Robustness, Semiparametric Estimation and Goodness-of-Fit of Latent Trait Models

Panagiota Tzamourani

London School of Economics and Political Science
University of London

Submitted in Fulfilment of the Requirement
for the Degree of Doctor of Philosophy

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr Martin Knott, for his guidance, help, encouragement and patience throughout my course of study. I would also like to thank Professor David J. Bartholomew who supervised the work described in the last chapter of my thesis.

The Statistics Department of the LSE has been a very stimulating and very friendly environment to work and study in and I would like to thank all its members for their friendliness and encouragement and particularly Yasmine Barlieb, Susannah Brown, Jane Galbraith and Pippa Smith.

I am also grateful to Irini Moustaki, for reading my thesis and most importantly for being a dear friend and always cheering me up.

I thank my parents for giving me so much and my sister, Marina, for her support and positive thinking.

Finally, I would like to acknowledge the Bank of Greece for giving me leave of absence to complete my studies.

Στους γονείς μου και στη Μαρίνα

Abstract

This thesis studies the one-factor latent trait model for binary data. In examines the sensitivity of the model when the assumptions about the model are violated, it investigates the information about the prior distribution when the model is estimated semiparametrically and it also examines the goodness-of-fit of the model using Monte-Carlo simulations.

Latent trait models are applied to data arising from psychometric tests, ability tests or attitude surveys. The data are often contaminated by guessing, cheating, unwillingness to give the true answer or gross errors. To study the sensitivity of the model when the data are contaminated we derive the Influence Function of the parameters and the posterior means, a tool developed in the frame of robust statistics theory. We study the behaviour of the Influence Function for changes in the data and also the behaviour of the parameters and the posterior means when the data are artificially contaminated.

We further derive the Influence Function of the parameters and the posterior means for changes in the prior distribution and study empirically the behaviour of the model when the prior is a mixture of distributions.

Semiparametric estimation involves estimation of the prior together with the item parameters. A new algorithm for fully semiparametric estimation of the model is given. The bootstrap is then used to study the information on the latent distribution than can be extracted from the data when the model is estimated semiparametrically.

The use of the usual goodness-of-fit statistics has been hampered for latent trait models because of the sparseness of the tables. We propose the use of Monte-Carlo simulations to derive the empirical distribution of the goodness-of-fit statistics and also the examination of the residuals as they may pinpoint to the sources of bad fit.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and Literature Review

1.1 Introduction

Latent variable models try to explain the association between observed variables (also called *manifest* variables) by the means of a set of unobserved variables - the *latent* variables or *factors.*

Both the observed and the latent variables may be categorical (measured on a nominal or ordinal scale) or metrical (measured on an interval or ratio scale).

When the latent variables are metrical and the manifest variables categorical the model is called a latent trait model. This thesis concentrates on latent trait models for binary manifest data and particularly on the logit/probit model with one latent variable.

The data, which could be responses to an ability test or to a questionnaire measuring attitudes, are often 'contaminated' by guessing, cheating, faking, or errors when recording the responses. In this thesis we will investigate the sensitivity of the model when such errors are present in the data. We will also investigate the sensitivity of the model if one of its main assumptions, the assumption about the form of the prior distribution, is violated. We will then propose a semi-parametric estimation method, by which the distribution of the latent variable is estimated together with the item parameters, and investigate the information that can be extracted about the form of this distribution. Finally we will examine new goodness-of-fit and diagnostic methods for the latent trait model.

In this Chapter we shall first introduce the model and then further illustrate the concept of robust statistics and give some robust statistics tools. We shall then review previous studies regarding robustness issues of the model and the sensitivity of the model to the form of the prior distribution, studies related to semiparametric estimation of latent trait models and finally studies on goodness-of-fit issues.

1.2 A latent variable model

Suppose

x is a vector of p manifest variables $x_1, ..., x_p$ and y is the latent variable.

The joint density of the x's is given by

$$f(\mathbf{x}) = \int_{R_y} h(y) g(\mathbf{x}|y) dy \tag{1.1}$$

where R_y is the range space of y. Our main interest is in what can be known about y after x has been observed. This is given by the conditional density,

$$h(y|\mathbf{x}) = h(y) g(\mathbf{x}|y) / f(\mathbf{x}) \tag{1.2}$$

Since h and g are not uniquely determined by (1.1) we do not yet have a complete specification of $h(y|\mathbf{x})$.

The classes of distributions to be considered can be narrowed down by assumptions and restrictions arising from statistical properties. Furthermore, the choice for g depends on the type of the manifest variables whereas to some extent the choice for h (discrete or continuous) will lead to different types of latent variable models.

An essential assumption for the general latent variable model is that of conditional independence, formulated as,

$$g(\mathbf{x} \mid y) = \prod_{i=1}^{p} g_i(x_i \mid y) \tag{1.3}$$

It says that if the dependencies among the x's are induced by the latent variable y, then when y is accounted for and held fixed the x's will be independent. Then $f(\mathbf{x})$ can be written as

$$f(\mathrm{x}) = \int_{R_y} [\prod_{i=1}^{p} g_i(x_i \mid y)]h(y)dy \tag{1.4}$$

for some h and g_i, where R_y is the range space of y.

1.3 Latent Trait Models for Binary Data

Binary data arise when questions are scored dichotomously, for example, in social surveys when the respondents are asked to agree or to disagree with an item of a questionnaire, or, in educational testing, when students answer an item right or wrong. Responses are usually coded with 1 for a 'positive' response and with 0 for a 'negative' response. If n respondents have answered a set of p binary items, the data will be in a form of an $n \times p$ matrix. Any row of that matrix will be the set of p responses from a given individual and will be called the *response* or *score pattern* for that individual. With p variables, each having two outcomes, there are 2^p different response patterns. It is common to summarise the data in a table showing the frequencies of each response pattern. If 2^p is large compared to n most of the frequencies of the response patterns may be 0 or 1. Then only the response patterns actually observed are listed.

Two approaches to the construction of models for binary data have been used in the past. The *response function approach* relates the probability of a positive response of an individual to the value for that individual of the latent variable y. The response function approach forms the basis of Item Response Theory (IRT), which provides methodology for the design, construction and evaluation of educational and psychological tests.

The *underlying variable approach* was developed in the line of traditional factor analysis and supposes that the binary x's have been produced by dichotomising underlying continuous variables. In Bartholomew and Knott (1999) it is shown that these two approaches are equivalent for binary data.

1.3.1 The response function approach

The response function $\pi_i(y)$ gives the probability of a positive response to item i of an individual given his/her value of the latent variable y. In educational testing $\pi_i(y)$ is referred to as the *item characteristic curve* (ICC) or *item response function* (IRF). The shape of the curve shows how the probability of a correct response changes with ability.

Several models for $\pi_i(y)$ have been proposed, the most frequently used of which are the normal and logistic (cumulative distribution) ogive models.

The normal ogive model Lawley (1943) and Lord and Novick (1968) proposed the normal ogive model:

$$\pi_i(y) = p(y, a_i, b_i) = \Phi(g_i(y)) = \int_{-g_i}^{\infty} \phi(t)dt = \int_{-\infty}^{g_i} \phi(t)dt \tag{1.5}$$

where

$$g_i(y) = a_i(y - b_i) \tag{1.6}$$

is a linear function of y involving two item parameters a_i and b_i, and $\phi(t)$ is the normal density function. b_i is called the difficulty parameter for item i and a_i the discrimination parameter.

The two-parameter logistic ogive model Birnbaum (1968) proposed the logistic ogive with two and three item parameters for the form of the response function.

For the two parameter model the probability of a positive response (using his notation for the parameters) is given by:

$$\pi_i(y) = \frac{\exp(D(a_i(y - b_i)))}{1 + \exp(D(a_i(y - b_i)))} \tag{1.7}$$

b_i is the item difficulty parameter, a_i is the item discrimination parameter as above and D is a unit scaling factor. Birnbaum (1968) took $D = 1.7$ to maximise quantitative agreement between the parameters of the normal and logistic models.

The three-parameter logistic ogive model The three-parameter model includes a constant so that the response function does not reach 0. The response function is given by:

$$\pi_i(y) = c_i + (1 - c_i)\frac{\exp(D(a_i(y - b_i)))}{1 + \exp(D(a_i(y - b_i)))} \tag{1.8}$$

As c_i allows subjects on the low end of the ability scale to answer correctly, it is called the guessing parameter.

The Rasch model Rasch (1960) simplified the two-parameter logistic model so that all items have the same discriminating power. Thus the response function becomes:

$$\pi_i(y_l) = \frac{\exp(y_l - b_i)}{1 + \exp(y_l - b_i)} \tag{1.9}$$

where $i = 1, ..., p$ indexes the items and $l = 1, ..., n$ indexes the individuals. y was considered by Rasch a subject specific parameter and not a latent variable and it was usually denoted by θ. b_i is the item difficulty parameter.

A general model Bartholomew (1980) gave a general approach to the choice of a suitable response function. That paper listed a set of properties which the family of response functions should possess and then proposed a class of linear models meeting these requirements of the form

$$G^{-1}\pi_i(y) = a_{0i} + a_{1i}H^{-1}(y), \; (i = 1, ..., p) \tag{1.10}$$

where $\pi_i(y) = P(x_i = 1|y)$ and y is uniformly distributed on (0,1). The functions G^{-1} and H^{-1} are arbitrary but such that their inverses G and H are distribution functions of random variables symmetrically distributed about zero.

Bartholomew and Knott (1999) show that if G^{-1} is chosen to be the cumulative logistic distribution function, then the 'sufficiency principle' can be fulfilled and thus the information about the unidimensional y, which is contained in the x's, can be conveyed by a unidimensional vector of observable quantities.

The distribution of the latent variable is essentially arbitrary. It was argued in Bartholomew (1980) that H should be the distribution function of a symmetrically distributed random variable, on the basis that the direction of measurement of the latent variable was arbitrary, but this still leaves a wide choice. However, the two functions which have been used most often in practice are the logistic and the normal. These are very similar in shape, and the choice between them is of little practical importance. Bartholomew and Knott (1999) argue that the normal function should be chosen when there is more than one latent variable because then the latent variables will remain independent under orthogonal rotation.

We shall later discuss in more detail the form of the distribution of the prior.

If we choose the logistic distribution function for the response function and the normal distribution for y, (1.10) can be written in terms of the uniformly distributed latent variables y as

$$\text{logit}\pi_i(y) = a_{0i} + a_{1i}\Phi^{-1}(y) \tag{1.11}$$

or, in terms of the normally distributed variable z, as

$$\text{logit}\pi_i(z) = a_{0i} + a_{1i}(z). \tag{1.12}$$

We shall call this model the *logit/probit* model or, more briefly, the *logit* model.

This model is equivalent to Birnbaum's two-parameter logistic model for

$$a_{0i} = -Db_i a_i$$

and

$$a_{1i} = Da_i$$

and equivalent to the Rasch model for

$$a_{0i} = -b_i$$

and

$$a_{1i} = 1.$$

Interpretation of the parameters The parameters a_{0i} and a_{1i} are the difficulty and discrimination parameters for item i. These parameters determine the position and shape of the item's response function.

The parameter a_{0i} defines the probability of a positive response to item i by an individual of median ability. Assuming the median of z is zero, that will be

$$\pi_i = P(x_i = 1|z = 0) = \frac{\exp(a_{0i})}{1 + \exp(a_{0i})}. \tag{1.13}$$

If, for example, a_{0i} is large and positive (negative), the probability of a positive (negative) response for a median individual will be close to 1.

The parameter a_{1i} is a discrimination parameter because, for two individuals a given distance apart on the z scale, the greater a_{1i} the greater the difference in their probabilities of responding positively to item i.

As a coefficient of z, a_{1i} can also be thought of as a factor loading. It thus indicates the weight to be attached to the various x's in the interpretation of the factor z, or the strength of the link of the x's with z.

Properties of the response function The choice of which of the two possible outcomes is to be regarded as 'positive' is entirely arbitrary. It ought not to matter whether the code 1 is attached to 'Yes' or 'No'. Thus $\pi_i(z)$ and $1 - \pi_i(z)$ should have the same form. For the one-factor logit model,

$$\pi_i(z) = \frac{\exp(a_{0i} + a_{1i}z)}{1 + \exp(a_{0i} + a_{1i}z)} \tag{1.14}$$

or

$$\pi_i(z) = \frac{1}{1 + \exp(-a_{0i} - a_{1i}z)} \tag{1.15}$$

and

$$1 - \pi_i(z) = \frac{1}{1 + \exp(a_{0i} + a_{1i}z)} \tag{1.16}$$

Both have the same form, the only difference is in the signs of the a's. This is as it should be since if increasing z increases the probability of answering 'Yes' it should decrease the probability of answering 'No' by the same amount.

The second property is the invariance of the form of the response function to the direction of measurement of the latent variables, which is again arbitrary. The choice of which end of the scale is 'positive' and which is 'negative' is of no significance. Changing the direction of measurement involves replacing z by $-z$. This is equivalent to changing the sign of a_{1i}, but it does not change the form of the model.

We assume for the one factor model that the response function is nondecreasing in the latent variable, i.e. a_{1i}, $i = 1, 2, ..., p$ is always non-negative. This ensures that the probability of response increases, or decreases, in step with changes in the latent variable. We can make all a_{1i}'s positive or zero by an appropriate choice of which outcome is to be regarded as positive.

1.3.2 The underlying variable approach

This approach assumes that underlying the observed binary variables there are continuous variables, which we shall denote by ξ. An observed variable takes the value of 1 when the underlying continuous variable is above a threshold and the value of 0 otherwise. Thus we may define

$$x_i = 1 \quad \text{if} \quad \xi_i > k_i \quad \text{and} \quad x_i = 0 \quad \text{if} \quad \xi_i \leq k_i \tag{1.17}$$

This approach was developed so that traditional factor analysis techniques for continuous data could be applied to dichotomised data. The model, with the factors coming at a second underlying level, can be defined as follows:

$$\xi = \mu + \mathbf{\Lambda z} + \mathbf{e} \tag{1.18}$$

where μ is the overall mean, $\mathbf{\Lambda}$ is the factor loadings vector, z the underlying latent variable and $\mathbf{e}$ the error term. If $\mathbf{z}$ and $\mathbf{e}$ are both assumed to be normal then ξ will be normal and their correlation coefficients can be estimated from the bivariate 2 × 2 tables of the $\mathbf{x}$'s. These are the so-called tetrachoric correlations which can be used as input to a standard factor analysis program. Unfortunately, correlation matrices estimated by these means are not always positive definite.

1.3.3 Related Models

Latent Class Models When both the manifest and latent variables are categorical, a latent class models is defined. The latent class model was originally proposed by Lazarsfeld (1950), while Goodman (1974) and Haberman (1979) extended its practical applicability.

The latent class model can be parameterised in two different ways. The classical formulation, given in Bartholomew and Knott (1999) is as follows:

Suppose the population can be divided into K distinct categories, or latent classes. Let π_{ij} be the probability of a positive response on variable i for a person in class j and let η_j be the prior probability that a randomly chosen individual is in class j $(\sum_{j=1}^{K} \eta_j = 1)$.

The conditional distribution of x_i given the individual belongs to latent class j is given by

$$g_i(x_i|j) = \pi_{ij}^{x_i}(1-\pi_{ij})^{(1-x_i)} \tag{1.19}$$

For the case of K classes the probability of a response pattern $\mathbf{x}$ is given by

$$f(\mathrm{x}) = \sum_{j=1}^{K} \eta_j \prod_{i=1}^{p} \pi_{ij}^{x_i}(1-\pi_{ij})^{(1-x_i)} \tag{1.20}$$

Latent Class and Latent Trait Models as Loglinear Models The latent class model can also be formulated as a loglinear model (Haberman 1979). For example, the log of the expected frequency of the cell for latent class j and categories 1 for items 1, 2 and 3 is given by:

$$\log m_{j111}^{z123} = u + u_j^z + u_1^1 + u_1^2 + u_1^3 + u_{j1}^{z1} + u_{j1}^{z2} + u_{j1}^{z3} \tag{1.21}$$

in which $m_{j111} = n\, g(111|j)$, n is the sample size and $g(111|j) = \pi_{1j}\pi_{2j}\pi_{3j}$. Subscripts in (1.21) denote categories and superscripts denote variables.

The equivalence between these two formulations is shown if the conditional response probabilities of the items given the latent class are formulated in terms of the log-linear parameters:

$$\pi_{ij} = \frac{\exp(u_1^i + u_{j1}^{zi})}{\exp(u_0^i + u_{j0}^{zi}) + \exp(u_1^i + u_{j1}^{zi})} \tag{1.22}$$

A latent trait model can be specified as a loglinear row effects model, where the item categories are the rows and the latent variable is the interval level column variable. The row effects model has parameter row scores whereas the column scores are fixed. Let z_j denote one of the distinct points/ nodes of z and μ_1 denote the row score for row (response) 1 of item i. The two-way interaction between z and an item i can be formulated with $u_{j1}^{zi} = \mu_1\, z_j$, which means that the two-way interaction is proportional to the point of variable z by the row effect.

The conditional probability of positive (1) response to item i for a person in class j is given by

$$\pi_{ij} = \frac{\exp(u_1^i + \mu_1\, z_j)}{\exp(u_0^i + \mu_0\, z_j) + \exp(u_1^i + \mu_1\, z_j)} \tag{1.23}$$

1.4 Estimation Methods for Latent Trait Models

1.4.1 Maximum Likelihood

Conditional maximum likelihood Conditional maximum likelihood is the method that maximises the likelihood for the structural parameters of the model which is formulated as a conditional probability on sufficient statistics for the nuisance parameters of the model. In the Rasch model the latent variable y is considered a parameter - one for each subject - and is a nuisance parameter since the number of these parameters increases with the sample size. Andersen (1973) showed that for the Rasch model sufficient statistics for y, independent of the item parameters, exist and so by conditioning on the sufficient statistics, estimates for the structural parameters of the model - the item parameters - can be obtained.

The sufficient statistic for the Rasch model is the total score

$$t_l = \sum_{i=1}^{p} x_{li} \quad l = 1, ..., n, \quad i = 1, ..., p \tag{1.24}$$

The inference can be based on the total score $t_1, ..., t_n$. By virtue of sufficiency of t_l the conditional distribution of $x_{l1}, ..., x_{lp}$ given t_l is independent of the latent variable y.

Hence the conditional likelihood is defined as

$$L(a_1, ..., a_p) = \prod_{l=1}^{n} f(x_{l1}, ...x_{lp} | t_l) \tag{1.25}$$

and it involves only the item parameters. The estimates for $a_1, ..., a_p$ are obtained as those that maximise $L(a_1, ..., a_p)$.

Marginal maximum likelihood Bock and Lieberman (1970) used the marginal maximum likelihood method to estimate the item parameters of the two-parameter latent trait model. The characteristic of the marginal maximum likelihood is that the probability for a response pattern is integrated over the range of the latent variable z. Bock and Lieberman (1970) considered the normal ogive model for the response function and assumed the latent variable to follow the standard normal distribution.

Assuming one latent variable, the likelihood to be maximised is the joint probability function for a response pattern of a randomly selected subject which is given by

$$f(\mathbf{x}) = \int_{-\infty}^{\infty} \prod_{i=1}^{p} g_i(x_i|z)h(z)dz \tag{1.26}$$

Then $L = \log f(\mathbf{x})$ is differentiated with respect to the item parameters. Bock and Lieberman (1970) solved the likelihood equations by the Newton-Raphson method, approximating the integral over z by a weighted summation over a set of finite number of points and weights. For the normal case they used Gauss-Hermite quadrature nodes. The points (quadrature nodes) and weights are given in Straud and Sechrest (1966).

They noted that due to heavy computational requirements this method could only be used for a small set of items.

Bock and Aitkin (1981) reformulated the Bock-Lieberman likelihood equations and gave a computationally feasible solution also for large number of items. Their method was closely related to the E-M algorithm as discussed by Dempster, Laird, and Rubin (1977). They used the normal ogive for the form of the response function and the $N(0,1)$ and other continuous distributions for the latent variable which they approximated with a set of finite number of points and weights.

Bartholomew and Knott (1999) gave a variation of the E-M algorithm for the logit response function, which can easily be adapted to fit any response function. We shall describe this variation of the E-M algorithm in more detail.

The E-M algorithm Suppose that $h(z)$ is approximated by a set of finite points, and their corresponding weights, so that z takes the values $z_1, z_2, \cdots, z_k$ with probabilities $h(z_1), h(z_2), \cdots, h(z_k)$. The marginal distribution (1.26) for individual l is written:

$$f(\mathbf{x}_l) = \sum_{t=1}^{k} g(\mathbf{x}_l \mid z_t)h(z_t)$$

where

$$g(\mathbf{x}_l \mid z_t) = \prod_{i=1}^{p} \pi_i(z_t)^{x_{li}}(1 - \pi_i(z_t))^{1-x_{li}}.$$

We then have to maximise:

$$L = \sum_{l=1}^{n} \log f(\mathbf{x}_l)$$

By differentiating the log-likelihood with respect to unknown parameters we get:

$$\frac{\partial L}{\partial \alpha_{ji}} = \sum_{t=1}^{k} \frac{\partial \pi_i(z_t)}{\partial \alpha_{ji}} \frac{\{r_{it} - N_t \pi_i(z_t)\}}{\pi_i(z_t)\{1 - \pi_i(z_t)\}} \qquad j = 0, 1. \tag{1.27}$$

Where,

$$\begin{aligned} r_{it} &= h(z_t) \sum_{l=1}^{n} x_{li} g(\mathbf{x}_l \mid z_t)/f(\mathbf{x}_l) \\ &= \sum_{l=1}^{n} x_{li} h(z_t \mid \mathbf{x}_l) \end{aligned} \tag{1.28}$$

and

$$\begin{aligned} N_t &= h(z_t) \sum_{l=1}^{n} g(\mathbf{x}_l \mid z_t)/f(\mathbf{x}_l) \\ &= \sum_{l=1}^{n} h(z_t \mid \mathbf{x}_l) \end{aligned} \tag{1.29}$$

The probability function $h(z_t \mid \mathbf{x}_l)$ is the probability that an individual l with response vector $\mathbf{x}_l$ is located at z_t.

The N_t could be interpreted as the expected number of individuals at z_t and r_{it} is the expected number of those predicted to be at z_t who will respond positively. The N_t and r_{it} are functions of the unknown parameters.

If the response function is taken to be the logit, as in (1.14), (1.27) becomes:

$$\frac{\partial L}{\partial \alpha_{ji}} = \sum_{t=1}^{k} z^j \{r_{it} - N_t \pi_i(z_t)\}, \qquad j = 0, 1. \tag{1.30}$$

We define the steps of an E-M algorithm as follows:

- *step 1* Choose starting values for α_{0i} and α_{1i}

- *step 2* Compute the values of r_{it} and N_t from (1.28) and (1.29)

- *step 3* Obtain improved estimates of the parameters by solving (1.30) for $j = 0, 1$ and $i = 1, 2, ..., p$ treating r_{it} and N_{it} as given numbers.

- *step 4* Return to step 2 and continue until convergence is attained.

There is a program called TWOMISS (Albanese and Knott 1992a) which gives maximum likelihood estimates via this modified E-M algorithm for the one- and two-factor latent trait model, using the logit model for the response function.

Bock and Aitkin (1981) suggested that a number of quadrature nodes between 3 and 7 will be satisfactory for estimating a model with more than one latent variable. However, Shea (1984) showed that many more nodes are needed in order to get a reasonable accuracy for the parameter estimates. TWOMISS allows the user to choose either 8, 16, 32 or 48 quadrature nodes and weights to fit the model.

1.4.2 Generalised Least Squares (GLS)

Generalised least squares methods assume that most of the relevant information in the sample data is contained in the first- and second-order margins. Christoffersson (1975), using the underlying variable approach described in Section 1.3.2 proposed a Generalised Least Squares estimator based on the marginal distributions of a single item and of pairs of items, i.e. on $\pi_i = P(x_i = 1)$ and $\pi_{ij} = P(x_i = 1, x_j = 1)$. For the sample marginal proportions corresponding to π_i and π_{ij} the model is

$$\begin{aligned} p_i &= \pi_i + \epsilon_i, \quad i = 1, ..., M \\ p_{ij} &= \pi_{ij} + \epsilon_{ij}, \quad i = 1, ..., M-1, \quad j = i+1, ..., M \end{aligned} \tag{1.31}$$

The estimator is defined by minimising

$$f(\mathrm{k}, \Lambda, \Phi) = \epsilon \mathrm{S}^{-1} \epsilon = (\Pi - \mathrm{P})\mathrm{S}^{-1}(\Pi - \mathrm{P}) \tag{1.32}$$

regarded as a function of the unknown parameters **k**, **Λ** and **Φ**. **k** is the vector of thresholds, **Λ** is the matrix of factor loadings, **Φ** is the covariance matrix of the factors and $\mathbf{S}^{-1}$ is the estimated dispersion matrix of **Π**. The differences in the parameter estimates and the standard errors of this estimator and the marginal ML estimator of Bock and Lieberman (1970) are negligible. As GLS requires heavy computations, Christoffersson (1975) proposed a simpler two-step estimator. In the first step the threshold levels are estimated and in the second step these are used to estimate Λ and Φ.

Muthén (1978) made further improvements on this method by substantially reducing the amount of numerical integration required.

1.5 Measurement of the Latent Variable

The simplest way to score an individual is using the *total score* (number of 1's):

$$t_l = \sum_{i=1}^{p} x_{li}. \tag{1.33}$$

In using the total score we would be treating all items as equally important.

Another way to score the response patterns is with the estimated *component score* (Bartholomew 1984), which is the sum of the parameter estimates $\hat{a}_{1i}$ of the model for the items with response 1:

$$CS_l = \sum_{i=1}^{p} \hat{a}_{1i} x_{li}$$

The response pattern 00110 would be scored $\hat{a}_{13} + \hat{a}_{14}$. This method of scoring is said by Bartholomew to be more informative and to give a better scoring than just using the total number of 1's.

We also may use the estimated *conditional mean* $E(z|\mathbf{x})$ of the latent variable given the response pattern **x**, that is

$$E(z|\mathrm{x}_l) = \sum_{t=1}^{k} z_t h(z_t|\mathrm{x}_l) \tag{1.34}$$

or

$$E(z|\mathrm{x}_l) = \sum_{t=1}^{k} z_t g(\mathrm{x}_l|z_t) h(z_t) / f(\mathrm{x}_l) \tag{1.35}$$

This is also called the Bayesian Expected A Posteriori (EAP) estimate of the distribution of z (Bock and Aitkin 1981).

The conditional standard deviation given the response pattern can be used as a measure of the information delivered by the latent variable.

Knott and Albanese (1993) and Albanese (1990) investigated the relation between component scores and conditional means and found that the latter maintains the advantages of the component score, but it is more stable. The component score is strongly dependent on the values of a_{1i} while the conditional mean depends on π_i, which does not vary much even if some $\hat{a}_{1i}$ are very large. This implies that significant differences in the posterior means do not always reflect different positions on the latent scale, according to the conditional mean $E(z|x)$, which is not desirable in most of the applications. They conclude that if the responses are complete (none missing) then both the component score and the posterior means will provide the same ranking of individuals on the latent scale, but if one wants to use the latent score in further analyses then the conditional mean is more informative and reliable than the component score, specially when one or more $\hat{a}_{1i}$ are large (bigger than 3.0).

For this reason we will be using the posterior means as the scoring method of individuals on the latent scale for the purposes of this thesis.

1.6 Robustness

Robust statistics studies the behaviour of statistical procedures, not only under strict parametric models, but also in smaller and in larger neighbourhoods of such parametric models. It studies deviations from the models since the assumptions needed for a parametric model to be valid are usually only approximately true.

The first theoretical approach to robust statistics was introduced by Huber (1964) and Huber (1981). Huber identified neighbourhoods of a stochastic model which are supposed to contain the 'true' distribution that generates the data. Then he found the minimax estimator that behaves optimally over the whole neighbourhood.

Another approach to robust statistics is through the Influence Function, originated by Hampel (1968). With this approach the effects of a very small (infinitesimal) violation of the model assumptions are studied and estimators are developed such that small deviations from the parametric model have small effects.

In robustness theory the terms 'model deviations' and 'violation of the model assumptions' are used interchangeably with the term 'data contamination', in the sense that, if there is a 'true' parametric model, data contamination arises when part of the data is not generated from this model.

Generally, deviations from strict parametric models may be due to

i) the occurrence of gross errors, ii) rounding and grouping, iii) violations of the distributional assumptions of the model.

i) the occurrence of gross errors. Gross errors are occasional values where something went wrong, like mistakes in copying or computation. Gross errors result in 'outliers', namely values which deviate from the pattern set by the majority of the data. Some gross errors may be harmless. On the other hand, outliers can also seriously change the fitted model.

In our case, where we have binary data, errors can only occur as an interchange of several 0's and 1's. And since we have multivariate data where an individual's response comprises a response pattern, such a transposition will cause the frequency of the correct response pattern to decrease by one and the frequency of another response pattern to increase by one. Of course, a gross error in recording the frequency is also possible, but is equivalent.

(ii) rounding and grouping. Mainly applicable to metric data, as continuous data are often rounded, grouped, or classified coarsely. This could lead to the result, for example, that a continuous distribution would not be a valid approximation to the data any more.

Unless the response is genuinely dichotomous, binary data are a result of a 'coarse' classification. There have been studies which are concerned with the effects of such a classification but for the purposes of the present study we will assume that responses are genuinely dichotomous or that such effects are not significant.

(iii) violations of the distributional assumptions of the model. In the context of metric data this would refer primarily to deviations from normality, or violation of the independence assumption, due to unsuspected serial correlations.

The assumptions needed for the latent variable models we use are given in Sections 1.2 and 1.3.1. We shall apply robust statistics methods to see whether the assumption about the distribution of the latent variable or data contamination influences the parameter estimates and the measurement of the latent variable.

The assumption of conditional independence is equivalent to the assumption that the number of latent variables are sufficient to explain the association between the observed variables. Whether this is true or whether more latent variables are needed in the model can be tested by goodness-of-fit measures and model choice criteria.

This thesis will explore the influence function approach and therefore we will now give its definition and some related concepts.

1.6.1 Basic concepts and definitions

We shall first give the basic concepts on which robust statistics measures are based as they appear in Hampel, Ronchetti, Rousseeuw, and Stahel (1986) and later we shall attempt to apply them to our model.

Suppose we have one-dimensional observations $X_1, ..., X_n$ which are independent and identically distributed (i.i.d.). The observations belong to some sample space $\mathcal{X}$, which is a subset of the real line $\Re$ (or may equal $\Re$). A parametric model consists of a family of probability distributions F_θ on the sample space, where the unknown parameter θ belongs to some parameter space Θ. In classical statistics, one then assumes that the observations X_i are distributed *exactly* like one of the F_θ, and undertakes to estimate θ based on the data at hand. In robustness theory the model is considered a mathematical abstraction which is only an *idealised approximation* of reality, and statistical procedures are constructed with the aim to behave fairly well under deviations from the assumed model.

The empirical distribution G_n of the sample $(X_1, ..., X_n)$ is given by $(1/n)\sum_{i=1}^{n}\Delta_{x_i}$, where Δ_x is the point mass 1 in x.

As estimators of θ we consider real-valued statistics $T_n = T_n(X_1, ..., X_n)$.

We consider estimators which are functionals [i.e., $T_n(G_n) = T(G_n)$ for all n and G_n] or can asymptotically be replaced by functionals. This means that we assume that there exists a functional T:domain $(T) \rightarrow \Re$ [where the domain of T is the set of all distributions of $F(\mathcal{X})$ for which T is defined] such that $T_n(X_1, ..., X_n) \overset{n\rightarrow\infty}{\rightarrow} T(G)$ in probability when the observations are i.i.d. according to the true distribution G in domain (T). We say that $T(G)$ is the asymptotic value of $\{T_n; n \geq 1\}$ at G.

Robustness theory examines the sensitivity of the estimators or the estimates in a neighbourhood of the true model and provides estimators which are stable in a specific neighbourhood. Consider the set of distributions

$$\{F_\epsilon | F_\epsilon = (1-\epsilon)F_\theta + \epsilon W\} \tag{1.36}$$

where W is an arbitrary distribution function. The data generated from F_ϵ are actually generated from the true parametric model F_θ with probability $(1-\epsilon)$ and from W with small probability ϵ. A particular case arises when $W = \Delta_x$, i.e. the distribution gives probability 1 to an arbitrary point x. In that case, the neighbourhood of the model, which will also be referred to as the contaminated model, is given by

$$\{F_\epsilon | F_\epsilon = (1-\epsilon)F_\theta + \epsilon \Delta_x\} \tag{1.37}$$

1.6.2 The influence function

To study the behaviour of the estimators when the assumptions of the parametric model may be violated several measures have been developed.

A very important such measure in robustness theory is the *influence function* which was invented by Hampel (1968)and was further developed by Hampel (1974) and Hampel et al. (1986).

Formally, the influence function (IF) of T at F is given by

$$\mathrm{IF}(x; T, F) = \lim_{t \to 0} \frac{T[(1-t)F + t\Delta_x] - T(F)}{t} \tag{1.38}$$

in those $x \in \mathcal{X}$ where this limit exists.

The influence function describes the effect of an infinitesimal contamination at the point x on the estimates, standardised by the mass of the contamination. It can be thought of as the relative change of the estimates caused by a small proportion of additional observations at x. One could say it gives a picture of the infinitesimal behaviour of the asymptotic value, so it measures the asymptotic bias caused by contamination in the observations.

In the multidimensional case, observations take values in an arbitrary space $\mathcal{X}$ and the parameters are vector-valued, so that $\Theta \subset \Re^p$. Functionals T can be defined on a suitable subset of the set of probability measures on $\mathcal{X}$, taking values in Θ. The p-dimensional influence function of a functional T at a distribution F is given as in (1.38).

A useful property of the influence function is the following: if some distribution G is 'near' F, then the first-order von Mises expansion of T at F (which is derived from a Taylor series) evaluated in G is given by

$$T(G) = T(F) + \int IF(x;T,F)d(G-F)(x) + remainder. \tag{1.39}$$

1.6.3 Robustness measures derived from the Influence Function

Hampel (1974) introduced some summary values of the Influence Function which describe local robustness properties of the estimators. These are the following: i) the *gross-error sensitivity* of T at F, which is defined by the supremum of the absolute value:

$$\gamma^*(T,F) = \text{sup}_x|\text{IF}(x;T,F)|. \tag{1.40}$$

The gross-error sensitivity measures the worst (approximate) influence which a small amount of contamination of fixed size can have on the value of the estimator. It may be regarded as an upper bound on the (standardised) asymptotic bias of the estimator. It is a desirable feature that $\gamma^*(T,F)$ be finite, in which case we say that T is *B-robust* at F. Here, the B comes from *Bias.*

ii) the *local-shift sensitivity*, which measures the worst (approximate and standardised) effect of shifting an observation slightly from the point x to some neighbouring point y and is given by

$$\lambda^* = \text{sup}_{x \neq y} \frac{|\text{IF}(y;T,F) - \text{IF}(x;T,F)|}{|y-x|}. \tag{1.41}$$

Note, that even an infinite value of λ^* may refer to a very limited actual change, because of the standardisation by $|y-x|$.

iii) the *rejection point*, which is defined for a symmetric F around 0 and is given by :

$$\rho^* = \inf\{r > 0; \text{IF}(x;T,F) = 0 \quad \text{when} \quad |x| > r\}. \tag{1.42}$$

The rejection point is the point after which the IF becomes zero. It is desirable that the IF becomes zero in some region, so that contamination in that region will not have any influence at all. Therefore, it is desirable that ρ^* is finite.

1.6.4 Global robustness measures - the breakdown point

The influence function is, by construction, an entirely *local* concept. Therefore, it must be complemented by a measure of the *global* reliability of the estimator, which describes up to what distance from the model distribution the estimator still gives some relevant information. In Hampel et al. (1986) it is explained that the term 'relevant information' refers to some arbitrary value of the Prohorov distance $\pi(G, F)$ (Prohorov 1956) between contaminated distributions G and the 'true' distribution F. The Prohorov distance of two probability distributions F and G is defined by

$$\pi(F,G) = \inf\{\epsilon; F(A) \leq G(A^{\epsilon}) + \epsilon \quad \text{for all events A}\}, \tag{1.43}$$

where A^{ϵ} is the set of all points whose distance from A is less than ϵ.

Formally, the breakdown point ϵ^* of the sequence of estimators $\{T_n; n \geq 1\}$ is defined by:

$$\begin{aligned} \epsilon^* = \sup\{\epsilon \leq 1; \quad \text{there is a compact set} \quad K_\epsilon \overset{\subset}{\neq} \Theta \quad \text{such that} \\ \pi(F,G) < \epsilon \quad \text{implies} \quad G(\{T_n \in K_\epsilon\}) \overset{n\to\infty}{\to} 1)\}. \end{aligned} \tag{1.44}$$

Intuitively, the breakdown point is the smallest fraction of gross errors which can make the statistic unbounded. It is the distance from the model distribution beyond which the statistic becomes totally unreliable and uninformative. It is therefore desirable that the estimator has a high breakdown point.

1.6.5 M-estimators

Huber (1964) introduced a flexible class of estimators called 'M-estimators' which are a generalisation of maximum likelihood estimators.

The maximum likelihood estimator (MLE) is defined as the value $T_n = T_n(X_1, ..., X_n)$ which maximises

$$\prod_{i=1}^{n} f_{T_n}(X_i), \tag{1.45}$$

or equivalently by

$$\sum_{i=1}^{n}[-\ln f_{T_n}(X_i)] = min_{T_n}! \tag{1.46}$$

where ln denotes the natural logarithm.

Huber (1964) proposed to generalise this to

$$\sum_{i=1}^{n} \rho(X_i, T_n) = min_{T_n}! \tag{1.47}$$

where ρ is some function on $\mathcal{X} \times \Theta$. Suppose that ρ has a derivative $\psi(x,\theta) = \partial/\partial\theta\rho(x,\theta)$, so the estimate T_n satisfies the implicit equation

$$\sum_{i=1}^{n} \psi(X_i, T_n) = 0. \tag{1.48}$$

Any estimator defined by (1.47) or (1.48) is called an M-estimator.

1.6.6 Influence function for M-estimators

If G_n is the empirical cumulative distribution function generated by the sample, then the solution T_n of (1.48) can also be written as $T(G_n)$, where T is the functional given by

$$\int \psi(x, T(G))dG(x) = 0 \tag{1.49}$$

for all distributions G for which the integral is defined.

Let us now replace G by $F_{t,x} = (1-t)F + t\Delta_x$ and differentiate with respect to t, so

$$0 = \int \psi(y, T(F))d(\Delta_x - F) + \int \frac{\partial}{\partial\theta}[\psi(y,\theta)]_{T(F)}dF(y)\frac{\partial}{\partial t}[T(F_{t,x})]_{t=0} \tag{1.50}$$

(if integration and differentiation may be interchanged). Making use of (1.38) and (1.49) we obtain

$$IF(x;\psi,F) = \frac{\psi(y,T(F))}{-\int(\partial/\partial\theta[\psi(x,\theta)]_{T(F)}dF(y)} \tag{1.51}$$

under the assumption that the denominator is nonzero.

Therefore ψ is B-robust at F if and only if $\psi(.,T(F))$ is bounded.

The maximum likelihood estimator is also an M-estimator, corresponding to $\rho(x,\theta) =$

$-\ln f_\theta(x)$, so:

$$\mathrm{IF}(x;T,F)=\frac{\partial/\partial\theta\ln f_\theta(x)}{-\int(\partial^2/\partial\theta^2)[\ln f_\theta(x)]dF(x)} \tag{1.52}$$

The expression

$$\partial/\partial\theta[\ln f_\theta(x)]_{\tilde\theta}=\partial/\partial\theta[f_\theta(x)]_{\tilde\theta}/f_{\tilde\theta}(x) \tag{1.53}$$

is referred to as the score function and will be denoted by $s(x,\tilde\theta)$. The MLE corresponds to $\sum_{i=1}^n s(X_i, MLE_n)=0$.

1.7 Data Contamination and Latent Trait Models

Many researchers in the field of latent trait models have observed that estimating latent trait models is hampered by guessing, inattention to easy questions, cheating, faking, or simply errors when recording or recoding the answers. There have been a lot of studies, particularly within the area of educational testing, aiming at the identification of 'aberrant' response patterns, and also studies that proposed various methods of estimating ability robustly, particularly in the Rasch model.

Waller (1974) proposed an estimation procedure for the latent trait model that uses the information contained in the interaction between a person and an item to remove most of the effects of random guessing from estimates of ability, and from estimates of the difficulty and discrimination parameters. This is accomplished by removing from the estimation procedure those item-person interactions characterised by the item being too difficult for the model and therefore likely to invite guessing. So, the probability of positive response for item i by individual l is given by

$$\pi_{il}=F(a_{0i},a_{1i}|\theta_l) \quad \text{if} \quad F\geq P_c \quad \text{or} \quad g_{il} \quad \text{if} \quad F\leq P_c \tag{1.54}$$

where, F is the cumulative probability from the normal or the logistic ogive, θ_l is the ability of subject l, a_{0i} and a_{1i} are the difficulty and discrimination parameters, $g_{il}=$ Pr (person l guesses item i correctly given $F\leq P_c$), and P_c is some small probability. So the estimate of any person's ability is based on only those items for which there is a reasonable chance that the person achieved the correct response through the interaction of his ability and the item characteristics.

A preliminary estimate of a subject's ability is obtained from the approximate transformation of his per cent correct, inverse normal or logistic. Then an iterative procedure starts, where the probability of correct response F is estimated for each subject's response to an item. Any item for which this estimated probability is less than some small probability, the cut-off point P_c, is omitted. The optimal value P_c is found by fitting different values and taking the one that provides the best fit. The whole estimation procedure is joint maximum likelihood, i.e. estimates of θ_l are obtained by directly maximising the likelihood with respect to them. The performance of the uncorrected 2-parameter model and the modified 2-parameter model were compared with simulated data. Non-guessing and guessing data sets were generated. Waller's model provided the best goodness-of-fit statistic when the cut-off point used in the estimation procedure was the one used to generate the data. In that case the model did also better in recovering the true difficulty parameters. Waller compared the information obtained with the uncorrected 2-parameter model, the modified 2-parameter model and the 3-parameter guessing model with real datasets. The modified 2-parameter model gave the largest average information, but all the models did better than the others on some part of the ability range, no model could outperform the other along the whole ability range.

Wright and Mead (1976, unpublished manuscript, see Wainer and Wright (1980)) proposed a method based on the residuals for each item's response for each person (the 'WIN' method) in the Rasch model. A t-statistic is calculated for the fit of the person's response pattern and if it is greater than some chosen value then all items more than two logits above the person's ability estimate are omitted from the person's response pattern and a new ability estimate is obtained, based upon the shortened test. The process is repeated until an acceptable t-value is achieved or the response pattern gets too short to work with.

Wright (1980) and Smith (1982) proposed three indices for the analysis of the residuals: the unweighted total fit statistic, which compares observed and expected responses for each item for the whole response pattern, the unweighted between fit statistic, which compares the person's predicted score with the observed score on any subset of items and the unweighted within-set fit statistic, which compares observed and expected response for each item in a single subset of items.

This systematic analysis of each person's response pattern is referred to as *person analysis.*

Trabin and Weiss (1983) examine the fit of individuals to item response models by the means of the 'person response curve'. To construct a person response curve the items are ordered by difficulty levels into strata and then the proportion correct for an individual on each stratum is plotted against the stratum. The person response curve provides a means to study testee response variability and the fit of individuals to the IRT model. The lower right-hand portion of the curve for each testee (proportion correct for the most difficult items) provides information on their guessing behaviour whereas the upper left-hand corner provides information on the carelessness of the testees. The curve will also show any deviation from a unidimensional response pattern, that is if a testee was answering correctly beyond the chance level some difficult items which were beyond his/her ability level. Trabin and Weiss also plotted the expected person response curves, by plotting the probability of positive response from an IRT model against the difficulty of the items.

Levine and Rubin (1979) and Levine and Drasgow (1983) acknowledge the fact that some examinees are affected in their answers by test anomalies and thus the ability measure obtained from a latent trait model will not be valid for those examinees. They develop techniques which they call 'appropriateness measurement' and which aim to identify 'inappropriate' test scores. Appropriateness measurement is implemented by statistics, called appropriateness indices, that measure the degree to which an examinee's answer pattern is 'unusual', that is, unlike the pattern expected from typical examinees. In appropriateness measurement studies, examinees are sorted into two groups: (a) examinees with very unusual answer patterns and (b) all other examinees. Levine and Rubin (1979) identified three types of appropriateness indices and reported positive empirical findings from simulation studies. Levine and Drasgow (1983) extended these studies by using actual data and the estimated parameters instead of the actual values and reported similar detection rates.

Wainer and Wright (1980) examined robust estimation of ability in the Rasch model, They proposed the use of the jackknife (Mosteller and Tukey 1977) and a robustified jackknife to estimate ability. The jackknifed estimate of ability is a weighted mean of the ability estimates obtained from standard Maximum Likelihood estimation using all items, and the ability estimates obtained from standard ML estimation obtained by omitting each item in turn (these are called the jackknifed ability estimates). The robustified jackknife, called 'Amjack', is a robust function (used instead of the mean) of the jackknifed ability estimates, which is the Sine M-estimator (Andrews, Bickel, Hampel, Huber, Rogers, and Tukey 1972). Wainer and Wright (1980) compare the jackknife and robustified jackknife to the standard ML estimates of ability, to the traditional guessing correction (putting a lower asymptote on the item characteristic curve) and Wright and Mead's WIN method. In their simulations Wainer and Wright found that the jackknife and robustified jackknife performed well in terms of accuracy and efficiency when there was mild guessing or other distortions. When there was a fair amount of guessing WIN did better. The lower asymptote method performed well when there was a lot of guessing, the test was long and the ability of the examinees was low.

Mislevy and Bock (1982) applied Tukey's *biweight* (Mosteller and Tukey 1977) in estimating ability, to allow for response disturbances. They applied this to the 2-parameter latent trait model where the probability of correct response is given by

$$\pi_{li} = \exp(a_{0i} + z_l a_{1i})/(1 + \exp(a_{0i} + z_l a_{1i})) \tag{1.55}$$

The probability of a response pattern is given by

$$\pi_l = \prod_{i=1}^{p} \pi_{li}^{x_{li}} (1 - \pi_{li})^{(1-x_{li})} \tag{1.56}$$

The items parameters are assumed to be known and so (1.56) can be regarded as the likelihood function of z_l given the response pattern. The maximum likelihood of z_l is the value which maximises (1.56) with respect to the observations. Tukey's biweight is given by

$$\tilde{x} = \sum w_i x_i / \sum w_i \tag{1.57}$$

where $w_i = (1 - u_i^2)^2$ if $|u_i| \leq 1$ and 0 otherwise, and $u_i = (x_i - \tilde{x})/cd$. Tukey defines

d as half the inter-quartile range, roughly comparable to a standard deviation, and c as an arbitrary constant. Data points are assigned decreasing weights as they depart from the biweight and non at all if they lie beyond c inter-quartile range units away. It is noted that the biweight must be computed iteratively, since the biweight depends on the weights and the weights depend on the biweight. Mislevy and Bock (1982) define $\tilde{z}_l$, the biweight estimate of latent ability, analogously, by solving the modified likelihood equation:

$$\sum_{li} w_{li}(x_{li} - \pi_{li})a_{1i} = 0 \tag{1.58}$$

but with

$w_{li} = (1 - u_{li}^2)^2$ if $|u_{li}| \leq 1$ and 0 otherwise,

and $u_{li} = (-a_{0i} - \tilde{z}_l \, a_{1i} \,) \,/\, \mathrm{c}$

The term u_{li} represents the distance between item i and the subject z_l multiplied by the discrimination of the item. In the Rasch one-parameter model each a_{1i} can be set equal to 1. In that case u_{li} is simply proportional to the difference between estimated ability and item difficulty.

The choice of the constant c is arbitrary. Large values mean that little trimming of data will occur, whereas small values would lead to a lot of trimming. An item's influence is greatest if $-a_{0i}$ is equal to the biweight estimate of ability for subject l × the discrimination parameter, as this will give $u_{li} = 0$; influence drops for items further from the subject, as the distance gets large for either easy items and high ability or for difficult items and low ability. It reaches zero when u_{li} gets equal to or exceeds 1.

Mislevy and Bock (1982) propose estimation with Newton-Raphson iteration. The algorithm converged, but one type of response pattern was troublesome: If a subject missed every item except one or two of the hardest items, the biweight procedure would confront a vector of all incorrect responses - a pattern which gives an infinite maximum likelihood ability estimate. Analysis with simulated data showed that the biweight estimates of ability have smaller bias, and although they are more dispersed than the maximum likelihood estimates the reduction of bias gives a smaller mean square error than the maximum likelihood estimator even when the measurement disturbances are mild.

Smith (1985) compared Mislevy and Bock's Biweight, with Wainer and Wright's Amjack and person analysis within the Rasch model using simulated data. Smith (1985) tested the estimation methods in the presence of guessing and in the presence of 'startup' disturbance, that is when subjects perform poorly in the beginning of the test, because of anxiety. He found that Biweight recovered the generating ability better than Amjack and ML, in both instances. Though when ML was combined with person analysis, it did better than the robust procedures.

The studies reviewed above except Wainer and Wright (1980) and Mislevy and Bock (1982) look for aberrant responses in the data according to some ad-hoc criteria and propose methods of downweighting or eliminating them. With simulation studies, i.e. by generating artificial data with and without contamination, they assess how well different estimation methods do in estimating the true ability parameters. Wainer and Wright's Amjack and Mislevy and Bock's Biweight are estimation methods that do not require the examination of each response pattern separately, since the estimation procedure will 'automatically' downweight unusual response patterns. Smith's study showed limitations for both methods.

In our study we will measure the sensitivity or robustness of the latent trait model when aberrant response patterns exist. We do so by calculating the Influence Function and other robust statistics tools. Large Influence Function values for some response patterns mean that additional or fewer observations on those response patterns will affect a lot the estimation of the model parameters and the scoring of the latent variable. Large Influence Function values are associated with outliers so they may point out unusual response patterns with inconsistencies in the responses. One should thus be aware of those and if possible check whether they can be regarded as valid observations or they are the result of gross errors or other nuisance factors.

1.8 Sensitivity of the latent trait model to the prior distribution

The specification of model (1.4) requires an assumption to be made about the form of the distribution of the latent variable y. This is called the underlying or the prior distribution of y. The distinction is sometimes made between 'underlying' and 'prior', depending on whether one refers to the distribution generating the responses or the distribution used to fit the model.

The assumption about the distribution of y is arbitrary, therefore it is important to know whether the parameter estimates are sensitive to the form of this distribution.

The effect of the form of the prior in the analysis has been investigated by Bock and Aitkin (1981), Bartholomew (1988), Bartholomew (1993), Seong (1990).

Bock and Aitkin (1981) proposed instead of making arbitrary assumptions about the distribution of ability, to estimate it as a discrete distribution over a finite number of points. As the estimate of the ability distribution at the point z_t they used the posterior density given the data, which they called the empirical distribution and is given by:

$$h^*(z_t) = \frac{\sum_l r_l f(x_l|z_t)h(z_t)}{\sum_t \sum_l r_l f(x_l|z_t)h(z_t)} \tag{1.59}$$

where r_l is the observed frequency of the response pattern x_l. The values obtained from (1.59) can be used in place of $h(z_t)$ in the likelihood equations, so effectively in (1.28) and (1.29) to obtain improved estimates of the parameters. As starting point for the estimation of the empirical distribution Bock and Aitkin proposed the use of the $N(0, 1)$ as this represents maximum uncertainty in a distribution with finite mean and variance.

They fitted the two parameter normal ogive model to two data sets using the N(0,1), the empirical and the rectangular distributions as priors. To free the estimates from the dependence on the mean and variance of the prior they imposed the following restrictions on the parameters: $\prod_i^p a_{1i} = 1$ and $\sum_i^p a_{0i} = 0$ (a_{1i} and a_{0i} are the parameters of the normal ogive model). The parameter estimates they obtained based on the nonnormal priors were almost indistinguishable from the ones obtained based on the N(0,1) distribution.

Bartholomew (1988) examined the sensitivity to the prior of the logit model by taking different shapes of standardised symmetric distributions and looking at their effect on the first and second order margins of the item response distribution.

He first briefly examines the literature, which suggests that methods which depend only on the first and second marginal proportions yield results very close to those which are based on the comparison of the observed and expected values of the response pattern frequencies. Therefore if the choice of the prior has negligible effect on the expected first and second order marginal proportions then it will have little effect on the numerical values of the estimates.

He calculated the first and second order expected margins for different sets of parameters, in the range of [-4,4], for different priors. The priors he considered were:

i) a three-point distribution with $P(z = +z_0) = 0$, $P(z = -z_0) = 0$ and $P(z = 0) = 1$,

ii) a three-point distribution with $P(z = +z_0) = 1/2$, $P(z = -z_0) = 1/2$ and $P(z = 0) = 0$,

iii) the normal,

iv) the logistic

and v) the rectangular distribution.

His results show that the logistic, normal and rectangular priors give very similar results with difference rarely amounting to more than one or two units in the second decimal. Of the extremal distributions (i,ii) the first is the poorer but still the change in the margins is relatively small (one or two units in the first decimal place for the first order margins and one unit in the first decimal for the second order margins) considering how unrealistic the distribution is. The greatest discrepancies occur when the parameters are relatively large. A prior with very long tails would then lead to marked effect but this is not practically reasonable. He therefore concludes that any symmetrical prior will predict essentially the same first and second order margins and he recommends the normal prior for general use since it has been most widely used in practice and it has rotational advantages in models with several latent variables (Bartholomew and Knott 1999).

Seong (1990) investigated the effect of the prior distribution on the item and ability parameter estimates together with the following factors: i) the number of quadrature points, ii) the number of examinees and iii) the type of the underlying ability distribution. His results were based on simulated data. He used the terms ‘underlying’ and ‘prior’ distinctly for the distribution of the latent variable, depending on whether he was referring to the one used in the generation of the response patterns or the one used in the estimation of the parameters. Both distributions were approximated by quadrature points and weights. He used the two-parameter normal ogive as a response function and the Bayesian EAP for the ability parameter estimation.

As measures of accuracy he used the square root of the mean squared difference between the estimates and the parameters (RMSE) and the mean of the absolute difference between estimates and parameters. In addition he employed a split-plot factorial design to investigate the effects of the above factors. The dependent variables in the ANOVA were the log of the average of the squared differences (LMSE) between estimates and the parameters, for item difficulty, item discrimination and ability. His results showed that the average difference measures were decreased when the two ability distributions were matched. Also, more than half of the total sum of squares of the LMSE for ability estimation was due to the specification of the prior ability distribution. The results suggested that it is important that users should consider the match between the type of underlying and prior ability distributions employed, but when users do not have an idea about the type of underlying distribution, it is better to use the normal prior distribution than to choose an inappropriate prior distribution.

Zwinderman and van den Wollenberg (1990) examined the effect of misspecification of the latent ability (θ) distribution on the accuracy and efficiency of marginal maximum likelihood (MML) item parameter estimates and on MML statistics to test sufficiency and conditional independence for the Rasch model. The results were compared to the conditional maximum likelihood approach (CML). In their simulations they generated data from a standard normal distribution and from exponential distributions with various means to have various degrees of skewness. They fitted the Rasch model with conditional maximum likelihood and marginal maximum likelihood assuming a normal distribution for ability. When the underlying distribution was the standard normal MML gave more accurate and efficient estimates than CML. When though a normal prior was used when the underlying distribution was skewed, MML estimators were less accurate and less efficient than CML estimators. The effects were not large, though they increased as the skewness of the underlying distribution increased. CML estimators were also affected as the skewness of the number-correct distribution increased together with the skewness of the underlying distribution. The statistic R_{1m} statistic (Glas 1989) to test the sufficiency assumption of the Rasch model in the MML approach, based on the observed and expected frequency for each total score, was severely affected by the misspecification of the prior distribution.

Bartholomew (1993) noted that the population mean and standard deviation are confounded in the parameter estimates. Suppose that the latent variable y is distributed with mean μ and standard deviation σ and let $z = (y - \mu)/\sigma$. For the one-factor logit model we will have:

$$\text{logit}\pi_i(y) = a_{0i} + a_{1i}y \tag{1.60}$$

We can transform this model from the scale of y to that of z obtaining

$$\text{logit}\pi_i(z) = a_{0i} + a_{1i}(\mu + \sigma z) = a_{0i} + a_{1i}\mu + a_{1i}\sigma z = a^*_{0i} + a^*_{1i}z \tag{1.61}$$

Usually we assume a standardised distribution, so it is the parameters a^*_{0i} and a^*_{1i} that we estimate in practice.

The confounding of the item parameters with the population characteristics has some intuitive meaning. A question which is a poor discriminator in a population where attitudes or an ability show little variation may be quite discriminating in one where they are more highly dispersed. Similarly, questions which appear conservative in a radical population may be radical to conservatives. Or, in the context of ability, a question will be regarded as easy or difficult depending on the underlying ability of the students it has been given to.

This presents no problems if we are considering a single population, but may present difficulties if we are interested in several populations, or the same one at different times. Since we regard the a's as intrinsic properties of the questions, if we fit the logit model to two populations, any systematic differences in the estimates can be attributed to differences in μ and σ. (1.61) shows that differences in location, measured by μ, will show up in a_{0i} and differences in σ will affect a_{1i}. A non-systematic set of differences in either the difficulty of the discrimination parameters could not be due to a population shift in either location or dispersion.

In our study we will further investigate the sensitivity of the latent trait model to the form of the prior through the Influence Function. The Influence Function will show the behaviour of the estimates for small changes in the prior. We will also examine gross changes in the prior by fitting mixtures of normals as priors.

1.9 Semiparametric estimation of the latent trait model

Bock and Aitkin (1981) use of an empirical prior has also been referred to as 'semiparametric' estimation of the latent trait model, since there is no assumed parametric form for the prior. Since then methods have been developed to estimate latent trait models, particularly the Rasch model, without restricting the nodes of the latent distribution. This has been called nonparametric estimation of the latent distribution, or 'fully semiparametric' estimation of the latent trait model.

Nonparametric estimation of the mixing distribution The latent distribution is in effect a mixing distribution so the results on estimating nonparametrically a mixing distribution apply to some extent to latent trait models.

Most of the research on this topic has been based on the results of Kiefer and Wolfowitz (1956). In their paper they proved that the maximum likelihood estimator of a structural parameter (a parameter that relates the explanatory variables to the outcome of interest) is strongly consistent, when the (infinitely many) incidental parameters (parameters that represent the effects of omitted variables) are independently distributed random variables with a common unknown distribution, F. Moreover, F is also consistently estimated, although it is not assumed to belong to a parametric class.

Laird (1978) examines nonparametric maximum likelihood estimation of a distribution function F with the following probability model: $x^T = (x_1, ..., x_n)$ is the observed data vector and $y^T = (y_1, ..., y_n)$ is an unobserved random sample, with distribution function $F(y)$. The sample space of the y_i's is and interval on the real line. Conditionally on y_i, each x_i is independently distributed with density $h_i(x_i|y_i)$. Marginally, each x_i has density

$$g_i(x_i|F) = \int h_i(x_i|y)dF(y) \tag{1.62}$$

The parametric form of $h_i(.|.)$ is assumed to be known. Laird shows that the nonparametric maximum likelihood estimate of the mixing distribution F is self-consistent and then she uses this property to prove that, under certain conditions, the estimate must be a step function with finite number of steps. Then the mixing distribution is characterised uniquely by the location of the steps and the amount of probability at each step and the estimation problem reduces to that of maximum likelihood estimation of the parameters of a mixture of k densities. k is taken to be the smallest integer such that the locations of the points are all distinct and their probabilities all positive. k is not known in advance but one can start with a small k and maximise the likelihood until all locations are distinct and the probabilities all positive or start with a large k and take k to be the number of distinct steps with positive probability.

Lindsay (1983) studied the geometry of the likelihood of the estimator of a mixture density Q and gave conditions on the existence, discreetness, support size characterisation and uniqueness of the estimator. The mixture density corresponding to the mixing distribution Q is defined with

$$f_Q(x) = \int f_\theta(x) dQ(\theta) \tag{1.63}$$

where the f_θ 's are called the atomic densities because they correspond to the atomic mixing distributions which assign probability one to any set containing θ.

Suppose the observation vector $\mathbf{x}$ has K distinct observations $y_1, ..., y_K$ and n_k is the number of x's which equal y_k. The atomic and mixture likelihood vectors are defined as $f_\theta = (f_\theta(y_1), .., f_\theta(y_K))$ and $f_Q = (f_Q(y_1), .., f_Q(y_K))$

Lindsay defined the function $D(\theta, Q)$

$$D(\theta, Q) = \sum_{k=1}^{K} n_k(\frac{f_\theta(y_k)}{f_Q(y_k)} - 1) \tag{1.64}$$

which is the directional derivative of the loglikelihood at Q towards θ, and used this to determine various properties of the estimator. He proved that the measure $\hat{Q}$ which maximises $\log L(Q)$ also minimises $\sup_\theta D(\theta, Q)$ and that $\sup D(\theta, \hat{Q}) = 0$. Moreover, the support of $\hat{Q}$ is contained in the set of θ for which $D(\theta, \hat{Q}) = 0$. The maximum number of support points is the number of distinct observations. Lindsay gave an algorithm to estimate the mixing distribution, the vertex direction method (VDM), which uses $D(\theta, Q)$, and also suggested the use of the VDM in alternation with the EM to guarantee convergence.

Nonparametric estimation of the Rasch model Results regarding the nonparametric estimation of a mixing distribution have been applied to the Rasch model. In the following we review the most important papers.

Cressie and Holland (1983) formulate the marginal probabilities of the response patterns in terms of the conditional odds of responding positively to an item, $V_i = \pi_i(z)/(1 - \pi_i(z))$, and without making any assumptions about the distribution function G of z, with $p(\mathbf{x}) = p(\mathbf{0}) \int [\Pi_i V_i(z)^{x_i}] dG(z)$, or $p(\mathbf{x}) = p(\mathbf{0}) E[\Pi_i V_i(z)^{x_i}]$, where $p(\mathbf{0}) = \int \Pi_i (1/(1 - \pi_i(z)) dG(z)$. The marginal probabilities are therefore shown to be proportional to the moments of a special type of positive random vector (z is defined to be positive). Using this result they reexpress $p(\mathbf{x})$ for the Rasch model as a loglinear model, $\log p(\mathbf{x}) = \eta + \sum_{i=1}^{p} x_i \beta_i$, some parameters of which are the first p moments of $u = \exp(z)$, where p is the number of items. If the loglinear parameters satisfy certain inequalities (Karlin and Studden 1966) then there exists a distribution function G and its first p moments are defined.

De Leeuw and Verhelst (1986) put the various versions of the Rasch model in a general framework, making the distinction between a functional model, where the individuals are considered parameters, and the structural model, where the individuals are characterised by a random variable with distribution F_i, which may be equal across individuals. They then focused on the nonparametric estimation of the Rasch model, that is when the common distribution F is completely unknown. They proved that the marginal maximum likelihood estimates of the item parameters in the structural model, in which $F_i = F$, are equal to the conditional maximum likelihood estimates when the Rasch model holds, that is when the fitted proportions of the score totals are equal to the observed group totals. These equations, where the fitted proportions under CML are put equal to the fitted proportions under MML, can be represented as power moment equations and can be solved for the latent distribution, using properties of Tchebyscheff systems (Karlin and Studden 1966). The solutions are step functions, with $(p+1)/2$ number of steps if p, the number of items, is odd, and $(p+2)/2$ if p is even, with the first point being equal to $-\infty$ in the latter case.

The same ideas were developed independently by Follmann (1988). He also applied the results of Karlin and Studden to the Rasch model to show that the class of all marginal logistic models is equivalent to the class of marginal logistic models based on discrete latent distributions with at most $(p+2)/2$ latent abilities.

Lindsay, Clogg, and Grego (1991) gave more general results regarding the estimation of the Rasch model. They first made a distinction between concordant and discordant cases: A case is concordant if the marginal frequencies of the total scores can be fitted exactly by the model; it is discordant in all other cases. They then stated that for concordant cases CML and MML estimates of the item parameters are identical, as De Leeuw and Verhelst (1986) had proved, but the latent distribution cannot be estimated uniquely in all cases. More specifically, in the case of 'borderline concordance', i.e. when at least one of the determinantal inequalities that arise in checking for concordance is an equality, the estimate is unique. Moving above a critical number of points will not improve the fit and the resulting distribution will degenerate to the one obtained with the critical number of points.

When all the inequalities hold strictly though (positive definite concordance) the latent distribution cannot be estimated uniquely. Increasing the number of critical points will not improve the fit but will give different distinct latent distributions.

In the discordant cases, CML and MML estimates are different. Then the estimate of the latent distribution is unique and the maximum number of support points needed is $p/2$.

Semiparametric estimation of other latent trait models Bock and Aitkin (1981) suggested to fit the 2-parameter model using an empirical distribution, i.e. a distribution which will be estimated from the data simultaneously with the parameters, as we saw in Section 1.8. In their estimation method they fixed a grid of approximating points and they estimated at each iteration of the E-M algorithm the weights of the points. This method will be described in detail in Chapter 4, Section (4.2).

Mislevy (1984) gives a non-parametric approach of estimating the latent distribution, 'a smooth continuous m-variate distribution with finite moments', without specifying a particular latent trait model. The continuous density is approximated by a discrete distribution on a finite number of points- in effect an m-dimensional histogram. Only the weights though are estimated from the data, the points have to be defined beforehand. He refers to Laird (1978) for the simultaneous estimation of points and weights in the unidimensional case but does not implement it for any particular latent trait model. He suggests the use of the points and weights to calculate moments of the latent distribution. In the case where the form of the latent distribution is known (if it is a parametric distribution) then the loglikelihood can be differentiated w.r.t. the parameters and the equations can be solved for those, iteratively and using quadrature or with Monte Carlo integration. Mislevy implements this for the multivariate normal and beta-binomial distributions.

Links with latent class analysis Semiparametric estimation makes latent trait models more similar to latent class models. In latent class analysis the response probabilities of the items are not linked to the latent classes through a parametric form, and also the positions of the classes in not restricted in any way. In fully parametric latent trait model the response probabilities are linked though a parametric function to the underlying continuous latent variable and the distribution of the latent variable is parametrically specified.

In semiparametric estimation there is no restriction on the form of the latent variable distribution. It turns out to be that this is a step function but the position of the nodes and their probabilities are estimated from the data. Thus one can consider them as latent classes, where the classes are nodes along a continuum.

Croon (1990) considered the question of how an ordered relation may be defined on the classes in a latent class model. He proposed to impose inequality restrictions on the item response and the cumulative item response probabilities to impose ordinality of the latent classes. So, the order relation on the set of latent classes is defined by imposing on each item that the probability of 'positive' response should be an increasing function of the latent class number, or equivalently, that the probability of a 'negative' response should decrease as a function of the latent class number. In the case of a dichotomous item i with ordered response categories 0 (for a negative response) and 1 (for a positive response) the definition of ordered latent classes amounts to the requirement that the probabilities of positive response belonging to class t $P(X_i = 1|t)$ satisfy a certain monotonicity condition that can be represented by the following system of inequalities: $P(X_i = 1|t) \leq P(X_i = 1|t + 1)$, for $t = 1, 2, ..., T - 1$. For the estimation of the ordered latent class models he discusses an E-M algorithm, where during the M(aximization) step the item response probabilities have to be re-estimated under the inequality restrictions given above. He shows that this constrained concave optimization problem is formally identical to the problem of estimating a sequence of stochastically ordered distribution functions, for which algorithms exist. In his examples he discusses similarities of the ordered latent class model with the 2-parameter latent trait model. From the ordered latent class analysis, clear differences among the items emerged with respect to the way in which their response probabilities and cumulative response probabilities vary as a function of the latent variable. The two most obvious aspects of this relationship are its overall level and its steepness. So, if one compares the way the response probabilities of two items vary as a function or the ordinal latent variable, one will observe that irrespective of which class one considers, the response probabilities of one item will always be lower. This is related to the 'popularity', or 'difficulty' of the items. The steepness, on the other hand, which is how drastically the probabilities of an item change as one moves along the latent classes, has to do with the discriminatory power of the item.

From the estimated proportions of respondents belonging to each latent class, one can get a rough idea of how the respondents are distributed along the latent continuum, for example, whether the respondents are concentrated at a central position or at the extremes.

Croon notes that one cannot choose the optimal number of latent classes on the basis of a statistical test, as the results concerning the asymptotic distribution of the likelihood ratio test statistic do not apply. He therefore used in his examples the Akaike's Information Criterion instead.

Relations between latent trait and latent class model were also examined by Haertel (1990). He gave conditions for the two-latent class and two-parameter normal ogive model to agree and presented relations between items and generalised these to continuous models with more than one latent trait and discrete models with more than two latent classes. In particular, he equated the response probabilities of the 2-latent class model to the conditional response probabilities of the 2 parameter normal latent trait model at the two points that had been used to estimate the prior, to obtain the estimates of the two item parameters, the threshold and the slope.

Haertel noted that one could get the parameters of the two parameter latent trait model from the response probabilities of the latent class model and not vice versa, because of the indeterminancy of item parameters due to location and scale of prior. Although, he said, there is no direct correspondence between the item parameters of the two models, he proposed the log-odds ratio,

$$w_i = \ln[p_{2i}(1 - p_{1i})/p_{1i}(1 - p_{2i})]$$

as an item 's discrimination index under the latent class model. He also noted that the 2LC model provides a parameter for the latent trait distribution whereas such parameters are introduced in the 2PN model only under ancillary assumptions to facilitate estimation.

Heinen (1996) studied the links between latent trait, latent class and loglinear models and gave equivalences between them. He discussed and gave examples of semiparametric estimation of the Rasch and the 2-parameter latent trait model, though for the 2-parameter logistic model the nodes had to be held fixed.

LEM, a computer program developed by Vermunt (1997), fits various types of latent class and loglinear models. The various models are specified by imposing restrictions on the conditional response probabilities or the loglinear parameters, for example, by having inequality restrictions on the conditional probabilities one can formulate an ordinal latent class model. In LEM, a latent trait model can be defined as a loglinear row-effects model, where the latent variable is the ordinal column variable and the rows are the items. If the scores for the column variable are not given, then the latent trait model is parameterised as a row column association model and LEM can estimate the scores of the column variable, so this makes it equivalent to the fully semiparametric estimation of the latent trait model. However, in case of binary data, there are not enough degrees of freedom to estimate a row column association model. Thus the above parameterisation can only be defined for polytomous items.

Nonparametric estimation of the mixing distribution in other areas of application Nonparametric estimation of the mixing distribution has been applied in the estimation of Generalized Linear Models (Aitkin 1996), in longitudinal analysis (Davies 1987) and in two-level variance component models (Hinde and Wood 1987), where the mixing distribution is used to model overdispersion. We summarise Aitkin's results as these seem more closely linked to the estimation of the latent trait model.

Aitkin (1996) gives a nonparametric maximum likelihood estimation method of the mixing distribution in the case of overdispersed Generalized Linear Models. The mixing distribution has been used in GLM's to take care of overdispersion. Specifying a parametric form for it may affect the parameter estimates, thus its non-parametric estimation is of great value. Aitkin's method is based on the paper by Hinde and Wood (1987) where they address the computational issues of NPML in the more general framework of two-level variance component models. The model is specified as follows: $x = (x_1, ...x_n)$ is a random sample from exponential distribution $f(x|\theta)$ with canonical parameter θ and mean μ, and explanatory variables $U = (u_1, ...u_n)$ related to μ through a link function $\eta_i = g(\mu_i)$ with linear predictor $\eta_i = \beta' u_i$. Overdispersion is modelled by incorporating an unobserved random effect z_i for the ith observation.

