Nonparametric Estimation of Item Response Functions Using the EM Algorithm

Natasha T. Rossi

Department of Psychology

McGill University, Montreal

July 2001

A thesis submitted to the Faculty of

Graduate Studies and Research in partial fulfilment

of the requirements of the degree of Master of Arts.

Abstract

Bock and Aitkin (1981) developed an EM algorithm for the maximum marginal likelihood estimation of parametric item response curves, such that these estimates could be obtained in the absence of the estimation of examinee parameters. Using functional data analytic techniques described by Ramsay and Silverman (1997), this algorithm is extended to achieve *nonparametric* estimates of item response functions. Unlike their parametric counterparts, nonparametric functions have the freedom to adopt any possible shape, making the current approach an attractive alternative to the popular three-parameter logistic model. A basis function expansion is described for the item response functions, as is a roughness penalty which mediates a compromise between the fit of the data and the smoothness of the estimate. The algorithm is developed and applied to both actual and simulated data to illustrate its performance, and how the nonparametric estimates compare to results obtained through more classical methods.

Résumé

Bock et Aitkin (1981) ont développé un algorithme EM pour l'estimation de vraisemblance marginale maximum des fonctions de réponse d'item paramétriques, telle que ces estimations puissent être obtenues en l'absence d'estimations de paramètres examinés. Utilisant des techniques analytiques de données fonctionnelles décrites par Ramsay et Silverman (1997), cet algorithme est élargi afin d'obtenir une estimation non-paramétrique des fonctions de réponse d'item. Contrairement à leur équivalent paramétrique, les fonctions non-paramétriques ont la liberté d'adopter n'importe quelle forme. Ceci rend cette alternative plus populaire que le modèle logistique à trois paramètres. Une expansion de fonction de base est décrite pour les fonctions de réponse d'item, comme l'est une pénalité de rudesse qui négocie un compromis entre la compatibilité de donnée et la fluidité de l'estimation. La performance de l'algorithme est illustrée pour la donnée simulée actuelle. Les extensions et les limites de la méthode sont abordées.

Acknowledgements

I wish to express my heartfelt thanks to my supervisor, Professor James O. Ramsay, for his unending encouragement, commitment and generosity during the course of my studies. I am greatly indebted to him for the many careful readings of earlier drafts of this thesis, and for providing me with invaluable advice.

I shall be forever grateful to Professor Michael Maraun for introducing me to the field of Psychometrics, and for inspiring me to pursue graduate work.

I would also like to thank the faculty, staff and fellow students of the Department of Psychology at McGill University for providing a warm and friendly environment in which to conduct my graduate studies.

Many thanks to Ms. Giovanna LoCascio for her kindness and for always having the answer.

Finally, I am very grateful to my parents, Tony and Renata Rossi, for their constant love, support and encouragement in all of my endeavors.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview of the Thesis

This thesis will describe a procedure for the nonparametric estimation of item response functions using the EM algorithm. Among the key advantages of this model is the nonparametric approach to item response function estimation. Presumptions regarding the shape features of these functions are avoided, as are problems associated with item parameter estimation. Thus, this model is an attractive alternative to the popular three-parameter logistic model. Employment of the EM algorithm has the advantage of eliminating the estimation of examinee ability parameters, and so allows for computational speed and simplicity. As the data are viewed as functions, functional data analytic techniques can be applied. This provides for smooth estimates of the item response functions and hence the availability of derivative information.

Chapter 1 provides an introduction to the central tenets of item response theory and the item parameters most commonly used to describe the shape features of item response functions. Several parametric models are described, and a distinction is made between these and nonparametric models. Functional data analysis is defined and its role in curve estimation is outlined. The remainder of the chapter briefly introduces the EM algorithm and describes how it is currently

being used to estimate the parameters of the three-parameter logistic model. In chapter 2, the details of how functional data analytic techniques are used to estimate nonparametric item response functions are presented. Here, the focus is shifted from estimating the functions directly to estimating the logit transformations. The concept of a basis function is discussed, and a basis function expansion is considered for the item response functions. Chapter 3 provides the details of how marginal maximum likelihood estimation and the EM algorithm are used to estimate the item response functions. The quadrature rule used to estimate the integral in the E-step is described, as is a roughness penalty which mediates a compromise between the fit of the data and the smoothness of the estimate. In Chapter 4, two example analyses using the algorithm are described. The first is a simulated three-parameter logistic test, such that the performance of the algorithm may be assessed in a situation where the data are in fact describable by a parametric model. The second analysis involves a set of real test data. Finally, Chapter 5 provides a summary of the advantages of the current procedure, its limitations and suggestions for the direction of future research.

1.2 Data and Notation

The model presented here is a unidimensional model of responses to dichotomous test items. The data to be analyzed are the responses of examinees, indexed by $a = 1, \ldots, N$ to a set of test items, indexed by $i = 1, \ldots, n$. The response to item i by examinee a is coded by the binary variable u_{ai}, which takes a value of 1 if the item is answered correctly and a value of 0 otherwise. The set of examinee a's responses to the n test items is denoted by the response vector $u_a = (u_{a1}, \ldots, u_{an})'$. The set of response vectors for all examinees can be organized into an $N \times n$ matrix $\mathbf{U}$, where u_a' is the a^{th} row of $\mathbf{U}$.

The notation $P_i(\theta)$ denotes the probability of responding correctly to item i

given ability level θ,

$$P_i(\theta) = \text{Prob}(u_i = 1|\theta).$$

Since the proposed model is concerned only with dichotomous items, the probability of responding incorrectly to item i can be denoted as $1 - P_i(\theta)$, or $Q_i(\theta)$.

1.3 Item Response Theory

Item response theory (IRT) is based on the assumption that examinee responses to test items can be accounted for by latent traits which are fewer in number than the test items. In most applications it is assumed that a *single* latent trait accounts for responses to test items, with the latent trait most commonly conceptualized as examinee ability level.

1.3.1 Basic concepts

Item response functions

The fundamental concept of item response theory is the *item response function.* The item response function plots the probability of responding correctly to an item as a function of the latent trait, denoted θ, underlying performance on the items of the test. Item response functions are often assumed to have an ogival shape, although they are not limited to be of this type. Figure 1.1 is an example of a typical item response function. Here, even the examinee of the lowest ability has probability of .18 of answering this item correctly, and high ability examinees will almost surely respond correctly. This item response function increases monotonically, with the most drastic changes in $P(\theta)$ occurring for $-1 \leq \theta \leq 1$.

According to Lord (1980), there are two ways in which one can correctly interpret the probability of a correct response for an item:

1. A subpopulation of examinees can be conceived of at each point on the latent trait scale (i.e., a collection of examinees all having the same θ value).

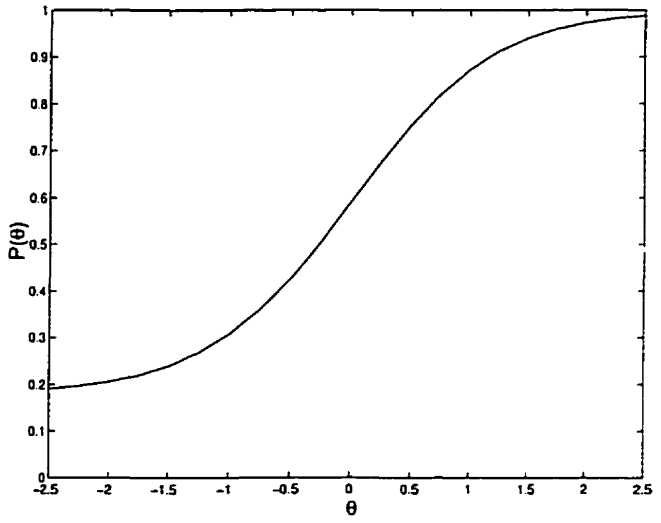

Figure 1.1: A typical item response function. Examinee ability level, θ, is represented on the horizontal axis. The vertical axis represents the probability of a correct response, $P(\theta)$.

Then the probability of responding correctly to an item is the probability that a randomly selected examinee from this homogeneous subpopulation will respond correctly to an item, or the proportion of these examinees who would respond correctly to an item.

2. A subpopulation of items all having the same item response function can be conceptualized. Then the probability of responding correctly is interpreted as the probability that a particular examinee will respond correctly to an item randomly chosen from the subpopulation of items.

Local independence and unidimensionality

Central to item response theory is the assumption that individual examinees respond independently to each and every test item, and independently of one another. That is to say, an examinee's response to item i is not influenced by the response to any other item, nor by the responses of other examinees to item i, nor the responses of any other examinee to any other item. The scores on two items

i and j are said to be *statistically independent* if the joint probability of a correct response to both items is equal to the product of the marginal probabilities, that is,

$$\text{Prob}(u_i = 1 \cap u_j = 1) = P_i(\theta)P_j(\theta). \tag{1.1}$$

If (1.1) does not hold, items i and j are said to be *statistically dependent.*

Item response theory uses the concepts of statistical independence and statistical dependence to describe the relationship between the latent trait, θ, and the probability of responding correctly, $P(\theta)$. The central concept of *unidimensionality* can be defined in terms of statistical dependence. Let it first be assumed that test items are statistically dependent in the population. Then the test is unidimensional if a single latent trait exists such that within each subpopulation of examinees homogeneous with respect to θ, the items are statistically independent. Since this independence holds only for a subpopulation of examinees located at a single point on the θ scale, it is called *local independence.*

It should be emphasized that unidimensionality and local independence are not the same thing. Unidimensionality is the assumption that a *single* latent trait accounts for the statistical dependence among items, i.e., the assumption of only one latent variable will lead to local independence. Local independence, however, may be achieved without unidimensionality. In general, the *dimensionality* of a test refers to the number of latent traits required to obtain local independence. It should also be emphasized that both unidimensionality and local independence are assumptions.

Item parameters

Several parameters may be used to describe the features of the item response functions. The left asymptote or guessing level, which can be denoted $P(-\infty)$, is relevant when the question format permits correct answers by guessing. This parameter allows examinees to have $P(\theta)$ greater than zero even at low values of θ. Multiple choice and true/false question formats are examples of items that are

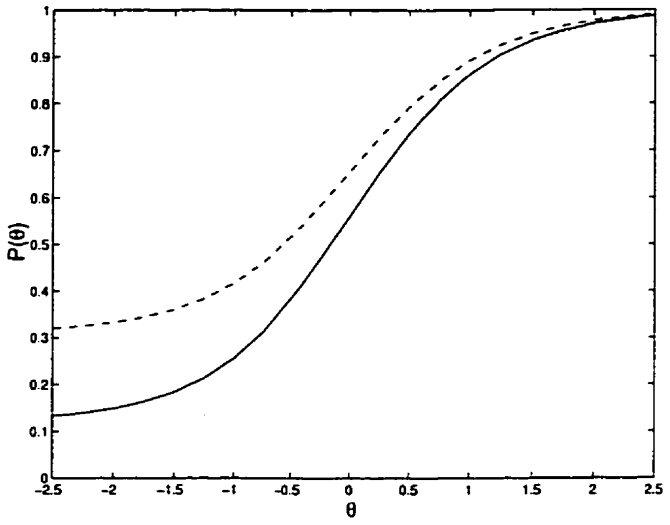

Figure 1.2: Three item response functions varying only with respect to the value of the left asymptote. For the dashed line $P(-\infty) = .31$, for the solid line $P(-\infty) = .12$, and for the dotted line $P(-\infty) = .03$.

likely to have a left asymptote greater than zero. On a multiple choice question with four response options, an examinee, regardless of his or her θ value, has a $1/4$ or .25 probability of choosing the correct response when guessing. As such it is not appropriate for $P(\theta)$ to approach zero as θ approaches $-\infty$. Figure 1.2 shows three item response functions that are identical except for the value of the left asymptote.

The item difficulty or location parameter, which may be denoted by $\bar{P}$, refers to the θ value midway between the guessing level and the right asymptote of unity, $\bar{P}(\theta) = (P(-\infty) + 1)/2$. For the case where $P(-\infty)$ is equal to zero, $\bar{P}$ is the value of θ at which $P(\theta)$ is equal to .5. Items with high $\bar{P}$ are difficult items, where $P(\theta)$ is high only for high ability examinees. Items with low $\bar{P}$ are easy items, where almost all examinees have a high probability of responding to the item correctly. Three items with varying levels of difficultly are shown in Figure 1.3. Ideally, a test should include items of varying difficulty.

Items may also differ from one another in terms of how they differentiate among examinees. The slope of the item response function at θ measures the extent to

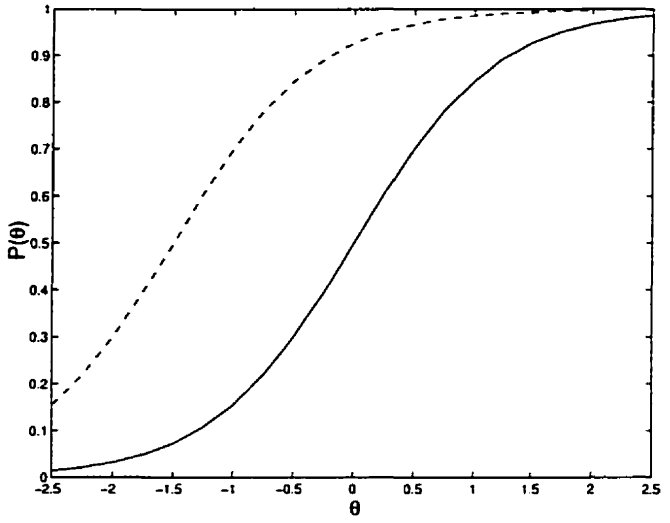

Figure 1.3: Three item response functions varying only with respect to the location parameter $\bar{P}$. For the dashed line $\bar{P} = -1.5$, for the solid line $\bar{P} = 0$, and for the dotted line $\bar{P} = 1.5$.

which the item discriminates among examinees on either side of θ. This can be computed by the derivative,

$$DP(\theta) = \frac{dP}{d\theta}.$$

The higher the value of $DP(\theta)$, the more sharply the item discriminates among examinees at θ. The overall discrimination power of the item, or the item discrimination parameter, can be measured by computing the maximum value of $DP(\theta)$. This maximum usually occurs at a point close to the item difficulty parameter.

Figure 1.4 shows a set of items which differ with respect to the degree of discrimination at $\theta = 0$. For the dashed line $DP(0) = .5$, for the solid line $DP(0) = .9$, and for the dotted line $DP(0) = 1.5$. Consider the range $-0.5 \leq \theta \leq 0.5$. For the dotted line, $P(\theta)$ varies from around .22 to .78 over this range, whereas for the dashed line, $P(\theta)$ varies from .39 to .62. Thus, over this range the dotted line better differentiates among examinees than do the dashed or solid lines. However, although items having high $DP(0)$ discriminate well among a subset of examinees, the trade-off is examinee discrimination over a decreased range of θ. The dotted line only distinguishes among examinees with $-1.5 \leq \theta \leq +1.5$,

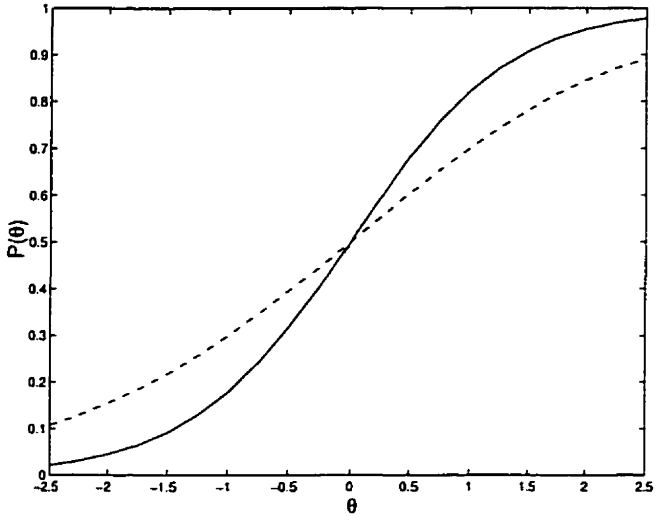

Figure 1.4: Three item response functions varying only with respect to the item discrimination parameter. For the dashed line $DP(0) = .5$, for the solid line $DP(0) = .9$, and for the dotted line $DP(0) = 1.5$.

whereas the dashed line discriminates almost equally over the entire range of θ.

