficient models. Section 2.1 defines the graded-response IRT model we use, and an estimation method. In Section 2.2 we introduce our extension to allow for varying coefficients, and we describe the estimation of a varying coefficient IRT model using (full-information) maximum likelihood (FIML), kernel smoothing and k-fold cross validation.

2.1 Definitions and model

First, we will define IRT models as applied to binary-answer tests. Then, we redefine the terminology so that we can use the model to analyze tests for general psychological traits.

Consider a test consisting of I questions (*items*) with binary answers. We will assume without loss of generality that 1 represents a correct answer, and 0 an incorrect answer. To define our model, we will assume several things: [Birnbaum, 1968]

1. The questions measure a single common uni-dimensional *trait* for each person, often called *skill*, denoted by F.

2. Given a randomly selected individual with a value f for F, the probabilities of answering the questions of a test correctly are independent. This is called *local independence.*

3. We can model the probability of giving a positive response on question i by using a probability function, which depends on f and has one or more item-specific parameters, where x_i equals 1 for a positive response, and 0 for a negative response:

$$P(x_i = 1|F = f) = g_i(f). \tag{1}$$

   This function g is called the *item response function.*

We will additionally assume that F is normally distributed in the population, $F \sim N(\mu, \sigma^2)$. This is required for our varying coefficient model, where we will allow both μ and σ to vary.

A model that corresponds to the three numbered assumptions is called an item-response model (*IRT model*). A simple item response function is the one corresponding to the one-parameter logistic model, the "Rasch model": [Rasch, 1960]

$$P(x_i = 1|F = f) = \frac{1}{1 + e^{\delta_i - f}}. \tag{2}$$

Here, δ_i can be interpreted as the difficulty of a question. The higher δ_i, it becomes less likely that the question will be answered positively, given some fixed f. Note that the following notation is equivalent:

$$\text{logit}(P(x_i = 1|F = f)) = f - \delta_i, \tag{3}$$

where $\text{logit}(q) = \log(\frac{q}{1-q})$.

Equation 3 says that the logit of the probability is linear in the difference between a person's skill, and the difficulty of a question. We also see in Equation 2 that when skill and difficulty are equal, the probability of answering a question correctly is 1/2. Furthermore, both the difficulty and ability are measured on the same scale; an additional quantity of difficulty can be compensated with an equal quantity of ability without changing the probability of a correct response, i.e. $(f + c) - (\delta_i + c) = f - \delta_i$. So, is a very intuitive model that measures skill and difficulty on a single scale.

Another model is the two-parameter logistic model (*2PL*), defined by the item response function [Birnbaum, 1968]

$$P(x_i = 1|F = f) = \frac{1}{1 + e^{\alpha_i(\delta_i - f)}}. \tag{4}$$

We call α_i the *discrimination* parameter. Note that the Rasch model defined in Equation 2 is a special case of the 2PL model, with α_i set to 1. The main difference is that this model allows the questions to not only vary in difficulty, but also in how much impact

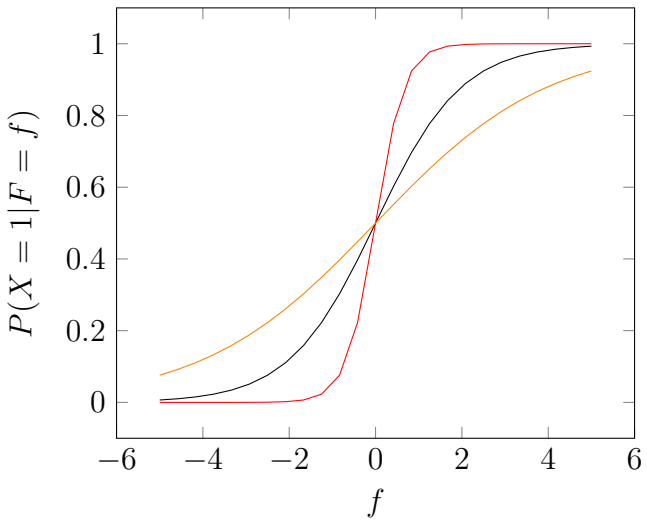

Figure 2: The probability curve for a 2PL model with difficulty $\delta_i = 0$ for various discriminations.

a small change in skill has on the probability of answering a question correctly: how well it discriminates between different levels of skill. For example, in Figure 2 the black line represents a question with discrimination $\alpha_i = 1$, the red line a question with discrimination $\alpha_i = 3$. With discrimination $\alpha_i = 3$, there is a very large difference in the probability of answering correctly between someone with skill -1 and skill 1.

2.1.1 Terminology

In the previous subsection we introduced the standard IRT terminology as applied to achievement tests. In this thesis however, we will study surveys that measure general psychological factors, such as introversion or inhibition. As such, the variable F no longer represents a skill, but it represents a one-dimensional measure of the target trait. We assume that the estimated trait is continuous and that for higher values of F it is more likely that a person responds to a question with a *trait-positive* response, either on a dichotomous or larger discrete scale.

2.1.2 Scoring polytomous items

The real-world data that we want to analyze consists of answers to questions with a polytomous answer scale. To be able to apply IRT, we need to extend our model from Section 2.1, which only dealt with dichotomous answers. The class of IRT models that deals with non-binary answers are polytomous item response models [Samejima, 1972].

There are various ways to model polytomous item-responses. We use a graded IRT model. It consists of sequential 2PL models [Samejima, 1997], and can be defined by:

$$P(x_i = k|F = f) = P(x_i \geq k|F = f) - P(x_i \geq k + 1|F = f), \tag{5}$$

for $k \in \{1,\ 2,\ \ldots,\ m\}$, where m is the number of points in our scale, so the number of answers to the questions. Here, each function $P(x_i > k|F = f)$ is either, for $k \in \{2,\ \ldots,\ m\}$ a 2PL model as in equation 4 or for $k = 1$ or $k = m + 1$ an edge case:

$$\begin{aligned} P(x_i \geq k|F = f) &= \frac{1}{1 + e^{\alpha_i(f - \delta_{ki})}} \\ P(x_i \geq m + 1|F = f) &= 0 \\ P(x_i \geq 1) &= 1, \end{aligned} \tag{6}$$

where α_i is a question-specific scale parameter, and δ_{ki} is an answer-specific difficulty parameter, for which we need the inequality $\delta_{2,i} < \delta_{\ldots,i} < \delta_{mi}$ to hold in order to ensure positive probabilities.

2.1.3 Model estimation

We can estimate the model using a marginal maximum likelihood[1] method [Bock and Aitkin, 1981] for the estimation of the item parameters.

For the estimation, we can assume the latent trait in the population to come from a standard normal distribution with $\mu = 0$, $\sigma = 1$. We can add a later step in the

[1]We can deal with missing items using Full Information Maximum Likelihood (FIML).

estimation process to allow for arbitrary and/or varying mean and standard deviation, by post-processing the estimated model. (See Section 2.2.1)

Given is a matrix $\mathbf{X}$ of scores from some survey or test taken by n people consisting of questions as described in the previous paragraphs, where x_{ni} is the response to question i given by person n. We assume that they are a function of unknown item parameters and a random latent trait. We can define the marginal (over the latent trait) log-likelihood of the full results as:

$$\log \mathcal{L}_{\mathbf{X}}(\zeta) = \sum_n \log\, p(\mathbf{x}_n|\zeta), \tag{7}$$

where ζ is a matrix consisting of vectors $\zeta_i = (\alpha_i, \delta_{i1} \dots \delta_{ik})$, one for each question. Probability density p is:

$$p(\mathbf{x}_n|\zeta) = \int \prod_i p(x_{ni}|\zeta_i, f)\phi(f)df, \tag{8}$$

with $\phi(\cdot)$ the standard normal density function. We can now maximize the (log-)likelihood to estimate the parameters of the model, given a matrix $\mathbf{X}$ of answers to questions:

$$\hat{\zeta}(\mathbf{X}) = \arg\max_{\zeta} \sum_n \log \int \prod_i p(x_{ni}|\zeta_i, f)\phi(f)df \tag{9}$$

If some columns $\mathbf{x}_n$ (so a column of answers corresponding to a person) of matrix $\mathbf{X}$ are thought to be more important, we can instead optimize a weighted (log-)likelihood, by defining weights w_n:

$$\hat{\zeta}(\mathbf{X}) = \arg\max_{\zeta} \sum_n w_n \log \int \prod_i p(x_{ni}|\zeta_i, f)\phi(f)df \tag{10}$$

To make it intuitively clear what such a weighting actually means, we can interpret a weighted model as equivalent to a larger unweighted model: for $w_n = \frac{a_n}{b_n} \in \mathbb{Q}$, the optimal parameters are equivalent to those in an unweighted model with repeated rows, with row n repeated $a_n \cdot \frac{z}{b_n}$ times for some z such that this number is integer. See Appendix B for a proof.

The weighted maximum likelihood calculation will be of use for the estimation of varying coefficient models.

2.2 Varying coefficients

In the previous section we introduced the graded response model. In this thesis, we are interested in extending the model to allow both model and population parameters to vary over some metric variable, such as time. In all sections of this thesis we will assume t to be from a finite equidistant discrete interval, simplifying to $t \in \{1, \ldots, T\}$ for a maximum value T. The described method is still valid for non-equidistant discrete intervals, since there need not be data for all values t. We can assume discreteness because real-world data sets are always finite in size, and can therefore be discretized.

With varying coefficients, the latent trait of a population at value t is no longer assumed to be distributed with a (standard) normal distribution, but along a distribution that varies through time, which we will denote by a superscript:

$$F^t \sim \mathcal{N}(\mu^t, \left(\sigma^t\right)^2). \tag{11}$$

Additionally, the 2PL models defining the likelihood of various graded responses also have time-dependent parameters:

$$P(x^t \geq k|\ f^t) = \frac{1}{1 + e^{\alpha^t (f^t - \delta_k^t)}}. \tag{12}$$

This addition allows for the use of IRT techniques for the analysis of longitudinal data. With such a model, one could study the variation of population traits over time, while simultaneously analyzing how item parameters change. For example, in a survey on social activity, a population might become more social, while at the same time indicating in the survey that they are spending less time on face-to-face contact (due to, for example, the rise of the internet). Ideally, our model should allow for the detection of both simultaneously.

2.2.1 Model estimation

Because we don't see a straight-forward way to directly extend the (marginal) maximum likelihood method to the case with varying coefficients, we have chosen to use kernel smoothing to estimate the parameters over time [Wand and Jones, 1994].

Lacking a reason to use something more complicated, we have chosen to use a Gaussian kernel:

$$K\left(z_0,\, z\right) = \frac{1}{b^2} \exp\left(-\frac{\left(z_0 - z\right)^2}{2b^2}\right), \tag{13}$$

where K represents our smoothing kernel, and b is the smoothing bandwidth.

First, we assume that the latent trait in the population is at all times distributed according to a standard normal distribution. We will correct for this assumption in a later step.

To estimate the model parameters, we use a weighted maximum likelihood estimation, with the algorithm described in Section 2.1.3 as basis, using weights from the Gaussian kernel smoother.

At time t_0, we use smoothing to generate weights $\{w_{t_n}\}_n$ for all subjects n (that is, for all unique times t_n):

$$\forall n \in \{1,\, \ldots,\, N\} :\, w_{t_n}^{t_0} = \frac{K(t_0,\, t_n)}{\sum_{t^* \in T} K(t_0,\, t_n^*)} \tag{14}$$

For every given item, we can now calculate the parameter estimates $\hat{\alpha}^{t_0}$ and $\hat{\delta}_k^{t_0}$ at t_0, with $\hat{\zeta}$ consisting of rows $\hat{\zeta}_i = (\hat{\alpha}_i, \hat{\delta}_{1i} \ldots \hat{\delta}_{ki})$ corresponding to questions as before:

$$\hat{\zeta}^{t_0}(\mathbf{X}) = \arg\min_{\zeta} \sum_n w_{t_n}^{t_0} \log \int \prod_i p(x_{ni} |\, \zeta_i,\, f) \phi(f) df. \tag{15}$$

Note that the only change compared to the version without varying coefficients is the weight. This allows us to re-use existing IRT estimation methods. The effect of this weight is that it allows us to calculate parameter estimates at any point t_0, even at points without observations. In addition, since the kernel smoothing function is smooth

(differentiable), this results in the parameter estimates being smooth over the metric variable as well.

We use the EM-algorithm [Bock and Aitkin, 1981] for the minimization of Equation 15, as implemented by the R-package MIRT [Chalmers, 2012a]. Internally, the minimization is done using BFGS [Broyden, 1970].

This calculation gives us parameter estimates, but still under the assumption that the latent trait is distributed in the population according to a standard normal distribution. A researcher is generally interested in the question whether the latent trait distribution changed over time, or if some item(s) changed in popularity/difficulty. We can exploit the relation between the parameters of the latent trait distribution and the model parameters to diagnose the degree to which one or the other changes, and to thereby correct for the assumption of standard normality of F.

