Combining the Generalized Linear Model and Spline Smoothing to Analyze Examination Data

Xiaohui Wang

Department of Mathematics and Statistics

McGill University, Montreal

November 1993

A Thesis submitted to the Faculty of
Graduate Studies and Research in partial fulfilment
of the requirements for the degree of M.Sc.

ISBN 0-315-94541-9

Combining the GLM and Spline Smoothing to Analyze Examination Data

To my grandmother

Abstract

This thesis aims to propose and specify a way to analyze test data. In order to analyze test data, it is necessary to estimate a special binary regression relation. called the item characteristic curve. The item characteristic curve is a relationship between the probabilities of examinees answering an item correctly and their abilities.

The statistical tools used in this thesis are the generalized linear models and spline smoothing. The method tries to combine the advantages of both parametric modeling and nonparametric regression to get a good estimate of the item characteristic curve. A special basis for spline smoothing is proposed which is motivated by the properties of the item characteristic curve. Based on the estimate of the item characteristic curve by this method, a more stable estimate of the item information function can be generated. Some illustrative analysis of simulated data are presented. The results seem to indicate that this method does have the advantages of both parametric modeling and nonparametric regression: it is faster to compute and more flexible than the methods using parametric models, for example, the three-parameter model in psychometrics, and on the other hand, it generates more stable estimate of derivatives than the purely nonparametric regression.

Résumé

Cette thèse a pour double objectif de proposer et de décrire une façon d'analyser les données provenant de tests. Dans le but d'analyser les données de tests, il s'avère nécessaire d'estimer une forme spéciale de régression binaire connue sous le nom de courbe caractéristique d'item. La courbe caractéristique d'un item exprime la relation entre la probabilité de répondre correctement à l'item et l'habileté des sujets.

Les instruments statistiques mis à profit dans cette thèse sont le modèle linéaire généralisé (GLM) et la fonction spline de lissage. La présente méthode tente de combiner à la fois les avantages de la modèlisation paramétrique et de la régression non-paramétrique afin d'obtenir un estimé plus fiable de la courbe caractéristique d'item. Etant donné les propriétés de la courbe caractéristique d'item une base spéciale pour la fonction spline de lissage est proposée. L'estimé de la courbe caractéristique d'item obtenu à l'aide de cette méthode produit un estimé plus stable de la courbe d'information d'item. Des analyses de données simulées sont présentées à des fins d'illustration. Les résultats semblent indiqués que cette méthode possède à la fois les avantages de la modèlisation paramétrique et de la régression non-paramétrique: elle s'avère moins longue à estimer, plus flexible que les méthodes basées sur des modèles paramétriques; et d'autre part, elle génère des dérivés plus stables que la méthode de régression non-paramétrique.

Acknowledgements

There are many people that deserve my heartfelt thanks. I only regret that I cannot mention them all by name here.

First, I want to thank my supervisor Professor James O. Ramsay for his care, diligence, guidance, constant encouragement, for providing me with financial support and for the careful reading of the manuscript. His comments improved my work considerably.

My thanks are also to Raymond Baillavgeon for his translation of the abstract from English to French.

I am greatly indebted to the Department of Mathematics and Statistics at McGill University, that provided me with a comfortable environments for my studies and thesis work.

I would like also to thank all other professors, staff and fellow students who helped me in any way in my work.

Last, but never at the least, I am very grateful to my parents for their encouragement.

Contents

1 Introduction 3

2 General Introduction to the Binary Response Model 6

2.1 Introduction to Binary Regression 6

2.2 Test Data 7

2.3 Generalized Linear Model (GLM) for Bernoulli Regression 9

2.4 Deviance: The Criterion for Fitting GLM 12

2.5 The GLM Algorithm 13

3 GLM and Nonparametric Regression 18

3.1 Combination of Nonparametric Regression and GLM 18

3.2 The Penalty Function $J(h)$ 19

3.3 Minimization of the Penalized Likelihood 21

4 Using Spline Smoothing to Estimate $h(\theta)$ 24

4.1 Function Space Defined by a Differential Operator L 24

4.2 Definition of Interpolating and Smoothing Spline 27

4.3 Computation of Smoothing Spline 30

5 Psychometric Theory 34

5.1 Notation 34

5.2 Background on Analyzing a Psychological or Educational Test . . 35

5.3 General Features of Item Characteristic Curves 36

5.4 Local Independence and Likelihood 38

5.5 Item Information Function 39

5.6 Test Information Function 41

5.7 The Three-Parameter Logistic Model 42

6 A Psychometric Model for $h(\theta)$ 44

6.1 A New Spline Model for $h(\theta)$ 44

6.2 Computing the Reproducing Kernels 49

6.2.1 Computing the Green's Function $g_0(\theta, w)$ for the Initial Value Constraint 51

6.2.2 Computing the Green's Function $g(\theta, w)$ for the Integral Constraint 52

6.2.3 Computing the Reproducing Kernel $k_2(\theta, t)$ 54

6.3 Computing the Operators Both in H_1 and H_2 54

6.4 Smoothing $P(\theta), (Dh)(\theta)$ and $I(\theta)$ 57

6.5 Summary of the Algorithm 57

7 Simulation, Discussion and Conclusion 60

7.1 Illustration with Simulated Data 60

7.2 Discussion and Conclusion 63

List of Figures

5.1 Six items responses of the real score from a final test of Introductory of Psychology. 37

5.2 One 3PL model with $P(\theta) = 0.2 + 0.8[e^{1.7(\theta+0.5)}/(1 + e^{1.7(\theta+0.5)})]$, the parameters a, b and c are equal to 1, -0.5 and 0.2 separately. . . 42

6.1 $P(\theta)$ generated by a 3PL model with $a = 1, b = 0, c = 0.2$. 45

6.2 $h(\theta)$ of 3PL model with $a = 1, b = 0, c = 0.2$. 46

6.3 $(Dh)(\theta)$ of 3PL model with $a = 1, b = 0, c = 0.2$. 47

7.1 Comparison between polynomial splines and psychometric splines by the mean square errors of $P(\theta), h(\theta), (Dh)(\theta)$ and $I(\theta)$ within simulated sample for different values of λ. 64

7.2 Comparison between polynomial splines and psychometric splines by mean square errors of $P(\theta), h(\theta), (Dh)(\theta)$ and $I(\theta)$ across simulated samples, and for different values of λ with the range of θ from -3 to 3. The solid line shows the result for the psychometric splines and the dotted line for the cubic splines 65

7.3 Comparison between polynomial splines and psychometric splines by mean square errors of $P(\theta)$, $h(\theta)$, $(Dh)(\theta)$ and $I(\theta)$ across simulated samples, and for different values of λ with the range of θ from -1.5 to 1.5. The solid line shows the result for the psychometric splines and the dotted line for the cubic splines 66

7.4 Comparison between the psychometric splines and GLM method with the basis $\{1, \theta, \ln(e^{\theta} + 1)\}$ for one simulated sample. P and h were estimated with $\lambda = 0.01$, and Dh and I were estimated with $\lambda = 0.1$. The solid line shows the true curve, the dotted line shows the estimate by the psychometric splines, and the dashed line shows the estimate by the 3 dimension linear regresion. 67

Chapter 1

Introduction

The purpose of exams is to figure out students' abilities. Therefore, it is very important to find a way to assess whether an exam is good or not. To assess an exam, the data coming from an exam administration are analyzed by statistical methods. This is one of the most important fields in psychometrics, and is called item response theory. Item response theory is a theory about the probability of examinees choosing an option, and probabilities should have some relationship to the ability of examinees. Within item response theory, the ability of examinees is usually viewed as unidimensional, and denoted as θ. The probability of a specific response to an item is assumed to depend only on θ, and denoted as $P(\theta)$.