Item and Test Information Functions

In order to construct a useful test, the initial step should be to determine the regions of the latent trait scale for which accurate discrimination among examinees is desirable. For instance, the Graduate Record Examination is designed to identify high ability examinees among all other examinees. Ideally, this test should contain items which discriminate highly among examinees in the middle to high range of the θ scale. It would not be useful to include items which are most discriminating at the lower end of the scale since the purpose of the test does not involve the assessment of low performing examinees.

One aims to construct a test consisting of items which discriminate highly among examinees with latent trait scores in the regions where the test is to be most informative. Since items provide different information about different regions on the latent trait scale, a measure of the amount of information provided by a particular item is useful. Good test items discriminate highly for some range of

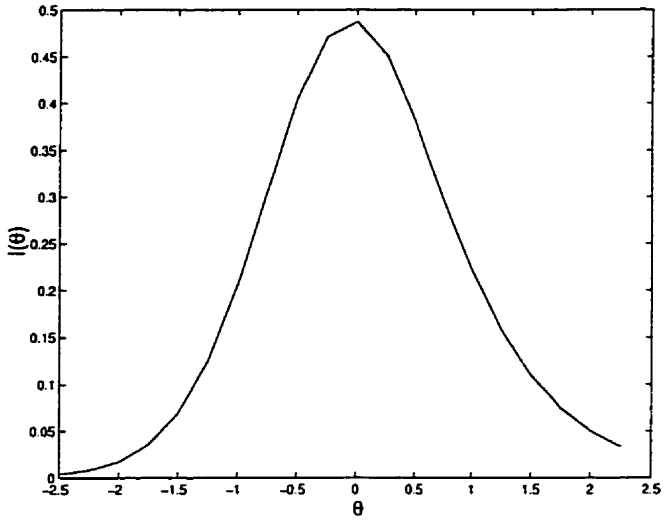

Figure 1.5: The information function for the item response function displayed in Figure 1.1. The horizontal axis represents examinee ability level, and the vertical axis represents, $I(\theta)$, the amount of information provided by the item about ability for examinees at ability θ.

ability values, of which the slope of $P(\theta)$ is an indicator. As the size of the slope of $P(\theta)$ at θ is measured by its derivative $DP(\theta)$, a measure of the amount of information provided by item i about ability for examinees at or near ability θ is given by

$$I_i(\theta) = \frac{[DP_i(\theta)]^2}{P_i(\theta)Q_i(\theta)}.$$

A value of $I_i(\theta)$ can be obtained for every value of θ, and the plot of $I_i(\theta)$ against θ is called the *item information function.* It is clear that information tends to be larger when the first derivative is larger so that $I_i(\theta)$ achieves its maximum at that value of θ where $P(\theta)$ discriminates most highly among examinees. Figure 1.5 displays the information function for the item response function displayed in Figure 1.1. According to the plot, this item is most informative about examinees at or near average ability $\theta = 0$ and provides little information for those examinees with very high or very low ability.

The *test information function* is the sum of the item information functions for

all items on the test, given by

$$I(\theta) = \sum_i I_i(\theta).$$

This measure is an indicator of the amount of information in the entire set of items about an individual examinee's value of ability θ.

1.3.2 Parametric IRT

Parametric item response theory refers to those applications of the theory in which the distribution of the item response function is specified except for the values of a finite number of parameters. *Nonparametric* methods apply in all other instances. As stated above, item response functions are often assumed to take on an ogival shape. Naturally, an example of such a function is the normal ogive. The normal ogive increases monotonically with a left asymptote of zero and an right asymptote of unity. Letting a be the discrimination parameter, b the difficulty parameter and $z = a(\theta - b)$, the equation for the normal ogive item response function is written as

$$P(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-z^2/2} d\theta.$$

Use of the normal ogive in practice has been replaced by the family of logistic models. For these models, the basis for the item response function is the cumulative logistic distribution function, which has the general form

$$P(\theta) = \frac{e^z}{1 + e^z}.$$

Like the normal ogive, the logistic item response function is ogival in shape and increases monotonically. Each of the three logistic models is a variation of this basic form, the models differing with respect to the number of parameters used. In practice, the logistic models are preferable to the normal ogive as the former require simpler computations. The difference between the two types of models is negligible if the basic form of the logistic model is modified as follows:

$$P(\theta) = \frac{e^{1.7z}}{1 + e^{1.7z}}.$$

By negligible it is meant that $P(\theta)$ for the logistic and normal ogive models does not differ by more than .01 over the θ scale.

For the two-parameter logistic (2PL) model,

$$z_i = a_i(\theta - b_i)$$

which gives the model

$$P_i(\theta) = \frac{e^{1.7a_i(\theta - b_i)}}{1 + e^{1.7a_i(\theta - b_i)}}. \tag{1.2}$$

The parameters a_i and b_i are indices of item discrimination and difficulty for item i, respectively. For the family of logistic curves, the item difficulty parameter refers to the point of inflection on the latent trait scale. That is, b denotes the θ value midway between the left asymptote $P(-\infty)$ and the upper asymptote of unity, $P(\theta) = (1 + P(-\infty))/2$. The discrimination parameter a is the slope of the item response function at the point of inflection. The 2PL model can be viewed as a three-parameter model (see below) where the guessing level is set to zero, implying that a low θ value could mean a $P(\theta)$ that is close to zero. Although this is not plausible for multiple-choice or true/false items, it may be the case with essay-type items where examinees are required to supply the complete response to a question instead of selecting the correct answer among a number of alternatives.

The one-parameter logistic model, also referred to as the *Rasch* model, is a special case of the 2PL model where all items have the same discrimination parameter. Since all items are equally discriminating, the subscript i may be dropped and this parameter may be referred to as the constant a. The difficulty parameter b, however, is not a constant, and so items may discriminate at different locations on the θ scale. The one-parameter model is written as

$$P_i(\theta) = \frac{e^{1.7a(\theta - b_i)}}{1 + e^{1.7a(\theta - b_i)}}.$$

Setting $a = 1$ and dropping the constant 1.7, the equation for the Rasch model becomes

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}. \tag{1.3}$$

With this formulation of the model it is more evident that $P(\theta)$ is a function of examinee ability and item difficulty.

The most general of the logistic models is the three-parameter logistic (3PL) model. With a and b defined as above and setting c as guessing level $P(-\infty)$, this model is defined by the equation

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{1.7a_i(\theta - b_i)}}{1 + e^{1.7a_i(\theta - b_i)}}. \tag{1.4}$$

Here, each item is free to vary with respect to the values of the three parameters. The 3PL model is the most commonly used with regards to multiple-choice and true/false question formats.

Estimation procedures

Maximum likelihood estimation (section 3.1.1) procedures are commonly used in item response theory. The computer program LOGIST (Wingersky, Patrick & Lord, 1988) uses a *joint maximum likelihood procedure* to simultaneously estimate the item parameters for all items and the latent trait scores for all examinees. However, there are a number of drawbacks when LOGIST is used with the 3PL model. For one, it is not known whether the estimates yielded by the program are consistent. An estimate is said to be *consistent* if as sample size increases, the value of the estimate approaches the true parameter value. Another limitation is that a large number of examinees are required for accurate estimation of the model parameters. Furthermore, the parameter estimates show large standard errors for the 3PL model, particularly for items with low difficulty. This results from a high positive covariance between the location parameter b and guessing level c (Thissen & Wainer, 1982). For these easy items there is little data with which to estimate c, therefore its standard error is made large. For more difficult items this effect is less severe.

The computer program BILOG (Mislevy & Bock, 1982) uses the *marginal maximum likelihood procedure* to estimate the parameters of the 3PL model (see section 1.6). By marginalizing over the individual ability parameters θ_a for the

N examinees, their estimation is avoided. As an improvement over LOGIST, the estimates produced by BILOG are believed to be consistent. In addition, it is possible to increase sample size N without simultaneously increasing the number of total parameters to be estimated (see section 1.5.1).

1.3.3 Nonparametric IRT

In the parametric approach, the focus is on the estimation of the model parameters. But a specific model presupposes that the test items are sufficiently represented by the features permitted by that model. For example, selection of the 3PL model assumes that a test's set of item response functions are all ogival in shape. Any items characterized by curves which deviate from this shape cannot be accommodated by this model. This includes curves which are either nonmonotonic, have a non-unit right asymptote or have multiple inflection points. Using more parameters in order to obtain greater flexibility would result in the overfitting of some items which require only a small number of parameters to describe them adequately, thus leading to poor estimators of the parameters actually needed. Furthermore, current estimation procedures tend to produce parameter estimates having strong positive covariance. Also, large amounts of data are required to estimate the 3PL model well, particularly for easy items. Parameter estimates for these items rely heavily upon data from low ability examinees. For a sample size of 500 and assuming that ability level is normally distributed, this leaves less than 34 pieces of data with which to estimate the model parameters in the region $\theta \leq -1.5$. In addition, programs such as LOGIST and BILOG are also computationally demanding.

The inspiration for *nonparametric* estimation within item response theory is the direct estimation of the item response functions. *Nonparametric* does not imply the absence of parameters to be estimated. (In fact, there are an arbitrary number of parameters and hence an arbitrary amount of flexibility can be achieved.) Instead, the term nonparametric suggests that the emphasis is on the

direct estimation of the curves and not on the estimation of the curve parameters. Hence, the problems associated with item parameter estimation are avoided. For example, direct estimation of the functions obviates any presumptions about the shape of the curves. Furthermore, the shape features of the functions, such as item discrimination and item difficulty, can still be described even though these values are not directly estimated.

1.4 Functional Data Analysis

With regards to test data, for each item there is available a set of responses from various examinees who vary in terms of their ability level. The set of responses to each item can be viewed as a function of examinee ability level, and functional data analytic techniques may be applied to derive estimates of these functions. *Functional data analysis* (FDA) can be defined as a set of techniques for the description and analysis of data where the observations are functions (Ramsay & Silverman, 1997). Aside from providing a set of useful techniques, FDA presents a conceptual framework with which to approach the current problem: the unit of interest is not the string of numbers representing examinee responses, but rather the *functions* $P(\theta)$.

There are two primary ways in which functional data analysis will play a leading role in this thesis. The first involves the assumption that the individual responses to an item reflect a continuum of ability level. The raw data are discrete, but are to be viewed as functions. Thus, the first step is to use FDA to represent the probability of a correct response, $P(u_{ai} = 1)$, as a function of θ. If the discrete data are error-free, they can be converted to a function using interpolation. However, in the case of test data, the measurement of some observational noise is presumed. Furthermore, if there is an interest in computing the derivatives of these functions, they must be represented by a smooth curve. Smooth functions can be derived using basis function methods, described in Chapter 2. Another FDA approach which will play a role here involves roughness penalty smoothing

(Chapter 3), which offers control of the smoothing by putting a limit on the total curvature of the estimate.

1.5 The EM Algorithm

This section will provide a brief introduction to the EM algorithm. A more detailed description of the algorithm will appear in Chapter 3.

1.5.1 EM and item response theory

Consider the three-parameter logistic model in (1.4) for a test consisting of 25 items administered to 500 examinees. Assuming that estimates of both the item parameters and examinees abilities are desired, there are 575 parameters to be estimated, three for each of the items and one for each examinee. The $N \times n$ matrix of observed responses, $\mathbf{U}$, then consists of 12,500 independent observations. In this context, sample size is considered to be not the total number of subjects, but rather the number of observations available to estimate each parameter. In other words, the focus is not on sample size defined as the number of examinees but instead a *data-to-parameter ratio* where a ratio of at least 50 or 100 is ideal. In this instance, there are 12,500 sample elements and 575 parameters to be estimated, which amounts to $12,500/575 \approx 22$ observations, a small sample size.

In certain circumstances only estimates of the item parameters, and not examinee abilities, are desired. The item parameters are referred to as *structural* parameters - these are the parameters whose estimates are of most significance. The ability parameters are referred to as *nuisance* parameters. They are not of primary interest to the investigator concerned with estimating item response functions, but due to their unobservable nature they must be estimated alongside the structural parameters. The number of structural parameters remains constant irrespective of sample size N, and using larger samples would seemingly increase the data-to-parameter ratio. But the number of ability parameters increases in proportion to N, so an increase in N to allow for better item parameter esti-

mation is accompanied by an increase in the number of ability parameters to be estimated. Thus, any attempt to improve item parameter estimates by an increase in N will unavoidably involve an increase in the total number of model parameters. In a situation where one could escape estimation of the examinee ability levels, the data-to-parameter ratio reduces to $12,500/75 \approx 167$, a rather impressive improvement in sample size. Furthermore, in addition to increasing the relative sample size, the chore of estimating parameters that are of no practical interest is avoided.

The *EM algorithm*, defined by Dempster, Laird and Rubin (1977), is an iterative procedure for finding maximum likelihood estimates in the presence of unobserved random variables in probability models. It was first applied within item response theory by Bock and Aitkin (1981) who, by working with marginal likelihoods, eliminated the estimation of the unobservable ability parameters in estimating item response functions (see section 1.6 for details). The algorithm works by alternating between two phases of analysis, the E (for Expectation) phase and the M (for Maximization) phase. A comprehensive review of the evolution of the EM algorithm for item parameter estimation can be found in Harwell, Baker and Zwarts (1988).

1.5.2 The expectation phase

In the E-phase, the marginal likelihoods are estimated for each examinee for a fixed set of item parameters ψ. Assuming local independence and letting $P(u_{ai})$ represent the probability of a correct response to item i by examinee a, the conditional likelihood of observing a particular response sequence can be written as

$$L(u_a|\theta_a;\psi) = \prod_{i=1}^{n} P(u_{ai}|\theta_a;\psi_i), \tag{1.5}$$

The latent variable θ is unknown, and in order to accommodate for this, the marginal, or average, likelihood of each observed response sequence is computed. The E-phase of the EM algorithm consists of taking the expectation of the con-

ditional likelihood,

$$ML(u_a|\psi) = E[L(u_a|\theta;\psi)] = \int L(u_a|\theta;\psi)g(\theta)d\theta. \tag{1.6}$$

Thus, through marginalization the task of estimating the nuisance parameters has been eliminated.

The final step in the E-phase is to compute the marginal likelihood for all examinees, or the grand marginal likelihood, ML(**U**). Assuming that examinees respond independently of one another, the grand marginal likelihood can be defined as

$$ML(\mathbf{U}|\psi) = \prod_{a=1}^{N} ML(u_a|\psi). \tag{1.7}$$

For simplification, the dependence of ML on ψ will be dropped from the notation, although it is always implied.

1.5.3 The maximization phase

In the M-phase of the EM algorithm, the grand marginal likelihood computed in the E-phase is maximized with respect to the item parameters. In other words, the values chosen as estimates of the item parameters are those that maximize the value of the grand marginal likelihood function, $ML(\mathbf{U})$. The grand marginal likelihood, rather than the individual marginal likelihoods, is maximized since the set of item parameters for any particular item affects the $ML(u_a)$'s for all examinees answering that item.

Once these parameter estimates are obtained, the E-phase is revisited and the marginal likelihoods for each examinee are recomputed using the parameter estimates from the previous M-phase. Thus, the EM algorithm is an iterative process of marginalizing over the likelihood function with respect to the nuisance parameters (E-phase), and then maximizing the function with respect to the structural parameters (M-phase). The algorithm iterates between these two stages until some convergence criterion is reached, usually when the change in parameter estimates is negligible, or the change in $ML(\mathbf{U})$ is small.