We will first demonstrate that a shift of the mean of the latent trait distribution results in a simple translation of all parameters δ_k. First observe that shifting the mean from μ to $\mu + c$ increases all f by c, and then rewrite:

$$\begin{aligned} P(x \geq k|f, \delta_k, \ldots) &= \frac{1}{1 + e^{\alpha(f-\delta_k)}} \\ &= \frac{1}{1 + e^{\alpha((f+c)-(\delta_k+c))}} \\ &= P(x \geq k|f + c, \delta_k + c, \ldots). \end{aligned} \tag{16}$$

An example of the transformation (with parameters changing over time) can be found in Figures 3 and 4. In Figure 3, all the δ_i parameters decrease over time, meaning that questions seemingly get easier. In this case, it seems that it is more likely that the population gets a higher trait value instead. We can use the transformation and assume μ to be linear from 0 to 2, as seen in Figure 4. Now the question parameters stay fairly horizontal over time, and the population trait mean μ increases instead.

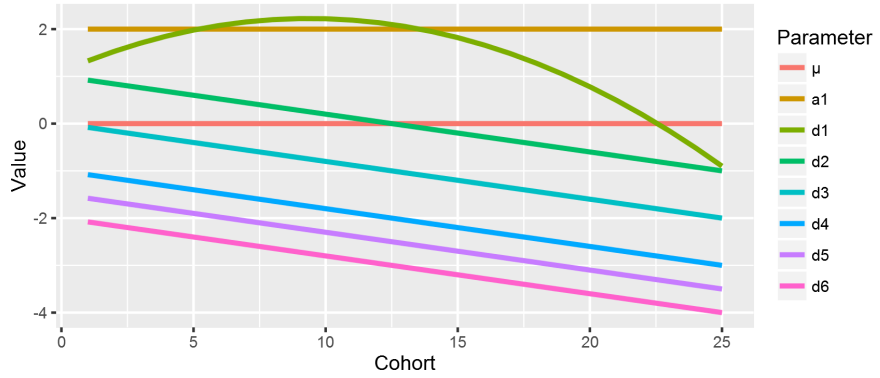

Figure 3: Sample parameters before transformation. We see that all δ_i decrease over time.

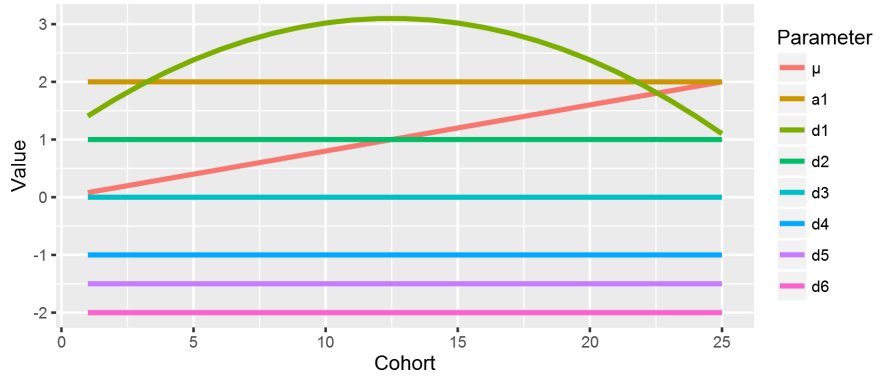

Figure 4: Sample parameters after a transformation that increases μ linearly over time. We see that the δ_i are now horizontal.

Similarly, we can change the standard deviation from σ to $c' \cdot \sigma$, noting that doing so results in scaling all f by c':

$$\begin{aligned} P(x \geq k|f, \delta_k, \alpha) &= \frac{1}{1 + e^{\alpha(f-\delta_k)}} \\ &= \frac{1}{1 + e^{\frac{\alpha}{c'}(c'f - c'\delta_k)}} \\ &= P(x_n \geq k|c'f_n, c'\delta_k, \frac{\alpha}{c'}). \end{aligned} \tag{17}$$

To allow the population trait mean μ and standard deviation σ to vary in our estimated varying coefficient IRT models, we can exploit this relationship by taking parameter estimates from equation 15, and defining two sum-of-squares loss functions that assume that item parameters should generally be constant over time:

$$L_1\left(\left\{\sigma^t\right\}_t\right) = \sum_{i=1}^{I}\sum_{t=1}^{T}\left(\log\left(\frac{\alpha_i^t}{\sigma^t}\right) - \frac{1}{T}\sum_{t'}^{T}\log\left(\frac{\alpha_i^{t'}}{\sigma^t}\right)\right)^2, \tag{18}$$

$$L_2\left(\left\{\mu^t\right\}_t\right) = \sum_{k=1}^{7}\sum_{i=1}^{I}\sum_{t=1}^{T}\left((\delta_{ki}^t + \mu^t) - \frac{1}{T}\sum_{t'}^{T}\left(\delta_{ki}^{t'} + \mu^{t'}\right)\right)^2. \tag{19}$$

Minimizing loss function L_1 allows σ^t to absorb variation in $\{\alpha_i^t\}_t$, while L_2 allows μ^t to absorb variation in $\{\delta_{ki}^t\}_t$. For example, if all estimated parameters $\{\delta_{ki}^t\}_t$ go down as t goes up, meaning people give higher responses to all questions over time, it is more likely that the population trait mean has shifted over time than it is for all questions to have become easier. We try to determine the optimal trait means μ^t such that the difficulty parameters $\{\delta_{ki}^t\}_t$ are as constant over time as possible. The same reasoning holds for loss function L_1 for σ^t in Equation 18, except it corrects for varying "spread" in estimated $\{\delta_{ki}^t\}_t$ parameters and fluctuations in $\{\alpha_i^t\}_t$ parameters.

Noting that the functions are independent, and keeping in mind Equation 17, we note that we can combine the loss functions and minimize a single value, resulting in $\check{\mu} = \{\check{\mu}^t\}_t$ and $\check{\sigma} = \{\check{\sigma}^t\}_t$:

$$\check{\mu},\, \check{\sigma} = \arg\min_{\{\mu^t\}_t, \{\sigma^t\}_t} L_1\left(\left\{\sigma^t\right\}_t\right) + L_2\left(\left\{\mu^t\right\}_t\right), \tag{20}$$

We can then modify the parameters as in Equations 16 and 17 to find the optimal adjusted parameter estimates, so that for all t, k and i:

$$\begin{aligned} \widetilde{\delta_{ki}^t} &= \check{\sigma}^t\delta_{ki}^t - \check{\sigma}^t\check{\mu}^t, \\ \widetilde{\alpha_i^t} &= \frac{\alpha_i^t}{\check{\sigma}^t} \end{aligned} \tag{21}$$

$$\begin{aligned}\widetilde{\mu^t} &= \check{\sigma}^t \check{\mu}^t \\ \widetilde{\sigma^t} &= \check{\sigma}^t\end{aligned}$$

Depending on our assumptions, we can simplify the process by assuming that σ does not change, reducing this process to just the minimization of L_2 and the adjustment step as in Equation 16. An example of such a transformation was illustrated in Figures 3 and 4, where the transformation that was performed was optimal, meaning it minimized the loss function.

To make this minimization tractable for large values of N, we use a 3-degree polynomial over t to best estimate σ^t with the constant coefficient set to 1, and a similar 3-degree polynomial for μ^t with the constant coefficient set to 0. Using such a polynomial reduces the minimization free variables from t (every time has its own mean trait) to just 3. If there is domain knowledge about the potential shapes over time of μ^t and its standard deviation over time, one could substitute another suitable function for the polynomial.

For the actual minimization of the loss function, we use a generic numeric optimizer. Special care should be taken to avoid local optima. We used the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [Broyden, 1970], and to prevent local optima we started from the best estimate from a 3-dimensional grid covering the most likely parameter values.

2.2.2 Bandwidth selection

In order to determine the bandwidth of the Gaussian kernel smoother defined in Section 2.2.1, we use 10-fold cross-validation. This consists of splitting the data into 10 equally-sized parts, and then predicting each part based on a model created from the other 9 parts. From those 10 estimations on the predicted data, we take the mean of the log-likelihoods.

We select the bandwidth based on the deviance[2], which we define as the mean log-likelihood of the cross-validated solutions at a certain bandwidth, divided by -2. The bandwidth with the lowest deviance is selected for use in all experiments.

[2]technically, the deviance relative to some constant saturated model.

Given that we have no exact formula for the optimal bandwidth, we cannot perform explicit power calculations. However, we can calculate very rough estimates of required sample sizes using a method laid out in Section 4.3.4.

2.2.3 Confidence bands

Having estimated model and population parameters for the full data as described in Section 2.2.1, we might be interested in the estimation errors of these parameters. To determine these, we use a bootstrap re-sampling algorithm: first, we create 50 re-sampled populations, each sampled from the original data (with replacement), and while keeping the number of people per moderator variable the same. We then calculate the parameter estimates for each of these populations at all points in time, that is, for all t_n. Then, for every point in time, we calculate the standard error, based on the 50 population estimates at the optimal bandwidths. Based on this standard error, we can, for example, draw point-wise 95% confidence bands around the kernel smoothed parameter estimates.

With regard to limitations of the interpretation of these confidence bands, please see the discussion in Chapter 5.

3 Simulated results

In order to test the power and accuracy of our model, we have created several simulated data-sets with parameters that vary in various ways. Our goal is to use varying coefficient models to recover as much information about the varying parameters as we can. In this chapter, we will describe the generation of this data, the use of our model, and we will look into the model performance. Additionally, we try to determine the influence of sample size on the optimal bandwidth size for the real-world data. We start from a simple data set and model, and gradually add complicating factors.

3.1 Data generation

We have chosen to generate simulated data that is similar to the data we use in Chapter 4. This means that we have a total of 8954 people, divided into 25 cohorts of varying sizes, as seen in figure 5. Of course, while we use cohorts for these simulations, please note that the method is equally valid if all people have a unique t_n.

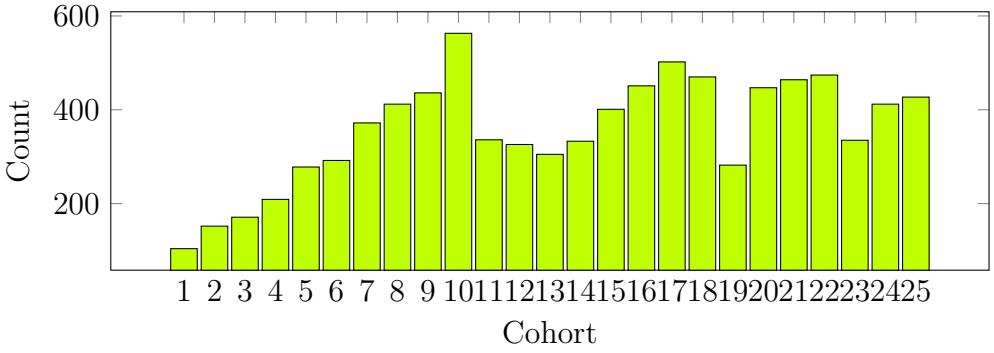

Figure 5: An overview of the cohort sizes.

The generation process of the answer matrix is an inverse transform sampling process. Given the parameters of the distribution of f_n and given the model parameters α_i and δ_{ki}:

1. For every person n, we generate a trait value f_n from a normal distribution with parameters specific to an experiment.

2. For each question i and for every person n, we pick a uniformly random value $0 \leq p_{ni} \leq 1$.

3. For every p_{ni}, determine where it lies in the intervals defined by Equations 6, given parameters f_n, α_i and δ_{ki}.

To illustrate step 3, consider the following parameters for some question:

α	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6
1	2	1	0	-1	-1.5	-2

If we generated $f = 0$ in step 1, we can fill these values in into Equations 6, and find the following probabilities:

$P(x \geq 8)$	$P(x \geq 7)$	$P(x \geq 6)$	$P(x \geq 5)$	$P(x \geq 4)$	$P(x \geq 3)$	$P(x \geq 2)$	$P(x \geq 1)$
0	0.12	0.18	0.27	0.5	0.73	0.88	1

Now, if in step 2 we generate a p between 0 and 0.12, we assign an answer of $x = 7$. If we generate a p between 0.12 and 0.18, we assign an answer of $x = 6$, and so on.

The parameters in the graded response models (i.e., the question difficulties) and the trait distributions are varied between our experiments.

3.2 Experiments

We have performed several experiments, generating data sets that vary in different ways, and applying our methodology to try and reproduce the generation parameters. We start with simple data sets to verify the basic principles of the method, and slowly continue to more complex experiments.

Experiment 1: no varying coefficients

For all cohorts the trait value is generated from the standard normal distribution: $F_n \sim \mathcal{N}(0, 1)$. Additionally, all of the graded response model parameters are the same for all cohorts and for all questions:

α	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6
1	2	1	0	-1	-1.5	-2

We first attempted to determine the optimal bandwidth. However, as Figure 6 shows, there was no optimal bandwidth: for growing bandwidths, the deviance converges downwards. This is understandable: since all of the parameters are constant over time, we can estimate the best model by using the full data for every cohort. This happens for a very large bandwidth. We proceeded by first applying the algorithm described in Section 2.1.3 with a very high bandwidth (100), and then with a lower bandwidth of 2.0, to see what we would find then.