The likelihood of an observation is given by integrating over the distribution of z. If one assumes a parametric form for z the integral is approximated by a sum over k mass-points z_k with masses p_k (appropriate for the chosen form).

The likelihood is thus the likelihood of a finite mixture of exponential family densities with known mixture proportions p_k at known mass-points z_k, with the linear predictor for the ith observation in the kth mixture component being $\eta_{ik} = \beta' u_i + \sigma z_k$, where σ is the scale parameter of the mixing distribution. One can regard this as the exact likelihood for this discrete mixing distribution for z. For the NPML estimation of the masses and mass-points, the masses and mass-points are treated as unknown parameters. The number K of points is also unknown but treated as fixed, and sequentially increased until the likelihood is maximised. Since the variance of the mixing distribution is a function of the unknown parameters, the scale parameter is dropped and the mass-point parameters are defined as α_k, with linear predictor $\eta_{ik} = \beta' u_i + \alpha_k$. α functions as an intercept parameter for the kth component: it can immediately be estimated simply by including a 'component' factor in the model with K levels instead of variable z_k (Hinde and Wood 1987). Estimates of the weights are obtained by differentiating the likelihood with respect to them. The model is estimated by the same EM algorithm, with an additional calculation in the M-step of the estimate of p_k from the posterior probabilities (the probabilities of the observations given the component).

Lindsay and Lesperance (1995) reviewed methodological developments in semiparametric maximum likelihood estimation of mixture models and gave examples in various areas of application, whereas Boehning (1995) reviewed algorithms for nonparametric estimation of the mixing distribution. Most of the algorithms are based on the directional derivative (1.64) and are all variants of the *vertex direction method* (VDM). The *vertex exchange method* (VEM) proposed by Boehning (1985) had the advantage over VDM that it could add a good support point and eliminate a bad support point at each iteration. An improved and faster algorithm, the *intra-simplex direction method* (ISDM), also a variant of VDM was proposed by Lesperance and Kalbfleisch (1992). Susko, Kalbfleisch, and Chen (1998) further developed this algorithm so that it can solve the problem of constrained estimation.

In this thesis we will further explore semiparametric estimation of the prior, using Bock and Aitkin's empirical prior method, and also fully semiparametric estimation with a new algorithm. We will also investigate on the information that can be obtained about the prior from a set of binary responses by measuring the variability of the estimated prior with bootstrap simulations.

1.10 Goodness-of-fit

1.10.1 Goodness-of-fit statistics

Goodness-of-fit indices Goodness-of-fit is tested by comparing the observed and expected frequencies of the 2^p possible response patterns. If n, the sample size, is reasonably large goodness-of-fit may be measured by either of

(a) the likelihood ratio statistic

$$G^2 = 2\sum_{i=1}^{2^p} O_i \log\left(O_i/E_i\right) \tag{1.65}$$

or

(b) the Pearson chi-squared statistic

$$X^2 = \sum_{i=1}^{2^p} \left(O_i - E_i\right)^2 / E_i = \sum_{i=1}^{2^p} \frac{O_i^2}{E_i} - n. \tag{1.66}$$

where O_i and E_i are, respectively, the observed and expected frequencies of the ith response pattern. Under the null hypothesis that the model fits, each statistic is distributed, approximately, like chi-squared with $2^p - 2p - 1$ degrees of freedom, $2p$ being the number of parameters to be estimated.

G^2 and X^2 belong to the family of 'power divergence' statistics (Read and Cressie 1988), all of which take the following form, differing only by the parameter λ:

$$2[\lambda(\lambda+1)]^{-1}\sum_{i=1}^{2^p} O_i[(O_i/E_i)^\lambda - 1] \tag{1.67}$$

The likelihood ratio statistic is obtained for $\lambda \to 0$. With $\lambda = 1$ we get the Pearson chi-square. Read and Cressie suggested $\lambda = 2/3$.

Another set of indices are indices that compare a fitted model with a base-line model. As a base-line we take the value of G^2 when the x's are assumed to be independent. This is obtained in the usual way for testing independence in p-way contingency tables and is equivalent to setting $\alpha_{i1} = 0$ for all i in (2). If we let G_0^2 be the value of G^2 for this special case and G_1^2 be the corresponding value for the logit/probit model then the percentage of G^2 explained is

$$\%G^2 = \frac{G_0^2 - G_1^2}{G_0^2} \times 100 \tag{1.68}$$

$\%G^2$ gives the extent to which the model explains the associations among the x's. An index similar to (1.68) has been used by Krebs and Schuessler (1987) for latent trait models. The notion of the amount of association explained is distinct from that of goodness-of-fit. It is quite possible for a model which fits well to account for only a small part of the association. Nevertheless a poorly fitting model will give rise to a larger than expected G_1^2 and this will have the effect of reducing the percentage of G^2 explained.

1.10.2 The problem of goodness-of-fit

The use of these goodness-of-fit indices is often problematic. The difficulty arises because 2^p increases exponentially with p so that the average expected frequency quickly becomes too small for the chi-squared approximation to the sampling distribution to be valid. For example, if $p = 10$, $2^p = 1024$ and therefore, even with $n = 1000$, there will be many response patterns with E_i's which are much less than 1. The problem may be partly overcome by pooling response patterns so that the expected values for the groups thus formed are large enough (> 5, say) to justify using the chi-squared approximation to the sampling distribution. However, this often leads to a situation where there are no degrees of freedom left. There are two other drawbacks to pooling which become more serious as p increases. First, pooling response patterns results in a loss of information. If $p = 20$, for example, 2^p is close to 1 million and it will then be necessary to pool about 5000 response patterns, on average, to achieve an expected frequency of 5. The test will reveal nothing about the deviations from the model within these groups. We may therefore expect to find an increasing loss of power as p increases and this is borne out by calculations reported later.

Secondly, in the calculation of X^2 we regard all of those response patterns for which $O_i = 0$ as pooled into a single group. This happens in the unpooled case as well, since their contribution is the sum of their expected frequencies and this is the sum of all expected frequencies of the 2^p table minus the sum of the expected frequencies of the NR observed response patterns: $\sum O_i = \sum E_j$, with $i = 1, ..., \mathrm{NR}$ and $j = 1, ..., 2^p$, and so the contribution of the unobserved response patterns to X^2 is equal to $\sum(0 - E_k)^2/E_k = \sum E_k = \sum O_i - \sum E_i$, where $k = 1, ..., 2^p - \mathrm{NR}$.

This observation alone is sufficient to show that the chi-squared approximation cannot be valid because it depends on the assumption that the joint distribution of the O_i's is multinomial. This cannot be so if the pooling is with regard to the values of the frequencies that are pooled.

As for the likelihood ratio statistic, only those response patterns for which $O_i > 0$ contribute to its calculation, so as p gets large, fewer cells out of the 2^p cells are contributing.

These considerations show (a) that there are serious questions about precisely how G^2 and X^2 should be calculated and (b) that the sampling distribution cannot have the χ^2 form if extensive pooling is involved.

Since it is not always clear which method is preferable, we shall calculate two versions of G^2 and X^2:

(i) using (1.65) and (1.66) without pooling

(ii) by pooling response patterns so that the expected frequencies in each group exceed 5. This will be done by proceeding down the list of those patterns for which O_i is non-zero in order of $E(z \mid \mathbf{x})$ and forming groups cumulatively until the expected size criterion is satisfied. Pooling is done on the basis of $E(z \mid \mathbf{x})$ to ensure that the groups so formed are as homogeneous as possible with respect to the latent variable.

To overcome the problems presented when the frequency tables are sparse, we will propose the use of Monte Carlo sampling to approximate the empirical distribution of (1.65) and (1.66) and other derived statistics and also the use of diagnostic methods based on the residuals, in order to draw inferences on the goodness-of-fit of a model.

1.10.3 Studies on the goodness-of-fit indices for categorical data

The statistical problem arising from the sparseness of the contingency table affects other models for data of this kind, and particularly latent class and loglinear models.

There have been a lot of studies that have investigated the behaviour of G^2 and X^2 in a variety of models and conditions but it appears to be difficult to predict whether X^2 and G^2 would behave liberal (too many incorrect rejections) or conservative (to few incorrect rejections). Larntz (1978) concluded that G^2 is liberal, due primarily to the influence that small expected cell frequencies have on this statistic. Koehler and Larntz (1980) found that if there were many expected frequencies less than one, then X^2 would be liberal, rejecting too often. They also found that if the ratio of the sample size n over the number of cells k was less than 0.5 G^2 would be conservative, but if n/k was greater than 0.5 G^2 would be liberal. Everitt (1988) showed that when there were few items in the latent class model G^2 was conservative and when there were many items G^2 was liberal.

In the following we will review in greater detail three more recent studies that have conducted Monte-Carlo simulations to examine the empirical distributions of X^2 and G^2 and then used the empirical distributions of these statistics to draw inferences on the goodness-of-fit of the models tested.

Collins, Fidler, Wugalter, and Long (1993) investigated this problem in relation to the latent class model. They carried out simulations to investigate how the sampling distributions of G^2, X^2 and the Read and Cressie (RC) index for $\lambda = 2/3$, would be affected by the number of items, the number of latent classes fitted, the size of the parameters (conditional response probabilities) and the different ratio of sample size to number of cells (the ratio varied from 1 to 16). They found substantial deviations between the expectation of the chi-squared distribution and the G^2 and the Read and Cressie distributions. In particular, the expectations of the empirical distributions of G^2 and RC were lower than the expectation of the chi-squared distribution if the conditional response probabilities were equal to 0.9, and higher than the expectation of the chi-squared if the conditional response probabilities were equal to 0.65. So if G^2 and RC were compared to a critical value from chi-squared would not be rejecting enough in the first case and they would be rejecting too often in the second case. The mean of the X^2 distribution was closer to the expectation of the chi-squared distribution but its standard deviation much larger than the standard deviation of the theoretical distribution. Collins et al. (1993) carried also simulations to demonstrate procedure of Monte-Carlo simulations to test goodness-of-fit. The behaviour of the indices was consistent with what they found above and the Monte-Carlo procedure gave a clearer picture to the relative fit of the models considered.

Reiser and VandenBerg (1994) used Monte-Carlo simulations to study the behaviour of G^2 and X^2 for the logit/probit latent trait model. In particular, they carried out simulations under a variety of conditions and measured the rate of Type I and II errors of these statistics for two estimation methods, the 'full information method', where the 2^p response pattern frequencies are used in the analysis and the 'limited information' method, where only the first- and second-order marginal frequencies are used in the estimation of the same model. Their results for the full information method showed that when the ratio of sample size to number of cells was smaller than 2, type I error of G^2 was 0, whereas X^2 performed well even with 10 manifest variables and degree of sparseness equal to 0.5. Regarding the power of the tests, G^2 performed satisfactorily when the number of items was up to seven but with more than 8 items (and ratio of sample size to number of cell 1.95) the test had no power against the null hypothesis. The Pearson statistic performed well up through eight variables, and then lost power. In general the limited information method had higher type I error rates but higher power.

Langeheine, Pannekoek, and van de Pol (1996) bootstrapped the G^2, X^2 and the Read and Cressie statistic of latent class, loglinear and latent Markov models fitted to data with various degrees of sparseness. Their analyses showed that the bootstrapped goodness-of-fit indices gave the same p values as the non-bootstrapped ones when there was no sparseness, but when there was sparseness then the bootstrapped p values gave a higher probability of accepting a model as compared with the nonbootstrapped p values. In some cases though there was disagreement between X^2 and G^2 which the bootstrap would not always resolve.

Whereas Collins et al. (1993) and Reiser and VandenBerg (1994) simulated data using the estimated parameters of the model, Langeheine et al. (1996) used a procedure for sampling from a multinomial distribution over the 2^p cells of the table with probabilities estimated from the model and n equal to the sample size.

We will use Monte-Carlo simulations to assess the goodness-of-fit of the Schuessler Social Life Feeling scales. The number of items of these scales vary from 5 to 12, thus allowing us to study the behaviour of the Monte-Carlo method for different number of items. We will also use some simulated datasets to assess the power of the test. We will also propose the use of diagnostic methods to complement the goodness-of-fit test.

1.11 The Data

The data used in this thesis are the 'Social Life Feeling' scales and ability tests from the National Foundation of Educational Research (NFER).

Social Life Feeling scales Schuessler's original data on social life feelings related to 12 scales using American respondents (Schuessler 1982). The study was then extended to German samples in collaboration with Krebs (Krebs and Schuessler 1987). In order to adapt the American scales to the German situation and, especially, to facilitate comparisons between the two countries the scales were further refined by omitting some questions and a few complete scales to yield what the authors called *inter-cultural scales.* There were 9 scales with the numbers of items varying between 5 and 12. There is no scale 7 in the intercultural scales though we will use the American scale 7 in some of our analyses. The questions for each scale are listed in the Appendix. The sampling schemes, described in Schuessler (1982) and Krebs and Schuessler (1987), were designed to yield about 2000 respondents in Germany and 1500 in the USA. The files supplied to us have complete sets of responses of 1490 individuals for Germany and 1416 individuals for the USA. In the case of Scale 4, which applies only to employed persons, the numbers are roughly halved in the German case.

National Foundation of Educational Research data We will use test 1 of the NFER tests for primary school boys (Gorman, White, L.Orchard, and A.Tate 1981).

The test comprises 21 items and the sample size is 566. The test is described in the Appendix.

Attitude to employment data The data are taken from Albanese and Knott (1992b) and Birkhoff (1991). They are the responses to 4 items chosen from 14 items concerning the attitude to work of 1915 German company employees in 1987. These items are part of an investigation about what the employees thought to be the strengths and weaknesses of the company and how they felt about their personal situation at the work place. The items are given in the Appendix.

1.12 Outline of the thesis

In Chapter 2 we will investigate the sensitivity of the parameters of the latent trait model and the posterior means when data are contaminated, using robust statistics tools and empirical data.

In Chapter 3 we will study the behaviour of the model when the assumption about the prior distribution is violated. Small violations will be studied using robust statistics tools and empirical data. We will also study gross discrepancies of the usual assumptions by fitting mixtures of normals as priors.

In Chapter 4 we will investigate semiparametric and fully semiparametric estimation of the latent trait model. We will also measure the information about the prior that can be obtained from fitting a latent trait model to a set of binary responses using bootstrap samples.

In Chapter 5 we will examine the fit of the logit/probit latent trait model to Schuessler's Social Life Feeling scales using Monte-Carlo simulations as an alternative to the usual goodness-of-fit methods. We will also give diagnostic procedures, based on the residuals, which give insight to reasons of poor fit.

In Chapter 6 we will draw the conclusions of this thesis.

Chapter 2

Sensitivity of the logit/probit model to contaminated data

2.1 Introduction

In this chapter we will investigate the effect of data contamination on the parameter estimates and the posterior means, using robust statistical methods and empirical analysis.

By data contamination we mean that some responses arose because of some mechanism operating other than the assumed latent variable. For example, in educational testing data contamination may be present because of guessing, cheating or carelessness. In attitude and personality questionnaires data contamination may be due to faking or carelessness. Mistakes may also occur in the recording of the data.

We will first derive the Influence Function for the parameters of the latent trait model, and study its behaviour for some datasets. We will then study the actual changes in the parameters for artificially contaminated datasets and find the amount of contamination for which maximum likelihood estimation is robust. Other types of contamination, like shifting observations from one response pattern to another, and increasing the probability of positive response for an item, will be investigated.

We will also study the effect of data contamination on the scoring of the latent variable, both theoretically, through the Influence Function, and empirically, through artificially contaminated datasets. At the end of the chapter we will study the behaviour of existing 'robust' estimation methods proposed in the literature for either the parameters or the scoring of the latent variable.

2.2 The Influence Function of the parameters

The influence function (IF) of a parameter at response $\mathbf{x}$ may be thought of as the rate of change of the parameter when a small extra probability is given to response $\mathbf{x}$. Rewriting (1.38) using the notation for the latent trait model, the influence function IF$(a_{ji};\mathbf{x})$ for the parameter a_{ji}, $j = 0, 1$, $i = 1, ..., p$ is given by

$$\lim_{\epsilon \to 0} \frac{a_{ji}(\epsilon) - a_{ji}}{\epsilon} \tag{2.1}$$

where a_{ji} is evaluated from $f(\mathbf{x})$ and $a_{ji}(\epsilon)$ is evaluated from a population of $\mathbf{X}$'s which follow $f(\mathbf{x})$ with probability $(1 - \epsilon)$ and take the value $\mathbf{x}$ with probability ϵ. $\mathbf{x}$ is a response pattern and the population of the $\mathbf{X}$'s the set of all possible response patterns for a given number of items.

We take the parameters a_{ji} for an arbitrary distribution over the responses as being defined by their making the expected value of the score function of the logit/probit model equal to 0. This will allow the influence function to tell us about the sensitivity of maximum likelihood estimates of the a_{ji} to changes in the distribution of responses when the logit/probit model is fitted.

The score function for a_{ji} is given by the following equations, which are expanded forms of (1.27)

$$\sum_{t=1}^{k}(x_i - \pi_i(z_t))z_t^j h(z_t|\mathbf{x}), \tag{2.2}$$

or,

$$\sum_{t=1}^{k}(x_i - \pi_i(z_t))z_t^j f(\mathbf{x}|z_t)h(z_t)/f(\mathbf{x}), \quad j = 0, 1, \quad i = 1, ..., p. \tag{2.3}$$

We arrange these elements (for each $\mathbf{x}$) in a vector $\mathbf{s}$.

Since we have a maximum likelihood estimator (giving maximum likelihood estimates (MLE)), which is a special case of an M-estimator, the vector influence function of the a_{ji} can be obtained from

$$\text{IF}[a_{ij};\mathbf{x}] = \mathcal{I}^{-1}\mathbf{s} \tag{2.4}$$

(Hampel et al. (1986), also Section 1.6.6) where $\mathcal{I}$ is the information matrix, given by

$$\mathcal{I} = \sum_{\text{all}\,\mathbf{x}} \frac{1}{f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial a_{ji}} \frac{\partial f(\mathbf{x})}{\partial a_{hm}} \tag{2.5}$$

where

$$\frac{\partial f(\mathbf{x})}{\partial a_{ji}} = \sum_{t=1}^{k} (x_i - \pi_i(z_t)) z_t^j f(\mathbf{x}|z_t) h(z_t) \tag{2.6}$$

In the equations above, $j = 0, 1$, $h = 0, 1$, $i = 1, ..., p$, $m = 1, ..., p$ and $\mathbf{x}$ a response pattern taking 2^p different values.

If p is not too large, it is easy to calculate the influence function from (2.4) for any given response $\mathbf{x}$.

From Equations (2.4), (2.5), (2.6) we see that the influence function of a_{0i} is a sum of product of probabilities and the influence function of a_{1i} is a sum of product of probabilities times z, which is a finite number. Thus, the influence function is flnite, in which case we say that the maximum likelihood estimator for this model is B-robust. This is expected since we have a distribution over a finite number of response patterns. But, although the rates of change of the parameters cannot get infinite, the rates or the actual values of the parameters may still get very large, changing the interpretation of the model and possibly the results of the analysis. Therefore, we would still like to study the behaviour of the estimates for some datasets, as the frequency distribution over the response patterns slightly changes.

2.3 Contamination on a Point: Some Influence Function Results

To show the use of the influence function and results that can be deduced from it we will apply it firstly to the first four items of Schuessler 's Social Life Feeling scale 7 and then to all 6 items of the same scale. Scale 7 is labelled 'People's Cynicism' and is applied to the American sample only. The questions are given in the Appendix.

We describe in detail the influence function (IF) for the data for the first 4 items of Schuessler's scale 7. We chose to work only with 4 items in order to have a small number of possible response patterns and thus to be able to observe the behaviour of the IF at each one of them. For the 6 item set we will calculate the influence function in order to see how the number of the items may affect the sensitivity of the parameters.

It has not been possible to estimate the IF for a larger number of items because the probabilities of some response patterns are very small which causes numerical problems.

2.3.1 Schuessler Social Life Feeling scale 7, first 4 items

We calculated the influence function for a model with parameter values set to MLE for a logit/probit model fitted to the first four items of Scale 7. The parameter values and their standard errors are given in Table 2.1 and the frequency distribution over the 16 response pattern is given in Table 2.2.

Since the IF is evaluated at a model, and measures the rates of change of the parameters in a neighbourhood of the model, we shall assume that these estimates are the true model parameters. (We never know the true model parameters. Our sample may have come from a family of models, with different parameters. Therefore, it is only important to know that we are 'close' to the 'true' model. We do not know when we obtain a sample whether it contains contaminated data or not, but at least when it contains a small proportion of contaminated data, we would like to obtain estimates of the model parameters that are 'close' to the estimates we would have obtained had our sample been 'clean').

Table 2.1: Scale 7, first 4 items, parameter values for model

i	a_{0i}	s.e.	a_{1i}	s.e.
1	2.19	0.16	1.38	0.19
2	-0.03	0.06	0.66	0.10
3	0.87	0.13	2.06	0.35
4	1.67	0.11	1.16	0.15

Values of the influence function for the different parameters at every response pattern are shown in Figures 2.1 and 2.2. These are the rates of change of the parameters as an infinitesimal amount of probability is placed on each response pattern in turn. For example, in Figure 2.1 the rate of change of the difficulty parameter of item 1, a_{01}, as response pattern 0000 carries extra probability, is around 2.

Rates of change for the difficulty parameter From Figure 2.1 we see that the influence function values range from around -30 to 6. The largest fluctuations are observed

Figure 2.1: Influence function for a_{0i}, $i = 1, ..., 4$, at all response patterns for scale 7

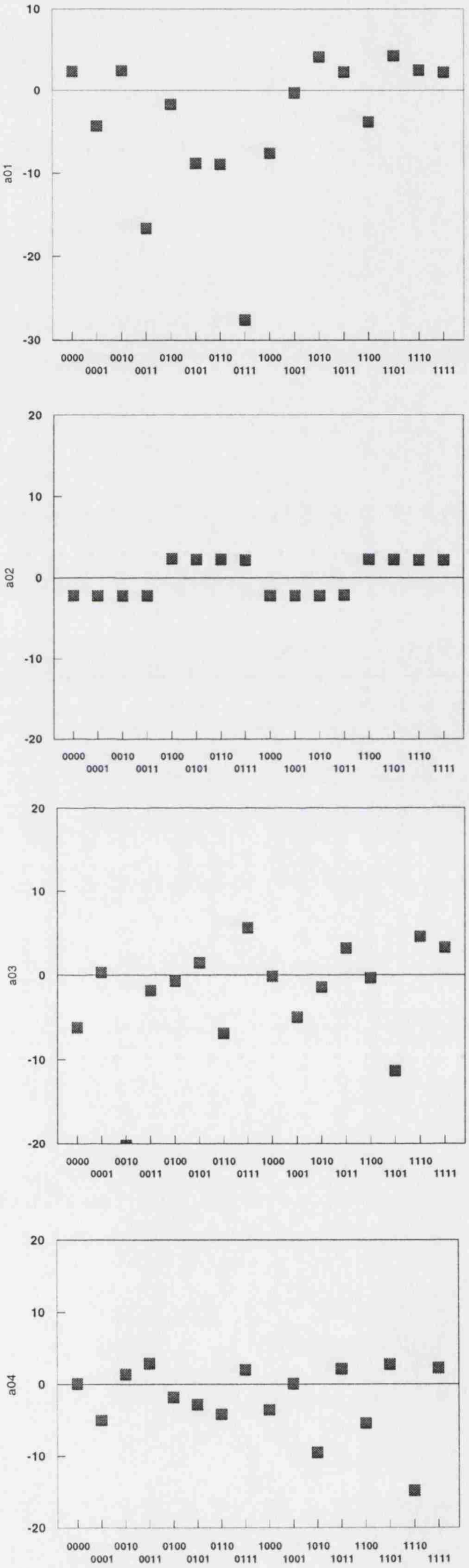


Figure 2.2: Influence function for a_{1i}, $i = 1, ..., 4$, at all response patterns for scale 7

Please note different range of y-scale for a_{13}

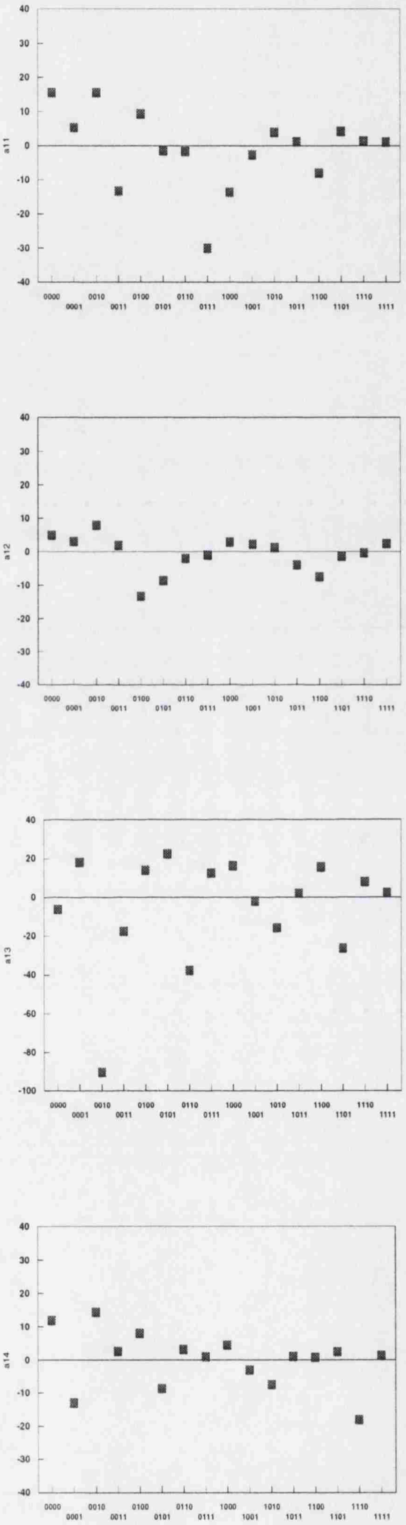


Table 2.2: Scale 7, first 4 items, Influence Function

	resp.	obs	a_{01}	a_{02}	a_{03}	a_{04}	a_{11}	a_{12}	a_{13}	a_{14}
1	0000	44	2.3	-2.2	-6.2	0.1	**15.5**	5.0	-6.0	11.9
2	0001	59	-4.2	-2.2	0.3	-4.9	5.4	3.1	18.3	-13.0
3	0010	13	2.4	-2.2	**-20.2**	1.4	15.5	**8.1**	**-90.2**	**14.3**
4	0011	31	-16.6	-2.2	-1.8	**2.9**	-13.4	2.0	-17.5	2.7
5	0100	17	-1.6	**2.4**	-0.7	-1.7	9.4	**-13.4**	14.1	8.1
6	0101	32	-8.8	2.3	1.5	-2.8	-1.5	-8.6	**22.6**	-8.7
7	0110	8	-8.9	2.3	-6.9	-4.2	-1.6	-1.9	-37.6	3.3
8	0111	23	**-27.6**	2.2	**5.6**	2.0	**-30.1**	-1.0	12.5	1.0
9	1000	80	-7.5	-2.2	-0.1	-3.5	-13.8	2.9	16.4	4.6
10	1001	145	-0.2	-2.2	-4.9	0.1	-2.6	2.2	-1.9	-2.9
11	1010	50	4.1	-2.2	-1.4	-9.5	4.0	1.2	-15.9	-7.5
12	1011	296	2.3	-2.1	3.2	2.1	1.2	-4.0	2.3	1.1
13	1100	39	-3.8	2.3	-0.3	-5.4	-8.0	-7.5	15.7	0.8
14	1101	113	**4.3**	2.2	-11.3	2.8	4.2	-1.4	-26.2	2.5
15	1110	43	2.5	2.2	4.6	**-14.8**	1.5	-0.4	8.1	**-18.0**
16	1111	423	2.3	2.2	3.3	2.2	1.2	2.4	2.4	1.4
min			-27.6	-2.2	-20.2	-14.8	-30.1	-13.4	-90.2	-18.0
25%			-7.8	-2.2	-5.3	-4.4	-3.9	-2.4	-16.3	-4.0
median			-0.9	0.0	-0.5	-0.8	1.2	0.4	2.4	1.3
75%			2.4	2.2	1.9	2.0	4.5	2.5	14.5	3.6
max			4.3	2.4	5.6	2.9	15.5	8.1	22.6	14.3

for the difficulty parameters of items 1 and 3, whereas the one for item 2 is very well behaved.

For the difficulty parameter a_{0i}, one would intuitively expect a negative change when $x_i = 0$ and a positive change when $x_i = 1$. As extra probability is placed on a response pattern with $x_i = 1$, the probability of a positive response for this item rises and thus a_{0i} must become larger. This behaviour is observed for the difficulty parameter of item 2, but the difficulty parameters of the rest of the items change sometimes in an unexpected way. For example, we would expect the rates of change of a_{01} to be negative at the first 8 response patterns and positive at the last 8 response patterns, but this is not so. Also, a_{03} has a fairly large negative rate of change at response pattern 0010, which we would expect to be positive.

Rates of change for the discrimination parameter The rates of change for the discrimination parameters (Figure 2.2) are generally larger than the ones for the difficulty parameter and a rate of change up to -90 is observed for a_{13}.

This suggests that parameter values (or MLE) for the discrimination parameter may change a lot for small changes in the data.

The largest negative rates of change are usually observed for the discrimination parameter of item i when extra probability is placed on a response pattern where the response for item i is different from the responses to the other items. The largest negative rates of change for a_{11} occurs at response pattern 0111, for a_{12} at 0100, for a_{13} at 0010 and for a_{14} at 1110. Since a_{1i} is a measure of the association of the item with the other items and with the latent variable, the negative rates of change at these response patterns mean that as the latter carry extra probability, the model will try to reduce the association of item i with the rest of the items and with the latent variable.

The largest positive rates of change (22 and 16) are observed for the discrimination parameter of item 3, at response patterns 0101 and 1000. The largest positive rate for a_{11}, for a_{12} and for a_{14} is observed at 0010. It seems that the discrimination parameter for item i increases, when extra probability is placed on response patterns where the response to item i is the same as to most of the items.

Large positive rates of change in the discrimination parameter may imply that the response function of an item may become easily a threshold function, by increasing or decreasing the probability of a response pattern, we cannot say though from the values of the rates of change whether this would be the case or not.

A very large negative change on the other hand may downweight unduly the item in the calculation of the posterior means.

2.3.2 Schuessler Social Life Feeling scale 7, all 6 items

It is interesting to see what the IF looks like when we have all 6 items of the scale, since large rates of change when there are only 4 items could be due to the small number of items.

The estimated parameters of Scale 7 with all 6 items are given in Table 2.3.

Table 2.3: Scale 7, 6 items, parameter values for model

i	a_{0i}	s.e.	a_{1i}	s.e.
1	2.19	0.14	1.38	0.15
2	-0.03	0.06	0.77	0.09
3	0.83	0.10	1.90	0.19
4	1.73	0.11	1.27	0.14
5	-0.16	0.08	1.50	0.15
6	1.81	0.12	1.46	0.15

Table 2.4 shows the IF for a_{0i} and a_{1i}. In the bottom of the Table some descriptive statistics for the IF are given. The minimum and maximum values of the IF for each parameter are shown in bold so that the response patterns for which they occur can be identified (except for a_{02} as its minimum and maximum values occur for many response patterns).

The range of the IF is generally smaller than the IF when there were only four items, both for a_{0i} and a_{1i}. The IF now ranges from -21 to 5 approximately for a_{0i} and from -47 to 14 for a_{1i}.

The pattern of the changes is consistent with what we had observed in the four item dataset. The IF of a_{0i} is generally negative if the frequency of a response pattern with item $i = 0$ increases and generally positive otherwise. The IF of a_{1i} is generally positive if the frequency of a response pattern increases with response to item i same as to most of the other items and negative otherwise. For example, increasing the frequency of 00100 will decrease a_{13} and increase all other a_{1i}'s. (It is not so easy to predict the direction of change of a_{1i} as it is with a_{0i}, and even for a_{0i} the direction of change is sometimes not the expected one.)

We observe that the most extreme changes occur at 'extreme' response patterns, i.e. response patterns for which at least five out of the six have the same response, either 0 or 1. Particularly 000010 is extreme since there is a positive response to the most difficult item and 001000 since there is a positive response to the most discriminating item.

The magnitude of the frequency does play a role in the magnitude of the IF but a secondary one. The largest rates of change do not necessarily occur at the response patterns with the smallest frequencies but large frequencies may prevent 'extreme' response patterns exhibiting very large rates of change. For example, response patterns 101111, 111110, 111111 are extreme in the sense that five items have response 1, but perhaps because of their large frequencies (169, 135 and 169 respectively) the IF of all parameters is quite small.

2.4 Actual rates of change of parameters and parameter estimates with contamination in the data

To evaluate the magnitude of the rates of change of the parameters given by the influence function, we calculated empirical rates of change which would have a similar interpretation as the influence function. Assuming that our original dataset is a 'clean' dataset generated from the model, the empirical rates of change show how fast and in what direction the parameters change when estimated from a dataset which contains some contamination at a particular response pattern. Thus the empirical rates of change are the sample analogue of the IF and will partly serve to verify our IF results. The main advantage of doing a sample sensitivity analysis though is to understand and get a feeling of the magnitude of the rates of change by comparing the actual parameter estimates obtained from different samples with the original ones.

We used the first four items of Schuessler's scale 7. The contaminated datasets were constructed by decreasing the observed frequencies of all response patterns by a percentage, in the following example by 3%, and adding the same percentage (3%) of the total frequency to a particular response pattern. This was repeated for all response patterns. We thus constructed 16 contaminated datasets, from which parameter estimates were

Table 2.4: Scale 7, all 6 items, Influence Function

	resp.p.	obs	a_{01}	a_{02}	a_{03}	a_{04}	a_{05}	a_{06}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}
1	000000	30	0.7	-2.3	-3.7	-0.8	**-3.0**	-0.8	**13.1**	4.2	3.9	**10.2**	2.8	10.3
2	000001	13	-1.4	-2.3	-2.4	-1.7	-2.9	-7.6	9.7	3.6	7.2	8.1	3.7	-16.8
3	000010	1	-0.7	-2.2	-0.2	-1.0	**5.2**	-0.5	10.7	**4.5**	**14.1**	9.3	**-39.4**	10.0
4	000011	0	-5.5	-2.2	-0.9	-3.6	4.4	-1.6	3.5	3.3	11.5	4.5	-24.9	-6.0
5	000100	20	-1.1	-2.3	-2.6	-5.5	-2.9	-1.3	10.1	3.6	6.5	-13.5	3.5	8.8
6	000101	31	-5.3	-2.3	-2.5	-1.7	-2.8	-3.0	3.7	2.8	6.4	-6.4	3.7	-8.6
7	000110	3	-4.8	-2.2	-0.8	-1.1	4.5	-3.1	4.5	3.5	12.0	-5.2	-26.7	5.4
8	000111	5	-11.3	-2.2	-3.1	1.0	3.8	0.9	-5.2	2.0	3.9	-1.1	-14.0	-1.4
9	001000	10	-0.9	-2.2	-8.9	-1.0	-3.0	-0.2	10.3	4.3	**-47.4**	9.2	**9.2**	**10.5**
10	001001	1	-6.8	-2.2	-3.2	-4.3	-2.9	-0.2	1.6	2.9	-24.0	3.2	6.6	-3.5
11	001010	1	-6.3	-2.2	-1.6	-3.8	4.1	-3.8	2.3	3.5	-18.3	4.0	-17.8	4.2
12	001011	1	-14.2	-2.2	1.5	-8.4	3.5	2.2	-9.6	1.3	-4.2	-4.6	-8.4	0.9
13	001100	6	-5.9	-2.2	-3.9	-0.1	-2.9	-3.8	2.8	3.1	-26.7	-3.3	7.0	4.3
14	001101	15	-13.3	-2.2	0.1	1.3	-2.8	1.4	-8.4	1.2	-9.7	-0.5	2.8	-0.4
15	001110	3	-13.0	-2.2	1.3	1.8	3.5	-9.1	-7.9	1.6	-5.4	0.4	-9.4	-5.0
16	001111	7	-22.5	-2.2	2.8	2.2	3.0	2.6	-22.3	-1.3	2.1	1.4	-2.0	1.8
17	010000	4	-0.1	2.4	-2.7	-1.0	-2.8	-0.8	11.7	**-13.6**	6.1	9.4	3.9	9.8
18	010001	9	-3.5	2.4	-2.1	-2.7	-2.8	-4.5	6.5	-9.1	7.7	6.1	4.4	-11.2
19	010010	2	-2.9	2.4	-0.2	-2.1	4.8	-1.8	7.4	-8.2	13.8	7.1	-30.9	7.7
20	010011	2	-8.7	2.4	-1.9	-5.3	4.0	0.2	-1.4	-4.6	8.3	1.2	-17.6	-2.7
21	010100	4	-3.0	2.4	-2.2	-3.2	-2.8	-2.3	7.2	-9.6	7.4	-9.2	4.3	6.9
22	010101	22	-8.4	2.4	-3.0	-0.3	-2.8	-1.0	-0.8	-5.8	4.6	-3.7	3.8	-4.8
23	010110	0	-7.9	2.4	-1.6	0.2	4.1	-5.3	-0.1	-5.1	9.4	-2.6	-19.2	1.6
24	010111	6	-15.3	2.3	-5.0	1.8	3.4	2.0	-11.3	-2.0	-2.8	0.4	-7.7	0.7
25	011000	1	-3.7	2.4	-5.4	-2.5	-2.9	-2.1	6.2	-7.1	-33.6	6.4	8.5	7.2
26	011001	4	-10.5	2.4	-0.8	-6.3	-2.8	1.0	-4.1	-3.9	-13.9	-0.7	5.0	-1.2
27	011010	1	-10.1	2.4	0.5	-6.0	3.7	-6.7	-3.5	-3.3	-9.0	-0.1	-12.0	-0.9
28	011011	2	-18.9	2.3	2.6	-11.2	3.1	2.6	-16.8	-0.8	1.2	-9.8	-3.8	1.9
29	011100	1	-9.5	2.4	-1.3	0.9	-2.9	-6.5	-2.6	-4.3	-15.9	-1.4	5.6	-0.4
30	011101	7	-17.9	2.3	1.7	1.8	-2.7	2.1	-15.2	-1.5	-2.8	0.6	-0.1	0.9
31	011110	5	-17.6	2.3	2.5	2.2	3.2	-12.8	-14.8	-1.1	0.7	1.3	-4.6	-11.6
32	011111	10	**-28.3**	2.2	3.3	2.4	2.7	2.8	**-31.1**	1.1	4.4	1.7	1.1	2.2
33	100000	37	-7.2	-2.3	-2.7	-1.7	-2.9	-1.5	-13.3	3.5	6.2	8.1	3.3	8.5
34	100001	28	-2.1	-2.3	-2.7	-4.0	-2.8	-2.8	-5.5	2.7	5.7	3.7	3.5	-8.1
35	100010	6	-1.4	-2.2	-1.1	-3.5	4.4	-3.5	-4.3	3.4	11.2	4.6	-25.8	4.7
36	100011	9	1.1	-2.2	-3.5	-7.2	3.7	1.0	-0.5	1.8	2.5	-2.3	-13.3	-1.3
37	100100	36	-2.7	-2.3	-2.7	-2.0	-2.8	-3.8	-6.4	2.8	5.9	-6.8	3.5	4.3
38	100101	70	0.3	-2.2	-4.2	0.3	-2.8	0.0	-1.8	1.5	0.1	-2.4	2.4	-3.0
39	100110	13	0.9	-2.2	-3.0	0.8	3.8	-7.7	-0.8	2.0	4.2	-1.4	-14.7	-2.5
40	100111	26	2.3	-2.2	-7.3	2.0	3.2	2.4	1.2	-0.1	-12.2	0.9	-4.0	1.3
41	101000	11	-0.1	-2.2	-3.6	-4.1	-2.9	-4.2	-2.4	2.9	-25.5	3.4	6.6	3.5
42	101001	14	1.5	-2.2	0.3	-8.4	-2.7	1.4	0.1	1.0	-9.1	-4.7	2.3	-0.4
43	101010	5	2.1	-2.2	1.4	-8.2	3.5	-9.7	1.0	1.4	-4.9	-4.2	-8.8	-6.2

Table 2.4 continued: Scale 7, all 6 items, Influence Function

	resp.p.	obs	a_{01}	a_{02}	a_{03}	a_{04}	a_{05}	a_{06}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}
44	101011	20	2.4	-2.2	2.9	-13.9	3.0	2.5	1.4	-1.6	1.9	-15.0	-1.6	1.6
45	101100	30	1.4	-2.2	-0.1	1.2	-2.8	-9.2	-0.1	1.3	-10.6	-0.7	3.1	-5.3
46	101101	77	2.1	-2.2	2.3	1.9	-2.5	2.1	0.9	-1.2	-0.8	0.7	-4.2	0.9
47	101110	20	2.5	-2.2	2.9	2.2	3.0	-16.5	1.6	-1.1	2.0	1.3	-2.2	-18.2
48	101111	169	2.3	-2.2	3.1	2.3	2.9	2.5	1.3	-5.3	2.6	1.4	2.1	1.4
49	110000	6	-4.1	2.4	-2.3	-2.7	-2.8	-2.6	-8.6	-9.3	7.0	6.2	4.1	6.4
50	110001	19	-0.4	2.4	-3.2	-5.6	-2.8	-0.9	-2.9	-5.5	3.6	0.7	3.5	-4.5
51	110010	2	0.3	2.4	-1.9	-5.2	4.0	-5.8	-1.8	-4.8	8.3	1.4	-18.4	0.8
52	110011	12	2.0	2.3	-5.5	-9.5	3.3	2.0	0.9	-1.8	-4.7	-6.6	-7.0	0.7
53	110100	8	-0.8	2.4	-3.0	-0.5	-2.8	-5.8	-3.5	-6.0	4.4	-4.0	3.7	0.8
54	110101	68	1.3	2.3	-5.6	1.2	-2.8	1.2	-0.2	-2.6	-5.4	-0.6	1.4	-0.7
55	110110	4	1.9	2.3	-4.8	1.7	3.4	-10.7	0.7	-2.1	-2.1	0.2	-8.2	-7.8
56	110111	33	**2.9**	2.3	**-10.4**	**2.6**	2.8	**3.1**	2.1	0.6	-24.0	2.0	1.2	2.6
57	111000	4	1.1	2.4	-1.1	-6.2	-2.8	-7.0	-0.6	-4.1	-15.0	-0.4	5.1	-1.4
58	111001	21	2.0	2.3	1.8	-11.1	-2.7	2.0	0.9	-1.3	-2.5	-9.7	-0.8	0.8
59	111010	5	2.6	2.3	2.6	-10.9	3.1	-13.5	1.7	-0.9	0.8	-9.3	-4.2	-12.9
60	111011	13	2.6	2.3	3.3	**-17.4**	2.8	2.7	1.6	1.2	4.0	-21.8	1.4	1.9
61	111100	5	2.0	2.3	1.6	1.8	-2.7	-12.8	0.9	-1.6	-3.3	0.5	0.3	-11.6
62	111101	135	2.3	2.3	3.2	2.2	-2.3	2.5	1.3	0.9	2.9	1.3	-9.8	1.4
63	111110	13	2.7	2.3	**3.4**	2.4	2.8	**-21.2**	1.9	1.0	4.7	1.7	1.1	**-26.7**
64	111111	270	2.4	2.3	3.2	2.4	2.9	2.5	1.3	2.7	2.2	1.5	3.3	1.5
min			-28.3	-2.3	-10.4	-17.4	-3.0	-21.2	-31.1	-13.6	-47.4	-21.8	-39.4	-26.7
25%			-7.4	-2.2	-3.0	-5.3	-2.8	-5.4	-3.6	-3.5	-5.4	-3.4	-8.5	-3.8
median			-1.2	0.0	-1.6	-1.4	0.2	-1.4	0.4	0.3	2.0	0.4	1.1	0.8
75%			1.6	2.3	1.6	1.4	3.4	1.6	1.9	2.7	5.9	2.3	3.6	2.8
max			2.9	2.4	3.4	2.6	5.2	3.1	13.1	4.5	14.1	10.2	9.2	10.5

obtained. The empirical rate of change of a parameter is the difference of the 'original' estimate from the new one, divided by the amount of contamination (0.03).

In Figures 2.3 and 2.4 (left column) we see the empirical rates of change for the difficulty and discrimination parameters of the items, for the first four items of scale 7, as each response pattern carries 3% of the total extra frequency.

Difficulty parameters The actual rates of change of the parameters at each response pattern can be compared with the IF, which represents the 'theoretical' rates of change.

We observe that the pattern of the influence function and the pattern of the actual rates of change across the response patterns are very similar (see Figures 2.1 and 2.3). Also the range of the values of the two are approximately the same, though the IF is for some cases larger than the actual rate of change, e.g. for a_{01} at 0111 and for a_{03} at 0010. Thus the 3% contamination, although it disturbs the frequency of the response patterns with small initial frequency quite a lot, it can still be considered small for our purposes and therefore, by looking at the magnitude of the actual values obtained from fitting the model with 3% contamination we can get some idea of the 'extremeness' of the magnitude

Figure 2.3: Actual values and rates of change of a_{0i}, $i = 1, ..., 4$, as each response pattern carries 3% extra frequency, Scale 7, first 4 items

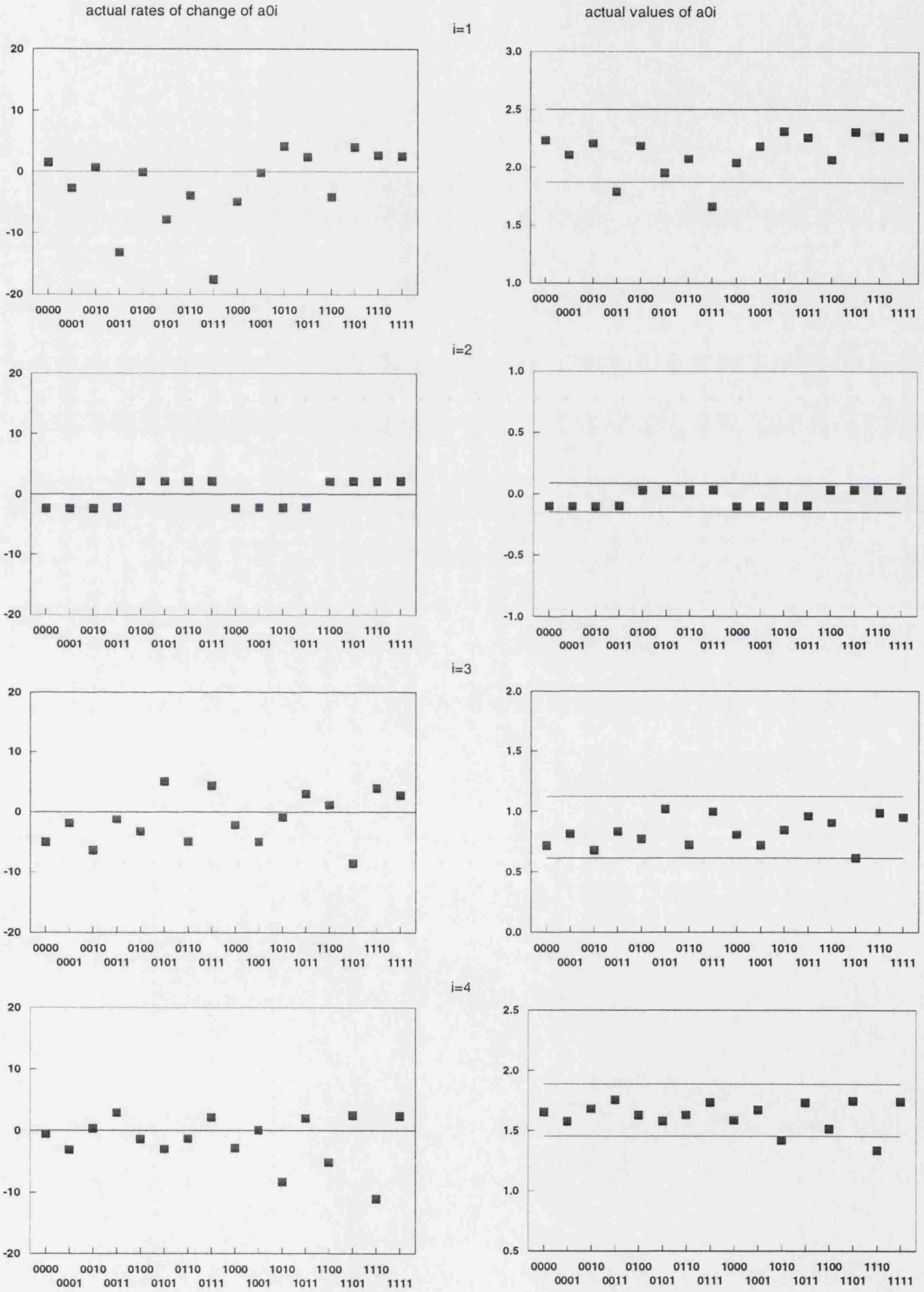
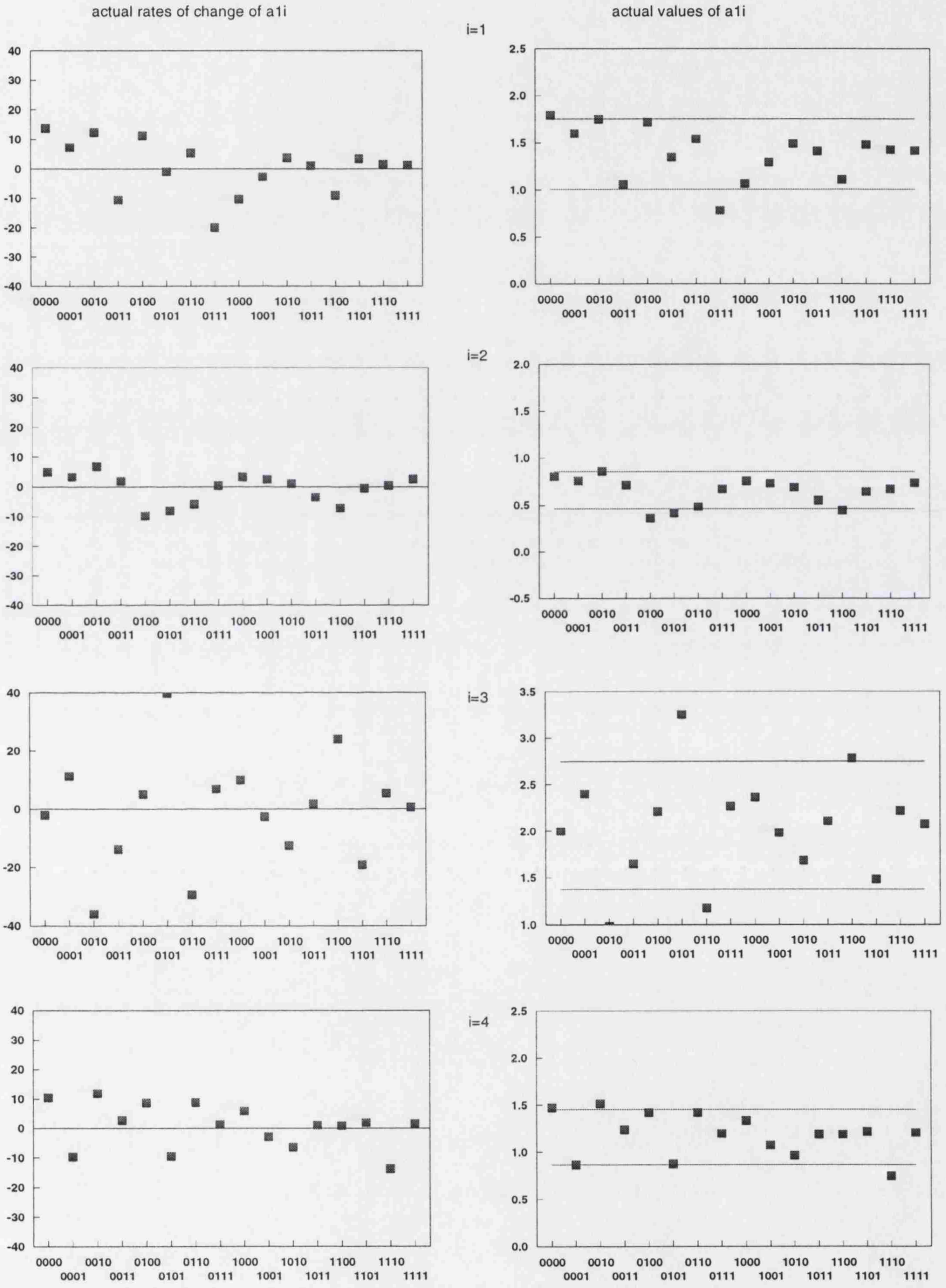


Figure 2.4: Actual values and rates of change of a_{1i}, $i = 1, ..., 4$, as each response pattern carries 3% extra frequency, Scale 7, first 4 items



of the IF or vice-versa. We will mainly judge the 'extremeness' of the parameter values by whether they fall in the confidence intervals of the 'original' parameter estimates. The latter are given in Table 2.5.

Table 2.5: Scale 7, first 4 items, 95% confidence interval of parameters

	a_{0i}		a_{1i}	
item	l.b.	u.b.	l.b.	u.b.
1	1.87	2.50	1.01	1.75
2	-0.15	0.09	0.46	0.86
3	0.61	1.12	1.37	2.75
4	1.45	1.89	0.87	1.45

The largest in absolute value rates of change of a_{01} correspond to values outside the confidence interval of the original estimates. These are at response patterns 0011 and 0111. 3% extra frequency at these response patterns will make item 1 a lot easier than would be expected from random variability.

Also, the two largest rates of change of a_{04}, which occur at response patterns 1010 and 1110, correspond to parameter values outside the confidence interval of the initial parameter values. For a_{02} and a_{03}, all parameter values are within the confidence interval of the initial parameter estimates.

We conclude that 3% contamination at a particular response pattern, can sometimes result to more extreme parameter values than the values that would be expected from random variability.

Discrimination parameters As with a_{0i}, we can compare the actual rates of change of the parameters under 3% contamination, with the influence function. The pattern of the two across the response patterns which carry the extra frequency is very similar (Figures 2.2 and 2.4). The range of the actual rates of change is smaller though, i.e. the largest IF values correspond to the largest actual rates of change but the latter are not as large (for example, actual rate of change of a_{13} at response pattern 0010 is around -20, not -90). The largest influence function values though do point to parameters that fall outside the 95% confidence interval of the 'original' parameter estimates, if the corresponding response patterns carry 3% extra frequency. The response patterns for which this occurs are: for a_{11} at 0111, for a_{12} at 0100 and 0101, for a_{13} at 0010, 0101, 0110 and 1100, for a_{14} at 0010 and 1110. We note that for 0010 a_{11}, a_{12} and a_{14} are on the top bound of the confidence intervals and a_{13} below the low bound. All the above response patterns have small frequencies, and as we also saw with a_{0i} large rates of change of a_{1i} are associated with response patterns with small frequencies, but this is not the only factor determining their size, since rates of change at response pattern 0010 are larger than the rates of change at response pattern 0110, which is the one with the smallest original frequency (8 observations).

What is perhaps more important in determining the size of the rates of change of the parameter is the initial size of the parameter. Here a_{13} was the largest discrimination parameter. It seems that large discrimination parameters may suddenly jump to a value that denotes a threshold response function and this is certainly undesirable.

2.5 Other types of contamination

The type of contamination considered so far was to add observations to a particular response pattern. In a real situation this could occur if a group of people were strongly influenced to respond in a particular way to all questions irrespectively of their latent trait modelled, for example, in an ability test, a group of students sitting close together may cooperate and give the same answers. This could also be a result of a gross error, if the frequency of a response pattern is misrecorded.

Another situation that would contaminate the data would be the following: a proportion of people who would have answered in a particular way, for some reason they responded to an item the other way round, for example, some individuals, instead of answering 1000 answered 1100. This could happen if, again in an ability testing situation, some students who would not answer any of the last three items were 'whispered' the correct answer to the second item. This would have the effect of some amount of frequency to shift from response pattern 1000 to 1100.

Another type of disturbance in the data would occur if the probability for a positive response for an item were determined to some extent by some external factor, so that the probability of a positive response for that item were slightly increased or decreased for some individuals or all individuals irrespective of their ability. For example, an item could be particularly difficult for students with mother tongue other than English, and that would create response patterns with an 'unnecessary' zero response for that item.

The rates of change of the parameters for such situations can be calculated directly from the influence function.

2.5.1 Increasing the probability of an item

The effect on the parameters of increasing the probability of a positive response to an item can be measured by averaging the IF over the response patterns with '1' for the particular item. For example, the rate of change of a_{ji} if we increase the probability of positive response to item 1, is given by averaging the IF of a_{ji} at the last eight response patterns (the ones with 1 for item 1).

The left side of Figures 2.5 and 2.6 show the rates of change of each parameter plotted against the item which carries extra probability.

The right side of the same Figures show the actual rates of change of each parameter, calculated from data contaminated in the following way: 3% of the total frequency was spread equally to the eight response patterns with response 1 to the item which was supposed to carry the extra probability. The 'theoretical' rates of change (calculated from the IF) and the actual rates of change are very close, verifying our results.

Regarding the direction of the changes of the parameters, we note the following: Increasing the probability of item i generally causes the difficulty parameter a_{0i} to increase, thus making item i easier. This is true for parameters a_{01}, a_{02} and a_{04} but not for a_{03}, which shows a counter-intuitive behaviour. And increasing the probability of item i causes the discrimination parameter a_{1i} to decrease, making the item less discriminating.

As to the magnitude of the rates of change we observe that they are very small, generally a lot smaller than the rates of change of the parameters when extra probability was placed on one response pattern at a time. A large rate of change is only observed for a_{13} when the probability of item 3 is increased, but again this is not as large as the rates of change of a_{13} when the probability of some response patterns was increased individually.

We thus expect the parameters to change little and smoothly as we progressively increase the probability placed on an item. To verify that, we progressively increased the probability of each item by spreading equally 3%, 5% and 10% of the total frequency to the eight relevant response patterns. The new difficulty parameters are shown in Table 2.6. They are generally close to the original parameters and they are all within the confidence intervals of the original estimates with 5% contamination. a_{01} falls below the

Table 2.6: a_{0i} obtained from contaminated data (increasing the probability of each item in turn), with different ϵ, Scale 7, first 4 items

ϵ	0	0.03	0.05	0.10
i	increasing P(x_1=1)			
1	2.19	2.21	2.22	2.22
2	-0.03	-0.03	-0.03	-0.03
3	0.87	0.83	0.81	0.77
4	1.67	1.58	1.52	1.38
increasing P(x_2=1)				
1	2.19	2.06	1.96	1.76
2	-0.03	0.03	0.08	0.18
3	0.87	0.84	0.82	0.75
4	1.67	1.59	1.54	1.43
increasing P(x_3=1)				
1	2.19	2.05	1.95	1.74
2	-0.03	-0.03	-0.03	-0.03
3	0.87	0.83	0.84	0.86
4	1.67	1.61	1.56	1.45
increasing P(x_4=1)				
1	2.19	2.03	1.92	1.69
2	-0.03	-0.03	-0.03	-0.03
3	0.87	0.84	0.83	0.80
4	1.67	1.69	1.70	1.75

low confidence interval bound when the probability of items 2,3 and 4 increases by 10%. a_{04} falls below the low confidence interval bound when the probability of items 1 and 2 increase by 10% whereas a_{03} exceeds the confidence interval bound when the probability of item 2 increases by 10%.

The discrimination parameters obtained are shown in Table 2.7. They are very close to the original ones, and fall outside the confidence intervals of the original estimates in only two cases, a_{12} when the probability of item 2 increases by 10% and a_{13} when the probability of item 3 increases by 10%.

We also note that the parameters may be more affected by increasing the probability of another item rather than 'their' item. For example, a_{01} and a_{11} are more affected when the probability of items 2, 3 or 4, rather than the probability of item 1, is increased. Although it is not clear why this happens, it partly explains some counter-intuitive directions of change when extra probability is placed on a response pattern (e.g. positive change for a_{01} when extra probability is placed on 0000).

Figure 2.5: Average rates of change of a_{0i}, $i = 1, ..., 4$, as extra probability is placed on each item, Scale 7, first 4 items

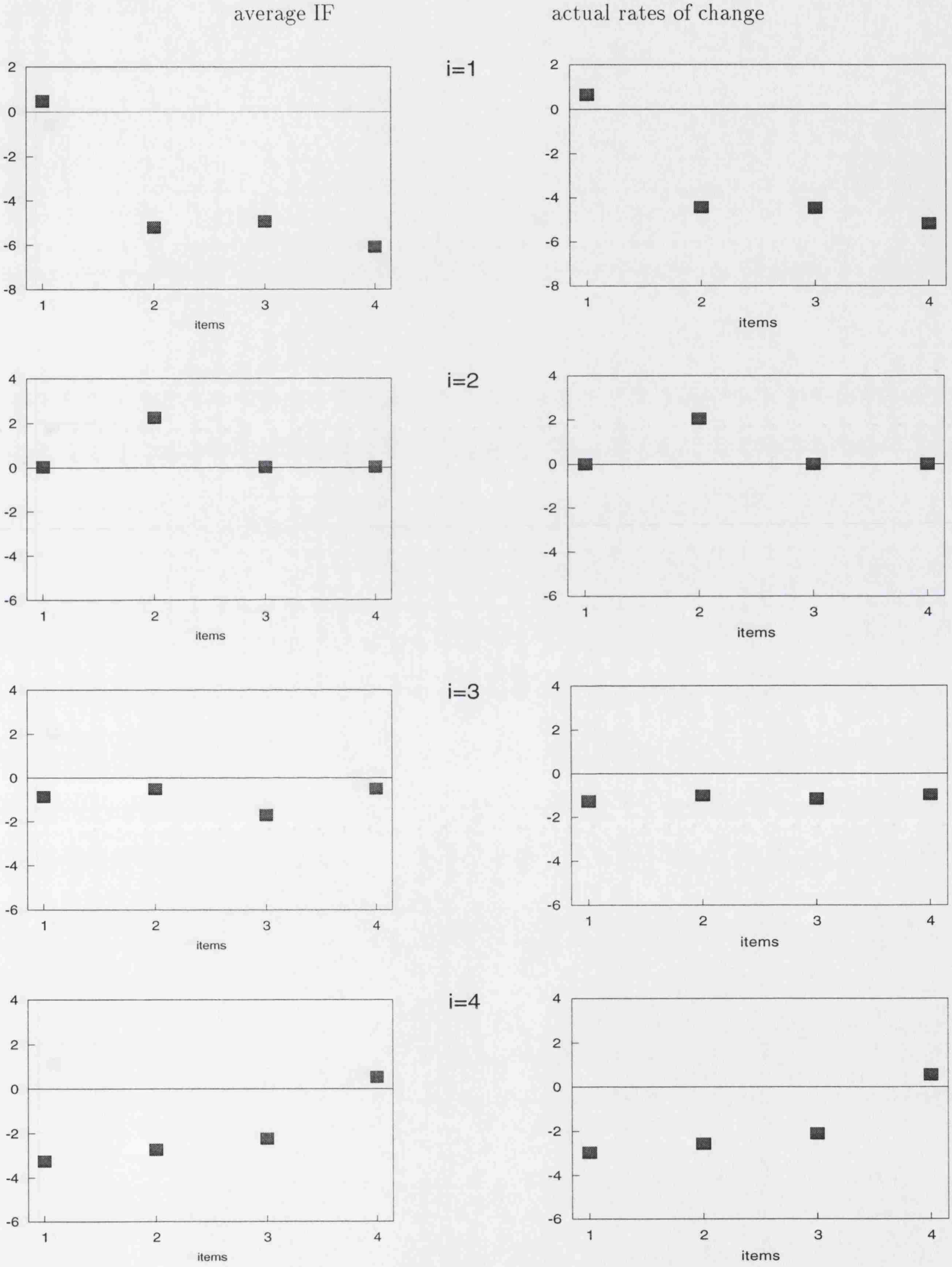


Figure 2.6: Average rates of change of a_{1i}, $i = 1, ..., 4$, as extra probability is placed on each item, Scale 7, first 4 items

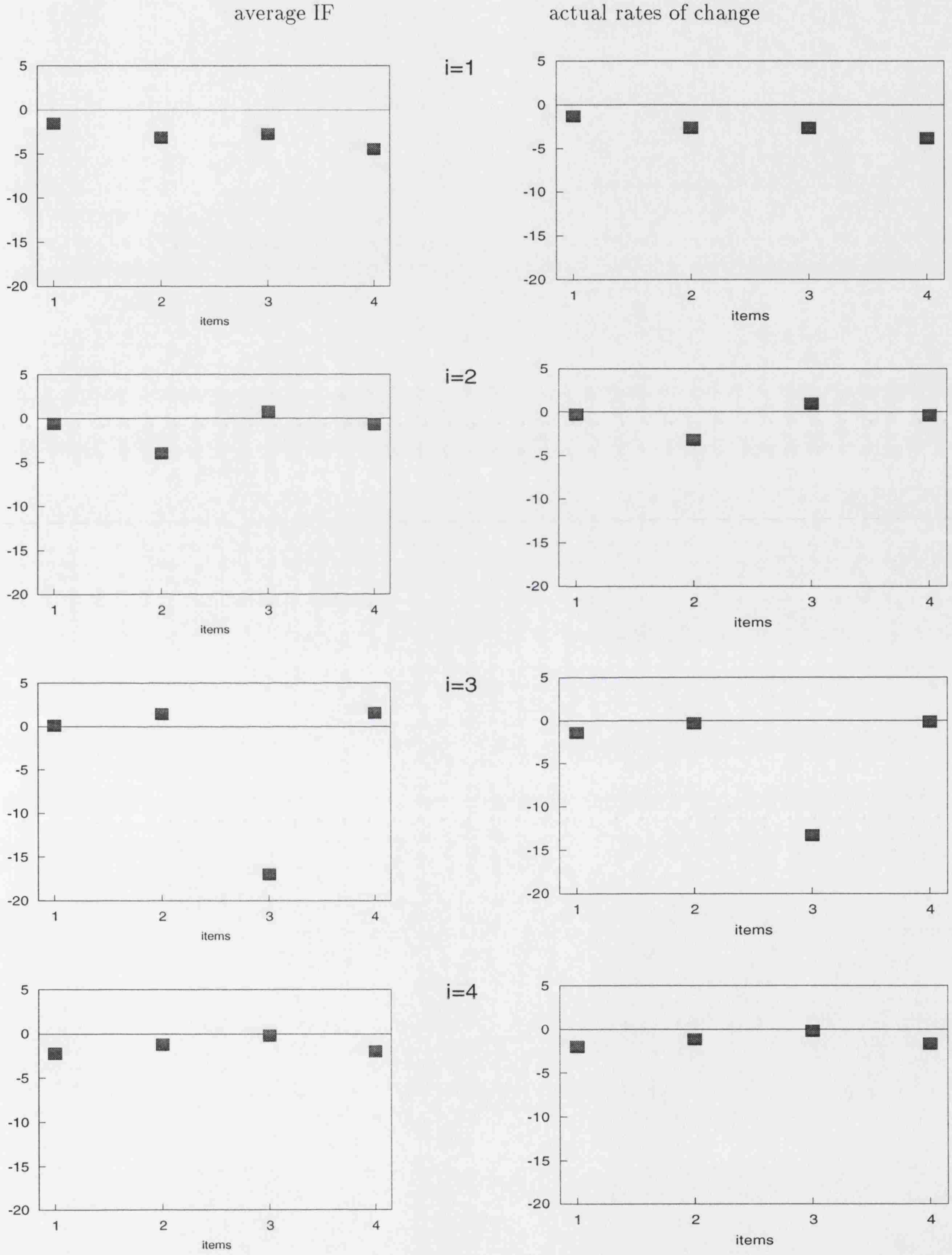


Table 2.7: a_{1i} obtained from contaminated data (increasing the probability of each item in turn), with different ϵ

ϵ	0	0.03	0.05	0.10
i	increasing P(x_1=1)			
1	1.38	1.34	1.31	1.22
2	0.66	0.65	0.64	0.61
3	2.06	2.02	2.01	2.00
4	1.16	1.10	1.06	0.97
increasing P(x_2=1)				
1	1.38	1.30	1.26	1.17
2	0.66	0.57	0.50	0.36
3	2.06	2.05	2.03	1.92
4	1.16	1.13	1.11	1.08
increasing P(x_3=1)				
1	1.38	1.30	1.25	1.13
2	0.66	0.69	0.70	0.71
3	2.06	1.66	1.46	1.14
4	1.16	1.16	1.15	1.12
increasing P(x_4=1)				
1	1.38	1.27	1.19	1.03
2	0.66	0.65	0.64	0.60
3	2.06	2.06	2.08	2.13
4	1.16	1.11	1.08	1.00

Table 2.8: Local shift sensitivities of the parameters of Scale 7, first 4 items

	a_{01}	a_{02}	a_{03}	a_{04}	a_{11}	a_{12}	a_{13}	a_{14}
	19.0	4.6	18.8	17.0	31.3	18.4	84.2	24.9
from	0011	0000	0010	1110	0111	0100	0010	0001
to	0010	0100	1010	1111	1111	0000	0000	0000

2.5.2 Shifting probabilities from one response pattern to another

Hampel et al. (1986) considered the effect of shifting an observation from one point x to a neighbouring point y, which they called 'wiggling' of observations. They said that intuitively the effect of such a shift of observations can be measured by IF$(y; T, F)$-IF$(x; T, F)$. The local shift sensitivity (Section 1.6.3) is a measure of the worst (approximate and standardised) effect of shifting observations from one point to another. When the data are metrical the difference of the influence functions is standardised by the difference of those points. In our case we cannot take a difference, but perhaps the analogue is to take the number of different items in these two response patterns. For example, moving from 0010 to 0101 for a_{13} the difference in the influence functions is 22.60-(-90.22)=112.82. Since 0010 and 0101 differ in three items, then dividing by three will give 37.61. Again for a_{13} shifting between 0010 and 0001 gives a standardised rate of change equal to 54.3 and between 0010 and 0000 a rate of change equal to 84.2. Since this is the maximum standardised difference this is the local shift sensitivity for a_{13}. The local shift sensitivities are the absolute values of the rates of change since symmetrical rates of change are predicted for shifts of observations between a pair of response patterns.

The local shift sensitivities for all the parameters are given in Table 2.8.

These are quite large, that is for some parameters are close or even above the absolute maximum value of their influence function, which is not surprising since they were constructed from the largest values of the influence function. Let us see how the empirical rates of change and the actual parameters behave as we move observations between these pairs of response patterns.

Table 2.9 shows the rates of change of the difficulty and discrimination parameters as we move 8 observations between the response patterns for which the largest rates of change will occur for one or more of the parameters.