In this thesis, some statistical tools are applied to the problem of estimating the function $P(\theta)$: the generalized linear model and spline smoothing adapted to this special problem. The goal is to combine the advantages of both. In fact, this approach is a combination of a parametric model and nonparametric regression. The advantages of such a combination are of:

1. rapid computation because the algorithm in the generalized linear model based on the scoring method converges rapidly,

2. more flexibility because it allows both parametric model and unspecified components to contribute to the estimate, and

3. a stable and efficient estimate of derivatives.

The detail of these advantages will be discussed in the later chapters.

The thesis aims only to propose and specify a method of analysis, although a small simulation study is included to indicate the feasibility of the method. A detailed evaluation of the method is left to further research.

For simplicity, only binary data are considered in this thesis. That means one only considers whether the item is answered correctly or not, i.e. that there are two responses to each item. The modelling of binary data requires an estimate of the binary regression of these data on the values of ability θ.

Chapter 2 recalls the basic knowledge of generalized linear models, especially on how to use it for binary regression. It introduces the main features of generalized linear models, and reviews the generalized linear model algorithm for calculating the estimated function $P(\theta)$ using the deviance criterion defined by Nelder and Wedderburn (1972).

Chapter 3 describes how to combine generalized linear models and nonparametric regression together with the penalized likelihood function.

Chapter 4 describes a special method of nonparametric regression, spline smoothing. It gives details of its use and techniques of computation.

Chapter 5 is an introduction to background of psychometric theory. It introduces the main features of item characteristic curve, which describes the relationship between the probability of an examinee choosing a correct answer and the ability of an examinee. Some assumptions and concepts are also introduced in this Chapter.

The main contributions of this thesis are in Chapter 6 and Chapter 7. In Chapter 6, a new basis adapted to this special problem for spline smoothing is put forward. The details of a modified algorithm based on the generalized linear model algorithm are described to show how to estimate binary regression.

Some small simulations are carried out in Chapter 7. Comparisons are made among the ordinary generaliz?d linear model, generalized linear model with cubic polynomial spline and the new method.

Chapter 2

General Introduction to the Binary Response Model

2.1 Introduction to Binary Regression

In the real world, there are many examples of observations only taking two possible states, for convenience denoted by 0 and 1. For example, to test a new medicine, a patient has either recovered, response denoted by 1, or not, response denoted by 0. If such a response is denoted by a random variable, Y, it can be expressed in the following way,

$$\mathrm{E}(Y) = \Pr(Y = 1) = P, \quad \Pr(Y = 0) = 1 - P \tag{2.1}$$

This means for a certain patient, the probability of his recovering from his illness now is P, the probability of his failure in recovering is $1 - P$. This random variable is said to follow a Bernoulli distribution with parameter P. And these kind of data are called *binary data.*

Generally, Y is called the response variable. There may be other variables used to explain or predict the variation of response variable Y, which are called the explanatory variables or covariates. Such a covariate value will be indicated by θ. Variable θ can be a vector.

Given data $\{\theta_i, y_i\}, i = 1, \ldots, n$, the purpose is to build up a relationship between $P(\theta)$ and θ, which is assumed known,

$$E(Y|\theta) = P(\theta).$$

Function $P(\theta)$ is called the binary regression function.

2.2 Test Data

The binary response of primary interest in this thesis comes from testing data. When an examinee takes an exam, his response to each item can also be denoted by 1 or 0. If he answers correctly, it is denoted by 1, otherwise, denoted by 0. Now P is the probability of his choosing a correct answer.

Assume there are n examinees taking an exam. For a certain item, an examinee's response can be viewed as a binary random variable if one is only interested in whether the response is correct or not. Because the probability of different examinees choosing a correct answer is different, this difference will be explained by examinees' ability. Ability, denoted by θ, is thought to be the total knowledge of examinees with respect to such an item. Usually θ is taken to be one dimensional and evaluated by a real number. Therefore each θ corresponds to a $P(\theta)$ which is the probability of a correct response conditional on θ. The relationship between $P(\theta)$ and θ is called as the item response model, and such a $P(\theta)$ is called item characteristic curve (ICC).

A good item should indicate the difference between examinees. This means that examinees with low ability should have a low probability to get a correct answer, and on the other hand, examinees with high ability should have more chance to answer correctly.

There are three main features of an ICC of interest.

1. A guessing level. Although $P(\theta)$ is bounded by 0 and 1, the lower bound of $P(\theta)$ is usually greater than 0. For example, considering an item for a multiple choice exam with M options, if an examinee knows nothing about this item, his probability of choosing a correct answer may be equal to $1/M$ by a random selection rule. The guessing level can be referred to as $P(-\infty)$ for a certain item.

2. The difficulty of an item for the whole group of examinees is also a very important concept. There are two ways to estimate this difficulty. One considers probability, the other considers ability. If one knows the average of the ability of this whole group, the probability of choosing a correct answer with respect to such a value of ability is a measurement of the difficulty for this item. If this value of probability is lower, it suggests that this item is more difficult for this group, otherwise, it means this item is easier. Another way is to calculate the average between the guessing level and 1 and find the value of θ corresponding to this average. The higher the value of θ, the easier this item.

3. One needs to define a quantity to measure the sensitivity of $P(\theta)$. Usually, the change in $P(\theta)$ over a standard small change of θ will measure it. Such a quantity is defined as *item discrimination* or *item discriminability.* If one has a differentiable ICC, the first derivative of ICC measures the sensitivity of $P(\theta)$, and the maximum value of the first derivative of $P(\theta)$ can be useful to measure the overall discriminating power of an item.

2.3 Generalized Linear Model (GLM) for Bernoulli Regression

Generalized linear models, introduced by Nelder and Wedderburn (1972), are statistical models for the analysis of data from exponential families. This is an extension of classical linear models, for classical linear models only allow data to come from normal distribution, which is obviously one special case of an exponential family distribution.

The data are assumed to come from an exponential family in the form

$$\{\theta_i, y_i\},\ i = 1, 2, \ldots, n,$$

in which y_i's are independent observations conditional on covariate values θ_i's. Let Y be the random variable taking values of y which can be equal to any y_i. According to the exponential family, the density function of Y has the form

$$f(y|h, \phi) = \exp[(yh - b(h))/a(\phi) + c(y, \phi)], \tag{2.2}$$

for some specific functions of a, b and c. Parameter h is called the canonical parameter. Because it is always a function of the mean, it is also called the location parameter. Parameter ϕ is a known scale parameter for a specific exponential family. Mean and variance of Y can be derived by

$$\mathrm{E}(Y) = \mu = (Db)(h) = \frac{db}{dh},$$

$$\mathrm{Var}(Y) = V = (D^2 b)(h)a(\phi) = \frac{d^2 b}{dh^2}.$$

For the Bernoulli distribution, $\phi = 1$. Consequently, the density of Y has the form

$$f(y|P) \ = \ P^y(1 - P)^{1-y}$$

$$= \exp\left(y \ln \frac{P}{1-P} - \ln \frac{1}{1-P}\right).$$

Comparing with (2.2), h can be written as

$$h = \ln \frac{P}{1-P},$$

and functions a, b and c for Bernoulli distribution are

$$a(\phi) = 1,$$

$$b(h) = \ln(e^h + 1),$$

$$c(y, \phi) = 0.$$

Therefore the density function of the Bernoulli distribution becomes

$$f(y|h) = \exp[yh - \ln(e^h + 1)].$$

Also mean and variance have the form

$$\mathrm{E}(Y) = \mu = e^h/(e^h + 1),$$

$$\mathrm{Var}(Y) = V = e^h/(e^h + 1)^2 = \mu(1 - \mu).$$

There are many approaches to study the dependency of $P(\theta)$, the probability of a binary response variable Y, on an explanatory variable value θ. The approach in this thesis is to use fairly simple empirical functions that express $P(\theta)$ in terms of θ and some unknown parameters in a reasonably flexible way, such that the parameters have a clear interpretation and preferably such that the resulting statistical analysis is straightforward.