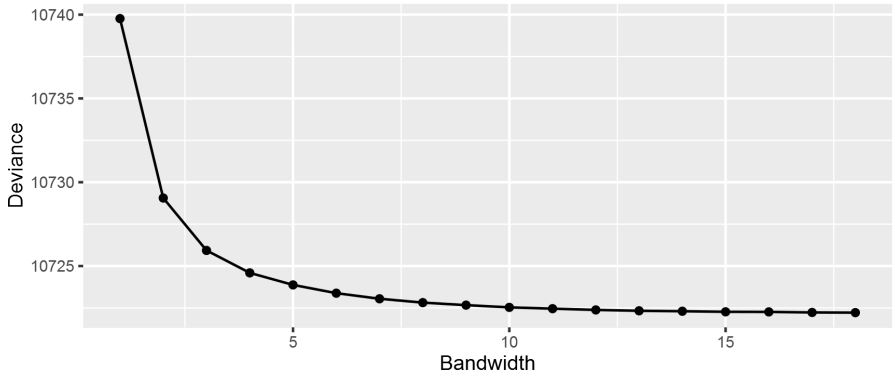

Figure 6: Deviance for various bandwidths in experiment 1. Since the coefficients are constant, it works best to assume a very high bandwidth, using all cohorts for every point estimate.

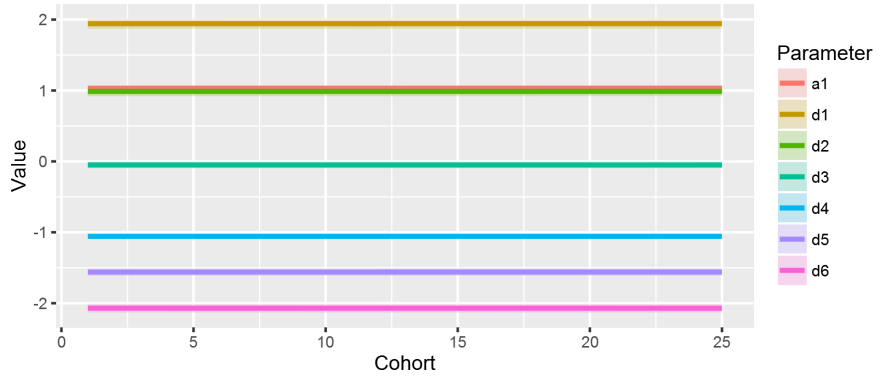

Figure 7: Varying coefficients for the first item in experiment 1, with bandwidth 100.0, which is high enough to get constant estimates. We see that all parameters are reproduced.

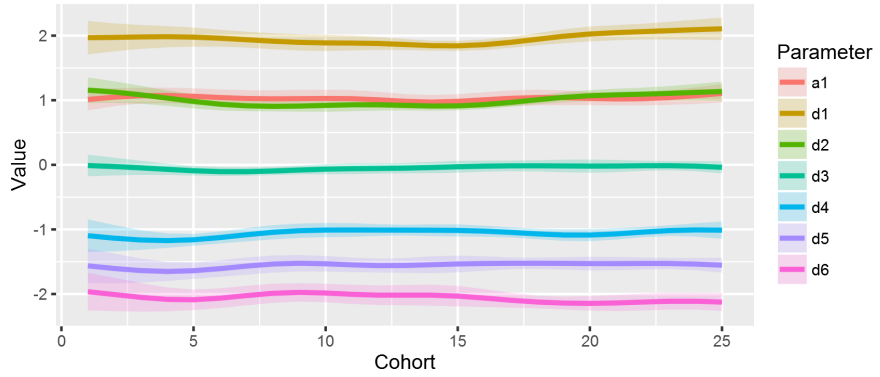

Figure 8: Varying coefficients for the first item in experiment 1, with (non-optimal) bandwidth 2.0. There is some variation, but the estimates are always close to the correct values.

For bandwidth 100.0 (Figure 7), we see straight lines with very narrow point-wise confidence bands. The parameters are all very close to their true values. Note that this is effectively an IRT model estimation without varying coefficients.

For bandwidth 2.0 (Figure 8), we see that all of the parameters are also estimated quite well: the true values are almost everywhere within the confidence bands. The confidence bands are relatively small, indicating that based on these outputs the true model parameters are likely constant. Using a sub-optimal bandwidth shows more

variance in the data, both in terms of the estimates and the confidence bands.

We also see that the confidence bands are wider in some regions than in others. This is largely the result of the smaller sample sizes on those regions. This is most significant in the boundaries, where we can only smooth in one direction, thereby weighing the central cohort relatively heavier.

Experiment 2: varying δ_1

For all cohorts the trait value is generated from the standard normal distribution: $F_n \sim \mathcal{N}(0, 1)$. The δ_1 parameter varies over time/cohort, as seen in the following table, where t is the cohort:

α^{t_n}	$\delta_1^{t_n}$	$\delta_2^{t_n}$	$\delta_3^{t_n}$	$\delta_4^{t_n}$	$\delta_5^{t_n}$	$\delta_6^{t_n}$
2	$1.1 + \frac{8}{625}(25 - t_n)t_n$	1	0	-1	-1.5	-2

First, we determined the optimal bandwidth. As seen in Figure 9, the bandwidth with the lowest deviance is around 2.4. Then, after estimating the full model using the algorithm described in Section 2.1.3 using that bandwidth, we looked at the estimated values for one of the questions (see Figure 11).

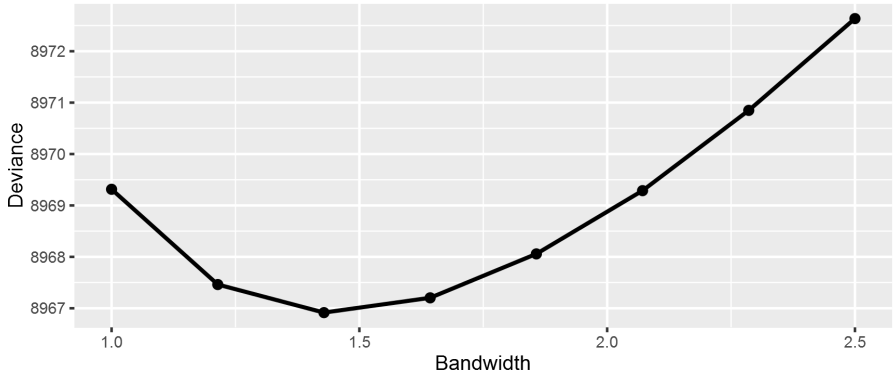

Figure 9: Model deviance for various bandwidths in experiment 2. We see (after additional more precise calculations) that the deviance is minimized at a bandwidth of approximately 1.5.

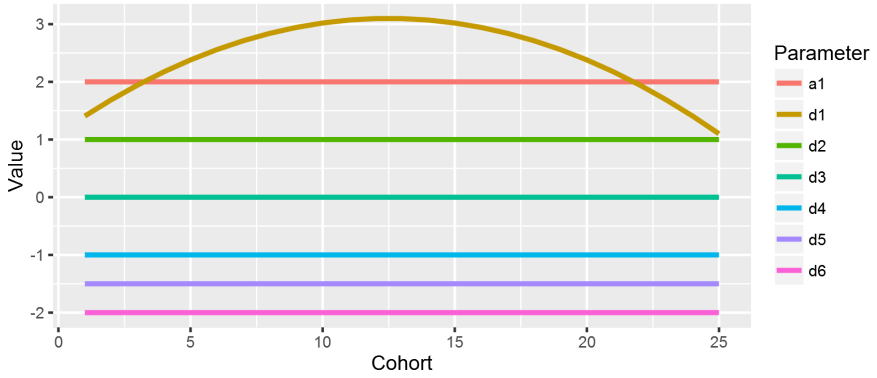

Figure 10: True parameters for all of the items in experiment 2, as used for data generation.

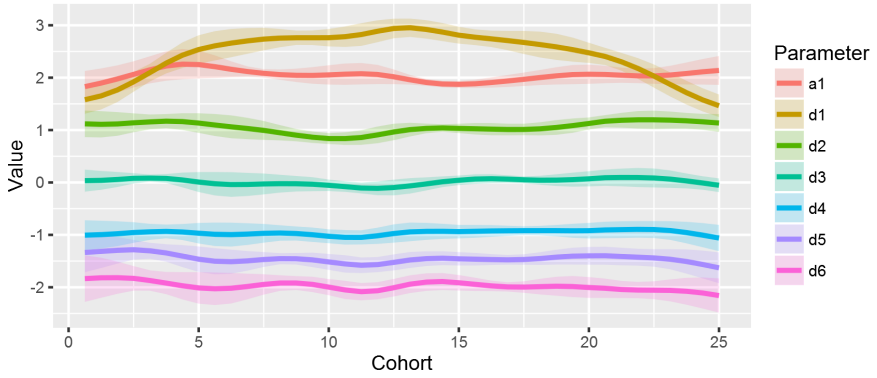

Figure 11: Varying coefficients for the first item in experiment 2, using the optimal bandwidth determined in Figure 9. The shape of δ_1 resembles the input shape.

We clearly reproduce the trend in the true model parameters. The top of δ_1 is a somewhat lower than the real parameter, but that's expected due to the kernel smoothing.

In order to determine the effect of sample size on the accuracy of the model parameters, we performed the full method various times, on data sets generated as in the rest of this chapter, but by scaling the number of participants in each cohort. Then, given model parameters, we compare the calculated parameters to the true values by calculating the square root of the mean square error (RMSE) of the δ_i parameters, after adjusting for the α_i's:

$$\text{RMSE} = \sqrt{\frac{1}{N(K-1)I} \sum_{n=1}^{N} \sum_{k=1}^{K-1} \sum_{i=1}^{I} \left(\hat{\alpha}_i \left(\hat{\delta}_{ki}^{\hat{t}_n} - \hat{\mu} \right) - \alpha_i \left(\delta_{ki}^{t_n} - \mu \right) \right)^2} \tag{22}$$

In simple words, it is the mean squared distance between the true and the estimated values of the model parameters. Table 1 and Figures 12 and 13 show the results. We see that for larger sample sizes, the estimated parameters get closer to the true values, since the RMSE goes down (with some variance).

Sample size	Optimal bandwidth	RMSE
2986	2.3	0.0494
4477	2.0	0.0432
5968	1.8	0.0447
8954	1.5	0.0383
13430	1.4	0.0371
17908	1.4	0.0381

Table 1: The optimal bandwidth and RMSE for various samples of data generated based on the parameters in experiment 2.

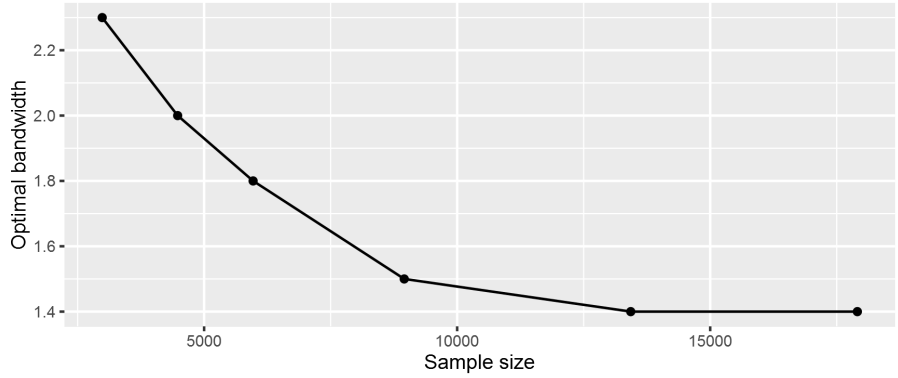

Figure 12: The optimal bandwidth for various samples of data generated based on the parameters in experiment 2.

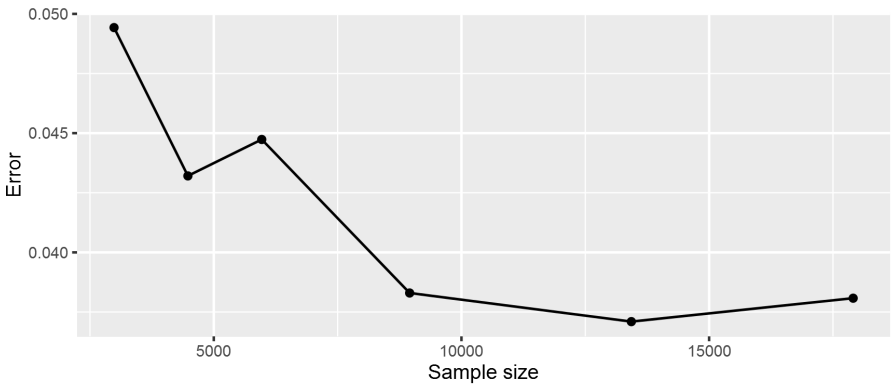

Figure 13: The RMSE for various samples of data generated based on the parameters in experiment 2.