1.6 Estimation of the 3PL Model with an EM Algorithm

This section will present the application of the EM algorithm to the estimation of the 3PL model item parameters as described by Bock and Atikin (1981) and more recently by Bock (1989). First, recall that under the assumption of local independence, the joint distribution of the set of item responses for the a^{th} examinee can be written as

$$P(u_a|\theta_a, \psi) = \prod_{i=1}^{n} P_i(\theta_a)^{u_{ai}} Q_i(\theta_a)^{1-u_{ai}}.$$

This is the probability of response vector u_{ai} conditional on a known value of θ and the item parameters. The likelihood of observing the matrix of the set of responses to all items from all examinees, denoted $\mathbf{U}$, represents the likelihood function and can be written

$$L(u_a) = \prod_{i=1}^{n} \prod_{a=1}^{N} P_i(\theta_a)^{u_{ai}} Q_i(\theta_a)^{1-u_{ai}}.$$

Under maximum likelihood estimation, the parameter estimates are those values of a, b and c which maximize the value of L. These estimates are found from the roots of the likelihood equations, which are obtained by setting the first derivatives of the likelihood equal to zero. For convenience, the log of the likelihood function is used:

$$\log L = \sum_{i=1}^{n} \sum_{a=1}^{N} [u_{ai} \log P_i(\theta_a) + (1 - u_{ai} \log Q_i(\theta_a)].$$

This leads to the system

$$\frac{\partial}{\partial a_i}(\log L) = 0, \quad \frac{\partial}{\partial b_i}(\log L) = 0, \quad \frac{\partial}{\partial c_i}(\log L) = 0.$$

If the θ_a are known, the parameters for the i^{th} item are estimated simultaneously using the above system of equations. With LOGIST, the initial θ values are treated as known while solving for the item parameters, then the process is reversed with the item parameters treated as known and the θ values estimated.

Among the many shortcomings associated with LOGIST (see section 1.3.2), the most significant is that the examinee abilities are nuisance parameters. The Bock and Aitkin (1981) solution to this uses an EM algorithm to replace $L(u_a)$ by its average over θ, or the marginal likelihood, $ML(u_a)$. First, some distribution $g(\theta)$ is assumed for the examinee ability parameters. Then the marginal likelihood of response vector u_a given item parameters ψ is the average of $L(u_a)$ over the prior distribution of abilities:

$$\begin{aligned} ML(u_a|\psi) &= E[L(u_a|\theta_a)] \\ &= \int_{-\infty}^{\infty} L(u_a|\theta)g(\theta)d\theta. \end{aligned}$$

Thus, the ability parameters have been eliminated by averaging them out, or marginalizing over them.

The grand marginal likelihood is now

$$ML(\mathbf{U}) = \prod_{a=1}^{N} ML(u_a)$$

and the log marginal likelihood is

$$\log ML(\mathbf{U}) = \sum_{a=1}^{N} \log ML(u_a).$$

The parameter estimates are chosen to be those values of a, b and c that maximize the log marginal likelihood.

The two general steps of the p^{th} cycle of the EM algorithm are:

1. E-step: compute the expectation of the likelihood, $E[\log L(u_a|\theta), \psi^p]$.

2. M-step: choose ψ^{p+1} such that the log marginal likelihood is maximized.

The process is repeated until some convergence criterion is satisfied. This procedure is used in BILOG, and overcomes several of the difficulties encountered by LOGIST. BILOG, however, produces parametric estimates of the item response functions, whereas nonparametric estimates allow for more flexibility (see section 1.3.3). In this thesis, the EM algorithm is applied to the nonparametric estimation of item response functions.

Chapter 2

Estimating Nonparametric Item Response Functions

2.1 Basis Functions

Smoothness is a desirable characteristic of an estimated item response function. This is essential if the item or test information function is to be computed, as it depends on the item first derivative functions. A common smoothing procedure is to represent the function to be estimated as a linear combination of a set of K linearly independent *basis functions* ϕ_k, with weight coefficients c_k:

$$P(\theta) = \sum_{k=1}^{K} c_k \phi_k(\theta). \tag{2.1}$$

The right side of (2.1) is called a *basis function expansion* for function $P(\theta)$. The degree of smoothness of function $P(\theta)$ is determined by the number K of basis functions. As K increases, the fit of the data improves but the estimate becomes less smooth. On the other hand, using a small K will yield a smooth function that does not fit the data closely. In section 3.3, a compromise between fit and smoothness will be discussed.

There are a number of options for the type of basis functions. The ideal situation is one in which a good approximation is obtained with a relatively small number of basis functions K. Preferably, they should possess features resembling those known to belong to the functions being estimated. Classic bases include the polynomials and the B-spline bases, to be covered in sections 2.1.1 and 2.1.2,

respectively. Another classic basis function expansion is provided by the Fourier series,

$$x(t) = c_0 + c_1 \sin \omega t + c_2 \cos \omega t + c_3 \sin 2\omega t + c_4 \cos 2\omega t + \ldots,$$

defined by the basis $\phi_0(t) = 1$, $\phi_{2r-1}(t) = \sin r\omega t$ and $\phi_{2r}(t) = \cos r\omega t$. This basis is periodic, with the period $2\pi/\omega$ determined by parameter ω.

2.1.1 Polynomial Bases

One possibility is to represent the function as a linear combination of the basis functions

$$\phi_k(x) = (x - \omega)^k, \qquad k = 0, \ldots, K, \tag{2.2}$$

known as the *monomial bases*. Parameter ω is a shift parameter. Any polynomial of degree $K - 1$ or less can be expressed as a linear combination of K fixed linearly independent polynomials, of which the monomials are a classic example. However, in order for the polynomials to fully capture local behavior a large value for K is needed. Even so, in this case the data may fit well in the center but is less satisfactory at the extremes, since the polynomial functions themselves exhibit wild behavior at the extremes (Ramsay & Silverman, 1997). Furthermore, although the derivatives of polynomial functions are easy to compute, they are rarely reasonable estimators of the true derivative. This is due to the rapid localized oscillation common to high-order polynomial fits (Ramsay & Silverman, 1997). Figure 2.1 displays a set of monomial basis functions for $K = 2$.

2.1.2 Regression Spline Bases

An alternative to the polynomial bases are polynomial splines, which offer greater flexibility and have the capacity to capture changing local behavior. In order to derive these functions, first the range $[a, b]$ of the function to be estimated is partitioned into n subintervals $[\tau_{i-1}, \tau_i]$, $1 \leq i \leq n$, where

$$a < \tau_0 < \ldots < \tau_n < b.$$

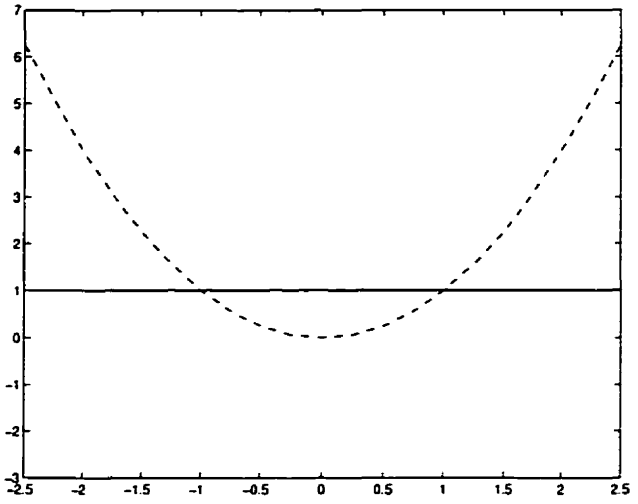

Figure 2.1: A set of monomial basis functions for $K = 2$ and $\omega = 0$. The solid line represents the basis function for which $k = 0$, the dotted line for $k = 1$, and the dashed line for $k = 2$.

The τ_i are referred to as *knots*, and there are $n + 1$ of them. Excluding the *boundary knots* τ_0 and τ_n, the remaining set of knots $\tau_1, \ldots, \tau_{n-1}$ are referred to as the *interior knots*. Now function f may be approximated by a polynomial *spline* S_n, where S_n is formed by connecting adjacent pairs of points (τ_i, y_i), $0 \leq i \leq n$, by a polynomial of degree at most $k \geq 1$, and forcing these polynomials to join smoothly at these knots. As a special case $y_i = f(\tau_i)$ may be chosen, where S_n interpolates the function, although a better estimate can be achieved by relaxing this restriction.

In the simplest case S_n is formed by connecting the (τ_i, y_i) with straight line segments, that is, polynomials of degree 1. In this case, spline S_n is referred to as a *first degree spline* (see Figure 2.2). In general, a spline S_n of degree k in $[a, b]$ is constructed by joining the intervals $[\tau_{i-1}, \tau_i]$, each of which contains a polynomial of degree at most $k \geq 1$. In order to give the spline a certain degree of smoothness it is further required that S_n adhere to certain continuity conditions at the interior knots. Specifically, S_n must have at least $k - 1$ derivatives which are continuous on $[a, b]$. For example, any first degree spline is a continuous function although

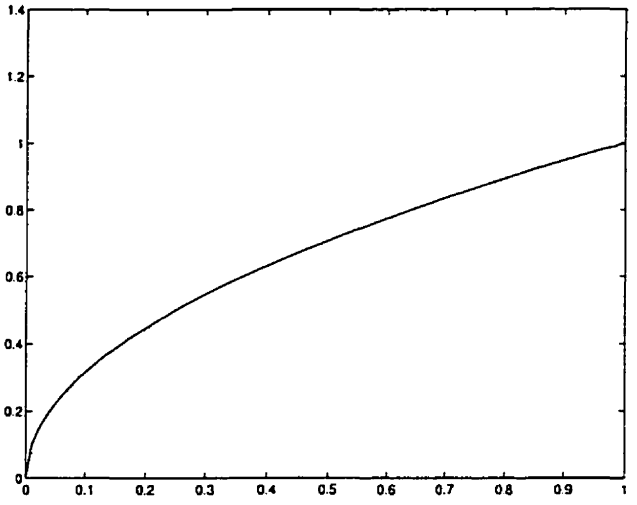

Figure 2.2: The solid line is the curve $y = \sqrt{x}$. The dotted line is an estimate of this curve by a first degree spline with a single interior knot.

it may have discontinuities in its first derivative, these discontinuities occurring at the knots. A *quadratic spline* ($k = 2$) will have a continuous first derivative, a *cubic spline* ($k = 3$) continuous first and second derivatives, and so on. The cubic splines, or *piecewise cubic polynomials*, are a popular choice for basis functions.

There is some question as to the number and position of the knots. For any given set of knots, the spline is computed by multiple regression on an appropriate set of basis vectors. By allowing more knots the spline becomes more flexible, although with too large a K one runs the risk of overfitting the data resulting in poor generalizability of the estimate. Another concern is the choice of basis functions for representing the splines for a given set of knots.

One possible choice for a basis for first degree splines is the following. To facilitate this discussion it is advantageous to extend the sequence of knots to an

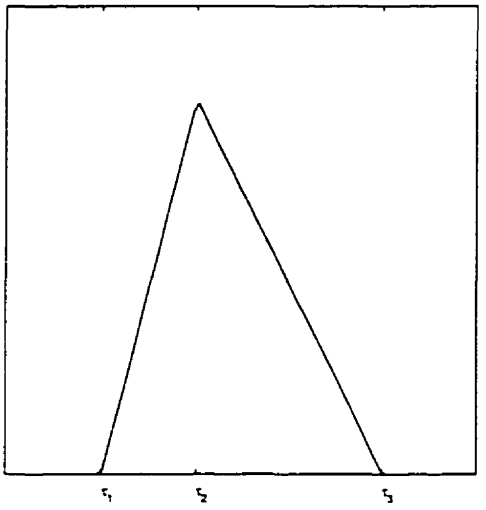

Figure 2.3: An example of a linear B-spline basis function. The dashed vertical lines represent the interior knots. The function has value 0 at all knots except τ_2.

infinite sequence $\ldots \tau_{-1}, \tau_0, \ldots, \tau_n, \tau_{n+1}, \ldots$. Then, for $-\infty < i < \infty$, define

$$B_i^1(x) = \begin{cases} 0, & x \leq \tau_i \\ \frac{x-\tau_i}{\tau_{i+1}-\tau_i}, & \tau_i < x \leq \tau_{i+1} \\ \frac{\tau_{i+2}-x}{\tau_{i+2}-\tau_{i+1}}, & \tau_{i+1} < x \leq \tau_{i+2} \\ 0, & \tau_{i+2} < x. \end{cases} \tag{2.3}$$

Notice that $B_i^1(\tau_{i+1}) = 1$ and $B_i^1(x) = 0$ at all other knots (see Figure 2.3). It is clear that the functions B_i^1 are linearly independent and hence form a basis for first degree splines. The spline can be expressed as

$$S_n(x) = \sum_{i=0}^{n} y_i B_{i-1}^1, \qquad \tau_0 \leq x \leq \tau_n$$

Later, it will be shown that B-splines of degree k, B_i^k, are a generalization of the first degree splines and hence form a basis for splines of degree k.

Truncated power basis

The simplest way to represent polynomial splines is as the monomial basis in (2.2) supplemented by a linear combination of the *truncated powers*. For coefficient

weights c_i and d_j, these splines are written as

$$S_n(x) = \sum_{i=0}^{k} c_i x^i + \sum_{j=1}^{n-1} d_j (x - \tau_j)_+^k. \tag{2.4}$$

For illustrative purposes, this polynomial spline of degree k can be constructed, for some fixed real number τ and working with one subinterval at a time, by augmenting the monomial basis (2.2) with the *truncated powers*

$$(x-\tau)_+^k = \begin{cases} (x-\tau)^k, & x \geq \tau \\ 0, & x < \tau \end{cases} \tag{2.5}$$

of degree $k \geq 0$. Notice that (2.5) is a spline of degree k with a single knot at $x = \tau$.

First, the spline of degree k on $[\tau_0, \tau_1]$ is written as

$$S_n(x) = \sum_{i=0}^{k} c_i x^i. \tag{2.6}$$

Now any polynomial function can be written as

$$\sum_{i=0}^{k} a_i (x - \tau_1)^i,$$

which may be adapted to

$$\sum_{i=0}^{k} a_i (x - \tau_1)_+^i, \tag{2.7}$$

in order to derive a function which is zero for $x < \tau_1$. Thus, to extend the spline S_n from $[\tau_0, \tau_1]$ to $[\tau_0, \tau_2]$ without disrupting its representation on $[\tau_0, \tau_1]$, the function in (2.7) can be added to the right side of (2.6). However, there is a smoothness condition to be satisfied at τ_1. Specifically, the continuity of S_n or its first $k-1$ derivatives must not be disrupted at this knot. To account for this, $a_0, \ldots, a_{k-1}$ must all be set to zero, and relabeling a_k as d_1, the polynomial spline over interval $[\tau_0, \tau_2]$ can be written as

$$S_n(x) = \sum_{i=0}^{k} c_i x^i + d_1 (x - \tau_1)_+^k. \tag{2.8}$$

To further extend S_n across all n intervals, a suitable truncated power is added for each interval while not disturbing the representation of S_n in the previous

intervals. This results in a polynomial spline valid on the whole interval $[\tau_0, \tau_n]$, which is written as in (2.4). Note that the total number of basis functions is $(k+1)+(n-1) = k+n$, the order of the piecewise polynomial plus the number of interior knots.

Although (2.4) is of theoretical interest, it is not used to produce an estimate for the function of interest. For evaluation purposes it is preferable to express the spline in terms of the B-spline basis functions, given that if there is more than a small number of knots, the truncated power basis tends to produce nearly singular cross-product matrices (Ramsay & Silverman, 1997).