But $P(\theta)$ must satisfy the following condition,

$$0 \le P(\theta) \le 1. \qquad (2.3)$$

This may cause some difficulties in fitting models directly, and one needs to find a way to fit models in which constraint (2.3) is automatically satisfied.

It was shown in McCullagh and Nelder (1989) that each exponential family distribution has a special transformation, which is a function of mean μ, and for which there exists a sufficient statistic, which is assumed to be a linear combination of some covariates $X' = (x_1, \ldots, x_p)$. This transformation will be called a *link function*, denoted by η, and it occurs for the Bernoulli distribution when $\eta(\mu) = \ln[\mu/(1-\mu)]$. Recall that

$$h(\theta) = \ln \frac{P(\theta)}{1 - P(\theta)},$$

therefore η maps the interval $(0,1)$ onto the whole real line and builds up a relationship between $P(\theta)$ and the location parameter $h(\theta)$. It follows that

$$h(\theta) = (\eta \circ P)(\theta) = \ln \frac{P(\theta)}{1 - P(\theta)},$$

which is also called *logit* transformation.

This suggests estimating $h(\theta)$ instead of estimating $P(\theta)$ directly. Notice that the original problem is trying to estimate $P(\theta)$ in terms of the covariate θ, so $X' = (x_1, \ldots, x_p)$ must be functions of θ, i.e. $X'(\theta) = (x_1(\theta), \ldots, x_p(\theta))$.

It is clear that the generalized linear model for the Bernoulli distribution has the following three-part specification:

1. The *random component*: the components of Y have independent Bernoulli distribution with $\mathrm{E}(Y) = P = \mu$.

2. The *link function*: between the random and systematic components:

$$\eta(P) = \ln[P/(1-P)].$$

3. The *systematic component*: covariates $X' = (x_1(\theta), \ldots, x_p(\theta))$ produce a linear predictor η given by

$$h(\theta) = (\eta \circ P)(\theta) = X'\beta = \sum_{k=1}^{p} x_k(\theta)\beta_k.$$

In this formulation, generalized linear models allow two extensions from classical linear models: first the distribution in the random component 1 may come from any exponential family other than the normal; such as in this situation where it comes from a Bernoulli distribution, and secondly the link function in component 2 may become any monotonic differentiable function.

The log likelihood may be written in two forms

$$\begin{aligned} l(P(\theta), y) &= \sum_{i=1}^{n}\{y_i h(\theta_i) - \ln[e^{h(\theta_i)} + 1]\} \\ &= \sum_{i=1}^{n}\left\{y_i \ln\left[\frac{P(\theta_i)}{1 - P(\theta_i)}\right] + \ln[1 - P(\theta_i)]\right\} \\ &= \sum_{i=1}^{n} l_i(P(\theta_i), y_i), \end{aligned} \tag{2.4}$$

where

$$\begin{aligned} l_i(P(\theta_i), y_i) &= y_i h(\theta_i) - \ln[e^{h(\theta_i)} + 1] \\ &= y_i \ln\left[\frac{P(\theta_i)}{1 - P(\theta_i)}\right] + \ln[1 - P(\theta_i)]. \end{aligned}$$

2.4 Deviance: The Criterion for Fitting GLM

Nelder and Wedderburn (1972) defined *deviance* to measure the discrepancy or goodness of fit in terms of the logarithm of likelihood ratio.

There are two extreme models which can be considered. One is the model just with one parameter, representing a common μ for all the y_i's which is called the *null model.* The other, *saturated model*, contains n parameters, one per observation, and μ's derived from it match the data exactly. Thus the null model consigns all the variation between the y_i's to the random component, but the saturated model leaves none for the random component.

In practice, the null model is too simple to give a systematic knowledge about the data, and the saturated model is meaningless because it does not summarize the data but merely repeats them in full. However, it suggests that the fitted model should be between these two models and with the number of parameters between 1 and n.

Denote all the values of y by a vector $\mathbf{y}$. It is convenient to express the log likelihood in terms of the mean-value parameter μ rather than the link function η. The maximum likelihood achievable in a saturated model with n parameters is $l(\mathbf{y}; \mu = \mathbf{y})$, which is ordinarily finite. The discrepancy of a fit is proportional to twice the difference between the maximum log likelihood achievable and that achieved by the model under investigation. If one denotes by $\hat{\eta} = \eta(\hat{\mu})$ under the estimated model and $\tilde{\eta} = \eta(\mathbf{y})$ under the saturated model, the discrepancy can be written as

$$\sum 2[y_i(\tilde{\eta}_i - \hat{\eta}_i) - b(\tilde{\eta}_i) + b(\hat{\eta}_i)] = D(\mathbf{y}, \hat{\mu})$$

$D(\mathbf{y}, \hat{\mu})$ is known as *deviance* for the current model and is a function of the data only.

The form of deviance for the Binomial distribution follows:

$$D(\mathbf{y}, \hat{\mu}) = 2\sum_{i=1}^{n}\{y_i \ln(y_i/\hat{\mu}_i) + (1 - y_i)\ln[(1 - y_i)/(1 - \hat{\mu}_i)]\}.$$

2.5 The GLM Algorithm

The algorithm for fitting GLM can be defined by considering the problem of fitting a single observation, for which the log-likelihood is

$$l = yh - b(h).$$

The maximum likelihood estimate $\hat{\beta}$ satisfies the equation,

$$\frac{\partial l}{\partial \beta}\Big|_{\hat{\beta}} = 0$$

It is shown by McCullagh and Nelder (1989) that such an estimate of the parameter β in the linear predictor η can be obtained by iterative weighted least squares. Considering one element β_j of β, the $\frac{\partial l}{\partial \beta_j}$ can be expressed as the following by the chain rule,

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \mu}\frac{\partial \mu}{\partial \eta}\frac{\partial \eta}{\partial \beta_j}.$$

Since

$$\frac{\partial l}{\partial \mu} = \frac{\partial l}{\partial h} \Big/ \frac{\partial \mu}{\partial h},$$

and

$$\frac{\partial l}{\partial h} = y - \mu,$$

note that $(Db)(h) = \mu$, and define $V = (D^2 b)(h) = (D\mu)(h)$. From $\eta = \sum \beta_j x_j$, it follows that,

$$\frac{\partial l}{\partial \beta_j} = (y - \mu)\frac{1}{V}\frac{d\mu}{d\eta}x_j.$$

With the definition of the quadratic weight

$$W = (V\frac{d\eta}{d\mu})^{-2}V, \tag{2.5}$$

$\frac{\partial l}{\partial \beta_j}$ becomes,

$$\frac{\partial l}{\partial \beta_j} = W(y - \mu)\frac{d\eta}{d\mu}x_j. \tag{2.6}$$

Therefore the maximum likelihood equation for β is given by

$$\sum W(y-\mu)\frac{d\eta}{d\mu}x_j = 0.$$

Using the Fisher's scoring method, the updated $\beta^* = \hat{\beta} + \delta b$ is defined by the solution to the linear equation

$$\frac{\partial l}{\partial \beta} = H(\hat{\beta})\delta b, \tag{2.7}$$

where H is the information matrix

$$H(\hat{\beta}) = -\mathrm{E}(D^2_{\beta} l)$$

and $\hat{\beta}$ is the current estimate of β. Consequently, using (2.6), the elements of H are

$$\begin{aligned}
H_{rs} &= -\mathrm{E}\left(\frac{\partial^2 l}{\partial\beta_r\partial\beta_s}\right) \\
&= -\mathrm{E}\sum\left[\frac{\partial l}{\partial\beta_s}\left(\frac{\partial l}{\partial\beta_r}\right)\right] \\
&= -\mathrm{E}\sum\left\{\frac{\partial l}{\partial\beta_s}\left[(y-\mu)W\frac{d\eta}{d\mu}x_r\right]\right\} \\
&= -\mathrm{E}\sum\left\{(y-\mu)\frac{\partial}{\partial\beta_s}\left(W\frac{d\eta}{d\mu}x_r\right) + W\frac{d\eta}{d\mu}x_r\frac{\partial}{\partial\beta_s}(y-\mu)\right\} \\
&= -\mathrm{E}\sum\left(-W\frac{d\eta}{d\mu}x_r\frac{\partial\mu}{\partial\beta_s}\right) \\
&= \sum W\frac{d\eta}{d\mu}x_r\frac{\partial\mu}{\partial\beta_s} \\
&= \sum W x_r\frac{\partial\eta}{d\beta_s} \\
&= \sum W x_r x_s.
\end{aligned}$$