Experiment 3: varying F_n in data, but fixed in model

In the previous experiments, we assumed that the mean of the distribution of F_n was constant throughout the years. We can, however, also let that vary. If we then estimate the model while still assuming that this mean is constant, that variation will be captured by other model parameters. Here, we chose the same model parameters as in experiment 1:

α	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6
1	2	1	0	-1	-1.5	-2

But now, $F_n \sim \mathcal{N}(-2 + \frac{8}{625} t_n \cdot (25 - t_n), 1)$. We determine the optimal bandwidth (Figure 14) and use it to calculate the varying coefficients (Figure 16). We see that, since the distribution of F_n is assumed to be fixed, the variation shows up in the other estimates. The RMSE is 0.20; this is very high because all estimates are far from the original values, as seen in the figures.

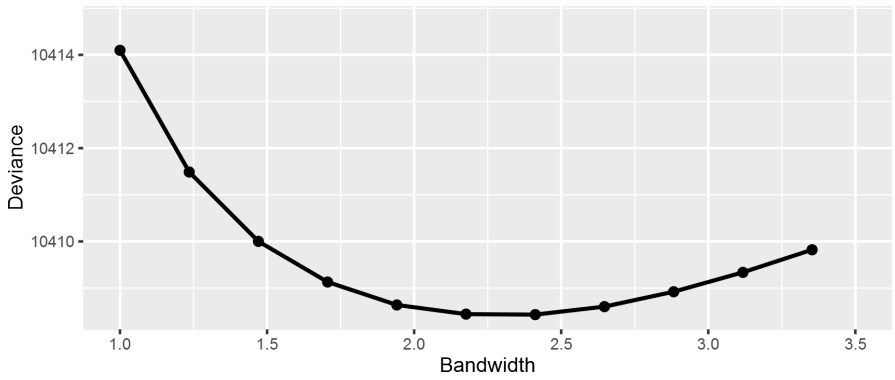

Figure 14: Model deviance for various bandwidths. The optimum is around 2.3.

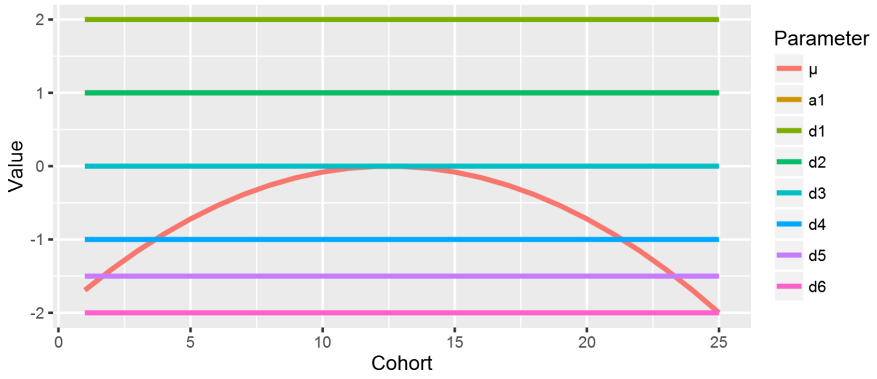

Figure 15: True parameters for all of the items in experiment 3, as used for data generation.

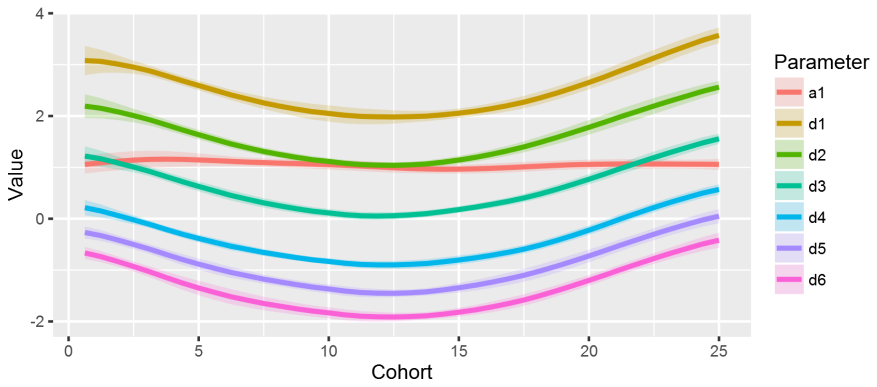

Figure 16: Varying coefficients fort he first item in experiment 3.

Experiment 4: varying F_n in data, varying in model

We now generate data with the skill variable distributed normally with a mean that varies over time. That is, for person n taking the test at time t,

$$F_n \sim \mathcal{N}(-1 + \frac{t_n}{12.5},\ 1), \tag{23}$$

The other model parameters are fixed:

α	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6
1	2	1	0	-1	-1.5	-2

We then estimate the model assuming that the mean trait value varies, as described in Section 2.2.1. This results in Figure 17. This is clearly an improvement over Figure 16, where the variation in μ was picked up by all of the other parameters. In this model, the drift of the mean population trait is estimated quite well, and the other parameters are fairly constant. There is some variation at the edges, because of the effect of boundaries on kernel smoothing. The RMSE is 0.032.

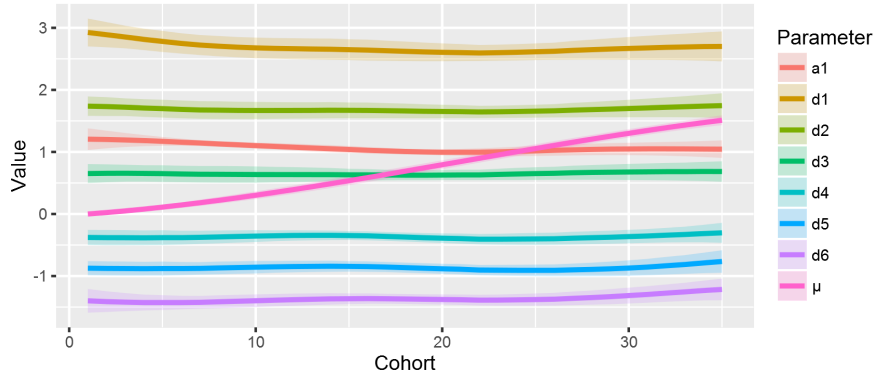

Figure 17: Varying coefficients for the first item in experiment 4. We used an optimal estimation bandwidth of 3.5.

Experiment 5: varying both F_n and model parameters

In this experiment, we again generate data with a distribution for F with a mean that varies over time:

$$F_n \sim \mathcal{N}(-1 + \frac{t_n}{12.5},\ 1). \tag{24}$$

But now, one of the parameters of the measurement model also varies:

α^{t_n}	$\delta_1^{t_n}$	$\delta_2^{t_n}$	$\delta_3^{t_n}$	$\delta_4^{t_n}$	$\delta_5^{t_n}$	$\delta_6^{t_n}$
1	2	1	0	-1	-1.5	$-3.6 + \frac{8}{625} t_n \cdot (25 - t_n)$

We then estimate the model under the assumption that the mean trait indeed varies over time, as described in Section 2.2.1. For this estimation, we assume that the standard deviation is still fixed at 1. This results in Figure 19. We see that the trend in F is clearly picked up by the model, and that the rest of the parameters are properly adjusted, like we saw in experiment 4 and 5. The variation in parameter δ_6 is retained after the adjustment. The RMSE is 0.052.

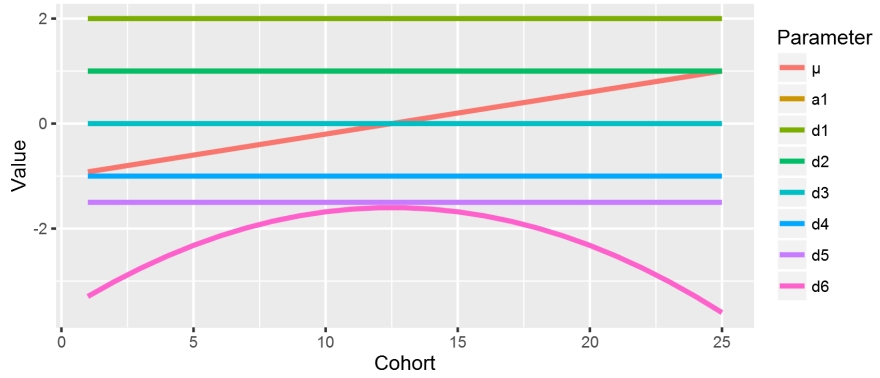

Figure 18: True parameters for experiment 5, as used for data generation. α is 1, and is invisible since δ_2 is drawn on top of it.

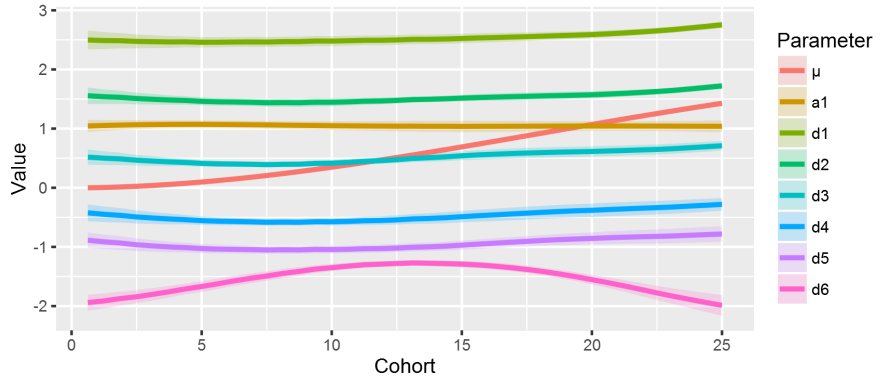

Figure 19: Varying coefficients for the first item in experiment 5. Estimation bandwidth was 3.5, so we can compare it to experiment 4.

Experiment 6: varying both F_n and individual model parameters

So far, all experiments have consisted of 14 identical questions, at least in terms of their parameters. Now, we will simulate a test with questions of varying difficulties. Let i be the question number, $0 \leq i \leq 13$.

We again generate data with a distribution for F_n with a mean that varies over time:

$$F_n \sim \mathcal{N}(-1 + \frac{t_n}{12.5},\ 1). \tag{25}$$

as it was in the previous two experiments. The model parameters now depend on both the time t and the question i:

$\alpha_i^{t_n}$	$\delta_{1,i}^{t_n}$	$\delta_{2,i}^{t_n}$	$\delta_{3,i}^{t_n}$	$\delta_{4,i}^{t_n}$	$\delta_{5,i}^{t_n}$	$\delta_{6,i}^{t_n}$
1	$2+\frac{i}{6}$	$1+\frac{i}{6}$	$0+\frac{i}{6}$	$-1+\frac{i}{6}$	$-1.5+\frac{i}{6}$	$-3.6+\frac{8}{625}t_n \cdot (25-t_n)$

Figure 20 shows that we are able to estimate all the question parameters for all questions, and that the estimate for the population mean is still accurate as well. The RMSE is 0.061.

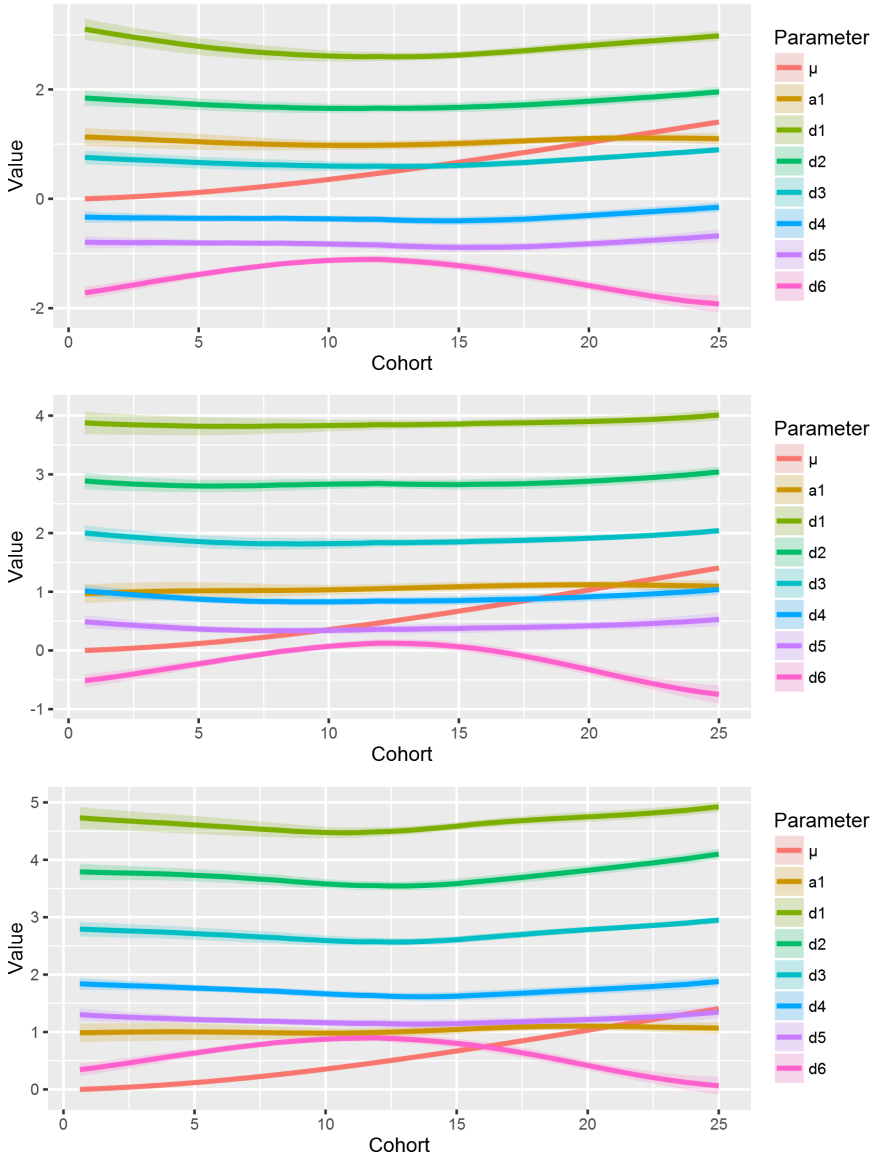

Figure 20: Varying coefficients for experiment 6. The first figure is for question 1, the second for question 8, and the third for question 13. Observe that for all questions the parameters are estimated quite accurately. The bandwidth was 3.2.