We observe that in most cases the rates of change are symmetric for shifts of observa-

Table 2.9: Scale 7, first 4 items, empirical rates of change of parameters as we move observations from one response pattern to another

from	to	a_{01}	a_{02}	a_{03}	a_{04}	a_{11}	a_{12}	a_{13}	a_{14}
0011	0010	19.9	0.0	-13.8	-1.6	29.3	5.6	-55.4	11.4
0010	0011	-17.7	0.0	26.7	1.9	-27.6	-8.5	101.6	-10.9
0000	0100	-4.2	4.5	6.5	-2.4	-6.4	-17.9	23.6	-4.8
0100	0000	4.3	-4.8	-6.1	2.3	6.5	20.3	-22.3	4.7
0010	1010	2.7	0.0	28.2	-10.8	-10.1	-9.4	106.9	-22.3
1010	0100	-2.5	0.0	-14.2	11.4	10.2	6.8	-56.9	22.0
1110	1111	-0.9	0.0	-1.4	18.9	-1.4	3.4	-6.4	21.7
1111	1110	0.5	0.0	1.3	-16.3	0.8	-2.9	5.8	-18.7
0111	1111	35.1	0.0	-2.3	-0.3	36.6	3.9	-9.9	-0.7
1111	0111	-27.3	0.0	2.0	0.0	-28.9	-3.2	8.8	0.1
0010	0000	0.4	-0.0	20.6	-1.5	0.7	-4.2	111.3	-2.7
0000	0010	-1.1	-0.0	-10.3	1.1	-1.9	3.6	-67.7	2.0
0001	0000	5.8	0.0	-5.7	5.9	8.8	2.4	-21.0	26.3
0000	0000	-6.0	0.0	7.2	-4.4	-9.3	-2.8	26.5	-24.1

tions between a pair of response patterns and fairly well approximated by the standardised differences in the influence functions (see Tables 2.8 and 2.2). There are a few cases though, for example when observations are moved from 0010 to 1010, where the rates of change for a_{03} and a_{13} are much larger than the rates of change for the opposite shift of observations and much larger than what the influence functions would predict.

The fact that positive changes in the parameters may be larger than the negative ones for symmetric shifts in frequencies may be due to the frequency distribution of these data. For example, shifting frequencies from 0010 to 0000, to 0010 and to 1010 causes larger rates of change than shifting frequencies in the opposite direction, but 0010 had initial frequency of 13 whereas 0000, 0011 and 1010 had initial frequencies 44, 31 and 50 respectively.

In Table 2.10 the new estimates of a_{0i} are shown, as we move 8, 12 and 14 observations between the above response patterns. a_{02} and a_{04} remain within the confidence interval of the original estimates, but a_{01} exceeds the upper bound of its confidence interval if 12 observations are shifted from 0111 to 1111 and a_{03} which exceeds the upper bound of its confidence interval and becomes a lot easier if 12 observations (or 13) are moved from 0010 to 0011, from 0010 to 1010 or if 13 observations are moved from 0010 to 0000 (the parameter estimates outside the confidence intervals of the original estimates are shown

Table 2.10: Scale 7, first 4 items, a_{0i} as we move observations from one response pattern to another

		8	12	14	8	12	14	8	12	14	8	12	14
from	to	a_{01}			a_{02}			a_{03}			a_{04}		
0011	0010	2.30	2.36	2.39	-0.03	-0.03	-0.03	0.79	0.77	0.75	1.66	1.66	1.66
0010	0011	2.09	2.04	2.03	-0.03	-0.03	-0.03	1.02	**1.17**	**1.23**	1.68	1.69	1.69
0000	0100	2.17	2.15	2.15	-0.01	0.01	0.01	0.91	0.92	0.93	1.66	1.65	1.65
0100	0000	2.21	2.22	2.23	-0.06	-0.07	-0.08	0.83	0.82	0.81	1.68	1.69	1.69
0010	1010	2.20	2.21	2.21	-0.03	-0.03	-0.03	1.03	**1.19**	**1.26**	1.61	1.58	1.57
1010	0100	2.17	2.17	2.17	-0.03	-0.03	-0.03	0.79	0.76	0.75	1.73	1.77	1.78
1110	1111	2.18	2.18	2.18	-0.03	-0.03	-0.03	0.86	0.86	0.85	1.78	1.84	1.87
1111	1110	2.19	2.19	2.19	-0.03	-0.03	-0.03	0.88	0.88	0.88	1.58	1.54	1.52
0111	1111	2.38	**2.51**	**2.58**	-0.03	-0.03	-0.03	0.86	0.85	0.85	1.67	1.67	1.72
1111	0111	2.04	1.97	1.94	-0.03	-0.03	-0.03	0.88	0.89	0.89	1.67	1.67	1.67
0010	0000	2.19	2.19	2.19	-0.03	-0.03	-0.03	0.98	1.09	**1.14**	1.66	1.66	1.66
0000	0010	2.18	2.18	2.18	-0.03	-0.03	-0.03	0.81	0.79	0.78	1.67	1.68	1.68
0001	0000	2.22	2.24	2.24	-0.03	-0.03	-0.03	0.84	0.82	0.82	1.70	1.72	1.73
0000	0000	2.15	2.14	2.13	-0.03	-0.03	-0.03	0.91	0.93	0.95	1.65	1.64	1.63

in bold).

Table 2.11 shows the new estimates of a_{1i} for the same changes in the frequencies of the response patterns.

Except for the discrimination parameters of item 4 the rest of the discrimination parameters fall outside the confidence interval of the original estimates for some changes in the frequencies of the response patterns. These cases are the following (values are shown in bold): a_{11} exceeds the upper bound of its confidence interval if 12 or 14 observations are shifted from 0111 to 1111;

a_{12} just exceeds the upper bound of its confidence interval if 14 observations are moved from 0100 to 0000;

a_{13} though is more affected from these changes and exceeds the upper bound of the confidence interval of the original estimate if we move 12 (or 13) observations from 0010 to 0011, to 1010 or to 0000.

2.6 The breakdown point

The concept of the breakdown point was described in Section 1.6.4. It is the largest fraction of gross errors which 'never can carry the estimate over all bounds' (Hampel et al. (1986), page 97) or 'the distance from the model distribution the estimator still

Table 2.11: Scale 7, first 4 items, a_{1i} as we move observations from one response pattern to another

		8	12	14	8	12	14	8	12	14	8	12	14
from	to	a_{11}			a_{12}			a_{13}			a_{14}		
0011	0010	1.54	1.63	1.67	0.69	0.70	0.71	1.75	1.64	1.59	1.22	1.25	1.27
0010	0011	1.22	1.15	1.13	0.61	0.58	0.57	2.63	**3.17**	**3.37**	1.09	1.06	1.05
0000	0100	1.34	1.32	1.32	0.56	0.52	0.49	2.19	2.26	2.29	1.13	1.11	1.11
0100	0000	1.42	1.43	1.44	0.78	0.84	**0.87**	1.93	1.87	1.84	1.18	1.19	1.20
0010	1010	1.32	1.29	1.28	0.61	0.58	0.57	2.65	**3.25**	**3.48**	1.03	0.97	0.95
1010	0100	1.44	1.47	1.48	0.70	0.71	0.72	1.74	1.63	1.58	1.28	1.34	1.37
1110	1111	1.37	1.36	1.36	0.68	0.69	0.70	2.02	2.00	1.99	1.28	1.35	1.38
1111	1110	1.38	1.38	1.39	0.65	0.64	0.63	2.09	2.10	2.11	1.05	1.00	0.98
0111	1111	1.58	**1.71**	**1.78**	0.68	0.70	0.70	2.00	1.97	1.98	1.15	1.15	1.21
1111	0111	1.22	1.15	1.11	0.64	0.64	0.63	2.11	2.13	2.14	1.16	1.16	1.16
0010	0000	1.38	1.38	1.38	0.64	0.63	0.62	2.68	**3.22**	**3.41**	1.14	1.13	1.13
0000	0010	1.37	1.36	1.36	0.68	0.69	0.70	1.68	1.53	1.47	1.17	1.17	1.17
0001	0000	1.43	1.45	1.46	0.68	0.68	0.68	1.94	1.89	1.86	1.30	1.38	1.42
0000	0000	1.33	1.30	1.28	0.65	0.64	0.63	2.21	2.30	2.35	1.02	0.96	0.93

gives some relevant information'.

Before attempting to find the breakdown point of the maximum likelihood estimator for the logit/probit model, we must define what the above mentioned 'bounds' would be for that model. Very large discrimination parameters may be considered meaningless or invalidating the model, because the ability estimates coming from that model (the posterior means) give deceptive results: in their calculation the information coming from the item with the large discrimination parameter prevails over the information coming from the other items, and thus the posterior mean depends mostly on the response to that item and very little on the responses to the rest of the items.

Very large can be considered anything larger than 4, since for such values the response function becomes a threshold function, though the effect of a large discrimination parameter on the posterior analysis will depend on the size of the other discrimination parameters as well.

A very large difficulty parameter does not have such a disturbing effect. If the difficulty parameter is large then the posterior means of all individuals will be pushed upwards, but their order will not change.

It would be extremely difficult to specify mathematically the largest proportion of contamination that the estimator could endure before giving too large (or too small) estimates.

Therefore, we shall see empirically how the parameters for the dataset we considered above change, as we gradually increase the amount of contamination in the data. We are interested in large changes in the parameter estimates, therefore we will look at the types of contamination that caused sometimes severe changes in the response patterns, namely increasing gradually the frequency of a response pattern and shifting observations between response patterns.

2.6.1 Increasing the frequency of a response pattern

We will consider changes in the first 8 response patterns for the parameters of the first three items and changes in the last 8 response patterns for changes in the parameters of item 4, as for some of those response patterns the parameters had the largest rates of change.

Difficulty parameters Figures 2.7 show the difficulty parameters a_{0i} for the four items obtained as we gradually increased the percentage of extra frequency a response pattern received, against that percentage.

The overall picture of the four graphs suggests that generally the parameters change linearly as the amount of contamination increases and they fall within their original confidence intervals with 5% contamination. Let us see what happens at each parameter individually:

a_{01} changes smoothly with increased contamination on most of the response patterns considered, but exceeds the lower bound of the confidence interval with 3% extra observations at 0011 and 0111.

a_{02} changes proportionally to the percentage of contamination and just reaches the confidence interval bounds at 5% contamination for the response patterns considered.

a_{03} changes smoothly, proportionally to the amount of contamination and reaches the confidence interval bounds at 5% contamination for all response patterns considered, except for 0101, where there is a very large increase in the parameter if the percentage of extra observations placed on the response pattern increases from 3% to 5%.

a_{04} displays smooth changes and remains within the original confidence interval even with 5% contamination.

Figure 2.7: a_{0i} as extra frequency is given on a response pattern, against the percentage of extra frequency

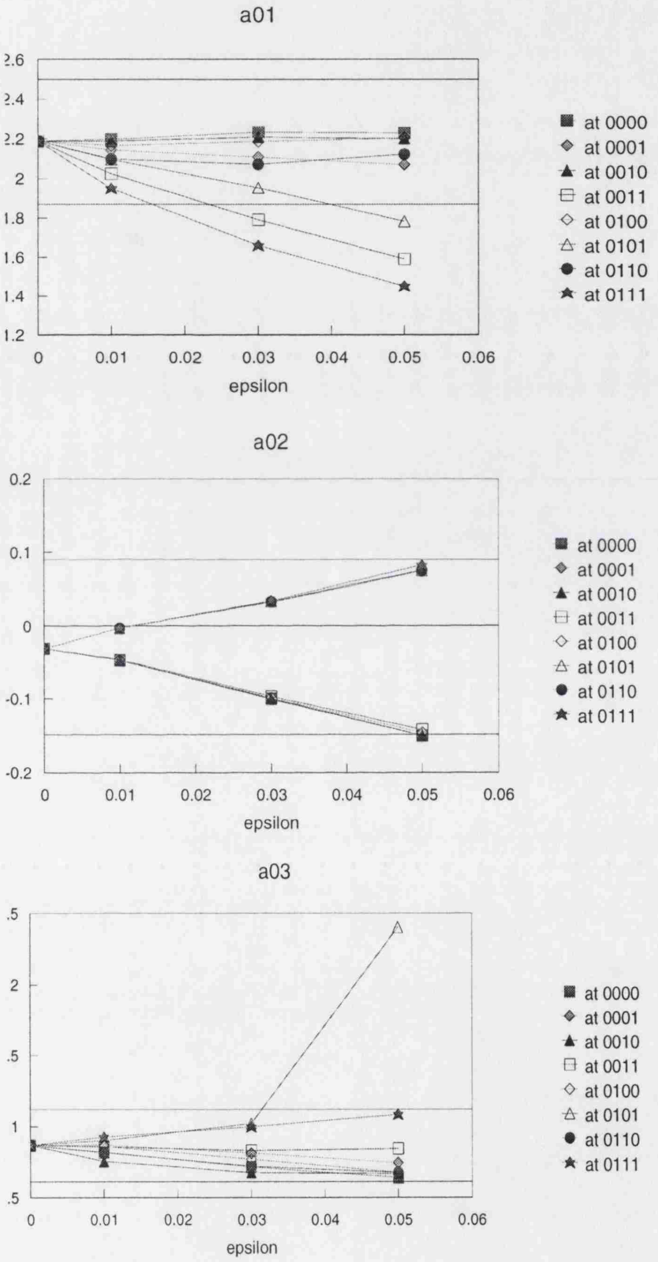
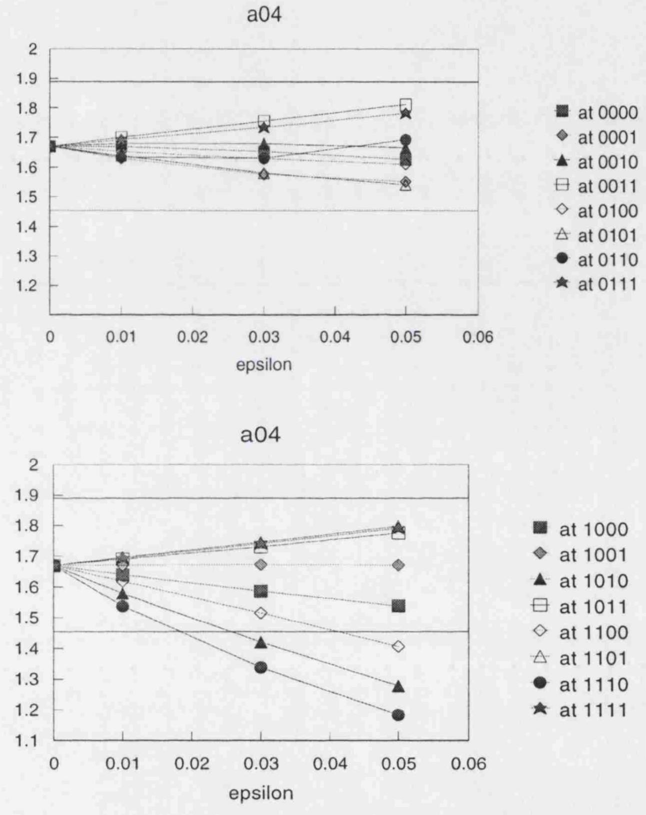


Figure 2.8: a_{04} as extra frequency is given on a response pattern, against the percentage of extra frequency



Discrimination parameters In Figures 2.9 the discrimination parameters are plotted against the percentage of extra frequency a particular response pattern receives. The 95% confidence interval bounds for the corresponding 'original' parameter are also plotted, as these may be used to assess the magnitude of the changes.

We observe that with 1% contamination at any of the response patterns considered, the new parameters lie in the confidence interval of the original parameters. With 3% contamination though, at some response patterns, the parameters reach or slightly exceed the confidence interval bounds. This happens for a_{11} at response patterns 0000 and 0111, for a_{12} at 0100 and 0101, for a_{13} at 0101, 0110 and 0010 and for a_{14} at 0010 and 0000.

With 5% contamination, the parameters exceed the confidence interval bounds more often, i.e. at more response patterns. We observe that for all parameters, at all of the response patterns considered, the change is generally linear, smooth. The parameters seem well behaved, as small distortions in the frequency distribution of the response patterns lead generally to small changes in the parameters, and the greater the distortion in the frequency distribution (in the sense defined so far) the greater the changes in the parameters.

The linearity of the curves for a_{11}, a_{12} and a_{14} shows that the influence function is a good approximation of the rate of change of these parameters at any level of contamination.

There is one exception to that: the change of a_{13} if we increase the extra frequency placed on response pattern 0101 from 3% to 5% is very abrupt and a_{13} becomes very large, in effect infinite. Thus placing 5% extra frequency to a response pattern can produce meaningless results, the estimator at that level of contamination, for the model considered here, breaks down.

2.6.2 Shifting observations between response patterns

We saw in Section 2.5.2 that even a very small number of observations, for example 12, can cause large changes to some parameters. We want to find the smallest number of observations for which large changes in the parameters occur. We will therefore consider changes in the frequencies of the response patterns as in Section 2.5.2, increasing the number of observations being shifted from 8 to 20. These numbers correspond to a proportion of 0.0056 and 0.014 respectively, so much smaller than the amounts considered

Figure 2.9: a_{1i} as extra frequency is given on a response pattern, against the percentage of extra frequency

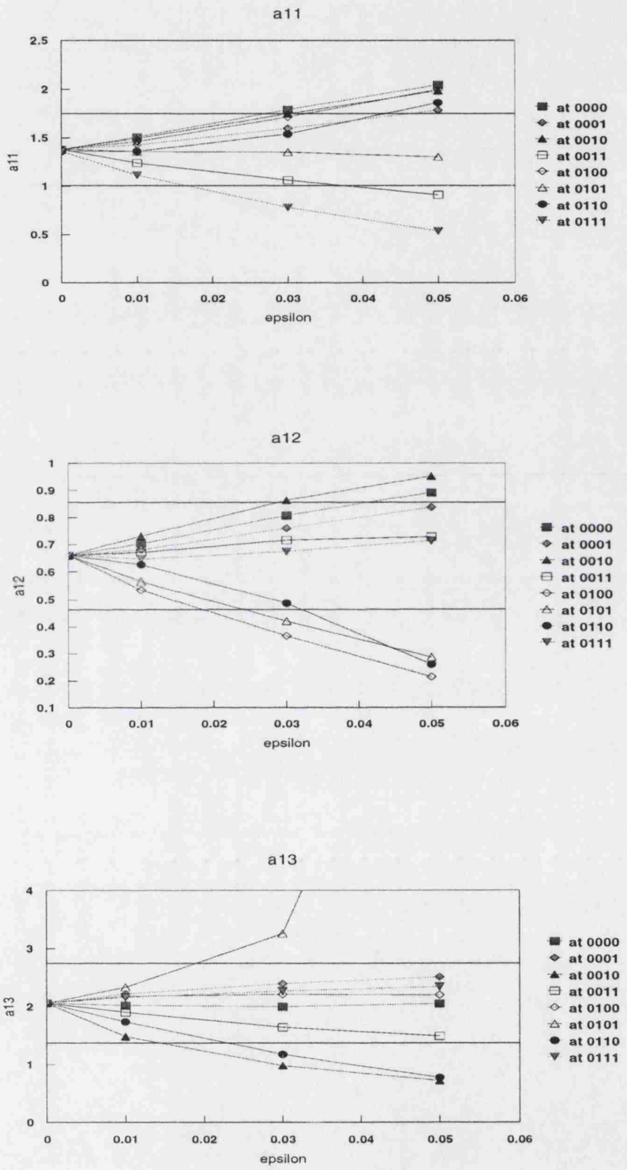
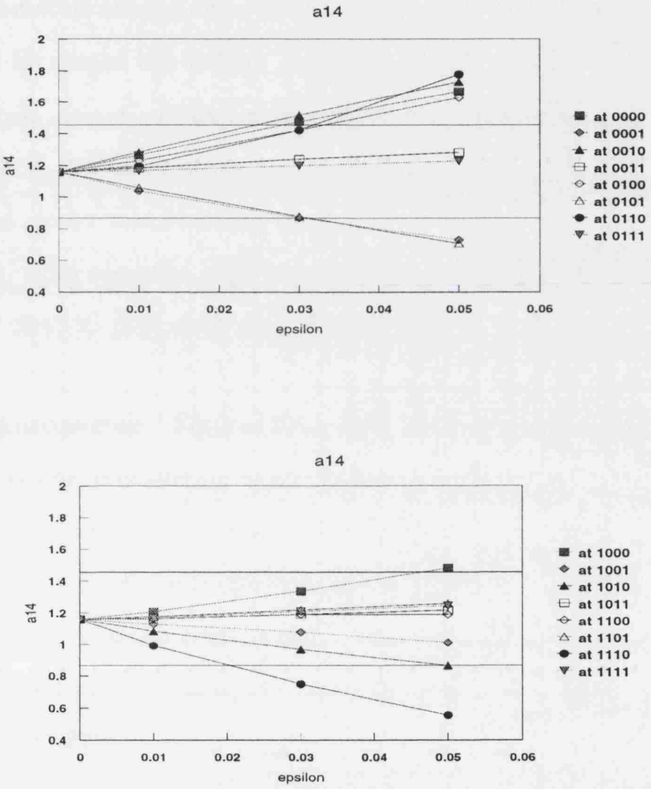


Figure 2.10: a_{14} as extra frequency is given on a response pattern, against the percentage of extra frequency



above. (In fact we could only remove 13 observations from 0010 and 17 from 0100. For the move of observations in the opposite direction we moved the number shown in the Figures - 16 or 20).

Difficulty parameter Figures 2.11 show the difficulty parameters as the number of observations shifted between response patterns increases. There are many straight lines which mean that the parameters change little even with 16 being shifted between response patterns. As we saw in Section 2.5.2, the parameters are more affected by changes in frequencies between particular pairs of response patterns. For changes at those response patterns some parameters change smoothly as the number of observations shifted between response patterns increases but others exceed the confidence intervals of the original parameter estimates. As we saw in Section 2.5.2 a_{01} and a_{03} reach the bounds of their confidence intervals if 12 observations are shifted between response patterns.

a_{04} exceeds the upper bound of its confidence interval with 16 observations shifted from 1110 to 1111. The opposite shift of observations also causes a large change in the opposite direction but the parameter remains within the confidence interval.

Discrimination parameter Figures 2.12 show the discrimination parameters as we increase the number of observations we move between the same response patterns. Changes seem smooth as the number of observations increases but a_{11} and a_{12} reach the bounds of the confidence intervals of the original estimates with 14 observations shifted between some response patterns and a_{13} even sooner, with 9 observations. In particular, a_{13} exceeds the upper bound of the confidence interval if 9 observations are shifted from 0010 to 1010, to 0011 or to 0000.

Although 9 observations are a large proportion of 13, which is the frequency of response pattern 0010, they make up only 0.6% of the total frequency. And 14 observations, which when shifted between some particular response patterns most parameters are affected, make up only 1% of the total frequency.

So the estimator can be quite sensitive to shifts in frequencies between response patterns.

It is hard to draw a line and say how much resistant the parameters of any model should be. Hampel et al. (1986) cite a lot of surveys which contained a large percentage,

Figure 2.11: a_{0i} as observations are shifted between response patterns, Scale 7, first 4 items

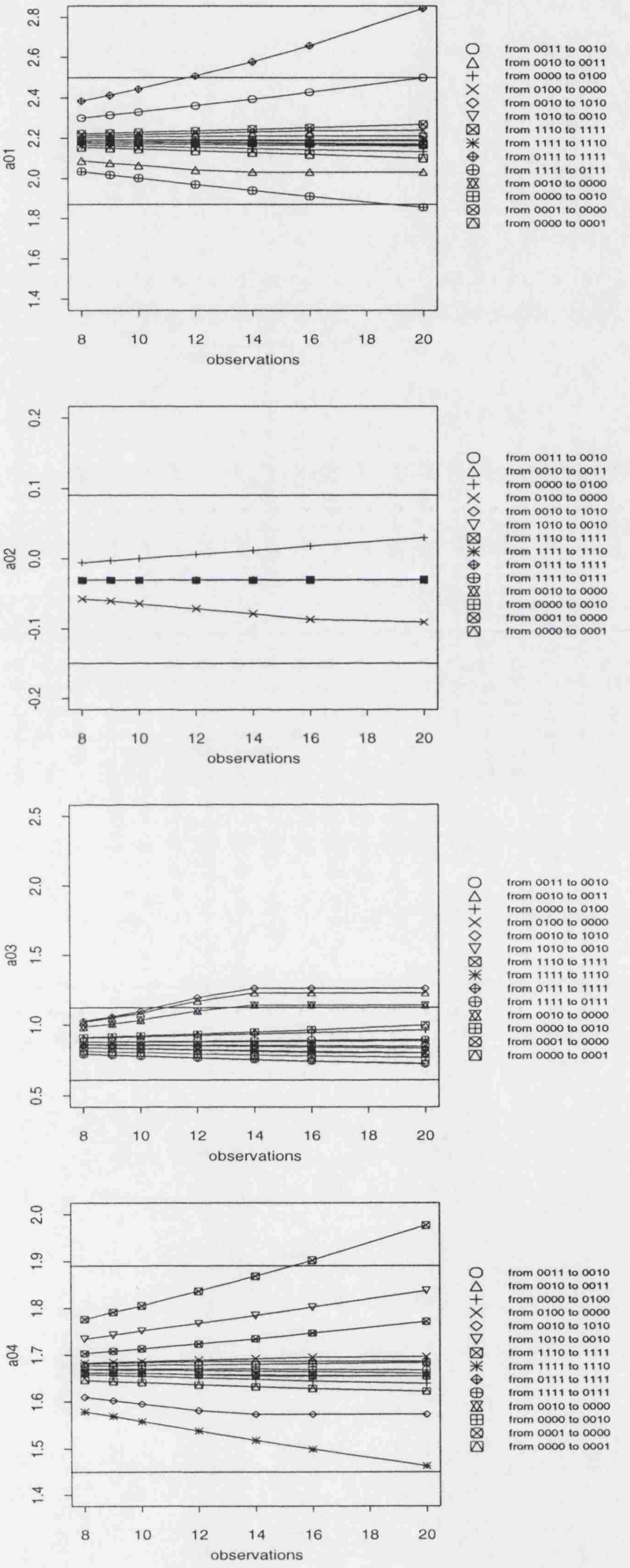
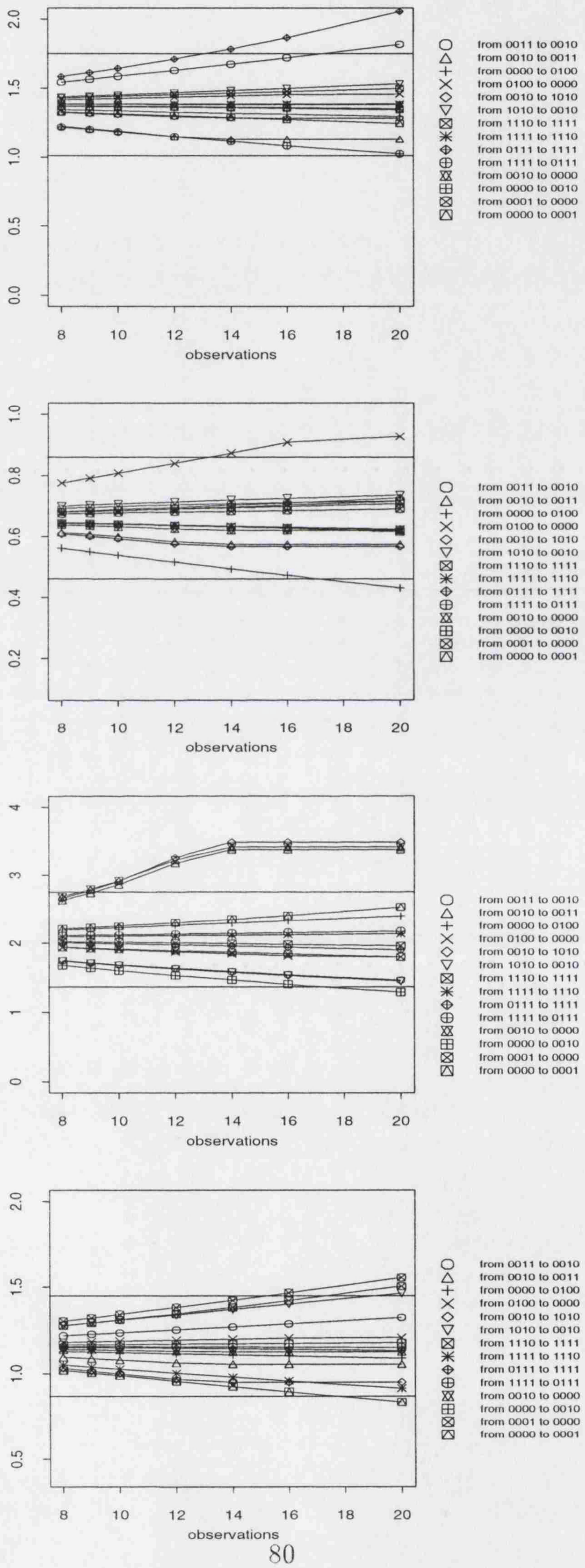


Figure 2.12: a_{1i} as observations are shifted between response patterns, Scale 7, first 4 items



up to 50%, of gross errors. But if the estimator was that 'insensitive' though, then it would not distinguish between the contaminated data and data coming from a different model. Thus the pattern shown above, a smooth change as the amount of distortion increases, is most desirable. The logit/probit model generally shows this behaviour, but occasionally a small distortion can cause a threshold model. The breakdown point seems to be 0.6% since with a larger amount of contamination large changes in the parameter estimates may occur.

2.7 Robustness properties of the posterior means

One of the aims of fitting a latent trait model, is to estimate the latent scores and place the individuals along the latent trait continuum. As we saw in Chapter 1, Section 1.5, the posterior means E(z|x) are the most appropriate means to score individuals on the latent scale.

We would like to see what effect a small contamination of the data will have on the scoring of the individuals. Therefore we examine changes in the posterior means for changes in the frequency distribution of the response patterns. As before, we will derive the Influence Function of the posterior means and the component score, and then observe the actual changes in the posterior means when the data are contaminated.

2.7.1 Influence Function of the posterior means

As we saw in Chapter 1, Section 1.5 the posterior mean is given by

$$E(z|\mathrm{x}_l) = \sum_{t=1}^{k} z_t g(\mathrm{x}_l|z_t)h(z_t)/f(\mathrm{x}_l) \tag{2.7}$$

Since the influence function is a derivative of the estimator the usual calculus properties apply and so the influence function of the posterior means is derived as follows:

$$\mathrm{IF}(E(z|\mathrm{x}_l)) = \sum_{t=1}^{k} z_t h(z_t) \frac{[\mathrm{IF}(g(\mathrm{x}_l|z_t))f(\mathrm{x}_l) - g(\mathrm{x}_l|z_t)\mathrm{IF}(f(\mathrm{x}_l))]}{f(\mathrm{x}_l)^2} \tag{2.8}$$

where

$$\mathrm{IF}(f(\mathrm{x}_l)) = \sum_{t=1}^{k} \mathrm{IF}(g(\mathrm{x}_l|z_t))h(z_t), \tag{2.9}$$

Table 2.12: Scale 7, first 4 items, Influence Function of E(z|x)

	x							
resp	0000	0001	0010	0011	0100	0101	0110	0111
0000	-0.2	2.1	-7.3	-4.7	0.4	2.6	-6.6	-3.9
0001	2.1	-3.7	8.4	1.1	3.4	-2.9	8.7	1.1
0010	-2.8	4.1	-35.7	-22.8	0.7	9.7	-28.4	-15.2
0011	1.3	4.2	-2.3	1.6	3.1	6.3	0.1	4.3
0100	1.2	2.7	3.2	4.1	-5.7	-4.2	-3.9	-3.4
0101	2.6	-0.8	11.6	6.7	-0.9	-4.6	7.3	2.0
0110	0.6	3.5	-11.1	-5.6	0.5	4.3	-9.6	-4.1
0111	3.4	6.1	12.0	13.7	4.3	6.6	12.3	14.3
1000	2.0	4.3	8.6	9.3	3.6	5.1	9.1	10.2
1001	0.5	0.0	1.1	0.5	1.9	1.4	2.4	1.8
1010	0.2	-2.0	-4.2	-5.4	1.2	-0.7	-2.6	-3.9
1011	-0.5	-0.3	0.0	0.3	-2.4	-2.0	-1.8	-1.5
1100	1.9	2.7	8.6	8.3	-1.1	-0.6	5.1	4.7
1101	-1.1	0.8	-9.5	-5.8	-1.4	1.3	-8.6	-5.0
1110	1.5	-5.1	6.3	-0.8	2.0	-4.7	6.5	-1.0
1111	-1.0	-0.9	-1.0	-1.1	-0.2	-0.3	-0.4	-0.3
min	-2.8	-5.1	-35.7	-22.8	-5.7	-4.7	-28.4	-15.2
25%	-0.3	-0.8	-5.0	-4.9	-1.0	-2.3	-4.6	-3.9
median	0.9	1.4	0.5	0.4	0.6	0.5	-0.2	-0.7
75%	1.9	3.6	8.4	4.7	2.3	4.5	6.7	2.6
max	3.4	6.1	12.0	13.7	4.3	9.7	12.3	14.3

$$\mathrm{IF}(g(\mathrm{x}_l|z_t)) = \sum_{i=1}^{p} \mathrm{IF}(\pi_i)^{x_i}(-\mathrm{IF}(\pi_i)^{(1-x_i)}) \frac{\prod_{i=1}^{p} \pi_i^{x_i}(1-\pi_i)^{(1-x_i)}}{\pi_i^{x_i}(1-\pi_i)^{(1-x_i)}} \tag{2.10}$$

and

$$\mathrm{IF}(\pi_i) = (\mathrm{IF}(a_{0i}) + \mathrm{IF}(a_{1i})\, z)\, \pi_i\, (1-\pi_i) \tag{2.11}$$

2.7.2 Influence Function results

We calculated the Influence Function for the posterior means obtained from fitting the first 4 items of Schuessler's scale 7. Table 2.12 shows the Influence Function values for all the posterior means as each response pattern carries extra probability. Figures 2.13 show the Influence Function of all the posterior means as response patterns 0000, 0001 and 0010 carry extra probability.

Table 2.12 cont., Scale 7, first 4 items, Influence Function of E(z|**x**)

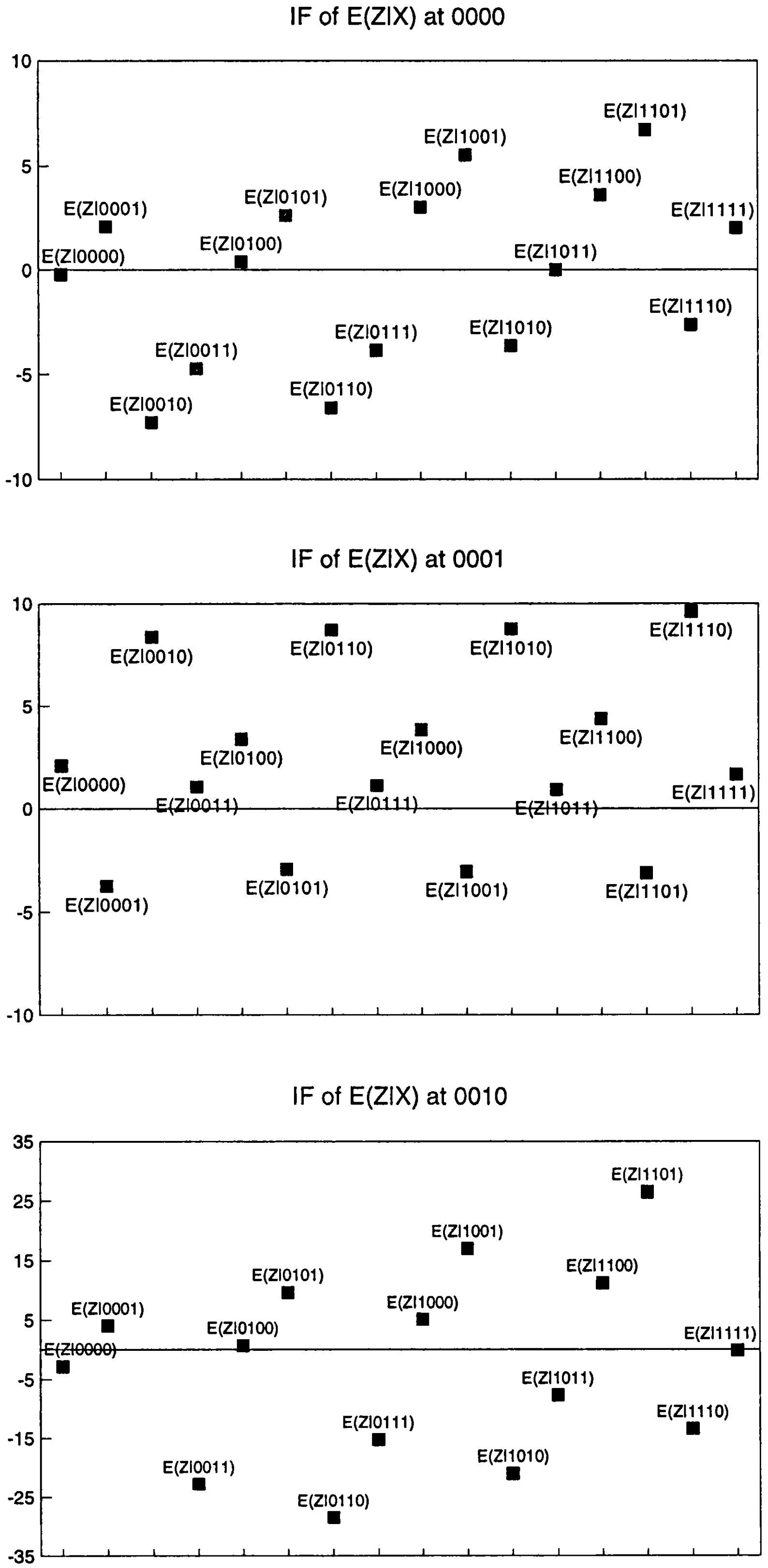
	x							
resp	1000	1001	1010	1011	1100	1101	1110	1111
0000	3.0	5.5	-3.6	0.0	3.6	6.7	-2.6	2.0
0001	3.9	-3.0	8.8	0.9	4.4	-3.1	9.6	1.7
0010	5.1	17.1	-20.9	-7.7	11.3	26.5	-13.4	-0.2
0011	-2.1	1.6	-4.9	-1.8	0.1	4.2	-2.9	0.1
0100	2.8	3.5	4.2	6.9	-4.2	-3.7	-3.3	-0.9
0101	2.1	-2.5	9.6	4.8	-1.9	-7.4	5.2	-0.4
0110	1.9	6.8	-7.0	-1.7	3.0	9.1	-5.7	-1.0
0111	-6.6	-4.8	0.3	0.6	-6.0	-5.2	-0.3	0.0
1000	-3.5	-2.5	1.0	1.2	-2.7	-2.4	1.1	1.8
1001	0.3	-0.1	0.8	-0.3	1.7	1.2	2.1	0.8
1010	3.0	1.8	0.0	-1.5	4.4	3.7	1.9	-0.1
1011	-0.3	-0.1	0.4	1.2	-2.1	-1.9	-1.5	-0.8
1100	-1.0	-1.1	4.4	4.2	-4.3	-5.1	0.3	-0.3
1101	1.9	5.3	-4.5	-0.9	2.5	6.7	-3.8	-0.6
1110	3.3	-3.6	7.8	-0.1	3.5	-3.8	8.4	-0.3
1111	-1.1	-1.2	-1.3	-1.3	-0.5	-0.6	-0.6	-0.2
min	-6.6	-4.8	-20.9	-7.7	-6.0	-7.4	-13.4	-1.0
25%	-1.0	-2.5	-3.9	-1.4	-2.2	-3.8	-3.0	-0.5
median	1.9	-0.1	0.3	0.0	0.9	-1.2	-0.4	-0.2
75%	3.0	4.0	4.3	1.2	3.5	4.8	2.0	0.3
max	5.1	17.1	9.6	6.9	11.3	26.5	9.6	2.0

Let us see how the posterior means change as extra probability is placed on a response pattern.

As the probability of 0000 increases, items 1, 2 and 4 become more discriminating, whereas item 3 becomes less discriminating. We observe that most of the posterior means that display negative changes are of response patterns that have response 1 for item 3 (i.e. response patterns 0010, 0110, 0111, 1010, 1110 but not 1011 and 1111). Since item 3 has lost some of its discriminating power, responding 1 to this item is less rewarding. The largest positive changes occur for response patterns 1001, 1101, 1100 and 1000. All have response 1 for item 1, and as it happens a_{1i} has the largest positive IF at response pattern 0000.

As probability of response pattern 0001 increases, items 1, 2 and 3 become more discriminating and item 4 less discriminating. Again, the posterior means of the response patterns that have response 1 for item 4 display negative rates of change, as responding 1 to item 4 is not so 'important'. Most of the posterior means move upwards, since the other three items become more discriminating and responding 1 to them pushes the

Figure 2.13: Influence Function of E(z|$\mathbf{x}$) when response patterns 0000, 0001 and 0010 carry extra probability



posterior mean higher.

We observe that the rates of change are not as extreme as the rates for the individual parameters. To get a clearer idea of what these values represent we shall study the actual rates of change as extra frequency is placed on each response pattern in turn, and also what transpositions in the ordering of the response patterns occur, since it is the ordering of the response patterns that essentially matters.

2.7.3 Actual rates of change

Graphs 2.14 show the actual rates of change of all response patterns as 3% extra frequency is placed on response patterns 0000, 0001 and 0010.

We observe that the pattern of the changes is similar to the IF but their range is smaller, i.e. the largest actual rates of change are not as large as predicted by the IF. Since the location and scale of the posterior means is determined by the prior distribution, it is only their ordering that matters. In the following we shall see whether these changes actually alter the ordering of the response patterns.

Graphs 2.15 and 2.16 show E(z|x) from the contaminated data and E(z|x) from the original data ordered according to the magnitude of E(z|x) of the original data, when some response patterns carry 3% of the total extra frequency.

When 0000 carries extra frequency, the ordering of the response patterns is preserved for most response patterns, except for the following transpositions: E(z|0010) is now lower than E(z|0101) and E(z|1100) so it is shifted two places down, E(z|0110) is now lower than E(z|1001) and E(z|1010) lower than E(z|1101) whereas before they were higher, so the last two are shifted one place down.

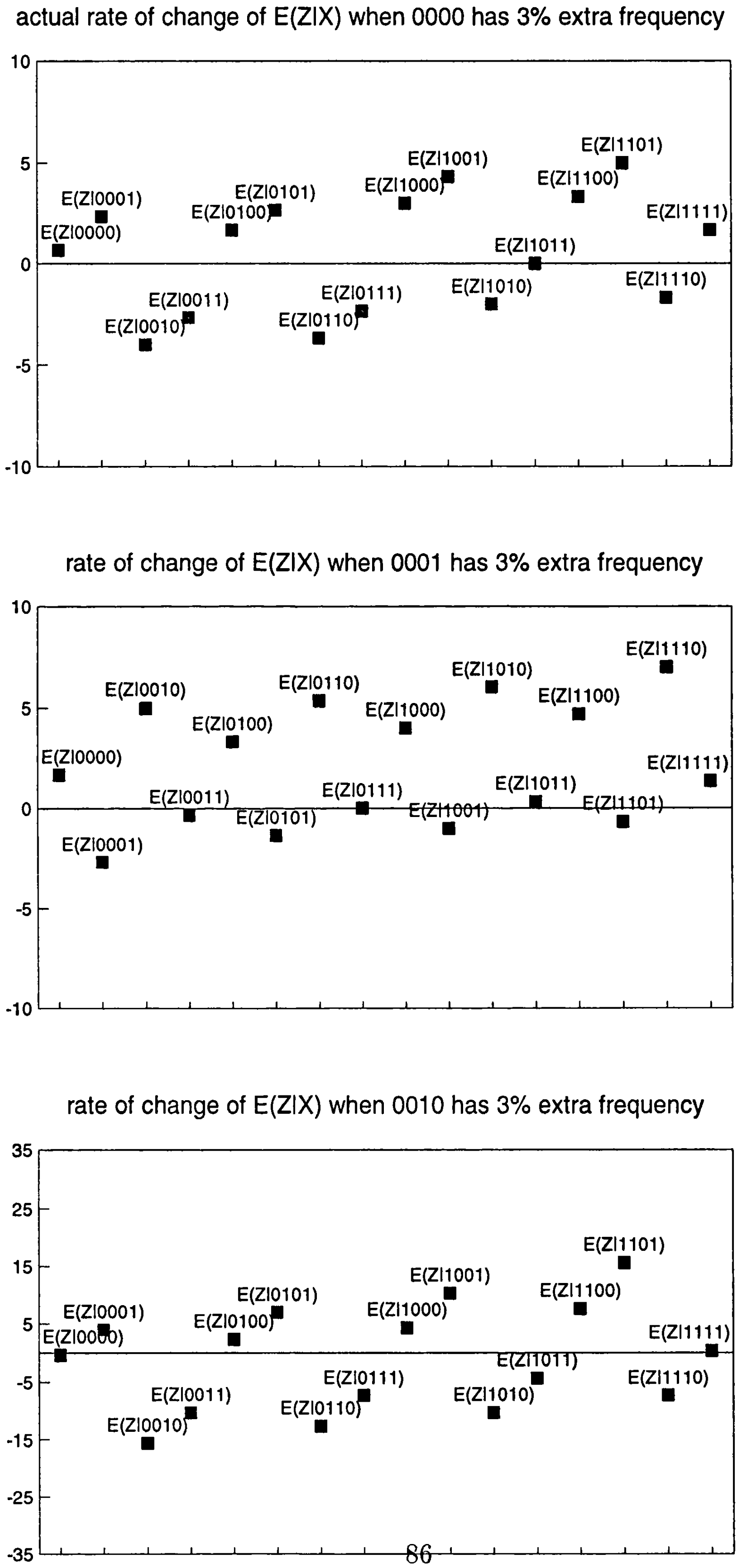
The ordering of the response patterns is preserved for all of them when 3% of the total frequency is placed on response pattern 0001 but more severe changes occur when extra frequency is placed on response pattern 0010. In the latter case, response pattern 0010 is shifted down four places, since item 3, the only one with response 1 in that pattern, has lost much of its discriminating power. Response pattern 0110, is also shifted down 3 places, below 1001, 1100 and 0101, whereas before it was above them.

Moreover, response pattern 1010 is now below 1101 and 1001 whereas 1101 is now above 1010, 0111 and 1110.

Changes as extra frequency is placed on response patterns 0101, 0110 and 0111 are

Figure 2.14: Actual rates of E(z|x) as 3% extra frequency is placed on response patterns 0000, 0001 and 0010

Please note different range of y scale for third graph



actual rate of change of E(ZIX) when 0000 has 3% extra frequency

rate of change of E(ZIX) when 0001 has 3% extra frequency

rate of change of E(ZIX) when 0010 has 3% extra frequency depicted in Graphs 2.16.

The largest transpositions occurred when extra frequency was placed on response pattern 0010, 0110 and 0111. These were the response patterns with the largest IF. Generally transpositions occur in the middle of the latent trait range. The ordering of the response patterns at the extremes of the range is more robust to changes in the data. We conclude that the IF corresponds to actual rates of change when the amount of contamination is small (for example 3%) and the large rates indicate changes that will affect the ordering of the response patterns.

2.7.4 Shifting observations from one response pattern to another

In this section we examine the effects of shifting observations from one response pattern to another on the posterior means.

Figures 2.17, 2.18 and 2.19 show the posterior means from contaminated data in the above way ordered according to the posterior means of the original data.

In Figure 2.17, top panel, we see that when we move 14 observations from 0000 to 0010 some transpositions in the ordering of the response patterns occur. In particular 0010 is shifted below 0101 and 1100, 0110 is shifted below 1001 and 0111 is shifted below 1101. This is due to the decrease in the discriminating power of item 3, as more observations are put on 0010. The pattern of changes is very similar to the changes when extra observations are put on 0010 but the changes are not so severe, since the percentage of contamination is smaller.

The reverse action, i.e. putting observations from 0010 to 0000 causes larger transpositions (Figure 2.17, bottom panel). Item 3 becomes even more discriminating and thus 0010 is pushed above 1001, whereas 1101 is pushed below 0010, 0110 and 0011. Here the changes are more severe than when extra observations are put on 0000.

In Figure 2.18, top panel we have transpositions of up to two places (i.e. the posterior means are shifted below or above two neighbouring posterior means), as 14 observations are shifted from 0101 to 0110. The pattern of changes is similar to the ones observed when extra observations are put on 0110 but again changes here are smaller. In Figure 2.18, bottom panel, we have the reverse shift in frequency but here only 7 observations are shifted. The pattern of changes is the opposite of the one observed above, but as expected changes are much smaller.

Figure 2.15: Posterior means from contaminated and original data ordered according to posterior means from original data: 3% contamination on response patterns 0000, 0001 and 0010

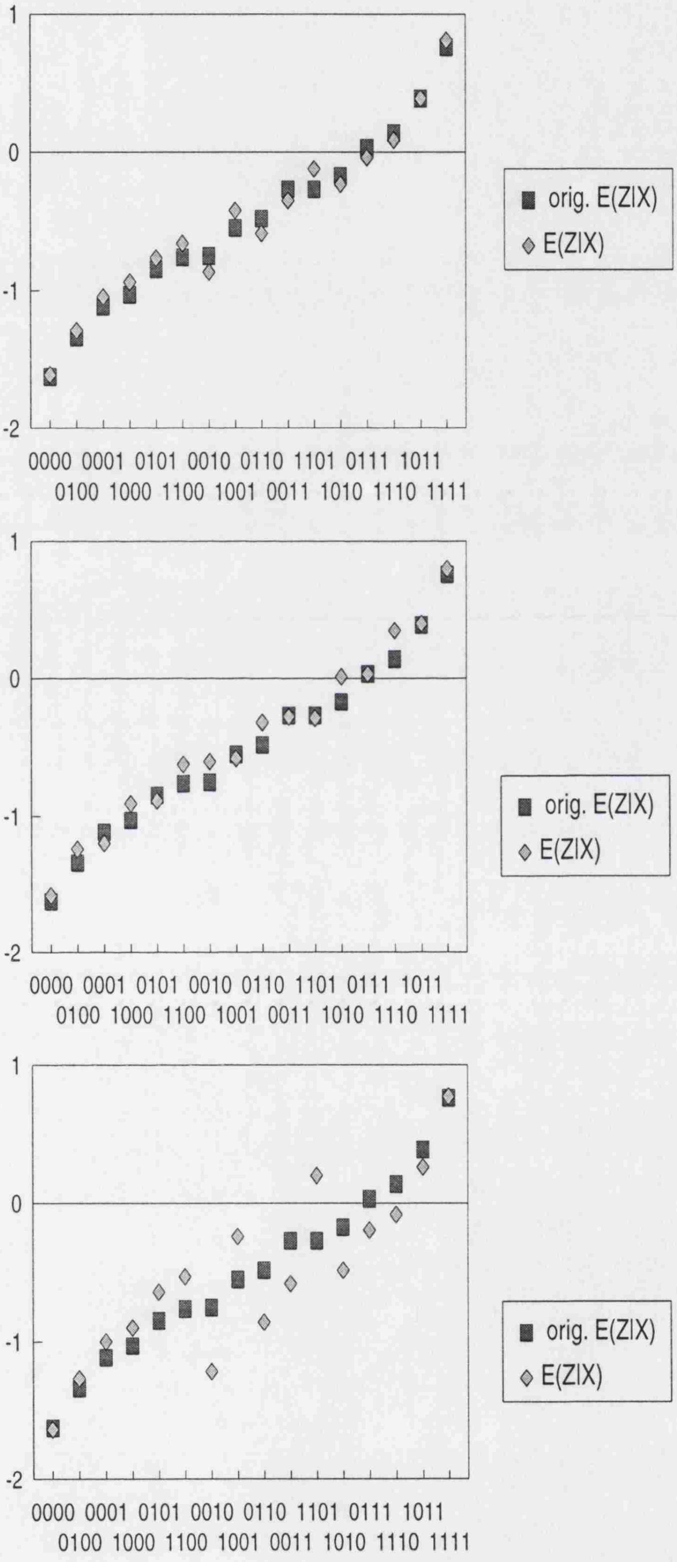
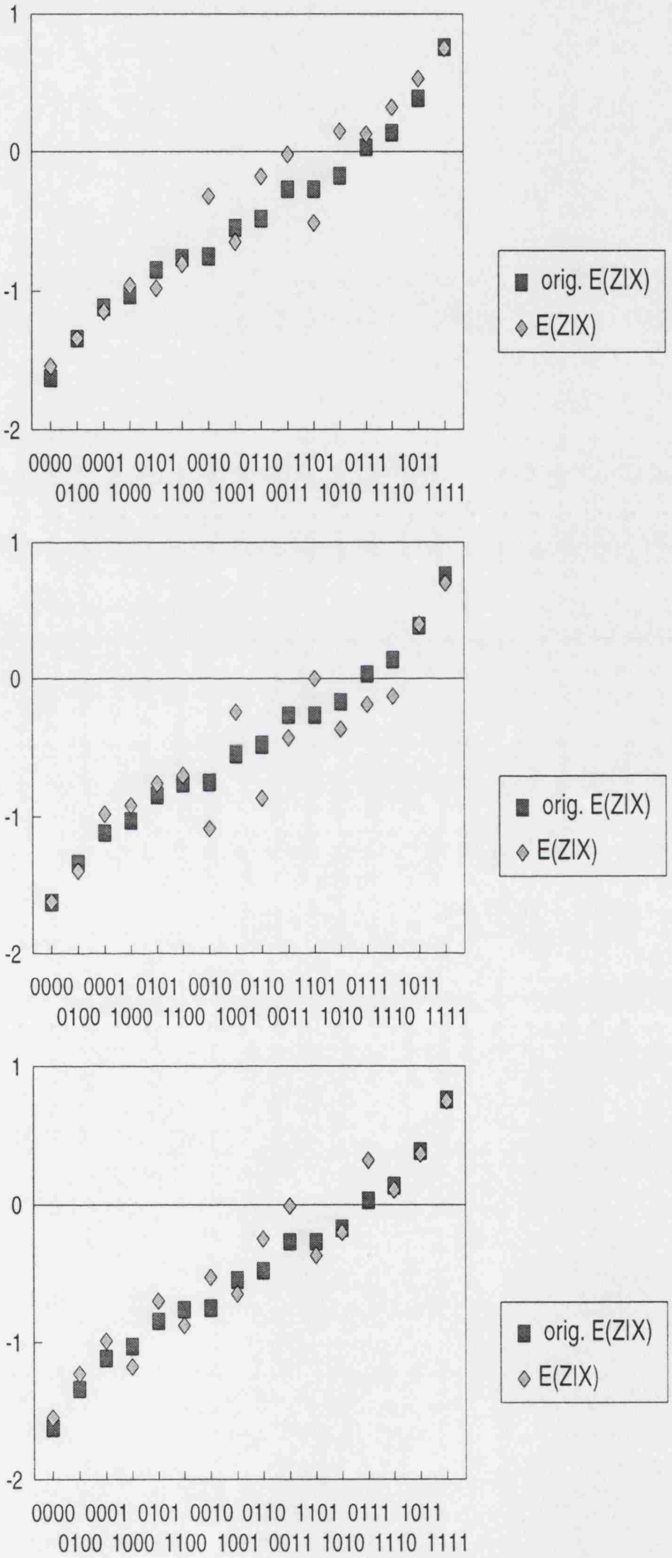
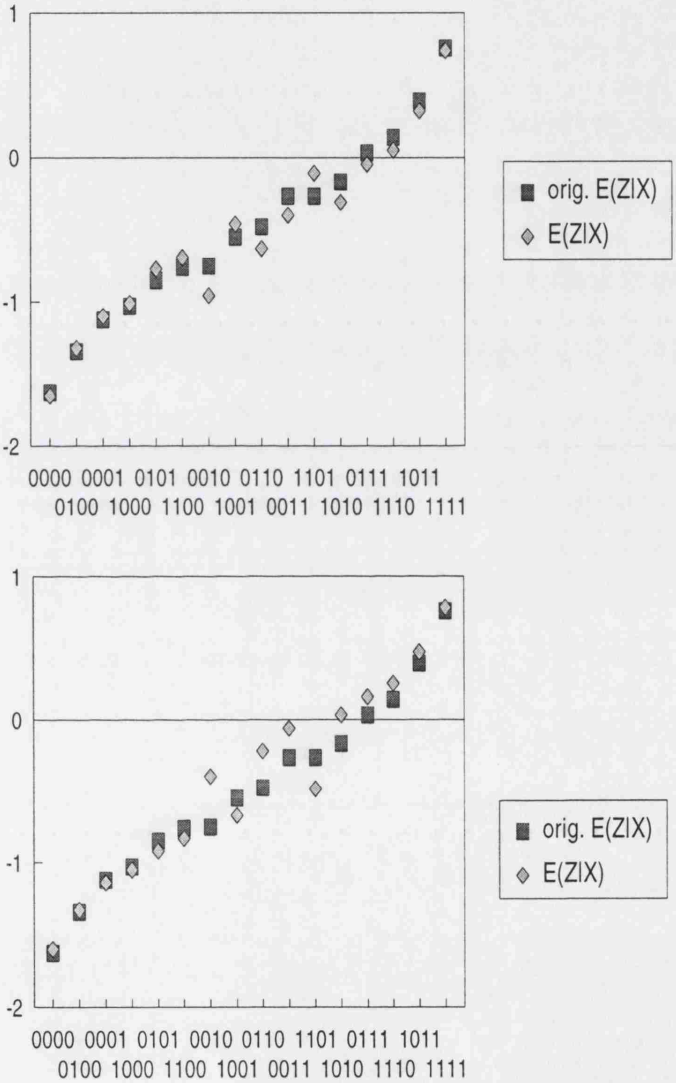


Figure 2.16: Posterior means from contaminated and original data ordered according to posterior means from original data: contamination on response patterns 0101, 0110, 0111



Figures 2.19 show one place or none transpositions (posterior means may only change with one neighbouring response pattern places) and this is also the typical picture of the posterior means when we shift observations between response patterns not shown here.

Figure 2.17: Posterior means from contaminated and original data ordered according to posterior means from original data; shifting 14 observations from 0000 to 0010 (top) and shifting 14 observations from 0010 to 0000 (bottom)



2.8 Existing 'robust' methods

In the following we will examine how methods that have been proposed in the literature as 'robust' behave when the data are contaminated by placing extra probability on a

Figure 2.18: Posterior means from contaminated and original data ordered according to posterior means from original data; shifting 14 observations from 0101 to 0110 (top) and shifting 7 observations from 0011 to 0101 (bottom)

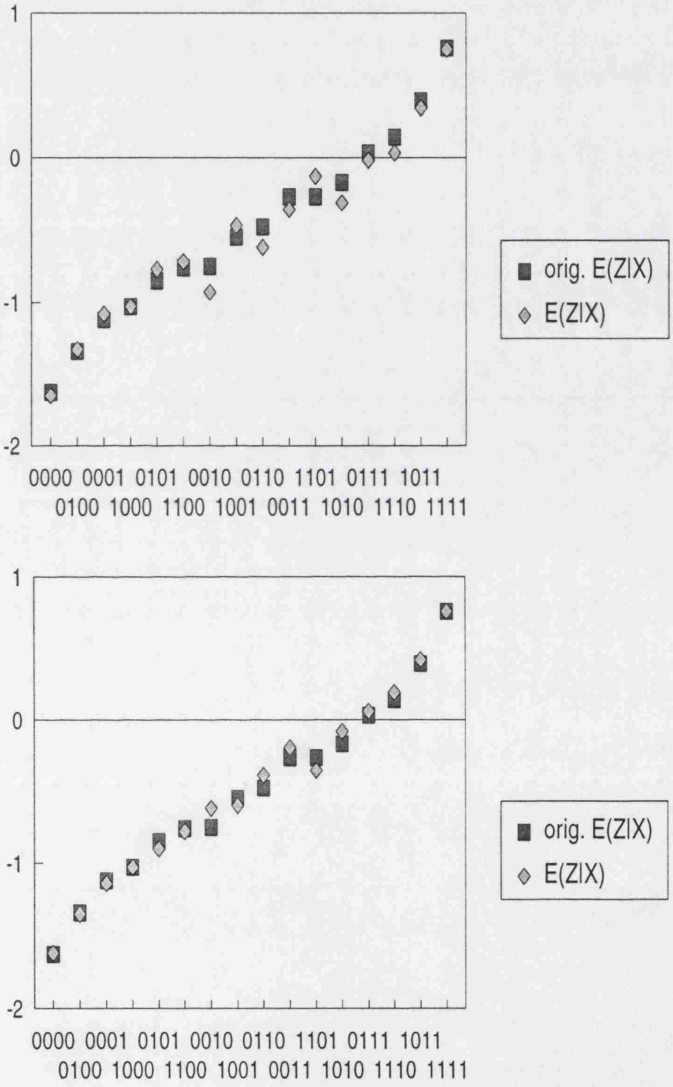
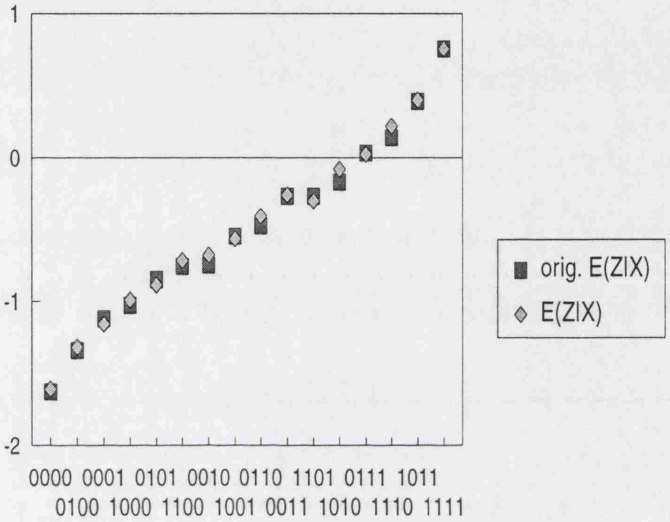
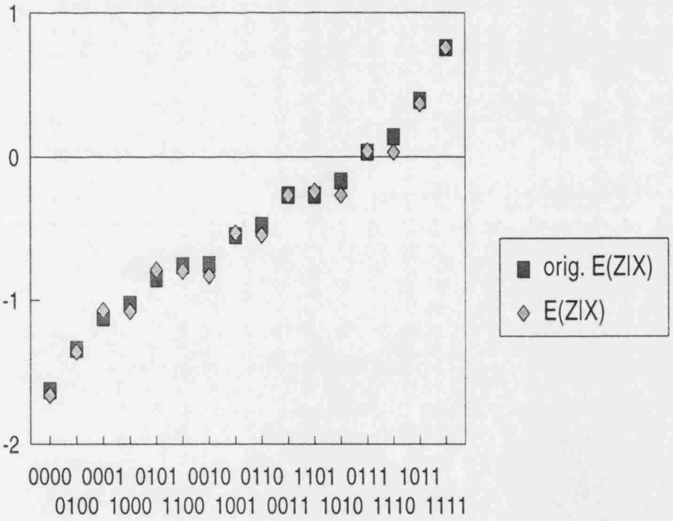


Figure 2.19: Posterior means from contaminated and original data ordered according to posterior means from original data; shifting 14 observations from 1111 to 1110 (top) and shifting 14 observations from 1110 to 1111 (bottom)





response pattern.

2.8.1 'Jackknifed' estimates of the parameters and the posterior means

Wainer and Wright (1980) (see Section 1.7) suggested a combination of the score obtained from the data and of the scores obtained from omitting each item in turn as a robust estimate of the score of the latent variable.

They called these estimates 'jackknifed' estimates.

Following their idea we will explore whether a combination of the scores obtained from $p-1$ item datasets will give smaller rates of change when the data are contaminated.

We create datasets omitting each item in turn, so for a dataset with p items we create p subsets, each consisting of $p-1$ items. For each one of these subsets we estimate the item parameters and the posterior means for each response pattern.

Item parameters As a robust estimate of the parameters of the p item dataset we take the median of parameter estimates obtained from the subsets of items.

We applied this procedure to the first four items of scale 7 and to all six items of scale 7, both for the original and contaminated data. The contamination we considered is giving 3% of the total frequency to response pattern 0010 for the 4 item set and the same frequency to response pattern 000010 for the 6 item set, as these were the response patterns for which the IF had the extreme values for most parameters.

In the upper part of Table 2.13 the 'robust' parameter estimates for the original and contaminated 4 item sets are given, together with the rates of change of the parameters when there is contamination. In the lower half of the same table there are the standard parameter estimates and rates of change for the same data .

We observe that the standard and 'robust' estimation procedures give very similar estimates, both for the original and contaminated data. The discrimination parameter for item 3 is somewhat smaller if the robust procedure is used.

The rates of change of the parameters of the robust procedure are slightly smaller for three of the four items, but very close to the rates of change obtained from the standard procedure. a_{13} still changes a lot if we place extra frequency on 0010.

In Table 2.14 we have the parameter estimates for the 6 item dataset, with the robust and standard estimation procedure, for the original and the contaminated data. Again

Table 2.13: Scale 7, first 4 items, 'robust' parameter estimates

median of estimates of omitted item subsets						
i	a_{0i}			a_{1i}		
	orig. data	3% at 0010	rate of ch.	orig. data	3% at 0010	rate of ch.
1	2.15	2.17	0.82	1.32	1.70	12.75
2	-0.03	-0.10	-2.37	0.69	0.88	6.25
3	0.83	0.67	-5.30	1.92	0.96	-31.85
4	1.68	1.67	-0.22	1.17	1.51	11.28
	standard estimates					
i	a_{0i}			a_{1i}		
	orig. data	3% at 0010	rate of ch.	orig. data	3% at 0010	rate of ch.
1	2.19	2.21	0.70	1.38	1.75	12.33
2	-0.03	-0.10	-2.33	0.66	0.86	6.80
3	0.87	0.68	-6.37	2.06	0.98	-35.90
4	1.67	1.68	0.33	1.16	1.52	11.90

we observe that the rates of change of the parameters when the robust procedure is used are very similar to the ones from the standard procedure and it is hard to say which estimator gives smaller rates of change.

Posterior means Using the $p-1$ item data we calculated 'robust' posterior means in two ways:

(a) We assigned the posterior means obtained from the $p-1$ item data to their corresponding response patterns of the p item data. For example, if item 1 was omitted, the posterior mean of the response pattern 000 was assigned to 0000 and 1000. If item 3 was omitted, the posterior mean of 010 was assigned to 0100 and to 0110. So, for the original 4 item dataset, each response pattern was assigned 4 posterior means. For each response pattern we took the mean and median of the four posterior means as robust estimates of the score of the latent variable. We calculated scores with this procedure for the contaminated data as well. The rates of change of the 'robust' scores, together with the rates of change of the standard posterior means, are shown if Figure 2.20. The rates of change are very similar, the rates of change of the 'robust' estimates are as large as the rates of change of the posterior means obtained from all items.

In Figure 2.21 we have the median of the posterior means from the reduced item data for the contaminated data plotted together with the same score estimates for the

Table 2.14: Scale 7, 6 items, 'robust' parameter estimates

median of estimates of omitted item subsets						
i	a_{0i}			a_{1i}		
	orig. data	3% at 000010	rate of ch.	orig. data	3% at 000010	rate of ch.
1	2.18	2.10	-2.67	1.36	1.61	8.33
2	-0.03	-0.10	-2.42	0.81	1.01	6.98
3	0.83	0.73	-3.42	1.94	1.99	1.82
4	1.72	1.69	-0.98	1.25	1.54	9.90
5	-0.16	-0.05	3.52	1.58	0.86	-23.88
6	1.83	1.89	1.98	1.50	1.96	15.35
	standard estimates					
i	a_{0i}			a_{1i}		
	orig. data	3% at 000010	rate of ch.	orig. data	3% at 000010	rate of ch.
1	2.19	2.14	-1.66	1.38	1.67	9.58
2	-0.03	-0.10	-2.32	0.77	0.96	6.35
3	0.83	0.70	-4.37	1.90	1.84	-1.83
4	1.73	1.72	-0.19	1.27	1.61	11.35
5	-0.16	-0.05	3.43	1.50	0.80	-23.26
6	1.81	1.89	2.41	1.46	1.94	16.00

original data, ordered according to the magnitude of the latter. It seems that these score estimates are more affected with the extra frequency on response pattern 0010, than the usual posterior means (Figure 2.15, bottom panel). We see that response pattern 0010 is now shifted six places down and 1101 is shifted four places up. (We note that 0010 for the original data is two places up, above 1100 and 1001 compared with its place when the posterior means were used to score response patterns on the latent variable).

(b) We took the median of the parameter estimates obtained from the four three-item subsets and calculated the posterior means for each response pattern for the original and the contaminated dataset. The rates of change of the posterior means when extra frequency is placed on response pattern 0010 are given in Figure 2.22. We observe that the rates of change of the 'robust' posterior means are equal to the rates of change of the usual posterior means for most response patterns and slightly smaller at response patterns 0010, 0110, 1010, 1101 and 1110.

These differences seem really too small to say that the 'robust' method is any better from the usual calculation of the posterior means. But we also need to see whether the ordering of the response patterns remains unchanged when the data are contaminated

Figure 2.20: Actual rates of change of posterior means as 3% extra probability is placed at response pattern 0010

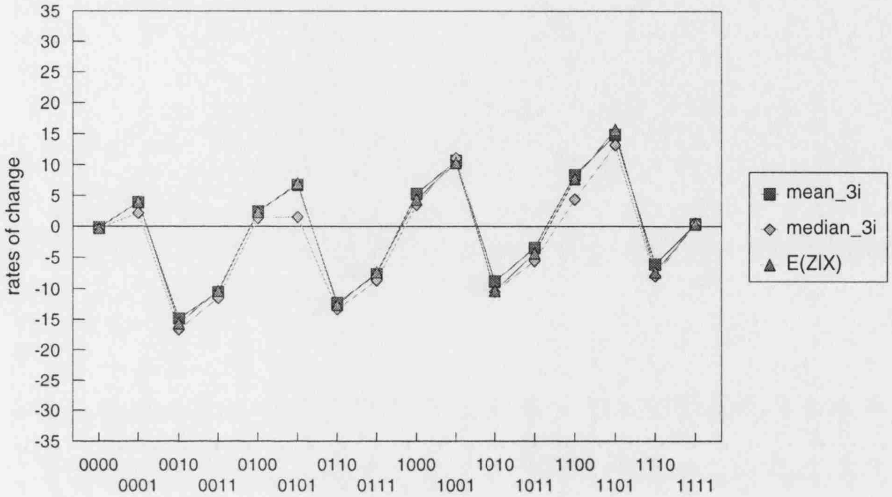
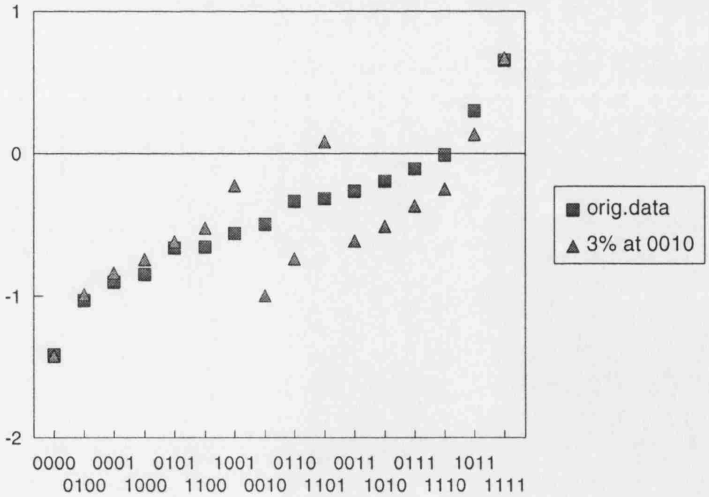


Figure 2.21: Scores equal to the median of the posterior means from the $p-1$ subsets for contaminated (3% at 0010) and original data ordered according to the scores for the original data



and we use the robust method. In Figure 2.23 we have the posterior means for the contaminated data against the posterior means for the original data, both calculated with the 'robust' procedure. There are still a lot of transpositions. These can be compared with the ones happening with the usual posterior means (the third panel of Figure 2.15). With the new estimates, 0010 is shifted down three places (as in Figure 2.15), except that now 0010 is below 1100 even for the original data.