The new estimate β^* which is obtained by adjusting β by δb satisfying equation (2.7), is defined to be

$$H\beta^* = H\beta + H\delta b = H\beta + \frac{\partial l}{\partial\beta}.$$

Now

$$(H\beta)_r = \sum_s H_{rs}\beta_s = \sum W x_r \eta.$$

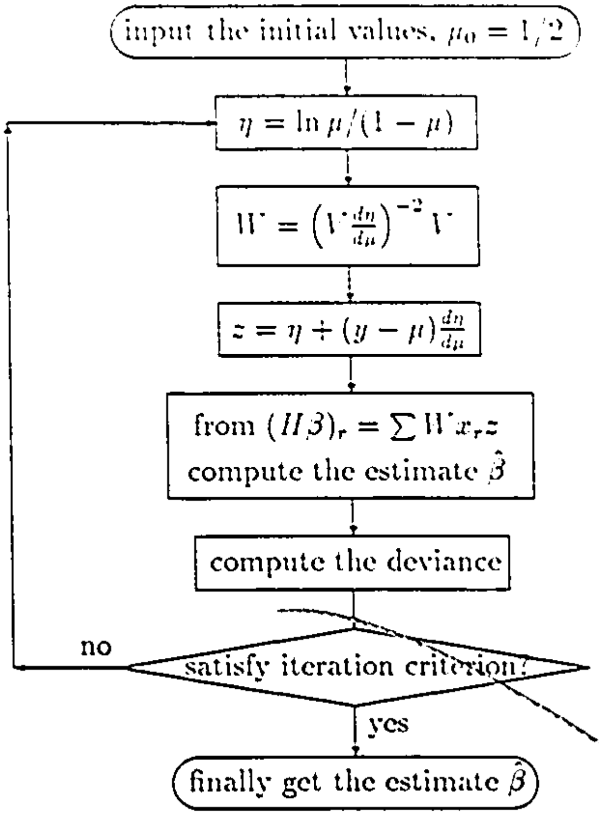
And

$$(H\beta^*)_r = \sum W x_r \left[\eta + (y - \mu)\frac{d\eta}{d\mu}\right]. \tag{2.8}$$

which is the equation of the form of linear weighted least squares with weight (2.5) and dependent variable

$$z = \eta + (y - \mu)\frac{d\eta}{d\mu}. \tag{2.9}$$

Consequently, one can see that in this regression, the dependent variable in a specific iteration is not y but z, a linearized form of the link function applied to y as in (2.9), and the weights are functions of the fitted values $\hat{\mu}$. The process is iterative because both the ***adjusted dependent variables*** z and the weight W depend on the fitted values, for which only current estimates are available. The procedure underlying the iteration is as follows:


input the initial values, $\mu_0 = 1/2$
$\eta = \ln \mu/(1 - \mu)$
$W = \left(V \frac{d\eta}{d\mu}\right)^{-2} V$
$z = \eta + (y - \mu)\frac{d\eta}{d\mu}$
from $(H\beta)_r = \sum W x_r z$
compute the estimate $\hat{\beta}$
compute the deviance
no
satisfy iteration criterion?
yes
finally get the estimate $\hat{\beta}$

Chapter 3

GLM and Nonparametric Regression

3.1 Combination of Nonparametric Regression and GLM

GLM models like other parametric models have stable curves, and stable derivatives. But they leave no room for an unspecified component of variation in the function $P(\theta)$. Consequently, these models can often fail to capture important features in the data. On the other hand, purely nonparametric regressions can have other disadvantages such as unstable derivatives.

In the GLM model described in Chapter 2, the mean is related to the linear predictor of variable Y or GLM regression surface via the transformation, which will be called h in the following chapters,

$$h_i = h(\theta_i) = (\eta \circ P)(\theta_i) = X'(\theta_i)\beta.$$

Therefore one can rewrite $P(\theta)$ as

$$P(\theta) = \frac{e^{h(\theta)}}{1 + e^{h(\theta)}}, \quad -\infty < \theta < \infty.$$

It is frequently easier to attempt to estimate h rather than P because h is in principle unbounded.

Function h will be estimated nonparametrically. Like Silverman (1978), Anderson and Blair (1982), Gu and Qiu (1993), and Gu (1993), function h will be estimated by the penalized likelihood method of Good and Gaskins (1971). The estimate of $h(\theta)$ is the function, $\widehat{h}_\lambda(\theta)$, that minimizes the penalized negative logarithm of the likelihood,

$$Q_\lambda(y_i, h(\theta_i)) = \frac{1}{n}\sum_{i=1}^{n} l_i(y_i, h(\theta_i)) + \lambda J(h), \tag{3.1}$$

where $l_i(y_i, h(\theta_i))$ is defined in (2.4). Function $J(h)$ is a penalty functional designed to incorporate prior notions, such as smoothness, about the behaviour of h. It will be discussed in detail below.

The estimate of $h(\theta)$ will depend on the parameter λ, which is called the *smoothing parameter.* Parameter λ controls the relative weighting of the penalty function in estimating h. When $\lambda \rightarrow 0$ the solution will be a function that maximizes the likelihood, whereas when $\lambda \rightarrow \infty$ the solution will be determined by the penalty function. Usually, a generalized cross-validation procedure is put forward for empirically assessing the parameter.

3.2 The Penalty Function $J(h)$

The transformation $h(\theta)$ is a function defined on the whole real line R. In practice, the domain of θ is always chosen as a subinterval of R, denoted by Ω. And the range of $h(\theta)$ is also the whole real line. The domain of the penalty functional, J, is all the possible functions h.

For any function $h(\theta)$ in a certain function space, some may be more smooth than others. The penalty functional, $J(h)$, is defined to be a measurement of smoothness of the function h with the appropriate properties required by a particular situation. From the mathematical view, the smoothness of a function means the existence of a certain number of derivatives. If a function has higher derivatives, it is said to be more smooth. Assume that one considers functions h with derivatives up to order m, $D^j h, j = 0, 1, \ldots, m$. In practice, it is useful to assume that $D^m h \in L^2$.

From the basic idea of smoothness, it is easy to understand why smoothness was defined classically in terms of D^m. Polynomials are groups of functions which are most smooth. For any polynomial p with order m, the smoothness of $h + p$ is considered to be just the same as that of h. Therefore, polynomials are regarded as *hyper-smooth*.

With the same idea of classical smoothness, one can consider a more general situation. Define a linear differential operator, L, as

$$L = w_0 D^0 + w_1 D^1 + \cdots + w_{m-1} D^{m-1} + D^m$$

where D^0 is the identity functional, and $w_0, w_1, \ldots, w_{m-1}$ may be functions of θ, which must satisfy certain conditions to be specified later. With this definition of L, a more general smoothness can be defined. Let $\ker L$ be a space containing these functions f which satisfy the condition $Lf = 0$. Any function f in such a space can be regarded as *hyper-smooth* in this general idea of classical polynomial smoothness. The advantage of such a procedure will be seen when such a special smoothness is used in Chapter 6. Therefore, smoothness is hereby defined as

$$J(h) = \int_{\Omega} (Lh)^2(\theta) d\theta = \|Lh\|^2_{L^2}. \tag{3.2}$$

3.3 Minimization of the Penalized Likelihood

To minimize the penalized likelihood, first consider how to find a $\hat{h}_\lambda$ which minimizes the penalized likelihood for a fixed value of λ.