Experiment 7: varying F_n and a fluctuating model parameter

This experiment has a question parameter that fluctuates over time. Our goal is to see if we can capture its fluctuation, or if it gets smoothed away.

We generate data with a distribution for F_n that has a mean that varies over time:

$$F_n \sim \mathcal{N}(-1 + \frac{t_n}{12.5},\ 1), \tag{26}$$

and δ_6 fluctuates over time between -4 and -2:

α^{t_n}	$\delta_1^{t_n}$	$\delta_2^{t_n}$	$\delta_3^{t_n}$	$\delta_4^{t_n}$	$\delta_5^{t_n}$	$\delta_6^{t_n}$
1	2	1	0	-1	-1.5	$-3 + \cos(\frac{\pi}{4} t_n)$

For clarity, these input parameters are illustrated in Figure 21.

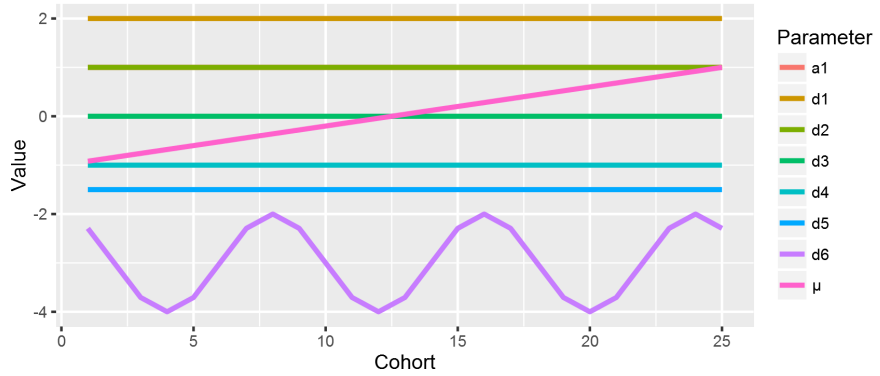

Figure 21: True parameter values as specified for experiment 7, over time.

We now estimate the parameters with various bandwidths, to see the effect of kernel smoothing. The fluctuations of the sixth parameter have a fairly small width, so we expect them to disappear quickly with increasing bandwidths. Figure 22 shows that this is indeed the case. Note that all values are shifted by 1 compared to the true parameter values in Figure 21; this is because our simulation assumes[3] the mean trait value starts at 0. We see that at the bandwidth 3.5, the fluctuations are smoothed away.

[3]Without loss of generality: starting at another value shifts all parameters by an equal amount.

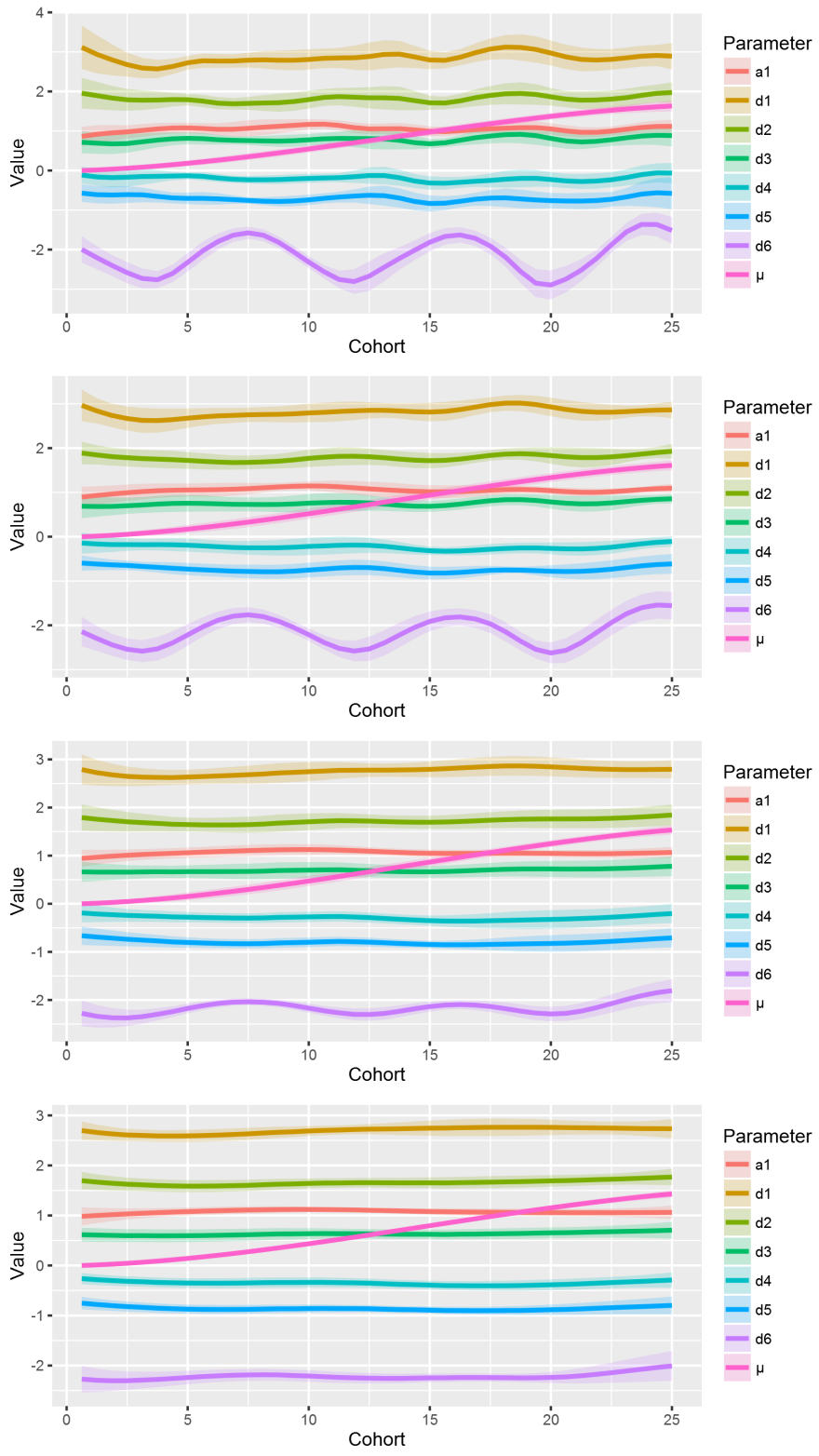

Figure 22: Varying coefficients for the first question in experiment 7. The first figure is for a bandwidth of 1.0 (RMSE 0.044), the second for 1.5 (optimal; RMSE 0.053), the third for 2.5 (RMSE 0.072), and the fourth for 3.5 (RMSE 0.082).

Experiment 8: varying F_n and σ^{t_n} in data, and varying in model

We now generate data with a distribution for F_n with both a mean and a standard deviation that vary over time:

$$F_n \sim \mathcal{N}\left(-1 + \frac{t_n}{12.5},\ \left(0.5 + \frac{t_n}{40}\right)^2\right). \tag{27}$$

The other model parameters are fixed, except for α_i, which varies based on the question, i:

α_i	$\delta_{1,i}$	$\delta_{2,i}$	$\delta_{3,i}$	$\delta_{4,i}$	$\delta_{5,i}$	$\delta_{6,i}$
$1 + \frac{i}{25}$	2	1	0	-1	-1.5	-2

This time, we no longer assume that σ^{t_n} is constant, but we estimate it freely as described in Section 2.2.1. The results are in Figures 23 and 24; the corrected estimation works well. We see that both μ^{t_n} and σ^{t_n} are recovered from the model, resulting in stable question parameters as intended. The RMSE is 0.057.

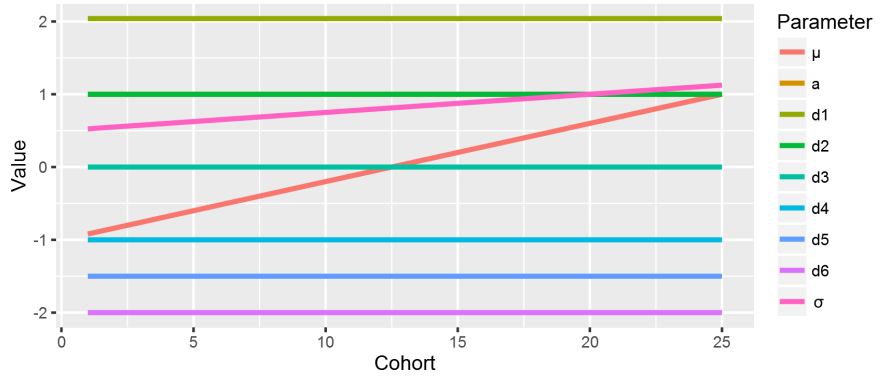

Figure 23: True parameter values as specified for question 1 of experiment 8, over time.

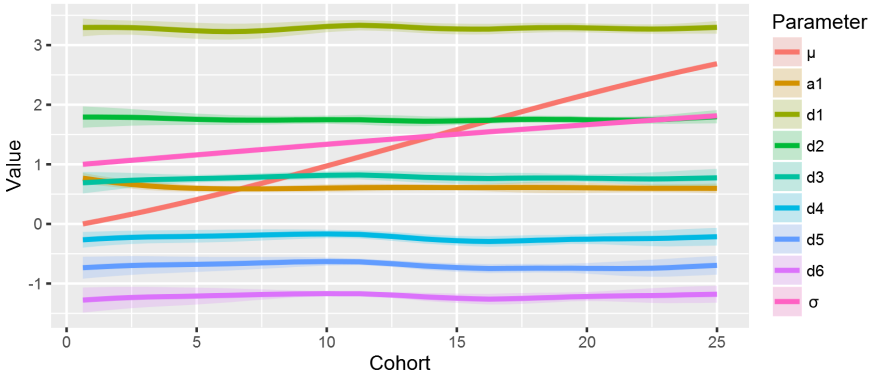

Figure 24: Varying coefficients for question 1 of experiment 8, over time. The bandwidth used was 3.5.

4 Application

4.1 Data-set

To apply varying coefficient models, we use the data from a large-scale personality test published by the University of Amsterdam [Smits et al., 2013]. The test consists of scores of 8954 psychology freshmen from the University of Amsterdam between the years 1982 and 2007, measuring the "Vijf Persoonlijkheidsfactoren" ("[Big] Five Personality-factors"). This includes the following factors:

- Extroversion;
- Agreeableness;
- Conscientiousness;
- Neuroticism;
- Openness to Experience.

The test consists of a total of 70 items, and for all the test takers we know the year (cohort) they took the test. The students are distributed over the cohorts as per Figure 25.

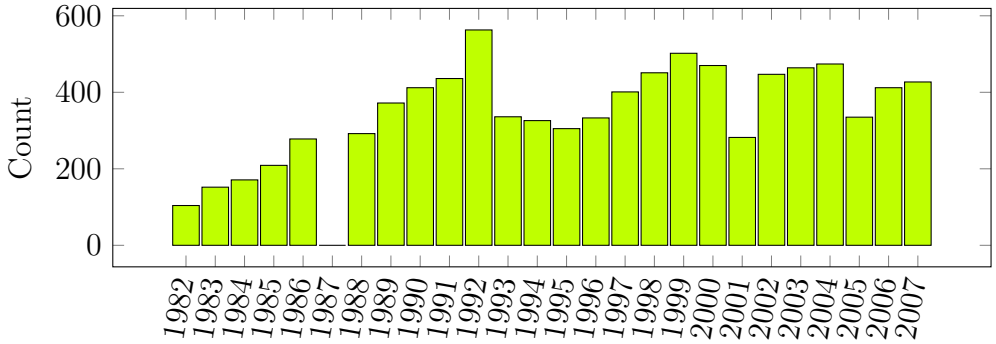

Figure 25: An overview of the cohort sizes.

For the experiments in this chapter, we focus on the items corresponding to one of the factors, namely extroversion. Results for the other factors can be found in the appendix.