0110 is shifted down three places, below 1001, 1100 and 0101, as with the standard posterior means.

0011 is shifted below 1100 and just below 1001. With the standard posterior means 0010 was also shifted below 1100 but remained just above 1001.

And as with the standard posterior means, response pattern 1010 has moved below 1101 and 1001 whereas 1101 has moved above 1010, 0111 and 1110.

So, although this 'robust' procedure attenuated a bit the effect of the large discrimination parameter of item 3 and moved 0010 below 1100 for the original data, it did not provide any robustness to the posterior means when the data were contaminated. We observed exactly the same transpositions between the posterior means of the contaminated and original data, calculated from the 'robust' parameters, as we had observed for the standard posterior means calculated straight from the data.

Figure 2.22: Rate of change of E(z|x) calculated with the usual and the 'robust' method (medians) when 3% extra frequency is placed on 0010

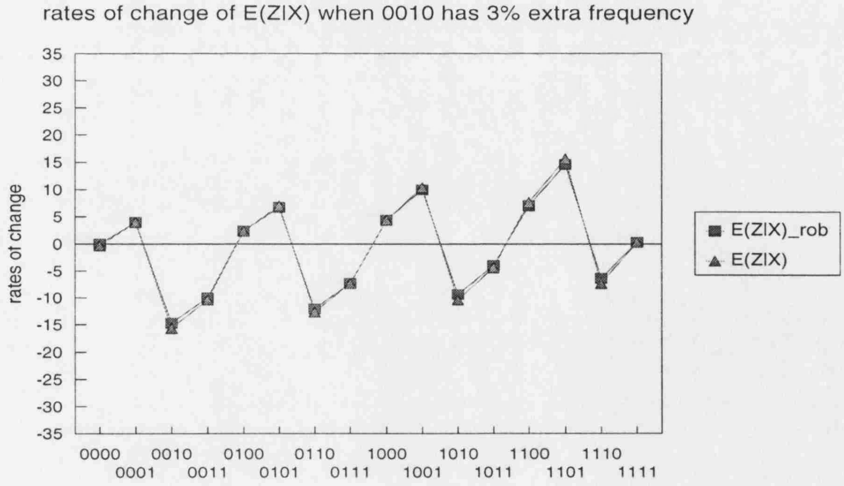
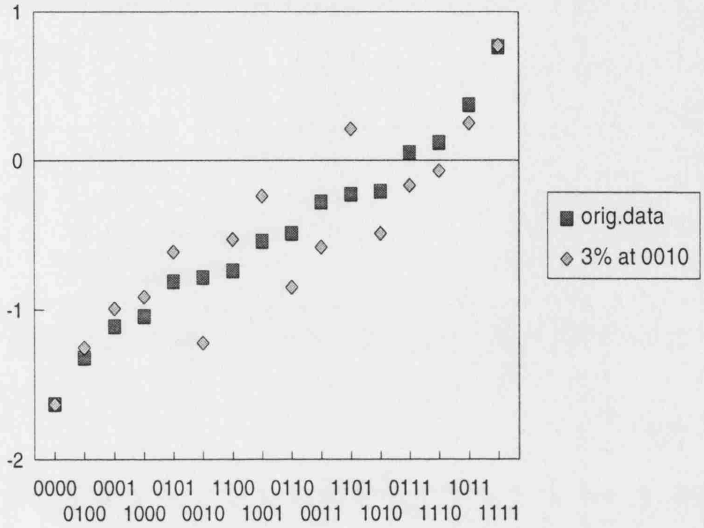


Figure 2.23: Posterior means calculated with 'robust' method from contaminated (3% at 0010) and original data ordered according to posterior means from original data

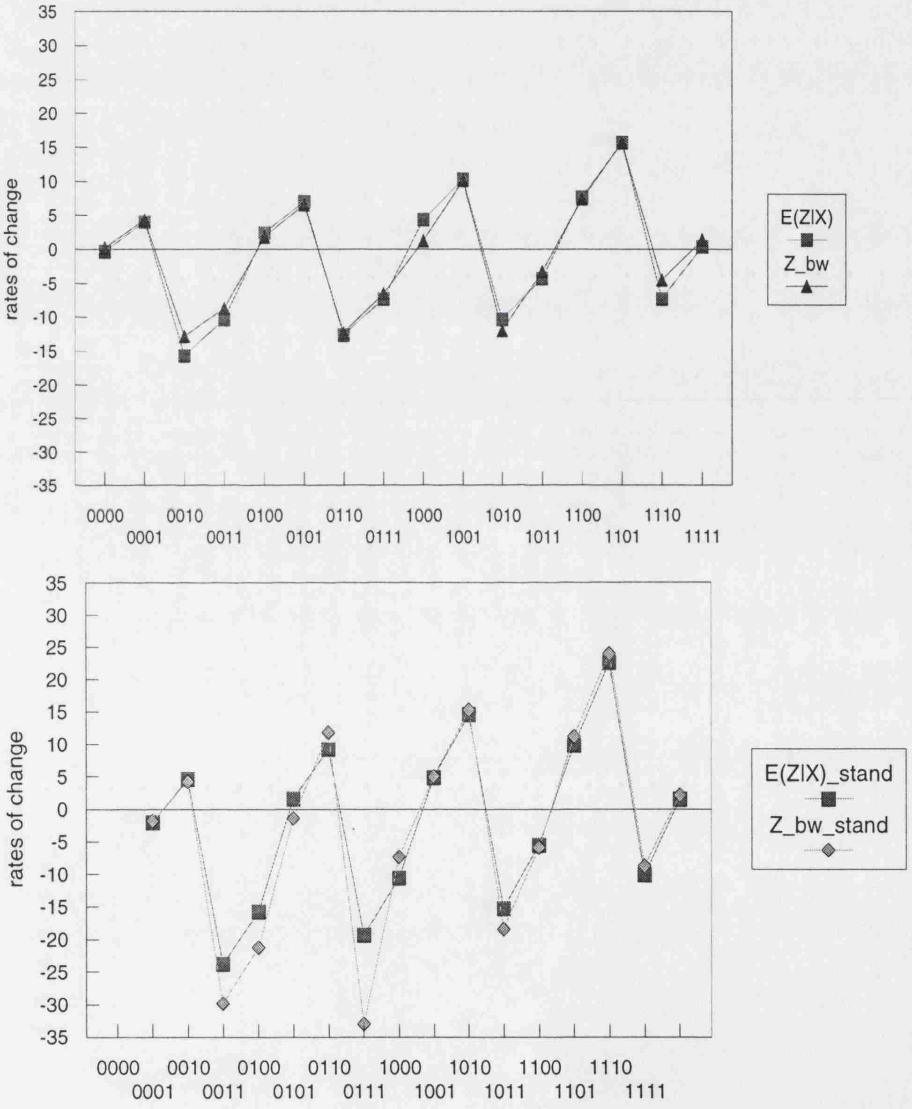


2.8.2 Biweight estimates of ability

In the following we will examine the behaviour of the biweight estimates of ability, developed by Mislevy and Bock (1982) (see Chapter 1, Section 1.7). Biweight estimates are claimed to be robust to isolated deviant responses. We would hope they might also be robust when the data are contaminated in the way we examined above, i.e. placing extra frequency on a response pattern. The biweight estimates are available as an option in the program BILOG (Mislevy and Bock 1990). Figure 2.24, top, shows the rates of change of the posterior means and the biweight estimates ('Z_bw') when 3% extra frequency is placed on response pattern 0010. We see that the rates of change are very similar, perhaps the rates of change of the biweight scores are slightly smaller.

Since the rates of change are dependent on the location and scale of the estimates, we rescaled the posterior means and the biweight scores so that the mean and standard deviation of the score estimates are 0 and 1 respectively. Again this in an option available in BILOG. The rates of change of the rescaled posterior means ('Z_bilog') and rescaled biweight scores ('Z_bw') when the data are contaminated in the way described above are shown in the lower Figure 2.24. We now see that the biweight scores have the largest rates of change for most response patterns. The biweight score estimates are no more robust to contamination in the data than the posterior means.

Figure 2.24: Rates of change of posterior means and biweight estimates when 3% extra frequency is placed on 0010 without scaling (top) and scaled to have mean=0 and sd=1 (bottom)



2.9 Conclusions

In this chapter we derived the Influence Function for the parameters of the latent trait model and the posterior means. We also observed the behaviour of the parameters and the posterior means in artificially contaminated data. Parameters and posterior means were generally well behaved, that is for small amount of contamination - extra observations - on most response patterns, the parameters changed little, in a predictable way and proportionally to the amount of contamination. There were cases though where relatively few extra observations on some response patterns caused large rates of change of some parameters, threw them out of their 'original' confidence interval and changed the ordering of the scores of the response patterns on the latent variable scale. Methods that have been proposed in the literature as 'robust' proved to be no more robust than the standard maximum likelihood estimates of the parameters or the posterior means for the data we examined.

Chapter 3

Sensitivity of the logit/probit model to the distribution of the latent factor

3.1 Introduction

In this chapter we shall investigate the robustness of the logit/probit model to changes in the prior distribution. We will first derive and examine the Influence Function of the parameters for changes in the prior and then investigate empirically the changes in the parameters when we fit priors which are mixtures of the standard normal and a small amount of probability on a point. We will also examine how the posterior means are affected when we fit the model with such priors.

We will then investigate changes in the parameters for gross changes in the prior. This consists of empirical analysis using mixtures of normals as priors. We use mixtures as this allows us to take a variety of shapes of distributions and see the effect of general changes in the form of the prior.

In the following we will need to distinguish (as in Seong (1990)) between the 'underlying' and the 'prior', as terms for the distribution of the latent variable of the population studied. Before proceeding to the main analyses, we shall first show the correspondence between changes in either of these two distributions.

3.1.1 The prior and the underlying distributions

The underlying distribution is in effect the generating mechanism of the response patterns. Given the item characteristics, the individual's position on the latent scale determines his/her probability of a positive response to an item. We do not know the form of the distribution but we can make an assumption about it in order to fit the model. The prior is the distribution used to fit the model. We need to clarify how the changes in the item parameters due to changes in the underlying distribution correspond to the changes in the item parameters due to changes in the prior distribution.

Let the 'true' item parameters be denoted with a_{ji} and the estimates obtained from fitting a model be denoted with a^*_{ji}.

Suppose that responses are generated from a population where the latent (underlying) variable y is distributed as $\mathrm{N}(\mu, \sigma^2)$ from the model

$$\mathrm{logit}\pi_i(y) = a_{0i} + a_{1i}y \tag{3.1}$$

If we fit with a $\mathrm{N}(\mu, \sigma^2)$ as prior then the estimated parameters will be $a^*_{0i} = a_{0i}$ and $a^*_{1i} = a_{1i}$ (approximately, apart from sampling variations).

But if we fit with the N(0,1) as the prior then the model fitted is

$$\mathrm{logit}\pi_i(z) = a^*_{0i} + a^*_{1i}z \tag{3.2}$$

But $z = (y - \mu)/\sigma$ so,

$$\begin{aligned} \mathrm{logit}\pi_i(y) &= a^*_{0i} + a^*_{1i}(y - \mu)/\sigma \qquad (3.3) \\ &= a^*_{0i} + a^*_{1i}y/\sigma - a^*_{1i}\mu/\sigma \end{aligned}$$

If we equate (3.1) and (3.3) then

$$a_{1i} = a^*_{1i}/\sigma \quad \text{or} \quad a^*_{1i} = a_{1i}\sigma \tag{3.4}$$

and

$$a_{0i} = a_{0i}^* - a_{1i}^* \mu / \sigma \tag{3.5}$$

or (3.6)

$$a_{0i}^* = a_{0i} + a_{1i}^* \mu / \sigma \tag{3.7}$$

$$= a_{0i} + a_{1i} \mu \tag{3.8}$$

So, if $\mu > 0$ and we fit a N(0,1) prior the difficulty parameter estimates will be larger than the true parameters. Intuitively, if we estimate a population with mean ability higher than the mean of the prior, then the difficulty parameters will appear easier. Similarly, if the true discrimination parameter is larger than 1 and we fit the model with a N(0,1) prior, the discrimination parameter estimates will be larger than the true ones, reflecting the fact that the population is more dispersed than a one having a standard normal distribution.

Bartholomew (1993) makes the point that the parameter estimates are confounded with the population parameters and thus if we fitted some items to two populations using the N(0,1) as prior then the items would appear easier in the population with the higher ability. Similarly, the items would appear more discriminatory in the population with the largest variance.

Now, if the underlying distribution is N(0,1) the 'true' model is

$$\text{logit}\pi_i(z) = a_{0i} + a_{1i} z. \tag{3.9}$$

If the data are fitted with a $N(\mu, \sigma^2)$ as prior, then the model fitted is

$$\text{logit}\pi_i(y) = a_{0i}^* + a_{1i}^* y \tag{3.10}$$

But $y = \mu + \sigma z$ so,

$$\begin{aligned} \text{logit}\pi_i(z) &= a_{0i}^* + a_{1i}^*(\mu + \sigma z) & (3.11) \\ &= a_{0i}^* + a_{1i}^* \mu + a_{1i}^* \sigma z \end{aligned}$$

Then,

$$a_{1i}^* = a_{1i}/\sigma \tag{3.12}$$

and

$$\begin{aligned} a_{0i}^* &= a_{0i} - a_{1i}^*\mu \\ &= a_{0i} - a_{1i}\mu/\sigma \end{aligned} \tag{3.13}$$

Therefore, if $\mu \geq 0$ then $a_{0i}^* \leq a_{0i}$ and if $\sigma > 1$ then $a_{1i}^* \leq a_{1i}$.

Let us see how changing the underlying distribution or the prior will affect the parameter estimates. If we multiply the standard deviation of the underlying distribution σ by a positive number, r, and fit a N(0,1) prior, the new parameter estimates will be multiplied by r. On the other hand, if we use as prior a distribution with standard deviation multiplied by r, then the new discrimination parameters will be r times smaller.

If we increase the mean of the underlying distribution (keeping its variance constant) and fit a N(0,1), then the new parameter estimates for the difficulty parameter will be larger by a_{1i} times the difference in the means. If we increase the mean of the prior though, keeping the variance constant, the new a_{0i}^*'s will be smaller by the estimated a_{1i} times the difference in the means.

3.2 Rates of Change for Priors - The Influence Function

In this section we will apply the influence function to measure the sensitivity of the parameters to the prior distribution.

Changes in the underlying distribution will lead to changes in the distribution of the response patterns, and so the rates of change in the parameters that result will be a weighted average of the values of the influence function at the different response patterns. We consider changes from the standard normal prior to a mixture of that with a little extra probability at z_o.

Suppose we distort the prior by adding a very small amount of probability, ϵ, on a point. Then the joint distribution of the x's will be given by:

$$f_\epsilon(\mathrm{x}) = \int g(\mathrm{x} \mid z)((1-\epsilon)h(z) + \epsilon\delta_{z_o})dz = (1-\epsilon)f(\mathrm{x}) + \epsilon f(\mathrm{x} \mid z_o) \tag{3.14}$$

This is close to $f(\mathrm{x})$ so using (1.39) we have

$$a_{(under f_\epsilon)} = a_{(under f)} + \epsilon \int \mathrm{IF}(\mathrm{x}, a)(f(\mathrm{x} \mid z_o) - f(\mathrm{x}))d\mathrm{x} \tag{3.15}$$

Since

$$\int \mathrm{IF}(\mathrm{x}, a)f(\mathrm{x})d\mathrm{x} = 0$$

the rate of change of a is given by

$$\int \mathrm{IF}(\mathrm{x}, a)f(\mathrm{x} \mid z_o)d\mathrm{x}, \tag{3.16}$$

or

$$\sum_{\mathbf{x}} \mathrm{IF}(\mathrm{x}, a)f(\mathrm{x} \mid z_o), \tag{3.17}$$

i.e. the rate of change is given by weighting the influence function at each $\mathbf{x}$ with the conditional distribution of $\mathbf{x}$ at the contaminated point z_o and averaging over $\mathbf{x}$.

From the above we see that the distortion on a point z_o is incorporated in $f(\mathbf{x})$. Thus we are measuring the distortions in the prior indirectly by measuring the sensitivity of the estimator to distortions in the joint distribution $f(\mathbf{x})$ which are due to distortions in the underlying distribution. A distortion on a point of the underlying distribution z_o will affect $f(\mathbf{x})$ and not just a point of the $\mathbf{x}$'s. By averaging over the rates of change of the estimator for all possible $\mathbf{x}$'s, we will obtain the rate of change due to a distortion in the underlying distribution of the latent variable.

3.3 Influence Function Results

We calculated the influence function for a model with parameter values set to MLE for a logit/probit model fit to the first four items of the Schuessler Social Life Feelings Scales 7, to all six items of Scale 7 and to the employment data. The employment data are the responses of 1915 individuals to 4 items regarding attitude to employment taken from Albanese and Knott (1992b) and Birkhoff (1991). The items are given in the Appendix.

We used a N(0,1) for the prior and it was approximated by a set of 16 quadrature points and weights. We calculated the influence function and the rate of change of each parameter on each quadrature point.

The parameter values for the subset of scale 7, the whole scale 7 and the employment data are in Tables 2.1, 2.3 and 3.1 respectively.

Table 3.1: Parameter values for model, employment data

i	a_{0i}	a_{1i}
1	0.87	0.97
2	0.81	1.97
3	1.43	1.83
4	0.86	0.27

Rates of change of the difficulty parameter Figures 3.1, 3.2 and 3.3 show the rate of change in the difficulty parameter estimates due to distortions in $f(\mathbf{x})$ which are due to having a small extra amount of probability at each quadrature point of the underlying distribution.

The IF of a_{0i}, for all i and all three datasets, does not seem very large. It is 's' shaped for most parameters, as it is generally negative when some extra probability is placed on a negative z and positive when some extra probability is placed on a positive z. We saw why this happens in Section 3.1.1. The changes are due to the small change in the location of the underlying distribution as we move along the range of z. As more probability is placed on the negative side the number of more 'negative' response patterns, i.e. patterns with a lot of zeroes, increases. The model tries to adjust to that by making the items more 'difficult', or a_{0i} smaller.

Not all parameters follow this pattern though. For example, a_{01} of scale 7, either within the four or six item set, has positive IF if extra probability is placed on the far end of the negative side of the latent variable distribution. Also, for some parameters the largest negative value does not occur in the extremes of the distribution but in one of the quadrature points closer to the middle. For example, a_{01} and a_{06} have their minimum IF

Figure 3.1: Influence Function for a_{0i} for changes in the underlying distribution, scale 7, first 4 items

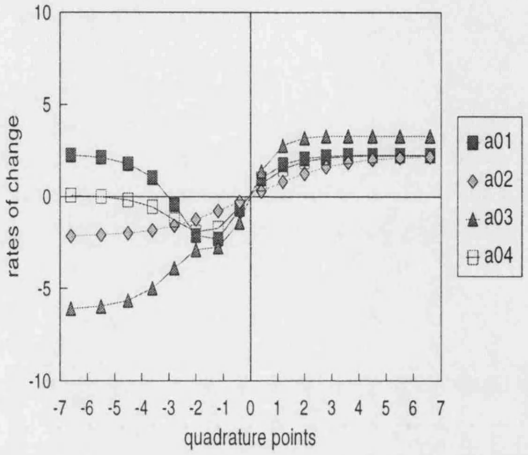


Figure 3.2: Influence Function for a_{0i} for changes in the underlying distribution, scale 7

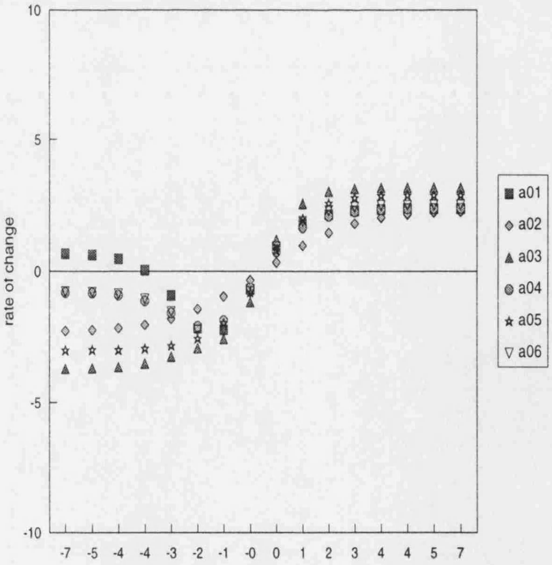
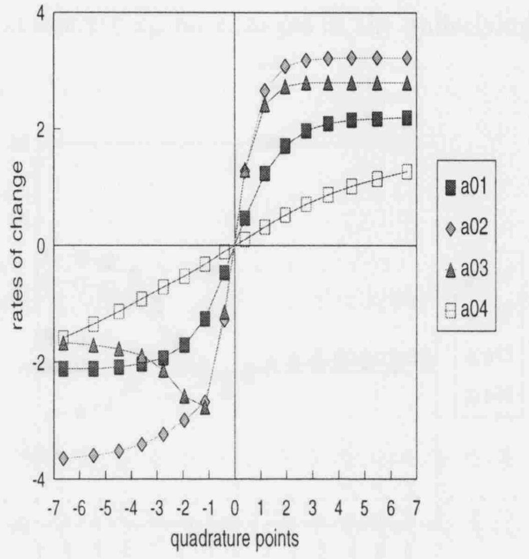


Figure 3.3: Influence Function for a_{0i} for changes in the underlying distribution, employment data



at -2.0. So there other factors that contribute to the way parameters change other than the location of the distribution.

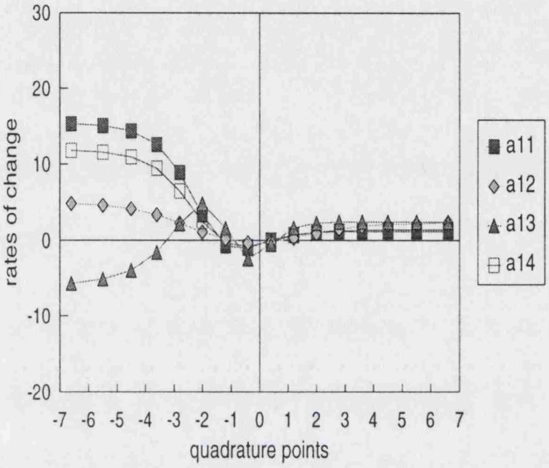
We also observe for both the 4 and the 6 item set of scale 7, the rates of change are larger and more variable for different items in the negative range of z, than in its positive range. This is attributed to the skewness of the distribution of the data. There are fewer 'negative' response patterns, patterns with a lot of zeroes than patterns with a lot of ones, so a change in the negative side of the underlying distribution, which will affect the distribution of the 'negative' response patterns will have a larger effect on the parameters.

For the employment data the IF is of the same order of magnitude whether changes occur in the positive or negative side of the distribution. Again we observe that a_{03} has its minimum negative IF not at the negative extremes of the distribution but at -2.0.

The rates of change seem to 'level off' in the extremes of z. As we move to the extremes of z the rate of change becomes stable, does not increase along with (the absolute value of) z. It seems the 'location' of the distribution that affects the parameters has more to do with where the main part of the distribution lies.

Rates of change of the discrimination parameter The rates of change of a_{1i} due to distortions in the distribution of the latent variable are shown Figures 3.4, 3.5 and 3.6.

Figure 3.4: Influence Function for a_{1i} for changes in the underlying distribution, scale 7, first 4 items



The rates of change of a_{1i} are negative on the central quadrature points. As extra probability is placed on one of the central quadrature points the underlying distribution shrinks, its variance decreases. The response patterns generated from such a distribution reflect this and thus the items appear less discriminatory.

The rates of change of the discrimination parameters are generally larger than the rates of change of the difficulty parameters. The first four items of scale 7 have similar rates of change, whether they are considered alone or part of the whole scale. The rates of change of the discrimination parameters of the employment data are smaller than the rates of change of the first 4 items of scale 7.

The rates of change of a_{1i} also level off for extreme values of z. This indicates that the parameters are not increasingly affected by small changes in the prior at the extremes underlying distribution, even if these progressively move out of the range of the main part of the distribution.

Figure 3.5: Influence Function for a_{1i} for changes in the underlying distribution, scale 7

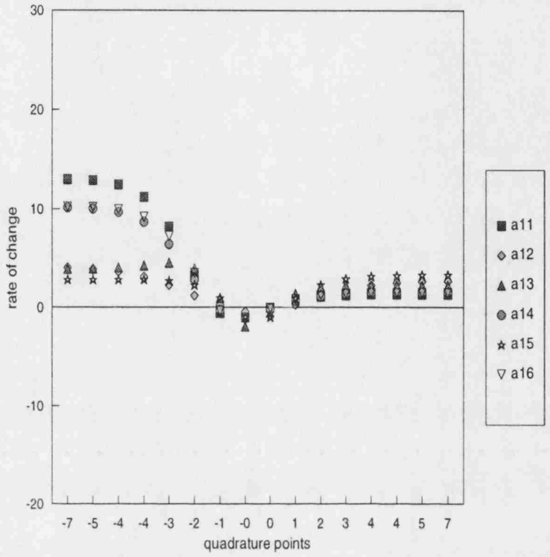
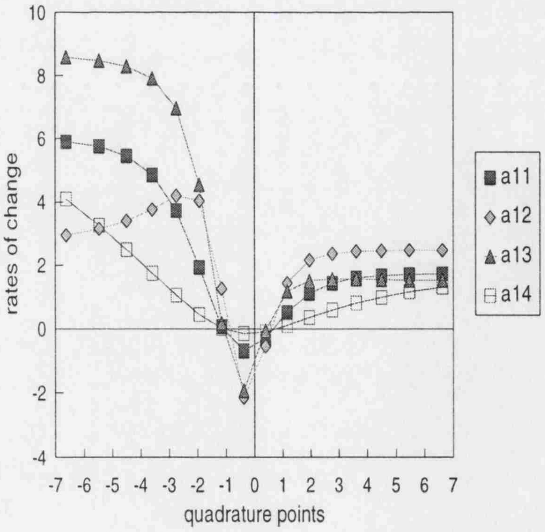


Figure 3.6: Influence Function of a_{1i} for changes in the underlying distribution, employment data



3.4 Empirical results

We fitted the latent trait model to the above data using a 'contaminated' prior, similar to the prior used for the Influence Function. So the contaminated prior is a mixture of the standard normal plus some probability added to a quadrature point. We used 0.01 and 0.03 probability to contaminate the prior in the above way.

We calculated the empirical rates of change in order to check the results from the Influence Function. The empirical rate of change or 'standardised' change is given by the difference of the 'new' parameter estimate from the one obtained with a N(0,1) as prior divided by the amount of change, i.e. 0.01 or 0.03. These standardised changes were calculated on all 16 quadrature points.

We expect from Section 3.1.1 the actual rates of change to have the opposite signs of the IF, since the IF corresponds to changes in the underlying distribution whereas for the actual changes we have modified the prior, the distribution used to fit the model.

We will also examine how the new estimates relate to the confidence intervals of the estimates from the N(0,1) prior, to get a feeling of the magnitude of the actual changes in the parameters.

Difficulty parameter In Graphs 3.7 and 3.8 the standardised changes of a_{0i} of Scale 7 and the employment data are plotted against the quadrature point which carries the extra weight.

We observe that rates of change of a_{0i} for scale 7 and the employment data obtained from the influence function and the contaminated prior have very similar shapes and similar range of values, except for the different signs. The empirical rates of change of a_{0i} are generally positive when the extra probability is added on the negative side of the prior distribution and vice versa.

We shall now look at the actual parameter values to get a feeling of their changes with the amounts of contamination we considered above. We will denote the estimates obtained from the contaminated priors with a_{ji}^{*}.

In Figure 3.9 we have the empirical rates of change and the actual values of the difficulty parameters for the first 4 items of scale 7, when we place 0.03 extra probability at each quadrature point. We have also drawn the 95% confidence interval lines for the

Figure 3.7: Empirical rates of change for a_{0i} for changes in the prior, scale 7

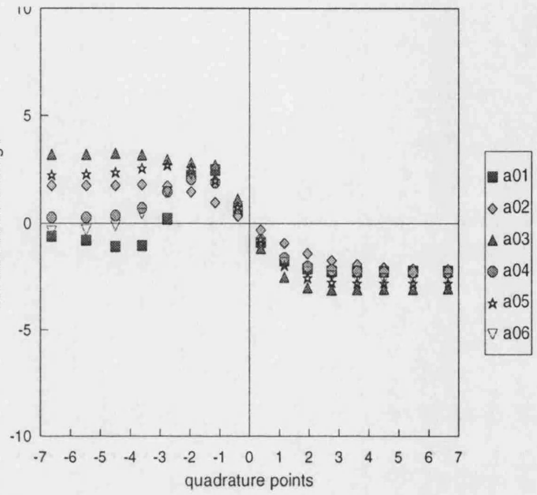
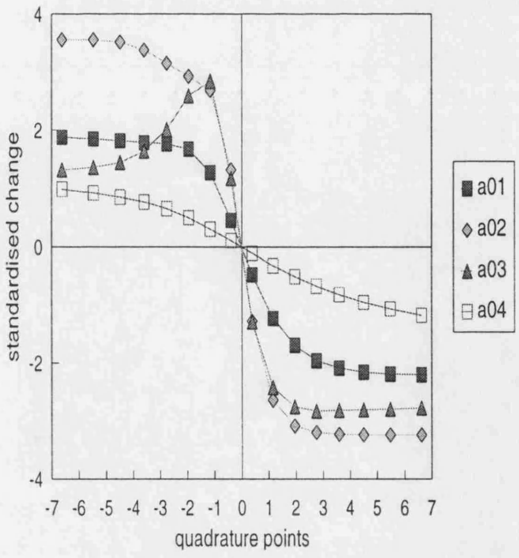


Figure 3.8: Empirical rates of change for a_{0i} for changes in the prior, employment data



parameters in order to see whether the new parameter estimates would fall within the confidence intervals of the parameters from the N(0,1) prior.

The empirical rates of change for a_{0i} are very similar to IF, except for the rate of change of a_{03} at -5.5 which is much larger. It seems that 0.03 is a larger distortion than the IF can predict for, which means that the change of the parameter is not proportional to the amount of contamination any more. The actual parameter values are generally well behaved and remain within the confidence intervals of the original parameters. The single exception is a_{03}^* which goes up to 1.3 from 0.87 with the N(0,1) prior, when 0.03 extra probability is placed at -5.5.

Figure 3.10 shows a_{02}^*, which has the largest rate of change of the difficulty parameter of the employment data, as we place extra probability on each quadrature point of the prior. We see that the new parameter values are well within the confidence interval of the parameter obtained with a N(0,1) prior.

We also examine a slightly larger amount of contamination. In Figure 3.11 we have the estimates of a_{01}^* as we increase the extra probability at a point from 0.01, to 0.03 and 0.05. All parameters remain within the confidence intervals of the ones from the N(0,1) prior.

Discrimination parameter The absolute values of the Influence Function and the empirical rates of change are very close for scale 7, both for the 4 and 6 item sets, and the employment data. The exception to that is the actual rate of change a_{13} of the 4 item Scale 7 data, at quadrature point -5.5, which is much larger than the IF. As with a_{03}, 0.03 contamination causes changes in the parameter much larger than the IF would predict, so the changes are not proportional to the amount of contamination, as we would expect them to be in a 'neighbourhood' of the model. For the rest of the items the Influence Function gives generally a good approximation of the rates of change we should expect when the data come from a 'contaminated' prior or they are fitted with one.

The actual rates of change of the parameters of the first 4 items of scale 7 are shown in Figure 3.14, together with the estimated discrimination parameters. In the graphs with the parameters we have also drawn the 95% confidence interval bands of the parameter estimates with the N(0,1) prior.

Looking at the discrimination parameters of the first 4 items of scale 7, we see that

Figure 3.9: Empirical rates of change for a_{0i} and parameter estimates for changes in the prior, Scale 7, first 4 items

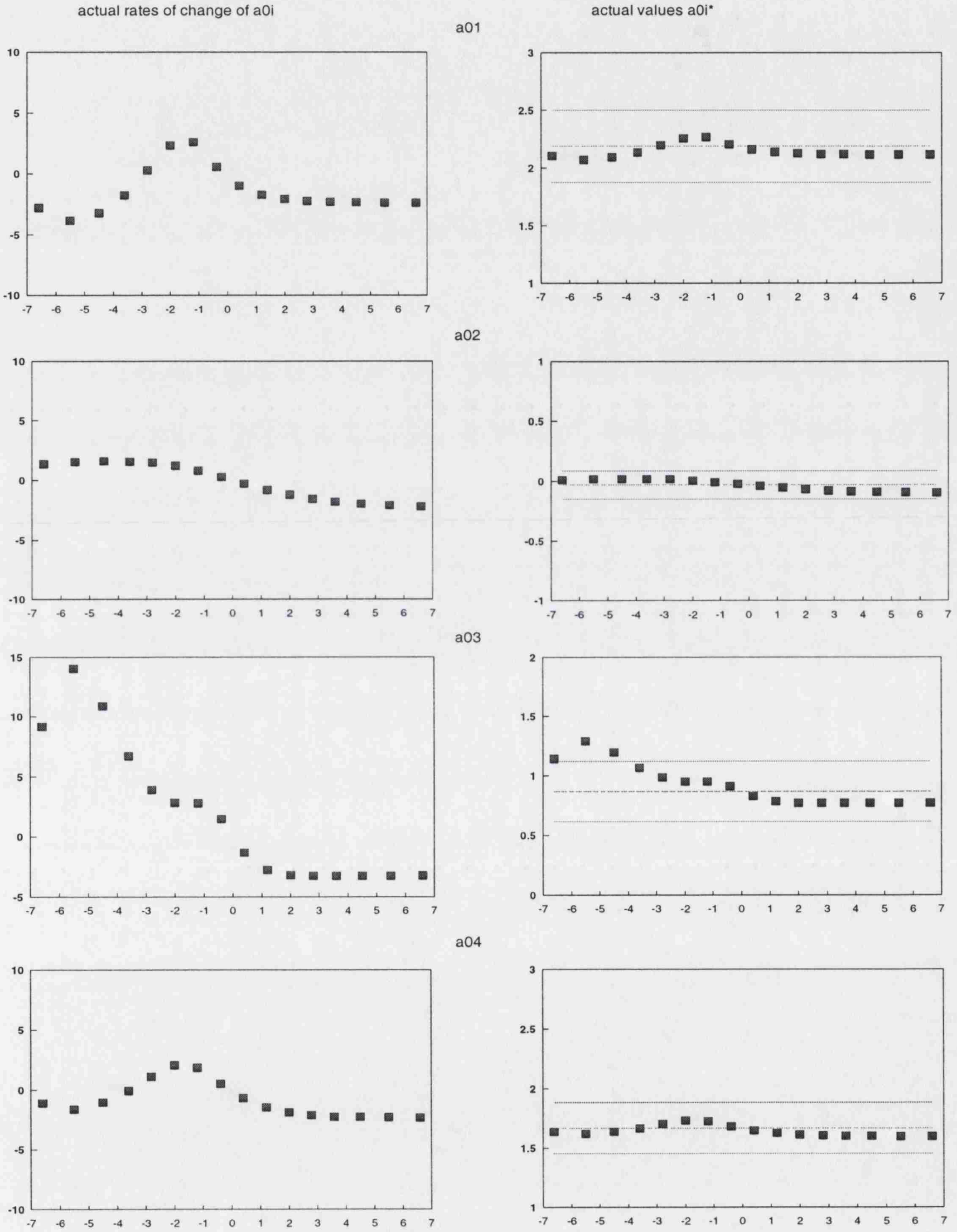


Figure 3.10: Employment data, a_{02}^* from contaminated prior (0.03 probability)

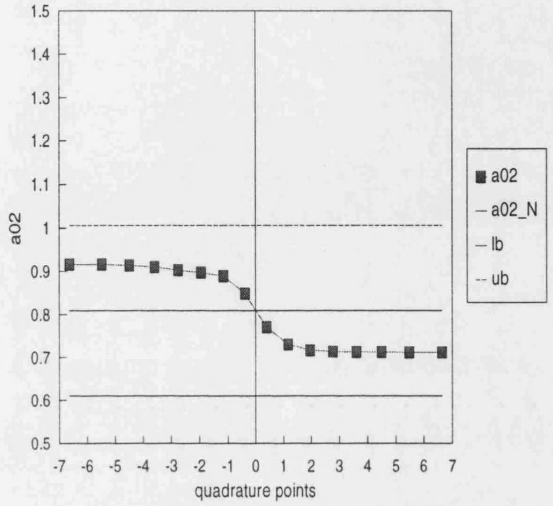


Figure 3.11: Scale 7, a_{01}^* as each quadrature point gets 0.01, 0.03 and 0.05 extra probability

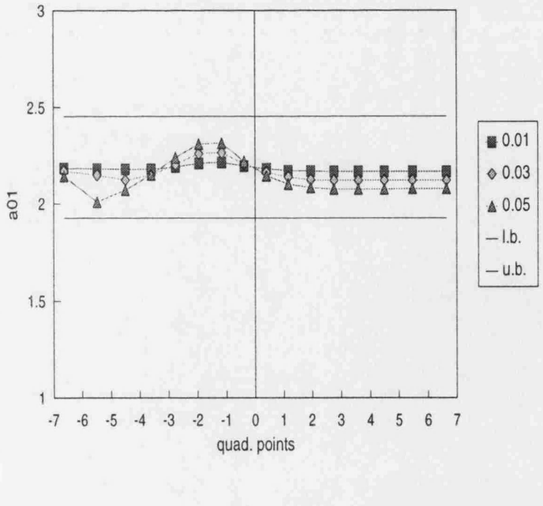


Figure 3.12: Empirical rates of change for a_{1i} for changes in the prior, scale 7

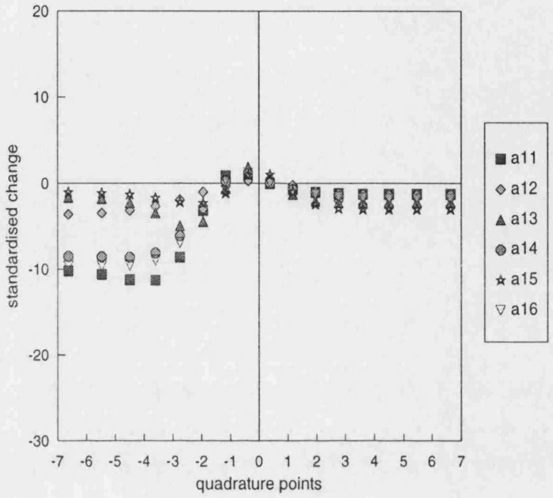


Figure 3.13: Empirical rates of change for a_{1i} for changes in the prior, employment data

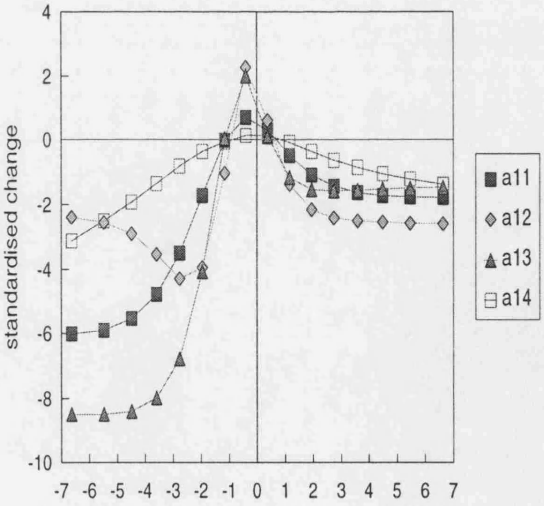
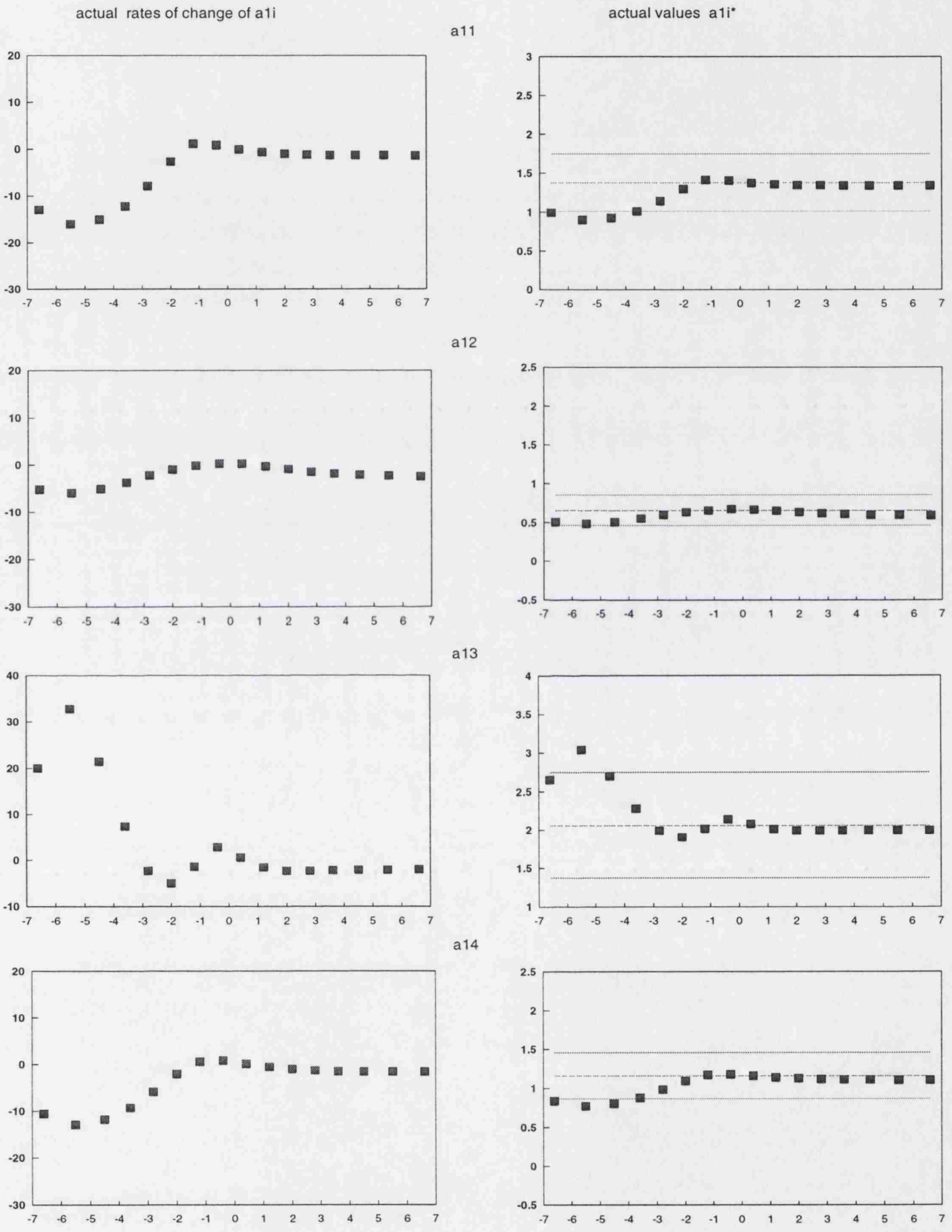


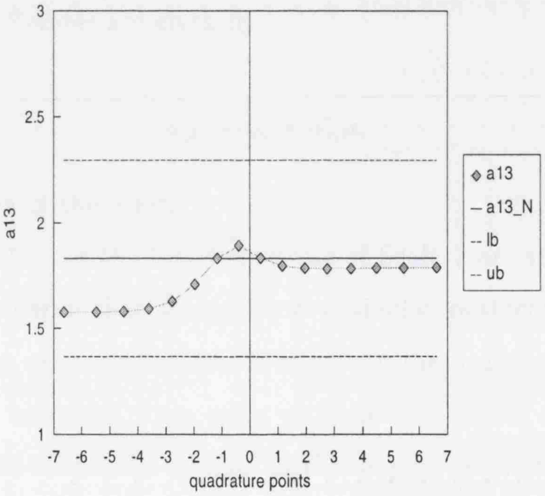
Figure 3.14: Empirical rates of change for a_{1i} and parameter estimates for changes in the prior, Scale 7, first 4 items



the parameters remain within the confidence interval bands when 3% extra probability is put on most of the quadrature points. a_{11}^* and a_{14}^* lie just outside the bands of their confidence intervals when extra probability is placed at -5.5 whereas a_{12}^* lies just on the bound of the interval. a_{13}^* definitely exceeds the bound at the 5.5 and lies just within the upper bound when extra probability is placed at -6.6 and -4.5.

The discrimination parameters of the employment data seem better behaved. Figure 3.15 shows a_{13}^* for the employment data as we place extra probability on each quadrature point. a_{13} is the discrimination parameter with the largest absolute rates of change but with 0.03 contamination in the prior it stays within the confidence interval of a_{13} estimated with a standard normal.

Figure 3.15: Employment data, a_{13}^* from contaminated prior (0.03 probability)



We will now also look at the discrimination parameters of the six item scale 7, with 0.01, 0.03 and 0.05 contamination.

In Figure 3.16 we see the estimates a_{11}^*, a_{13}^* and a_{14}^* of scale 7 obtained as we increased the amount of extra probability on each quadrature point.

We see that a_{13}^* remains within the confidence intervals of the initial estimates even with 0.05 extra probability at the quadrature points, whereas a_{11}^* and a_{14}^* reach the border of their initial confidence interval with 0.03 contamination and get outside their confidence interval with 0.05 contamination at quadrature points -5.5, -4.5 and -3.6.

The rest of the parameter estimates remain within their confidence intervals with 0.05

contamination on the quadrature points.

3.5 Standardised changes of the parameters

As the changes in the parameters we observed in Section 3.4 are largely due to the different location and scale of the contaminated prior from the N(0,1), in the following we will standardise the new parameter estimates obtained from the contaminated prior and see how far the standardised estimates are from the original estimates obtained with a N(0,1) prior.

Difficulty Parameter The estimates obtained from the contaminated prior are denoted with a^*_{0i}. We will first standardise a^*_{0i} using (3.13), in order to get the estimates that we would have got had we used a prior with mean equal to 0 and standard deviation equal to 1. The new estimates are given by

$$a'_{0i} = a^*_{0i} + a^*_{1i}\mu, \tag{3.18}$$

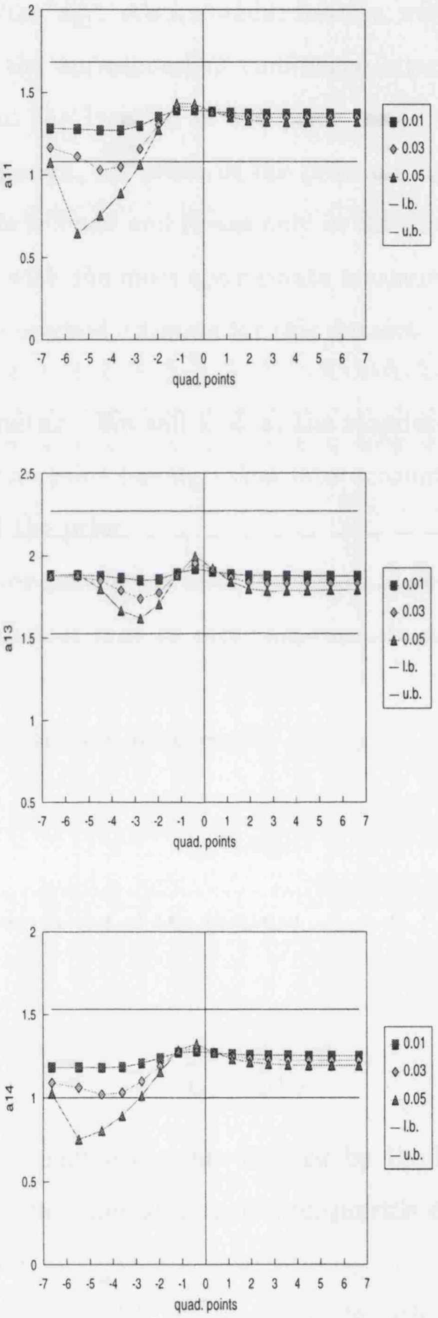
where μ is the mean of the prior.

The standardised a^*_{0i}'s for the first four items of Scale 7 are shown in the left panel of Figure 3.17. We observe that all a'_{0i} show a similar pattern: they are below the original estimates when extra probability is added on the negative range of z and they rise above the original estimates when extra probability is added on the positive side of z. a'_{03} is now within the confidence interval of the original estimate at $z = -5.5$, but on three of the extreme quadrature points the standardised estimates reach the confidence interval bounds of the original estimates and at 6.6 a'_{03} exceeds the upper bound. It seems that standardising in this way 'over-standardises' the parameter estimates. As the prior distribution is very skewed when extra probability is added on the extremes of the prior, the mean is not an appropriate measure of location of the prior. As alternative measure of location we will use the median. The new estimates are given by

$$a''_{0i} = a^*_{0i} + a^*_{1i}\text{median} \tag{3.19}$$

The mean and median of the distributions considered are given in Table 3.5. The standardised estimates from (3.19) are shown in the right panel of Figure 3.17. We

Figure 3.16: Scale 7, a^*_{11}, a^*_{13} and a^*_{14} as each quadrature point gets 0.01, 0.03 and 0.05 extra probability



observe that in most cases these standardised estimates are closer to the original ones than the unstandardised or the standardised with the mean estimates, and in most cases the differences between them and the original ones are negligible. Standardisation did not work for a_{01}^*, in the negative range of z, as it showed counter-expected behaviour, but $a_{0i}^{''}$ is closer to a_{0i} than $a_{01}^{'}$. Also, standardisation with the median did not bring $a_{03}^{''}$ at $z = -5.5$ within the corresponding confidence interval but the large change of a_{03} at -5.5 is only due to the location of the prior, since the change of a_{03} at -6.6 is smaller. On the whole though, the effect of the prior on the difficulty parameter apart from its location and scale is small and it was only in this single case that a standardised parameter (standardised with the most appropriate measure of location) fell outside the confidence interval of the original estimate for this dataset.

Discrimination Parameter We will look at the standardised discrimination parameter to see the effect of the prior having taken into account changes in the parameters attributed to the scale of the prior.

To standardise the prior we will use both the standard deviation and the interquartile ratio of the prior, as the latter may be more appropriate measure of dispersion for the distributions we consider.

So, the standardised estimates are given by

$$a_{1i}^{'} = a_{1i}^{*}\sigma \tag{3.20}$$

where σ is the standard deviation of the mixture

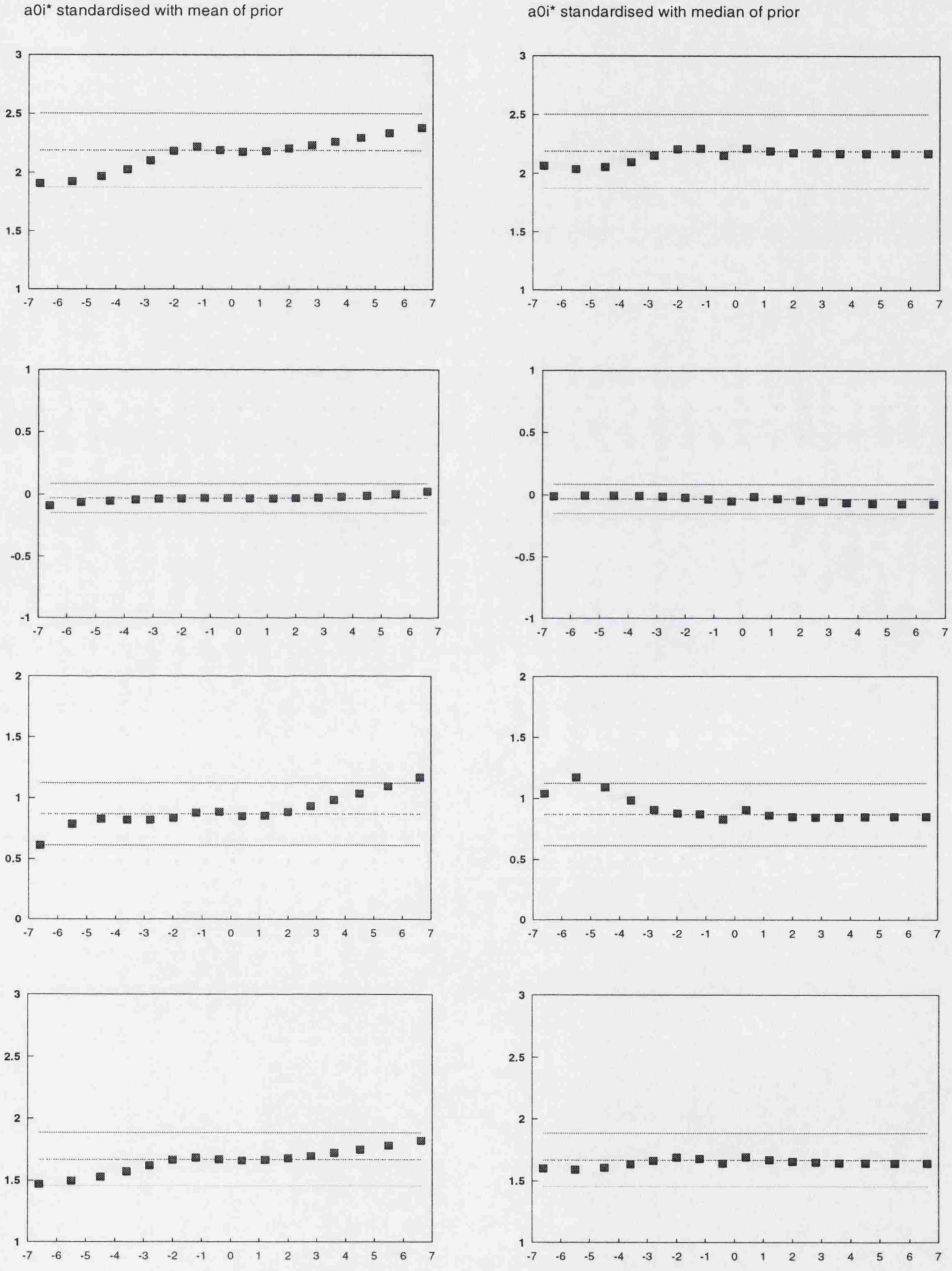
and

$$a_{1i}^{''} = a_{1i}^{*}\text{iqr} \tag{3.21}$$

where 'iqr' is the interquartile distance divided by the interquartile distance of the N(0,1) (=1.349). The standard deviation and interquartile distance of the priors considered are given in Table 3.5.

The estimates standardised with the standard deviation for the first four items of Scale 7 are given in the left panel of Figure 3.18. We see that standardisation works well in the middle range of z, though this is only noticeable for a_{13}^* as the unstandardised

Figure 3.17: Standardised a^*_{0i} with the mean and the median of the prior, Scale 7, first 4 items



estimates are also almost indistinguishable from the original ones in the middle range of z. In the negative range of z, where there were the largest discrepancies between the original and the unstandardised estimates, standardisation brought a^*_{11} and a^*_{14} within the confidence interval of the original estimates. Since a^*_{13} increased instead of decreasing compared to a_{13}, standardisation brought a'_{13} further away from the original estimate.

Regarding the positive range of z, all a^*_{1i}'s were very close to the original estimates. Standardisation with the standard deviation of the prior 'over-standardised' the estimates so that the standardised estimates got further away from the original ones at the outer points of z, sometimes even exceeding the corresponding confidence interval.

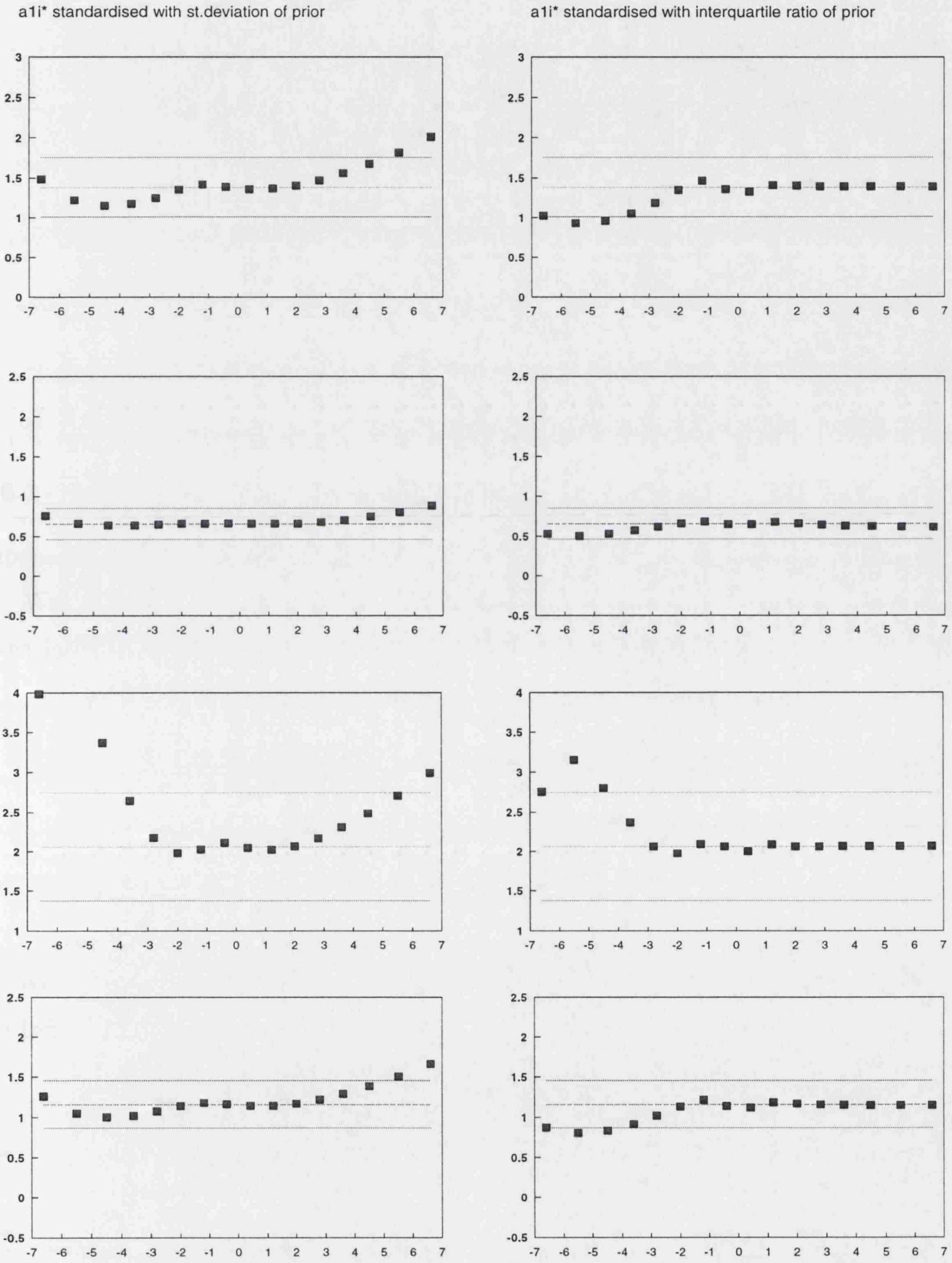
The estimates standardised with the interquartile ratio are shown in the right panel of Figure 3.18. Although a''_{11} and a''_{14} are still outside the confidence intervals of the original estimates for some z_0, more a''_{1i}'s than a'_{1i}'s are closer to the original estimates a_{1i}'s. We feel that the interquartile ratio is a more appropriate measure of scale for the priors we considered.

Overall, the estimates standardised with the interquartile ratio are very close to the original estimates from the N(0,1), in most cases indistinguishable from them. In the negative range of z, discrepancies from the original estimates are more obvious, and some of the standardised parameters fell outside the confidence intervals of the original estimates. Changes are attributed to other characteristics than the dispersion of the prior which affect these data, as the changes are larger at $z = -5.5$ than at $z = -6.6$ and also a^*_{13} gets larger instead of smaller when extra probability is put on those points.

3.6 Sensitivity of the posterior means to the prior distribution

The parameter changes showed some relatively large differences at some points of the prior distribution, similar to the changes we have seen when extra probability was placed on a response pattern. In the following we examine how contaminating the prior by placing extra probability on a point may affect the posterior means. We will first derive the Influence Function of the posterior means for changes in the prior distribution. We will then compare the IF and the empirical rates of change of the posterior means for a particular dataset.

Figure 3.18: Standardised a_{1i}^* with the standard deviation and the interquartile ratio of the prior, Scale 7, first 4 items



a1i* standardised with st.deviation of prior

a1i* standardised with interquartile ratio of prior

Table 3.2: Statistics of the mixture 0.97 N(0,1) + 0.03 at z_0

z_0	mean	median	st.dev.	intq.dist.
-6.63	-0.1989	-0.03855	1.4998	1.3994
-5.47	-0.1642	-0.03855	1.3569	1.3994
-4.49	-0.1348	-0.03855	1.2479	1.3994
-3.60	-0.1080	-0.03855	1.1607	1.3994
-2.76	-0.0828	-0.03855	1.0916	1.3994
-1.95	-0.0586	-0.03855	1.0397	1.3994
-1.16	-0.0349	-0.03855	1.0047	1.3994
-0.39	-0.0116	-0.03855	0.9870	1.3004
0.39	0.0116	0.039	0.9870	1.3004
1.16	0.3491	0.039	1.0047	1.3994
1.95	0.0586	0.039	1.0397	1.3994
2.76	0.0828	0.039	1.0916	1.3994
3.60	0.1080	0.039	1.1607	1.3994
4.49	0.1348	0.039	1.2479	1.3994
5.47	0.1642	0.039	1.3569	1.3994
6.63	0.1989	0.039	1.4998	1.3994

3.6.1 Influence Function of the posterior means

Suppose we distort the prior by adding a very small amount of probability, ϵ, on a point z_0. We denote the posterior means estimated from such a prior $E_\epsilon(z|\mathrm{x})$ and they are given by:

$$\begin{aligned} E_\epsilon(z|\mathrm{x}) &= \frac{\int zg(\mathrm{x}\mid z)h_\epsilon(z)dz}{f_\epsilon(\mathrm{x})} \\ &= \frac{\int zg(\mathrm{x}\mid z)((1-\epsilon)h(z)+\epsilon\delta_{z_o})dz}{f_\epsilon(\mathrm{x})} \qquad (3.22) \\ & \qquad (3.23) \end{aligned}$$

We will derive the Influence Function of the posterior means using the definition of the Influence Function (1.38) given in Section 1.6.2:

$$\begin{aligned} \mathrm{IF} &= \lim_{\epsilon\to 0} \frac{E(z|\mathrm{x})_{\text{from } h_\epsilon} - E(z|\mathrm{x})_{\text{from } h}}{\epsilon} \\ &= \lim_{\epsilon\to 0} \frac{\int zg(\mathrm{x}\mid z)h_\epsilon(z)/f_\epsilon(\mathrm{x}) - \int zg(\mathrm{x}\mid z)h(z)dz/f(\mathrm{x})}{\epsilon} \qquad (3.24) \end{aligned}$$

The numerator of (3.24) becomes

$$
\begin{aligned}
& \int zg(\mathrm{x} \mid z)h_\epsilon(z)/f_\epsilon(\mathrm{x}) - \int zg(\mathrm{x} \mid z)h(z)dz/f(\mathrm{x}) \\
= & \frac{\int zg(\mathrm{x} \mid z)[(1-\epsilon)h(z) + \epsilon\delta_{z_0}]dz}{(1-\epsilon)f(\mathrm{x}) + \epsilon g(\mathrm{x}|z_0)} - \frac{\int zg(\mathrm{x}|z)h(z)dz}{f(\mathrm{x})} \\
= & \frac{\int zg(\mathrm{x} \mid z)(1-\epsilon)h(z)dz}{(1-\epsilon)f(\mathrm{x}) + \epsilon g(\mathrm{x}|z_0)} + \frac{\int zg(\mathrm{x} \mid z)\epsilon\delta_{z_0}dz}{(1-\epsilon)f(\mathrm{x}) + \epsilon g(\mathrm{x}|z_0)} - \frac{\int zg(\mathrm{x}|z)h(z)dz}{f(\mathrm{x})} \\
= & \frac{\int zg(\mathrm{x} \mid z)(1-\epsilon)h(z)f(\mathrm{x})dz}{[(1-\epsilon)f(\mathrm{x}) + \epsilon g(\mathrm{x}|z_0)]f(\mathrm{x})} + \frac{z_0 g(\mathrm{x} \mid z)\epsilon}{(1-\epsilon)f(\mathrm{x}) + \epsilon g(\mathrm{x}|z_0)} \\
& - \frac{\int zg(\mathrm{x}|z)h(z)[(1-\epsilon)f(\mathrm{x}) + \epsilon g(\mathrm{x}|z_0)]dz}{[(1-\epsilon)f(\mathrm{x}) + \epsilon g(\mathrm{x}|z_0)]f(\mathrm{x})} \\
= & \frac{-\int zg(\mathrm{x} \mid z)\epsilon g(\mathrm{x}|z_0)h(z)dz}{[(1-\epsilon)f(\mathrm{x}) + \epsilon g(\mathrm{x}|z_0)]f(\mathrm{x})} + \frac{z_0 g(\mathrm{x} \mid z)\epsilon}{(1-\epsilon)f(\mathrm{x}) + \epsilon g(\mathrm{x}|z_0)} \qquad (3.25)
\end{aligned}
$$

The ϵ's in the numerators of (3.25) cancel out with the ϵ in the denominator of (3.24) So,

$$
\mathrm{IF} = \lim_{\epsilon \to 0} \frac{-\int zg(\mathrm{x} \mid z)g(\mathrm{x}|z_0)h(z)dz}{[(1-\epsilon)f(\mathrm{x}) + \epsilon g(\mathrm{x}|z_0)]f(\mathrm{x})} + \frac{z_0 g(\mathrm{x} \mid z)}{(1-\epsilon)f(\mathrm{x}) + \epsilon g(\mathrm{x}|z_0)} \qquad (3.26)
$$

Evaluated at $\epsilon = 0$

$$
\begin{aligned}
\mathrm{IF} &= \frac{-g(\mathrm{x}|z_0)\int zg(\mathrm{x}|z)h(z)dz}{f^2(x)} + \frac{z_0 g(\mathrm{x}|z_0)}{f(\mathrm{x})} \\
&= -\frac{g(\mathrm{x}|z_0)E(z|\mathrm{x})}{f(\mathrm{x})} + \frac{z_0 g(\mathrm{x}|z_0)}{f(\mathrm{x})} \qquad (3.27)
\end{aligned}
$$

Scale 7, four items The Influence Function of the posterior means for the first 4 items of Scale 7 are given in Table 3.3. We observe that apart from a few very large values the IF is generally quite small and in many cases zero.

The IF is very large for the posterior means of 0000 when extra probability is placed on the quadrature points which are on the negative tail of the prior. The extra probability on the extremes of the negative range of the prior draws the posterior means of the response patterns with many zeroes, and particularly 0000, further to the left of the negative range of the posterior distribution.

(We note that the IF here gives the rate of change for changes in the prior used to fit

Table 3.3: Scale 7, first 4 items, Influence Function of E(z|x) for changes in the prior

	x							
q.point	0000	0001	0010	0011	0100	0101	0110	0111
-6.6	-138.5	-0.4	-0.0	0.0	-4.9	-0.0	0.0	0.0
-5.5	-103.8	-1.0	-0.0	0.0	-8.0	-0.1	-0.0	0.0
-4.5	-73.1	-2.3	-0.1	-0.0	-11.0	-0.2	-0.0	0.0
-3.6	-44.1	-4.2	-0.4	-0.0	-12.5	-0.8	-0.1	-0.0
-2.8	-18.6	-5.5	-1.3	-0.1	-10.1	-2.0	-0.4	-0.0
-2.0	-2.7	-3.6	-2.1	-0.6	-3.9	-2.5	-1.1	-0.2
-1.2	1.1	-0.1	-1.0	-1.0	0.5	-0.9	-1.2	-0.7
-0.4	0.4	0.7	0.5	-0.2	0.6	0.6	0.2	-0.6
0.4	0.0	0.2	0.4	0.6	0.1	0.3	0.6	0.5
1.2	0.0	0.0	0.1	0.4	0.0	0.1	0.2	0.8
2.0	0.0	0.0	0.0	0.2	0.0	0.0	0.1	0.6
2.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3
3.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
4.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00
5.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.05
6.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.06
min	-138.5	-5.5	-2.1	-1.0	-12.5	-2.5	-1.2	-0.7
25%	-25.0	-1.4	-0.2	-0.0	-5.6	-0.4	-0.0	-0.0
median	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
75%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2
max	1.1	0.7	0.5	0.6	0.6	0.6	0.6	0.8

the model, as the calculation of the posterior means involves the use of this prior. The sign of the actual rates of change will therefore be the same of the IF).

When extra probability is placed on the extreme positive quadrature points the rates of change of the posterior means with a lot of ones are positive, as these posterior means are pushed further to the right of their distribution.

Changes in the middle range of the prior affects the posterior means that are in the middle of their range, though the IF is then very small.

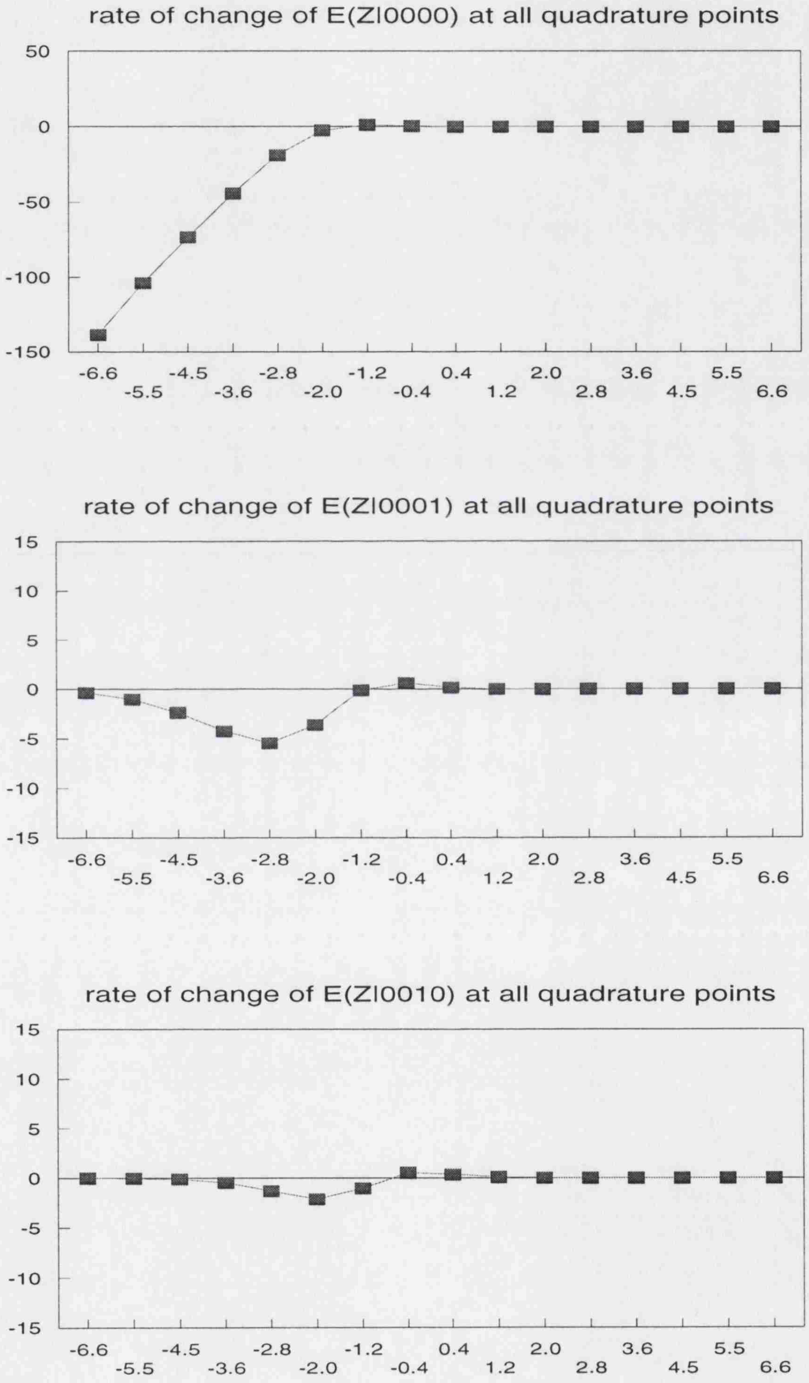
3.6.2 Empirical changes of the posterior means

Scale 7, four items We calculated the actual rates of change of the posterior means as 0.03 extra probability was placed on each response pattern in turn. The actual rates of change for the posterior means of the response patterns 0000, 0001, 0010, 0100, 1110 and 1111 are shown in Figures 3.21 and 3.22.