The penalty functional, $J(h)$, becomes

$$J(h) = \int (Lh)^2(\theta)d\theta.$$

For simplicity, $\Omega = [-T, T]$ will be used in this thesis. If $\int$ is unspecified, $\int_{-T}^{T}$ will be implied. The penalized likelihood for generalized linear models can be written as

$$Q_\lambda(y_i, h(\theta_i)) = \frac{1}{n}\sum_{i=1}^{n} l_i(y_i, h(\theta_i)) + \lambda \int (Lh)^2(\theta)d\theta. \qquad (3.3)$$

It will be shown in Chapter 5 that $\hat{h}_\lambda$ is a linear combination of two sets of linear functions. The first is a set of m functions, $u_i, i = 1, \ldots, m$, which span the function space $\ker L$, and therefore generate all *hyper-smooth* functions. The second is a set of n functions $q_j, j = 1, \ldots, n$ corresponding to a set of n discrete values of θ.

Consequently, the linear component of the link function will be composed of $u_i, i = 1, \ldots, m$ and $q_j, j = 1, \ldots, n$. Let $x_i, i = 1, \ldots, p$ with $p = m + n$, instead of u_i and q_j, i.e.

$$x_i = u_i, i = 1, \ldots, m,$$

$$x_{j+m} = q_j, j = 1, \ldots, n,$$

then h can be written as

$$h(\theta) = \sum_{k=1}^{p} x_k(\theta)\beta_k = X'(\theta)\beta.$$

With the linear expression for h, one can rewrite the penalized negative likelihood (3.1) as

$$Q_\lambda(\beta) = \frac{1}{n}\sum_{i=1}^{n} l_i(y_i, X'\beta) + \lambda J(X'\beta).$$

From the definition of J in (3.2), let Σ_{rs} be

$$\Sigma_{rs} = \int (Lx_r)(\theta)(Lx_s)(\theta)d\theta.$$

Then

$$\begin{aligned} Q_\lambda(\beta) &= \frac{1}{n}\sum_{i=1}^{n} l_i(y_i, X'\beta) + \lambda\beta'\Sigma\beta \\ &= l_s(\beta) + \lambda\beta'\Sigma\beta. \end{aligned} \tag{3.4}$$

To find $\hat{h}_\lambda$ which minimizes the penalized negative likelihood (3.1) is equivalent to find $\hat{\beta}$ which minimizes (3.4).

Recall the algorithm in the last chapter. To find the maximum likelihood estimate of β is equivalent to find β which is the linear weighted least squares (2.8) with weight (2.5) and dependent variable (2.9). That means $\hat{\beta}$ minimizing (2.8) is the same as which will minimize

$$\frac{1}{n}\sum_{i=1}^{n} w_i(z_i - X'\beta)^2 + \lambda\beta'\Sigma\beta. \tag{3.5}$$

With the same definition of W and z in the last chapter, applying the Newton-Raphson minimization procedure with the GLM algorithm technique will lead to a way to find a sequence of approximations, $\{\beta^k\}$, which will converge to $\hat{\beta}$ when k tends to infinity. The information matrix H can be calculated in each iteration as

$$H(\beta^k) = E\{[\partial^2 l_s(\beta^k)/\partial\beta_r\partial\beta_s]|\eta_i = X'(\theta_i)\beta\},$$

and therefore

$$\beta^{k+1} = \beta^k - [H(\beta^k) + 2n\lambda\Sigma]^{-1} D_{\beta^k}(Q_\lambda).$$

The mean and variance functions, μ_i and V_i, are both evaluated as though $X'\beta^k$ were the true values of $h(\theta_i)$. It follows that this technique is equivalent to an iteratively weighted ridge regression procedure.

To assess the smoothing parameter λ, a standard way is to use the cross-validation or generalized cross-validation. The detail of this technique can be found in Wahba and Wold (1975) and Craven and Wahba (1979).

Chapter 4

Using Spline Smoothing to Estimate $h(\theta)$

4.1 Function Space Defined by a Differential Operator L

Let S be a function space in which for any function f, $D^m f \in L^2$. It is necessary to define a Hilbert space H_0 on S by defining an inner product. For details one can refer to Ramsay and Dalzell (1991).

Within the function space S, there is a subspace, $\ker L$, which contains all the hyper-smooth functions, i.e. for any functions, $f_1, f_2 \in \ker L$,

$$\int (Lf_1)(\theta)(Lf_2)(\theta)d\theta = 0.$$

Let hyper-smooth functions $u' = (u_1, \ldots, u_m)$ span $\ker L$. On the other hand, it is needed to keep the form $\int(Lh_1)(Lh_2)$ as part of the inner product of H_0 for any functions $h_1, h_2 \in H_0$ in order to measure the smoothness of functions not in $\ker L$.

To achieve this objective, a complementary *constraint operator* B is defined whose kernel space is exactly the function space for which $\int(Lh_1)(Lh_2)$ is an inner product. Such a kernel space with the constraint operator B must be orthogonal to $\ker L$ if it has an inner product of the form, $\int(Lh_1)(Lh_2)$. Therefore

1. $\ker L \bigcap \ker B = 0$.

2. $\ker L \oplus \ker B = S$.

There are many ways to define the constraint operator B. But however B is defined, there is a common property: $\ker B$ must be of dimension ∞ and codimension m. This can be achieved by letting B be a mapping from S to R^m. Two kinds of B introduced here are often used in practice. One is called the initial value constraint, and is

$$(Bh)(\theta) = \{h(0), (Dh)(0), \ldots, (D^{m-1}h)(0)\}^t.$$

The other is the integral constraint operator, with the form:

$$(Bh)(\theta) = \left\{\int u_1(\theta)h(\theta)d\theta, \int u_2(\theta)h(\theta)d\theta, \ldots, \int u_m(\theta)h(\theta)d\theta\right\}^t.$$

With a well defined constraint operator B, the inner product of $\ker B$ will have the form $\int(Lh_1)(Lh_2)$ and the inner product of $\ker L$ can be defined as $(Bh_1)^t(Bh_2)$. Denote $\ker L$ as H_1 and $\ker B$ as H_2, so that $H_0 = H_1 \oplus H_2$, with the inner product, for any $h_1, h_2 \in H_0$, defined to be

$$\begin{aligned} \langle h_1, h_2\rangle_0 &= \langle h_1, h_2\rangle_1 + \langle h_1, h_2\rangle_2 \\ &= (Bh_1)^t(Bh_2) + \int(Lh_1)(Lh_2)d\theta. \end{aligned} \qquad (4.1)$$

It will be assumed that the inner product spaces H_0, H_1 and H_2 are complete, and hence Hilbert spaces of functions (see Adams 1975 for a complete treatment of Sobolev spaces.)

With the partition of the Hilbert space of functions H_0, any function h in H_0 can also be partitioned into two functions according to the two orthogonal subspaces H_1 and H_2. Let

$$h = f + r,$$

where $f \in H_1$, $r \in H_2$. One can see that f and r satisfy $Lf = 0$ and $Br = 0$.