B-spline basis

It was stated above that the B_i^1 form a basis for first degree splines. In this section it will be shown that a generalization of these, the B_i^k, form a basis for splines of degree k. First, it is necessary to express the B_i^1 in terms of simpler functions, specifically, the B_i^0, defined as

$$B_i^0(x) = \begin{cases} 1, & \tau_i < x \leq \tau_{i+1}, \\ 0, & \text{otherwise.} \end{cases}$$

The B_i^0 are piecewise constant functions, and they constitute a basis for the set of all piecewise constant functions.

It should be noted that any function which takes the value y_i on interval $\tau_i < x \leq \tau_{i+1}$ for $0 \leq i \leq n-1$ and the value 0 elsewhere can be written as

$$\sum_{i=0}^{n-1} y_i B_i^0(x).$$

Now the functions B_i^1 can be defined in terms of the B_i^0 by taking

$$B_i^1(x) = \left(\frac{x-\tau_i}{\tau_{i+1}-\tau_i}\right) B_i^0 + \left(\frac{\tau_{i+2}-x}{\tau_{i+2}-\tau_{i+1}}\right) B_{i+1}^0. \tag{2.9}$$

The validity of this equation can be verified by comparing the right side of (2.9) to the definition of B_i^1 given in (2.3). Functions B_i^0 and B_i^1 are referred to as

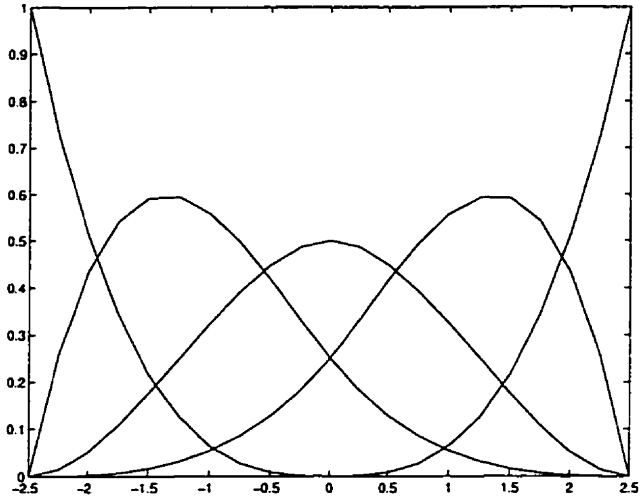

Figure 2.4: A set of cubic B-spline functions. Each basis function is nonzero over at most four adjacent intervals.

B-splines of degree 0 and 1, respectively. A B-spline of degree k can be defined recursively by

$$B_i^k(x) = \left(\frac{x - \tau_i}{\tau_{i+k} - \tau_i}\right) B_i^{k-1} + \left(\frac{\tau_{i+k+1} - x}{\tau_{i+k+1} - \tau_{i+1}}\right) B_{i+1}^{k-1}, \tag{2.10}$$

for each i and for $k = 1, 2, \ldots$. Function B_i^k is called a B-spline of degree k. For justification that B_i^k is indeed a spline, see Phillips and Taylor (1996).

The advantage of the B-splines over the truncated power bases in evaluating a spline S_n is that the B-splines have *compact support*, meaning that the function values are zero everywhere except over a finite interval. This implies that the resulting regression matrix is banded, overcoming the problem of singular cross-product matrices often encountered when using the truncated power bases. For splines of degree $k = 3$, the B-splines are themselves piecewise cubics with support on the interval $[\tau_{k-2}, \tau_{k+2}]$ and shorter support on the ends (see Figure 2.4). Thus, these $B_i(x)$ are nonzero over at most four adjacent intervals.

2.2 The Logistic Reformulation of $P(\theta)$

The model (2.1) has a major structural defect. Probabilities must fall between 0 and 1, whereas linear functions can take values over the entire real line, depending on the size of the coefficients c_k. Thus, unless restrictions are imposed on $c_1, \ldots, c_K$, the model (2.1) will yield $-\infty < P(\theta) < \infty$ so that to express $P(\theta)$ as such a linear combination would be inconsistent with the laws of probability. Model (2.1) can be valid over a finite range of θ values for which $0 \leq P(\theta) \leq 1$. However, using ordinary least squares to fit the model is problematic as the conditions making least squares estimators optimal are not satisfied. For one, the variance $V(u)$ of binary response variable u, $P(\theta)[1 - P(\theta)]$, is not constant, but rather depends on θ through its influence on P. As P moves towards zero or unity, $V(u)$ moves towards zero. Furthermore, as it is a binary variable, the variance of u cannot be assumed to be normal, so that the sampling distributions for the ordinary estimators are not applicable (Agresti, 1990).

The problem can be avoided by using some transformation $h[P(\theta)]$ which maps the unit interval $(0, 1)$ onto the real line $(-\infty, \infty)$ so that

$$h[P(\theta)] = W'(\theta) = \sum_{k=1}^{K} c_k \phi_k(\theta).$$

In order to derive the appropriate transformation or *link function* $h[P(\theta)]$, it is first realized that as with any function bounded by 0 and 1, $P(\theta)$ can be reformulated as

$$P(\theta) = \frac{\exp(\sum_{k=1}^{K} c_k \phi_k(\theta))}{1 + \exp(\sum_{k=1}^{K} c_k \phi_k(\theta))}. \tag{2.11}$$

With this reformulation, the condition $0 < P(\theta) < 1$ is satisfied. Although the possibilities of $P(\theta) = 1$ and $P(\theta) = 0$ are lost, this is often considered to be of no practical consequence.

Now for model (2.11) the odds of making response 1 are

$$\frac{P(\theta)}{1 - P(\theta)} = \exp(\sum_{k=1}^{K} c_k \phi_k(\theta)),$$

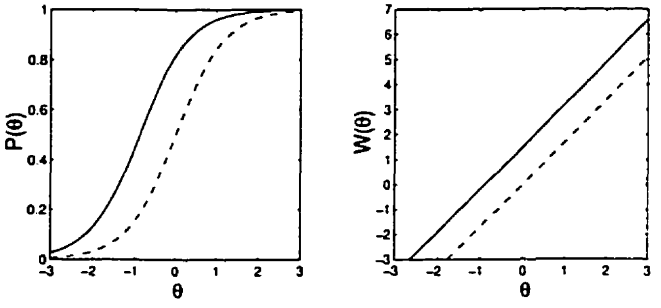

Figure 2.5: In the left panel are three Rasch item response functions $P(\theta)$ with varying values for the location parameters. In the right panel are the corresponding $W(\theta)$, which all have a slope of unity but vary with respect to the y-intercepts.

and the log-odds has the linear relationship

$$\log\left[\frac{P(\theta)}{1-P(\theta)}\right] = \sum_{k=1}^{K} c_k \phi_k(\theta) = W(\theta).$$

Thus the log-odds transformation, or the *logit*, is the appropriate link function.

The logistic reformulation of item response function $P(\theta)$ such that function $W(\theta)$ may instead be estimated greatly simplifies the task. This results primarily from $W(\theta)$ being an unconstrained function, which makes it an ideal candidate for a basis function expansion. For example, the logistic reformulation of the item response function for the 2PL model in (1.2) amounts to

$$W(\theta) = 1.7a(\theta - b),$$

and that for the Rasch model in (1.3) is

$$W(\theta) = \theta - b.$$

Thus, the $W(\theta)$'s for Rasch item i will be a straight line with a slope of unity and a y-intercept equal to b_i. For the 2PL model, the slope of the logit transformation

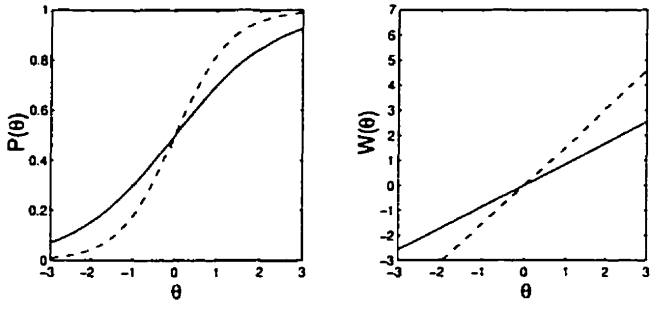

Figure 2.6: In the left panel are three 2PL item response functions $P(\theta)$ with identical values for the location parameters but varying values for the discrimination parameter. In the right panel are the corresponding $W(\theta)$, which vary in terms of both the slopes and y-intercepts.

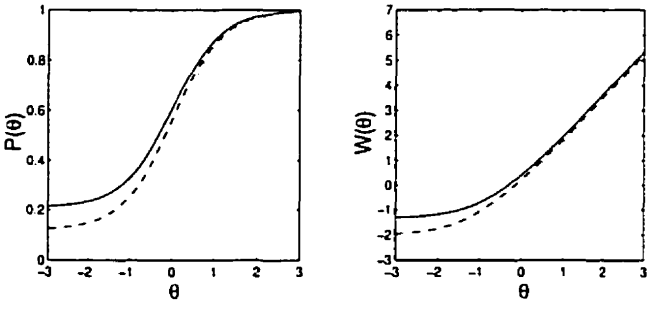

Figure 2.7: In the left panel are three item response functions $P(\theta)$ with identical values for the discrimination and location parameters, but varying values for the guessing levels. In the right panel are the corresponding $W(\theta)$, which all approach an upper right asymptote at a 45 degree angle, but vary with respect to the value of the lower left asymptote.

for item i is $1.7a_i$ and the y-intercept is $1.7a_ib_i$. The $W(\theta)$'s for the 3PL model are nonlinear functions which go to a lower asymptote on the left, and approach an upper asymptote at a 45 degree angle on the right. Figures 2.5, 2.6 and 2.7 show sets of item response functions $P(\theta)$ and the corresponding log odds transformations $W(\theta)$ for the Rasch, 2PL and 3PL models, respectively.

2.3 A Basis Function Expansion for $P(\theta)$

In order to obtain estimates of $P(\theta)$ that are intrinsically smooth, a basis function expansion for $W(\theta)$ is considered. A set of K basis functions are chosen and $W(\theta)$ is expressed as a weighted linear combination of these functions,

$$W(\theta) = \sum_{k=1}^{K} c_k \phi_k(\theta).$$

The K basis functions $\phi_k(\theta)$ are chosen to be the B-spline basis functions. Choosing a larger K will result in a more flexible and hence less smooth curve, whereas a smaller value for K will produce a smoother curve with less flexibility.

Chapter 3

The EM Algorithm

3.1 Preliminaries

3.1.1 Maximum likelihood estimation

Likelihood, denoted L, is proportional to the probability of the observed data given a proposed model. Assume some model, $P_i(\theta|\psi_i)$, of examinee performance on dichotomously scored item i given a set of item parameters ψ_i and ability level θ. Let u_a denote the response vector of examinee a to a set of n items. Assuming that the elements of u_a are independently distributed conditional on θ and with θ_a denoting the ability level of examinee a, the likelihood of response vector u_a is

$$L(u_a) = \prod_{i=1}^{n} P(u_{ai} = 1|\theta_a; \psi_i).$$

Similarly, since all N examinees are assumed to behave independently, the likelihood of the entire observed data matrix $\mathbf{U}$ can be computed as

$$L(\mathbf{U}) = \prod_{a=1}^{N} L(u_a) = \prod_{a=1}^{N} \prod_{i=1}^{n} P(\theta_a|\psi_i).$$

Now, should a specific set of values be considered for the item parameters, and L computed for the observed data and these particular parameter values, what results is the likelihood of the observed matrix $\mathbf{U}$ given item parameters ψ. *Maximum likelihood estimation* is a method of estimation which chooses as estimates those parameter values that maximize the value of the likelihood function. These

estimates may be obtained by setting the first derivative of $L(\mathbf{U})$ to zero and solving the equation with respect to the individual item parameters.

Computing the derivatives of products can be an enormous task. Fortunately, both the likelihood function and the log likelihood function,

$$\log L(\mathbf{U}) = \sum_{a=1}^{N} \log L(u_a),$$

are maximized for the same parameter values, and the computational burden is greatly alleviated in working with the derivatives of sums as opposed to products in maximizing $\log L(\mathbf{U})$.

3.1.2 Marginal MLE

Section 1.5.1 stated that a data-to-parameter ratio of at least 50 or 100 would constitute a sufficient sample size for obtaining estimates of item parameters. Although increasing sample size N will increase the data-to-parameter ratio, it will be accompanied by an increase in the number of ability parameters to be estimated. If it were possible to eliminate the estimation of the examinee ability parameters, not only would the relative sample size increase, but the task of estimating parameters which are of no practical interest would also be avoided.

This may be accomplished by computing the likelihood not for each θ value, but the likelihood obtained by averaging over all possible values of θ. This requires the assumption that examinees represent a random sample from a population where ability is distributed according to some known density function $g(\theta)$. Since the family of item response functions is highly flexible for nonparametric methods, the choice for this distribution is arbitrary. By tradition, the standard normal distribution has been employed, as ability level is considered to have a normal distribution. Also, since $g(\theta)$ can be chosen at will, it is preferable to select a distribution whose mathematical properties make it convenient for computational purposes.

One is then working with the *marginal likelihood*, denoted ML. Where $L(u_a|\theta)$

is the likelihood of a response vector viewed as a function of θ, called the *conditional likelihood*, the marginal likelihood of response vector u_a is defined as

$$ML(u_a) = E[L(u_a)] = \int L(u_a|\theta)g(\theta)d\theta = \int \prod_{i=1}^{n} P_i^{u_{ai}}(\theta)Q_i^{1-u_{ai}}(\theta)d\theta. \qquad (3.1)$$

The marginal likelihood $ML(u_a)$ is the average of $L(u_a)$ across the values of θ, so that for each examinee $ML(u_a)$ is a single number. Thus, by averaging over θ the task of estimating the ability level for each examinee has been eliminated. In other words, the estimation of nuisance parameters can be avoided by taking the expectation of the likelihood function with respect to these parameters.

3.2 Using the Algorithm

3.2.1 The E-phase

The ultimate objective of the E-phase is the computation of the marginal likelihood for all examinees, called the *grand marginal likelihood*. Assuming that examinees respond to items independently of one another, the grand marginal likelihood is

$$ML(\mathbf{U}) = \prod_{a=1}^{N} ML(u_a).$$

In order to compute this quantity, it is first necessary to evaluate the integral in (3.1). To do so requires the application of some numerical method for approximating an integral. More specifically, some *quadrature rule* must be invoked so as to replace the integral in (3.1) by a weighted sum. With ω representing some fixed weight function, this sum is of the form

$$\int \omega(x)f(x)dx \approx \sum_{q=1}^{Q} w_q f(x_q).$$

Quadrature weights w_q and quadrature points x_q can be selected in a number of ways, with the intention of finding a satisfactory approximation to the integral. The details of selecting a quadrature rule specific to the present case are deferred to section 3.2.4. With the application of the quadrature rule, the integral in (3.1)

translates to

$$ML(u_a) \approx \sum_{q=1}^{Q} w_q L' u_a|\theta_q).$$

For the purpose of computational simplification, the log grand marginal likelihood is preferable to $ML(\mathbf{U})$, obtained by summing the log individual likelihoods,

$$\log ML(\mathbf{U}) = \sum_{a=1}^{N} \log ML(u_a).$$

Invoking the quadrature rule yields

$$\begin{aligned}
\log ML(\mathbf{U}) &\approx \sum_{a=1}^{N} \log \sum_{q=1}^{Q} w_q L(u_a|\theta_q) \\
&= \sum_{a=1}^{N} \log \sum_{q=1}^{Q} w_q \prod_{i=1}^{n} P_i^{u_{ai}}(\theta) Q_i^{1-u_{ai}}(\theta) \\
&= \sum_{a=1}^{N} \ln \sum_{q=1}^{Q} w_q \exp[\sum_{i=1}^{n} u_{ai} \ln P_i(\theta_q) + (1-u_{ai}) \ln Q_i(\theta_q)],
\end{aligned}$$

the evaluation of which completes the E-phase.