Table 3.3 cont.: Scale 7, first 4 items, Influence Function of E(z|x) for changes in the prior

	x							
q.point	1000	1001	1010	1011	1100	1101	1110	1111
-6.6	-0.1	0.0	0.0	0.0	-0.0	0.0	0.0	0.0
-5.5	-0.4	-0.0	0.0	0.0	-0.0	0.0	0.0	0.0
-4.5	-1.1	-0.0	0.0	0.0	-0.1	-0.0	0.0	0.0
-3.6	-2.5	-0.1	-0.0	0.0	-0.5	-0.0	-0.0	0.0
-2.8	-4.0	-0.5	-0.1	-0.0	-1.3	-0.1	-0.0	0.0
-2.0	-3.3	-1.3	-0.4	-0.0	-2.1	-0.6	-0.1	-0.0
-1.2	-0.4	-1.2	-0.9	-0.3	-1.0	-1.1	-0.6	-0.1
-0.4	0.7	0.3	-0.4	-0.7	0.5	-0.2	-0.7	-0.6
0.4	0.2	0.5	0.6	0.0	0.4	0.6	0.4	-0.5
1.2	0.0	0.2	0.5	1.1	0.1	0.4	0.9	0.8
2.0	0.0	0.0	0.3	1.6	0.0	0.2	0.8	3.0
2.8	0.0	0.0	0.1	1.6	0.0	0.0	0.5	5.7
3.6	0.0	0.0	0.0	1.3	0.0	0.0	0.3	8.8
4.5	0.0	0.0	0.0	1.0	0.0	0.0	0.1	12.0
5.5	0.0	0.0	0.0	0.7	0.0	0.0	0.1	15.5
6.6	0.0	0.0	0.0	0.4	0.0	0.0	0.0	19.6
min	-4.0	-1.3	-0.9	-0.7	-2.1	-1.1	-0.7	-0.6
25%	-0.6	-0.0	-0.0	-0.0	-0.2	-0.0	-0.0	-0.0
median	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
75%	0.0	0.0	0.0	1.0	0.0	0.0	0.3	6.5
max	0.7	0.5	0.6	1.6	0.5	0.6	0.9	19.6

Figure 3.19: Influence Function of the posterior means of 0000, 0001 and 0010 as extra probability is placed on each quadrature point

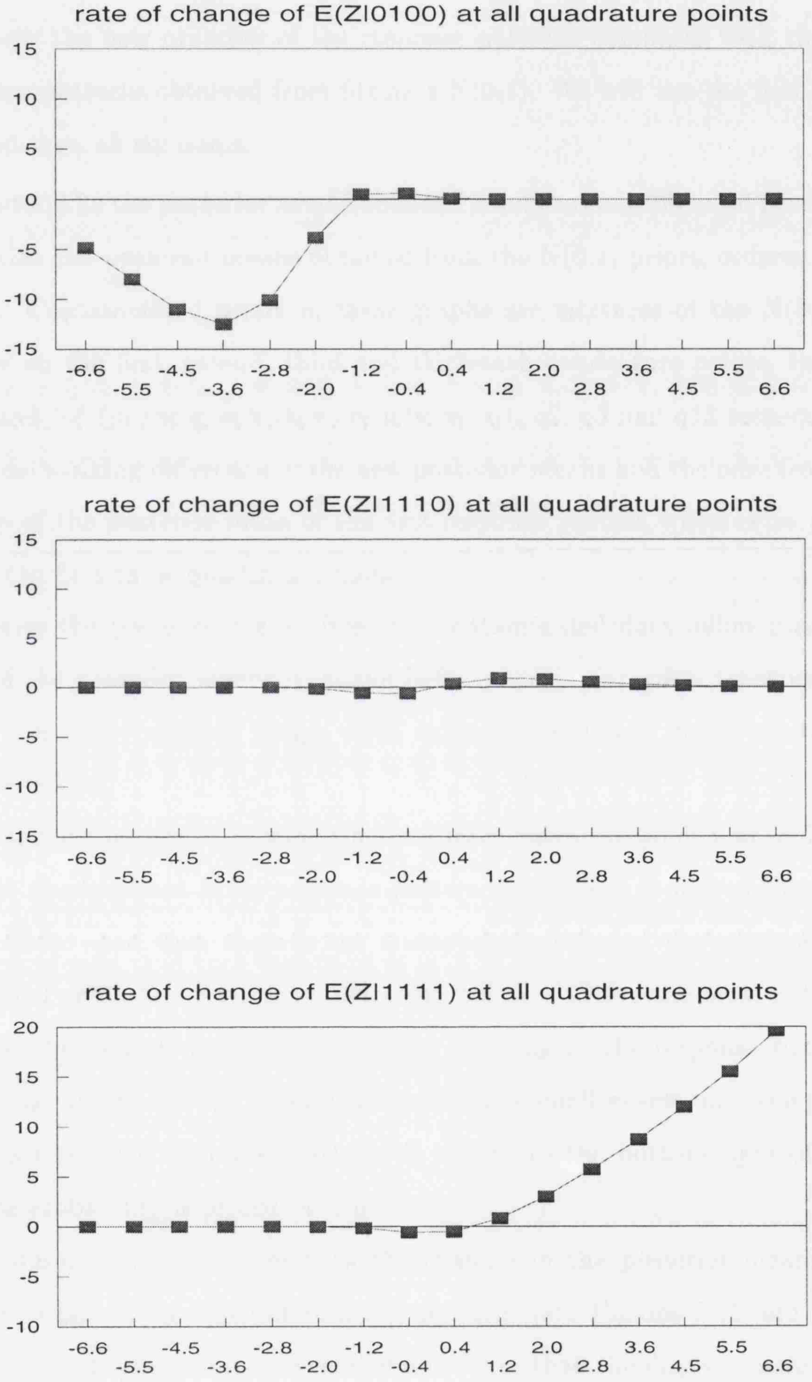


rate of change of E(Z|0000) at all quadrature points

rate of change of E(Z|0001) at all quadrature points

rate of change of E(Z|0010) at all quadrature points

Figure 3.20: Influence Function of some posterior means of 0100, 1110 and 1111 as extra probability is placed on each quadrature point


rate of change of E(Z|0100) at all quadrature points

rate of change of E(Z|1110) at all quadrature points

rate of change of E(Z|1111) at all quadrature points

The pattern and the range of the actual rates of change is very similar to the Influence Function. Although there are no rates of change equal to zero, the actual rates of change that correspond to IF=0 are very small.

Since it is the ordering of the response patterns that essentially matters, we will examine how the new ordering of the response patterns compares with the ordering of the response patterns obtained from fitting a N(0,1). We will use the first four items of Scale 7 and then all six items.

In Figures 3.23 the posterior means obtained from the contaminated priors are plotted together with the posterior means obtained from the N(0,1) priors, ordered according to the latter. Contaminated priors in these graphs are mixtures of the N(0,1) with 0.03 probability on the first, second, third and thirteenth quadrature points, that is on -6.6, -5.5, -4.5 and 3.6 (in the graph they are labelled q1, q2, q3 and q13 respectively).

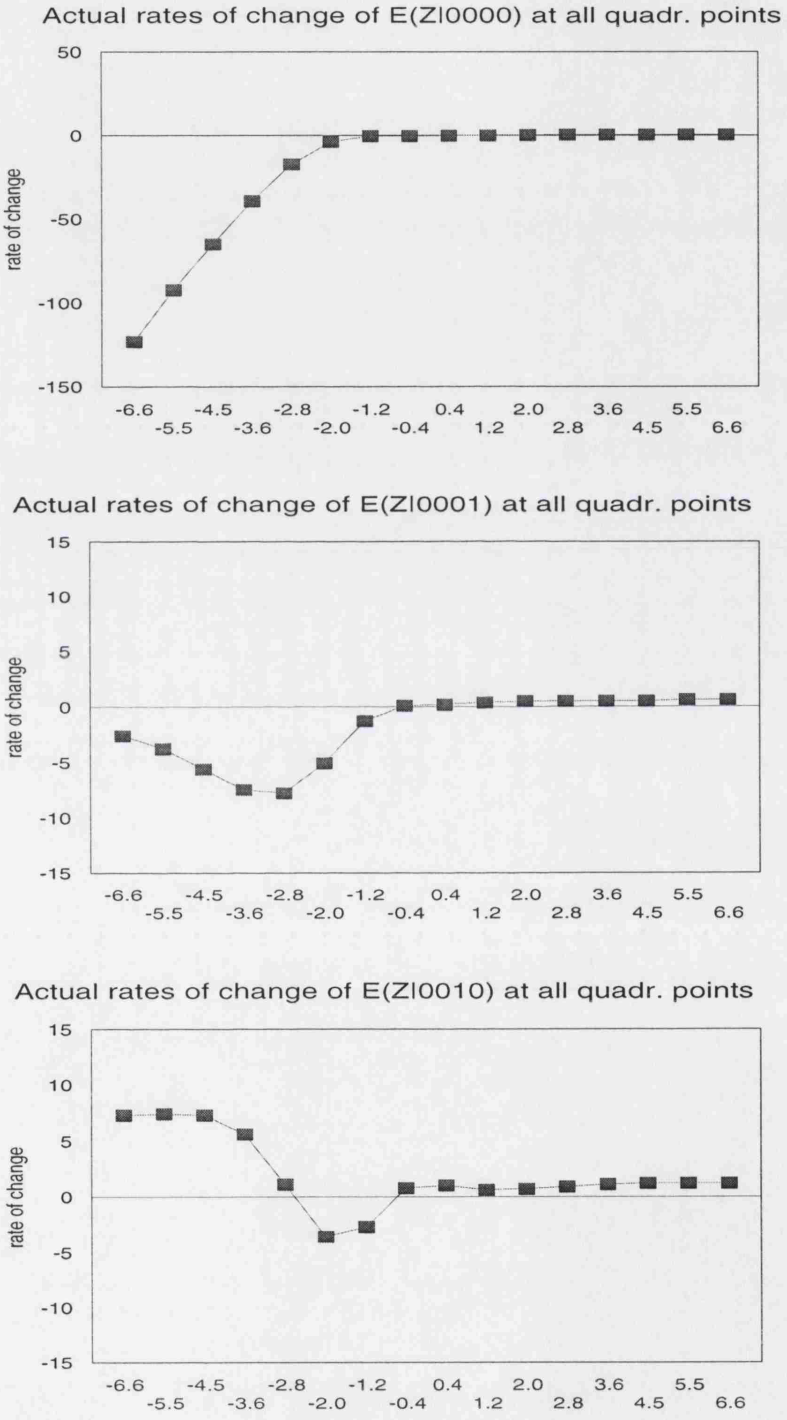
The most striking difference in the new posterior means and the ones from the N(0,1) is the drop of the posterior mean of the first response pattern when extra probability is placed on the first three quadrature points.

Otherwise the posterior means from the contaminated data follow more or less the ordering of the posterior means from the N(0,1), except for a few transpositions in the middle of the latent variable range, when extra probability is placed on the first three quadrature points.

The transpositions are the same when we place extra probability at -6.5, -5.5 or -4.5. The largest displacement is for response pattern 1101 which is moved then below 0011, 0110 and 0010. And then there is the transposition between the consecutive response patterns 0010 and 1001. At these quadrature points we have the most extreme rates of change and the largest displacements in the ordering of the response patterns. Extra probability on the rest of the quadrature points has small effects, and the picture of the 'new' against the 'old' response patterns is similar to the bottom right of Figure 3.23, when extra probability is placed on 3.6.

If we compare these changes with the changes in the posterior means when extra probability is placed on different response patterns (see Figures 2.15 and 2.16), we observe that these displacements are smaller and fewer than the displacements that happen when extra observations are placed at response patterns 0010, 0110 or 0111. In fact the transpositions in the middle of the latent variable range when extra probability is put on

Figure 3.21: Actual rates of change of the posterior means of 0000, 0001 and 0010 as extra probability is placed on each quadrature point

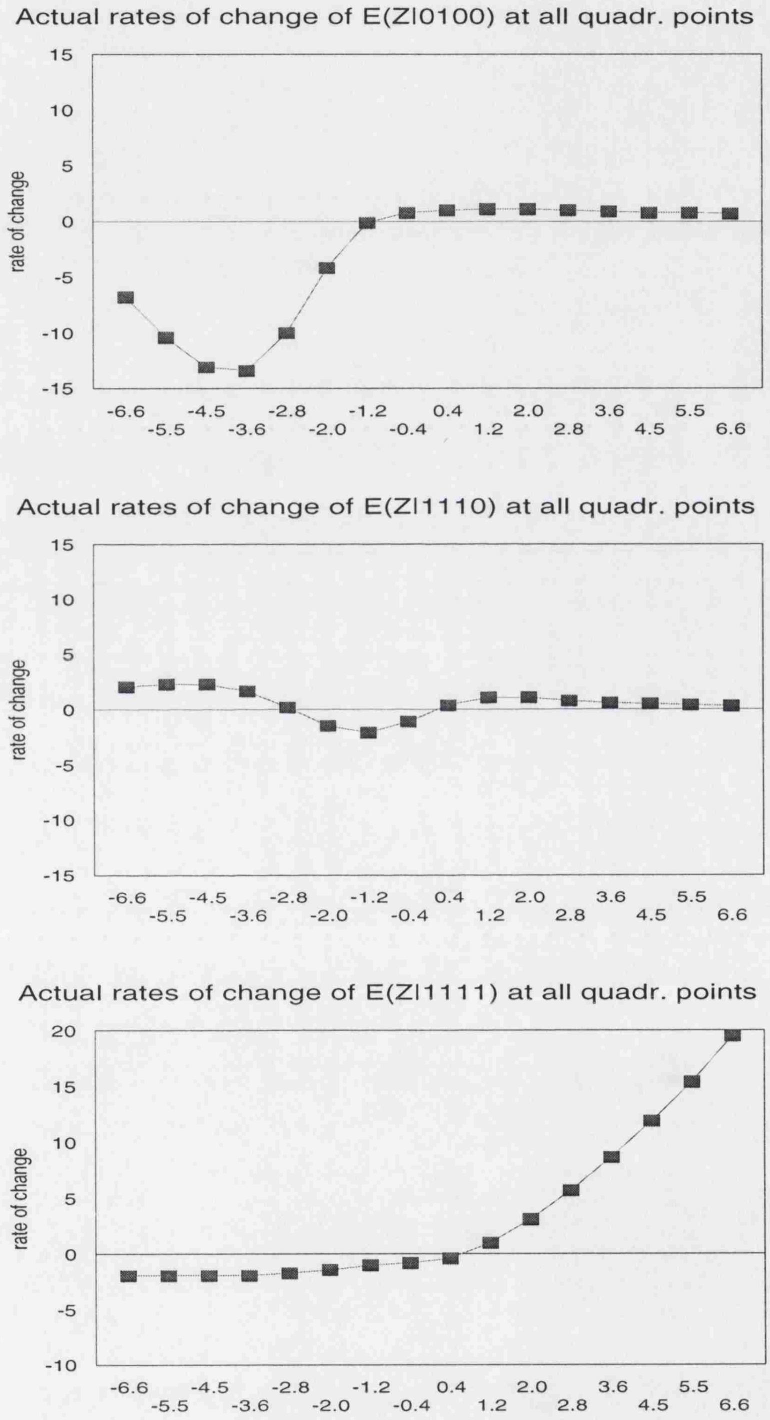


Actual rates of change of E(Z|0000) at all quadr. points

Actual rates of change of E(Z|0001) at all quadr. points

Actual rates of change of E(Z|0010) at all quadr. points

Figure 3.22: Actual rates of change of the posterior means of 0100, 1110 and 1111 as extra probability is placed on each quadrature point

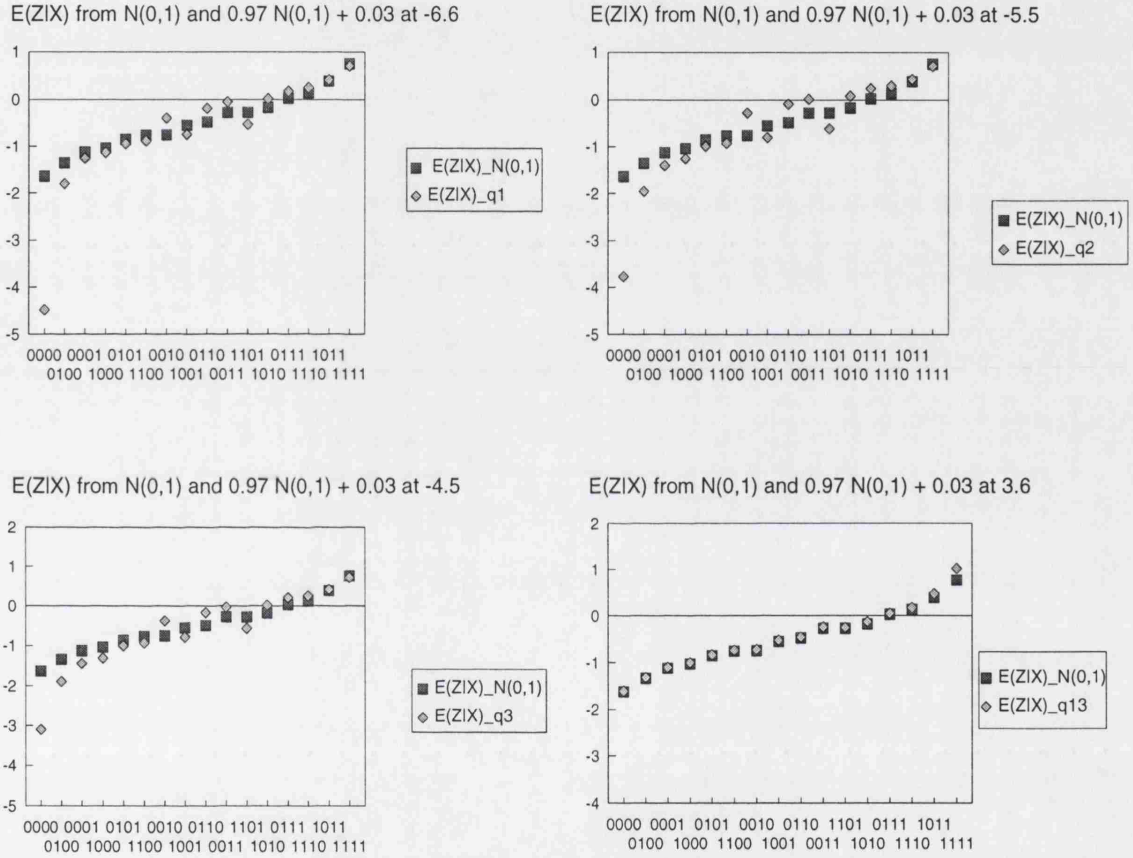


Actual rates of change of E(Z|0100) at all quadr. points

Actual rates of change of E(Z|1110) at all quadr. points

Actual rates of change of E(Z|1111) at all quadr. points the first three quadrature points are very similar to the transpositions that happen when extra frequency is placed on response pattern 0101. (Please note difference in the scale of the graphs, which is due to the large negative value of E(Z|0000) here).

Figure 3.23: Scale 7, 4 items, posterior means from N(0,1) and from contaminated priors with 0.03 extra probability, ordered according to posterior means from N(0,1)



E(ZIX) from N(0,1) and 0.97 N(0,1) + 0.03 at -6.6

E(ZIX) from N(0,1) and 0.97 N(0,1) + 0.03 at -5.5

E(ZIX) from N(0,1) and 0.97 N(0,1) + 0.03 at -4.5

E(ZIX) from N(0,1) and 0.97 N(0,1) + 0.03 at 3.6

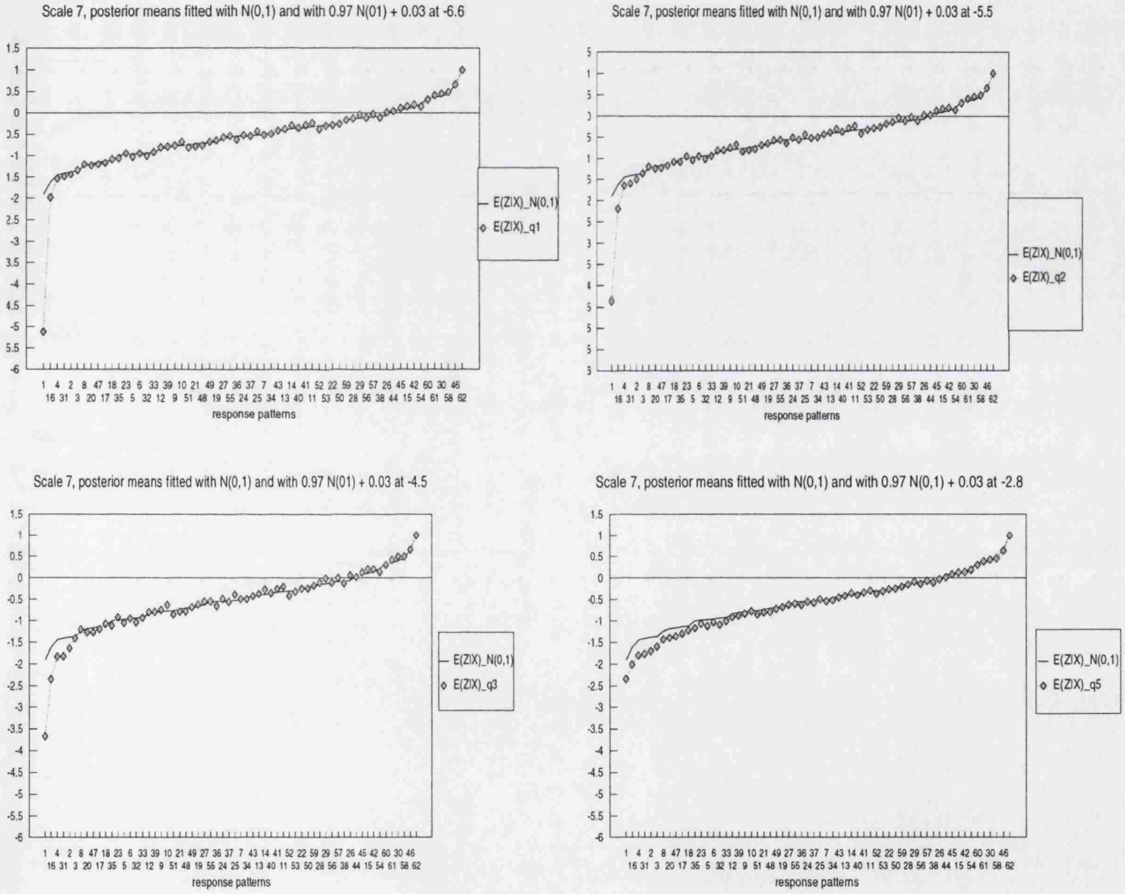
Scale 7, six items We repeat this analysis for the posterior means obtained if we use all six items of scale 7.

Figure 3.24 shows the posterior means as we place 0.03 probability at -6.6, -5.5, -4.5 and -2.8 together with the posterior means obtained from the N(0,1) prior, ordered according to the latter.

There are a few transpositions between response patterns, usually between two consecutive response patterns, or larger ones, when one response pattern is moved up to four places below or above from its previous place, but the order between the ones it is 'jumping' is preserved. We must also take into account that here we have sixty-two response patterns so a rank change of four response patterns is not so crucial as when we have sixteen. The more severe transpositions - larger jumps - happen when we place extra probability at the second and third quadrature points.

When extra probability is placed at -2.8, the fifth quadrature point, the initial order of the posterior patterns is preserved, except for a few one-place transpositions.

Figure 3.24: Scale 7, posterior means from N(0,1) and from contaminated priors with 0.03 extra probability, ordered according to posterior means from N(0,1)



Scale 7, posterior means fitted with N(0,1) and with 0.97 N(01) + 0.03 at -6.6

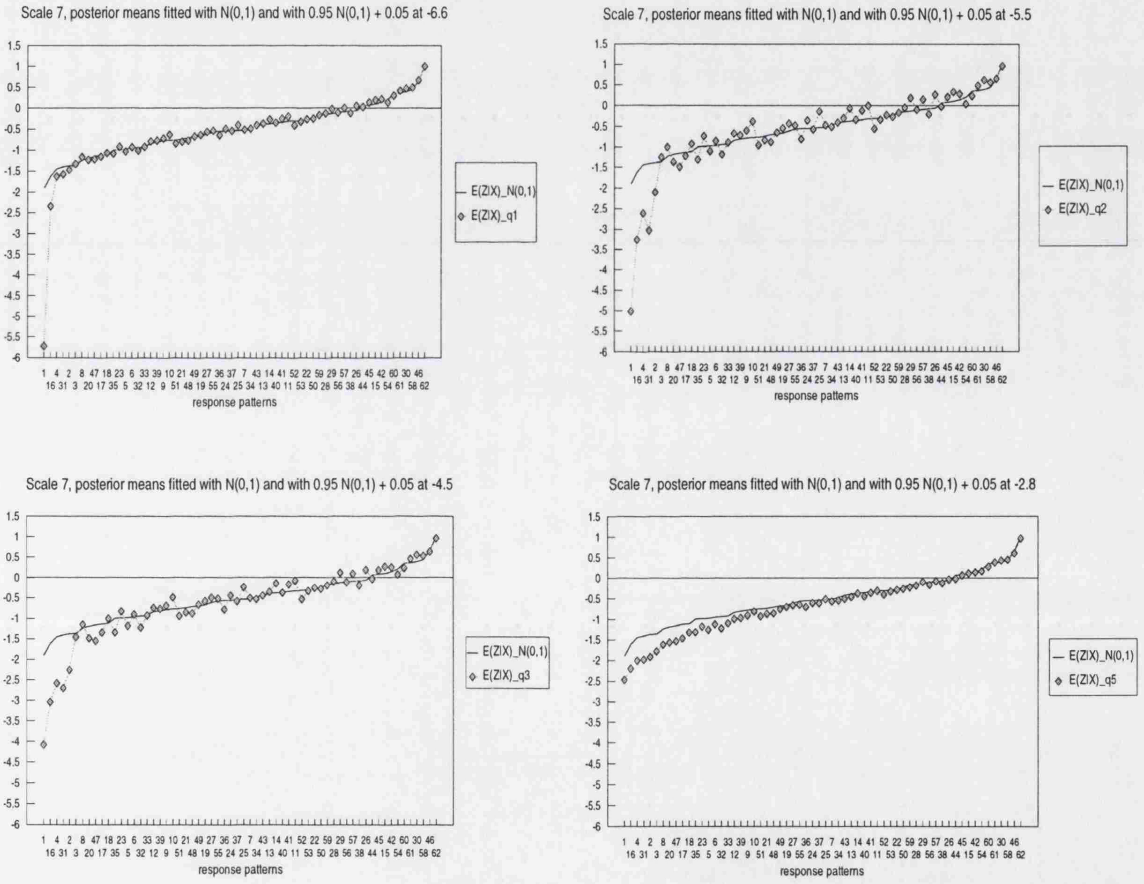
Scale 7, posterior means fitted with N(0,1) and with 0.97 N(01) + 0.03 at -5.5

Scale 7, posterior means fitted with N(0,1) and with 0.97 N(01) + 0.03 at -4.5

Scale 7, posterior means fitted with N(0,1) and with 0.97 N(0,1) + 0.03 at -2.8

There are naturally more transpositions when we place 0.05 extra probability at points of the prior. In Figure 3.25, we see the posterior means as we place 0.05 probability at -6.6, -5.5, -4.5 and -2.8. There is quite a lot of jumping up and down the 'original' posterior means line when extra probability is placed on -5.5 or -4.5, though these are the quadrature points with the most extreme rates of change. The bottom one of Figure 3.25 shows the posterior means as we place probability on -2.8. There only one place transpositions between the posterior means (i.e. changes in the ranking with posterior means that are immediately before of after them only) and this is a more typical picture when extra probability is placed on the rest of the quadrature points.

Figure 3.25: Scale 7, posterior means from N(0,1) and from contaminated priors with 0.05 extra probability, ordered according to posterior means from N(0,1)



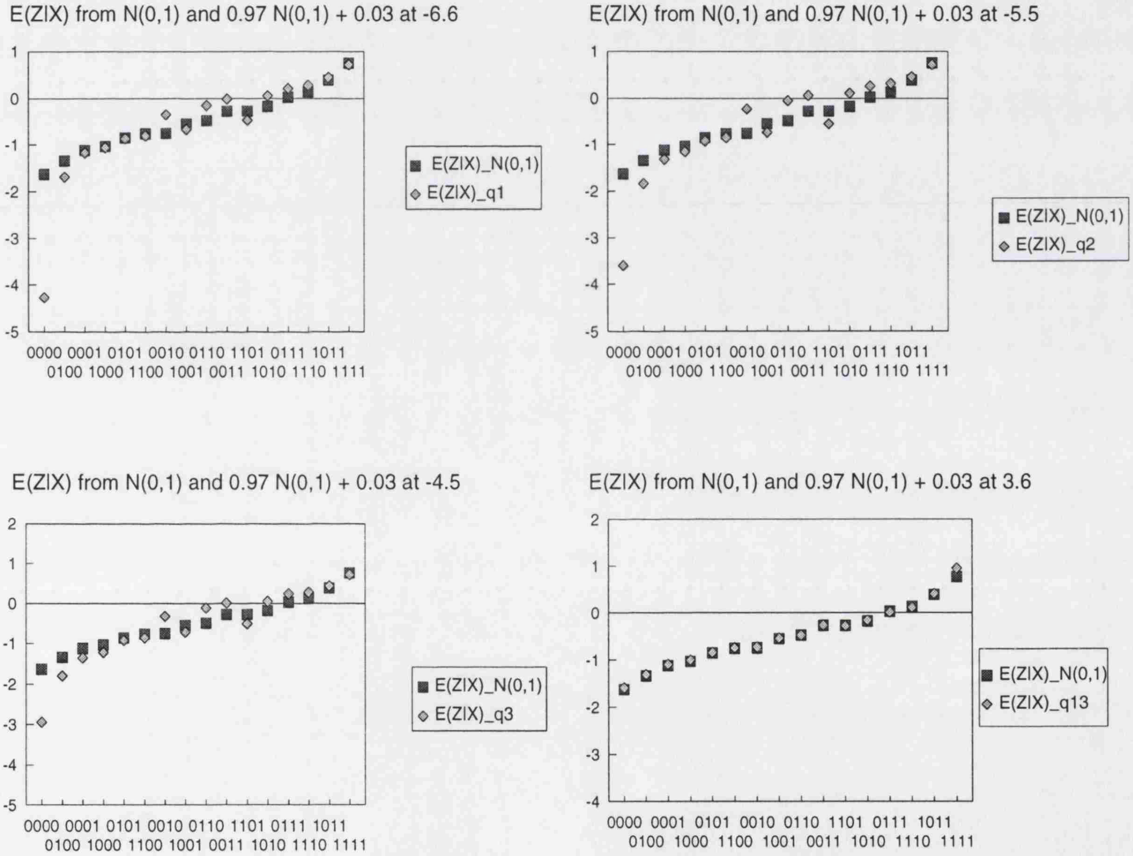
Scale 7, first 4 items, posterior means from standardised estimates As some of the changes in the posterior means are due to changes in the location and scale of the prior, we will also look at the posterior means obtained from the standardised parameters. We used the parameters standardised with the median and interquartile ratio of the prior as these were closer the parameters obtained from the N(0,1) in most cases.

Figures 3.26 show the posterior means obtained from the standardised parameters when extra probability is put at each approximation point of the prior.

We see that these Figures are very similar to Figures 3.23. The same transpositions happen between the posterior means of the response patterns, though in some cases the new posterior means are slightly closer to the original ones. These transpositions happen for changes in the negative range of z. We saw in Section 3.5 that the changes in the parameter estimates when extra probability was added on a negative point of z could not be fully attributed to the change in the location and dispersion of the prior relative to the N(0,1) and the standardised estimates were still different from the ones obtained from a N(0,1) prior. This explains why we still see transpositions in the posterior means as when the latter are calculated from the unstandardised estimates. We must note though that the numerical differences between consecutive response patterns are very small.

When extra probability is added on the positive range of z, the posterior means whether they have been calculated from the unstandardised or the standardised estimates preserve the ordering of the posterior means from the N(0,1) prior.

Figure 3.26: Scale 7, 4 items, posterior means from N(0,1) and from contaminated priors with 0.03 extra probability, ordered according to posterior means from N(0,1). The parameters used in the calculation of the posterior means from the contaminated prior have been standardised.



3.7 Fitting mixtures of normals

In the following we shall investigate the effect of fitting a one factor latent trait model taking mixtures of normals of the form $(1-\delta)$ N(0,1) + δ N(μ_2, σ_2^2) as the prior distribution.

The mixtures we considered are

i) $\delta = 0.05$, $\sigma_2^2 = 1$ and $\mu_2 =$ 1,2,3,4,5, and 6.

ii) $\delta = 0.50$, $\sigma_2^2 = 1$ and $\mu_2 =$ 1,2,3,4,5 and 6.

iii) $\delta = 0.25$, $\mu = 0$ and $\sigma_2^2 =$ 2, 2.5 and 3.

iv) $\delta = 0.50$, $\mu = 0$ and $\sigma_2^2 =$ 2, 2.5 and 3.

These parameters determine very different shapes of distributions. The first group comprises skewed distributions, with a small lump on the side, at different locations. The second group comprises symmetric distributions consisting of two equal parts, which may overlap or be almost completely separate. The third and fourth groups comprise unimodal distributions with inflated variance.

3.7.1 Calculations

The prior was approximated by a set of 16 quadrature points and their corresponding weights, 8 quadrature pairs were used for each component of the mixture. The points and the weights were obtained by the NAG subroutine D01BBF. The weights were then normalised (so that they added up to 1) and then again multiplied by the weight given to each component of the mixture. The model was fitted with TWOMISS. For this study we used the employment data.

3.7.2 Results

The parameter estimates obtained by fitting the logit model with a mixture as a prior to the employment data are given in the first part of Tables 3.6, 3.4, 3.7, 3.5, 3.9, 3.8 and are denoted by a_{ji}^*. These can be compared with the parameter estimates obtained with a N(0,1), given in the left hand side of the same Tables and denoted by a_{ji}, $j = 0, 1 \quad i = 1, \ldots 4$.

We observe that a_{0i}^* are smaller than a_{0i} if $\mu_2 > 0$. This follows from (3.13). Since $\mu > 0$ the subjects are assumed to have a higher job satisfaction than when fitting a

N(0,1). When modelling the same response patterns, once with $\mu > 0$ and once with $\mu = 0$ the items must appear more 'difficult' to respond positively to in the first case.

The more the distribution is shifted to the right, the smaller the difficulty parameter becomes. The largest discrepancies are observed when $\delta = 0.5$ and μ_2 large.

The difficulty parameter remains unaffected though by the change of the dispersion and overall shape of the prior; the a^*_{0i}'s coming from the mixtures with the inflated variance are identical to the 'original' a_{0i}'s (see Tables 3.9 and 3.8).

As we saw in Section 3.1.1, changes in the variance of the prior distribution will manifest themselves in the discrimination parameter. Since the variance of all the mixtures we considered is larger than 1, the response patterns are modelled as coming from a more variable population regarding attitude to work than a N(0,1) population, and therefore an easier to discriminate population. When we model the same set of responses once as coming from a more variable population and once as coming from a less variable population the items should appear less discriminatory in the first case. As expected, the discrimination parameters a^*_{1i} are smaller than when fitting the data with the N(0,1) as prior. The largest changes in a_{1i} are observed with the 0.5 N(0,1)+ 0.5 N(μ_2, 1) distribution.

In order to free the parameter estimates from the location and scale of the prior fitted and see whether other characteristics of the prior affect the parameter estimates, we standardised them using Equations (3.18), (3.19), (3.20) and (3.21).

The estimates standardised with the mean and standard deviation $a^{'}_{ji}$ and the estimates standardised with the median and interquartile ratio $a^{''}_{ji}$ are given in Tables 3.6, 3.4, 3.7, 3.5, 3.9, 3.8.

The mean, median, standard deviation and the interquartile distance (iqd) of the priors are given in Table 3.10.

Let us see how the standardised estimates compare with the estimates obtained with the N(0,1).

0.5 N(0,1) + 0.5 N(μ_2,1) In Table 3.4 we see that the estimates standardised with the mean, $a^{'}_{0i}$, are very close to a_{0i}, in fact indistinguishable from the N(0,1) estimates, even when $\mu_2 = 6$. Since the mean is equal to the median for these distributions, $a^{''}_{0i}$ is equivalent to $a^{'}_{0i}$. Plots of the standardised estimates are shown in Figure 3.27. The straight lines are drawn at the estimates from the N(0,1) prior and at bounds of their confidence intervals.

In Figure 3.28 we have the new standardised discrimination estimates, both with the standard deviation and the interquartile ratio against the mean of the second component of the mixture, μ_2. We see that the estimates standardised with the standard deviation are smaller than the N(0,1) parameters, but remain within the confidence intervals of the N(0,1) parameters even with $\mu_2 = 6$. Standardisation with the interquartile ratio however makes the estimates larger than the N(0,1) estimates and keeps a_{12} and a_{14} within the confidence interval of the N(0,1) parameters for all values of μ_2, but throws a_{11} and a_{13} outside the confidence intervals for $\mu_2 \geq 4$. The standardised estimates are also given in Table 3.5.

For this type of mixture standardisation with the standard deviation gives better results.

0.95 N(0,1) + 0.05 N(μ_2,1) In Figure 3.29 and Table 3.6 we have the standardised difficulty parameters plotted against μ_2. Except for a'_{04}, the other standardised with the mean estimates exceed the confidence interval bounds of the N(0,1) estimates for some value of μ_2. Standardisation with the median gives far better results, as the standardised parameters are within the confidence interval bands, though they are consistently smaller than the N(0,1) estimates.

Regarding the discrimination parameter a'_{1i} (Figure 3.30, Table 3.7), for items i=1,2,3 it is within the confidence interval of the corresponding a_{1i}'s up to μ_2 equal to 4, or 5 for item 2, but the standardised estimates are too high for larger values of μ_2. a''_{1i} though is very close to a_{1i}, and well within the confidence intervals of the N(0,1) parameters.

0.5 N(0,1) + 0.5 N(0,σ_2^2) The standardised estimates with the standard deviation and the interquartile ratio of the mixture are shown in Figure 3.31 and in Table 3.8. Both standardisations give estimates close to the ones from the N(0,1) prior and well within the confidence intervals of the latter. It seems the standard deviation works better for a_{11} and the interquartile ratio for a_{12}, but differences are so small that they could be attributed to numerical imprecision.

0.75 N(0,1) + 0.25 N(0,σ_2^2) The new and standardised estimates are given in Table 3.9. As with the estimates from 0.5 N(0,1) + 0.5 N(0,σ_2^2), the standardised estimates

Figure 3.27: 0.5 N(0,1) + 0.5 N(μ_2,1), standardised a_{0i}^{*} with the mean of the mixture

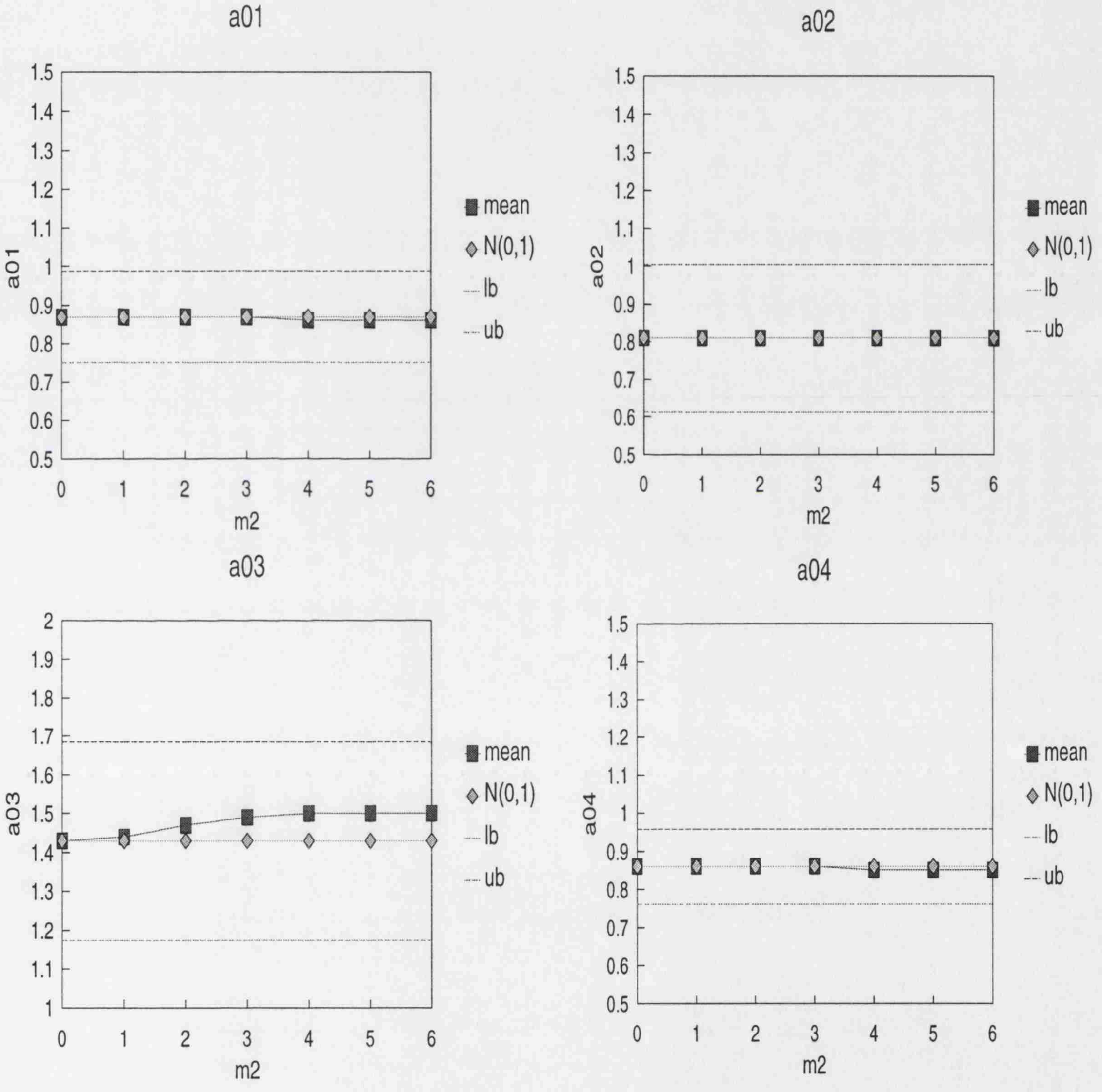


Figure 3.28: 0.5 N(0,1) + 0.5 N(μ_2,1), a^*_{1i}, standardised with the standard deviation and interquartile ratio of the mixture

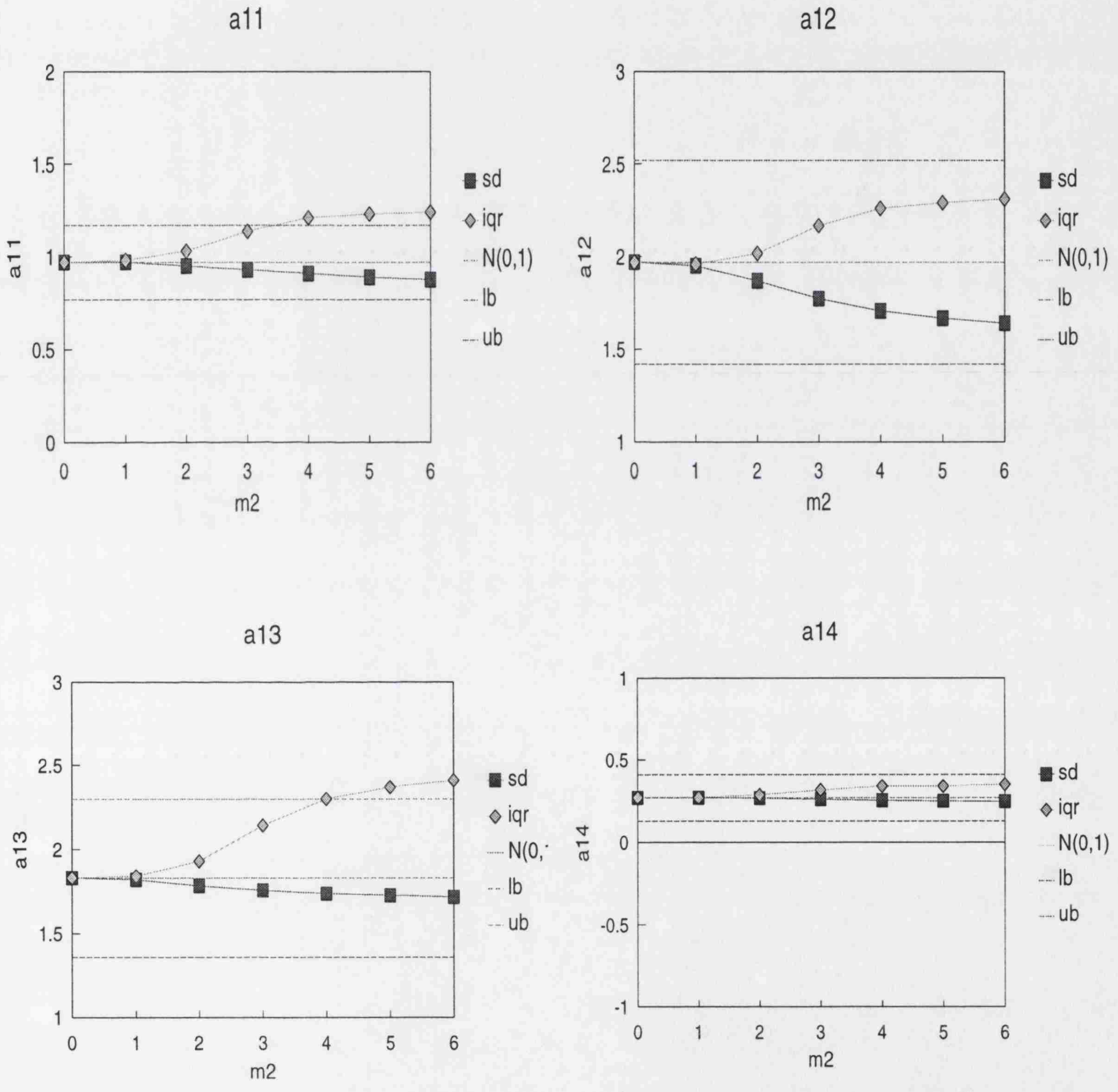


Figure 3.29: 0.95 N(0,1) + 0.05 N(μ_2,1), standardised a^*_{0i} with the mean and median of the mixture

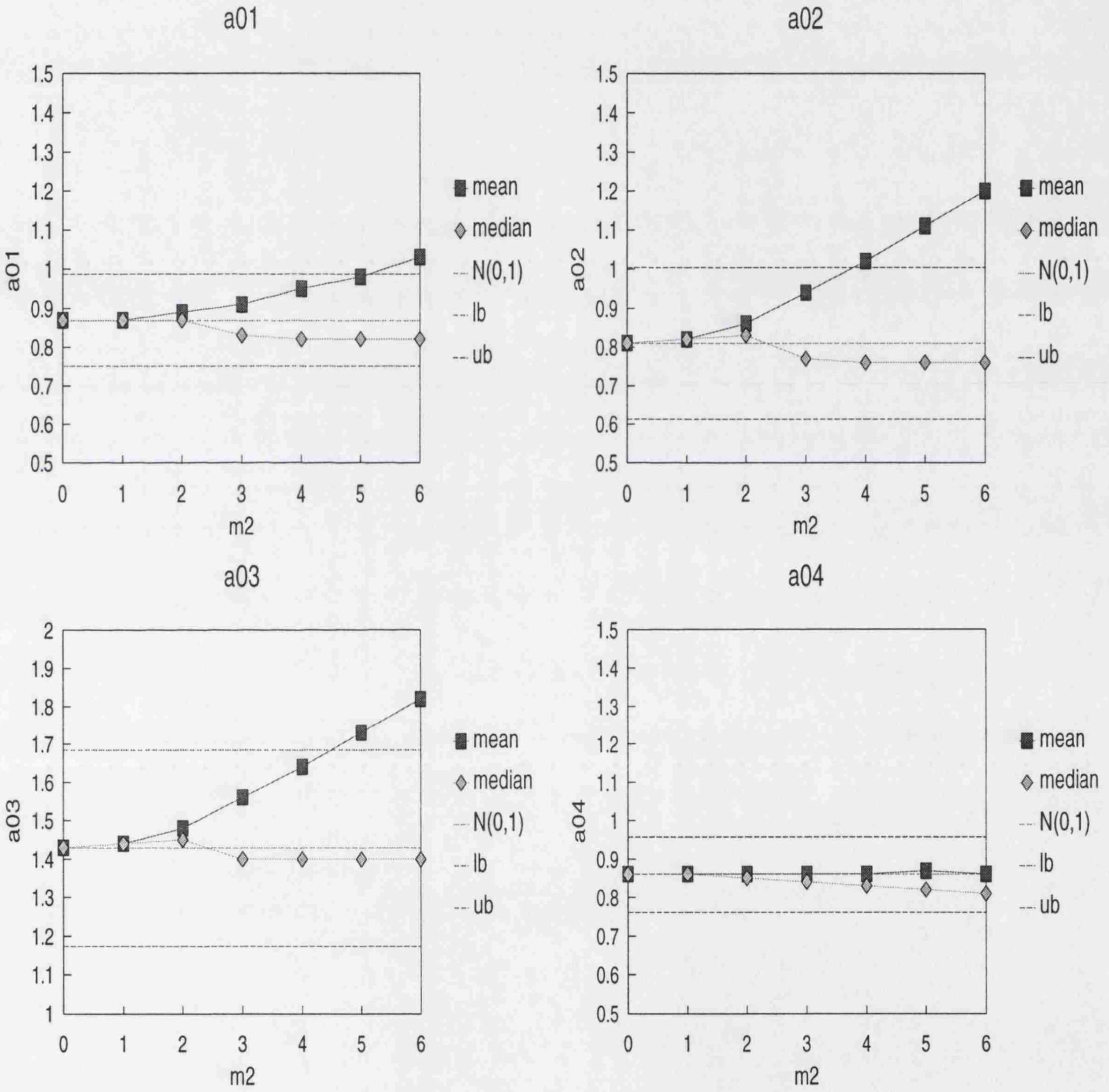


Figure 3.30: 0.95 N(0,1) + 0.05 N(μ_2,1), a^*_{1i} standardised with the standard deviation and interquartile ratio of the mixture

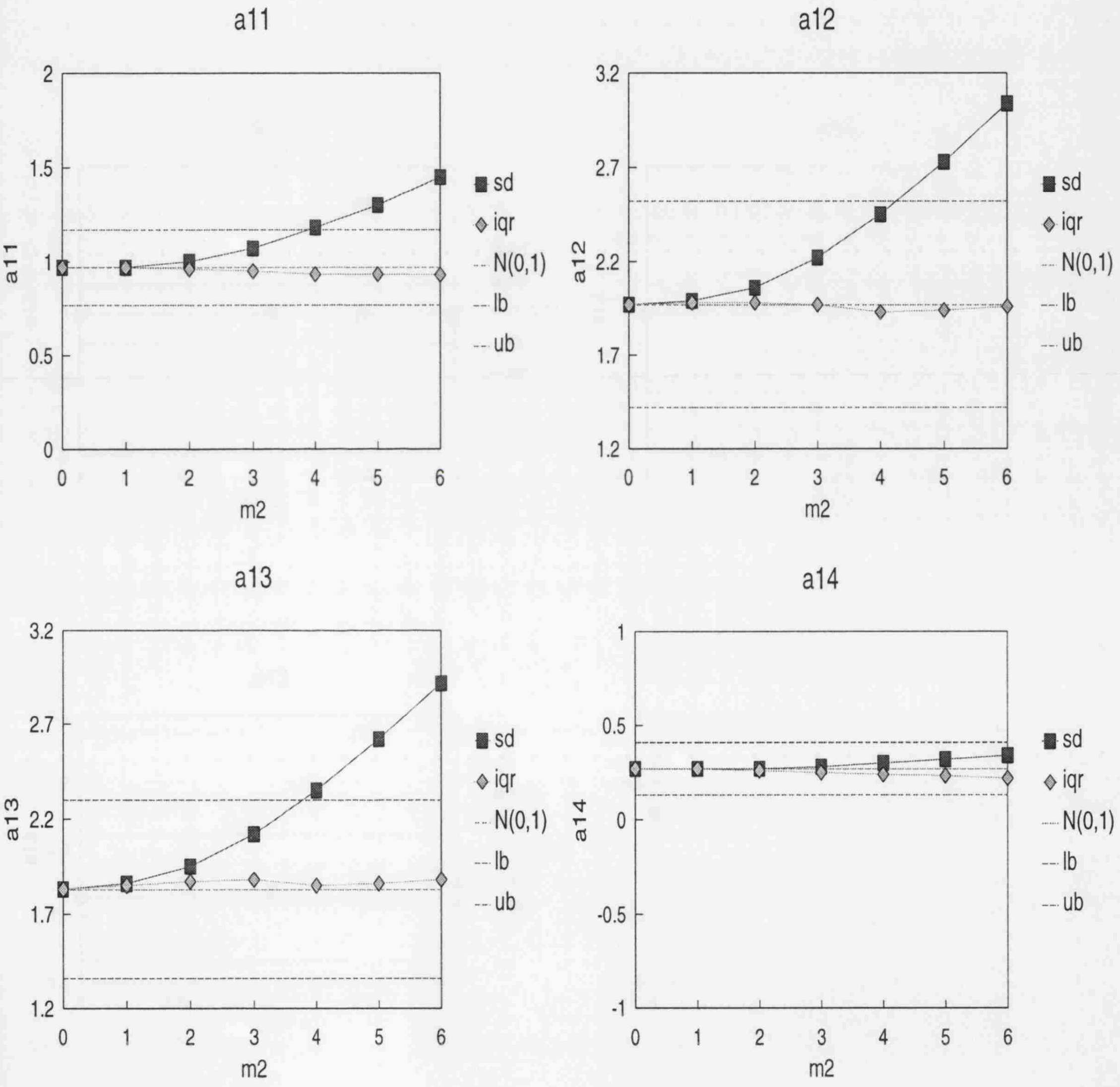


Figure 3.31: 0.5 N(0,1) + 0.5 N(0,σ^2), a^*_{1i} standardised with the standard deviation and interquartile ratio of the mixture

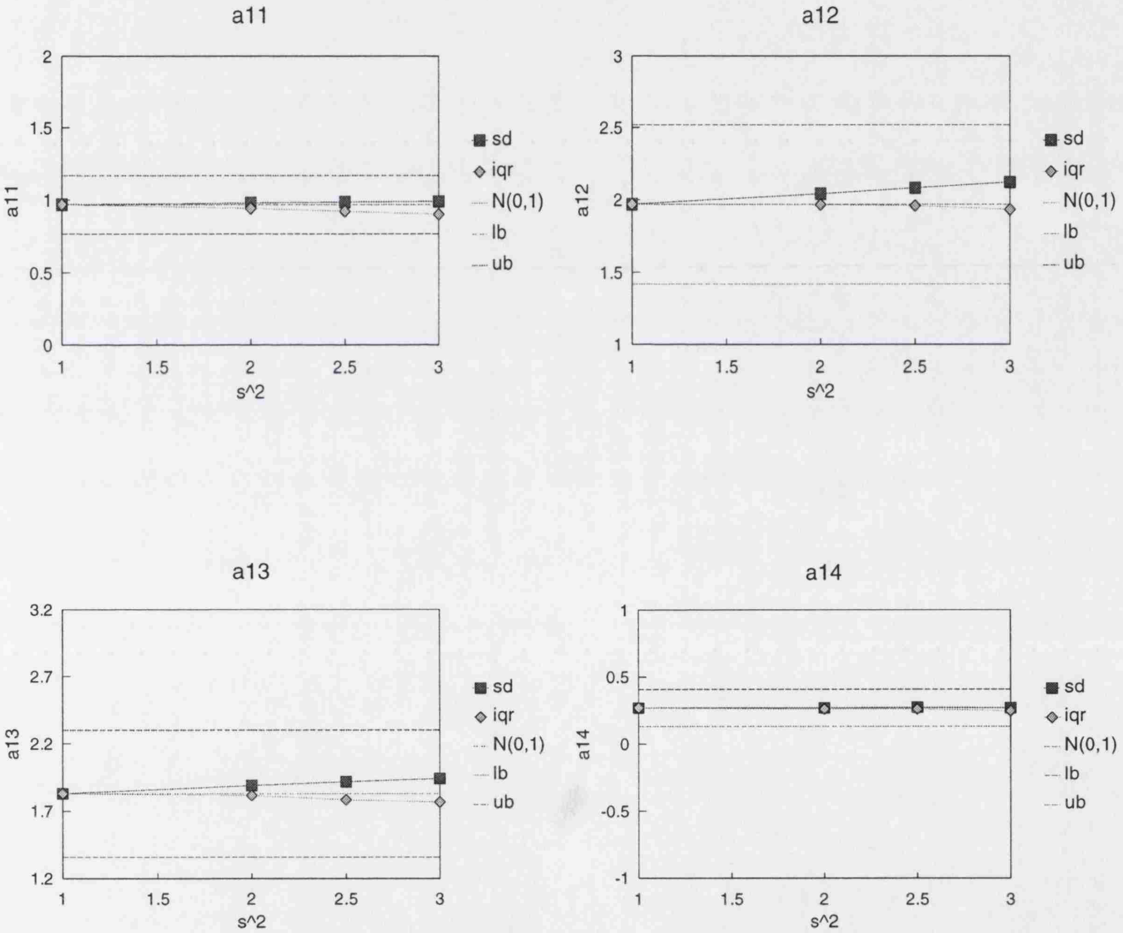


Table 3.4: a_{0i}, prior: 0.5 N(0,1) + 0.5 N(μ_2,1)

$\mu_2 = 0$		μ_2=1	μ_2=2	μ_2=3	μ_2=4	μ_2=5	μ_2=6
a_{0i}	s.e.	a^*_{0i}					
0.87	0.06	0.44	0.2	0.09	0.05	0.03	0.03
0.81	0.10	-0.07	-0.51	-0.66	-0.71	-0.73	-0.74
1.43	0.13	0.61	0.18	0.01	-0.06	-0.10	-0.12
0.86	0.05	0.74	0.67	0.64	0.63	0.62	0.62
a_{0i}	s.e.	$a'_{0i} = a^*_{0i} + a^*_{1i}\ \mu$					
0.87	0.06	0.87	0.87	0.87	0.87	0.86	0.86
0.81	0.10	0.81	0.81	0.81	0.81	0.81	0.81
1.43	0.13	1.43	1.44	1.47	1.49	1.50	1.50
0.86	0.05	0.86	0.86	0.86	0.86	0.85	0.85

with either the standard deviation or the interquartile ratio are very close to the estimates from the N(0,1).

We conclude that standardising with some appropriate measure of location and dispersion of the prior gives very similar results to those obtained by fitting a N(0,1) as a prior. So, assuming any distribution as the prior, will give essentially the same parameter estimates apart from a location and scale transformation. When the distribution assumed is very skewed a robust measure of location and scale is needed to recover the estimates that would have been obtained with a N(0,1). It seems that other characteristics of the shape of the prior distribution than location and scale have little effect on parameter estimates. The logit model seems fairly robust to the form of the prior.

3.8 Conclusions

In this chapter we examined the sensitivity of the parameter estimates and the posterior means to changes in the prior distribution. We first examined the effect of the parameters to small changes in the prior. We derived the Influence Function for the prior, measured at each of the quadrature points, which gives the rates of change of the parameters when a small amount of probability is placed at each point. The Influence Function seemed generally well behaved, levelling off at the extremes of the prior distribution. We also fitted the prior using a mixture of a N(0,1) plus a small amount of probability on each quadrature point in turn. Most of the parameters showed small changes and remained

Table 3.5: a_{1i}, prior: 0.5 N(0,1) + 0.5 N(μ_2,1)

$\mu_2 = 0$		μ_2=1	μ_2=2	μ_2=3	μ_2=4	μ_2=5	μ_2=6
a_{1i}	s.e.	a^*_{1i}					
0.97	0.10	0.87	0.67	0.52	0.41	0.33	0.28
1.97	0.28	1.74	1.32	0.98	0.76	0.62	0.52
1.83	0.24	1.63	1.26	0.97	0.78	0.64	0.54
0.27	0.07	0.24	0.19	0.14	0.11	0.09	0.08
a_{1i}	s.e.	$a'_{1i} = a^*_{1i}\ \sigma$					
0.97	0.10	0.97	0.95	0.93	0.91	0.89	0.88
1.97	0.28	1.95	1.87	1.77	1.71	1.67	1.64
1.83	0.24	1.82	1.78	1.76	1.74	1.73	1.72
0.27	0.07	0.27	0.27	0.26	0.25	0.25	0.25
a_{1i}	s.e.	$a''_{1i} = a^*_{1i}$ iqr					
0.97	0.10	0.98	1.03	1.14	1.21	1.23	1.24
1.97	0.28	1.96	2.02	2.17	2.26	2.29	2.31
1.83	0.24	1.84	1.93	2.15	2.30	2.37	2.41
0.27	0.07	0.27	0.29	0.32	0.34	0.34	0.35

Table 3.6: a_{0i}, prior: 0.95 N(0,1) + 0.05 N(μ_2,1)

$\mu_2 = 0$		μ_2=1	μ_2=2	μ_2=3	μ_2=4	μ_2=5	μ_2=6
a_{0i}	s.e.	a^*_{0i}					
0.87	0.06	0.83	0.80	0.78	0.77	0.76	0.76
0.81	0.10	0.73	0.67	0.66	0.65	0.65	0.65
1.43	0.13	1.35	1.31	1.29	1.29	1.29	1.29
0.86	0.05	0.85	0.83	0.82	0.82	0.81	0.80
a_{0i}	s.e.	$a'_{0i} = a^*_{0i} + a^*_{1i}\ \mu$					
0.87	0.06	0.87	0.89	0.91	0.95	0.98	1.03
0.81	0.10	0.82	0.86	0.94	1.02	1.11	1.20
1.43	0.13	1.44	1.48	1.56	1.64	1.73	1.82
0.86	0.05	0.86	0.86	0.86	0.86	0.87	0.86
a_{0i}	s.e.	$a''_{0i} = a^*_{0i} + a^*_{1i}$ median					
0.87	0.06	0.87	0.87	0.83	0.82	0.82	0.82
0.81	0.10	0.82	0.83	0.77	0.76	0.76	0.76
1.43	0.13	1.44	1.45	1.40	1.40	1.40	1.40
0.86	0.05	0.86	0.85	0.84	0.83	0.82	0.81

Table 3.7: a_{1i}, prior: 0.95 N(0,1) + 0.05 N(μ_2,1)

$\mu_2 = 0$		μ_2=1	μ_2=2	μ_2=3	μ_2=4	μ_2=5	μ_2=6
a_{1i}	s.e.	a^*_{1i}					
0.97	0.10	0.95	0.92	0.90	0.89	0.88	0.88
1.97	0.28	1.95	1.89	1.86	1.85	1.85	1.84
1.83	0.24	1.82	1.78	1.77	1.77	1.77	1.77
0.27	0.07	0.27	0.25	0.24	0.23	0.22	0.21
a_{1i}	s.e.	$a'_{1i} = a^*_{1i}\ \sigma$					
0.97	0.10	0.97	1.00	1.07	1.18	1.30	1.45
1.97	0.28	1.99	2.06	2.22	2.45	2.73	3.04
1.83	0.24	1.86	1.95	2.12	2.35	2.62	2.92
0.27	0.07	0.27	0.27	0.28	0.30	0.32	0.34
a_{1i}	s.e.	$a''_{1i} = a^*_{1i}$ iqr					
0.97	0.1	0.97	0.96	0.95	0.93	0.93	0.93
1.97	0.28	1.98	1.98	1.97	1.93	1.94	1.96
1.83	0.24	1.85	1.87	1.88	1.85	1.86	1.88
0.27	0.07	0.27	0.26	0.25	0.24	0.23	0.22

Table 3.8: a_{ji}, prior: 0.5 N(0,1) + 0.5 N(0,σ^2)

$\sigma^2 = 1$		σ^2=2	σ^2=2.5	σ^2=3
a_{0i}	s.e.	a^*_{0i}		
0.87	0.06	0.87	0.87	0.87
0.81	0.10	0.82	0.82	0.82
1.43	0.13	1.43	1.43	1.43
0.86	0.05	0.86	0.86	0.86
a_{1i}	s.e.	a^*_{1i}		
0.97	0.10	0.80	0.75	0.70
1.97	0.28	1.67	1.57	1.50
1.83	0.24	1.54	1.45	1.37
0.27	0.07	0.22	0.21	0.19
a_{1i}	s.e.	$a'_{1i} = a^*_{1i}\ \sigma$		
0.97	0.10	0.98	0.99	0.99
1.97	0.28	2.04	2.08	2.12
1.83	0.24	1.89	1.92	1.94
0.27	0.07	0.27	0.27	0.27
$a_{1,i}$	s.e.	$a''_{1i} = a^*_{1i}$ iqr		
0.97	0.10	0.94	0.92	0.90
1.97	0.28	1.97	1.93	1.93
1.83	0.24	1.82	1.78	1.77
0.27	0.07	0.26	0.26	0.25

Table 3.9: a_{ji}, prior: 0.75 N(0,1) + 0.25 N(0,σ^2)

$\sigma^2 = 1$		σ^2=2	σ^2=2.5	σ^2=3
a_{0i}	s.e.	a^*_{0i}		
0.87	0.06	0.87	0.87	0.87
0.81	0.10	0.82	0.82	0.82
1.43	0.13	1.43	1.43	1.43
0.86	0.05	0.86	0.86	0.86
a_{1i}	s.e.	a^*_{1i}		
0.97	0.10	0.88	0.84	0.82
1.97	0.28	1.82	1.77	1.72
1.83	0.24	1.69	1.63	1.59
0.27	0.07	0.24	0.23	0.22
a_{1i}	s.e.	$a'_{1i} = a^*_{1i}\ \sigma$		
0.97	0.10	0.98	0.99	1.00
1.97	0.28	2.04	2.07	2.11
1.83	0.24	1.89	1.92	1.95
0.27	0.07	0.27	0.27	0.27
$a_{1,i}$	s.e.	$a''_{1i} = a^*_{1i}$ iqr		
0.97	0.10	0.95	0.93	0.92
1.97	0.28	1.97	1.97	1.94
1.83	0.24	1.83	1.81	1.79
0.27	0.07	0.26	0.26	0.25

Table 3.10: Statistics of the mixture distributions

<table>
<tr><td colspan="7">Prior: 0.95 N(0,1) + 0.05 N(μ_2, 1)</td></tr>
<tr><td>statistic</td><td>μ_2=1</td><td>μ_2=2</td><td>μ_2=3</td><td>μ_2=4</td><td>μ_2=5</td><td>μ_2=6</td></tr>
<tr><td>μ
median
σ
iqd</td><td>0.05
0.05
1.02
1.37</td><td>0.10
0.08
1.09
1.41</td><td>0.15
0.06
1.20
1.43</td><td>0.20
0.06
1.33
1.41</td><td>0.25
0.06
1.48
1.42</td><td>0.30
0.06
1.65
1.43</td></tr>
<tr><td colspan="7">Prior: 0.5 N(0,1) + 0.5 N(μ_2, 1)</td></tr>
<tr><td>statistic</td><td>μ_2=1</td><td>μ_2=2</td><td>μ_2=3</td><td>μ_2=4</td><td>μ_2=5</td><td>μ_2=6</td></tr>
<tr><td>μ,/median
σ
iqd</td><td>0.50
1.12
1.52</td><td>1.00
1.41
2.06</td><td>1.50
1.80
2.97</td><td>2.00
2.24
4.00</td><td>2.50
2.69
5.00</td><td>3.00
3.16
6.00</td></tr>
<tr><td colspan="7">Prior: 0.75 N(0,1) + 0.25 N(0,σ^2)</td></tr>
<tr><td>statistic</td><td colspan="2">σ^2=2</td><td colspan="2">σ^2=2.5</td><td colspan="2">σ^2=3</td></tr>
<tr><td>σ
iqd</td><td colspan="2">1.12
1.46</td><td colspan="2">1.17
1.50</td><td colspan="2">1.22
1.52</td></tr>
<tr><td colspan="7">Prior: 0.50 N(0,1) + 0.50 N(0,σ^2)</td></tr>
<tr><td>statistic</td><td colspan="2">σ^2=2</td><td colspan="2">σ^2=2.5</td><td colspan="2">σ^2=3</td></tr>
<tr><td>σ
iqd</td><td colspan="2">1.22
1.59</td><td colspan="2">1.32
1.66</td><td colspan="2">1.41
1.74</td></tr>
</table>

within the confidence interval of the estimates obtained with a N(0,1) prior. There were also though large unexpected changes in some parameters as 0.03 probability was placed at some quadrature points, which remained even after we had standardised the parameters for the change in the location and the scale of the prior.

The effect on the posterior means of 0.03 probability on some quadrature points caused some transpositions between them, but these were not so extreme as when extra probability was placed on some response patterns.

We then fitted mixtures of normals as priors to examine the effect of gross changes in the prior on the parameter estimates of the latent trait model. We saw that when the parameter estimates were standardised with an appropriate measure of scale and location, for example the median and the interquartile ratio respectively when the priors were very skewed, the estimates differed very little from the ones obtained with a N(0,1) prior, and remained within the confidence intervals of the latter.

Chapter 4

Semiparametric estimation of the latent trait model

4.1 Introduction

This chapter focuses on the amount of information about the distribution of the latent variable, the 'prior' distribution, that can be retrieved from a set of n responses to p dichotomously scored questions.