For each value of θ, there exists a function $k_0(\theta, \cdot)$ which represents the evaluation of h at θ, $h(\theta)$, as follows:

$$\langle k_0(\theta, \cdot), h(\cdot) \rangle = h(\theta), \quad \forall h \in H_0.$$

The function $k_0(\theta, \cdot)$ is known as a representor of this evaluation, and its existence and uniqueness are assured by the Riesz representation theorem and the continuity of the evaluation functional in S, (see Aubin 1979). Corresponding representor of evaluation in H_1 and H_2 also exist, and are denoted as $k_1(\theta, \cdot)$ and $k_2(\theta, \cdot)$. These representors viewed as bivariate functions are called *reproducing kernels*. Function $k_0(\theta, \cdot)$ is the reproducing kernel for H_0, $k_1(\theta, \cdot)$ for H_1 and $k_2(\theta, \cdot)$ for H_2. Since H_0 is partitioned into two complementary subspaces, $H_0 = H_1 \oplus H_2$, each subspace has its own inner product as well as the reproducing kernels, and $k_0 = k_1 + k_2$. The reproducing kernel, $k_0(\theta, \cdot)$ of the Hilbert space H_0, belongs to the space H_0 when either argument is fixed. These spaces are called *reproducing kernel Hilbert spaces*, abbreviated to RKHS. Real functions in a RKHS are smooth in the sense that two functions which are arbitarily close together in the sense defined by the inner product must also have values which are close.

4.2 Definition of Interpolating and Smoothing Spline

The reason why one needs to do smoothing is that within each iteration of the algorithm described in last chapter, at the beginning, one only has the discrete values of θ and the corresponding values of h. The real function of h is unknown. The aim is to find a smoothing function which captures the trend of the original data. It is true that there is no knowledge or control over what either the original function $\tilde{h}$ or the smoothing function h does between arguments values. It seems reasonable to ask that $\tilde{h}$ minimize the criterion

$$Q_\lambda(\mathrm{h},\tilde{\mathrm{h}}|w) = \frac{1}{n}(\mathbf{z}-\tilde{\mathrm{h}})'\mathbf{W}(\mathbf{z}-\tilde{\mathrm{h}}) + \lambda\|\tilde{h}\|^2 \qquad (4.2)$$

where $\mathbf{z} \in R^n$ and is defined as in (2.9), $\theta \in R^n$, $\tilde{\mathrm{h}} \in R^n$, with

$$\tilde{\mathrm{h}} = (\tilde{h}(\theta_1), \tilde{h}(\theta_2), \ldots, \tilde{h}(\theta_n)),$$

and $\mathbf{W}$ is the weight matrix, $\mathbf{W} \in R^{nn}$. The norm of $\|h\|$ is defined as

$$\|h\|^2 = \int (Lh)^2(\theta)d\theta.$$

In practice, one always deals with data with discrete values. These data can be viewed as a set of n function values, $h_i = h(\theta_i)$, $i = 1, \ldots, n$ corresponding to argument values θ_i, $i = 1, \ldots, n$. The interpolating function should be equal to the function values when evaluated at the corresponding arguments and have the norm $\|h\|$ as small as possible so that h is as conservative or cautious as one can make it between arguments. It can be shown that the minimum norm h is a linear combination of the n reproducing kernels $k_0(\theta_i, \cdot)$, $i = 1, \ldots, n$ (Wahba 1991),

$$h(\cdot) = \sum_{i=1}^{n} c_i k_0(\theta_i, \cdot).$$

However, it is unwise to fit a function that passes exactly through the data $\{\theta_i, h_i\}$ in statistics because there are always some observational errors. Therefore, to smooth the data means to fit a function which can capture the trend of the data and also be as close as possible at the same time. It can be shown that a smoothing function h minimizing the criterion (3.3) is of the form

$$h(\cdot) = \sum_{i=1}^{m} d_i u_i(\cdot) + \sum_{i=1}^{n} c_i k_2(\theta_i, \cdot). \tag{4.3}$$

This means that it is essential to calculate the reproducing kernels k_1 and k_2.

The reproducing kernel k_1 is easily computed. Defining the matrix $\mathbf{U}_1$ to be the order m symmetric matrix with values,

$$\langle u_i, u_j \rangle_1, \;\; i, j = 1, 2, \ldots, m,$$

then,

$$k_1(s, t) = u(s)' \mathbf{U}_1^{-1} u(t). \tag{4.4}$$

Computing the reproducing kernel k_2 is in general much more difficult, and details are given by Dalzell and Ramsay (1993). The remainder of the section gives the steps required to compute k_2.

In order to distinguish the initial value constraint and the integral constraint, let B denote the former and $\bar{B}$ denote the latter.

The first step involves computing the Green's function $g_0(s, w)$ for the initial value constraint. In the particular case where h is a real-valued function, the Green's function associated with the differential operator L and the initial value condition $B(h) = 0$ satisfies

$$\forall h \;\mid\; B(h) = 0, \; h(\theta) = \int g_0(\theta, w)(Lh)(w)dw.$$

In order to calculate the initial value Green's function $g_0(\theta, w)$, one needs first to calculate the Wronskian matrix. The Wronskian matrix is the order m matrix-valued function with elements

$$(D^{j-1}u_i)(\theta), i, j = 1, 2, \ldots, m.$$

It is assumed that $W(\theta)$ is nonsingular for all θ. The adjoint functions,

$$u^* = (u_1^*, u_2^*, \ldots, u_m^*),$$

are defined by $u^* = (W^{-1})'_m$ where $(W^{-1})_m$ is the m^{th} row of W^{-1}. The Green's function with the initial value condition is

$$g_0(\theta, w) = \begin{cases} u(\theta)'u^*(w) = \sum_{i=1}^m u_i(\theta)u_i^*(w), & 0 \leq w \leq \theta \\ -u(\theta)'u^*(w) = -\sum_{i=1}^m u_i(\theta)u_i^*(w), & \theta \leq w \leq 0 \\ 0, & \text{otherwise} \end{cases}$$

Once $g_0(\theta, w)$ is computed, one can modify g_0 to obtain $g(\theta, w)$, which is the Green's function associated with the differential operator L and the integral constraint $\tilde{B}h = 0$ satisfies

$$\forall h \mid \tilde{B}(h) = 0, \; h(\theta) = \int g(\theta, w)(Lh)(w)dw.$$

Define the vector valued functional b_0 by

$$b_0(w) = (\tilde{B}g_0)(\cdot, w).$$

Let order m matrix $\mathbf{B}^*$ have elements $b_i u_j$, where b_i is the i^{th} functional in B: that is $\mathbf{B}^* = Bu'$. If $\mathbf{B}^*$ is nonsingular and the functional b_i commutes with integration in H_0, then the Green's function for the condition $\tilde{B}h = 0$ is

$$g(\theta, w) = g_0(\theta, w) - u(\theta)'\mathbf{B}^{*-1}b_0(w). \qquad (4.5)$$

The proof can be found in Dalzell and Ramsay (1993).

The integral constraint Green's function can be derived simply from k_2 since it follows from $(Lk_2)(\theta, \cdot) = g(\theta, \cdot)$, where L is applied to k_2 as a function of w for a fixed θ. Conversely, the reproducing kernel k_2 can also be computed from g as,

$$\begin{aligned} k_2(\theta, t) &= \langle k_2(\theta, \cdot), k_2(t, \cdot)\rangle_2 \\ &= \int (Lk_2)(\theta, w)(Lk_2)(t, w)dw \\ &= \int g(\theta, w)g(t, w)dw. \end{aligned}$$

4.3 Computation of Smoothing Spline

Using the matrix notation, one can have a form of h similar to (4.3),

$$h = \mathbf{T}d + \mathbf{K}c$$

where $\mathbf{T}$ is a $n \times m$ matrix, with

$$t_{ij} = u_i(\theta_j), \quad i = 1, 2, \ldots, n, \; j = 1, 2, \ldots, m,$$

and $\mathbf{K}$ is the symmetric $n \times n$ matrix, with entries

$$k_{ij} = k_2(\theta_i, \theta_j), \quad i, j = 1, 2, \ldots, n.$$

By the above notation, Q_λ (4.2) can be written as:

$$Q_\lambda = \frac{1}{n}(h - \mathbf{T}d - \mathbf{K}c)'\mathbf{W}(h - \mathbf{T}d - \mathbf{K}c) + \lambda c'\mathbf{K}c. \tag{4.6}$$