3.2.2 The M-phase

In the M-phase, the grand marginal likelihood computed in the E-phase is maximized with respect to the set of item parameters ψ_i. There is a reason for working with the grand marginal likelihood as opposed to the individual marginal likelihoods $ML(u_a)$. Any set of item parameters ψ_i affects $ML(u_a)$ for all examinees. Thus, in order to obtain reasonable estimates of the item parameters ψ_i, information is required on all examinees, and this is provided by the grand marginal likelihood $ML(\mathbf{U})$.

The item parameters are estimated using a maximum likelihood estimation procedure, so that those values which maximize the value of $ML(\mathbf{U})$ are chosen as estimates of the item parameters. In practice this amounts to the same thing as maximizing the mathematically simpler $\log ML(\mathbf{U})$ function over all n parameter values. For a particular item j, the maximizing values are those for which the slope of $\log ML(\mathbf{U})$ will be zero, that is, for some fixed item index j the solution

to the equation

$$\frac{\partial \log ML}{\partial \psi_j} = 0 \tag{3.2}$$

is sought.

Now

$$\frac{\partial \log ML}{\partial \psi_j} = \sum_a \frac{1}{ML(u_a)} \sum_q w_q \frac{\partial L(u_a|\theta_q)}{\partial \psi_j}. \tag{3.3}$$

Rearranging the partial derivative $\partial L(u_a|\theta_q)/\partial \psi_j$ on the right side and recognizing that the parameter vector ψ_j affects only the jth item, all items except the jth can be taken outside of the product in (3.1) to obtain

$$\frac{\partial L(u_a|\theta_q)}{\partial \psi_j} = [\prod_{i \neq j}^{n} P_i^{u_{ai}}(\theta_q) Q_i^{1-u_{ai}}(\theta_q)] \frac{\partial P_j^{u_{aj}}(\theta_q) Q_j^{1-u_{aj}}(\theta_q)}{\partial \psi_j}.$$

Now the product taken over all items can be recovered by multiplying both sides by

$$1 = \frac{P_j^{u_{aj}}(\theta_q) Q_j^{1-u_{aj}}(\theta_q)}{P_j^{u_{aj}}(\theta_q) Q_j^{1-u_{aj}}(\theta_q)}$$

and then simplifying to get

$$\begin{aligned} \frac{\partial \log ML}{\partial \psi_j} &= [\prod_{i=1}^{n} P_i^{u_{ai}}(\theta_q) Q_i^{1-u_{ai}}(\theta_q)] \frac{\partial [P_j^{u_{aj}}(\theta_q) Q_j^{1-u_{aj}}(\theta_q)]}{\partial \psi_j} / [P_j^{u_{aj}}(\theta_q) Q_j^{1-u_{aj}}(\theta_q)] \\ &= L(u_a|\theta_q) \frac{\partial [P_j^{u_{aj}}(\theta_q) Q_j^{1-u_{aj}}(\theta_q)]}{\partial \psi_j} / [P_j^{u_{aj}}(\theta_q) Q_j^{1-u_{aj}}(\theta_q)] \end{aligned} \tag{3.4}$$

Ignoring the quantity $L(u_a|\theta_q)$ for the moment and expanding the partial derivative in the fraction yields

$$\begin{aligned} &\frac{\partial [P_j^{u_{aj}}(\theta_q) Q_j^{1-u_{aj}}(\theta_q)]}{\partial \psi_j} / [P_j^{u_{aj}}(\theta_q) Q_j^{1-u_{aj}}(\theta_q)] \\ &= (1 - u_{aj}) P_j^{u_{aj}}(\theta_q) Q_j^{-u_{aj}}(\theta_q) \frac{\partial Q_j(\theta_q)}{\partial \psi_j} \\ &\quad + u_{aj} P_j^{u_{aj}}(\theta_q) Q_j^{-u_{aj}}(\theta_q) P_j^{-1}(\theta_q) Q_j(\theta_q) \frac{\partial P_j(\theta_q)}{\partial \psi_j}. \end{aligned}$$

Cancelling out like terms gives

$$(1 - u_{aj}) \frac{\partial Q_j(\theta_q)}{\partial \psi_j} Q_j^{-1}(\theta_q) + u_{aj} P_j^{-1}(\theta_q) \frac{\partial P_j(\theta_q)}{\partial \psi_j}.$$

Substituting $1 - P_j(\theta_q)$ for $Q_j(\theta_q)$ in the partial derivative and simplifying the resulting expression leads to

$$\frac{\partial P_j(\theta_q)}{\partial \psi_j}\left[(u_{aj}-1)Q_j^{-1}(\theta_q) + u_{aj}P_j^{-1}(\theta_q)\right] = \frac{u_{aj}-P_j(\theta_q)}{P_j(\theta_q)Q_j(\theta_q)}\frac{\partial P_j(\theta_q)}{\partial \psi_j},$$

so that

$$\frac{\partial[P_j^{u_{aj}}(\theta_q)Q_j^{1-u_{aj}}(\theta_q)]}{\partial \psi_j}/[P_j^{u_{aj}}(\theta_q)Q_j^{1-u_{aj}}(\theta_q)] = \frac{u_{aj}-P_j(\theta_q)}{P_j(\theta_q)Q_j(\theta_q)}\frac{\partial P_j(\theta_a)}{\partial \psi_j}.$$

Substituting the expression on the right into (3.3) results in

$$\frac{\partial \log ML}{\partial \psi_j} = \sum_a \frac{1}{ML(u_a)}\sum_q w_q L(u_a|\theta_q)\frac{u_{aj}-P_j(\theta_q)}{P_j(\theta_q)Q_j(\theta_q)}\frac{\partial P_j(\theta_q)}{\partial \psi_j} = 0,$$

and exchanging the order of the two summations yields the expression

$$\begin{aligned}\frac{\partial \log ML}{\partial \psi_j} &= \sum_q w_q \frac{\partial P_j(\theta_q)}{\partial \psi_j}\frac{1}{P_j(\theta_q)Q_j(\theta_q)} \\ &\quad \times \sum_{a=1}^{N}\left[\frac{u_{aj}L(u_a|\theta_q)}{ML(u_a)} - \frac{P_j(\theta_q)L(u_a|\theta_q)}{ML(u_a)}\right] = 0. \qquad (3.5)\end{aligned}$$

Two quantities within (3.5) have a natural interpretation. What shall be denoted as

$$N_q = \sum_{a=1}^{N}\frac{w_q L(u_a|\theta_q)}{ML(u_a)}, \qquad (3.6)$$

can be interpreted as the expected number of examinees associated with θ_q, since the quantity being summed is an estimate of the probability of examinee a having θ_q. Second, the quantity which shall be indicated by

$$f_{jq} = \sum_{a=1}^{N}\frac{w_q u_{aj} L(u_a|\theta_q)}{ML(u_a)} \qquad (3.7)$$

is the expected frequency of right answers for item j for examinees associated with θ_q.

Substituting for these two quantities, (3.5) simplifies to

$$\frac{\partial \log ML}{\partial \psi_j} = \sum_q [f_{jq} - P_j(\theta_q)N_q]\frac{\partial W_j(\theta_q)}{\partial \psi_j} = 0. \qquad (3.8)$$

According to the binomial distribution, the log likelihood of getting f_{jq} successes in N_q trials given probability of success P_{jq} is

$$\log L_j = \sum_{q=1}^{Q} f_{jq} \log P_{jq} + (N_q - f_{jq}) \log Q_{jq}, \tag{3.9}$$

and taking the derivative of this expression with respect to W_{jq} and setting it to zero yields exactly (3.8). Thus, (3.8) essentially describes a binomial sampling experiment.

The M-phase then involves solving equation (3.8) for each item j, regarding N_q and f_{jq} as fixed. As soon as estimates of the parameter vectors ψ_j are obtained in this M-step, the E-phase is revisited and the marginal likelihoods for each examinee are recomputed using the most recent parameter estimates. The algorithm iterates between the E- and M-phases until neither the marginal likelihoods nor the parameters ψ_j change significantly from one iteration to the next.

3.2.3 Starting values for the item response functions

The first iteration of the EM algorithm requires some provisional estimates of the item parameters and the item response functions. Estimates of $P(\theta_a)$ can be obtained as follows:

1. Compute N standard normal quantiles $z_a = \Phi^{-1}[1/(N+1)]$.
2. For each quantile, compute the index t_a such that $\theta_{q-1} + \theta_q \leq 2z_a < \theta_q + \theta_{q+1}$, or assign the index of 1 or Q as appropriate.
3. Compute the total scores $x_a = \sum_i u_{ai}$.
4. Sort the examinees with respect to the total scores x_a.
5. For every i and q, set $\hat{P}_{iq}$ to the average of that item's scores, which are either 1 or 0, for the score-sorted examinees with index $t_a = q$.

Estimates with values of 0 or 1 are replaced by values such as $1/2N$ and $1 - 1/2N$, respectively.

3.2.4 Approximation by quadrature

As stated above, some quadrature rule must be employed to estimate the integral

$$\int L(u_a|\theta)g(\theta)d\theta$$

in the E-step. The question is which rule will yield a satisfactory approximation to the integral.

Gaussian quadrature considers formulas of the form

$$\int_a^b f(x)dx \simeq \sum_{q=1}^{Q} w_q f(x_q), \tag{3.10}$$

where the quadrature points $x_1, x_2, \ldots, x_Q$ and weights $w_1, w_2, \ldots, w_Q$ are chosen to minimize the expected error obtained when performing the approximation for some arbitrary function f. Thus, Gaussian quadrature chooses the points for evaluation in an optimal manner.

To determine the accuracy of the rule in (3.10), it is typically assumed that the best choice of values is that producing the exact result for the largest class of polynomials. Now (3.10) is exact for the class of polynomial functions if and only if it is exact for the monomials $f(x) = 1, x, \ldots, x^Q$ (Phillips & Taylor, 1996). To be exact for $f(x) = x^j$, it is required that

$$\int_a^b x^j dx = \sum_{q=1}^{Q} w_q x_q^j. \tag{3.11}$$

Now the left side of (3.11) is known, which implies that by taking $j = 0, 1, \ldots, 2Q-1$, $2Q$ equations can be set up to solve for the $2Q$ unknowns w_q and x_q, $q = 1, \ldots, Q$. If these equations have a solution, then the resulting quadrature rule will be exact for all polynomial functions of degree $2Q-1$ or less.

There also exist Gaussian quadrature formulas of the form

$$\int_a^b \omega(x)f(x)dx \simeq \sum_{q=1}^{Q} w_q f(x_q). \tag{3.12}$$

As above, weights w_q and points x_q can be found so that (3.12) is exact for all polynomial functions of degree $2Q-1$ or less. As this paper is concerned with an

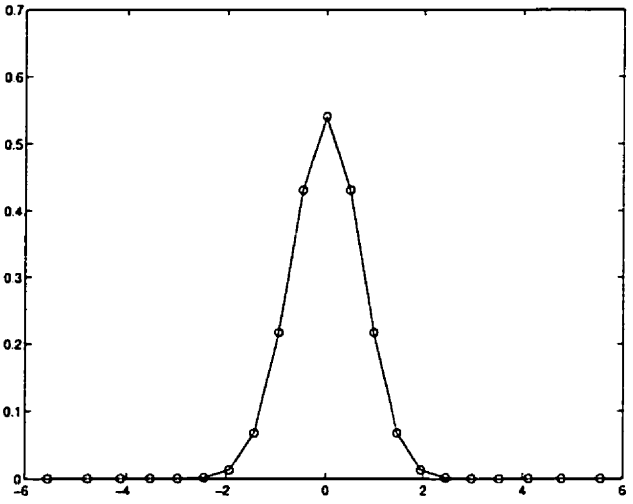

Figure 3.1: Gauss-Hermite quadrature weights and points. This quadrature rule provides a number of quadrature points in areas where there are no data available to estimate the integral. In this case, $-5.5 \leq x_q \leq 5.5$, whereas θ has been fixed to range from -2.5 to 2.5.

infinite integral, the Gaussian rule takes the form

$$\int_{-\infty}^{\infty} \omega(x) f(x) dx \simeq \sum_{q=1}^{Q} w_q f(x_q). \tag{3.13}$$

The selection of $\omega(x) = \exp(-x^2)$ gives the Gauss-Hermite quadrature rules. The Gauss-Hermite rules have been used by Bock and Aitkin (1981) to estimate the marginal likelihood, and so they were employed here as well.

The optimal weights w_q and points x_q obtained with this method are shown in Figure 3.1. It is clear than in this case $-5.5 \leq x_q \leq 5.5$, whereas θ has been fixed to range from -2.5 to 2.5. Thus, the Gauss-Hermite quadrature rule provides a number of quadrature points in areas where there are no data available to estimate the integral, and using these weights and points did not produce reasonable estimates of the item response functions. As a result, the Gauss-Hermite quadrature rule was abandoned in exchange for a rule that allows for control over the location of the quadrature points so that these points can be restricted to be equally spaced about the desired range.

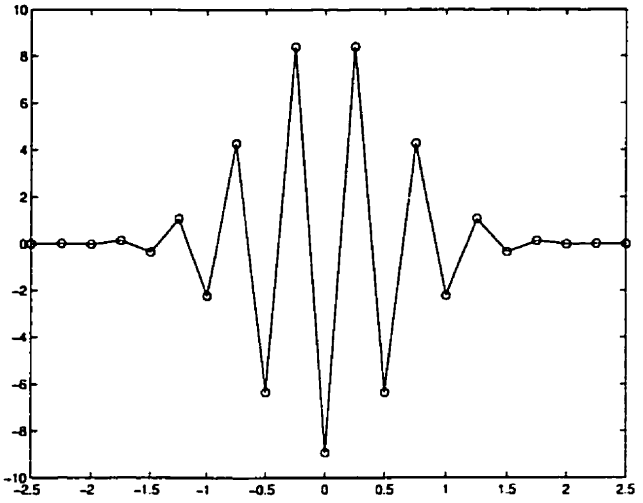

Figure 3.2: Optimal weights for fixed x_q. In comparison to Gauss-Hermite, these quadrature points are bounded by ± 2.5. However, with the application of this set of weights, the convergence of the algorithm became unstable.

In order to eliminate the extreme quadrature points, the points were forced to be equally spaced and only the weights were optimized with respect to the set of Q polynomial functions. That is, for each *test function* $f_i(x) = x^{i-1}, i = 1, \ldots, Q$, the relation

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} f_i(x) dx = \sum_{q=1}^{Q} w_q f_i(x_q) \tag{3.14}$$

was satisfied. The weights obtained by this rule are shown in Figure 3.2. As an improvement, the quadrature points are bounded by ± 2.5. However, with this choice of weights the convergence of the algorithm became unstable.

The reason for this instability is that the set of weights used are optimal with respect to the class of *polynomial* functions. Thus, although they may work well with the polynomials, this may not be the case with other classes of functions. In particular, the current application involves integration over the conditional likelihood functions, which do not look like polynomials (see Figure 3.3). For a particular examinee, the conditional likelihood function is a single-peaked curve with its maximum at the examinee's ability level θ_a and tails quickly approaching

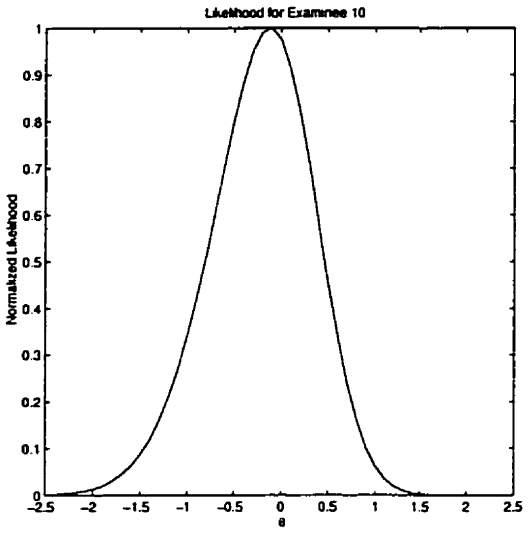

Figure 3.3: The likelihood for a simulated examinee's data in a 3PL model test, rescaled to have a maximum of one. This function more closely resembles a B-spline than a polynomial.

zero in both directions. Thus the collection of N conditional likelihood functions will be a set of peaks whose locations vary across the θ scale.