Bock and Aitkin (1981) suggested the use of an 'empirical' prior for the latent trait model, that is a prior estimated together with the item parameters, in order to free the estimation from any assumptions about the prior distribution. The use of an empirical prior, has also been called *nonparametric* estimation of the prior, or *semiparametric* estimation of the latent trait model, since one still assumes a parametric form for the response function. Bock and Aitkin (1981) noted, that because the likelihood is so insensitive to the shape of the prior, one can't estimate accurately its finer features.

In this chapter we try to assess the information about the prior by measuring the variability of its semiparametric estimate, by comparing it, for artificial data, with the distribution used to generate the responses, and by comparing it with N(0,1), as this is usually the assumed form of the prior. Moreover, we investigate the effect of a different number of support points and different initial locations and weights of these points on the shape of the estimated prior and the scores of the latent variable.

We then explore the *fully semiparametric estimation* of the model, that is the estimation of both the nodes and the weights simultaneously with the item parameters. Although much research has been undertaken on the nonparametric - fully semiparametric estimation of the Rasch model, we do not know of any similar work on the 2-parameter latent trait model. The estimation method we propose is an expansion of the EM algorithm. Again we investigate the variability of the estimated prior, and the advantages of this method over the simple semiparametric approach.

Finally we compare the results from fitting the model semiparametrically with latent class and ordered latent class analysis.

4.2 Semiparametric Maximum Likelihood Estimation

If the latent trait model is estimated semiparametrically, then the prior is estimated together with the item parameters. We shall first describe the method where the nodes are defined on a grid where the bulk of the distribution is expected to lie and only the weights are estimated from the data. This method was given by Bock and Aitkin (1981).

Estimating the weights involves a small modification on the EM algorithm given in Chapter 1, Section 1.4.1. At the end of each Maximisation step, the posterior weights $h^*(z_t)$ are computed as follows:

$$h^*(z_t) = \sum_{l=1}^{n} h(z_t \mid \mathrm{x}_l)/n \tag{4.1}$$

Then (1.28) and (1.29) are calculated with $h^*(z_t)$ in place of $h(z_t)$. We take the nodes z and the posterior weights $h^*(z_t)$ obtained at the last iteration as the estimates of the approximation - or support - points of the empirical prior distribution.

Since the parameters are confounded with the location and scale of the prior distribution, we standardise the nodes of the prior at each iteration so that the mean of the prior is 0 and its variance 1. The mean and the variance are computed from the estimated nodes and weights with $\mu = \sum_{t=1}^{K} z_t h^*(z_t)$ and $\sum_{t=1}^{K}(z_t - \mu)^2 h^*(z_t)$.

4.2.1 Results from the Semiparametric Estimation of the Latent Trait Model

We shall first look at the estimated points and weights that result from different sets of starting points. These may be equally spaced or not, and have equal or unequal probabilities.

As noted in Section 1.9, many researchers have investigated the question of the optimal or sufficient number of points needed for the prior when this is estimated empirically. De Leeuw and Verhelst (1986) and Lindsay, Clogg, and Grego (1991) gave the maximum number of points needed for a prior for the Rasch model as $(p+1)/2$ if p, the number of items, is odd, and $(p+2)/2$ if p is even, with the first point being equal to $-\infty$ in the latter case.

We don't have equivalent results for the 2-parameter logistic model, and also this is a slightly different problem since the points are fixed on a grid, so we will investigate empirically the effect of the number of points, by estimating the prior using different number of points and comparing the results.

Since Heinen (1996) noted an effect of the shape of the starting prior on the goodness-of-fit of the model, we also take different starting grids for the same number or points to see their effect.

The data we are going to use to illustrate our results are the NFER test 1 for primary school boys, the intercultural scale 9, both for the American and the German samples, and some artificial data. NFER test 1 comprises 21 items and the sample size of the boys data is 566. The test is described in the Appendix. Scale 9, consists of 12 items (given in the Appendix) that have to do with 'Future Outlook'. The sample sizes are 1416 and 1490 for the American and German samples respectively.

We use as starting points sets of 16, 10, 8, 6, 4, 3 and 2 points, between -4 and 4. For most of these sets the points are equidistant and the weights uniform, but we use also sets for which the weights are skewed to the right or to the left. We also use a set of 4 not equidistant points. In all the cases, the points are standardised so that their mean is equal to 0 and their variance equal to 1.

The left plot of Figure 4.1 shows plots of the priors obtained from the sets of 16, 10 and 8 starting points (only the points with weight greater than 0.0001 are depicted). All the sets give the same overall picture, a distribution skewed to the left. From the 16 point prior, 7 points came out with zero weight (≤ 0.00001), from the 10 point prior 3 came out with zero weight and from all the different 8 point priors 2 points resulted with zero weight. It seems that not so many points are needed to approximate the prior.

The uniform prior in (-3.5, 3.5), the uniform in (-1.75, 1.75) and the skewed to the right prior gave exactly the same set of final points and weights. The skewed to the left prior and the N(0,1) gave only slightly different points and weights, so, as can been seen in the plot, all three resulting distributions are very similar to each other.

The right panel of Figure 4.1 shows the approximation points and weights for the priors with 6 or fewer support points. Two different sets of 6 points gave the same solution for the prior, which has a similar shape to the one obtained from 8 points. One of the points resulted in zero weight. Of course, this does not necessarily mean that 6 points are too many, but that the point may have been forced to a location where the probability was zero.

It seems that the magnitude of the starting weights, or the shape of the prior once the location of the points has been fixed, does not matter. We note though that all of the above sets of points are equidistant, except for the ones that approximate the N(0,1) distribution, where the central points are slightly closer together than the points further out on the sides of the distribution.

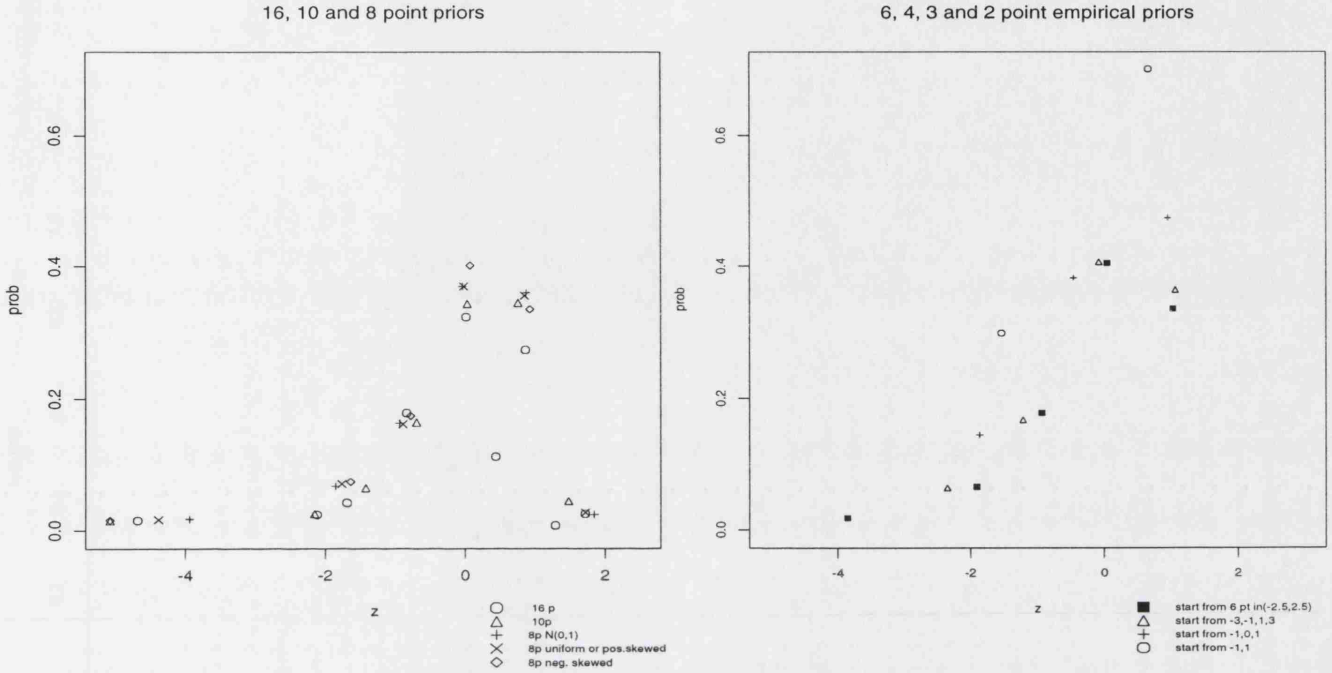
Let us see what happens when the distances between the points are more different.

We fitted the empirical prior to the American scale 9 data with two sets of four starting points, -3, -1, 1, 3 and -3, 0, 1, 3 and two sets of three starting points, -1, 0, 1 and -1, 1, 2. The resulting distributions are shown in Figure 4.2. We see now that these are quite different. For the four point priors, we see that the prior from the second set (-3,0,1,3) is a two modal distribution, because the two middle points are two close together and one had to carry a low probability. Of course it is hard to compare one set of points with another and judge how much the distributions really differ.

The prior as a step function The resulting prior can either be considered a step function, with number of steps equal to the number of points with non zero weight, or a smooth function, with the support points and weights approximating this function.

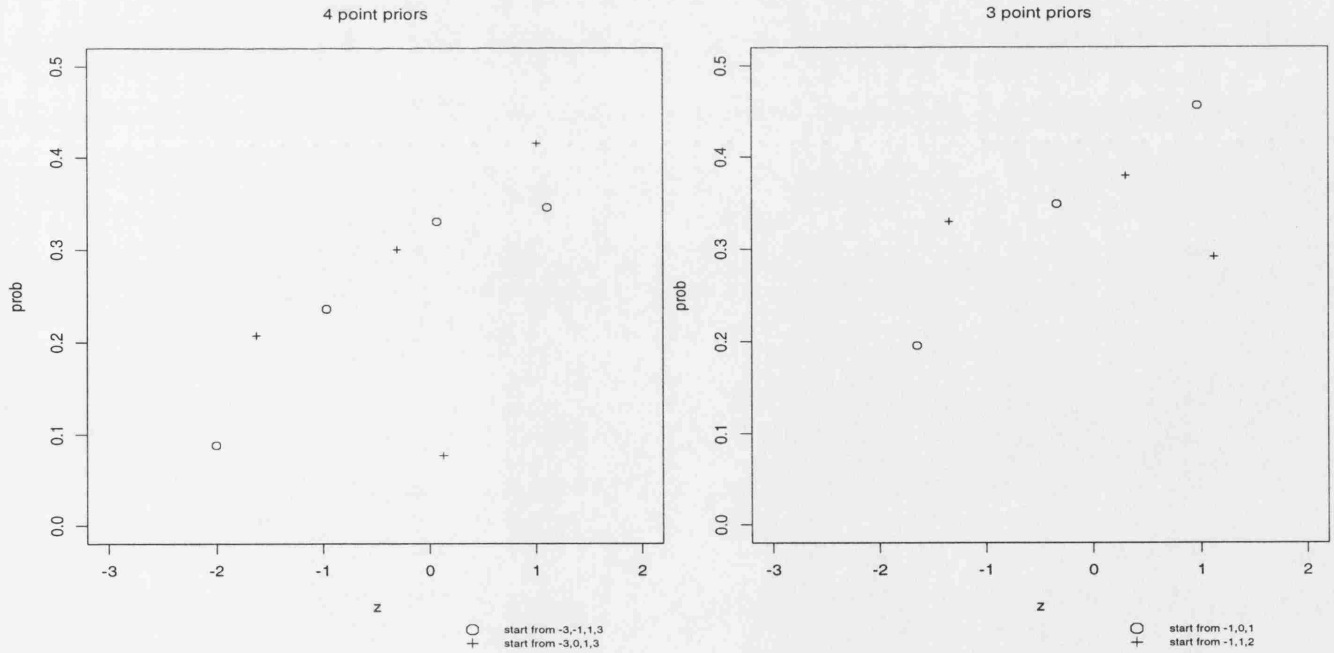
In the case of the step function, the jumps occur at each support point and they are

Figure 4.1: Boys, test 1, support points and weights for empirical priors obtained from different sets of starting points


16, 10 and 8 point priors

6, 4, 3 and 2 point empirical priors

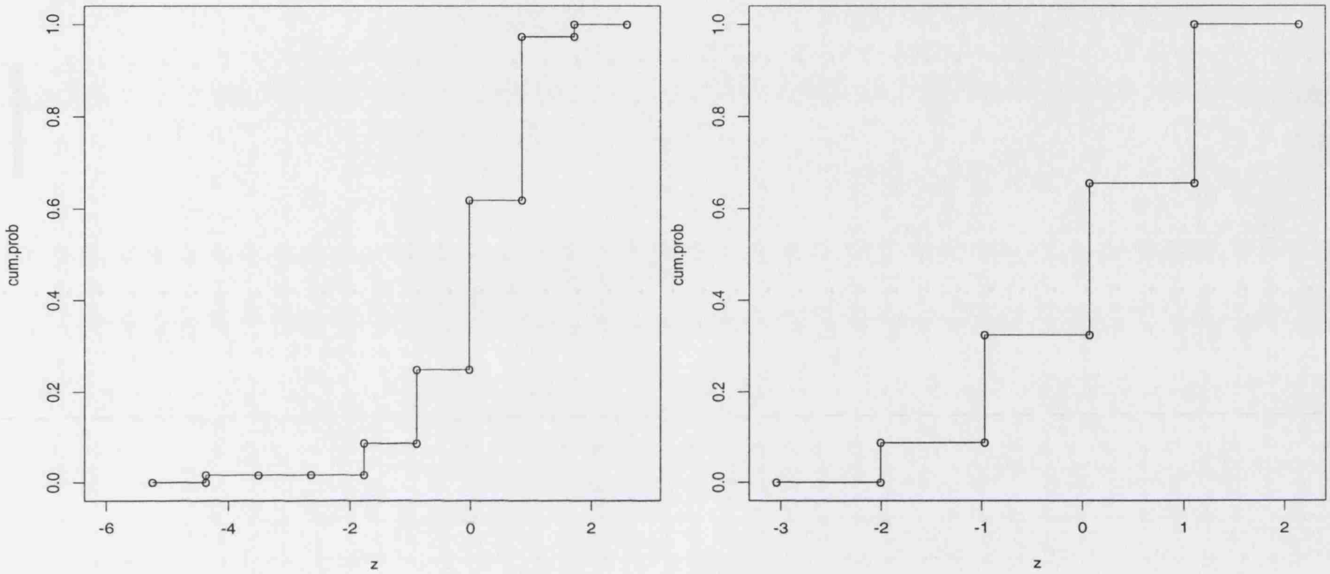
Figure 4.2: American scale 9, support points and weights for empirical priors obtained from different sets of starting points (equidistant and not-equidistant points)


4 point priors

3 point priors

equal to the weight of the point. The cumulative distribution functions (cdf) of the 8 and 4 point priors for the Boys, test 1 data and the American scale 9 data are plotted in Figure 4.3.

Figure 4.3: Cumulative distribution functions for the 8 point empirical prior for the Boys, test 1 data (left) and 4 point empirical prior for the American scale 9 (right)

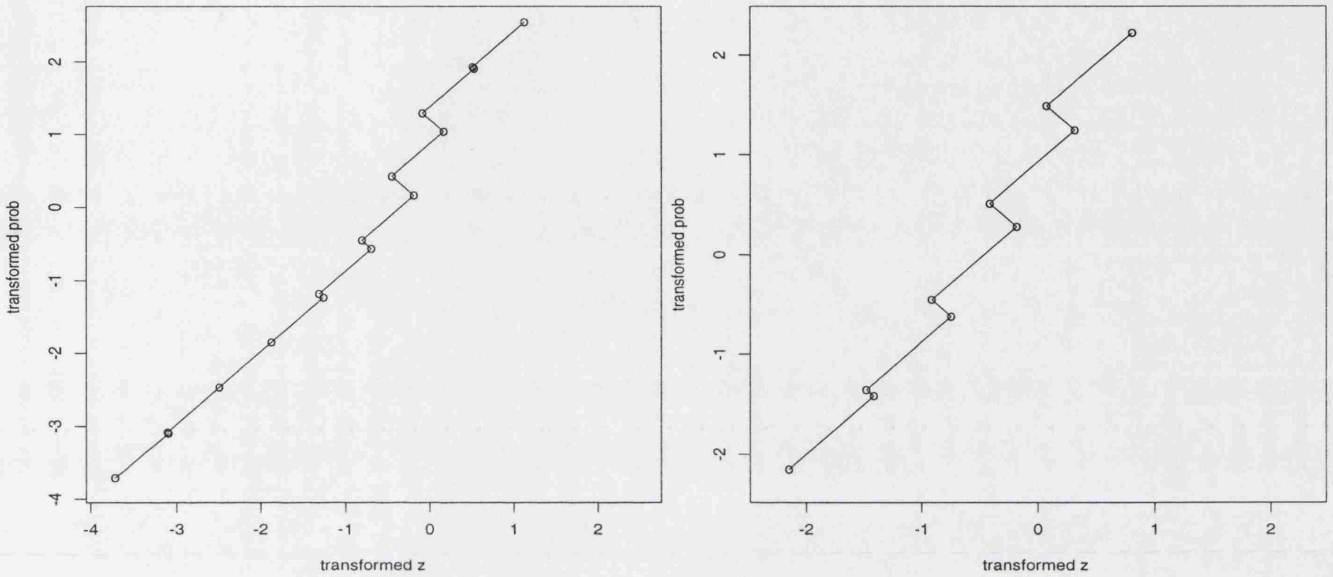


It is difficult to compare cdfs from different number of starting points, or even same number but with different shape. It would be useful to have percentiles and compare those, but again the percentiles are not unique, as ranges of probability correspond to same value.

A solution to this problem is to rotate the cdf through 45 degrees, and take the percentiles by interpolating from fixed points of the original y-axis to the x-axis, as one would do with a continuous cdf. An example of a rotated cdf, which is the 8 point prior for the Boys, test 1 data is plotted in Figure 4.4. The interpolated points will then have to be rotated back to the original axes.

The prior as a continuous function If the prior is considered a continuous distribution, the support points can be taken as the centres of histogram bins with probabilities equal to the weight of the points. From the histogram the ogive for each prior can be

Figure 4.4: Rotated cumulative distribution functions for the 8 point empirical prior for the Boys, test 1 data (left) and for the 4 point empirical prior for the American scale 9 (right)



constructed.

By looking at the smoothed ogives it will be easier to compare priors that were approximated with different numbers of support points. Moreover, from the ogive one can easily interpolate the percentiles of the prior and see how much these differ.

The ogives of the priors that were obtained for the different sets of 8 points and weights, are depicted in the left Figure 4.5, whereas the right panel of Figure 4.5 shows the ogives of the 6, 4, 3 and 2 point priors.

The ogives from the 8 point priors lie all very closely together. The ogives from the fewer point priors seem to differ in the first half of the distribution.

Figure 4.6 shows the ogives obtained from the two sets of 4 points and the two sets of 3 points (equidistant and non-equidistant points). These seem more different than the ogives obtained from the various 8 point priors but they are still similar.

It is still difficult to judge how much the ogives - or the percentiles - really differ, since we do not know how much variability is expected. In the following we shall use the bootstrap to measure the variability of the prior.

Figure 4.5: Boys, test 1, ogives of empirical priors obtained from different sets of starting points - equidistant and not equidistant points

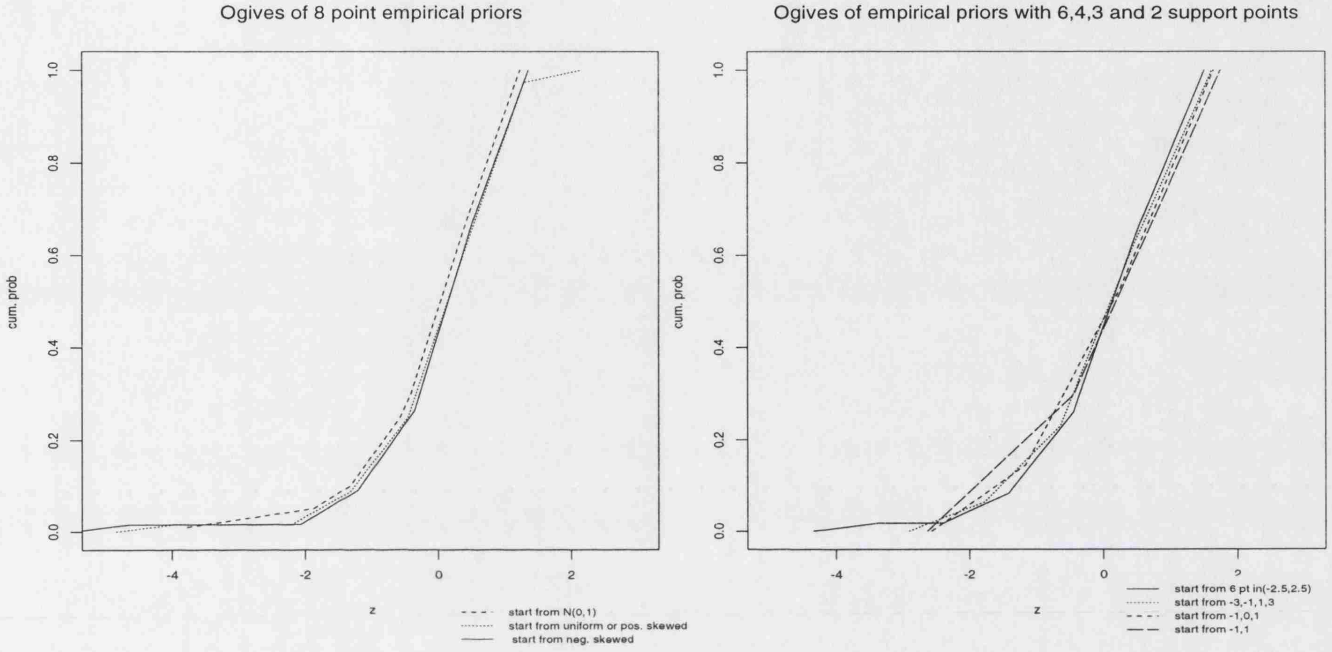
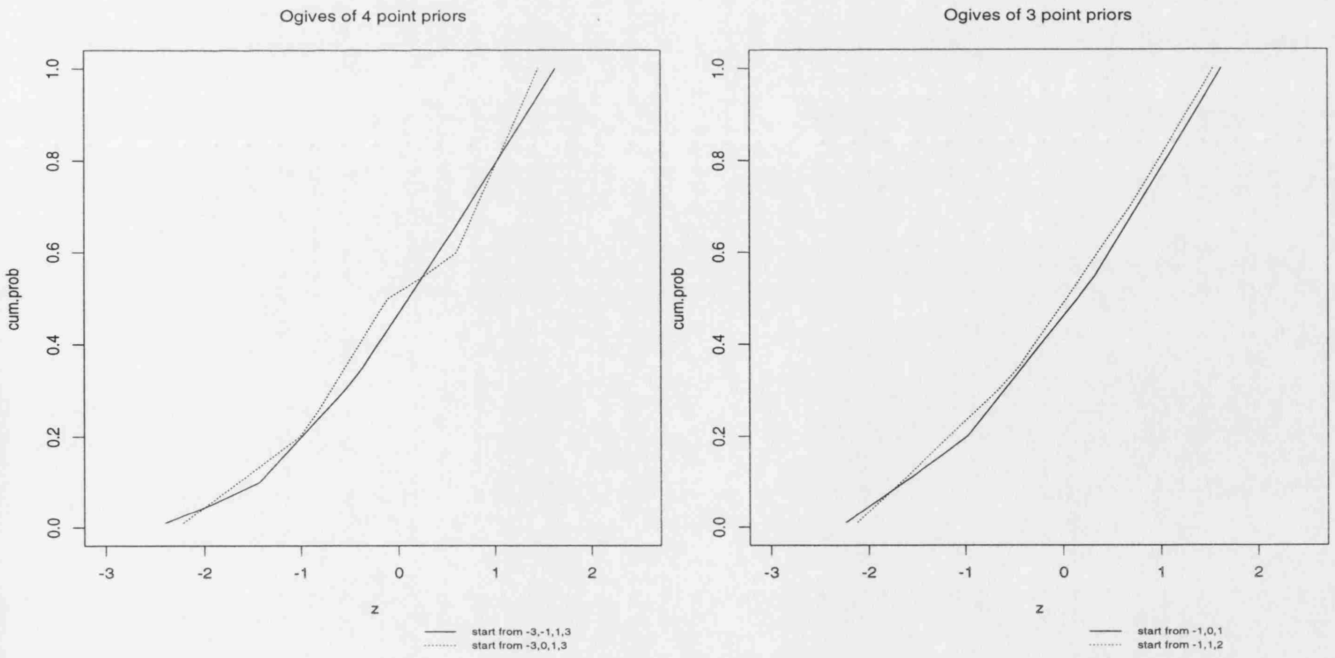


Figure 4.6: American scale 9, ogives of empirical priors obtained from different sets of starting points - equidistant and not equidistant points



4.3 Bootstrapping the prior - Nonparametric Bootstrap

The bootstrap (Efron and Tibshirani 1993) creates new samples from the original sample in order to measure the variability of the estimators. The variability in the sample is supposed to reflect the variability in the population the sample came from, so drawing new samples from the original sample will give an estimate of this variability.

For each estimated prior, we measured the variability of its percentiles. We took the percentiles that correspond to the cumulative probabilities:

$0.01, 0.05, 0.10, 0.15, ..., 0.90, 0.95, 0.99, 1.00.$

We constructed confidence intervals for the percentiles in the following way: we took 300 bootstrap samples, estimated the approximation points and weights, smoothed or rotated the prior and calculated the percentiles for each sample. We thus obtained a bootstrap distribution of the percentiles. We then constructed two types of confidence intervals for them: 1) Confidence intervals based on the percentiles of the bootstrap distribution (Efron and Tibshirani 1993). 2) Confidence intervals based on the Kolmogorov-Smirnov method (Mood, Graybill, and D.C.Boes (1963), pages 508-512).

The first type of confidence intervals (90% confidence) of the empirical prior percentiles are constructed taking the 5th and 95th percentiles of their bootstrap distribution as the lower and upper bound respectively of their confidence interval.

The second type of confidence intervals, which are based on the Kolmogorov-Smirnov method, are constructed in the following way: for each bootstrap sample we calculate the differences of the 5th to the 95th percentile from their respective original percentiles. (We disregard the 1st, 99th and 100th percentile because their values are extreme and their variability a lot larger than for the other percentiles. If we considered them the resulting confidence intervals would be so large that they would be hardly meaningful.) We record the maximum difference for each bootstrap sample and formed the bootstrap distribution of the maximum differences. We then take the 90th percentile of the maximum differences' distribution and add it to, and subtract it from each original percentile, for the lower and upper bound of its confidence interval. This method should give 90% confidence for the intervals for all percentiles considered (5th to 95th) simultaneously.

The ogive and the confidence intervals of the percentiles of the 8, 6, 4 and 3 point smoothed priors for the Boys, test 1 data, are depicted in Figure 4.7.

We observe that the confidence intervals seem to follow closely the shape of the original ogive (except at the extremes of the distribution) and so it seems there is some information about the prior distribution coming through the set of binary responses.

Figure 4.7: Boys, test 1, ogives of empirical priors with different numbers of support points and confidence bands of their percentiles

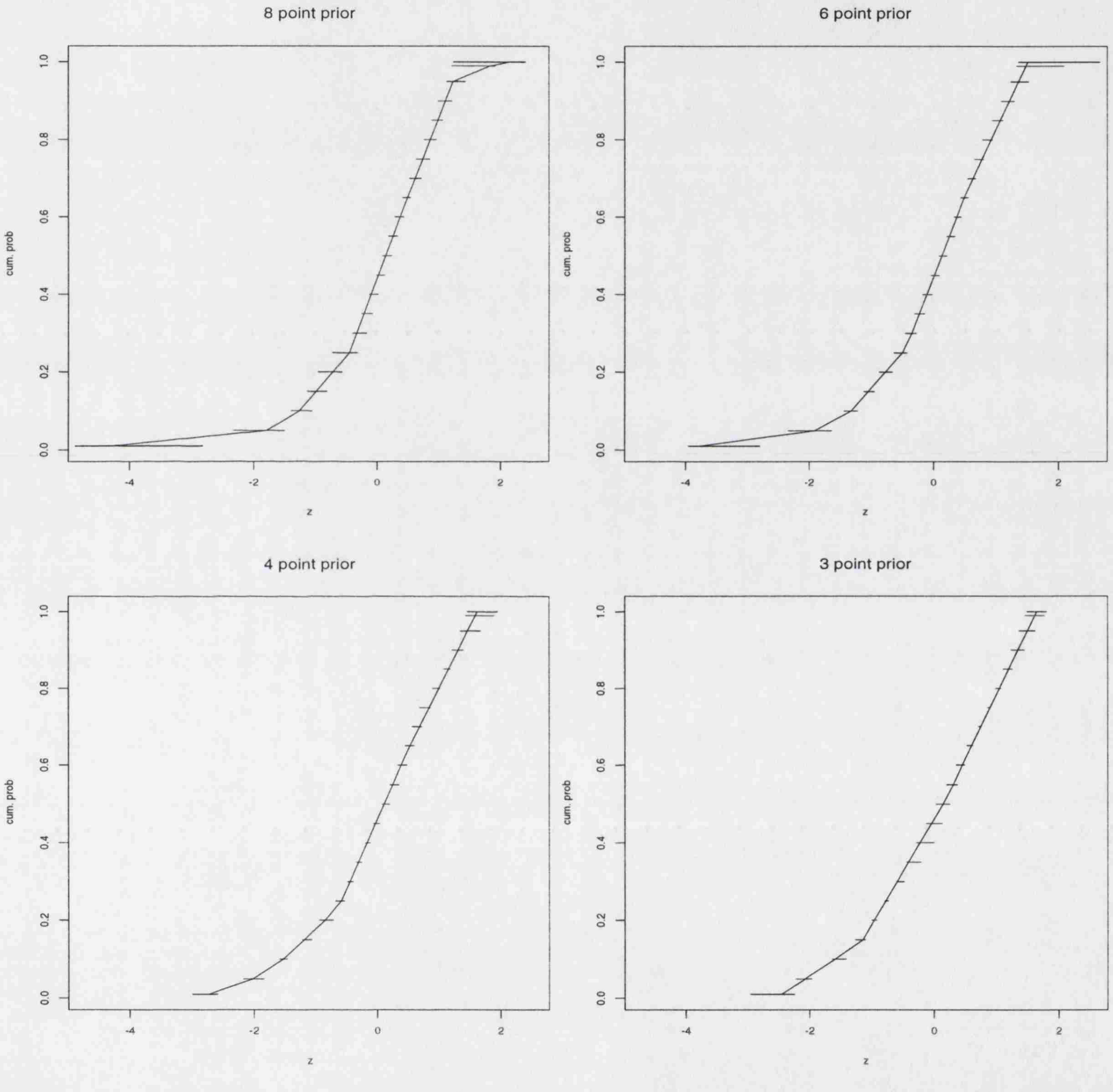
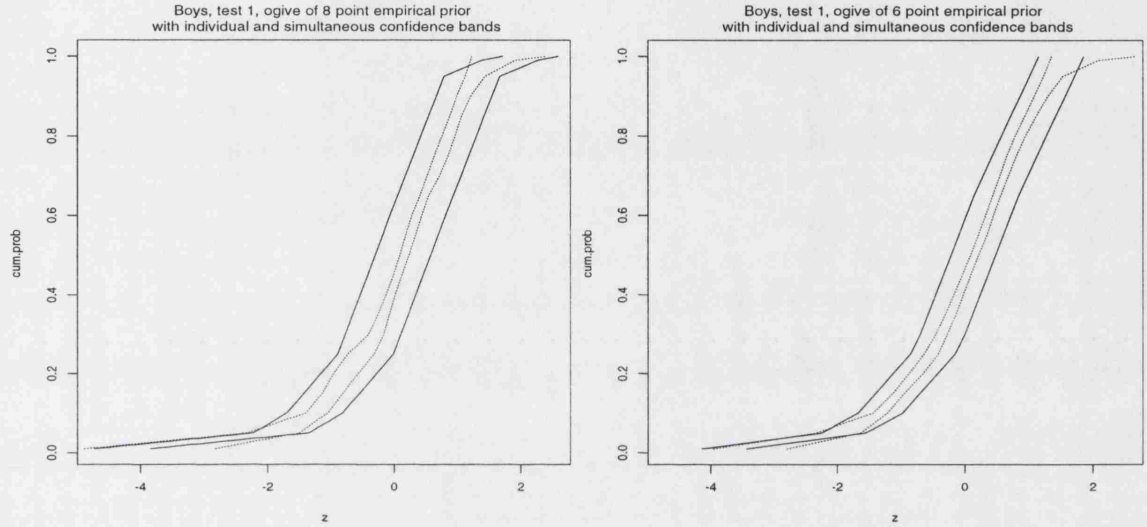


Figure 4.8 shows the confidence bands for the smoothed 8 and 6 point priors, based on Kolmogorov-Smirnov method ('simultaneous' confidence), and based on the percentiles ('individual' confidence) for comparison. We joined the lower and upper bounds of these intervals to form envelopes of the prior. This facilitates comparison between distributions,

although one should be aware that the level of confidence of the percentile confidence intervals does not apply to all of them simultaneously. As expected, the Kolmogorov-Smirnov confidence intervals are much wider.

Figure 4.8: Boys, test 1, Kolmogorov-Smirnov and percentile confidence bands of the 8 and 6 point prior ogives



We shall use these confidence intervals to see whether priors that have been approximated with different number of points have the same shape, to see how well the prior recovers the shape of the distribution used to generate the data, to compare the estimated priors with the N(0,1) and to compare priors between groups that have answered the same set of questions. In the following we will report results for the smoothed priors.

4.3.1 Approximation with different number of points

Figure 4.9 shows the envelopes of the ogives of the 8, 6, 4, and 3 point priors, in pairs, for the NFER test 1, boys data. The envelopes were formed by joining the bounds of the percentile confidence intervals.

We see that there is a complete overlap of the envelopes of the 8 point and 6 point priors, so the 6 point prior can still adequately capture the shape of the prior distribution. The envelopes of the 8 and 4 point priors do not overlap in the whole range of the distribution, which means that some of their percentiles differ significantly, though the distributions are still close to each other. There are more departures between the 8 and 3 point priors. There is a bit of overlap in the ogives at the very beginning, around the middle and a bit at the very end of the distributions, but this only suggests that the middle percentiles may coincide, and the distributions start and end at about the same locations, but this has to do only with the locations of the distributions and these have been held fixed. Therefore 3 points are too few to represent adequately the shape of a continuous prior.

In Figure 4.10 we have the envelopes of the 8, 6, 4 and 3 point priors, for the Boys, test 1 data, based on the Kolmogorov-Smirnov confidence intervals. As these are much wider there is an overlap of the envelopes even of the 8 and 3 point priors (except for the first percentile).

4.3.2 Normality

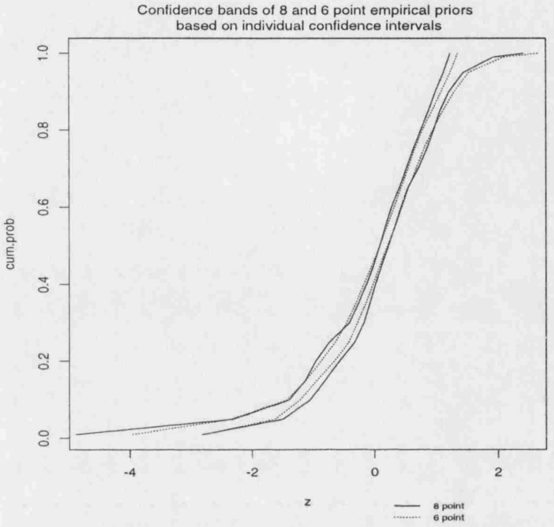
We would like to see how the shape of the prior compares with the N(0,1) particularly because a common assumption about the prior is that it is normally distributed. We therefore plotted the percentiles of the N(0,1) against the percentiles of the empirical prior (QQ-plot) and also against the (joined) bounds of their confidence intervals to see how much they would deviate from a straight line.

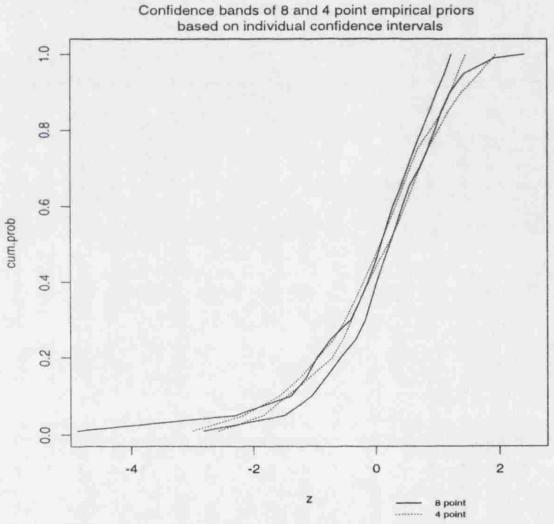
Figure 4.11 shows QQ-plots for the 8 point prior, with its confidence bands based on the percentiles (left) and on the Kolmogorov-Smirnov method (right), for the Boys, test 1 data.

The QQ-plot of the ogive shows a distribution which has a longer left tail than the N(0,1), is skewed to the left and is more contained in the right hand side. We observe that if we consider individual percentile confidence intervals then many of these will not include the corresponding N(0,1) percentile. If we consider the simultaneous confidence intervals then their envelope will include the straight line for the greater part of the distribution, and only the confidence bands of the 0.01, 0.95 and 0.99 percentiles will not include the corresponding N(0,1) percentiles.

The 4 point prior is more skewed than the 8 point one, so now the first 10% of the distribution lies further out to the left than expected from a N(0,1) (Figure 4.12, left). The 3 point prior, Figure 4.12, right, can hardly be said to approximate the N(0,1), since the confidence bands of the percentiles do not include at all the respective N(0,1)

Figure 4.9: Boys, test 1, envelopes of ogives of priors of different number of support points based on individual confidence intervals





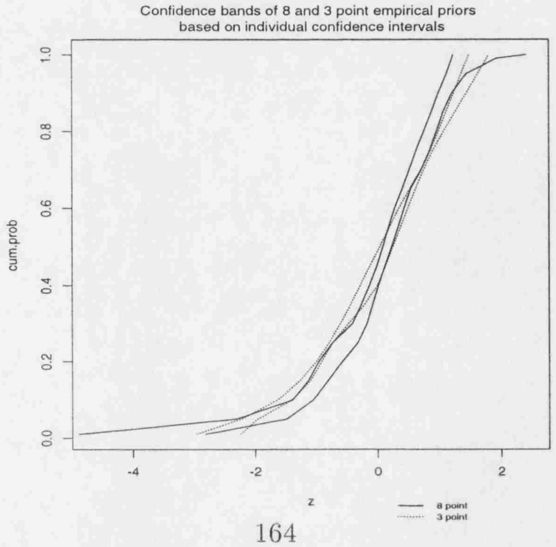
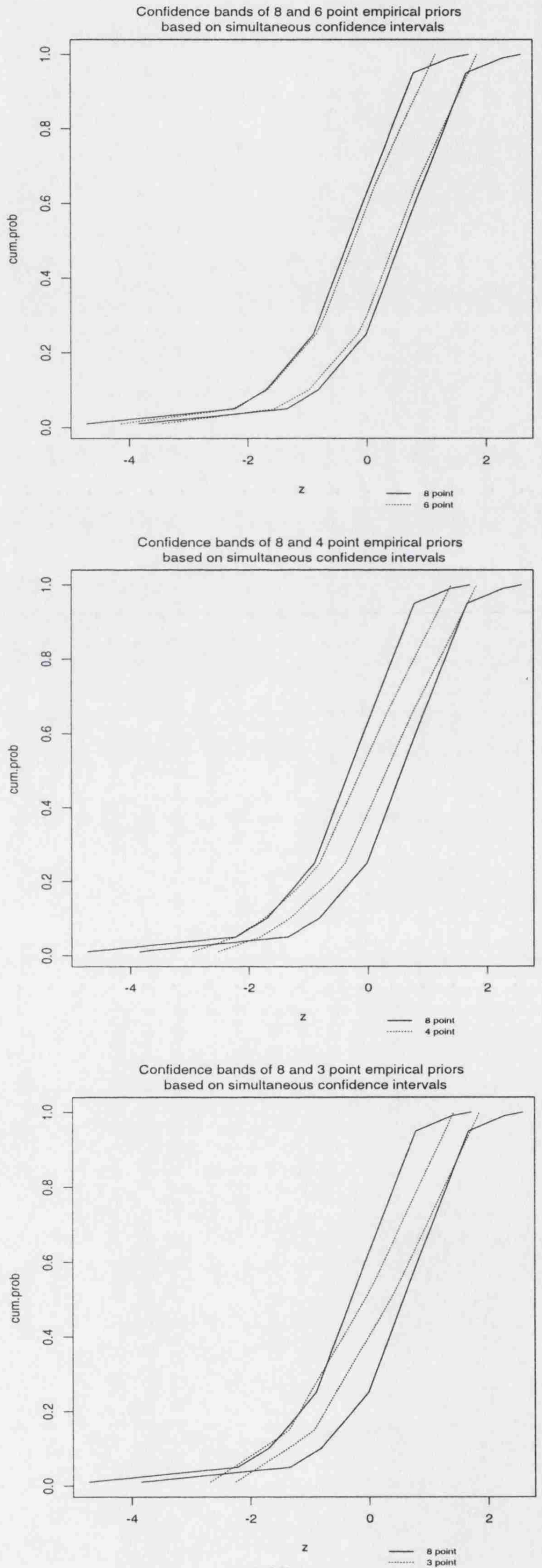


Figure 4.10: Boys, test 1, envelopes of ogives of priors of different number of support points based on simultaneous confidence intervals

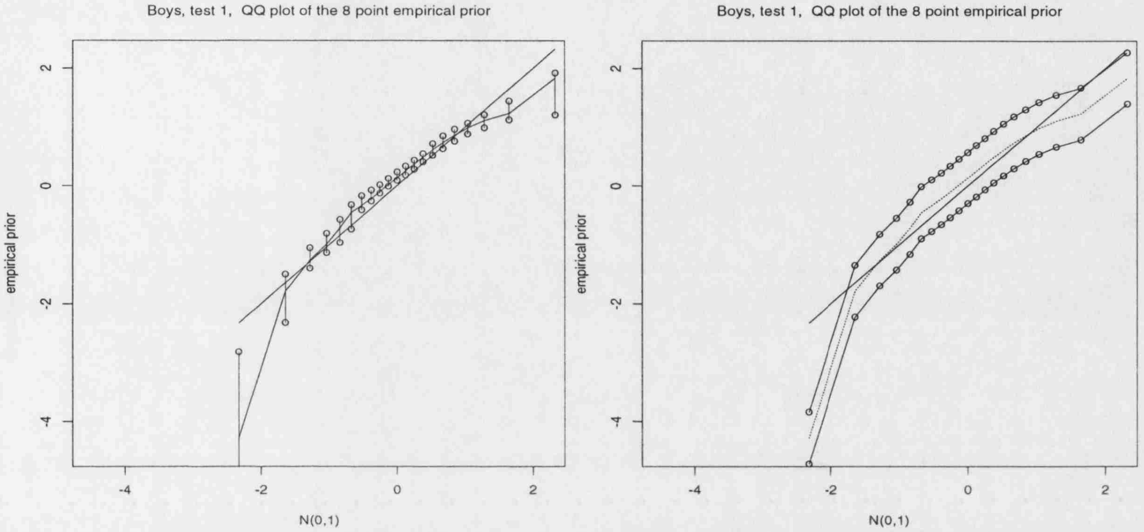


Confidence bands of 8 and 6 point empirical priors based on simultaneous confidence intervals

Confidence bands of 8 and 4 point empirical priors based on simultaneous confidence intervals

Confidence bands of 8 and 3 point empirical priors based on simultaneous confidence intervals

Figure 4.11: Boys, test 1, QQ-plots of 8 point empirical prior with percentile and Kolmogorov-Smirnov bootstrap confidence bands



percentiles or the latter lie just on the borders of the intervals. (Note, in Figure 4.12 only the Kolmogorov-Smirnov confidence intervals are depicted).

4.3.3 Artificial data-mixtures

The information about the underlying distribution that comes through a set of binary responses can also be measured by the extent to which the estimated prior approximates the distribution that generated the data, if the latter is known. Therefore, we generated responses from mixtures of normals and estimated the prior. We will report results for the mixture 0.5 N(0,1) + 0.5 N(3,1) though similar results have been obtained for the 0.5 N(0,1) + 0.5 N(4,1) mixture as well.

As usual, the prior was standardised during the estimation procedure in order to avoid indeterminancies with the parameter estimates.

To compare the estimated prior with the underlying distribution, we plotted the percentiles of the prior, and their bootstrap confidence intervals, against the percentiles of the standardised random deviates used to generated the responses (QQ plots) (Figure 4.13).

We observe that for the 8 point prior, most of the confidence intervals of the percentiles include the straight line, two lie just outside it, and the last two percentiles fall further up

Figure 4.12: Boys, test 1, QQ-plots of 4 and 3 point empirical prior with Kolmogorov-Smirnov bootstrap confidence bands

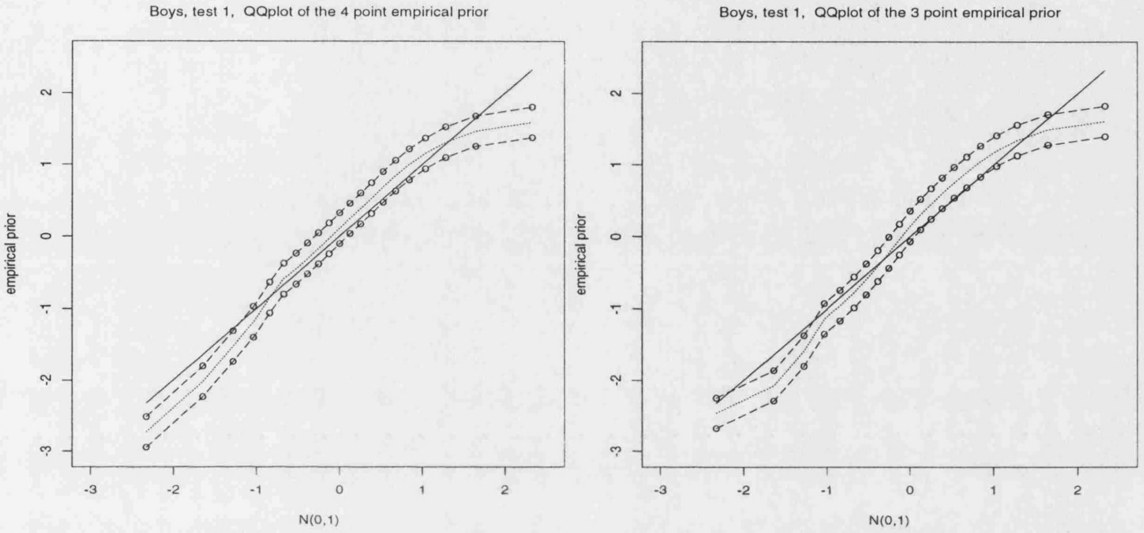
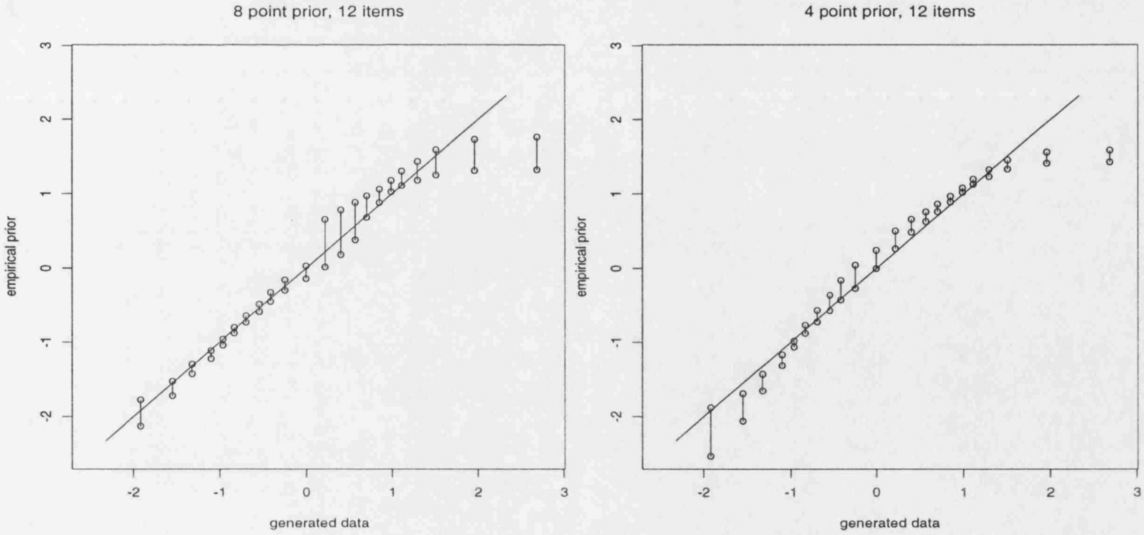


Figure 4.13: Artificial data, QQ-plots of 8 and 4 point empirical priors against the actual latent distribution

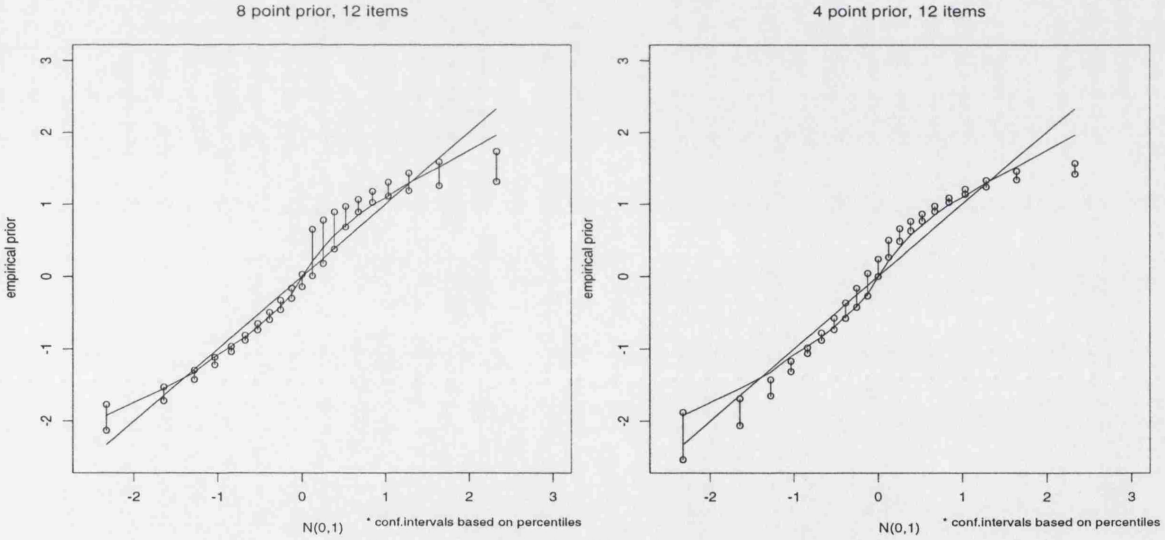


to the right of the 45 degrees line. This shows that the estimated prior has very similar shape to the distribution that generated the data, and only at the upper end - the last 5% of the distribution - it is more contained than the original distribution.

The 4 point prior is not as good as the 8 point prior to approximate the original distribution, with the 45 degree line lying mostly on the bounds of the confidence intervals of the percentiles. (If we take the simultaneous confidence intervals though, the confidence interval bands will include all the original percentiles, except the very last one.)

The QQ plots against the N(0,1) percentiles in Figure 4.14 are more suggestive about the success of the prior in revealing the original distribution of the data. We see that the estimated priors follow the curves of the distribution used to generate the data and they are only deeper than those of the original distribution, particularly if only 4 points are used to estimate it.

Figure 4.14: Artificial data, QQ-plots of actual latent distribution and empirical prior against the N(0,1)


8 point prior, 12 items

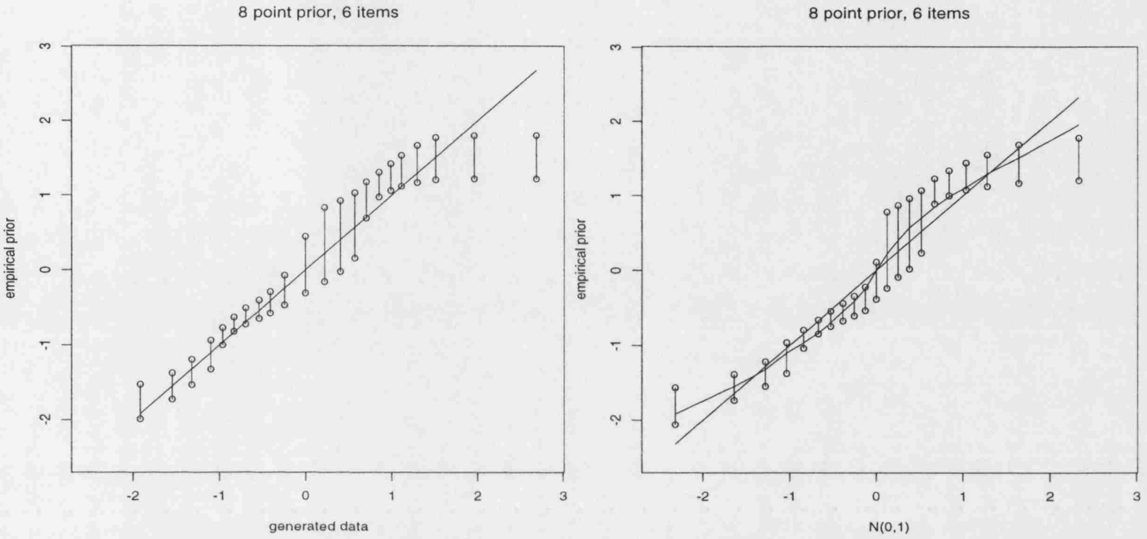
4 point prior, 12 items

* conf.intervals based on percentiles

* conf.intervals based on percentiles

We would also like to see the effect of the number of items in retrieving the form of the latent distribution. Figure 4.15 shows QQ plots of the empirical prior when only 6 items have been used in estimating the model, against the original 'generated' data and the N(0,1). Although the variability of the percentiles is larger, the overall picture is the same, so even with 6 items the form of the distribution that generated the data can be reasonably retrieved.

Figure 4.15: Artificial data, QQ-plots of 8 point empirical prior estimated from 6 items against the actual latent distribution and the N(0,1)



4.3.4 Parametric Bootstrap

With the parametric bootstrap samples are drawn assuming the latent distribution is the estimated discrete prior and the parameters of the model that generate the responses are the estimated parameters. We implemented parametric bootstrapping only for the unsmoothed prior.

The variability of the prior is again inferred from the variability of the percentiles, which have been obtained from the rotated cdf.

The results from the parametric bootstrap are compared with the results from the nonparametric bootstrap, again considering the prior as a discrete distribution. This serves to check the accuracy and efficiency of the two methods.

Figure 4.16 shows the cumulative distribution functions of 8, 4 and 3 point priors for the Boys, test 1 data, together with 90% percentile confidence bands from the parametric and non-parametric bootstrap.

We observe that the bands follow the shape of the cumulative distribution function, for example the jumps of bootstrap distributions are around the jumps of the original distribution. The confidence bands of the parametric bootstrap are narrower than those of the non-parametric bootstrap, they follow the original distribution more closely.

Therefore the non-parametric bootstrap provides relevant and valid information on the distribution of the prior.

The shape of the 8 point cdf is similar to the smoothed ogive of the same distribution. For the 4 and 3 point priors the steps are now more obvious, so this representation is more accurate.

The 3 point prior displays a lot of variability around the steps, and smaller at the jumps, whereas the variability of the 8 point cdf seems more or less the same along the range of the distribution.

Figure 4.17 shows confidence bands of the 8 and 4 point cdfs, the 8 and 3, and the 4 and 3 point cdfs overlaid. The difference in the shapes of the distributions is now more obvious than when looking at their smoothed ogives, but, if we look at the confidence intervals of the percentiles, they will all overlap.

The differences in the cumulative distribution functions are more apparent if we look at their parametric bootstrap confidence bands in Figure 4.18. Some of the confidence intervals of the percentiles of the 8 and 3 point priors don't overlap, for example those corresponding to cumulative probability of between 0.35 and 0.45. The 4 and 3 point cdfs are even more different, and there is less overlap in the confidence intervals of their percentiles, for example the percentiles corresponding to between 0.3 and 0.5 probability (approximately). Also, the cumulative probability corresponding to the z range of about 0.3 and 0.6 is different for the two distributions. It seems that 3 points are too few to approximate a continuous distribution.

4.4 Scoring of the latent variable

In this section we examine how the posterior means obtained from an empirical prior compare to the posterior means obtained from a N(0,1) and also how the number of points of the prior affects the distribution of the posterior means.

Figure 4.19 shows plots of the posterior means obtained from the different empirical priors against the posterior means obtained from the N(0,1) prior, for Boys test 1.

For all priors, most of the posterior means lie between -2 and 2, with only a few reaching out to -3. The posterior means obtained from the models with the 16 and 8 point priors are almost perfectly linearly related with the posterior means obtained with

Figure 4.16: Boys, test 1, cumulative distribution functions of the 8, 4 and 3 point priors, with parametric and non-parametric bootstrap confidence bands

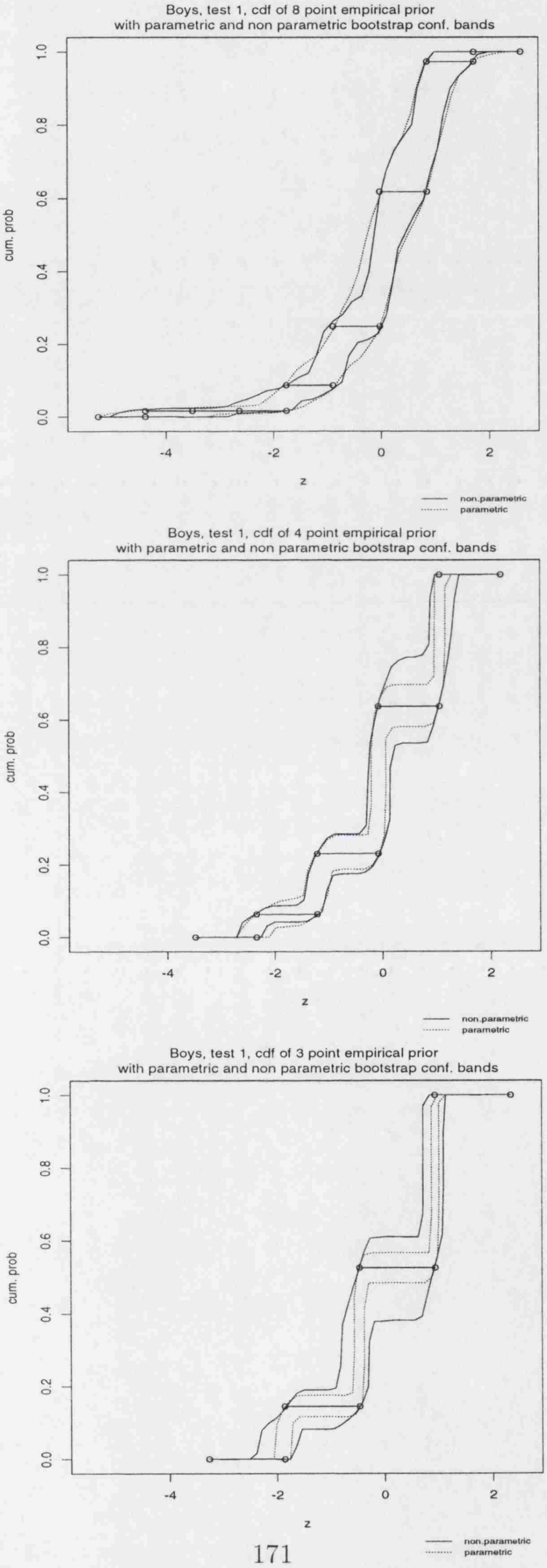
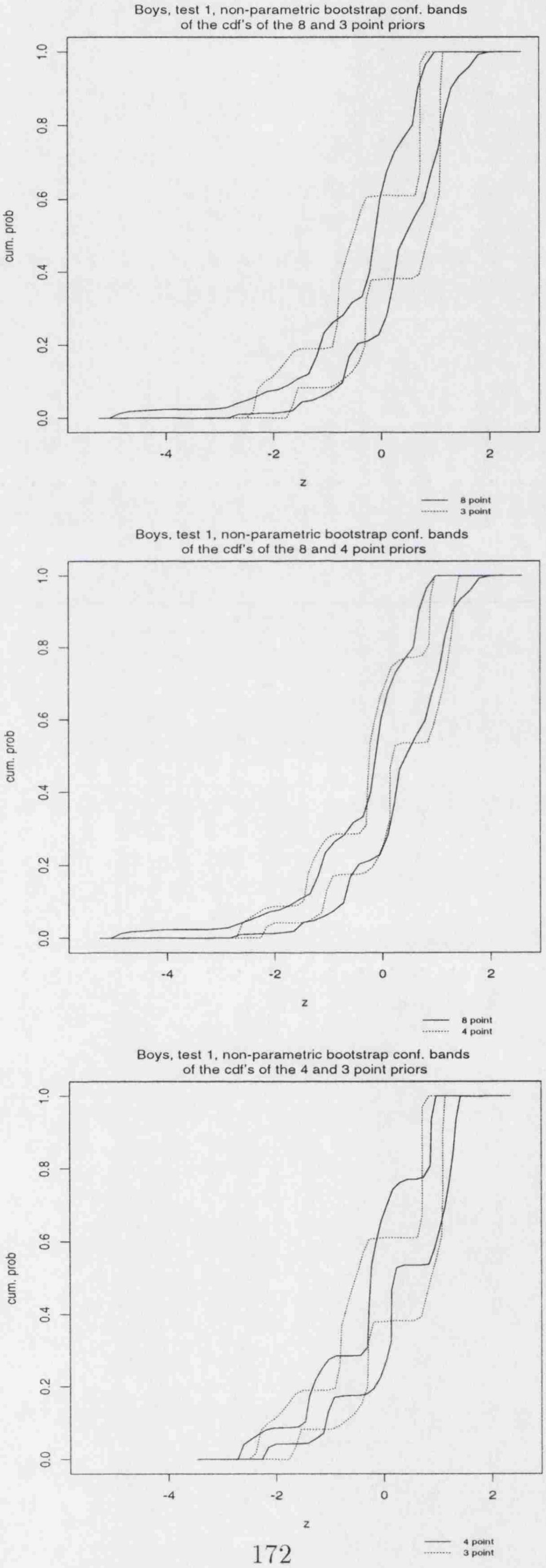


Figure 4.17: Boys, test 1, cumulative distribution functions with nonparametric bootstrap confidence bands overlaid for different priors



Boys, test 1, non-parametric bootstrap conf. bands of the cdf's of the 8 and 3 point priors

Boys, test 1, non-parametric bootstrap conf. bands of the cdf's of the 8 and 4 point priors

Boys, test 1, non-parametric bootstrap conf. bands of the cdf's of the 4 and 3 point priors

Figure 4.18: Boys, test 1, cumulative distribution functions with parametric bootstrap confidence bands overlaid

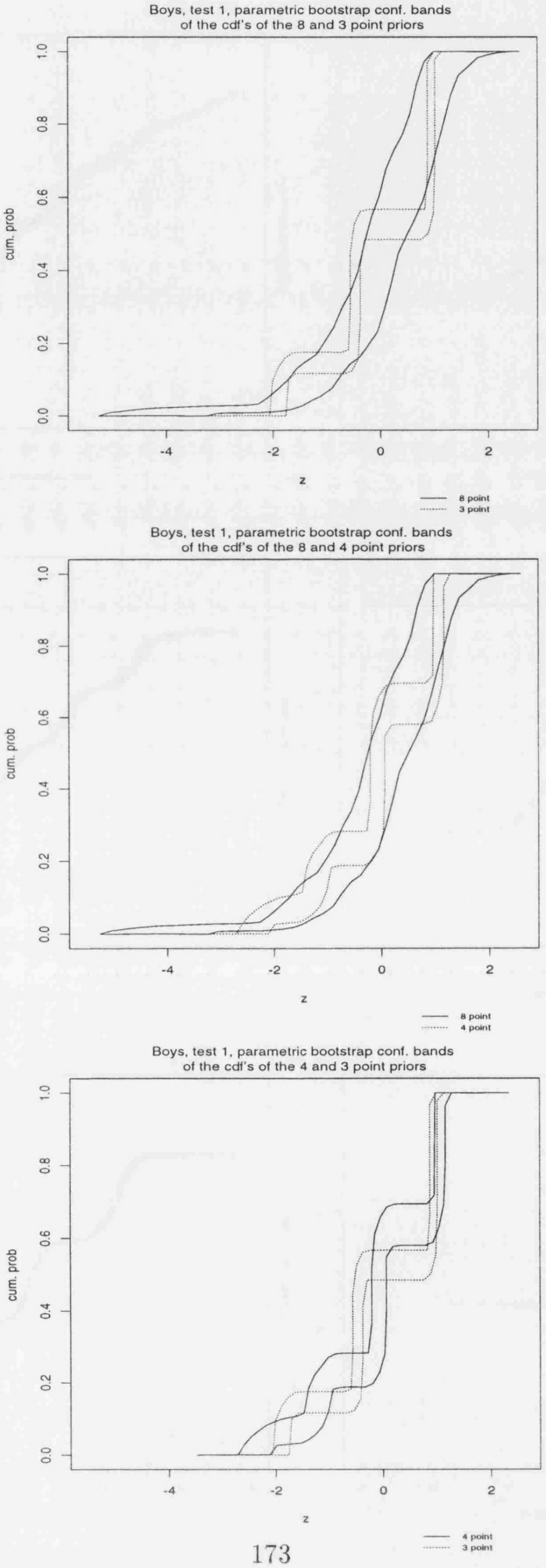
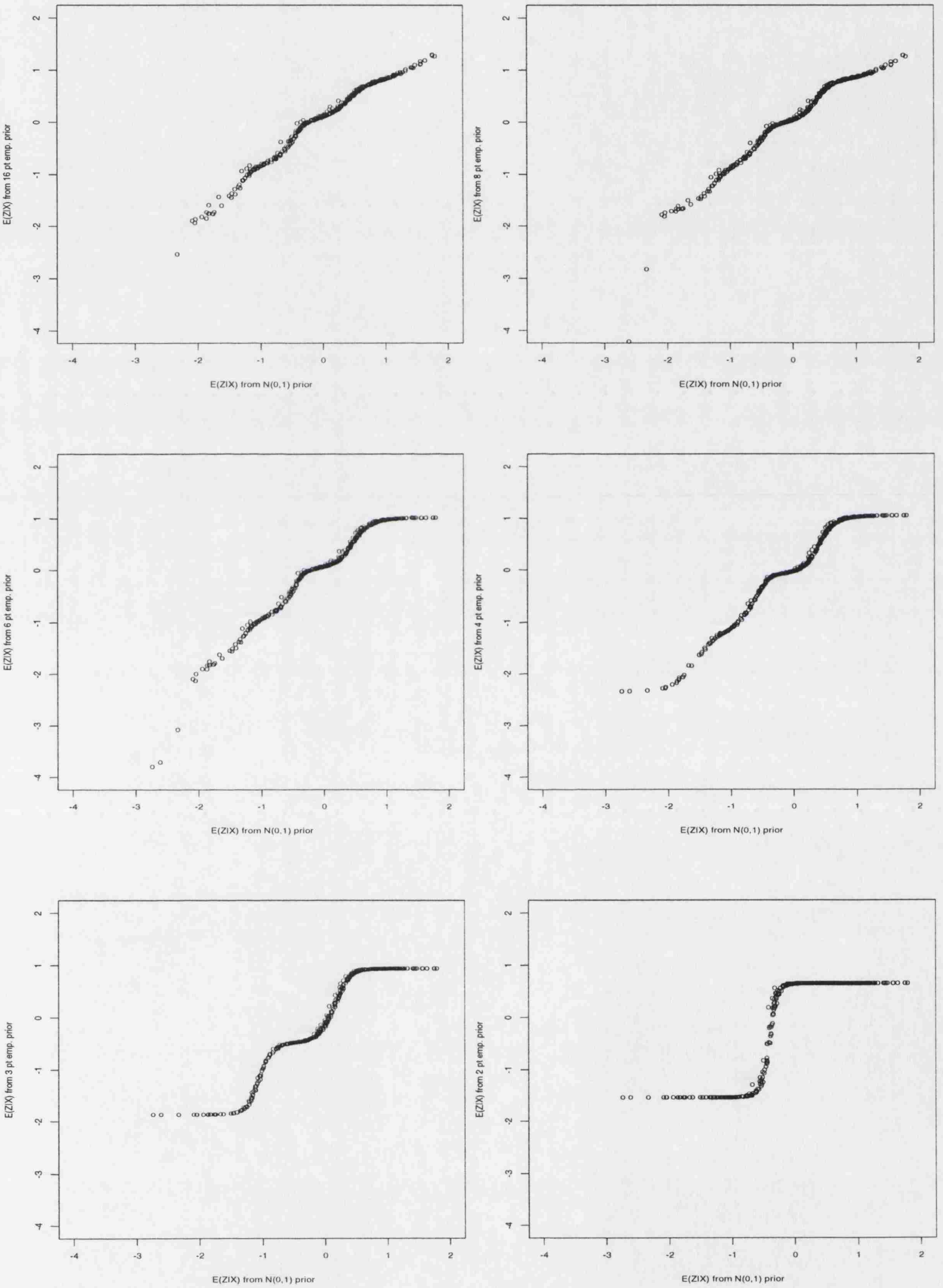


Figure 4.19: Boys, test 1, posterior means from different empirical priors against the posterior means from N(0,1) prior



the N(0,1). This almost true for the posterior means obtained from the 6 point prior, though the distribution is more contained at the right end compared to the one obtained from the N(0,1) or the priors with more approximation points. With the 4 point prior, small 'plateaus' are becoming apparent, which indicates clusters of individuals with very similar posterior means. This is much more obvious with the posterior means obtained from the 3 point and the 2 point priors. We see that there are 3 groups and 2 groups respectively, having almost the same posterior means, whereas the N(0,1) prior or a prior with more approximation points would have spread them out.

We observe the same for the American scale 9 data. Figure 4.20 shows scatterplots of the posterior means obtained with 8, 4, 3 and 2 point priors against the posterior means obtained with a N(0,1) prior. Figure 4.21 shows the cumulative distribution functions of the posterior means obtained with a N(0,1) prior and 16, 10, 6, 4, 3 and 2 priors, and then their quantiles against the N(0,1) quantiles. We see that the posterior means obtained with a N(0,1) are closest to the N(0,1) curve. The posterior means obtained with a 10 point and 6 point prior are quite close together and closer to the posterior means from the N(0,1) than the posterior means obtained from the other priors.

We therefore conclude that more than 4 points are needed if one wants an ordering of the posterior means on a continuous line; with fewer support points clusters around points are formed.

4.5 Comparing populations

Fitting a latent trait model using an empirical prior rather than a fixed one may have advantages when one aims to compare samples that have taken the same test or have answered the same questionnaire. Having a fixed location and scale for the prior, is essential, so as to allow differences in location and scale between the samples to come through the parameter estimates. The empirical prior, a standardised distribution, will then only reveal other characteristics or differences between the latent distributions.