4.2 Adjustments to data-set

The resulting data-set consists of the answers to 14 questions (items) by 8954 students. For copyright reasons, we are unable to reproduce the item text here; see the original publication [Smits et al., 2013] for details. Answers were coded 1-7, representing (in original Dutch, and a literal translation):

1. "absoluut niet" / "absolutely not"
2. "tamelijk slecht" / "somewhat bad"
3. "meer niet dan wel" / "more negative than positive"
4. "middenpositie" / "middle position"
5. "wel enigszins" / "somewhat"
6. "vrij goed" / "quite good"
7. "goed" / "good"

We interpreted these answers as a 1-7 Likert scale.

We found some miscoded answers, where the score on the supposed Likert scale was not within 1-7, but was 8 or 9. The code book accompanying the data set did not mention this possibility. Since we did not see a clear pattern in the occurrences, we chose to treat such values as NA, and assume that they were meant as special cases of NA, as is often the case with such values.

4.3 Results

We applied our implementation from section 2.2.1 to the Extroversion factor of the Big Five Personality test, assuming that all parameters and the population trait mean could vary.

First, we determined the optimal bandwidth. Then we looked at the resulting parameter estimates. Finally, we look at the influence of sample size on the optimal bandwidth.

4.3.1 Bandwidth determination

We performed this 10-fold cross-validation on the extroversion data set. The optimal bandwidth we found was around 3.0, as seen in Figure 26. The deviance graph follows a hockey-stick form as expected: very low bandwidths do not smooth at all, resulting in models that are extremely over-fitted; very high bandwidths smooth all observations into a single model, losing any parameter variation over time.

This bandwidth is of course only optimized for this specific data set; as it is dependent on the presence of trends and variation in the underlying data. In Section 4.3.4 we look at the influence of sample size on the optimal bandwidth size.

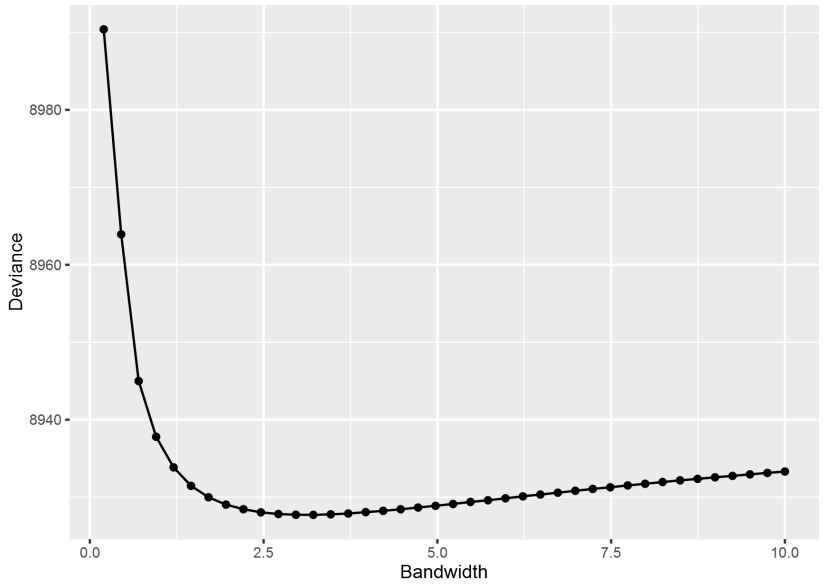

Figure 26: Deviance calculated using 10-fold cross-validation on the real-world data set for various bandwidths. The optimal bandwidth is around 3.0, where the deviance is minimized.

4.3.2 Parameter estimates

Using the kernel smoothing bandwidth 3.0 as found in the previous section, we calculated the model parameters. The results can be found in Figure 27. We see a very small

downward tendency for μ^t (see Figure 28 for a larger view), indicating that the mean extroversion *increased* a small amount over time. σ^t seems to be constant over time.

Most question parameters are fairly constant (or have a very low sample size, such as the extremely varying variants of δ_1); we do however see some narrowing of question 11, meaning that people over time tended to give less extreme answers.

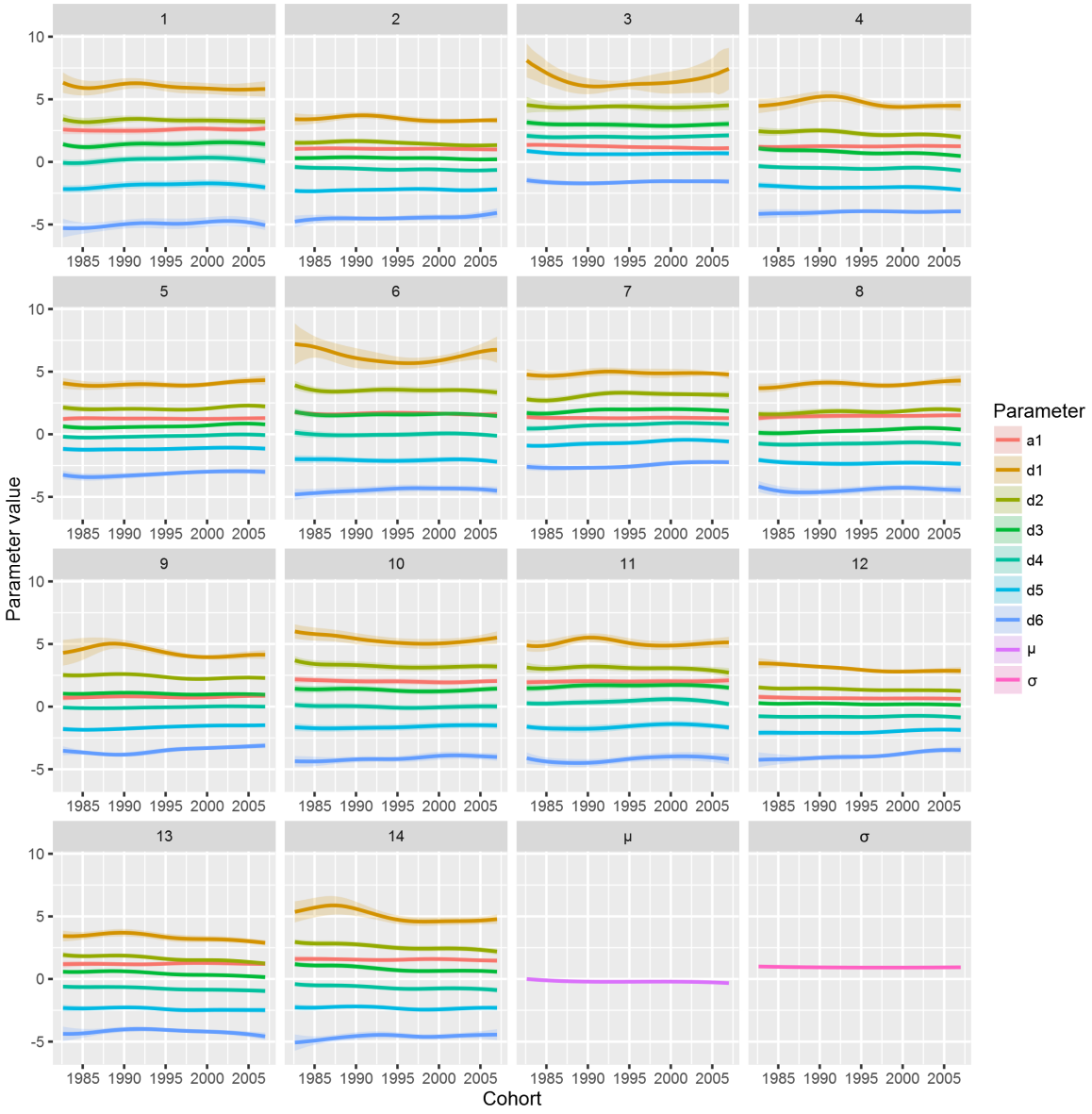

Figure 27: Estimates for all parameters. The RMSE of this fit is unknown, since we don't know the true values.

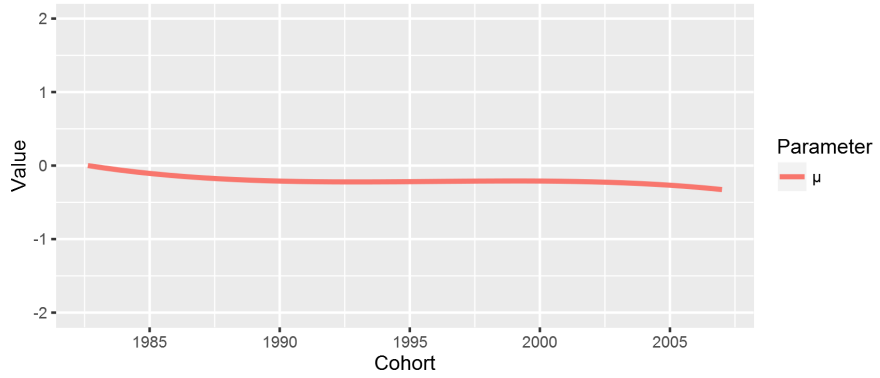

Figure 28: A zoomed-in view of μ^t.

In Figure 29 we zoomed in on question 3 Figure 27, to look at the values and confidence bands in detail.

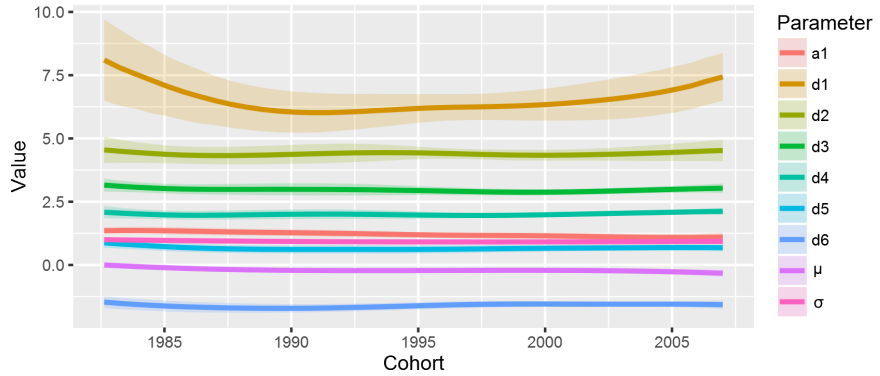

Figure 29: Parameter estimates for the third question, with bootstrapped confidence bands. The wideness of the bands around δ_1 indicates that the estimates for that parameter are inaccurate.

Based on the Figure 27 we suspected that the wild behavior of δ_1 was not significant; the larger version strengthens that suspicion, given how wide the confidence bands are. We can see that the parameter estimates other than δ_1 are fairly exact: the bands are very narrow. So, given how flat the estimates are, it seems that answers to this question given some trait value stayed constant over time, and that the variation of δ_1 is due to a small sample size, meaning that very few people chose the first option in question 3.

Question 11 was also interesting: there seems to be a parameter drift in δ_6; calculating the confidence bands might show whether this is real or coincidental. Figure 30 shows that the lowest parameter does indeed drift somewhat significantly, and so does the highest; answers on this question became less "extreme" over time, in both directions, after correcting for any variation in F.

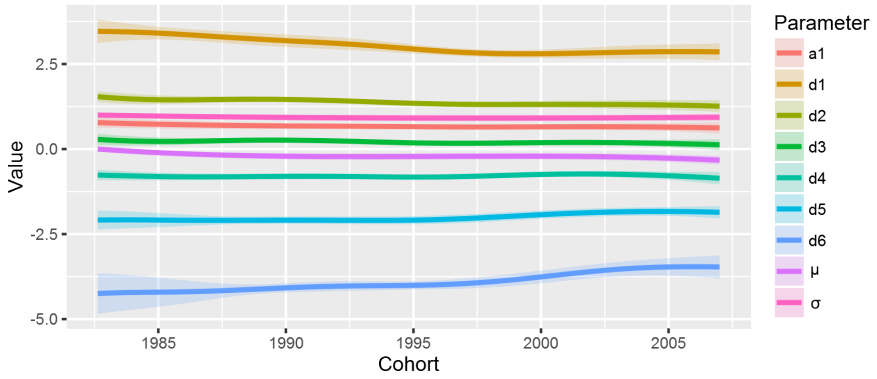

Figure 30: Parameter estimates for the 12th question, with bootstrapped confidence bands.

Looking at question 6 (Figure 31), we see a pattern similar to question 3, except that this question is more discriminatory, given that its parameters are wider apart. The bands around δ_1 are also narrower than in question 3.

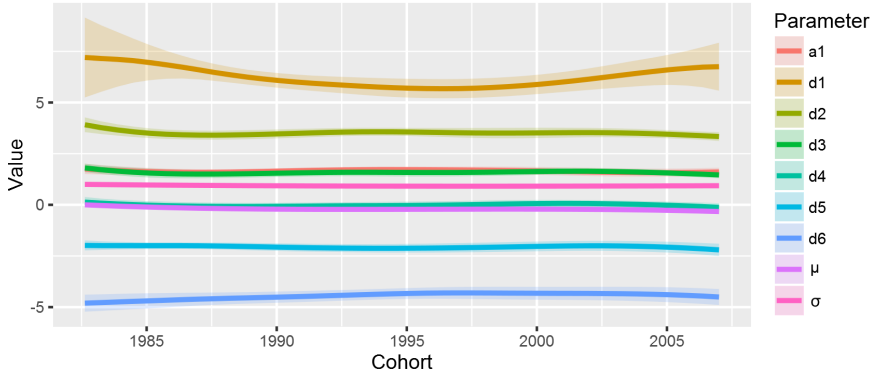

Figure 31: Parameter estimates for the 6th question, with bootstrapped confidence bands.