What is desired is a set of weights that work well with a set of test functions resembling the functions which will be integrated. The class of B-spline basis functions serves this purpose well (see Figure 2.4). Using the quadrature rule (3.14) where the $f_i(x)$ are the B-spline basis functions, the weights shown in Figure 3.4 were obtained. As with the optimal weights, these are bounded by ± 2.5 but with the advantage that convergence of the algorithm became stable.

3.3 Regularizing the Fit

When using a basis function expansion to derive smooth estimates of the item response functions, there is another issue which merits consideration. This is a consequence of the relationship between the number of basis functions and the degree of fit to the data. In particular, as the number K basis functions increases the fit to the data improves. However, the fitted item response function also

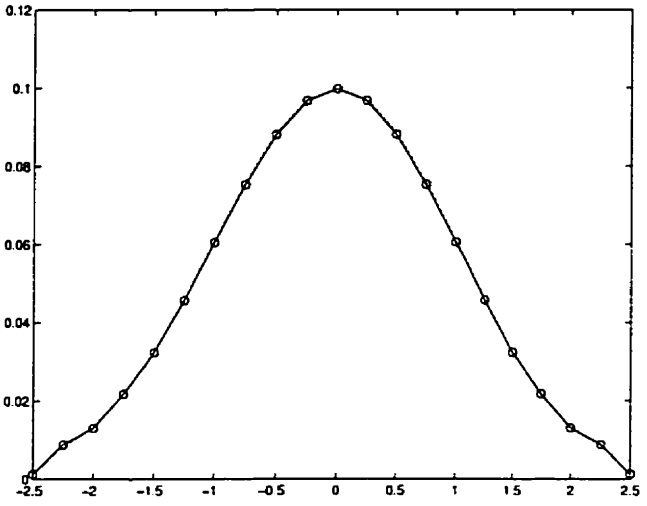

Figure 3.4: Quadrature weights using B-spline basis test functions. The dotted line represents the quantiles of the standard normal distribution. As with the optimal weights, these weights are bounded by ± 2.5 but with the advantage that convergence of the algorithm became stable.

becomes less smooth. What is needed is a compromise between fit to the data and the smoothness of the estimate.

Ramsay and Silverman (1997) describe the *regularization* or *roughness penalty* for forcing a high-dimensional basis expansion to be smooth. A common measure of the roughness of a function is given by its integrated squared second derivative,

$$\mathrm{PEN}_2(f) = \int \{D^2 f(x)\}^2 dx, \tag{3.15}$$

where $D^m f(x)$ is the mth derivative of $f(x)$. This quantity assesses the degree of curvature in function f, or equivalently the degree to which f deviates from a straight line. Functions with a high degree of curvature will manifest large values of $\mathrm{PEN}_2(f)$ since their second derivatives are large across the range of interest.

Establishing a compromise between fit and smoothness then amounts to modifying the model fitting criterion $\log ML(\mathbf{U})$ to the following *penalized negative log likelihood* function:

$$F_\lambda(\mathbf{U}) = -\log ML(\mathbf{U}) + \lambda \sum_{i=1}^{n} \int [D^m W_i(\theta)]^2 d\theta, \tag{3.16}$$

where λ is a *smoothing parameter*. For the case where $m = 2$, $\int [D^2 W_i(\theta)]^2 d\theta$ is a measure of the total curvature of $W_i(\theta)$. The more variable the function the larger this quantity is going to be, and the closer the function to a straight line the closer the value of the total curvature is to zero. The degree to which the fitting criterion is to be penalized by the curvature of the W's is controlled by smoothing parameter λ. For λ close to zero, the penalty is relaxed and so the data is fit without any regularization. However, as λ increases the curvature of the W's becomes exceedingly more significant in determining the value of $F_\lambda(\mathbf{U})$. Thus there needs to be less curvature in the W's in order to obtain its minimum until ultimately, as λ approaches infinity, the W's are forced to be linear. In the IRT case, linear W's are equivalent to the two-parameter logistic model.

The latter statement describes an important problem when smoothing with D^2. In penalizing the second derivative of the W_i's, the functions are forced to a straight line as λ approaches infinity. For linear $W(\theta)$, $P(\theta)$ corresponds to the two-parameter logistic model. The properties of actual test items, however, are often insufficiently described by the 2PL model (see section 1.3.2). In particular, the 2PL model assumes that the item response curves have a left asymptote equal to zero. Responses to real test items, however, may be a result of guessing, in which case the item response function will manifest a left asymptote which differs from zero. Thus it would be more appropriate to apply a roughness penalty that, when applied heavily, smooths $P(\theta)$ towards the three-parameter logistic model.

In smoothing towards a 3PL model, there are important features of these curves that must be captured by any expansion of the $W(\theta)$'s. The functions are monotone increasing with right asymptote 1 and left asymptote c, thus the possibility that $dW/d\theta = 0$ for large negative and large positive θ must be accommodated. Furthermore, in regions where there is likely to be little data, in particular large positive and large negative values of θ, $W(\theta)$ should be linear, or alternatively, $dW/d\theta$ should be constant.

In order to accommodate these features, Wang (1993) used the basis

$$\begin{aligned}\phi_1(\theta) &= 1 \\ \phi_2(\theta) &= \theta \\ \phi_3(\theta) &= \ln(e^\theta + 1).\end{aligned} \tag{3.17}$$

Function $W(\theta)$ is then expressed as

$$W(\theta) = \sum_{k=1}^{3} c_k \phi_k(\theta). \tag{3.18}$$

Figure 3.5 displays three item response functions $P(\theta)$ generated by the 3PL model and the corresponding $W(\theta)$'s. The proximity of the approximated $W(\theta)$'s to the actual logit of the $P(\theta)$'s demonstrates the appropriateness of the basis functions in (3.17) for capturing the essential characteristics of these curves. For example, for large positive values of θ, $\ln(e^\theta+1)$ approaches θ and so $W(\theta)$ is asymptotically linear on the right and hence the derivative of $W(\theta)$,

$$\frac{dW}{d\theta} = c_2 + c_3 \frac{e^\theta}{e^\theta + 1},$$

is constant. Similarly, for large negative values of θ, $dW/d\theta$ is asymptotically c_2.

The roughness penalty $F_\lambda(\mathbf{U})$ that will be used is of a more general form (see Heckman & Ramsay, 2000), replacing the D^2 in the original penalty term (3.15) with a linear differential operator L,

$$LW = \alpha_0(\theta)W(\theta) + \alpha_1(\theta)DW(\theta) + \alpha_2(\theta)D^2W(\theta) + D^3W(\theta). \tag{3.19}$$

For the purpose of smoothing towards the 3PL model, weighting functions $\alpha_j(\theta)$ are chosen such that any function which is a linear combination of the basis functions (3.17) will yield a value of LW equal to zero. This amounts to choosing $\alpha_j(\theta)$ such that, for $j = 1, \ldots, 3$,

$$\alpha_0(\theta)\phi_j(\theta) + \alpha_1(\theta)D\phi_j(\theta) + \alpha_2(\theta)D^2\phi_j(\theta) + D^3\phi_j(\theta) = 0.$$

The weight functions are then determined by setting up a system of three linear equations, one for each basis function $\phi_j(\theta)$, and solving for the $\alpha_j(\theta)$. The result

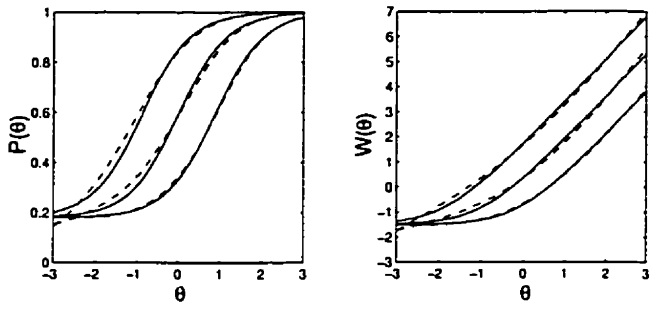

Figure 3.5: The solid lines in the left panel are three 3PL item response functions $P(\theta)$ with $a = 1$, $c = .18$ and varying values for location parameter $b = -1.5, 0, 1.5$. The solid lines in the right panel are the corresponding $W(\theta)$. For each curve, the nearest dashed line indicates the approximation based on the three basis functions in (3.17).

is

$$\begin{aligned} \alpha_0(\theta) &= \alpha_1(\theta) = 0 \\ \alpha_2(\theta) &= \frac{e^\theta - 1}{e^\theta + 1}. \end{aligned} \tag{3.20}$$

Thus, the new roughness penalty is

$$\mathrm{PEN}_L(W) = \int_{-\infty}^{\infty} \left[\frac{e^\theta - 1}{e^\theta + 1} D^2 W + D^3 W \right]^2 d\theta \tag{3.21}$$

and the new penalized negative log likelihood function,

$$F_\lambda(\mathbf{U}) = -\log ML(\mathbf{U}) + \lambda \sum_{i=1}^{n} \int_{-\infty}^{\infty} \left[\frac{e^\theta - 1}{e^\theta + 1} D^2 W_i + D^3 W_i \right]^2 d\theta, \tag{3.22}$$

will, when applied heavily, force the $W_i(\theta)$ to conform to something like the 3PL model.

Chapter 4

Example Analyses

The examples in this chapter are designed to show how the nonparametric estimation of item response functions performs in practice. The first section illustrates its performance for simulated data, and the second presents results for an actual set of testing data. In both instances, modest sample sizes are involved in order to display results in a demanding environment.

4.1 A Simulated 3PL Test

For the simulated data, the goal is to assess the performance of the algorithm with respect to three variables:

1. the number of test items n,
2. the number of examinees N, and
3. the value of smoothing parameter λ.

These factors are considered to be the most important for fitting the data well. As the number of items increases, more information is available regarding an individual examinee's ability level θ_a allowing for a better estimate of the item response function at θ_a. A similar argument can be made for increases in the number of examinees. Smaller values for smoothing parameter λ imply that the shape of the curve is more and more dependent on the actual data so that in the

extreme case, the curve fits the pseudo-probabilities f_{jq}/N_q as close as the basis system will allow. On the other hand, larger values for λ result in smoother curves that depend on the data less and less. In the latter case, the curves become closer to 3PL curves as λ increases. Simulating a 3PL test, it is expected that higher values for λ will produce curves that are closer to the true item response curves.

Three levels of the number of items ($n = 25, 50, 100$) and the number of examinees ($N = 500, 1000, 2000$) parameters were examined, along with four levels of the smoothing parameter ($\lambda = .01, 1, 10, 100$). The examinee ability levels θ were generated according to the standard normal distribution. Three sets of 3PL item parameters were randomly generated which produced three tests, one each of a 25, 50 and 100-item test. These item parameter remained fixed across all replications. For each test, 25 simulations were run within each combination of the levels of N and λ, resulting in a total number of 900 simulations. For each simulation, 14 B-spline basis functions of order 5 were used with 21 quadrature points and the roughness penalty described in section 3.3 was applied.

Goodness of fit is assessed as the square root of the average squared difference between the estimated and actual curve values. This quantity shall be referred to as the root mean square error, or RMSE:

$$\text{RMSE}_i = \sqrt{[P_i(\theta) - \hat{P}_i(\theta)]^2},$$

Taking this average over the entire θ range would yield a global measure of fit, whereas a more informed decision regarding the performance of the algorithm should assess the fit of the estimates at various locations along the θ scale, as it is expected that the curve estimates will be better in some areas than in others. For instance, some deviation of the estimate from the true curve is expected in the region of the lower ability levels. A breakdown of the algorithm in the estimation of the left asymptote can be attributed to the assumption that ability level is normally distributed. In the instance of 500 examinees, this amounts to the availability of approximately 12 pieces of data with which to estimate the lower tail end of the curve (i.e., in the region $\theta \leq -2$). Although a similar argument

Table 4.1: ANOVA Table for $\theta = -2$.

Source	SS$\times 10^3$	d.f.	MS$\times 10^3$	F	Prob>F
λ	42.03	3	14.01	164.8	0.000
N	277.52	2	138.76	1632.4	0.000
n	65.90	2	32.95	387.6	0.000
$\lambda \times N$	1.68	6	0.28	3.3	0.003
$\lambda \times n$	5.69	6	0.95	11.2	0.000
$N \times n$	3.85	4	0.96	11.3	0.000
$\lambda \times N \times n$	0.90	12	0.08	0.9	0.565
Error	73.44	864	0.09		
Total	471.02	899			

Table 4.2: ANOVA Table for $\theta = -1$.

Source	SS$\times 10^3$	d.f.	MS$\times 10^3$	F	Prob>F
λ	6.21	3	2.07	197.1	0.000
N	47.02	2	23.51	2236.3	0.000
n	0.43	2	0.21	20.2	0.000
$\lambda \times N$	0.28	6	0.05	4.5	0.000
$\lambda \times n$	0.20	6	0.03	3.2	0.004
$N \times n$	0.05	4	0.01	1.1	0.363
$\lambda \times N \times n$	0.28	12	0.02	2.2	0.010
Error	9.08	864	0.01		
Total	63.54	899			

may be made for the region where $\theta \geq 2$, less deviation is anticipated here since the right asymptote is constrained to a value of unity.

The RMSE between the true curve $P(\theta)$ and the estimate $\hat{P}(\theta)$ was obtained at five different θ values: $-2, -1, 0, 1, 2$. An ANOVA was performed on this measure for each of these θ values. The results are given in Tables 4.1 through 4.5.

Factor standard deviations

To assess the importance of the significant effects, estimates of the factor standard deviations were calculated for each effect A at each level of θ. Where $\alpha_j = \mu_j - \mu$ are the factor effects and a is the number of levels within effect A, the factor

Table 4.3: ANOVA Table for $\theta = 0$.

Source	SS$\times 10^3$	d.f.	MS$\times 10^3$	F	Prob>F
λ	0.69	3	0.23	25.6	0.000
N	33.63	2	16.82	1875.5	0.000
n	1.31	2	0.65	72.8	0.000
$\lambda \times N$	0.07	6	0.01	1.3	0.237
$\lambda \times n$	0.04	6	0.01	0.7	0.676
$N \times n$	0.30	4	0.08	8.4	0.000
$\lambda \times N \times n$	0.09	12	0.01	0.9	0.590
Error	7.75	864	0.01		
Total	43.87	899			

Table 4.4: ANOVA Table for $\theta = -1$.

Source	SS$\times 10^3$	d.f.	MS$\times 10^3$	F	Prob>F
λ	1.47	3	0.49	45.1	0.000
N	39.37	2	19.68	1811.6	0.000
n	1.39	2	0.07	64.1	0.000
$\lambda \times N$	0.36	6	0.06	5.6	0.000
$\lambda \times n$	0.18	6	0.03	2.7	0.012
$N \times n$	0.16	4	0.04	3.7	0.005
$\lambda \times N \times n$	0.15	12	0.01	1.2	0.309
Error	9.39	864	0.01		
Total	52.47	899			

Table 4.5: ANOVA Table for $\theta = 2$.

Source	SS$\times 10^3$	d.f.	MS$\times 10^3$	F	Prob>F
λ	10.54	3	3.51	96.0	0.000
N	68.64	2	34.32	937.5	0.000
n	5.81	2	2.91	79.4	0.000
$\lambda \times N$	2.38	6	0.40	10.8	0.000
$\lambda \times n$	1.71	6	0.29	7.8	0.000
$N \times n$	0.57	4	0.14	3.9	0.004
$\lambda \times N \times n$	1.09	12	0.09	2.5	0.003
Error	31.63	864	0.04		
Total	122.37	899			

Table 4.6: Factor Standard Deviations $\times 10^2$.