Comparison of priors of different groups Figure 4.22 shows the confidence interval bands (based on the percentiles) of the ogives of the 8 and 6 point empirical priors of the American and German samples that have answered the items of scale 9. We observe that for both the 8 and 6 point priors, the American ogive is more to the left in the lower half

Figure 4.20: American scale 9, E(Z|X) from 8, 4, 3 and 2 point empirical priors against E(Z|X) from N(0,1) prior

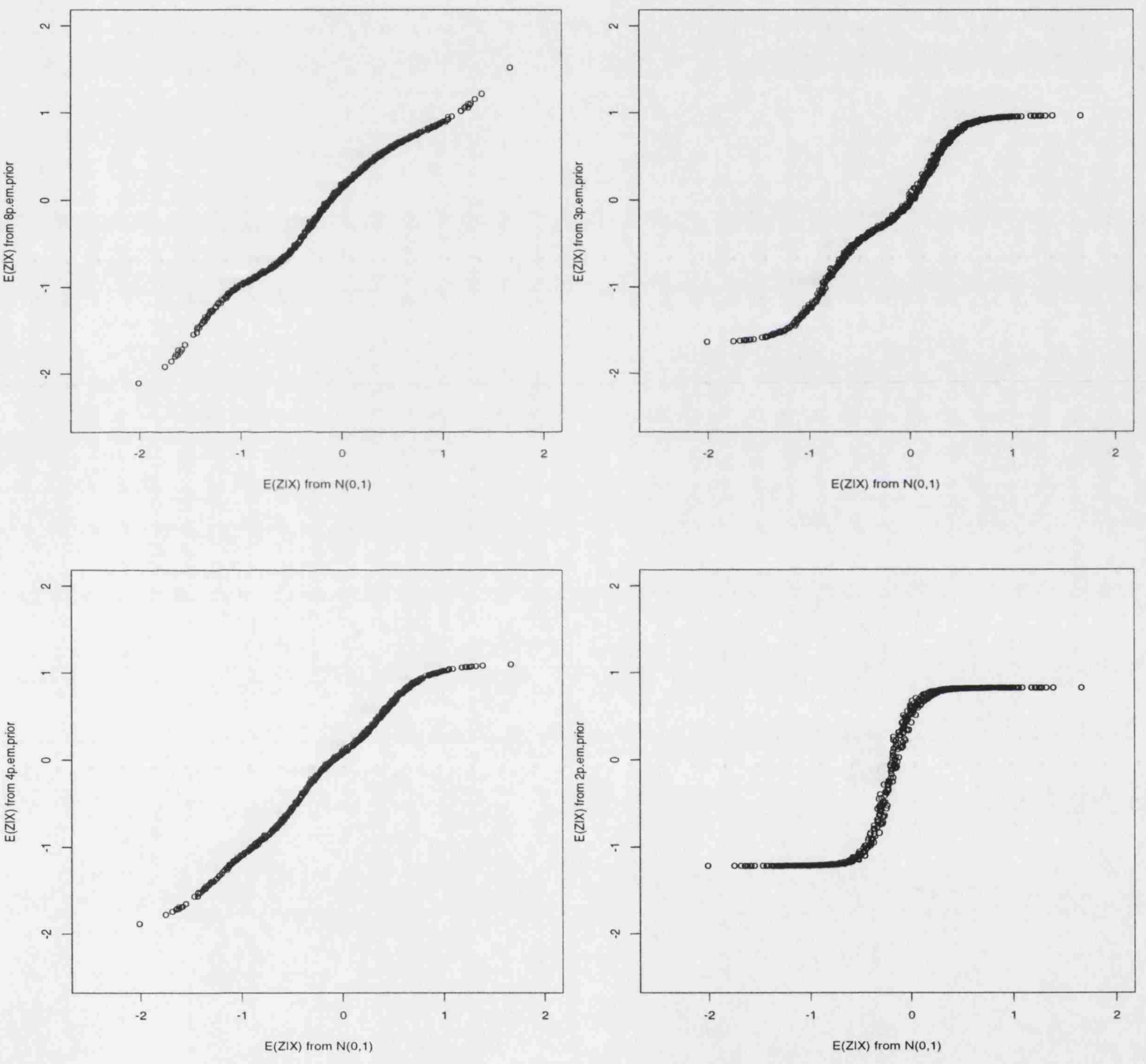
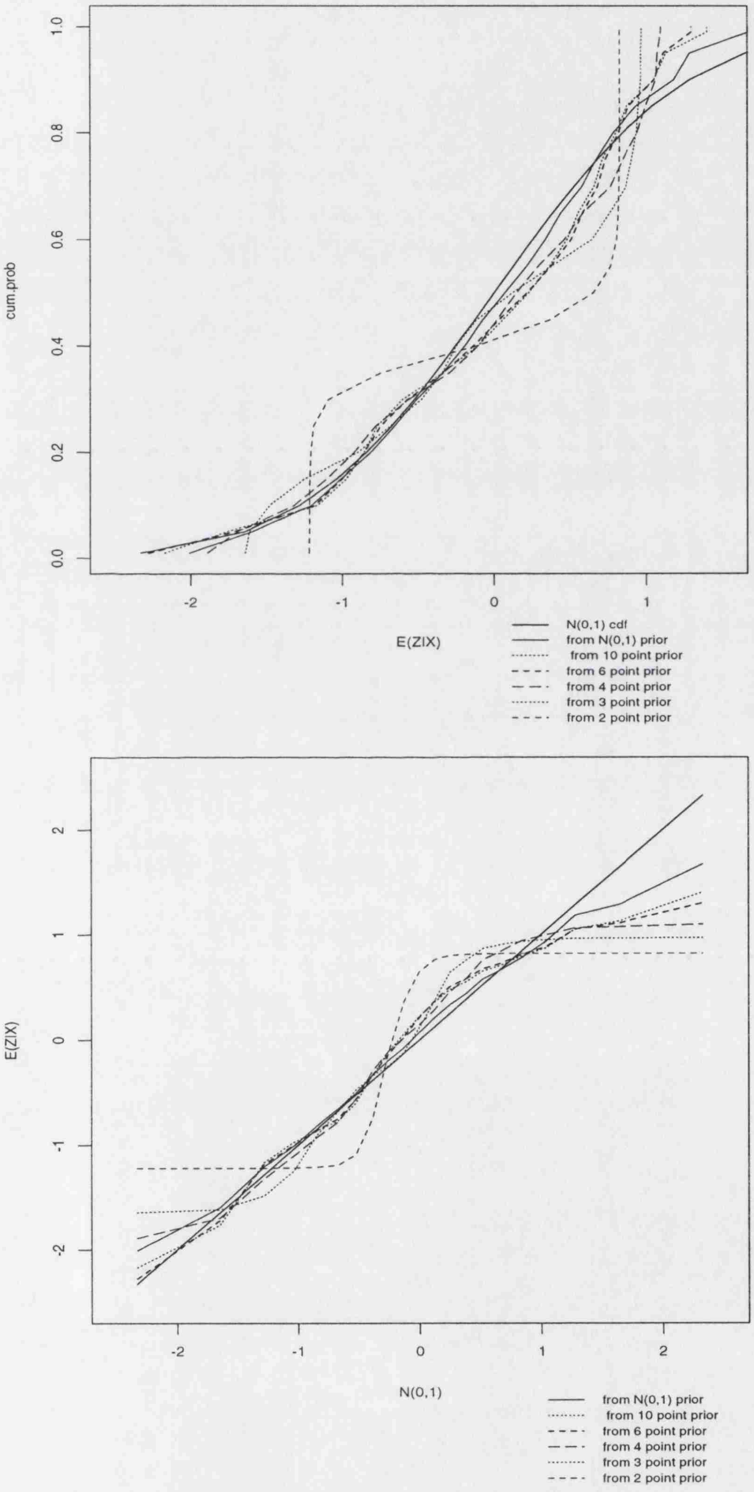
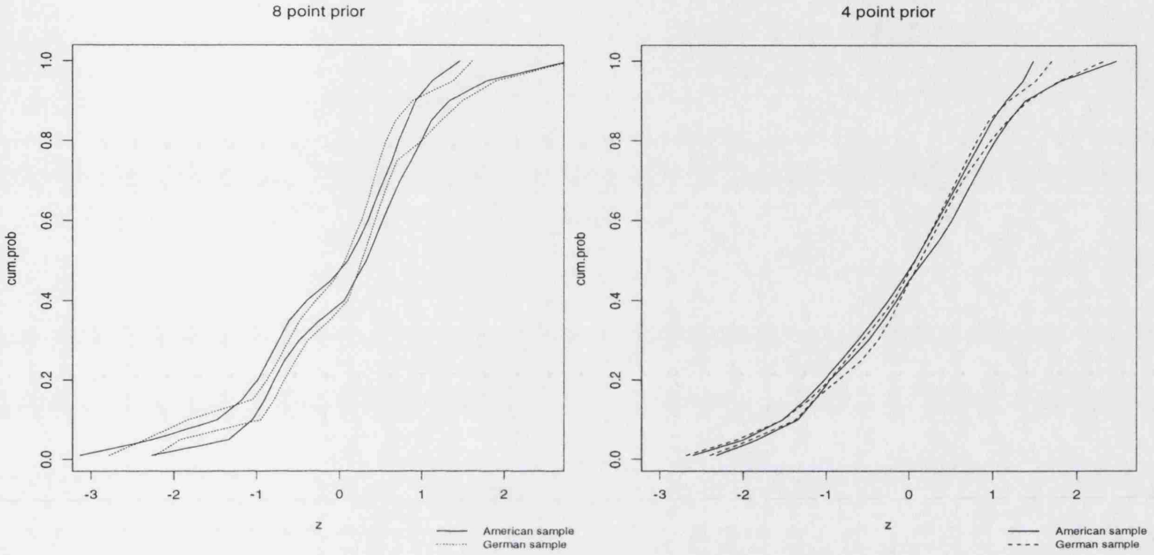


Figure 4.21: American scale 9, ogive and QQ plot of E(Z|X) obtained from various priors



and more to the right in the upper half (disregarding the tails), which implies that the distribution of the American sample is a bit heavier in its left half than the German one whereas the German is heavier in the right half.

Figure 4.22: Ogives of 8 and 4 point priors, for the American and German scale 9

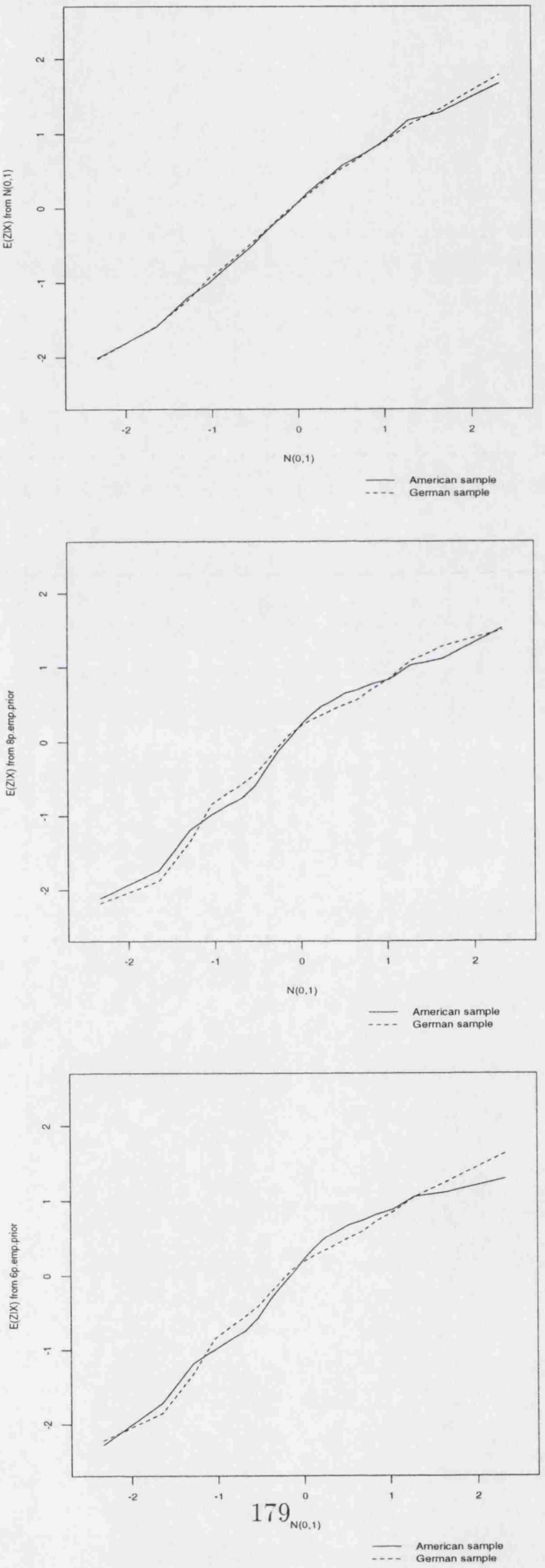


Comparison of posterior means These differences are reflected in the posterior means obtained with an empirical prior.

Figure 4.23 shows QQ plots of the quantiles of the posterior means obtained with a N(0,1) prior, and with an 8 and a 6 point empirical prior. The posterior means obtained with the N(0,1) are close together and both close to the N(0,1). The posterior means obtained with the empirical priors are not as close together. We see that to the left of the centre, the German quantiles are above the American ones, so the distribution of the American sample is heavier than the German one there, whereas the opposite is happening to the right of the centre of the distribution.

So, fitting an empirical prior allows one to compare the shape of the latent distribution of each group, in addition to the information on the location and scale of the distribution which can be obtained from the parameters or the posterior means.

Figure 4.23: American and German scales 9, E(Z|X) from N(0,1) and from 8 and 6 point empirical priors



4.6 Goodness-of-fit

The semiparametric estimation of the latent trait model should improve its goodness-of-fit, since the latent distribution is also being modelled and any discrepancies due to misspecification of the prior should be alleviated. The goodness-of-fit statistics can also help one to decide on the points needed to adequately capture the shape of the prior.

Tables 4.1 and 4.2 show goodness-of-fit statistics for parametric and semiparametric estimation of latent trait model, for the American scale 9 and Boys, test 1 respectively. The definitions of G^2, X^2 and $\%G^2$ are given in Section 1.10. The Akaike Information Criterion (AIC) is given by AIC $= -2l + 2r$ where l is the loglikelihood and r is the number of estimated parameters. The number of estimated parameters is the number of item parameters plus the number of points with non-zero weight ($>$ 0.0000009) minus the number of restrictions on the points and weights, which are 3).

We see that for both the Boys, test 1 data and the American scale 9 data semiparametric estimation with 4 points or more gives a better $\%G^2$ and a larger loglikelihood. For the American scale 9 data G^2 and X^2 are also better for the semiparametric models with more than 3 points, but for the Boys test 1 data at least 8 points are needed to beat the G^2 and X^2 from the N(0,1) prior.

The AIC suggests that the 4 point prior is best for the American scale 9 and the 10 point prior is best for the test 1 data.

Another indication as to how many points are needed to estimated the prior is the relative change in the loglikelihood as the number of points increase. Figures 4.24 and 4.25 show that likelihood increases a lot as we increase the number of points when this is small - less than 4 - but with 4 or more points the changes in the loglikelihood become smaller.

This is consistent with what we have observed so far, that is at least 4 and not more than 8 points are needed to capture the original shape of the generating distribution, and also about so many to have adequate variation in the posterior means.

Table 4.1: American scale 9: Goodness-of-fit statistics

	N(0,1)	10pt st uni	8pt st. uni	6 point	4pt uni	3pt	2pt
$\%G^2$	51.04	51.65	51.65	51.63	51.46	51.04	48.15
log-likelihood	-9876.26	-9862.42	-9862.45	-9863.00	-9866.86	-9876.27	-9941.91
G^2	1597.85	1538.33	1543.82	1537.55	1557.64	1608.30	1770.12
X^2	1316.40	1285.25	1294.98	1281.42	1309.91	1403.48	1754.16
d.f.	96	96	98	96	99	100	111
no.of param.	24	31	28	27	25	24	24
AIC	19800.52	19786.85	19780.90	19780.00	19783.73	19800.54	19931.81

Table 4.2: Boys, test 1: Goodness-of-fit statistics

	N(0,1)	10pt st uni	8pt st. uni	6 pt	4pt uni	3pt	2pt
$\%G^2$	24.09	24.82	24.80	24.73	24.31	23.33	20.51
log-likelihood	-6126.83	-6101.77	-6102.51	-6104.87	-6119.24	-6153.13	-6249.73
G^2	3827.15	3919.07	3681.93	3864.82	3900.83	3888.91	4022.07
X^2	23159.47	29082.41	19402.92	27129.65	27342.44	25219.49	24891.24
no.of param.	42	46	46	45	43	42	42
AIC	12333.67	12295.54	12297.02	12299.74	12324.48	12390.26	12583.46

The degrees of freedom were all negative because of the pooling of the response patterns.

Figure 4.24: American scale 9, loglikelihood against number of points used in simple semiparametric estimation

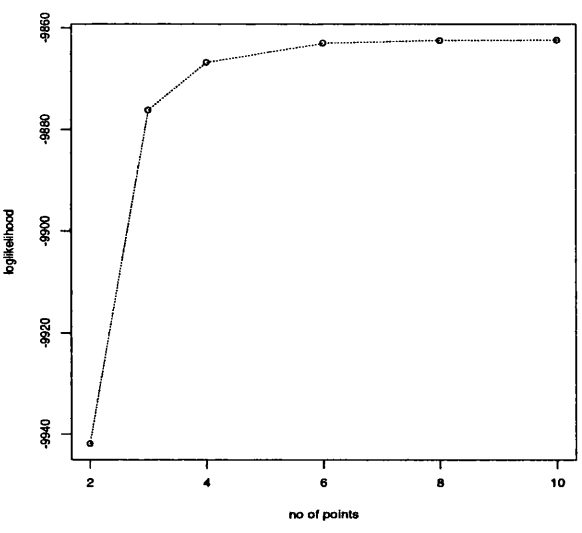
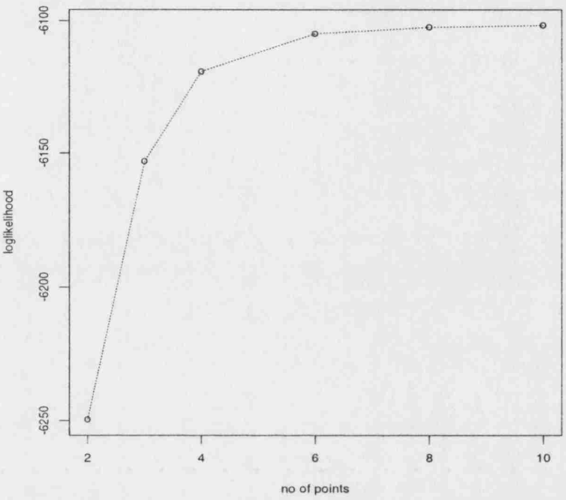


Figure 4.25: Boys, test 1, loglikelihood against number of points used in simple semiparametric estimation



4.7 Fully Semiparametric Estimation of the Latent Trait Model

In this section we will give an EM algorithm for the fully nonparametric Maximum Likelihood estimation of the latent trait model and examine the information that comes out about the prior distribution. The attribute 'fully', which is due to Heinen (1996), emphasises that also the nodes of the prior are estimated from the data.

The estimation of the approximation nodes is quite important, as it frees entirely the estimation from any predefined assumptions or starting values (unless there are local maxima). For example, we saw in Section 4.2.1 that two sets of four points, one with equidistant and one with non-equidistant points, gave different approximation points for the prior. Estimating both the location and the weights of the points will provide a better approximation for the form of the latent distribution.

4.7.1 Maximum Likelihood Estimation

In Chapter 1, Section 1.4.1 we gave the E-M algorithm that is being used to estimate the unknown parameters when the prior has a known parametric form. Now, in order to estimate the z's we break the estimation into two parts, that is there are two E- and M-steps: the first E- and M-steps apply to the estimation of the item parameters, whereas the second E- and M-steps apply to the estimation of the latent nodes.

To update the nodes we move each node by a proportion of the one-dimensional derivatives of the loglikelihood w.r.t. the relevant node.

As in Section 4.2, we standardise the prior after each M-step, and then adjust the parameters accordingly, so that we avoid the indeterminacy of the location and scale of the prior.

In order to avoid maximization under restrictions, we will differentiate the loglikelihood calculated with the standardised $\tilde{z}$'s w.r.t. the unstandardised z's and we will update the $\tilde{z}$'s with these quantities.

Then the loglikelihood will be again calculated with the standardised $\tilde{z}$'s and standardised α's, since the likelihoods under standardised prior and parameters and unstandardised prior and parameters are equivalent.

In the following we will denote the weight of node z_t with p_t, and with $\bar{z}$ and σ_z the mean and standard deviation of the prior respectively.

By differentiating the log-likelihood with respect to the unknown z's we get:

$$\begin{aligned}
\frac{\partial L}{\partial z_t} &= \sum_{h=1}^{n} \frac{\partial \ln f(\mathrm{x}_h)}{\partial z_t} \\
&= \sum_{h=1}^{n}\sum_{v=1}^{k} \frac{\partial \ln f(\mathrm{x}_h)}{\partial \tilde{z}_v}\frac{\partial \tilde{z}_v}{\partial z_t} \\
&= \sum_{h=1}^{n}\sum_{v=1}^{k} \frac{\partial f(\mathrm{x}_h)}{\partial \tilde{z}_v}\frac{1}{f(\mathrm{x}_h)}\frac{\partial \tilde{z}_v}{\partial z_t} \\
&= \sum_{h=1}^{n}\sum_{v=1}^{k}\sum_{j=1}^{k} \frac{\partial g(\mathrm{x}_h|\tilde{z}_j)h(\tilde{z}_j)}{\partial \tilde{z}_v}\frac{1}{f(\mathrm{x}_h)}\frac{\partial \tilde{z}_v}{\partial z_t}
\end{aligned} \tag{4.2}$$

Now,

$$\begin{aligned}
\sum_{j=1}^{k} \frac{\partial g(\mathrm{x}_h|\tilde{z}_j)}{\partial \tilde{z}_v} &= \frac{\partial \ln g(\mathrm{x}_h|\tilde{z}_v)}{\partial \tilde{z}_v} g(\mathrm{x}_h|\tilde{z}_v) \\
&= \frac{\partial \sum_{i=1}^{p} x_{ih} \ln \pi_i(\tilde{z}_v) + (1 - x_{ih}) \ln(1 - \pi_i(\tilde{z}_v))}{\partial \tilde{z}_v} g(\mathrm{x}_h|\tilde{z}_v) \\
&= \sum_{i=1}^{p} \frac{(x_{ih} - \pi_i(\tilde{z}_v))}{\pi_i(\tilde{z}_v)(1 - \pi_i(\tilde{z}_v))}\frac{\partial \pi_i(\tilde{z}_v)}{\partial \tilde{z}_v} g(\mathrm{x}_h|\tilde{z}_v)
\end{aligned}$$

$$\begin{aligned} &= \sum_{i=1}^{p} \frac{(x_{ih} - \pi_i(\tilde{z}_v))}{\pi_i(\tilde{z}_v)(1-\pi_i(\tilde{z}_v))} \alpha_{1i}\pi_i(\tilde{z}_v)(1-\pi_i(\tilde{z}_v))g(\mathbf{x}_h|\tilde{z}_v) \\ &= \sum_{i=1}^{p}(x_{ih} - \pi_i(\tilde{z}_v))\alpha_{1i}g(\mathbf{x}_h|\tilde{z}_v) \qquad (4.3) \end{aligned}$$

because

$$\frac{\partial \pi_i(\tilde{z}_v)}{\partial \tilde{z}_v} = \alpha_{1i}\pi_i(\tilde{z}_v)(1-\pi_i(\tilde{z}_v)) \qquad (4.4)$$

and,

$$\begin{aligned} \frac{\partial \tilde{z}_v}{\partial z_t} &= \frac{\partial(z_v - \bar{z})/\sigma_z)}{\partial z_t} \\ &= \frac{\partial((z_v - \bar{z})/\partial z_t)\sigma_z - (z_v - \bar{z})(\partial \sigma_z/\partial z_t)}{\sigma_z^2} \qquad (4.5) \end{aligned}$$

Using

$$\begin{aligned} d_t &= \partial \sigma_z/\partial z_t \\ &= -\sum_{j=1, j\neq t}^{k} p_j p_t z_j + p_t(1-p_t)z_t \\ &= -\sum_{j=1}^{k} p_j p_t z_j + p_t z_t \qquad (4.6) \end{aligned}$$

and evaluating the derivative at $\bar{z} = 0$ and $\sigma_z = 1$, we have
if $v \neq t$

$$\begin{aligned} \frac{\partial \tilde{z}_v}{\partial z_t} &= -\partial \bar{z}/\partial z_t - z_v \partial \sigma_z/\partial z_t \\ &= -p_t - z_v d_t \qquad (4.7) \end{aligned}$$

and if $v = t$

$$\begin{aligned} \frac{\partial \tilde{z}_v}{\partial z_t} &= 1 - \partial \bar{z}/\partial z_t - z_t \partial \sigma_z/\partial z_t \\ &= 1 - p_t - z_t d_t \qquad (4.8) \end{aligned}$$

So, (4.2) becomes

$$\begin{aligned}
\frac{\partial L}{\partial z_t} &= \sum_{h=1}^{n}\sum_{v=1}^{k}\sum_{i=1}^{p}(x_{ih} - \pi_i(\tilde{z}_v))\alpha_{1i}g(\mathbf{x}_h|\tilde{z}_v)h(\tilde{z}_v)\frac{1}{f(\mathbf{x}_h)}\frac{\partial \tilde{z}_v}{\partial z_t} \\
&= \sum_{v=1}^{k}\sum_{i=1}^{p}\sum_{h=1}^{n}(x_{ih} - \pi_i(\tilde{z}_v))\alpha_{1i}h(\tilde{z}_v|x_h)\frac{\partial \tilde{z}_v}{\partial z_t} \\
&= \sum_{v=1,v\neq t}^{k}\sum_{i=1}^{p}\alpha_{1i}(r_{iv} - \pi_i(\tilde{z}_v)N_v)(-p_t + z_v d_t) + \sum_{i=1}^{p}\alpha_{1i}(r_{it} - \pi_i(z_t)N_t)(1 - p_t + z_t d_t) \\
&= \sum_{v=1}^{k}\sum_{i=1}^{p}\alpha_{1i}(r_{iv} - \pi_i(\tilde{z}_v)N_v)(-p_t + z_v d_t) + \sum_{i=1}^{p}\alpha_{1i}(r_{it} - \pi_i(z_t)N_t) \qquad (4.9)
\end{aligned}$$

where

$$r_{iv} = \sum_{h=1}^{n} x_{ih}h(\tilde{z}_v \mid \mathbf{x}_h) \qquad (4.10)$$

$$N_v = \sum_{h=1}^{n} h(\tilde{z}_v \mid \mathbf{x}_h) \qquad (4.11)$$

and $d_t = \partial\tilde{z}_v/\partial z_t$.

The weights are calculated by

$$p_t = \sum_{h=1}^{n} h(z_t|\mathbf{x}_h)/n \qquad (4.12)$$

We define the steps of an EM algorithm as follows:

- **step1** Choose starting values for α_{i0} and α_{i1} and the approximating points and weights of the prior

- **step 2: E-step 1** Compute the values of r_{it} and N_t from (1.28)and (1.29)

- **step 3: M-step 1** Obtain improved estimates of the parameters by solving (1.30)

- **step 4: E-step 2** Compute the values of r_{iv} and N_v from (4.10) and (4.11)

Table 4.3: American scale 9: Estimated prior

z_t	$h(z_t)$
-2.287	0.063
-0.937	0.273
0.042	0.264
0.842	0.364
2.184	0.037

- **step 5: M-step 2** Obtain improved estimates of the nodes by solving (4.9) and calculate weights by solving (4.12).

  Standardise the nodes so that the mean of the prior equals 0 and its standard deviation equals 1. Adjust the parameters appropriately. Calculate the likelihood and check convergence. Return to step 2 and continue until convergence is attained.

4.7.2 Results

American scale 9 Table 4.3 shows the estimated approximation points and weights for the American scale 9. The scale consists of 12 items. It is interesting to see how many approximation points are needed or can be estimated from this number of items.

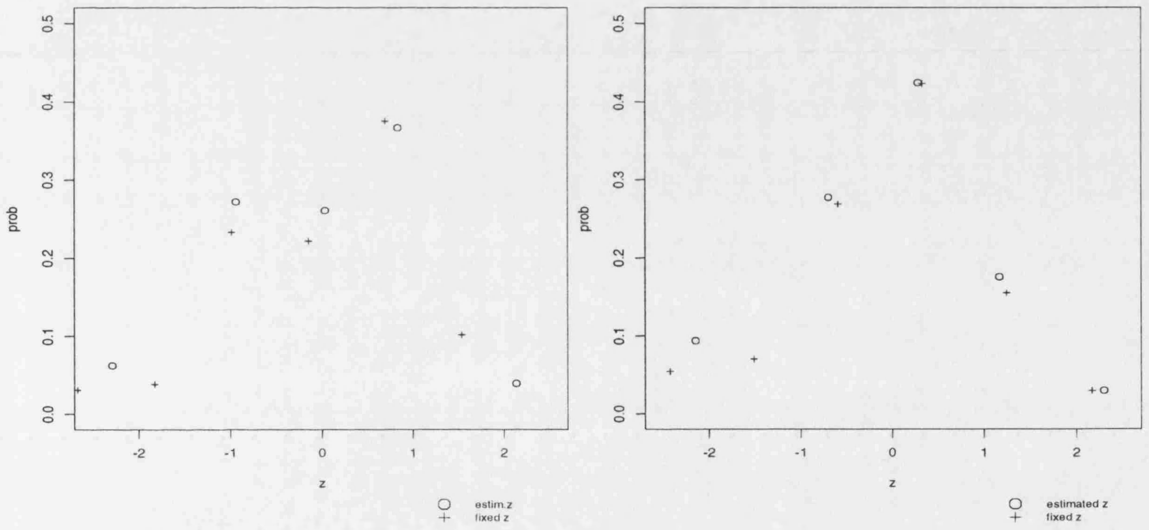
We started with 10, 8, 6 and 5 points. All these sets gave 5 points with non-zero weight. The 5 point solution from the 5 starting points is the same as the 5 point solution from the 6 points, which indicates that there is a unique 5 point approximation for the latent distribution.

We also fitted a 4 point prior, starting from two different sets of points, one with equidistant and one with non-equidistant points (as in Section 4.2.1). Both sets gave the same final solution, indicating again that there is a unique 4 point solution. The semiparametric estimation had produced different results for the two different starting values, which is understandable since the points were fixed and their number too small for the weights to adjust between them.

The same procedure was repeated for German scale 9 and the Boys, test 1 data. The German scale 9 also gave 5 distinct points for the prior (identical solution with when we started with 6 points) and the Boys, test 1, which consists of 21 items, gave 6 distinct points for the prior, a lot fewer than the number of items.

We can see how the points obtained from fully semiparametric estimation compare with the ones obtained from simple semiparametric estimation. In Figure 4.26 we see the points and weights obtained from a simple 6 point semiparametric solution and a 5 point fully semiparametric solution for the American and German scale 9. (We are not comparing with the 5 point simple semiparametric solution because one of the points came out with zero weight). Both solutions seem very similar. The last 4 points and weights are very close together. The last two points of the fixed point prior are an approximation to the first point of the fully semiparametric prior. Since the fully semiparametric prior is more flexible to adjust, it is more economical with points.

Figure 4.26: Points and weights of 5 point fully semiparametric prior and 6 point semiparametric prior, American scale 9, left, and German scale 9, right



Relationship of the number of items with the number of approximation points

We saw in Section 4.2.1 that there have been results (De Leeuw and Verhelst 1986), (Lindsay, Clogg, and Grego 1991) establishing the maximum number of points needed for the prior for the Rasch model as $(p+1)/2$ if p, the number of items, is odd, and $(p+2)/2$ if p is even, with the first point being equal to $-\infty$ in the latter case.

We investigate empirically whether there is a connection between the number of support points with the number of items of the test. We would like to see whether there are

Table 4.4: Sets of items of German scale 9: Number of points of prior

no of items	no of points needed	max.no of points expected
12	5	7
11	5	6
10	4	6
9	5	5
8	4	5
7	4	4
6	3	4
5	3	3
4	3	3
3	3	2

results for the two-parameter logistic model, similar to those for the Rasch model.

We constructed sets of items by reducing the number of items sequentially for three datasets, and estimated the prior with different number of points, until all distinct points had non-zero weight. The results for the German scale 9, the mixture of normals and the Boys test 1 data appear in Tables 4.4, 4.5 and 4.6 respectively.

Both German scale 9 and the mixture 'n30' have 12 items and 1490 responses. Boys test 1 has 21 items and 566 responses. For the first two datasets results are the same except for the 6 item solution. The number of points with non-zero weight are in most cases fewer than the maximum number of points suggested by De Leeuw and Verhelst (1986). The only case that exceeds the number of points expected is the case of the first 3 items of German scale 9, where 3 points came out with non-zero weight.

For the Boys test 1 data fewer points are being estimated for the prior, which is probably due to the small number of responses.

4.7.3 Optimality criteria

Lindsay (1983) gave the conditions for the estimated mixing distribution to be optimal, as we saw in Section 1.9. The D function (1.64) will be in our case:

$$D(z,f)=\sum_{l}^{\mathrm{NR}} n_l\left(\frac{f(x_l|z)}{f(x_l)}-1\right) \tag{4.13}$$

where n_l is the frequency of the distinct response pattern x_l and NR the number of

Table 4.5: Sets of items of 'Mixture of Normals': Number of points of prior

no of items	no of points needed	max.no of points expected
12	5	7
11	5	6
10	4	6
9	5	5
8	4	5
7	4	4
6	4	4
5	3	3
4	3	3

Table 4.6: Sets of items of Boys test 1: Number of points of prior

no of items	no of points needed	max.no of points expected
21	6	11
15	5	8
12	4	7
8	3	5

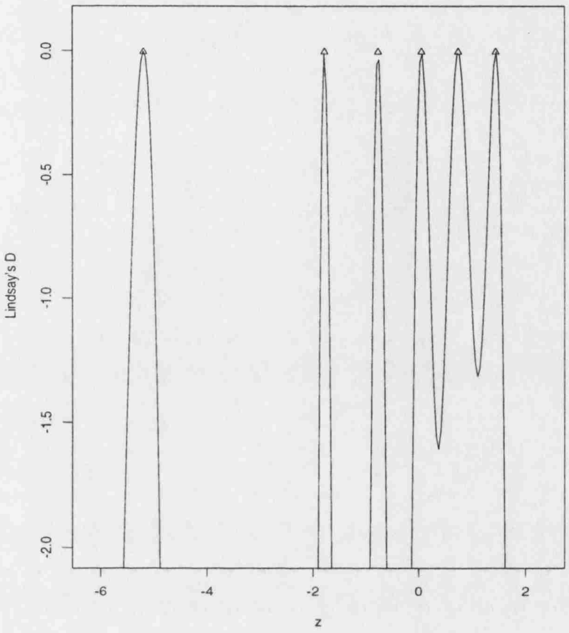
distinct response patterns. For fixed item parameter estimates, $D(z, \hat{f})$ should be equal to zero at each estimated support point z. Moreover $\sup_z D(z, \hat{f}) = 0$, if $\hat{f}$ is from an optimal prior.

Results Boys, test 1. After successively fitting the model with different number of points, we obtained 6 distinct points with non-zero weight for the prior. Figure 4.27 shows a plot of Lindsay's D against the latent variable, and the support points along the zero line. We see that the estimated points fall on the peaks of the D curve, these are at 0, and the D curve does not cross the zero line. This shows that the solution for the prior is optimal.

German scale 9, 12 items. The estimated prior has five support points. Figure 4.28, left shows Lindsay's D curve and the support points. All five points fall exactly on the peaks of the curve, which are at zero, and again the D curve does not go above the zero line. So this solution seems to be the optimal one.

Let us see what happens if we fit a 4 point prior. The D curve, Figure 4.28 right, shows clearly that this is not optimal. There are peaks around the support points going

Figure 4.27: Boys, test 1, Lindsay's D for the 6 point prior



above zero, indicating that an extra point is needed to bring them down.

For the sets consisting of 4 to 11 items the prior came out with 3 to 5 approximation points, the number being equal or smaller than the maximum number of approximation points needed for the Rasch model. For the 3 items set though, we obtained 3 distinct points. Function D is plotted in Figure 4.29 for the 3 and 2 point priors. In the 3 point solution, the line is flat and goes along 0 after about z=-0.2, so both the second and third points have D=0. With the 2 point prior D also remains at zero above -0.2 approximately) so also this solution is optimal. In this case the number of points cannot be determined uniquely.

4.7.4 Scoring of the latent variable

Figure 4.30 shows a QQ plot of the posterior means obtained from a N(0,1) prior, from a 8 point empirical (simple semiparametric) prior and from the 5 point optimal fully-semiparametric prior. Posterior means obtained from the 5 point fully semiparametric estimation are identical to the posterior means obtained from the 8 point simple semiparametric estimation (the two lines coincide). So, simple and fully semiparametric estimation methods can provide the same results, but with the simple semiparametric estimation more points are needed so that the weights can adjust optimally along these points.

Figure 4.28: German scale 9, Lindsay's D for the 5 and 4 point fully-semiparametric prior

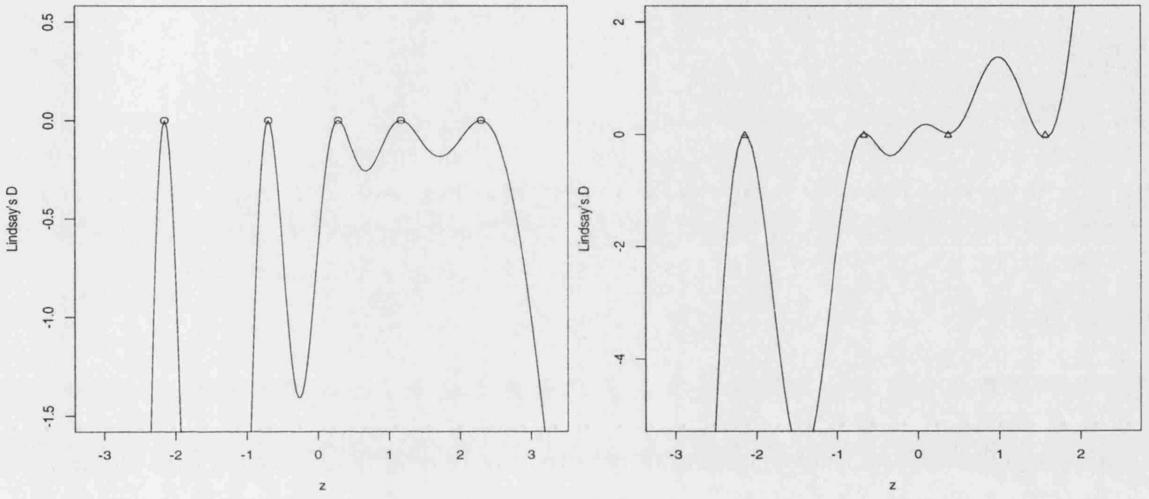


Figure 4.29: German scale 9, first 3 items, Lindsay's D for the 3 and 2 point fully-semiparametric prior

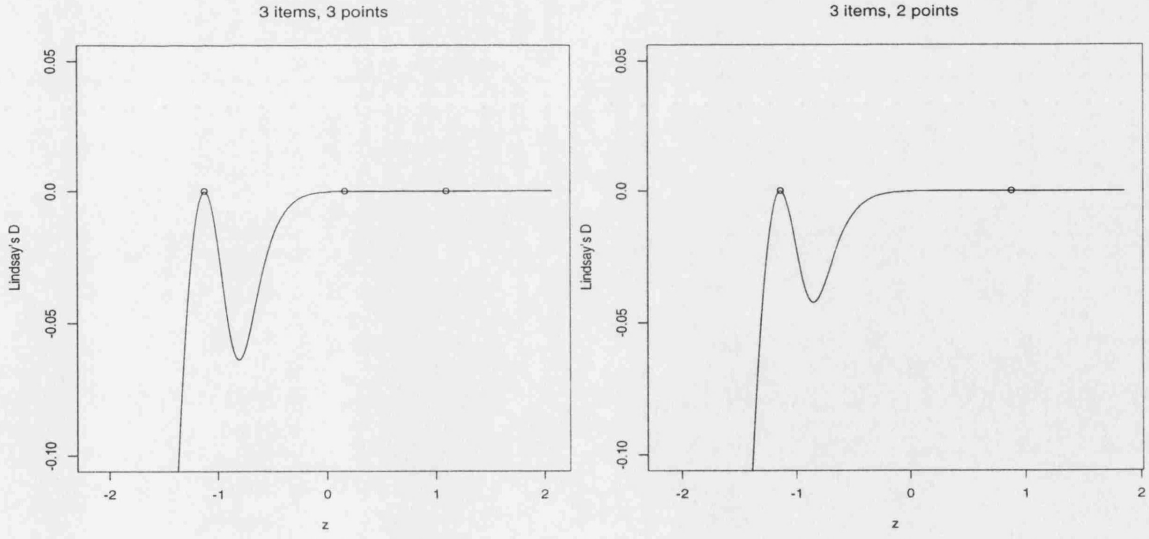
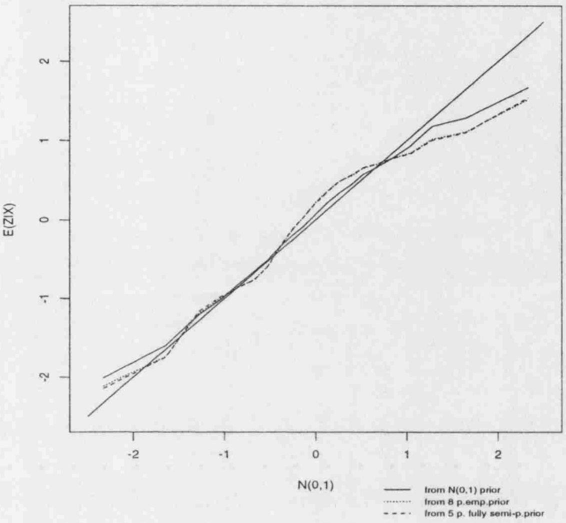


Figure 4.30: American scale 9, QQ plots of E(Z|X) from N(0,1), semiparametric and fully semiparametric estimation



The distribution of the posterior means obtained with a N(0,1) is closer to the N(0,1) than the distribution of the posterior means when the model is fitted semiparametrically, so the N(0,1) prior imposes some structure on the distribution of the posterior means.

4.8 Bootstrapping the estimated prior

We again use the nonparametric bootstrap, as in Section 4.3 to construct confidence intervals of the percentiles of the prior. We smoothed or rotated the estimated prior, calculated its percentiles and then formed 300 bootstrap samples. For each percentile of the estimated prior we constructed confidence intervals based on the percentiles of their bootstrap distribution.

We will first use the bootstrap confidence intervals to see how well the estimated prior approximates the distribution that generated the data.

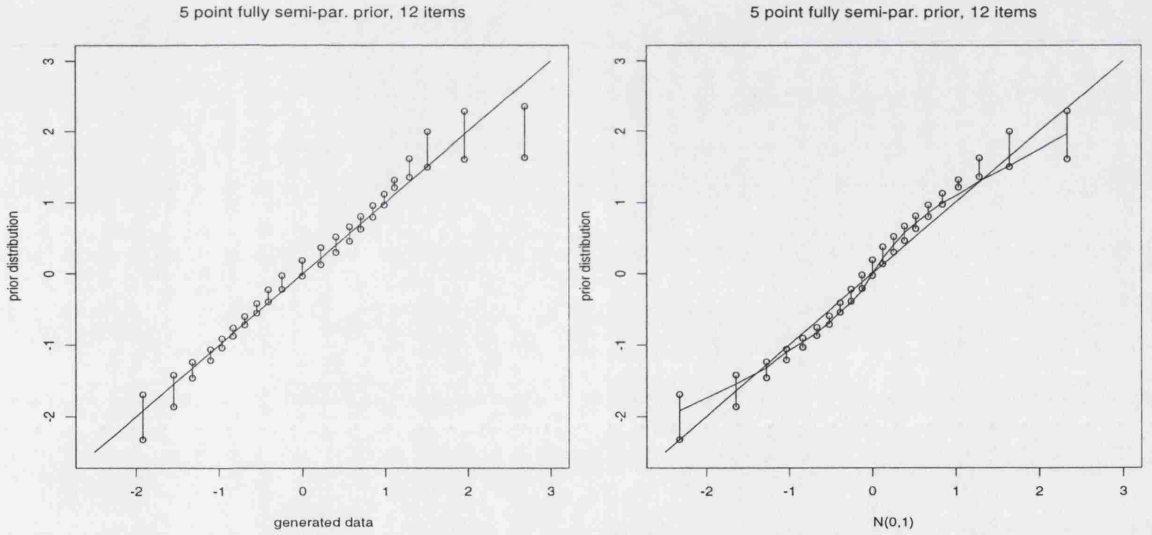
4.8.1 Artificial data

In this section we use the 0.5 N(0,1) + 0.5 N(3,1) mixture, which we will call 'n30'. The estimated prior came out with 5 support points. We smoothed the prior and constructed bootstrap confidence intervals of its percentiles.

Figure 4.31 shows QQ plots of the percentiles of the estimated prior and their bootstrap confidence intervals against the percentiles of the distribution used to generate the data (left) and against the N(0,1) (right).

In the left plot we see that most of the percentiles include the 45 degree line, or they are just off it (except the very last percentile which is further to the right). So the 5 point prior is a very good approximation of the distribution that generated the data. The same can be inferred from the QQ plot against the Normal. If we compare it with the simple empirical prior, with the fixed nodes, we can say that it does as well as the 8 point empirical prior in covering the original curve and even better at the extremes of the distribution, and also the confidence intervals are generally smaller. So the nonparametric prior is more economical and more efficient.

Figure 4.31: Artificial data: QQ plot of fully semiparametric prior against original data and against the N(0,1)



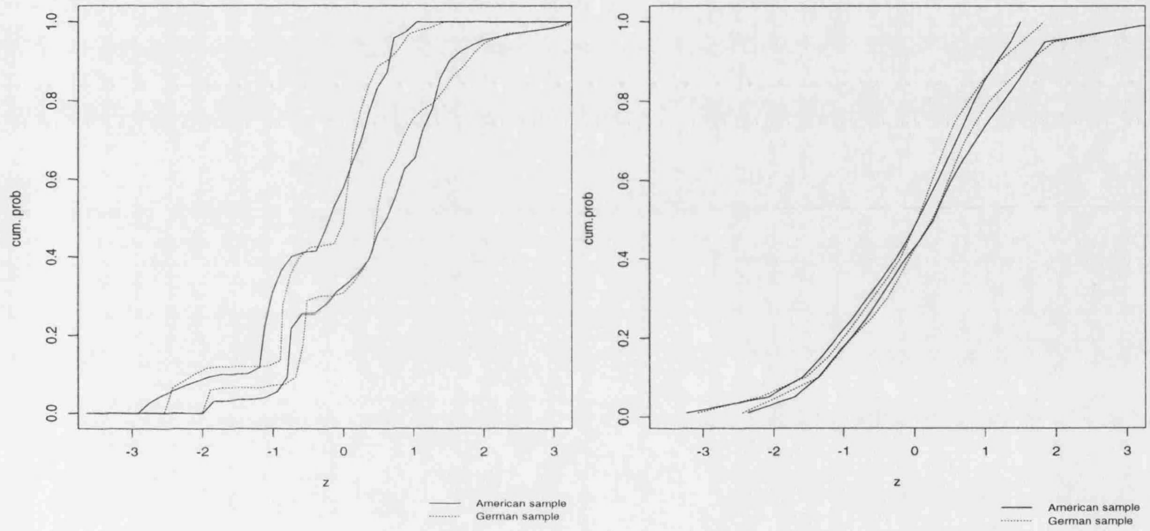
4.8.2 Comparison of Populations

In the following we will compare the priors and the posterior means of the American and German scale 9 obtained from the fully semiparametric model.

Figure 4.32 gives the bootstrap confidence bands of the 5 point fully semiparametric prior, considered either as a discrete distribution or an approximation to a continuous distribution, for the American and German scale 9. The pictures are very similar to those of Figure (4.22), Section 4.5, and again they show that the distribution of the American data is a bit heavier in the left half than the German one.

Figure 4.33 gives the ogives of the posterior means for American and German scale 9, obtained from the fully semiparametric prior. The picture is as expected from the ogives above and very similar with the ogive obtained from the 8 point simple semiparametric prior. So the fully semiparametric prior allows more information to come through the posterior means than a N(0,1) prior.

Figure 4.32: American and German scale 9, bootstrap bands of cdf of nonparametric prior and of smoothed nonparametric prior



4.9 Goodness-of-fit

It is interesting to see whether fitting a nonparametric prior improves the goodness-of-fit of the latent trait model. We expect that it will, since some discrepancies between the model and the data could be attributed to an inappropriate prior.

Tables 4.7 and 4.8 give the goodness-of-fit statistics for models with the N(0,1) prior, the 8 point simple semiparametric prior and the 5 point fully semiparametric prior.

For the American scale 9 data, G^2 and X^2 are similar for the simple and fully semiparametric model and both are smaller than the ones obtained with a N(0,1) prior.

The AIC suggests the 8 point simple semiparametric model is the best one but the

Figure 4.33: American and German scale 9, ogives of posterior means

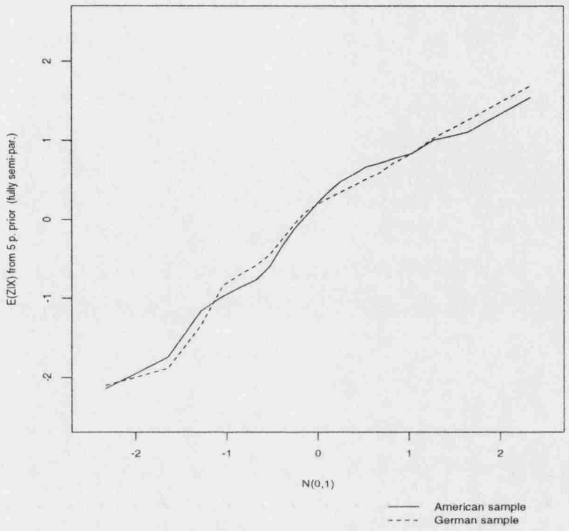


Table 4.7: American scale 9: Goodness-of-fit statistics

	N(0,1)	8pt semipar	5 pt fully semipar.
$\%G^2$	51.04	51.65	51.66
loglikelihood	-9876.26	-9862.45	-9862.18
G^2	1597.85	1543.82	1542.98
X^2	1316.40	1294.98	1295.65
d.f.	96	98	97
no.of param.	24	28	31
AIC	19800.52	19780.90	19786.36

likelihood is larger for the fully-semiparametric model.

For the Boys test 1 data, the fully semiparametric prior provides a larger $\%G^2$ and a larger likelihood, but the X^2 and G^2 indices are worse than the models from the N(0,1) or the 8 point semiparametric priors. The AIC suggests the 8 point simple semiparametric model is the best one (also compared with the other simple semiparametric models examined earlier).

Table 4.8: Boys, test 1: Goodness-of-fit statistics

	N(0,1)	8pt semipar.	6 pt fully semipar.
$\%G^2$	24.09	24.80	24.87
loglikelihood	-6126.83	-6102.51	-6100.05
G^2	3827.15	3681.93	3904.73
X^2	23159.47	19402.92	28629.67
no.of param.	42	46	51
AIC	12337.67	12297.02	12302.10

4.10 Comparison with latent class analysis

Fitting a latent trait model using an empirical prior is very similar to latent class analysis. In latent class analysis, the probabilities of belonging to a class are set out arbitrarily in the beginning and estimated and used again at each iteration of the EM algorithm. When fitting an empirical prior, the probabilities of being at a particular point are set out arbitrarily in the beginning and are estimated at each iteration. The difference is that in the latent trait model there is an underlying continuum and the probabilities are attached to some points on that continuum, whereas in latent class analysis the probabilities are not attached to any points. Because of the latter there isn't any ordering of the classes, whereas when we fit a latent trait model we expect to get the individuals ordered along the continuum representing the latent trait.

Croon (1990) proposed an ordered latent class model where the ordered relations between the classes are defined by imposing inequality restrictions on the item response probabilities (see also Section 1.9). He considered this model as a nonparametric version of the latent trait model, or a model that is between the latent class model, where the latent variable is measured on a nominal level and the latent trait model, where the latent variable is measured on an interval scale level.

The semiparametric latent trait model seems to be very close to the ordered latent class model. In the following we shall compare the results of fitting Boys test 1 data and the American scale 9 data with latent class models, ordered latent class models and fully-semiparametric latent trait models.

Comparison of class probabilities with prior distribution The class probabilities of the two latent class solutions are exactly the same as the probabilities on each of the two nodes of the prior distribution when two nodes are used. The nodes for the estimated prior can be deduced from their weights, once the mean and variance of the prior have been determined: the weights and nodes are subject to the restrictions

$$h(z_1)z_1 + h(z_2)z_2 = 0$$

and

$$h(z_1)z_1^2 + h(z_2)z_2^2 = 1.$$

From these we have

$$z_2^2 = 1/[h(z_2)^2/h(z_1) + h(z_2)]$$

and

$$z_1 = -h(z_2)z_2/h(z_1).$$

In Table 4.9 we see that for the Boys test 1 data the class probabilities of the 3 latent class solution are almost equal to the probabilities of the 3 point prior distribution (unrestricted solutions same as restricted one, no restrictions were necessary), and the 4 latent class probabilities, both from the unrestricted and restricted solution are again very similar to the probabilities of the 4 point prior distribution, any differences are less than 0.02.

In Table 4.10 the class probabilities of the 2, 3, 4 and 5 latent class and ordered latent class are shown, together with the weights of the nodes of the estimated priors, for the American scale 9 data. The weights of the estimated 3 point prior are very close to the latent class probabilities, and closer to the unrestricted than the restricted ones.

The weights of the 4 point prior are closer to the ordered latent class probabilities, except for the probability on the first node, which is about half the size of the first latent class probability.

For the 5 point prior /5 latent class solution we observe that the probabilities of the first three nodes are very close to the probabilities of the first three ordered latent classes, the last two nodes / classes though differ considerably.

Comparison of conditional response probabilities In the following we shall compare the item conditional response probabilities given a latent class with the conditional

Table 4.9: Boys, test 1: Estimated prior and latent class probabilities

	2 points / 2 classes			
nodes	-1.534	0.652		
weights	0.298	0.702		
l.c. probs	0.298	0.702		
	3 points / 3 classes			
nodes	-1.863	-0.462	0.939	
weights	0.144	0.382	0.474	
l.c. probs	0.109	0.388	0.503	
ord.l.c.probs	0.109	0.388	0.503	
	4 points / 4 classes			
nodes	-2.344	-1.212	-0.079	1.053
weights	0.064	0.167	0.406	0.364
l.c. probs	0.037	0.179	0.404	0.381
ord.l.c.probs	0.054	0.180	0.400	0.365

Table 4.10: American scale 9: Estimated prior and latent class probabilities

	2 points / 2 classes				
nodes	-1.219	0.820			
weights	0.402	0.598			
l.c. probs	0.402	0.598			
s.e	0.017	0.017			
	3 points / 3 classes				
nodes	-1.623	-0.289	0.982		
weights	0.206	0.350	0.444		
l.c. probs	0.195	0.345	0.460		
ord.l.c. probs	0.193	0.209	0.597		
	4 points / 4 classes				
nodes	-2.224	-0.848	0.348	1.271	
weights	0.074	0.306	0.394	0.226	
l.c.probs	0.210	0.197	0.182	0.411	
ord.l.c. probs	0.168	0.235	0.302	0.296	
	5 points / 5 classes				
nodes	-2.286	-0.936	0.042	0.842	2.184
weights	0.063	0.273	0.264	0.364	0.037
l.c.	0.310	0.153	0.171	0.132	0.233
ord.l.c. probs	0.065	0.267	0.230	0.229	0.208

response probabilities given a latent node of an estimated prior with the same number of nodes as the number of classes.

The item response probabilities in class 2 are all larger than the response probabilities in class 1. This is expected when there are just two classes and if it doesn't happen it can be achieved by recoding the items.

Haertel (1990) gave approximate relations between the latent trait model fitted with a 2 point fixed prior and the 2 latent class model, by setting $P(x_{hi}|z_j)$ under the latent trait model equal to $P(x_{hi} = 1|j)$ under the latent class model. He equated: $\Phi[a_i(z-b_i)]_{z=-1} = \pi_{1i}$ and $\Phi[a_i(z-b_i)]_{z=1} = \pi_{2i}$, where Φ is the standard normal cdf, π_{ji} are the conditional response probabilities given class j and -1 and +1 are the values for the quadrature nodes z_1 and z_2. From these equations he solved for a_i and b_i and noted that these parameters can only be determined up to a linear transformation. The indeterminacy in the 2-parameter latent trait model arises because the parameters are affected by changes in the location and scale of the prior. The latent class probabilities have no similar scale indeterminacy.

In our results we see that for the two latent class solution, the conditional probabilities of giving a positive response to item i given the latent class are the same as the conditional probabilities of giving a positive response to item i at each node of the two-point prior distribution.

Let use look at the marginal probabilities of a response pattern under the two models:

The probability of a response pattern $\mathbf{x}_l$ under the 2 point latent trait model is given by:

$$f(\mathbf{x}_l) = \sum_{j=1}^{2} \prod_{i=1}^{p} \pi_i(z_j)^{x_{li}} (1 - \pi_i(z_j))^{1-x_{li}} h(z_j) \tag{4.14}$$

For the case of 2 latent classes the probability of a response pattern $\mathbf{x}_l$ is given by

$$f(\mathbf{x}_l) = \sum_{j=1}^{2} \eta_j \prod_{i=1}^{p} \pi_{ij}^{x_i} (1 - \pi_{ij})^{(1-x_i)} \tag{4.15}$$

The two formulations so far are the same. So under maximum likelihood $h(z_j)$ should be equal to η_j and the conditional probabilities of the items should also be the same. The difference is that the latent trait model uses two item parameters to estimate the conditional probabilities on the two nodes. The latent class model directly estimates the two conditional probabilities for each item. The item parameters are adjusted to the scale and location of the nodes so that the conditional probabilities are the same as those coming from the latent class model (since those will give the same likelihood).

There is no apparent link between the latent trait and latent class conditional probabilities when there are more than two classes, but in the following we shall compare the conditional probabilities in the latent class and latent trait models when there are more than two classes for the test 1 and the American scale 9 data.

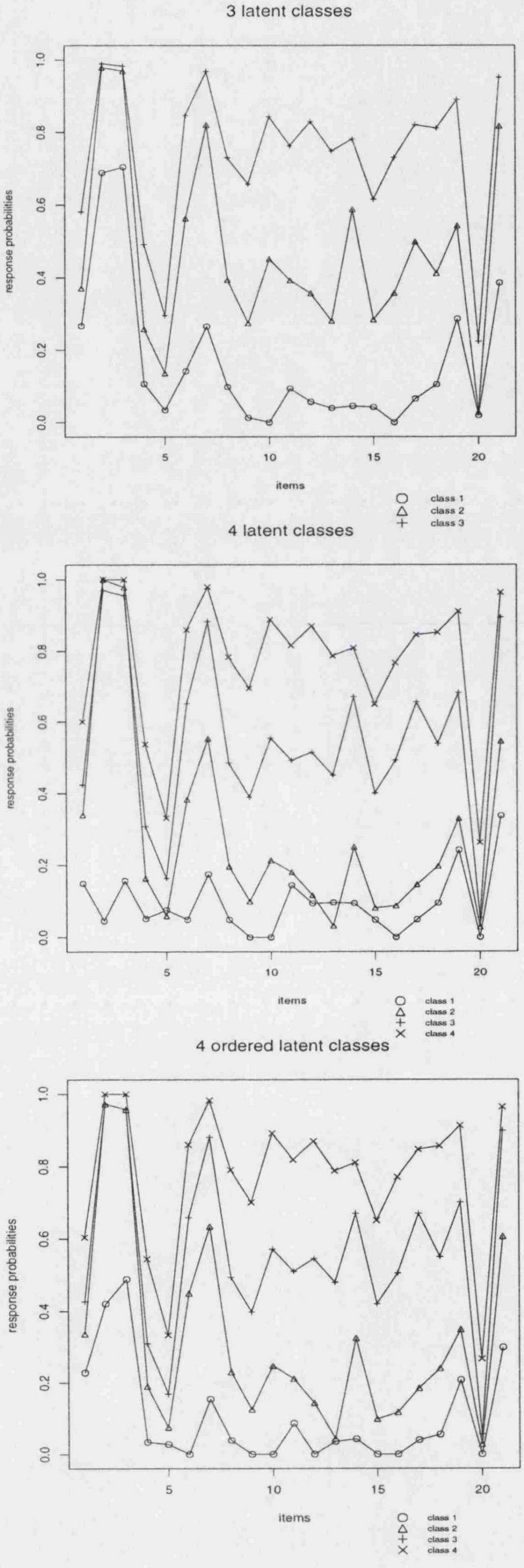
Boys, test 1 data Figure 4.34, shows the items' conditional response probabilities for the three and four unrestricted and restricted latent class model. We see that in the three class model (top of Figure) the response probabilities are all ordered, i.e. the response probability of positive response of an item given class j is greater than the response probability given class i, if $j > i$, for all items, indicating the ordering of the latent classes along a latent continuum.

In the four class model, the response probabilities are again ordered for most items, except for some small inversions. For items 1 and 2, where the response probability given class 2 is greater than the response probability given class 3, and for items 5 and 13 where the response probability given class 1 is greater than the response probability given class 2 (middle of Figure 4.34). These differences though are very small and could easily be attributed to estimation inaccuracy. The ordered latent class solution, same Figure, bottom, imposes such restrictions so that these inversions in the probabilities do not occur.

Figure 4.35 shows scatterplots of the conditional probabilities in the kth latent class against the conditional response probabilities on the kth estimated approximation point of the latent trait distribution for the three latent class model and the latent trait model estimated with a three-point prior, and Figure 4.36 shows the same scatterplots for the four latent class model and four point empirical prior.

We observe that for the three latent classes / three point prior, the conditional probabilities in classes 2 and 3 are not only linearly related but also almost numerically equal to the conditional probabilities at the second and third approximation points. The conditional probabilities in class one and on the first node are positively linearly related but

Figure 4.34: Boys, test 1: Conditional response probabilities of the items for the 3 and 4, unconstrained and constrained latent class model



there are some discrepancies from the 45 degree line.

Thus there seems to be not only an inherent ordering of the latent classes but also the latent classes seem to be placed at or around the estimated approximation points of the prior.

Similarly, for the four latent class / four point empirical prior model, the conditional probabilities in the first class are positively linearly related but generally different from the conditional probabilities on the first node, but the conditional probabilities in the other 3 classes are close, almost equal, to the conditional probabilities at the other nodes.

Although the ordering of the latent classes has been established for these data, we also observe a very close correspondence between the two sets of response probabilities. This shows that the spacing between the classes corresponds to the spacing between the estimated latent nodes of the latent trait distribution is such that it gives the same response probabilities as the latent class model. (The location and scale of the latent distribution have been constrained in the estimation of the latent trait model, but the parameters adjust for that so that they give the same response probabilities at relative points of the distribution, for example at the middle of the distribution). So in the latent trait model the model fits the response probabilities by imposing a structure through the items' difficulty, the items' discrimination and the spacing of the prior distribution. In the latent class model there is a parameter for each conditional probability in each class. There is an implicit indication of the item parameters in the latent class model as well though. As Croon (1990) noted, the overall level of the conditional response probabilities of an item as compared with the level of the response probabilities of another item, whatever the latent class, has to do with the difficulty of the item whereas the steepness of the conditional probabilities curve along the ordered latent classes has to do with the discrimination of an item.

When there are few classes / nodes the models use approximately the same number of parameters to model the response probabilities so the maximum likelihood solutions from both models should be very similar (and exactly the same when there are two nodes or classes). As the number of classes increases, the latent class model leaves much more freedom for the conditional probabilities to vary, whereas the conditional probabilities in the latent trait model are determined by the item parameters.

Figure 4.35: Boys, test 1: response probabilities of latent classes against response probabilities at each corresponding node of latent distribution, 3 nodes / classes

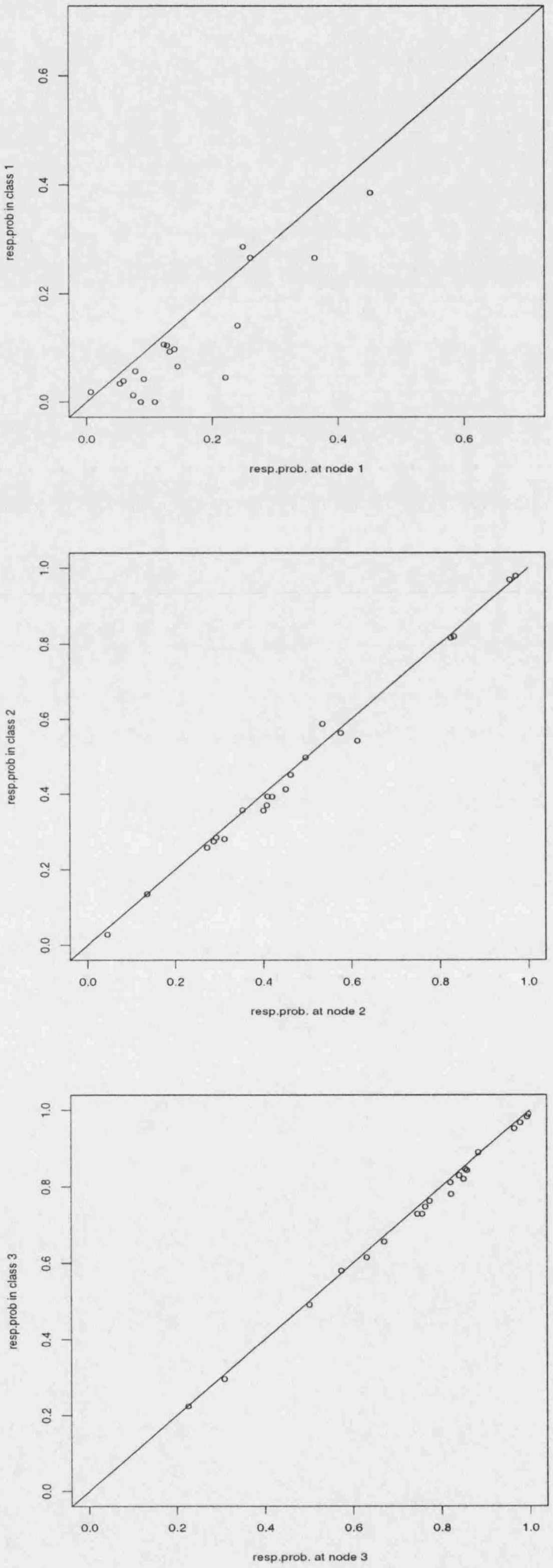
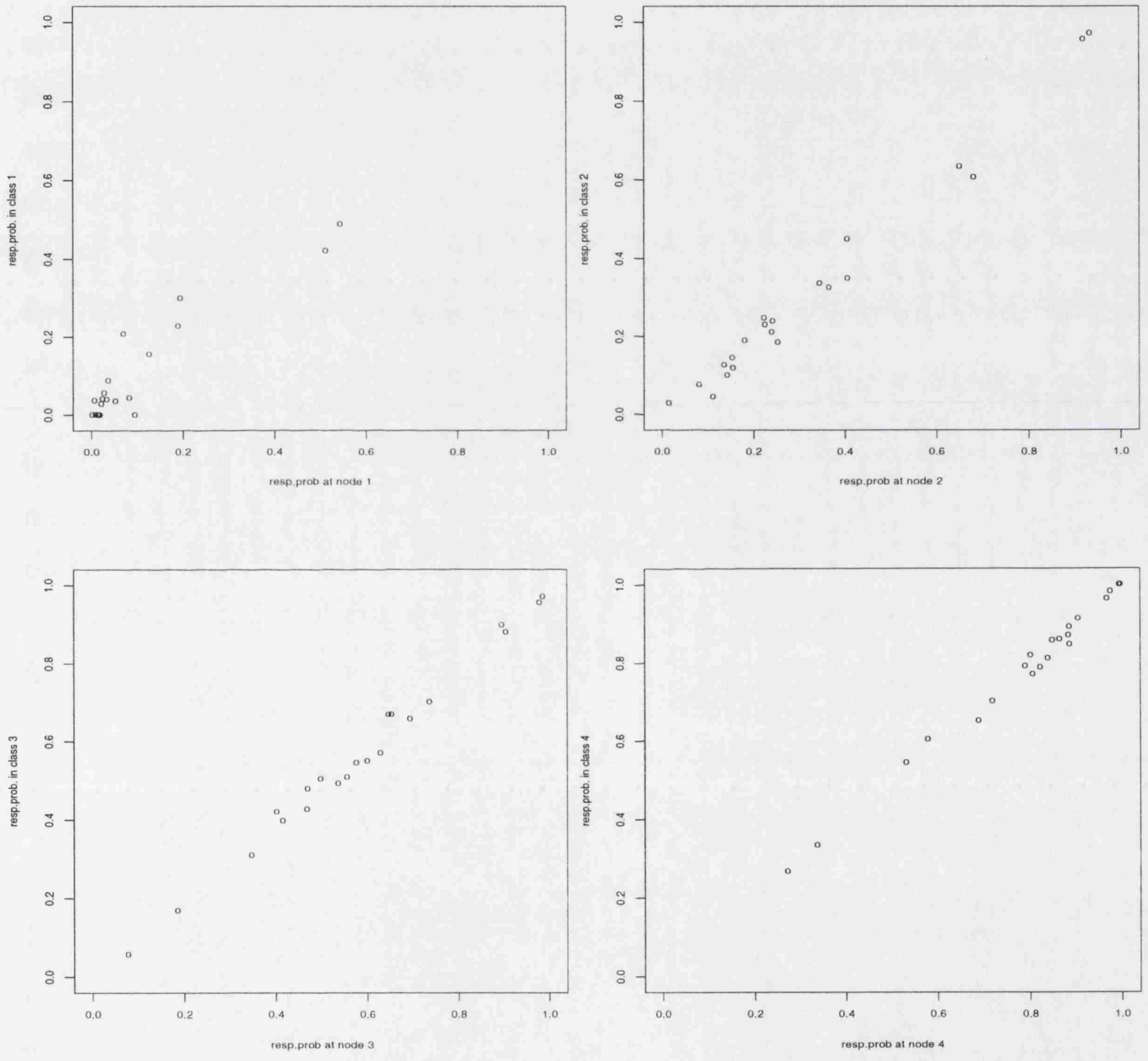


Figure 4.36: Boys, test 1: response probabilities in ordered latent classes against response probabilities at each corresponding node of latent distribution, 4 nodes / classes



American scale 9 Figures 4.37, 4.38 and 4.39 show scatterplots of the conditional response probabilities given a latent class against the corresponding node of the estimated prior, for the 3 latent class / 3 point prior, 4 latent class / 4 point prior and 5 latent class / 5 point prior respectively. Figure 4.40 presents the same information in a different way, it shows line plots of the conditional probabilities of the 5 point latent trait and 5 latent class model.