4.3.3 Exploratory conclusions

We did not find any strong trends in the data set on a macro level. The mean extroversion in the population stays fairly constant. Some questions changed over time: for example, question 11 became less discriminatory. The lack of large effects does not surprise us: we did not expect the mean extroversion in our population to change greatly in a short period of time, nor did we expect the answers to questions to change very much.

4.3.4 The effect of sample size on optimal bandwidth

In section 4.3.1 we have determined the optimal kernel smoothing bandwidth for a given real-world data set. Now, we're interested in determining the influence of sample size on the optimal bandwidth. We expect that, for lower sample sizes, larger bandwidths will be optimal, to smooth out the variance over time. For very large sample sizes, we would expect the optimal bandwidth to go towards 0, since we no longer need kernel smoothing to trade bias against variance, having very accurate local estimates.

In our experiment, we followed the procedure described in Section 2.2.2 on increasingly smaller subsets of the original data set. The results are shown in Figure 32 and in Table 2.

Now, if we wanted to be able to use a bandwidth of 2.5 on our data set (thus getting estimates that are smoothed less by the kernel smoothing), we could calculate a very rough estimate of the required sample size by extending Figure 32. We see a decrease of 0.5 of the optimal bandwidth between a sample size of 7163 and 8954. So, we estimate that approximately 11000 individuals would be required for the optimal bandwidth to be 2.5.

From the experiments in Chapter 3, one could get an idea of the bandwidth required to be precise enough to show the effect one has in mind.

Sample size	Optimal bandwidth
8954 (100%)	3.0
8506 (95%)	3.1
8058 (90%)	3.2
7163 (80%)	3.5
5969 (66%)	4.2
4477 (50%)	5.1

Table 2: Optimal bandwidth for various sample sizes, using randomly sampled parts of the real data.

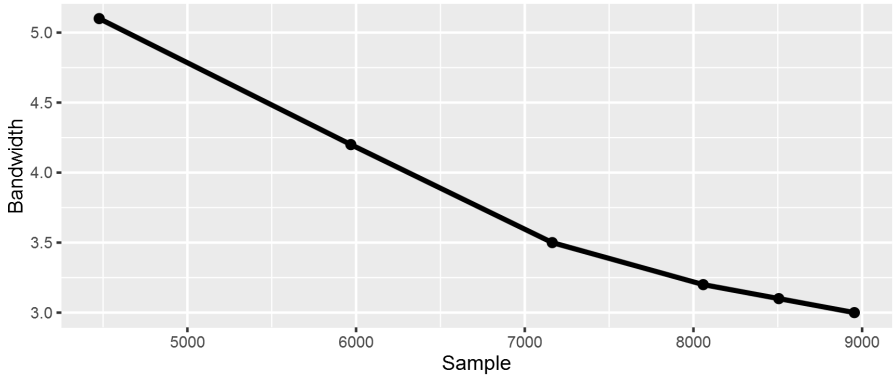

Figure 32: Optimal bandwidth for various sample sizes. For very large sample sizes, we expect the optimal bandwidth to go towards 0. We see that according to this approximation, our current sample size is not close to being large enough to use a multi-group model without kernel smoothing.

5 Discussion

We have demonstrated a method for the estimation of varying coefficient IRT models. We have shown its working on a both a simulated and a real-world data-set, and have seen that it works well for capturing and visualizing varying coefficients in IRT models on simulated data sets. The visual overview of all parameters make it easy to detect trends and come up with hypotheses about varying coefficients. We were able to estimate both model parameters and population parameters simultaneously, and we found sensible results using the restrictions we set on the population parameter.

Our method can calculate confidence intervals. While these are interesting, there are some reasons not to use them. For one, calculating the confidence intervals increases the computational time 50-fold. Also, the interpretation of the results can be difficult: while the confidence intervals do somewhat indicate the accuracy of the parameter estimates, the kernel smoothing makes it a lot less clear what exactly that accuracy means. For example, for a non-constant parameter, increasing the smoothing bandwidth will both increase estimation bias (the resulting parameter will be estimated to be more constant over time), but also reduce the confidence interval size, due to the local increase in sample size. The intervals might be a useful indicator of local accuracy: cohorts with less or less accurate data will have wider bands than larger and more accurate regions.

In real-world data, we found that most of the effects are subtle, after applying kernel smoothing with optimal parameters. This can mean two things: either the trends in the underlying data are very subtle, or our method is not sensitive enough. Larger sample sizes would allow us to use a smaller bandwidth, which would allow us to pick up smaller effects.

To conclude: if a longitudinal parameter is available in data in an IRT framework, we feel that it would indeed seem useful to look at a varying coefficient IRT model, to make sure you're not trying to fit a fixed (measurement-invariant) model to a population whose traits and/or interpretation of the items are changing.

Further research should be done by applying the method to different data sets; if possible, to data sets where shifts in some parameters are known. Varying coefficient

IRT models might also be used to quickly adapt tests and surveys in response to a changing population.

References

[Birnbaum, 1968] Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M. and Novick, M. R., editors, *Statistical theories of mental test scores*, pages 397–479. Addison-Wesley, Reading, MA.

[Bock, R. D. & Aitkin, M., 1981] Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the em algorithm. *Psychometrika*, 46:443–459.

[Broyden, 1970] Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90.

[Chalmers, 2012a] Chalmers, R. P. (2012a). Mirt: a multidimensional item response theory package for the r environment. *Journal of Statistical Software*, 48(6):1–29.

[Chalmers, 2012b] Chalmers, R. P. (2012b). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6):1–29.

[Fletcher, 1970] Fletcher, R. (1970). A new approach to variable metric algorithms. *The computer journal*, 13(3):317–322.

[Forero and Maydeu-Olivares, 2009] Forero, C. G. and Maydeu-Olivares, A. (2009). Estimation of irt graded response models: limited versus full information methods. *Psychological Methods*, 14(3):275.

[Golembiewski et al., 1976] Golembiewski, R. T., Billingsley, K., and Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by od designs. *The Journal of Applied Behavioral Science*, 12(2):133–157.

[Hambleton, 1991] Hambleton, R. K. (1991). *Fundamentals of item response theory*, volume 2. Sage publications.

[Hastie and Tibshirani, 1993] Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):pp. 757–796.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The elements of statistical learning*, volume 2. Springer.

[Lawley and Maxwell, 1962] Lawley, D. N. and Maxwell, A. E. (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 12(3):209–229.

[McDonald, 1999] McDonald, R. P. (1999). *Test theory: A unified treatment*. Psychology Press.

[Muraki, 1990] Muraki, E. (1990). Fitting a polytomous item response model to likert-type data. *Applied Psychological Measurement*, 14(1):59–71.

[R Core Team, 2014] R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Rasch, 1960] Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche, Copenhagen.

[Raudenbush et al., 2003] Raudenbush, S. W., Johnson, C., and Sampson, R. J. (2003). A multivariate, multilevel rasch model with application to self–reported criminal behavior. *Sociological methodology*, 33(1):169–211.

[Samejima, 1972] Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph Supplement*.

[Samejima, 1997] Samejima, F. (1997). Graded response model. In *Handbook of modern item response theory*, pages 85–100. Springer.

[Smits et al., 2013] Smits, I., Dolan, C., Vorst, H., Wicherts, J., and Timmerman, M. (2013). Data from 'cohort differences in big five personality factors over a period of 25 years'. *Journal of Open Psychology Data*, 1(1).

[Spearman, 1904] Spearman, C. (1904). "general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292.

[Thurstone, 1929] Thurstone, L. L. (1929). The measurement of psychological value. *Essays in Philosophy by Seventeen Doctors of Philosophy of the University of Chicago. Chicago: Open Court*, pages 157–174.

[Wand and Jones, 1994] Wand, M. P. and Jones, M. C. (1994). *Kernel smoothing*. Crc Press.

[Weiss and Yoes, 1991] Weiss, D. J. and Yoes, M. E. (1991). Item response theory. In *Advances in educational and psychological testing: Theory and applications*, pages 69–95. Springer.

[Zhang and Wang, 2014] Zhang, X. and Wang, J.-L. (2014). Varying-coefficient additive models for functional data. *Biometrika*, page asu053.

A Results for the other 4 factors

We have shown the results of the method as applied to Extroversion. Table 3 contains the optimal bandwidth for the other 4 factors, and Figures 33 - 36 show the corresponding resulting models.

Factor	Optimal bandwidth
Extroversion	3.0
Neuroticism	3.1
Agreeableness	3.4
Conscientiousness	2.6
Openness	4.5

Table 3: Optimal bandwidth for all 5 factors.

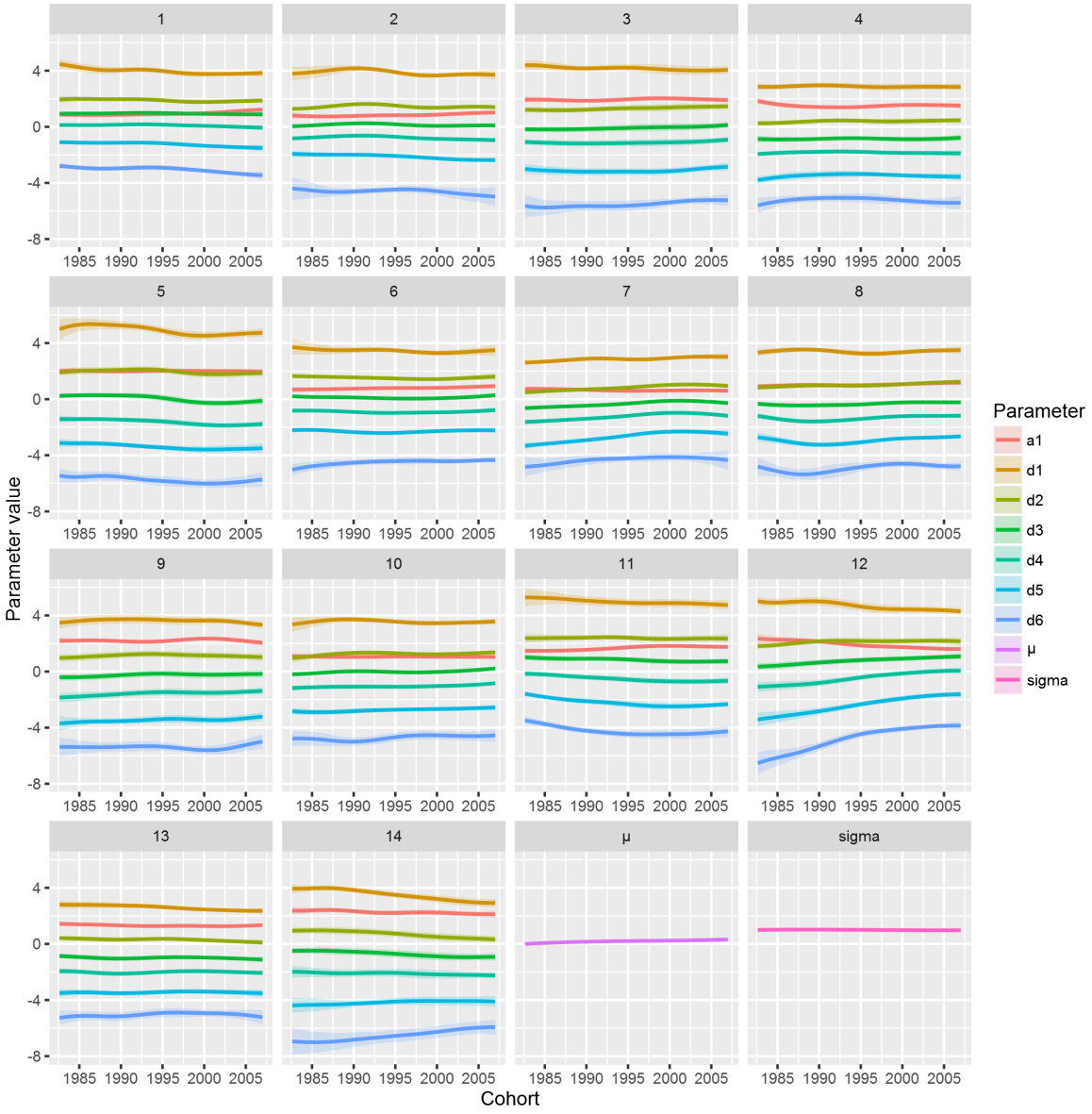

Figure 33: Estimates for all parameters for Neuroticism, using bandwidth 3.1.