Effect	θ				
	-2	-1	0	1	2
λ	0.6813	0.2621	0.0858	0.1264	0.3404
N	1.7555	0.7226	0.6111	0.6612	0.8729
n	0.8546	0.0670	0.1196	0.1235	0.2525
$\lambda \times N$	0.1142	0.0491	$\emptyset$	0.0576	0.1549
$\lambda \times n$	0.2400	0.0391	$\emptyset$	$\emptyset$	0.1287
$N \times n$	0.1975	$\emptyset$	0.0542	0.0362	0.0685
$\lambda \times N \times n$	$\emptyset$	0.0412	$\emptyset$	$\emptyset$	0.0852

standard deviations are defined as

$$\delta_A = \sqrt{\frac{\sum_{j=1}^{a} \alpha_j^2}{a}}.$$

The estimates of the standard deviations for each effect and for each value of θ are given in Table 4.6, magnified by a factor of 10^2. The estimates for nonsignificant effects are denoted by $\emptyset$.

It is the three main effects that appear to have the largest standard deviations across all values of θ. The parameter having the greatest effect is the number of examinees N, particularly when $\theta = -2$. In this case, the standard deviation for the number of examinees is more than two times greater than that for the number of items and smoothing parameter λ.

Across the various effects, the standard deviations are greatest for $\theta = -2$, which was expected. With only 500 examinees, precise estimation of $P(\theta)$ is less likely in this region since there are not more than 12 or 13 pieces of data available on the average. Increases in smoothing parameter λ force the estimates to look more and more like 3PL curves, and so possibly to look like the 3PL curves that generated the data. Any increases in sample size or number of items would provide more information on $P(\theta)$ at this ability level, and so greatly effect the estimation of the curve. Although this argument would seem to imply a similar finding for the effect sizes for $\theta = 2$, these values do not match those for $\theta = -2$, although they are the next largest in magnitude. Unlike the left asymptote, the

right asymptote does not vary but instead is forced to unity as λ increases. Thus there is less variation in the estimation of the right asymptote than there is for the left asymptote.

For most effects, the factor standard deviations are lowest at $\theta = 0$. Having randomly sampled ability levels θ from a standard normal distribution, with 500 examinees there are roughly 191 pieces of data available between $\theta = -0.5$ and $\theta = 0.5$ with which to produce an estimate of the item response function within this region. With this amount of data, it is likely that the estimate of the item response function here is quite good to begin with so that changes in the number of items, number of examinees or the degree of smoothing would not produce an improvement in fit as drastic as in the extremes of θ. This notion is further supported in observing the factor standard deviations of the main effects for $\theta = -1$ and $\theta = 1$. In general, the magnitudes of these values are greater than at $\theta = 0$ (where there is the greatest amount of data available), but less than their respective extremes (where the least amount of data is available). An argument similar to that made regarding the disparity among the standard deviations for $\theta = -2$ and $\theta = 2$ can be used to explain the greater standard deviations for $\theta = -1$ as compared to $\theta = 1$.

Interaction effects

In observing the factor standard deviations, it is clear that the main effects show more variation than do the interaction effects. However, the latter are a valuable source of information regarding the performance of the algorithm. Since all three variables are assumed to contribute considerably to the fit, it is important to examine the interactions among them so that appropriate decisions can be made regarding, for example, the value of λ. For instance, is there a particular value of λ that should be applied generally, or should the value of this parameter be adjusted according to sample size N? In this section, all interactions which were found to be significant are discussed, such that questions of this sort may be addressed.

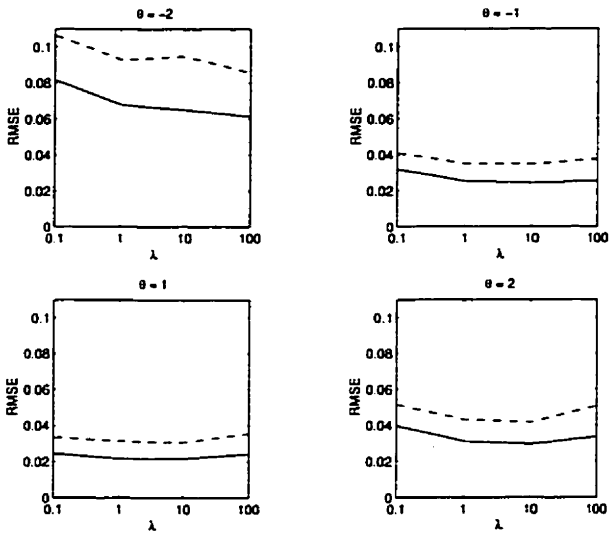

Figure 4.1: Plots of the two-way interaction between smoothing parameter λ and sample size N at $\theta = -2, -1, 1$ and 2. The dashed line represents an N of 500, the solid line an N of 1000, and the dotted line an N of 2000.

The $\lambda \times N$ two-way interaction was significant at all levels of θ except $\theta = 0$. Figure 4.1 shows the plots of the interactions at the relevant θ levels. The plots vary in terms of the y-axis labelling. The RMSE is greatest at $\theta = -2$ and so there is a greater range for this plot as compared to the others. The range is also slighter greater for $\theta = 2$, whereas the plots for $\theta = -1$ and $\theta = 1$ are roughly the same. Aside from the disparity in range, there are some similarities among the graphs. Most obvious is that for each line representing a particular number of examinees N, the greatest decrease in RMSE occurs in moving from a λ of .1 to a value of unity. A value of $\lambda = .1$ appears to be too permissive a value for this parameter, and the data is undersmoothed.

Another similarity across the θ range is the pattern for $N = 2000$. For all θ, it achieves its greatest decrease in RMSE in moving from $\lambda = .1$ to 1, after which RMSE does not further achieve a significant decrease. Similarly, for $N = 1000$, the greatest decrease in RMSE occurs when moving from $\lambda = .1$ to 1, and there is no further significant decrease from $\lambda = 1$ to 100. For $N = 500$, a significant decrease in RMSE occurs from $\lambda = .1$ to 1, and there is no significant change from

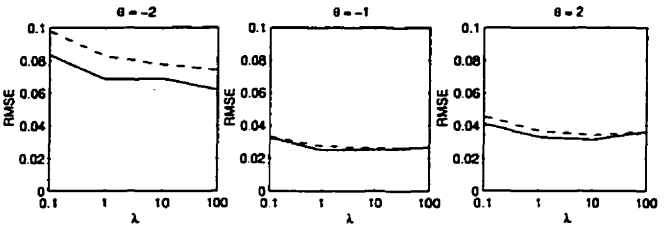

Figure 4.2: Plots of the two-way interaction between smoothing parameter λ and number of test items n at $\theta = -2$, -1 and 2. The dashed line represents an n of 25, the solid line an n of 50, and the dotted line an n of 100.

a λ of 1 to 10. The real variation in $N = 500$ comes in moving from $\lambda = 10$ to 100. For $\theta = -2$, the error decreases, but for all other θ values, the RMSE increases.

One possible explanation as to why an increase in λ would worsen the fit only at the upper end of the θ scale is as follows. For all N, the most troublesome region in which to obtain a close fit to the true curves is in the vicinity of $\theta = -2$. This is due to the variability permitted in the value of the left asymptote. It may vary considerably from zero. In contrast, the right asymptote does not differ from unity. Increasing λ has the effect of forcing $\hat{P}(\theta)$ to look more and more like a 3PL curve, and so in the upper regions all curves are forced to have an right asymptote of one. In the region where $\theta = -2$, not only is there little data with which to estimate $\hat{P}(\theta)$, but the curve may also vary greatly from zero. Thus, a high degree of smoothing actually serves to improve the proximity of $\hat{P}(\theta)$ to $P(\theta)$. At the other extreme, $\hat{P}(\theta)$ is already close to $P(\theta)$ when $\lambda = 10$, since there is little allowance for variation at the right asymptote. As λ increases, the estimated curve is oversmoothed and so made to be flatter than the target $P(\theta)$, resulting in a increase in RMSE as λ increases from 10 to 100.

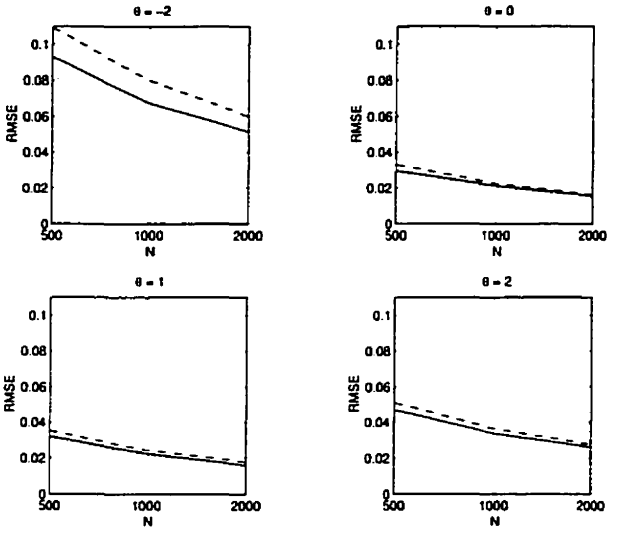

Figure 4.3: Plots of the two-way interaction between sample size N and number of test items n at $\theta = -2, 0, 1$ and 2. The dashed line represents an n of 25, the solid line an n of 50, and the dotted line an n of 100.

The $\lambda \times n$ interaction was found to be significant only at the θ values of -2, -1 and 2. Figure 4.2 displays the relevant graphs. Again, there is a greater range (reflecting a worse fit) for the extreme θ values as compared to $\theta = -1$, with the greatest range at $\theta = -2$. As with the number of examinees parameter, for all the values of n the greatest decrease in RMSE occurs with a move from $\lambda = .1$ to $\lambda = 1$, suggesting once again that $\lambda = .1$ is too loose a fitting criterion. The main differences in pattern here are similar to those noted for the interaction between λ and N. Specifically, for $\theta = -2$, there is an overall decrease in RMSE as λ increases, with the best fit for each distinct test occurring for $\lambda = 100$ (although none of the fits at $\lambda = 100$ are significantly better than those at $\lambda = 1$). A smoothing parameter value this high for $\theta = -1$ and $\theta = 2$, however, has in adverse effect on the closeness of the estimate to the true curve. In these regions, there is no significant reduction in RMSE beyond $\lambda = 1$. Thus, as with N, an increase to $\lambda = 100$ improves the fit in regions where there is more opportunity for variability (i.e., $\theta = -2$) but has the effect of oversmoothing the data in regions allowing for less variability (e.g., where $\theta = 2$).

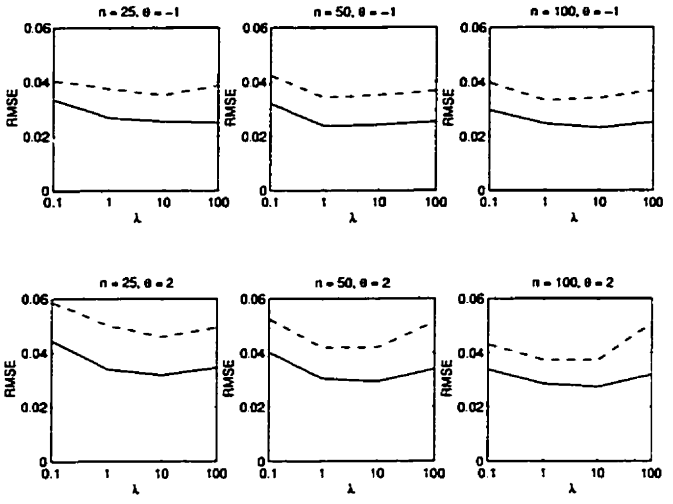

Figure 4.4: Plots of the three-way interaction between smoothing parameter λ, sample size N and number of test items n at $\theta = -1$ and 2. The dashed line represents an N of 500, the solid line an N of 1000, and the dotted line an N of 2000.

The $N \times n$ interaction effect was significant at all θ levels except $\theta = -1$. Figure 4.3 displays the relevant graphs. As with the interactions discussed above, the ranges for both $\theta = -2$ and $\theta = 2$ are greater than the others.

For all values of θ, the greatest decrease in RMSE comes with an increase from 500 to 1000 examinees. The decrease is particularly great for the 25 item test, and is least noticed for the 100 item test. This difference is due to the increase in the amount of information available on θ_a. With more items there is more information available with which to estimate $\hat{P}(\theta)$. The fit improves further for all tests in moving from 1000 to 2000 examinees, although the change is slightly less drastic as the move from $N = 500$ to $N = 1000$. As with the number of items, increases in the number of examinees also provides more information about θ_a.

The three-way interaction was found to be significant at $\theta = -1$ and $\theta = 2$. The plots of these interactions are shown in Figure 4.4, where the interaction between λ and N is displayed at each level of n. As it seems to be common throughout, the greatest decrease in RMSE occurs when λ moves from .01 to 1,

and this is true for all levels of n.

The pattern for $N = 2000$ varies little across n or at the differing values of θ. There is some decrease in RMSE as λ changes from .1 to 1, but there is little change in fit with any further increases in λ, which is not surprising. Given such a large sample size, there is a sufficient amount of data with which to estimate $P(\theta)$, and increases in λ and/or the number of items can contribute little to an already close fit.

The pattern for $N = 1000$ resembles that for $N = 2000$. It is the sample size of 500 which seems to be the source of the three-way interactions. For both values of θ, the best fit occurs at $\lambda = 10$ if $n = 25$. However, if $n = 50$ or 100, the best fit occurs when $\lambda = 1$. It might be the case that with $N = 500$ and $n = 25$, there is minimal information available regarding θ_a, and so a more stringent smoothing parameter has a beneficial effect. But when $n = 50$ or 100, there is sufficient information on θ_a such that increasing smoothing parameter λ from 1 to 10 does little to decrease the RMSE. In fact, increasing λ actually significantly worsens the fit when $\theta = 2$, presumably due to oversmoothing in the asymptote region, where a certain amount of curvature is appropriate. In contrast, this oversmoothing effect is significant at $\theta = -1$ only when $n = 25$.

4.1.1 Examples of Estimated Item Response Functions

Figure 4.5 shows various estimates of the item response functions for two items with fixed λ and test length and varying values for the number of examinees. These items in particular were chosen as they appear to best represent the sort of variation found throughout the various simulated tests. When sample size $N = 500$, the left asymptote for both curves is poorly estimated. There is also some discrepancy between the true curve and the estimate in the center region of the θ scale. For both items, the estimate improves when $N = 1000$, although the lower ends of the curves are not approximated as well as the center and upper regions. There is a slight improvement when $N = 2000$, although this increase in

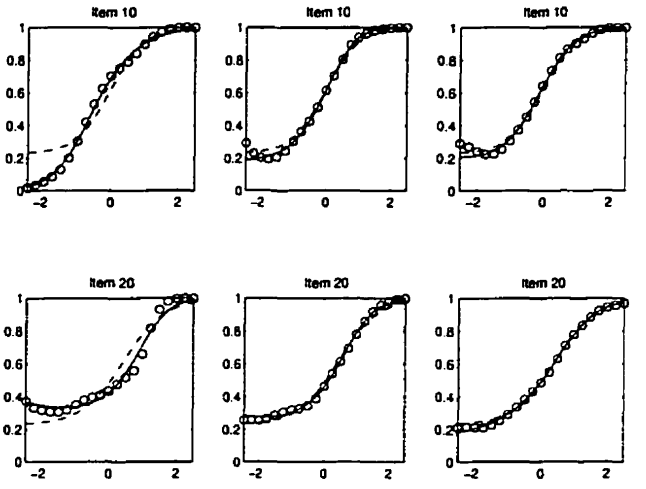

Figure 4.5: The estimated item response functions for items 10 and 20 varying across the number of examinees with $\lambda = 1$ and $n = 50$. The dashed curve represents the true item response function, the solid curve is the estimated function, and the circles are the probabilities f_{jq}/N_q. The estimates in the first column are based on a sample size of 500 examinees, those in the second column on 1000 examinees, and those in the third column on 2000 examinees.