All scatterplots show linear relationships between the two corresponding sets of response probabilities, with a small dispersion around the 45 degree line. The conditional probabilities in the middle classes, for example classes 2 and 3 in the 4 and 5 class models are closer to the 45 degree line than the conditional probabilities of class 1 in the 4 class model and class 5 in the 5 class model. In the latter cases the response probabilities given the latent class are more dispersed than the response probabilities given the latent node, which is expected since the latent class models give more freedom for the response probabilities to vary.

We note that for the Boys, test 1 data the 3 (and the 4) latent class solutions resulted in unidentified parameters (LEM (Vermunt 1997), which was used to fit the latent class models, gave a warning message). For the American scale 9 data, fitting an ordered latent class model with 4 ordered latent classes resulted in unidentified parameters.

Therefore, a latent class model with the optimal number of nodes fitted suggested from the gradient D (Section 4.7.3) for either datasets cannot be fitted with a latent class model.

4.11 Conclusions

In this chapter we explored semiparametric and fully semiparametric estimation of the latent trait model. We gave an algorithm for the fully semiparametric estimation of the latent trait model so that both the points and the weights of the prior can be estimated simultaneously with the item parameters, and investigated the amount of information about the prior that can be obtained from a sample of binary responses to a set of items. Regarding the simple semiparametric estimation, we showed that when there are enough (six to eight) points, the shape of the distribution that generated the data can be retrieved successfully. When the points are not held fixed but estimated from the data fewer points

Figure 4.37: American scale 9: 3 latent classes /3 point estimated prior: response probabilities of ordered latent classes against response probabilities at each corresponding node of latent distribution

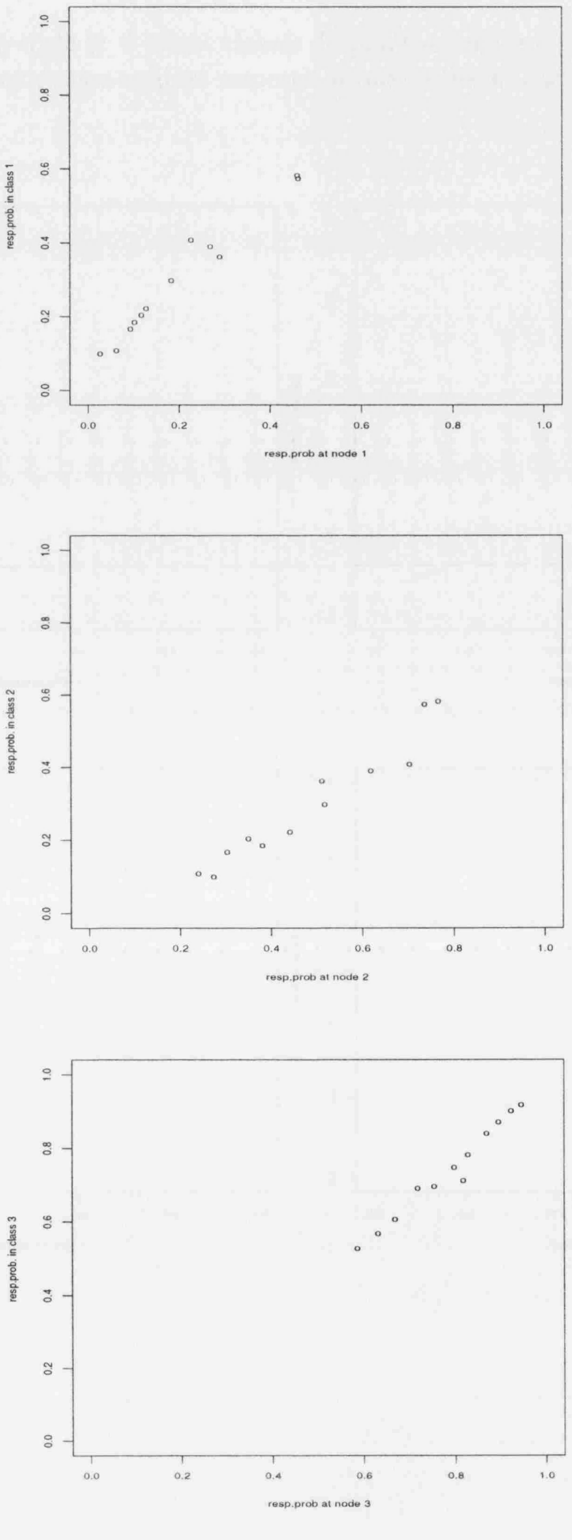


Figure 4.38: American scale 9: 4 latent classes /4 point estimated prior: response probabilities of ordered latent classes against response probabilities at each corresponding node of latent distribution

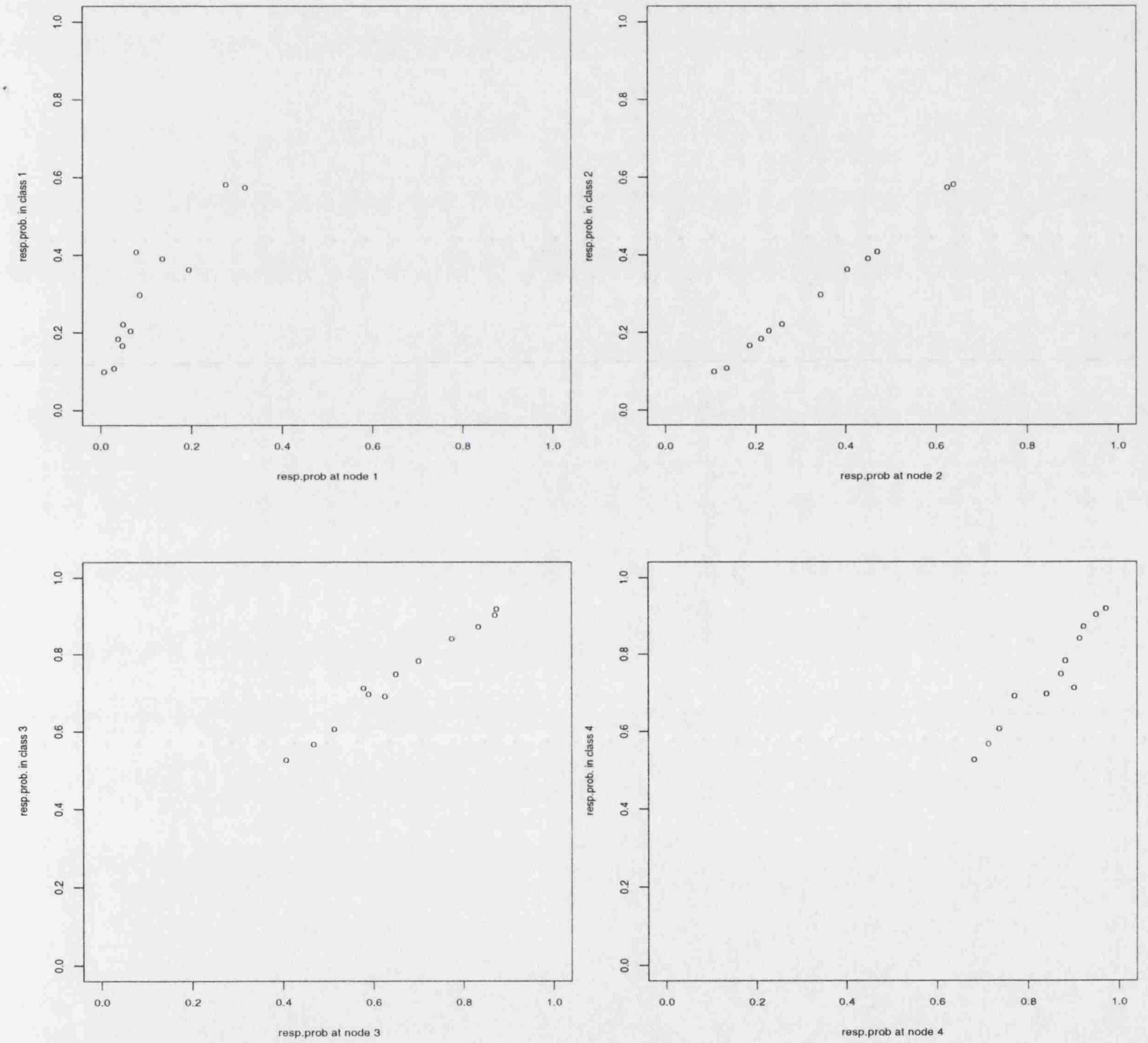


Figure 4.39: American scale 9: 5 latent classes /5 point estimated prior: response probabilities of ordered latent classes against response probabilities at each corresponding node of latent distribution

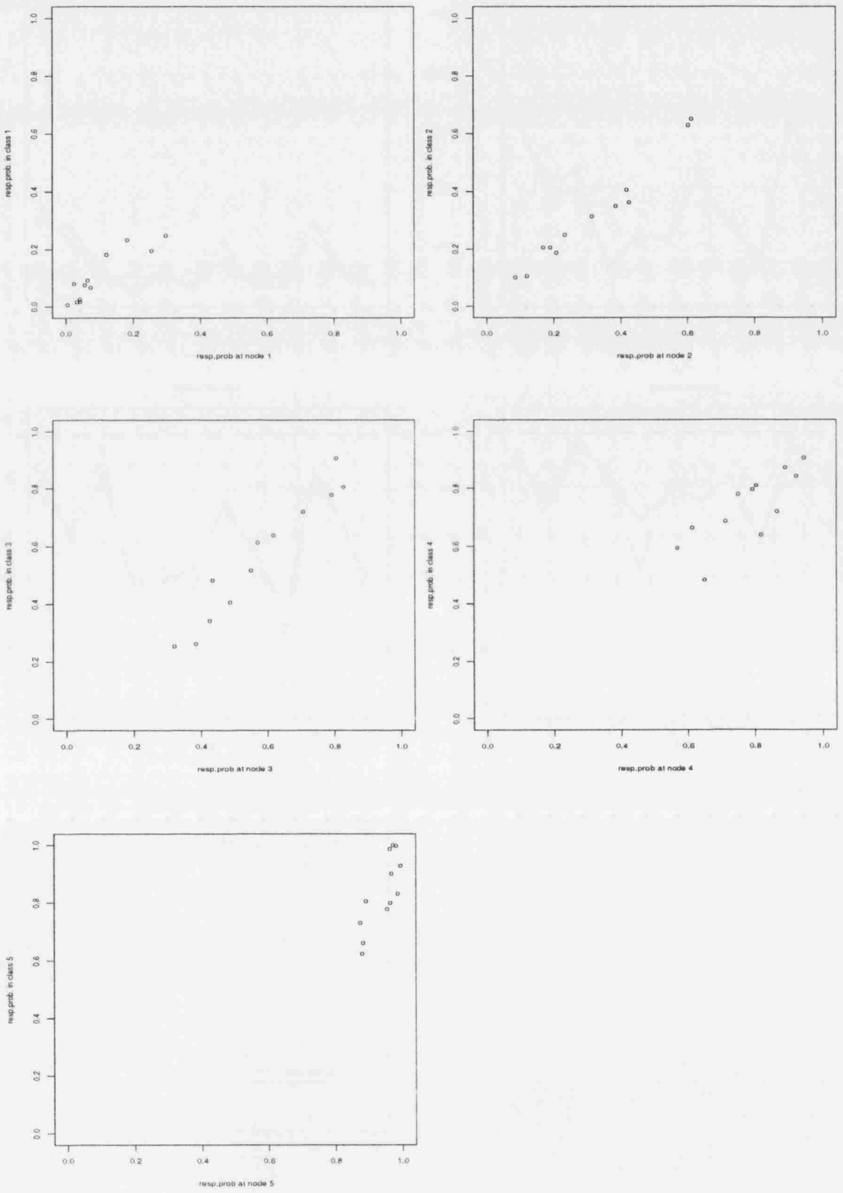
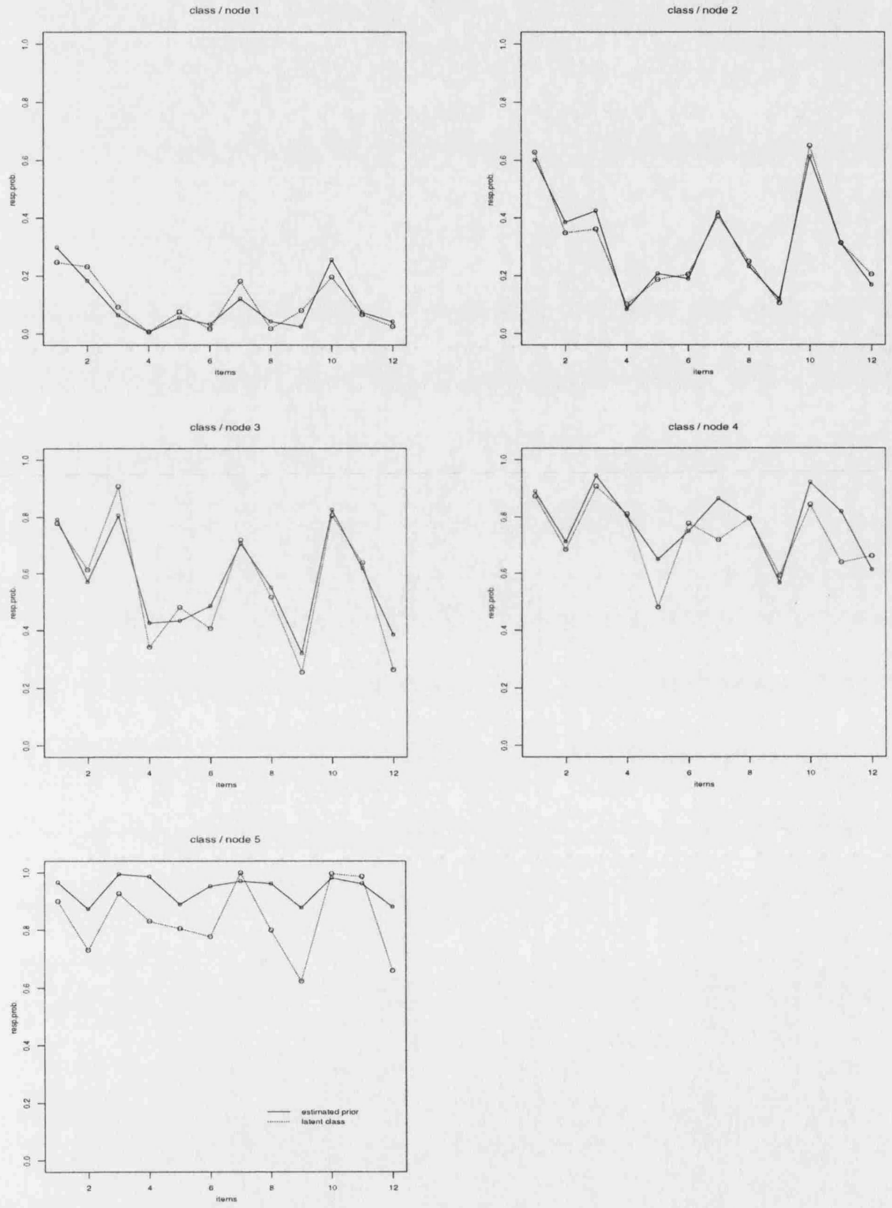


Figure 4.40: American scale 9: 5 latent classes /5 point estimated prior: response probabilities of ordered latent classes and response probabilities



may be sufficient.

We showed that it is feasible to estimate the latent trait model without any assumptions about the prior, apart from fixing its location and scale, and all this in the standard framework of estimating latent trait models. The optimality of the solution can be checked by formal criteria.

This makes the 2-parameter latent trait model far more attractive than the Rasch model, because it has the advantage over the Rasch model that no assumptions are needed about the ability distribution, and yet it is more flexible, since it allows a discrimination parameter.

Apart from allowing us to get rid of a restricting assumption, this estimation procedure offers information about the latent distribution. We investigated empirically, by means of the bootstrap and simulated data, the variability of the prior and how well the prior can recover the distribution that generated the data. We saw that the prior distribution may be different from the normal, which is the usual assumption about the prior. And although the parameters are not affected very much from the shape of the distribution, a parametric prior imposes its shape on the posterior means. The posterior means are therefore more informative when a nonparametric prior is fitted.

A nonparametric prior is also useful when one wants to compare different populations that have answered the same set of questions, since the shape of the latent distribution can also be compared, in addition to the location and scale comparisons that can be made from the parameters.

The fully semiparametric latent trait model is very similar to an ordered latent class model. It has the advantages over the latter that is more parsimonious and optimality criteria can be applied to determine the number of classes /nodes needed.

Chapter 5

The Goodness-of-Fit of Latent Trait Models

5.1 Introduction

One of the main problems with the use of latent trait models in attitude and educational measurement has been judging the goodness-of-fit of the chosen model. This arises because of the sparseness of the multi-way contingency tables to which the response patterns give rise. Many of the expected frequencies are typically very small and this invalidates the traditional tests of fit. The main purpose of this chapter is to investigate the fit of the logit/probit latent trait model to the data sets used by Krebs and Schuessler (1987). We shall propose the use Monte-Carlo simulations to approximate the empirical distributions of the goodness-of-fit statistics used. Apart from the global tests, we will further use diagnostic procedures, based on residuals, which give greater insight into the reasons for a poor fit and so suggest ways in which the scales may be improved. Some of this material has been presented in Bartholomew and Tzamourani (1999).

5.2 Goodness-of-Fit and Derived Statistics

Goodness-of-fit statistics The most common used statistics to measure the fit of a latent trait model are the chi-squared statistic X^2 and the loglikelihood ratio statistic G^2 (Chapter 1, Section 1.10). We saw in that Section that as the number of items, p, gets large the contingency table for these items becomes very sparse and thus the chi-squared distribution is no longer a valid approximation to the sampling distribution of G^2 and X^2. To avoid the problem of cells having very small expected frequency, pooling of response patterns can be used. Pooling may result in a loss of power of the statistics though, therefore we will calculate G^2 and X^2, both with pooling together response patterns with expected frequency less than 5 and without pooling.

In addition to the pooled and unpooled versions of G^2 and X^2 we will use three other statistics derived from them. Since the pooling of categories in TWOMISS is automatic, the number of degrees of freedom will vary from one sample to another. A way of standardising the results is to divide the pooled G^2 (or X^2) by its degrees of freedom, f say to give the 'amount of G^2 (or X^2) per degree of freedom'. The third statistic is the $\%G^2$ defined in (1.68). To calculate $\%G^2$ we use the unpooled G^2.

As a poorly fitting model will give a smaller percentage of G^2 explained than a well fitting one, if we use (1.68) as a test of goodness-of-fit we must look in its empirical distribution for values significantly smaller than would be expected if the model is correct.

5.3 The Monte Carlo Test

We shall carry out the test as follows. For any data set we first estimate the parameters of the logit/probit model. Next we generate N independent samples of x's treating the estimated α's as the true values. This is an example of the parametric bootstrap, since we generate data from the model parameters. For each such sample we compute G^2, X^2 and the derived statistics described above. We then judge the significance of the observed value by reference to the empirical sampling distribution. The critical values of G^2, X^2, G^2/f, X^2/f are the 95% values of their empirical sampling distribution, whereas the critical value of $\%G^2$ is the 5% value of its empirical sampling distribution.

We investigate the empirical sampling distributions of these statistics for the intercultural social life feelings' scales described in Chapter 1, Section 1.11, for the American and German samples. There are 9 scales with the numbers of items varying between 5 and 12. We have retained the original numbering of the scales to facilitate comparisons with earlier work but the reader should note that there is no Scale 7 in our analysis. The questions for each scale are listed in the Appendix.

Table 5.1: Goodness-of-fit statistics for the intercultural scales

	Pooled statistics						Unpooled statistics		
Scale	p	G^2	X^2	f	G^2/f	X^2/f	G^2	X^2	$\%G^2$
	American sample								
1	12	2151.7	2089.8	85	25.3	24.6	2613.7	4193.4	45.6
2	6	117.4	120.5	41	2.9	2.9	123.9	130.1	92.2
3	9	373.3	265.4	82	4.6	3.2	520.3	521.7	79.4
4	8	213.3	157.2	67	3.2	2.3	272.6	237.4	58.9
5	7	132.1	111.7	64	2.1	1.7	166.0	153.3	85.2
6	8	167.0	116.1	54	3.1	2.2	252.1	257.1	76.8
8	8	223.9	188.3	105	2.1	1.8	307.4	296.7	73.4
9	12	1602.2	1332.1	97	16.5	13.7	2220.9	4274.2	51.0
10	5	30.4	28.9	14	2.2	2.1	47.1	46.4	90.3
	German sample								
1	12	2067.6	1827.4	105	19.7	17.4	2661.1	4357.2	37.3
2	6	71.8	72.9	37	1.9	2.0	93.1	96.1	92.6
3	9	507.7	384.1	76	6.7	5.1	645.1	663.0	73.3
4	8	152.2	113.0	30	5.1	3.8	215.3	228.2	75.1
5	7	132.9	104.0	45	3.0	2.3	171.5	152.6	81.8
6	8	221.7	159.3	52	4.3	3.1	299.5	279.5	67.7
8	8	188.7	149.7	102	1.9	1.5	272.7	277.7	74.2
9	12	1367.8	1065.1	108	12.7	9.9	2063.5	4562.8	50.4
10	5	35.0	32.9	16	2.2	2.1	39.1	38.9	92.3

5.4 Results: Global Tests

5.4.1 Goodness-of-fit statistics for the inter-cultural scales

The results of the various tests based on G^2 and X^2 for the American and German scales are given in Table 5.1.

Using the chi-squared approximation the fit as judged by G^2 and X^2 (the pooled versions) is poor for all scales in both countries. All reach the 1% significance level and most greatly exceed it. We note that the degree of significance appears to increase with p. Also, the amount of G^2 explained declines markedly as p increases.

Let us see how the degree of sparseness varies between these scales, as measured by the $n/2^p$, i.e. the sample size over the number of possible response patterns. In Table 5.2 we give the number of possible response patterns 2^p, the ratio of the sample size n over 2^p and the actual number of distinct response patterns observed, NR, for each of the American and German scales. We see that even with 8 items, the average number

Table 5.2: Degree of sparseness of the intercultural scales

			American sample		German sample	
Scale	p	2^p	$n/2^p$	NR	$n/2^p$	NR
1	12	4096	0.3	822	0.4	905
2	6	64	22.1	61	23.3	64
3	9	512	2.8	304	2.9	296
4	8	256	5.5	185	2.7*	132
5	7	128	11.1	121	11.6	104
6	8	256	5.5	170	5.8	170
8	8	256	5.5	215	5.8	199
9	12	4096	0.3	729	0.4	691
10	5	32	44.3	31	46.6	32

* n=688 for German scale 4

per cell, if subjects were distributed uniformly over the cells, is 5.5. Since this is not the case, there will be a lot of cells with observed - and expected - frequency smaller than 5. The extent of pooling undertaken can be measured by comparing the number of distinct response patterns NR with the degrees of freedom in each scale, given in Table 5.1.

5.4.2 Monte-Carlo test results for the inter-cultural scales

The Monte Carlo test was carried out by computing each statistic listed in Table 5.1 on 1000 samples. The significance was then judged by reference to the empirical sampling distribution and Table 5.3 gives the estimated P-values.

This table presents a very different picture to that in Table 5.1. Virtually all of the tests yield P-values towards the upper tail but only in the case of scales 2, 5 and 10 is the evidence for rejection of the model unequivocal for both countries. In addition Scale 3 fits poorly in the German but not the American case. For Scale 8 all indices are below the 5% significance level for the American sample, except for G^2/f and $\%G^2$ which are just above 5%. Other more marginal cases are provided by Scale 4 (American) and Scale 6 (German). And for the Scale 9, the German sample, only the unpooled G^2 is significant. Here, especially, we might hope the residuals to throw some further light on the matter.

It is worth noting that likelihoods for this model are often fairly flat in the neighbourhood of the maximum so there are other sets of parameter values for the model which could be seriously entertained in each case and it is possible that some of these might give better fits. (G^2 is, of course, based on the likelihood and so should indicate the best

Table 5.3: P-values of Monte Carlo tests of fit for the intercultural scales

	Pooled statistics				Unpooled statistics		
Scale	G^2	X^2	G^2/f	X^2/f	G^2	X^2	$\%G^2$
	American sample						
1	0.204	0.350	0.218	0.319	0.140	0.205	0.399
2	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.496	0.527	0.418	0.449	0.298	0.226	0.300
4	0.056	0.094	0.030	0.039	0.201	0.488	0.344
5	0.005	0.005	0.008	0.012	0.004	0.005	0.011
6	0.634	0.733	0.440	0.583	0.382	0.219	0.398
8	0.018	0.006	0.053	0.018	0.015	0.011	0.063
9	0.281	0.349	0.268	0.327	0.258	0.200	0.459
10	0.030	0.019	0.036	0.019	0.000	0.001	0.003
	German sample						
1	0.271	0.209	0.293	0.230	0.215	0.041	0.396
2	0.002	0.001	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	0.001	0.000	0.000	0.000	0.001
4	0.216	0.140	0.222	0.147	0.314	0.579	0.355
5	0.000	0.001	0.000	0.001	0.000	0.004	0.006
6	0.041	0.068	0.065	0.082	0.019	0.062	0.076
8	0.283	0.295	0.432	0.448	0.122	0.061	0.185
9	0.106	0.152	0.141	0.152	0.019	0.082	0.170
10	0.009	0.007	0.025	0.017	0.014	0.011	0.015

fit possible but this is not necessarily so for the pooled version.) It is evident from the tables that there is some association between the P-values and the number of items. In particular, Scales 2, 5 and 10 have the fewest items and the poorest fits in both countries. There is, of course, no reason why this should not happen by chance because these are real data and we have no independent evidence on whether the scales are truly unidimensional. Nevertheless, the scales were all constructed on the same principles and we have already given reasons for expecting power to decline as p increases. We have, therefore carried out Monte Carlo experiments with artificial data to see whether this is likely to be the case. Before presenting those results we will first present some experiments that investigated the size of the test.

5.4.3 Measuring the size of the Monte-Carlo test

The experiment we describe below aims to measure the size of the Monte-Carlo tests of the various goodness-of-fit indices - we find how many times the Monte-Carlo tests reject the model when the data are generated from the model.

We used three sets of parameters, which were the estimated parameters for scales 1, 5, and 10, all estimated from the American sample. The number of items for each set was 12, 7 and 5 respectively.

From each set of parameters we generated 100 samples. To each sample we fitted the 1-factor latent trait model and recorded the parameters and the goodness-of-fit statistics. We then conducted the Monte-Carlo test for these statistics, for each of the 100 samples, as described in Section 5.3 but using only 200 simulations, and recorded their P-value.

Table 5.4 shows the frequencies of the P-values of the goodness-of-fit statistics (only two categories, values between 0.0 and 0.05 and values between 0.05 and 0.10), for each set of initial parameters.

Since the Monte-Carlo test is conducted for data generated from the model, we would expect the distribution of P-values to be uniform between 0 and 1 and the intervals 0.0 to 0.5 or 0.5 to 0.10 to each include approximately 5 values.

For the 5 item sets, we see that G^2, X^2, ungrouped G^2 and ungrouped X^2 have either 4 or 5 values between 0.0 and 0.05. So based on these statistics the Monte-Carlo test rejects the model 4 or 5 times out of a hundred, if we test at 5%. If we test at 10% the Monte-Carlo test of G^2, X^2, ungrouped G^2 and ungrouped X^2 rejects the model 7, 8, 9,

Table 5.4: Frequencies of P-values of the Monte-Carlo tests, conducted for 100 sets of artificial data generated from the 1-factor model

statistics	P-values between	12 items	7 items	5 items
G^2	0.0-0.05	1	3	4
	0.05-0.10	1	4	3
X^2	0.0-0.05	1	4	4
	0.05-0.10	1	4	4
G^2/f	0.0-0.05	2	2	2
	0.05-0.10	4	5	5
X^2/f	0.0-0.05	1	6	2
	0.05-0.10	2	1	5
G^2 ungr.	0.0-0.05	1	4	5
	0.05-0.10	4	1	4
X^2 ungr.	0.0-0.05	3	4	4
	0.05-0.10	4	5	7
$\%G^2$	0.0-0.05	0	2	1
	0.05-0.10	0	2	7

and 11 times out of the 100 respectively. G^2/f, X^2/f, and $\%G^2$ would reject the model fewer times, only 2, 2 and 1 times respectively if we tested at the 5% level and 7, 7 and 8 respectively if we tested at the 10% level.

With the 7 item sets we get approximately the same results. The most different figure is the frequency of $\%G^2$ between 0.05 and 0.10 which is only 2.

With the 12 item sets though we get quite different results. The frequencies in all cells are smaller, which means that the Monte-Carlo test for most indices would reject the model fewer times than expected from the nominal significance level. The best behaved index is the ungrouped X^2 with 3 P-values between 0.0 and 0.05 and 4 P-values between 0.05 and 0.10.

5.4.4 Measuring the power of the Monte-Carlo test

First of all we created artificial data sets with different numbers of items generated by a logit/probit model with two factors. The probability of positive response for item i for the two factor model is given by

$$\pi_i(z_1, z_2) = \frac{\exp(a_{0i} + a_{1i}z_1 + a_{2i}z_2)}{1 + \exp(a_{0i} + a_{1i}z_1 + a_{2i}z_2)} \tag{5.1}$$

Table 5.5: Model parameter values used to generate artificial data

i	a_{0i}	a_{1i}	a_{2i}
1	0.30	1.50	1.00
2	0.20	1.50	1.00
3	0.10	1.50	1.00
4	0.10	1.50	1.00
5	-0.10	1.50	1.00
6	-0.10	1.50	1.00
7	-0.20	1.50	1.00
8	-0.30	1.50	1.00
9	0.30	1.50	-1.00
10	0.20	1.50	-1.00
11	0.10	1.50	-1.00
12	0.10	1.50	-1.00
13	-0.10	1.50	-1.00
14	-0.10	1.50	-1.00
15	-0.20	1.50	-1.00
16	-0.30	1.50	-1.00

where z_1 and z_2 are the two latent factors and a_{1i} and a_{2i} their loadings for item i. We then carried out the Monte Carlo test to see whether the departure from the one-factor model would be detected. The number of items varied between 10 and 20. In each case all values of α_{i1} were set equal to the fairly typical value of 1.5. For the second factor, half of the items were given a loading of +1 and the remainder a loading of -1. The α_{i0}'s were given values varying between +0.3 and -0.3. For example, the parameter values used to generate the 16-item set are given in Table 5.5.

The Monte Carlo goodness of fit tests for a one-factor model were then applied to each artificial data set using 1000 replications. The estimated P-values are given in Table 5.6.

It is clear that there is a tendency for all tests to yield more highly significant results with the lower values of p than with the higher but the effect is hardly detectable, at this level of accuracy, until p exceeds 14 or 15.

We gave reasons in Chapter 1, Section 1.10 why the pooled versions of G^2 and X^2 lose power because of the pooling of the response patterns. Why should the unpooled G^2 though lose power as well whereas the unpooled X^2 seems not to be affected? As p gets very large, there will be more response patterns with frequency equal to 1. The expected

Table 5.6: P-values of Monte Carlo tests of fit for the artificial data

	Pooled statistics				Unpooled statistics		
p	G^2	X^2	G^2/f	X^2/f	G^2	X^2	$\%G^2$
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000
12	0.000	0.000	0.000	0.000	0.000	0.000	0.000
14	0.000	0.000	0.000	0.000	0.000	0.000	0.023
16	0.000	0.000	0.043	0.014	0.000	0.000	0.085
18	0.432	0.377	*	*	0.046	0.000	0.296
20	0.919	0.883	*	*	0.233	0.000	0.425

* These statistics could not be computed because the degrees of freedom were negative.

frequency of these response patterns is very small, and the individual contributions to G^2, i.e. $\ln(1/E_i)$ are smaller than the individual contributions to X^2, $(1-E_i)/E_i$, if $E_i < 0.5$. For values closer to 1, say between 0.6 and 1, which are perhaps more probable for best fitting models, then the contributions to G^2, $\ln 1/E_i$, are larger than the contributions to X^2, $(1-E_i)/E_i$. Therefore, if p is large, the best fitting models will also yield high G^2's and the G^2 obtained from a non-fitting model will not be so much larger, as they would be for X^2. In addition, as p gets large, fewer cells out of the 2^p cells are contributing to G^2. For the calculation of X^2 though, the frequencies of the unobserved response patterns are taken into account: as we saw in Section 1.10 the contribution of the unobserved response patterns to X^2 is equal to $\sum(0-E_k)^2/E_k = \sum E_k = \sum O_i - \sum E_i$, where $k = 1, ..., 2^p - \text{NR}$ and $i = 1, ..., \text{NR}$.

For two of the above sets of parameters, for the 8 and 12 parameter set we conducted a more thorough experiment on the lines of Section 5.4.3.

We generated 100 samples from the parameters in Table 5.5, for the 8 and 12 item sets, which indicate a 2-factor model. We fitted the 1-factor model to each of the samples, and recorded the goodness-of-fit statistics.

We then conducted the Monte-Carlo test for these statistics, for each of the 100 samples, as described in Section 5.4.3 and recorded its P-value.

We would like to see how many times the Monte-Carlo tests correctly reject the model.

For the 8 item set, all 100 P-values, for all of the goodness-of-fit statistics considered, are 0.00. This means that the Monte-Carlo tests for all indices always - correctly - rejected the model.

For the 12 item set, all 100 P-values for G^2, X^2, G^2/f, X^2/f, ungrouped G^2 and ungrouped X^2 were 0. As for the $\%G^2$, only three P-values were equal to or larger than 0.

For these artificial data the Monte-Carlo tests proved to be very powerful but the data have been generated from parameters that are very high for the second factor. Even for data generated from such parameters the behaviour of the Monte-Carlo tests deteriorates with increasing number of items as the results of Table 5.6 indicated above, but the enormous amount of time and resources to conduct the more thorough experiments prohibited us of conducting them for sets with larger number of items.

For real datasets, loss of power may diminish with fewer items than the number in the artificial datasets, as the departure from the unidimensionality in these datasets is very strong and will not be typical of real datasets. This fact makes the need for some kind of supplementary test the more necessary.

5.5 Results: Residuals

If the global test indicates a significant departure, we can examine the individual contributions to the total value of the statistic to see whether the significance arises from particular response patterns. The X^2 test lends itself to this kind of decomposition and the individual terms, $(O_i - E_i)^2/E_i$, can be used as residuals whose values should be in the neighbourhood of 1 if the fit is adequate. Unless p is very small, this interpretation breaks down because of the fact that O_i must be an integer and E_i may be very small indeed. This means that the corresponding residual is either very small or very large. We find that it is more informative to look at residuals constructed from pairs or triplets of responses. TWOMISS enables us to compare the observed and expected frequencies of observing all possible responses to pairs of questions and prints out the values of $(O - E)^2/E$ in each case. The same information can be obtained for all triplets. The pair-wise comparisons are of particular interest. If these residuals are small it means that the associations between all pairs of responses are well predicted by the model. A model which was successful in this respect might be judged adequate even if it was less good at accounting for the higher order associations.

There is no precise theoretical guidance on how to judge the significance of individual residuals. They are not independent and the chance of large values occurring by chance presumably increases with p. If we treat them as individually distributed like the square of a standard normal variable, values of the order of 4 or more would be suggestive of a real departure from the model. In practice, if several values of this order occur and if they involve a common variable, or pair of variables, it is worth examining the questions involved for a possible explanation.

The simulations reported above about the power of the global test also provide a pointer to what residuals we might expect to find for departures from unidimensionality of the kind considered there. The larger second order residuals were typically between 4 and 10: third order margins tended to be much larger.

In the following we will look at the two- and three-way residuals for each scale and try to identify in some cases any systematic effects that cause them.

Scale 1 In the case of Scale 1 there are two residuals greater than 4 in the American sample and five in the German sample. In particular, for the American sample, the two 'large' residuals are: for response 0,1 to items 5 and 4 the residual is 4.0 and for response 1,1,1 to items 10,11 and 12 the residual is 4.1. For the German sample, the large three-way residual is for response 1,1,1 to items 4,5,6 and it is equal to 4.96.

The two-way residuals with supporting data are set out as 2 × 2 tables in Table 5.7. When we set out the residuals in this form it is easy to see the nature of the deviations from the model's predictions. By referring to the questions at the same time (given in the Appendix) we could perhaps explain their occurrence.

For items 5 and 6 there is more disagreement between the items, i.e. more responses (0,1) and (1,0) than the model would predict, and less agreement than the model would predict. Responses (0,1) give rise to a large residual. Looking at the items, we see that both of them relate to chance influencing life, but item 5 has a positive attitude whereas in item 6 this is expressed with some resignation. The opposite directions of these items could perhaps explain the excess disagreement between the two.

For the pairs of items (4, 11) and (10, 12) there is less disagreement between the responses to these items and more agreement than the model would predict.

We see that items 10 and 12 are very similar, both referring to having 'control' or 'influence' over ones's life, and thus indicating that there is more of a direct link between these items than the underlying attitude.

Table 5.7: Scale 1, residuals for the pairs of questions (5,6), (4,11) and (10,12), German sample

	Pair	(5,6)		(4,11)		(10, 12)	
	Response	1	0	1	0	1	0
O		357	461	294	640	354	231
E	1	323.6	494.2	312.8	621.1	319.3	265.0
$(O-E)^2/E$		3.4	2.2	1.1	0.6	3.8	4.4
O		128	544	102	454	253	652
E	0	160.9	511.3	82.6	473.5	287.0	618.7
$(O-E)^2/E$		6.7	2.1	4.5	0.8	4.0	1.8

The rows relate to the first question in each pair and the columns to the second.

Table 5.8: Scale 2 residuals for the pairs of questions (1, 2), (1, 4) and (4, 5) (American sample)

	Pair	(1, 2)		(1, 4)		(4, 5)	
	Response	1	0	1	0	1	0
O		635	267	417	485	381	118
E	1	614.6	287.9	434.7	467.9	357.5	142.3
$(O-E)^2/E$		0.7	1.5	0.7	0.6	1.5	4.2
O		94	420	82	432	298	619
E	0	115.5	397.9	65.2	448.2	322.2	593.9
$(O-E)^2/E$		4.0	1.2	4.3	0.6	1.8	1.1

The rows relate to the first question in each pair and the columns to the second.

Items 4 and 11 look quite different and thus the excess disagreement between the two is difficult to explain. Of course, we note again that the single residual greater than 4 in the 2 × 2 table is not very large.

Scale 2 All versions of the global test give a highly significant result for both countries. In the American case this appears to arise from 3 pairs of questions: (1, 2), (1, 4), and (4, 5). The relevant residuals and frequencies are again set out as 2×2 tables in Table 5.8.

We observe that there is a greater tendency to respond positively to questions 1 and 2 and to 4 and 5 than the model predicts. In the case of 1 and 4 the reverse is true.

Reference to the Appendix shows that questions 1 and 2 are very similar. It may be that some respondents recognised this and tended to give the same answer to both. This amounts to saying that there is a direct link between answers to the questions in addition to the indirect link via the latent variable. Only the latter is provided for by the model.

Questions 4 and 5 both begin with the words 'most people'. It is possible that this pre-disposes some respondents to agree out of a desire to 'go with the crowd'. This phenomenon is known to occur in survey work where such people are described as 'yea-sayers'. If this is the case it means that there is a second factor at work which operates independently of the sentiment of the questions.

Questions 1 and 4, like 1 and 2, are very similar except, in this case, one puts essentially the same proposition negatively and the other positively.

Reiser, Wallace, and Schuessler (1986) investigated the effect of the direction-of-wording of social life feeling items and observed that individuals are generally more likely to endorse a negative item than disagree with a positive item, all else being equal. This may be the reason why the rate of agreeing with item 1 is much higher than the rate of disagreeing with item 4 but it is not clear how we can interpret this effect for pairs of responses.

These conclusions are corroborated by inspection of the three-way residuals. The triplet (1, 2, 4), with positive answers to 1 and 2 and a negative answer to 4, occurred 285 times against a prediction of 251.8 giving a residual of 4.4.

Since the German data also gave a significant departure overall, we might expect the same patterns in the residuals to emerge. A similar analysis to that given above shows that questions 4 and 5 are again implicated but this time in the pairs (3, 4) and (5, 6). The questions in this case were asked in German and it is not clear how far the arguments used above would apply in a different cultural context. Questions 3 and 4 are similar in the view they express as are 5 and 6. Again there is a hint that there may be a direct link between these variables.

None of the deviations we have noted for Scale 2 is unduly large and we would not press the particular interpretations we have put upon them. The logit/probit model does provide reasonable predictions of almost all of the two and three-way marginal responses. The main practical conclusion of the analysis is the warning to avoid questions with similar wording especially if they occur close together in the sequence. Similarly, one should avoid phraseology which might encourage respondents to agree (or disagree) regardless of the sentiment expressed. In the case of Scale 2 it seems unlikely that either of these features is important enough to diminish the value of the scale.

Scale 3 There are no residuals greater than 4 either for the American or German sample, although the Monte-Carlo tests reject the model for the German sample.

Scale 4 In the American sample, the two-way margins are all small and there are only two greater than 4 three way residuals, in particular: response 1,1,1 to items 1, 4 and 7: 6.34 and the same response to items 2,3 and 6: 4.93. The Monte-Carlo test of all indices accept the model except for G^2/f and X^2/f which were above the 5% significance level. The German sample has no residuals greater than 4 and all Monte-Carlo tests accept the model.

Scale 5 There were no large two- or three-way residuals for either the American or the German sample, although the Monte-Carlo tests reject the model for both samples.

Scale 6 All two- and three-way residuals are small for the American sample and all Monte-Carlo tests accept the model. In the German sample there is one relatively large residual, the three-way margin to items 2,7 and 9 is 6.1. Most of the Monte-Carlo tests accept the model except for the pooled and unpooled G^2.

Scale 8 In the case of Scale 8, in the American sample, there are no two-way residuals greater than 4. However, the Monte-Carlo test rejects the model. In the German sample, there is only one two-way residual greater than 4 and no three-way residuals greater than 0.89 so again there is little evidence of a departure from the model. The Monte-Carlo tests accept the model.

Scale 9 The American sample for scale 9 fitted the model well, since all residuals are small and all Monte-Carlo tests accept the model. The case of the German sample is more interesting because the largest two-way residuals all involve question 5. These are as follows:

$$(2, 5), 5.6; \quad (5, 7), 5.4 \text{ and } \quad (5, 12), 7.7$$

Table 5.9: Scale 10 residuals for variables 1 and 5

		American		German	
		1	0	1	0
O		81	65	101	94
E	1	60.8	85.5	85.8	109.1
$(O-E)^2/E$		6.7	4.7	2.7	2.1
O		248	1022	319	976
E	0	268.3	1001.8	334.3	960.7
$(O-E)^2/E$		1.5	0.4	0.7	0.2

The rows relate to question 1 and the columns to question 5.

When we turn to the three-way residuals, there are several in the range 7-10 and all involve question 5. The scale in question is named 'Future Outlook' and we note that question 5 has no future reference. It is also worth noting that the phrase 'this country is sick' is an American expression which may have had a different connotation in the German context.

If question 5 is omitted all three-way residuals are small and the only two-way residuals which are greater than 4 relate to different pairs. When Monte Carlo tests for the reduced set of 11 items were carried out, the significance levels vary between just below 5% to just above 1%. Overall we might regard the fit as just acceptable. In view of our earlier simulations it is interesting that the global test on the full set of 12 items failed to detect the effect of the anomalous question 5.

Scale 10 In this case there is evidence for a significant departure in the American data. The source of the trouble appears to lie with the pair (1, 5). The residuals are high both for this pair and the triplets into which it enters. The position for the two-way margins for the American and the German data is set out in Table 5.9.

In the German case the residuals are not large enough to attract attention on their own but the deviations are in the same direction as those yielding the much larger American residuals. Reference to the Appendix shows that there is no obvious link in form or content between the questions. It may be that there is a second latent variable operating. This view is supported by the observation that, although the scale is labelled 'economic self-determination', question 5 has no economic reference. The sense of personal autonomy which it expresses may have a broader connotation.

Looking at all scales, for most cases the rejection of the model by the Monte-Carlo tests was confirmed by large two- and three-way marginal residuals and the acceptance of the model confirmed by small residuals. However, there were some discrepancies between the Monte-Carlo tests and our expectations from the analysis of the residuals. In particular, for five scales (American scales 5 and 8 and German scales 3 and 5 and 10) the Monte-Carlo tests rejected the model although there were no large two- or three-way marginal residuals (for scale 10 though, the Monte-Carlo test agrees with the result from the chi-squared distribution which should be valid in that case). This shows that any deviations from the model are not such as to affect the expected pair-wise associations.

For the cases where the Monte-Carlo test failed to detect departure from the model but had a few high residuals (German scales 1 and 9), the failing of the Monte-Carlo test could be attributed to the loss of power due to the large number of items. On the other hand, the analysis of the residuals does not constitute a formal test and also, as p increases residuals might be expected to get larger.

5.6 Discussion

Our results show that the G^2 and X^2 tests for goodness-of-fit based on the chi-squared approximation are liable to grossly over-state the degree of significance. As an alternative we recommend the use of Monte-Carlo tests supplemented by an examination of the residuals. Although our simulations are based on 1000 replications, two hundred or so, samples are sufficient to obtain a good general indication of the quality of the fit and this is computationally feasible. The variety of statistics we have considered give broadly similar results.

Reiser and VandenBerg (1994) also concluded that both G^2 and X^2 over-state the significance for small values of p up to about 7 but that as p increases beyond that the type I errors fall rapidly to zero. This implies that it is hardly ever possible to get a value of G^2, for example, which is significant. This is different from what we observed in our data but other studies (see Section 1.10) noted differences in the way G^2 and X^2 behave when the number of items is large. These differences arise from the degree of sparseness and the number of empty cells. The latter greatly influence the calculation of the degrees of freedom. If the table is sparse but with no empty cells then the degrees of freedom will be 2^p- the number of parameters, which will be a very large nulmber, so G^2 and X^2 may not come out significant. On the other hand, if there are many eimpty cells the degrees of freedom will be $n-$ the number of parameters, a much smaller number.

Both Collins et al. (1993) and Langeheine et al. (1996) conclude that the chi-squared approximations to the G^2 and X^2 tests are inadequate for sparse tables and that Monte Carlo tests are the only viable option. Langeheine et al. (1996) use the term "nonnaive' bootstrap for what we have called the Monte Carlo test and argue that this is to be preferred to the naive bootstrap. The distinction is made in terms of whether we sample from the 'data' (naive) or the 'model' (nonnaive) but their method is different from ours. They give a procedure for sampling from a multinomial distribution over the 2^p cells of the table with probabilities estimated from the model and 'n' equal to the sample size. In contrast, we follow Collins et al. (1993) and simulate response patterns for each individual in the sample using the estimated parameters of the model. Individuals are then allocated to cells of the table according to the outcomes. The two methods are equivalent. Ours, of course, is specific to the logit/probit model but has advantages when 2^p becomes very large. This is because we generate n standard normal variates (z's) and p Bernoulli variates (x_i's) for each trial. The method of Langeheine et al. (1996) involves generating $2^p - 1$ binomial variates, it having first been necessary to estimate the probabilities for that number of cells. If p is 15 or 20, say, this becomes a formidable undertaking as these authors note.

Langeheine et al. (1996) and Collins et al. (1993) consider X^2 and G^2 as special cases of the Read and Cressie index, which is a wider class of measures of fit known as power divergence statistics, but they do not consider derived statistics such as $\%G^2$ or G^2/f.

The Monte-Carlo test though will not provide an adequate solution if p is very large either. Simulations with artificial data generated from the 1-factor model showed that the size of the test for all indices diminishes as p gets large. The best behaved index was the unpooled X^2. Further simulations with artificial data generated from a 2-factor model showed that the power of all tests except of the unpooled X^2 also diminishes when the number of items gets large, in our examples greater than 14, though this has also to do with how far the data deviate from the model.

We complemented the Monte-Carlo tests by the analysis of the two- and three-way margins of the residuals. The examination of residuals may help to identify the reasons for a poor fit. For example, our observations on question wording suggested that there might be direct links between responses to some questions.

Chapter 6

Conclusions

This thesis focused on the 2-parameter latent trait model for binary data. It examined the robustness of the model under violations of its assumptions, in particular when the data are contaminated, and when the assumption about the prior distribution is violated. It further examined the semiparametric estimation of the model, that is estimating the model without assuming a parametric form of the prior, and investigated the information on the distribution of the latent variable that can be retrieved from a set of binary responses. Finally it examined the goodness-of-fit of the model using Monte-Carlo simulations and diagnostic methods based on the residuals.

In Chapter 1 we reviewed estimation methods of latent trait models, and studies on the robustness of latent trait models. We presented the tools from robust statistics theory that were used in this thesis to examine the robustness of the 2-parameter model. We also reviewed studies on the semiparametric estimation of the Rasch model and mixing distributions, as the semiparametric estimation of the model we proposed is based on these methods. We finally presented the goodness-of-fit problem which arises from sparse tables and affects the latent trait model and reviewed studies using Monte-Carlo methods as an alternative to traditional methods for latent class and loglinear models.

In Chapter 2 we investigated the robustness of the item parameters and the posterior means when the data are contaminated. The data that can be analysed with a latent trait model are often the results of ability tests, psychometric tests or attitude questionnaires. The responses are often influenced by other factors than the assumed latent variable, for example cheating, faking, or misrecording the answers. Since the data are binary, a misrecorded answer for an item causes the frequency of a response pattern to decrease by one and the frequency of another to increase by one. We examined the behaviour of the item parameters and the posterior means when extra probability was placed on a response pattern, when the probability of an item was increased and when observations were shifted between response patterns. To do this we first derived the Influence Function and examined its behaviour. The Influence Function measures the effect on the estimator when extra probability is placed on a response pattern. We complemented the IF analysis by examining the behaviour of the parameter estimates and the posterior means when artificially contaminating the data in several ways. The effect of putting extra observations on a response pattern was in most cases confined, that is the parameters changed smoothly with increasing amount of contamination and remained within the confidence intervals of the original parameter estimates when the amount of contamination was small. In some cases though, putting observations on a response pattern would throw some parameters out of the original confidence interval for the data examined. In those cases the ordering of the posterior means would also be affected. Increasing the probability of positive response of an item had small or no effects on the parameters. On the other hand, shifting observations between response patterns would cause larger effects than increasing the frequency of a response pattern alone. For the data examined, shifting less than 1% of the total observations between some response patterns brought some parameters outside the confidence intervals of the original estimates, indicating a very small breakdown point for the estimator. Such changes in the frequency distribution of the response patterns caused also large changes in the ordering of the response patterns according to the posterior means.

It would be interesting to investigate in the future gross changes in the frequency distribution, for example missing out a frequency of a response pattern or transposing frequencies between two response patterns.

Methods to robustify the Maximum Likelihood estimation of the latent variable suggested in the literature, such as 'jackknifing' the posterior means (or the parameters) and the biweight estimation of ability proved no more robust than the usual procedures. Part of further research will be to develop a robust estimator by placing bounds on the Influence Function, as suggested by Huber (1981).

The parametric estimation of the latent trait model requires an assumption about the form of the latent variable, the prior distribution. This has been usually assumed to be the standard normal. In Chapter 3 we examined the sensitivity of the model when the assumption of normality of the prior distribution is violated. We derived the IF for the parameters and the posterior means for small changes in the prior. We also examined the behaviour of the parameters and the posterior means when the data were fitted with a prior contaminated in a similar way, i.e. with a mixture of a N(0,1) and a small probability on a point. Changes in the parameters and the posterior means seemed small, certainly smaller than when the same percentage of probability was placed on a response pattern, which was expected, since contamination on the prior spreads on several response patterns, thus averaging out changes in the parameters or the posterior means. We also fitted mixtures of normals as priors to measure the effect of large changes in the prior distribution. The parameters showed very small deviations from the original parameters after they had been standardised with an appropriate measure of location and dispersion, i.e. the mean and standard deviation respectively for symmetric distributions and the median and interquartile ratio for skewed distributions.

In Chapter 4 we explored the semiparametric and fully semiparametric estimation of the latent trait model, i.e. making no assumption about the form of the prior but having either fixed nodes on a grid and estimating the weights (simple semiparametric estimation) or estimating both the points and the weights simultaneously with the item parameters (fully semiparametric estimation). The estimation is achieved by an EM algorithm. Bock and Aitkin (1981) proposed the first method. We extended the algorithm by adding an extra E- and M-step to accommodate the simultaneous estimation of the nodes. The solution is checked with optimality criteria proposed by Lindsay (1983). Estimating the model fully semiparametrically makes the use of the 2-parameter latent trait model more attractive, since there is no need to make an assumption about the form of the latent distribution. It also brings the model closer to latent class analysis, and particularly when restrictions on the order of the latent classes are placed, but the latent trait model has the advantages that it is more parsimonious and formal criteria on the number of nodes / classes needed can be used. We further used the bootstrap to measure the variability of the estimated prior and thus assess the information that can be retrieved from a set of binary responses on the distribution of the latent variable. The envelopes formed by joining the bootstrap percentile confidence intervals of the percentiles of the prior distribution followed closely the shape of the smoothed prior, indicating that there is information about the form of the distribution in the data. The shape was, for some datasets, different from the N(0,1) which is the usual assumption about the prior. Furthermore, the empirical prior estimated for artificial data seemed to recover satisfactorily the form of the distribution that generated the data. A limitation of our algorithm is that it converges slowly and it could be perhaps improved by incorporating the use of the gradient D given by Lindsay (1983) to speed up the search for the optimal solution.

In Chapter 5 we examined the goodness-of-fit of the 2-parameter latent trait model using Monte-Carlo methods. Since the G^2 and X^2 statistics do not follow the chi-squared distributions if the cross-classification table of the data is sparse, we used Monte-Carlo simulations to approximate the empirical distribution of G^2, X^2 and other derived statistics. The Monte-Carlo test though seemed to lose power as the number of items increased and therefore we complemented the results based on the empirical distribution of the statistics with analyses of the 2- and 3-way marginal residuals. These are more informative as they may also indicate the source of any discrepancies of the model.

As the Monte-Carlo test provides a more formal means to test the model it would be worth examining further the properties of the test and also finding ways to improve its performance when the number of items get large.

Appendix

Intercultural 'Social Life Feeling' Scales

Schuessler conducted a survey comprising 114 questions on 'Social Life Feelings' on a sample of 1416 Americans (Schuessler 1982). From these questions twelve scales comprising 95 items were formed. The survey was later repeated in Germany, reaching a sample of 1490. After further analysis of the questions more items and a whole scale (scale 7) were eliminated. The resulting scales are the so called intercultural scales (Krebs and Schuessler 1987).

The data are coded in a way that '1' conveys a 'negative' feeling, such as 'doubt' or 'cynicism' in the scales below.

Scale 1: 'Doubt about Self-Determination'.

1. There are few people in this world you can trust, when you get right down to it.

2. If the odds are against you, it's impossible to come out on top.

3. The average person can get nowhere by talking to public officials.

4. The future is too uncertain for a person to plan ahead.

5. Nowadays a person has to live pretty much for today and let tomorrow take care of itself.

6. What happens in life is largely a matter of chance.

7. I've had more than my share of troubles.

8. The world is too complicated for me to understand.

9. I regret having missed so many chances in the past.

10. I have little influence over the things that happen to me.

11. For me one day is no different from another.

12. I sometimes feel that I have little control over the direction my life is taking.

Scale 2: 'Doubt about Trustworthiness of People'.

1. It is hard to figure out who you can really trust these days.

2. There are few people in this world you can trust, when you get right down to it.

3. Strangers can generally be trusted.

4. Most people can be trusted.

5. Most people don't really care what happens to the next fellow.

6. Many people are friendly only because they want something from you.

Scale 3: 'Feeling Down'.

1. At times I feel that I am a stranger to myself.

2. I sometimes feel forgotten by friends.

3. Out of place.

4. That my life is not very useful.

5. I feel somewhat apart even among friends.

6. Very lonely or remote from other people.

7. Depressed or very unhappy.

8. Bored.

9. Vaguely uneasy about something without knowing why.

Scale 4: 'Job Satisfaction'.

1. I am satisfied with the work I do.
2. I would like more freedom on the job.
3. People feel like they belong where I work.
4. There is too little variety in my job.
5. My job gives me a chance to do what I do best.
6. I have too small a share in deciding matters that affect my work.
7. My job means more to me than just money.
8. There must be better places to work.

Scale 5: 'Faith in Citizen Involvement'.

1. The average person has considerable influence on politics.
2. The average person has much to say about running local government.
3. Taking everything into account, the world is getting better.
4. The average person has a great deal of influence on government decisions.
5. People like me have much to say about politics.
6. By taking part in political and social affairs the people can control world events.
7. I am usually interested in local elections.

Scale 6: 'Feeling Up'.

1. Anyone can raise his standard of living if he is willing to work at it.
2. That my life is not very useful.
3. I have a great deal in common with most people.
4. Things get better for me as I get older.

5. When I make plans, I am almost certain that I can make them work.

6. I get a lot of fun out of life.

7. There is much purpose to what I am doing at present.

8. I am satisfied with the way things are working out for me.

Scale 8: 'Disillusionment with Government'.

1. In my opinion this country is sick.

2. Our local government costs the taxpayer more than it is worth.

3. Our country has too many poor people who can do little to raise their standard of living.

4. Most politicians are more interested in themselves than in the public.

5. We are slowly losing our freedom to the government.

6. The average person has much to say about running local government.

7. Most supermarkets are honestly run.

8. I have little confidence in the government today.

Scale 9: 'Future Outlook'.

1. Many people will be out of work in the next few years.

2. Although things keep changing all the time, one still knows what to expect from one day to another.

3. The future of this country is very uncertain.

4. The future looks very bleak.

5. In my opinion this country is sick.

6. The future is too uncertain to plan ahead.

7. We are slowly losing our freedom to the government.

8. The lot of the average man is getting worse, not better.

9. It's unfair to bring children into the world with the way things look for the future.

10. Many things our parents stood for are going down the drain.

11. I have little confidence in the government today.

12. The future looks very bright to me.

Scale 10: 'Economic Self Determination'.

1. Anyone can raise his standard of living if he is willing to work at it.

2. Our country has too many poor people who can do little to raise their standard of living.

3. Individuals are poor because of the lack of effort on their part.

4. Poor people could improve their lot if they tried.

5. Most people have a good deal of freedom in deciding how to live.

American scale

Scale 7:'People cynicism'

1. In a society where almost everyone is out for himself, people soon come to distrust each other.

2. Most people know what to do with their lives.

3. Too many people in our society are just out for themselves and don't really care for anyone else.

4. Many people in our society are lonely and unrelated to their fellow human beings.

5. Many people are friendly only because they want something from you.

6. Many people don't know what to do with their lives.

NFER data

Test 1 is a reading test and corresponds to a story entitled 'King Lion' (NFER workbook number 0247). The story is written as a fable, in which a small animal, a squirrel, outwits a more powerful one, King Lion. The lion announces that in order to save the animals work in fetching his food, he will eat one of them every day, in an order they choose. They are left to decide how to put his suggestions into practice. The squirrel saves everybody's life by leading the lion to a deep pool where, he alleges, a strange creature is waiting for him. On seeing his reflexion on the water, the lion jumps into the pool and drowns.

The questions are the following:

1. In the story it says that the lion would roar for food and his servants came running. Why did they come running?

2 and 3. Write down the names of two of Lion's servants.

4. On the day when the story takes place, Lion was in an atrocious temper. 'Atrocious' means....

5. When Lion looked at his servants 'with a cold eye', what sort of look do you think this was?

6. Why did Lion's servants squeak when they spoke to him?

7. The lion said, 'It's been inconsiderate of me to make all of you come with the food. You 're overworked'. Was the lion really worried that the animals might be overworked?

8. How can you tell?

9. 'In order to improve the service I'm going to make a small change in the way things are done. It's been inconsiderate of me to make all of you come with the food. You 're overworked'. Does this show Lion was kind hearted and cared about his servants? Explain why or why not.

10. The Lion said he was going to make a small change in the way things were run. What was the change that he wanted to make?

11. Why were the animals silent for the first hour of their meeting and why did they look at each other out of the corners of their eyes?

12. What was the problem with Wildebeeste's plan to choose who went first?

13. What was wrong with Wart Hog's plan to decide who went first?

14. If you had been at that meeting would you have thought of a better plan to decide

who the lion would eat first? What would it have been?

15. Why did all the animals shout 'WHAT?' when Ground Squirrel said 'I'll go'?

16. Did you notice anything about the squirrel's attitude to the lion? Is it different from that of the other animals?

17. How did Squirrel persuade King Lion to go with him down to the river?

18. If Squirrel had said 'King Lion, there's a creature by the water. You must go down and see it immediately', what do you think Lion would have said?

19 and 20. King Lion and the Ground Squirrel go down to the deepest pool on the river. Explain, in your own words, what happens after this.

21. Do you think Lion really deserved what happened to him?

Attitude to employment data

The data we used were taken from Albanese and Knott (1992b) and Birkhoff (1991). They are the responses to 4 items chosen from 14 items concerning the attitude to work of 1915 German company employees in 1987. These items are part of an investigation about what the employees thought to be the strengths and weaknesses of the company and how they felt about their personal situation at the work place. For each item the respondents were asked if they agree or disagree with each of the following statements:

1. My work is interesting because I have the feeling that I am needed.

2. In my work I find self-assurance and appreciation.

3. My work is fun.

4. Sometimes I have the feeling that I am exerting myself day after day and still I see no success.

For items 1 to 3 'agree' was coded as 1 and 'disagree' was as 0. Item 4 conveys a feeling opposite to that of the other items, so 'agree' was coded as 0, and 'disagree' as 1. Thus a positive response (1) indicates that the respondent has a positive attitude towards his/her work, is generally satisfied with it, whereas a negative response indicates the opposite.

References

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing 6*, 251–262.

Albanese, M.-T. (1990). *Latent Variable Models for Binary Response.* Ph. D. thesis, University of London.

Albanese, M.-T. and M. Knott (1992a). TWOMISS: A computer program for fitting a one- or two- factor logit-probit latent variable model to binary data when observations may be missing. Technical report, Statistics Department, London School of Economics and Political Science.

Albanese, M.-T. and M. Knott (1992b, August). A latent variable for attitudes to employment. In *Proceedings of the 10th Simpósio Nacional de Probobilidade e Estatistica*, Rio de Janeiro.

Andersen, E. (1973). Conditional inference and multiple choice questionnaires. *British Journal of Mathematical and Statistical Psychology 26*, 31–44.

Andrews, D., P. Bickel, F. Hampel, P. Huber, W. Rogers, and J. Tukey (1972). *Robust Estimates of Location.* Princeton University Press.

Bartholomew, D. J. (1980). Factor analysis for categorical data. *Journal of the Royal Statistical Society B 42*(3), 293–321.

Bartholomew, D. J. (1984). Scaling binary data using a factor model. *Journal of the Royal Statistical Society, Series B 46*(1), 120–123.

Bartholomew, D. J. (1988). The sensitivity of latent trait analysis to choice of prior distribution. *British Journal of Mathematical and Statistical Psychology 41*, 101–107.

Bartholomew, D. J. (1993). The statistical approach to social measurement. In D. Krebs and P. Schimdt (Eds.), *New Directions in Attitude Measurement.* Berlin, New York: Walter de Gruyter.

Bartholomew, D. J. and M. Knott (1999). *Latent Variable Models and Factor Analysis* (2nd ed.). London: Arnold.

Bartholomew, D. J. and P. Tzamourani (1999). The goodness-of-fit of latent trait models in attitude measurement. *Sociological Methods and Research 27*(4), 525–546.

Birkhoff, B. (1991). *Thesis on attitudes to employment.* Ph. D. thesis, University of London.

Birnbaum, A. (1968). *Some Latent Trait Models and their Use in Inferring an Examinee's Ability.* Reading, MA: Addison-Wesley.

Bock, R. and M. Aitkin (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika 46*(4), 443–459.

Bock, R. and M. Lieberman (1970). Fitting a response model for n dichotomously scored items. *Psychometrika 35*(2), 179–197.

Boehning, D. (1985). Numerical estimation of a probability measure. *Journal of Statistical Planning and Inference 11*, 57–69.

Boehning, D. (1995). A review of reliable maximum likelihood algorithms for semiparametric mixture models. *Journal of Statistical Planning and Inference 47*, 5–28.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika 40*(1), 5–31.

Collins, L. M., P. L. Fidler, S. E. Wugalter, and J. D. Long (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research 28*(3), 375–389.

Cressie, N. and P. W. Holland (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika 48*(1), 129–141.

Croon, M. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology 43*, 171–192.

Davies, R. (1987). Mass point methods for dealing with nuisance parameters in longitudinal studies. In R. Crouchley (Ed.), *Longitudinal Data Analysis.* Aldershot, Hants: Avebury.

De Leeuw, J. and N. Verhelst (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics 11*(3), 183–196.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of The Royal Statistical Society B 39*, 1–38.

Efron, B. and R. Tibshirani (1993). *An Introduction to the Bootstrap*, Volume 57 of *Monographs on Statistics and Applied Probability.* New York: Chapman & Hall.

Everitt, B. (1988). A Monte-Carlo investigation of the likelihood ratio test for number of latent classes in latent class analysis. *Multivariate Behavioral Research 23*, 531–538.

Follmann, D. (1988). Consistent estimation in the Rasch model based on nonparametric margins. *Psychometrika 53*(4), 553–562.

Glas, C. A. (1989). RIDA, computer program for Rasch incomplete design analysis. Technical report, Arnhem, The Netherlands: National Institute for Educational Measurement (CITO).

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika 61*, 215–231.

Gorman, T., J. White, L.Orchard, and A.Tate (1981). Language performance in schools, primary survey report no.1. Technical report, Assessment of Performance Unit, Department of Education and Science.

Haberman, S. (1979). *Analysis of Qualitative Data, Vol.1. Introductory Topics.* New York: Academic Press.

Haertel, E. H. (1990). Continuous and discrete latent structure models for item response data. *Psychometrika 55*(3), 477–494.

Hampel, F. (1968). *Contributions to the theory of robust estimation.* Ph. D. thesis, University of California, Berkeley.

Hampel, F. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association 69*(346), 383–393.

Hampel, F., E. Ronchetti, P. Rousseeuw, and W. Stahel (1986). *Robust Statistics, The Approach Based on Influence Functions.* Wiley Series in Probability and Mathematical Sciences. New York: Wiley.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences.* London: Sage Publications.

Hinde, J. and A. Wood (1987). Binomial variance component models with a non-parametric assumption concerning random effects. In R. Crouchley (Ed.), *Longitudinal Data Analysis.* Aldershot, Hants: Avebury.

Huber, P. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics 35*, 73–101.

Huber, P. J. (1981). *Robust Statistics.* New York: Wiley.

Karlin, S. and W. Studden (1966). *Tchebyscheff systems: with applications to analysis and statistics.* New York: Wiley.

Kiefer, J. and J. Wolfowitz (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics 27*, 887–906.

Knott, M. and M.-T. Albanese (1993). Conditional distributions of a latent variable and scoring for binary data. *Revista Brasileira de Probabilidade e Estatistica 6*, 171–188.

Koehler, K. and K. Larntz (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association 75*, 336–344.

Krebs, D. and K. F. Schuessler (1987). *Soziale Empfindungen : ein interkultureller Skalenvergleich bei Deutschen und Amerikanern.* Monographien: Sozialwissenschaftliche Methoden. Frankfurt/Main, New York: Campus Verlag.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association 73*(364), 805–811.

Langeheine, R., J. Pannekoek, and F. van de Pol (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods and Research 24*(4), 492–516.

Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *Journal of the American Statistical Association 73*, 253–263.

Lawley, D. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh 61*, 273–287.

Lazarsfeld, P. (1950). The logical and mathematical foundation of latent structure analysis. In S. Stouffer (Ed.), *Measurement and Prediction.* New York: Wiley.

Lesperance, M. and J. Kalbfleisch (1992). An algorithm for computing the nonparametric MLD of mixing distribution. *American Statistical Association 87*(417), 120–126.

Levine, M. and F. Drasgow (1983). Appropriateness measurement:validating studies and variable ability models. In D. Weiss (Ed.), *New Horizons in Testing.* New York: Academic Press.

Levine, M. V. and D. B. Rubin (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics 4*, 269–290.

Lindsay, B., C. Clogg, and J. Grego (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association 86*(413), 96–107.

Lindsay, B. and M. Lesperance (1995). A review of semiparametric mixture models. *Journal of Statistical Planning and Inference 47*, 29–39.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: A general theory. *The Annals of Statistics 11*(1), 86–94.

Lord, F. M. and M. R. Novick (1968). *Statistical theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Mislevy, R. (1984). Estimating latent distributions. *Psychometrika 49*(3), 359–381.

Mislevy, R. and R. Bock (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement 42*, 725–737.

Mislevy, R. and R. Bock (1990). BILOG 3, item analysis and test scoring with binary logistic models.

Mood, A., F. Graybill, and D.C.Boes (1963). *Introduction to the Theory of Statistics.* New-York: McGraw-Hill.

Mosteller, F. and J. F. Tukey (1977). *Data Analysis and Regression.* Philippines: Addison-Wesley.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika 43*(4), 551–560.

Prohorov, Y. (1956). Convergence of random processes and limit theorems in probability theory. *Theor. Prob. Appl. 1*, 157–214.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: The Danish Institute for Educational Research.

Read, T. R. C. and N. A. C. Cressie (1988). *Goodness-of-fit statistics for discrete multivariate data.* Springer Series in Statistics. New York: Springer-Verlag.

Reiser, M. and M. VandenBerg (1994). Validity of the chi-square test in dichotomous variable factor analysis when expected frequencies are small. *British Journal of Mathematical and Statistical Psychology 47*, 85–107.

Reiser, M., M. Wallace, and K. F. Schuessler (1986). Direction-of-wording effects in dichotomous social life feeling items. *Sociological Methodology 16*, 1–25.

Schuessler, K. F. (1982). *Measuring Social Life Feelings.* The Jossey-Bass Social and Behavioral Science Series. Jossey-Bass.

Seong, T.-J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement 14*(3), 299–311.

Shea, B. (1984). FACONE: A computer program for fitting the logit latent variable model by maximum likelihood. Technical report, Statistics Department, London School of Economics and Political Science.

Smith, R. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement 45*, 433–444.

Smith, R. M. (1982). *Detecting Measurement Disturbances with the Rasch Model.* Ph. D. thesis, University of Chicago.

Straud, A. and D. Sechrest (1966). *Gaussian Quadrature Formulas.* Prentice-Hall.

Susko, E., J. Kalbfleisch, and J. Chen (1998). Constrained nonparametric maximum-likelihood estimation for mixture models. *Canadian Journal of Statistics 26*(4), 601–617.

Trabin, T. E. and D. J. Weiss (1983). The person response curve: Fit of individuals to item response theory models. In D. Weiss (Ed.), *New Horizons in Testing.* New York: Academic Press.

Vermunt, J. K. (1997). LEM: A general program for the analysis of categorical data. Computer manual, Tilburg University.

Wainer, H. and B. Wright (1980). Robust estimation of ability in the Rasch model. *Psychometrika 45*, 373–391.

Waller, M. I. (1974). Removing the effects of random guessing from latent trait ability estimates. Research report, Educational Testing Service, Princeton, New Jersey.

Wright, B. (1980). *Afterword in probabilistic models for some intelligence and attainment tests* by G.Rasch. University of Chicago Press.

Zwinderman, A. H. and A. L. van den Wollenberg (1990). Robustness of marginal maximum likelihood estimation in the Rasch model. *Apllied Psychological Measurement 14*(1), 73–81.