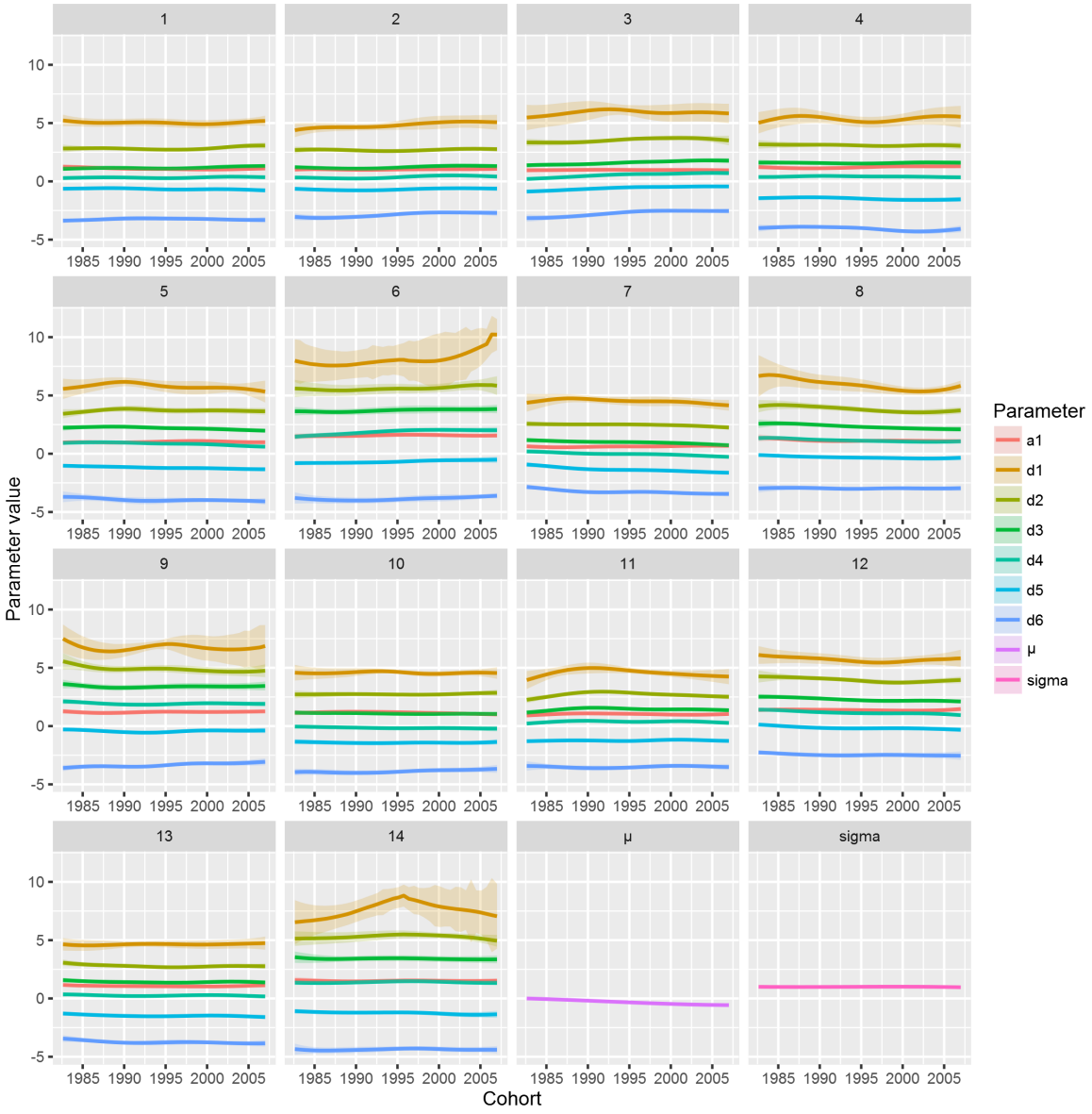

Figure 34: Estimates for all parameters for Agreeableness, using bandwidth 3.4.

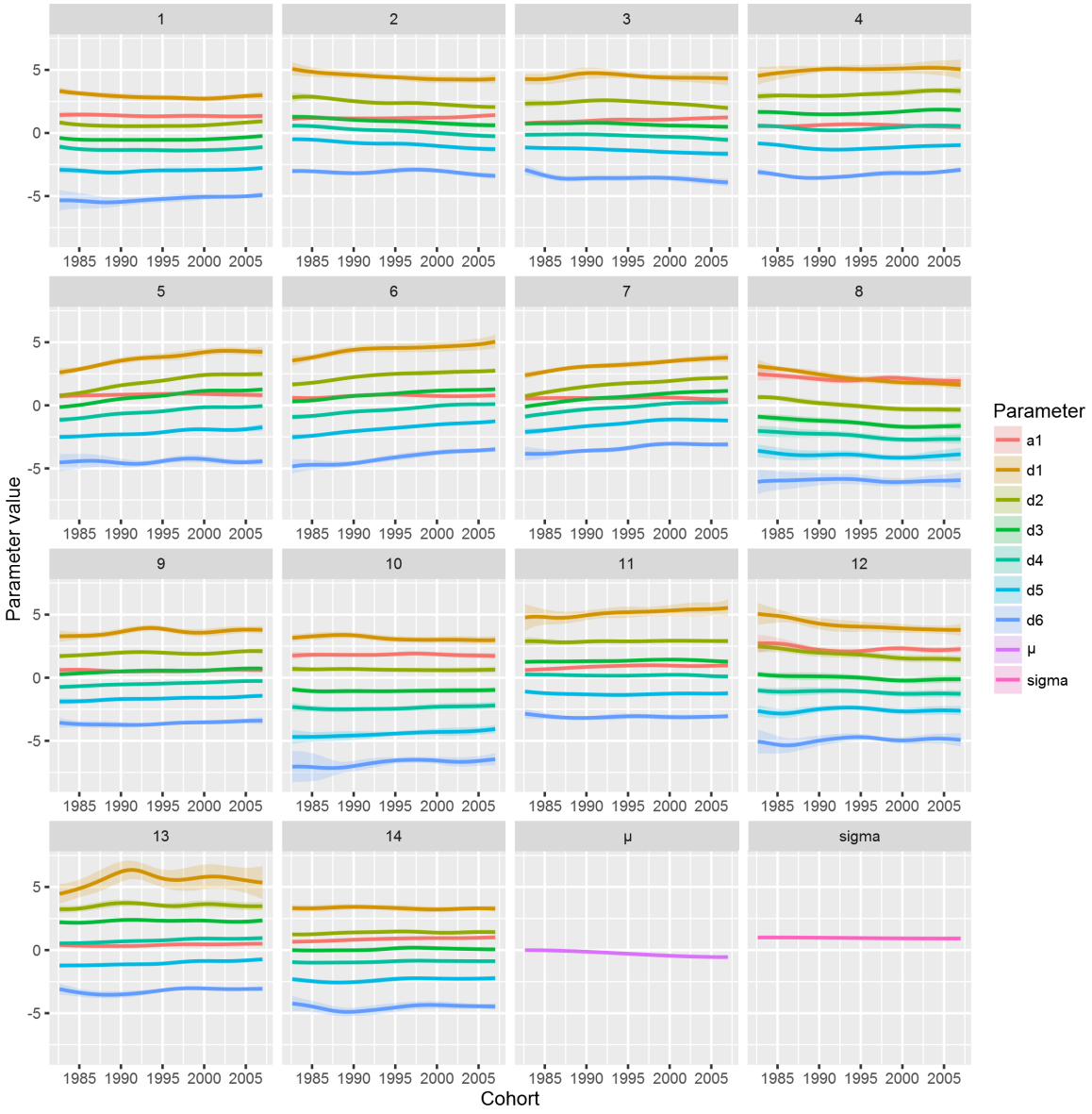

Figure 35: Estimates for all parameters for Conscientiousness, using bandwidth 2.6.

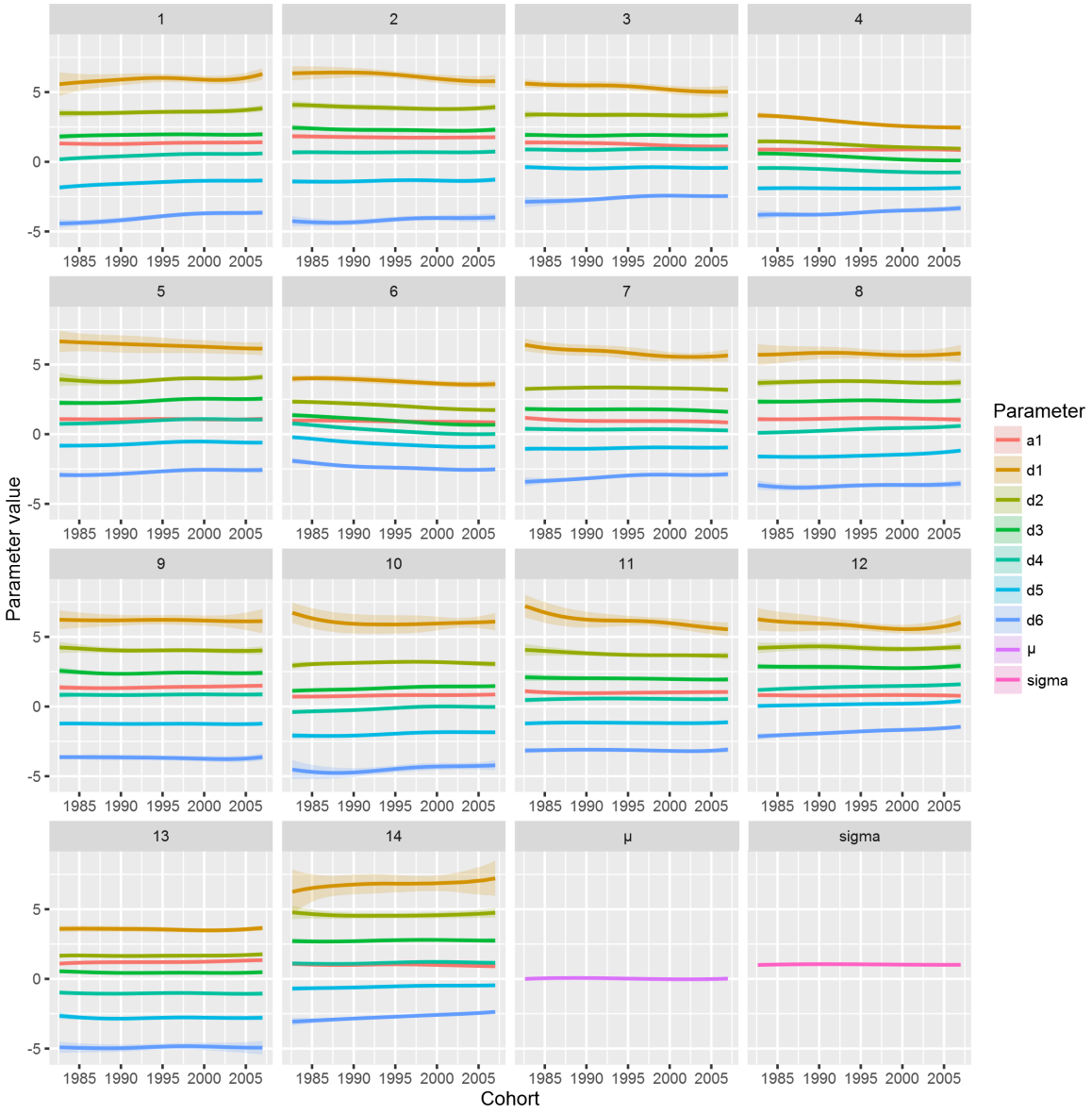

Figure 36: Estimates for all parameters for Openness, using bandwidth 4.5.

B Proof for 2.1.3

We can interpret a weighted model as equivalent to a larger unweighted model: for $w_n \in \mathbb{Q}$, the optimal parameters are equivalent to those in an unweighted model with repeated rows.

Let $w_n = \frac{a_n}{b_n}$ be fractions, so with $a_n, b_n \in \mathbb{N}_{\geq 0}$. Now, let $z = \text{LCM}\{b_n : 1 \leq n \leq N\}$. We can write $w_n = \frac{a_n \cdot \frac{z}{b_n}}{z}$. Because z is a multiple of b_n, we know that both numerator and denominator are integer. So, we can fill this in for w_n in Equation 10:

$$\begin{aligned}
\hat{\zeta}(\mathbf{X}) &= \arg\max_{\zeta} \sum_n w_n \log \int \prod_i p(x_{ni}|\zeta_i, f)\phi(f)df \\
&= \arg\max_{\zeta} \sum_n \frac{a_n \cdot \frac{z}{b_n}}{z} \log \int \prod_i p(x_{ni}|\zeta_i, f)\phi(f)df \\
&= \arg\max_{\zeta} \frac{1}{(n \cdot z)} \sum_n \left(a_n \cdot \frac{z}{b_n}\right) \log \int \prod_i p(x_{ni}|\zeta_i, f)\phi(f)df
\end{aligned}$$

Now, because z is a product of b_n it follows that $a_n \cdot \frac{z}{b_n}$ is also integer, so we see the equivalence to a larger matrix, where row n from matrix $\mathbf{X}$ is repeated $a_n \cdot \frac{z}{b_n}$ times.