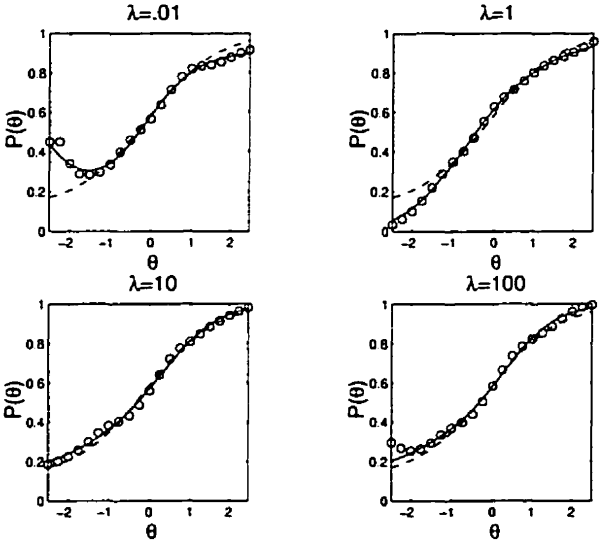

Figure 4.6: The true and estimated curves for item 17 for varying values of λ with $N = 1000$ and $n = 50$.

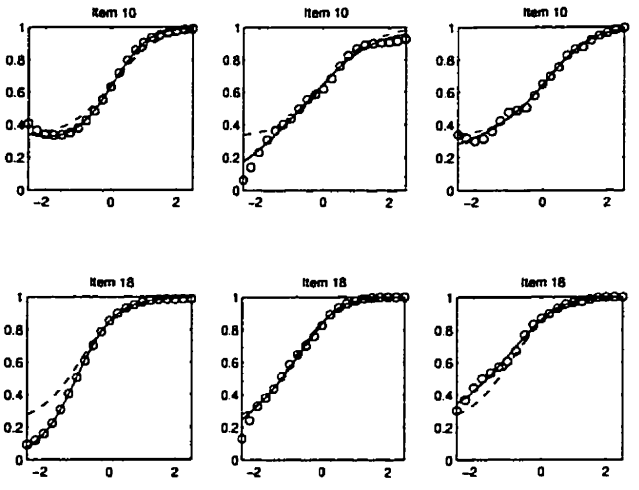

Figure 4.7: The true and estimated curves for items 10 and 18 for various test lengths, with $\lambda = 1$ and $N = 1000$. The estimates in the first column are based on a 25 item test, those in the second column on a 50 item test, and those in the third column on a 100 item test.

sample size adds little to the already excellent estimates when $N = 1000$.

Figure 4.6 displays the estimated item response functions for an item for different degrees of smoothness with fixed sample size and test length. The closeness of the estimate to the true curve appears to increase as λ increases, the closest fit achieved when $\lambda = 10$. It seems as though the data is over-smoothed at $\lambda = 100$, where the right asymptote is pulled away from the true item response function and closer to unity.

Figure 4.7 displays the item response functions for two items estimated at various test lengths for fixed sample size and λ. As the number of test items increases, there is more information available on θ, hence it is expected that the RMSE decreases as n increases. However, Figure 4.7 seems to tell a different story. Clearly the 100-item test results in the best estimate for item 10, but it is questionable whether the 50-item test is an improvement over the 25-item test or vice versa. For item 18, the 50 item test provides an excellent estimate of $P(\theta)$ and obviously an improvement over the 25-item test. But the 100-item test does

Table 4.7: Confidence Intervals $\times 10^4$.

θ	Mean Comparison	Lower bound	Mean Difference	Upper Bound
−2	$\mu_{..1} - \mu_{..2}$	102.43	124.59	146.75
	$\mu_{..1} - \mu_{..3}$	186.11	208.27	230.43
	$\mu_{..2} - \mu_{..3}$	61.52	83.68	105.84
−1	$\mu_{..1} - \mu_{..2}$	2.09	9.88	17.68
	$\mu_{..1} - \mu_{..3}$	8.96	16.75	24.55
	$\mu_{..2} - \mu_{..3}$	-0.92	6.87	14.66
0	$\mu_{..1} - \mu_{..2}$	10.29	17.49	24.69
	$\mu_{..1} - \mu_{..3}$	22.12	29.32	36.51
	$\mu_{..2} - \mu_{..3}$	4.63	11.83	19.02
−1	$\mu_{..1} - \mu_{..2}$	14.09	22.02	29.94
	$\mu_{..1} - \mu_{..3}$	21.34	29.26	37.18
	$\mu_{..2} - \mu_{..3}$	-0.68	7.24	15.17
−2	$\mu_{..1} - \mu_{..2}$	12.44	26.98	41.52
	$\mu_{..1} - \mu_{..3}$	47.53	62.07	76.61
	$\mu_{..2} - \mu_{..3}$	20.55	35.09	49.63

not improve on the 50- item test, even though it has more information with which to estimate the true curve.

Although the ANOVA revealed a significant effect for the number of items parameter across all values of θ, the curves in Figure 4.7 suggest that the marginal means for the number of items variable may not differ significantly from one another. Furthermore, the curves also bring to light the possibility that RMSE may not have a negative linear relationship with the number of test items. To address these issues, multiple comparisons among the marginal means for the number of items variable were tested at each level of θ. Alpha was set to .05 and adjusted using the Bonferroni procedure to compensate for the multiple tests.

The confidence intervals for the mean differences are displayed in Table 4.7. All but two mean differences were significant. In addition, the estimated values of the means decreased (i.e., RMSE decreased) as the number of test items increased, and this was true for all values of θ (see Table 4.8).

Table 4.8: Mean Estimates for $n \times 10^2$

θ	$\mu_{..1}$	$\mu_{..2}$	$\mu_{..3}$
-2	8.31	7.06	6.22
-1	2.86	2.76	2.69
0	2.38	2.21	2.09
1	2.58	2.36	2.28
2	3.82	3.55	3.20

4.2 GMAT Data

The data being analyzed here came from the quantitative subscale of the Graduate Management Admission Test (GMAT) administered to 2735 individuals. The subscale consists of 25 multiple choice items, each with four response options. As with the simulated tests, this test was analyzed using 14 B-spline basis functions of order 5 with 21 quadrature points. The analysis was performed four times, with the value of smoothing parameter λ varying for each iteration. The results for several items where $\lambda = 1$ are shown in Figure 4.8. The solid line represents the estimated item response function, the dashed line represents the starting values used for the algorithm, and the circles are the probabilities f_{jq}/N_q.

The estimated curve for item 10 possesses properties of an ideal item response function. The curve is monotone increasing, so the probability of responding correctly to this item increases with ability level. The slope of the curve is highest among average ability $\theta = 0$, thus the item discriminates best among examinees of average ability. However, this item provides little information about examinees with ability levels greater than 1 and less than -1, regions where the slope of the curve is shallow. Similarly, item 2 discriminates best among examinees of lower than average ability since the slope of this curve is highest in the vicinity of $\theta = -1.5$, but provides no information for examinees having $\theta > -0.5$.

The properties of items 5 and 21 are less than ideal. Item 5 has a high guessing level and may be labelled as uninformative. Even examinees of the lowest ability

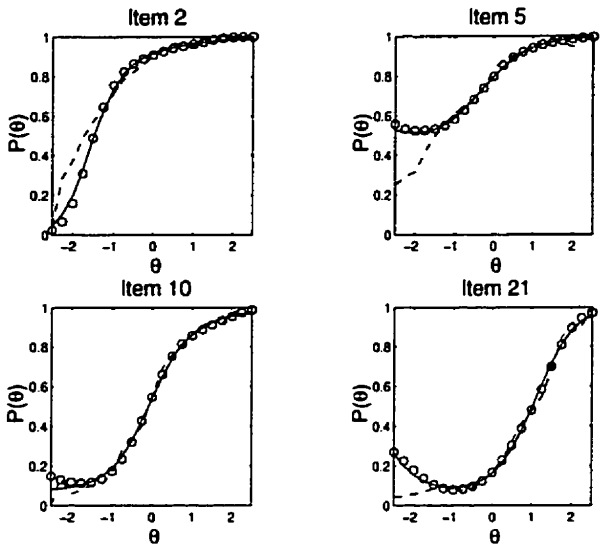

Figure 4.8: Items 2, 5, 10 and 21 of the GMAT quantitative subscale for $\lambda = 1$. The solid line represents the estimated item response function, the dashed line represents the starting values used for the algorithm, and the circles are the probabilities f_{jq}/N_q.

have more than a 50% chance of choosing the correct response to this item. The nonmonotonic nature of the estimated curve for item 21 also renders this item uninformative. Initially, the probability of responding correctly *decreases*, attains its minimum at $\theta = -1$. Thus for at least part of the population, the more knowledgeable the examinee, the less likely he or she is to respond correctly to this item. At $\theta = 1$ the curve changes direction and $P(\theta)$ increases over the remainder of the θ range. Thus from $\theta = -2.5$ to about $\theta = 0.5$, examinees can not be distinguished on the basis of $P(\theta)$ alone since, for example, $P(\theta = -2.5)$ is approximately the same as $P(\theta = 0.5)$.

Figure 4.9 displays the item response function for item 10 estimated at four different levels of smoothing parameter λ. As is evident in the wild behavior of the curve for $\theta < 1$, the function is insufficiently smoothed at $\lambda = .01$. Clearly, more smoothing is necessary in order for the estimate to be reasonable in this region. The problem is resolved by decreasing the total curvature in setting $\lambda = 1$. The curve now resembles a typical 3PL curve and does not appear to require further

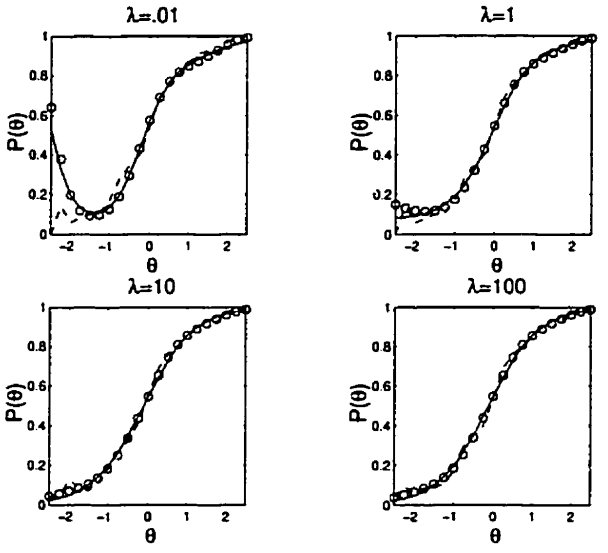

Figure 4.9: Item 10 of the GMAT quantitative subscale, estimated at four different levels of λ.

smoothing. Additional smoothing only reduces the value of the left asymptote.

Chapter 5

Discussion and Conclusions

The approach to item response function estimation described in this thesis has a number of benefits over current approaches. For one, the algorithm involves nonparametric estimation of functions. This draws the focus away from item parameter estimation, and more appropriately places it on the actual estimated curves. Needless to say, item response theory should be concerned with the function representing the relationship between ability level θ and the probability of a correct response, $P(\theta)$. Focussing on parameter estimation detracts one from the essence of the theory.

Furthermore, nonparametric estimation does not accommodate any preconceived notions of item response function behavior. There is relief from the assumption that all items of a single test have the same shape features, permitting greater flexibility and variability in the functions across items. Any particular item is free to manifest a shape that is either unusual or unexpected. For example, the current approach can accommodate nonmonotonic item response curves, whereas the 3PL model would not.

Apart from its ability to fit arbitrary complexities of curves, the basis function method allows for the user to control the smoothness of a result. This can be accomplished either by adjusting the number of basis functions used, or by adjusting smoothing parameter λ as a control over some predetermined roughness penalty. The benefits of the inclusion of a roughness penalty are twofold. First,

it offers a reasonable compromise between the closeness of the estimate to the data and the smoothness of this estimate. Both qualities are desirable, and so a tradeoff between the two must be established. Second, the roughness penalty may be modified so as to allow for "intelligent" smoothing towards a low dimensional baseline model that can be regarded as a sensible default. In the applications presented here, the 3PL model was considered to be an appropriate default.

Also, as an alternative to current estimation procedures, this approach avoids the problem of large sampling covariance between parameter estimates encountered with BILOG and the uncertaintly regarding the consistency of paramater estimates, as well as their large standard errors when using LOGIST. However, as in BILOG, use of the EM algorithm eliminates the estimation of the individual examinee abilities. This has the advantage of increasing the data-to-parameter ratio, or in other words, increasing the amount of information on θ with a corresponding increase in the number of parameters to be estimated. In addition, this method is not computationally demanding. All of the analyses presented in this thesis were performed in less than one minute.

The primary limitation of the current approach is that small sample sizes (≤ 500) suffer from poor item response function recovery, particulary at the low end of the θ scale. Also, the presence of interactions among the three variables examined here presents a challenge in deciding upon a suitable level for λ. In many of the situations a value of $\lambda = 1$ seems appropriate, whereas in others this value should be increased. Still, there are fits that improve little over what is achieved when $\lambda = .1$. Perhaps the most significant drawback is that differing values of λ are optimal for different regions on the θ scale for a single test or item. Often, a curve is well estimated when $\lambda = 1$ for all θ except $\theta \leq 2$.

A future direction of research then is to examine the possibility of specifying λ separately for different regions of a curve. With this added flexibility, λ can be increased in areas where the data tends to be undersmoothed, and decreased where oversmoothing takes precedence. At the very least, it should be possible to

adjust λ separately for each item of a particular test.

This may be important for expansion of the current model to accommodate polytomous data. For a polytomous item, m curves will need to be estimated, one for each of m possible options for that item. Options selected by low ability examinees will require more smoothing than those favored by examinees with average ability. Among other considerations, the roughness penalty may need adjustment in the polytomous case, as smoothing to a 3PL model may not be appropriate.

Despite its limitations, the future of the procedure described here is promising. It overcomes several of the shortcomings of the parametric models, allows for much user-controlled flexibility, and has the potential to provide these benefits beyond the realm of dichotomous test items.

References

Agresti, A. (1990). *Categorical Data Analysis.* New York: John Wiley & Sons, Inc..

Bock, R. D. (1989). Addendum - Measurement of human variation: A two-stage model. In R. D. Bock (Ed.), *Multilevel Analysis of Educational Data*, 319-342. New York: Academic Press.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika. 46*, 443-459.

Cox, D. R., & Snell, E. J. (1989). *Analysis of Binary Data.* New York: Chapman and Hall.

Crocker, L. M., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* Fort Worth, T.X.: Harcourt Brace.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. *Journal of Educational Statistics, 13*, 243-271.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized Additive Models.* London: Chapman & Hall.

Heckman, N. E., & Ramsay, J. O. (2000). Penalized regression with model-based penalties. *The Canadian Journal of Statistics, 28*, 241-258.

Lord, F. M. (1980). *Application of Item Response Theory to Practical Testing Problems.* Hillsdale, N.J.: Lawrence Erlbaum.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models.* London: Chapman & Hall.

Mislevy, R. J., & Bock. R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [Computer Program]. Mooresville, IN: Scientific Software.

Phillips, G. M., & Taylor, P. J. (1996). *Theory and Applications of Numerical Analysis.* San Diego, C.A.: Academic Press.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611-630.

Ramsay, J. O. & Silverman, B. W. (1997). *Functional Data Analysis.* New York: Springer-Verlag.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47*,397-412.

Wang, X. (1993). *Combining the Generalized Linear Model and Spline Smoothing to Analyze Examination Data.* McGill University: Unpublished Masters Thesis.

Wingersky, M. S., Patrick, R., & Lord, F. M. (1988). LOGIST Users Guide. Princeton, N.J.: Educational Testing Service.