To find vectors d and c which minimize the criterion Q_λ, the first step is to take the first derivatives of Q_λ according to d and c,

$$D_d Q_\lambda = -\frac{2}{n}\mathbf{T}'\mathbf{W}(h - \mathbf{T}d - \mathbf{K}c),$$

$$D_c Q_\lambda = -\frac{2}{n}\mathbf{K}\mathbf{W}(h - \mathbf{T}d - \mathbf{K}c) + 2\lambda\mathbf{K}c.$$

Letting these two equations equal to zero, and solving them, d and c will be found by,

$$\begin{cases} \mathbf{T}'\mathbf{W}(h - \mathbf{T}d - \mathbf{K}c) = 0 \\ \mathbf{K}\mathbf{W}(h - \mathbf{T}d - \mathbf{K}c) - n\lambda\mathbf{K}c = 0 \end{cases} \tag{4.7}$$

Assuming $\mathbf{K}$ is nonsingular and multiplying the second equation of (4.7) by $\mathbf{T}'\mathbf{K}^{-1}$, one gets

$$\mathbf{T}'c = 0. \tag{4.8}$$

So the stationary equations are now:

$$\mathbf{K}\mathbf{W}(h - \mathbf{T}d - \mathbf{K}c) = n\lambda\mathbf{K}c,$$

$$\mathbf{T}'c = 0.$$

If the weight matrix $\mathbf{W}$ is of full column rank, then from the first equation of (4.7) one can deduce the following,

$$\begin{aligned} d &= (\mathbf{T}'\mathbf{W}\mathbf{T})^{-1}\mathbf{T}'\mathbf{W}(h - \mathbf{K}c) \\ &= \mathbf{T}^{-}(h - \mathbf{K}c), \end{aligned}$$

where $\mathbf{T}^{-}$ is called the least squares generalized inverse of $\mathbf{T}$,

$$\mathbf{T}^{-} = (\mathbf{T}'\mathbf{W}\mathbf{T})^{-1}\mathbf{T}'\mathbf{W}.$$

Correspondingly,

$$\mathbf{T}d = \mathbf{T}\mathbf{T}^{-}(h - \mathbf{K}c) = \mathbf{P}(h - \mathbf{K}c), \tag{4.9}$$

where $\mathbf{P} = \mathbf{T}\mathbf{T}^{-}$ has the following properties: $\mathbf{P}^2 = \mathbf{P}$. Therefore, $\mathbf{P}$ is a projection matrix. By substituting (4.9) into the second equation of (4.7), one can see more clearly the properties of d and c,

$$\mathbf{K}\mathbf{W}[h - \mathbf{P}(h - \mathbf{K}c) - \mathbf{K}c] = \lambda n\mathbf{K}c.$$

Letting $\mathbf{Q} = \mathbf{I} - \mathbf{P}$, this equation can be written as

$$\mathrm{KWQ}(h - \mathrm{K}c) = \lambda n \mathrm{K}c. \tag{4.10}$$

Rewrite (4.10) as the following,

$$\mathrm{K}(\mathrm{WQK} + n\lambda \mathrm{I})c = \mathrm{KWQ}h. \tag{4.11}$$

and c can be obtained by

$$c = [\mathrm{K}(\mathrm{WQK} + n\lambda \mathrm{I})]^{-1}\mathrm{KWQ}h. \tag{4.12}$$

Now one can find both d and c which minimize the penalty likelihood Q_λ.

If we define $\mathbf{M}_\lambda = \mathbf{WQK} + n\lambda\mathbf{I}$, d is also able to be solved by

$$\begin{aligned} d &= \mathbf{T}^-[\mathbf{I} - \mathbf{K}(\mathbf{WQK} + n\lambda\mathbf{I})^{-1}\mathbf{WQ}]h \\ &= \mathbf{T}^-(\mathbf{I} - \mathbf{K}\mathbf{M}_\lambda^{-1}\mathbf{WQ})h, \end{aligned}$$

and with the result of d, c is solved by

$$c = \mathbf{M}_\lambda^{-1}\mathbf{WQ}h.$$

Therefore rewrite $h(\mathit{0})$ by the results of d and c,

$$\begin{aligned} \hat{h}(\mathit{0}) &= \mathbf{T}d + \mathbf{K}c \qquad (4.13) \\ &= \mathbf{P}[\mathbf{I} - \mathbf{K}(\mathbf{WQK} + n\lambda\mathbf{I})^{-1}\mathbf{WQ}]h + \mathbf{K}\mathbf{M}_\lambda^{-1}\mathbf{WQ}h \\ &= (\mathbf{P} + \mathbf{QK}\mathbf{M}_\lambda^{-1}\mathbf{WQ})h. \qquad (4.14) \end{aligned}$$

Here $\hat{h}$ is the smoothed values of h. The above expression of $\hat{h}(\mathit{0})$ in (4.14) means projecting h into two subspaces which are orthonormal. Since $\mathbf{P}$ is a projection matrix and $\mathbf{Q} = \mathbf{I} - \mathbf{P}$, these two orthonormal spaces combined together is not the whole space of H_0, except for $\lambda = 0$. More clearly, when $\lambda = 0$, $\mathbf{M}_\lambda$ is not a full rank matrix, if generalized inverse is applied here,

$$\mathbf{QKM}_\lambda^{-}\mathbf{WQ} = \mathbf{Q}.$$

now,

$$\begin{aligned}\hat{h}(\theta) &= \mathbf{P}h + \mathbf{Q}h \\ &= \mathbf{P}h + (\mathbf{I} - \mathbf{P})h \\ &= h_1 + h_2 \\ &= h\end{aligned}$$

where h_1 is the component in the space by the projection $\mathbf{P}$, denoting such a subspace by $\mathbf{P}_h$, and h_2 is the component of $\mathbf{Q}_h$. If $\lambda \neq 0$, $(\mathbf{QKM}_\lambda^{-}\mathbf{WQ})_h$ is a subspace of $\mathbf{Q}_h$. It is clear what a role λ plays. When λ becomes larger, the $\hat{h}$ becomes more smooth. Therefore, λ controls the degree of smoothness.

Chapter 5

Psychometric Theory

5.1 Notation

The purpose of this section is to introduce notation. In this thesis, the interest is only in the response to each item of an exam correct or not. These kind data are called dichotomous.

θ: This indicates the ability of examinees. It stands for the trait or skill of examinees, and is a latent or nonobservable variable.

n: This is the total number of levels for the ability, usually equal to the total number of examinees. It indicates how many levels there are in distinguishing the ability θ. Variable θ has levels $\theta_i, i = 1, \ldots, n$.

y_i: This is an indicator variable which is the response of an individual with i^{th} level of ability, i.e. $y_i = 1$ indicates this examinee getting a correct answer, $y_i = 0$, means his failure in choosing a correct answer.

$P(\theta_i)$: This is the probability of an examinee with the i^{th} level of ability choosing a correct answer, i.e.

$$P(\theta_i) = \Pr(y_i = 1|\theta_i)$$

The data always have this form $\{\theta_i, y_i\}, i = 1, \ldots, n$. Let Y be a random variable, following Bernoulli distribution with parameter $P(\theta)$,

$$P(\theta) = \Pr(Y = 1|\theta).$$

All the above notation is for only one specific item. If some information between items need to be considered, the following indices will be used:

N: This is the total number of items in an exam.

y_{ai}: This is an indicator variable to item a. It means the response of an examinee with the ability level θ_i. If $y_{ai} = 1$, meaning that the examinee with the ability level θ_i gets a correct answer of item a, if $y_{ai} = 0$, meaning that such an examinee gets a wrong answer of item a.

5.2 Background on Analyzing a Psychological or Educational Test

A psychological or educational test is a sample of behaviour. Usually the behaviour is quantified in some way to obtain a numerical score. Commonly, a test consists of separate items and the test score is a (possibly weighted) sum of item scores. In this case, statistics describing the test scores of a certain group of examinees can be expressed algebraically in terms of statistics describing the individual item scores for the same group.

In practical test development work, one needs to be able to describe the statistical and psychometric properties of any test that may be built. Also it is needed to describe the items by item parameters and the examinees by examinee parameters in such a way that one can describe as well as possible the response of any examinee to any item.

Furthermore, a main objective is to infer the examinee's ability level or skill which has some relationship with the testing scores. In order to do this, one must know something about how ability or skill determines the response to an item. The item response theory starts with a mathematical statement as to how response depends on the level of ability or skill θ. That means to find the relationships between the testing scores and the individual's real ability. This relationship is given by the *item response function*, also called the item characteristic curve (ICC). Throughout this thesis the ICC is simply the probability $P(\theta)$ of a correct response to the item conditional on ability level θ.

5.3 General Features of Item Characteristic Curves

Figure 5.1 displays six correct-answer item characteristic curves for six different items on a final exam in a course on Introductory of Psychology. This was actually a multiple choice test, consisting of 100 items, each with 4 options, and given to 379 examinees. The curves were estimated using the kernel smoothing approach of Ramsay (1991, 1993).

The first and perhaps most obvious aspect of the ICC is that it indicates the probability that an examinee of ability θ will get the item right. The low value of θ stands for the low ability and the high value of θ stands for the high ability. ICC, $P(\theta)$, tends to increase over whole range as θ increases, although it may decrease locally in some parts. If $P(\theta)$ is near zero, then we can conclude that this is a difficult item for an examinee with this ability level. Values near one indicate an easy item for an examinee with corresponding ability levels, and values near 0.5 suggest an item of intermediate difficulty.

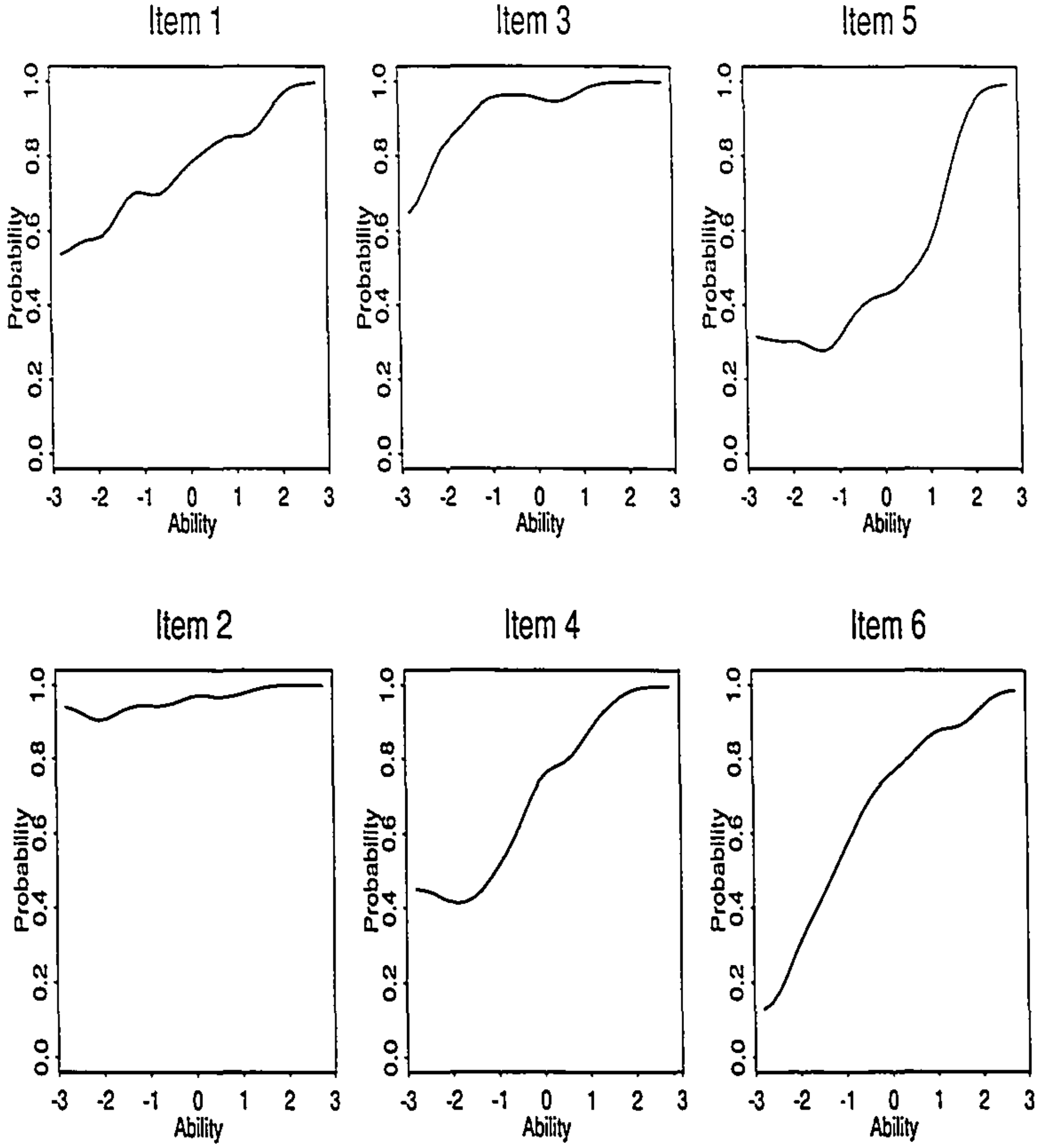


Figure 5.1: Six items responses of the real score from a final test of Introductory of Psychology.

In Figure 5.1, obviously item 2 is the easiest one because even the individual with a very low ability still has a high probability to get a correct answer. Item 1 and 3 are also very easy although they seem not as easy as item 2. There is a little difficulty to compare the item 4, 5 and 6. If one fixes his view on the average ability level, he may say, item 4 and item 6 are of the intermediate difficulty, and item 5 is tough.

From Figure 5.1, it can be seen that the probability of a correct response for examinees with extremely low ability values is not always zero. In fact, by using some random selection rules, every examinee can expect to average about a correct response one out of M tries. This probability is called as before, "guessing level".

5.4 Local Independence and Likelihood

In Chapter 2.2, a log-likelihood of all examinees within one item was introduced. If it is denoted by $l_a^{(1)}$, another kind of log-likelihood will be introduced here, denoted by $l_i^{(2)}$, which is the log-likelihood of all items for a certain examinee. And let L denote the whole log-likelihood of summary of $l_i^{(2)}$ over all examinees. One can think of the log-likelihood as a measure of how well the model fits the data, i.e. trying to use the maximum likelihood rule to estimate θ, such values of θ will give the largest value of L for all response sequences.

Now, $l_i^{(2)}$ can be written as:

$$l_i^{(2)} = \ln[\Pr(y_{ai} = 1, a = 1, \ldots, N|\theta_i)].$$

Local independence means that for a fixed level of θ, responses to two different items are independent. This is a very strong assumption, but if it holds, one can say that responses given or conditional on the information incorporated into such a model $P(\theta)$ are independent events.

Supposing that local independence really holds, then

$$\Pr(y_{ai} = 1, y_{bi} = 1|\theta_i) = \Pr(y_{ai} = 1|\theta_i)\Pr(y_{bi} = 1|\theta_i),$$

so

$$l_i^{(2)} = \sum_{a=1}^{N} \ln P(y_{ai} = 1|\theta_i) \tag{5.1}$$

for a fixed level of θ. Further,

$$L = \sum_{i=1}^{n}\sum_{a=1}^{N} \ln P(y_{ai} = 1|\theta_i) = \sum_{i=1}^{n} l_i^{(2)} = \sum_{a=1}^{N} l_a^{(1)}. \tag{5.2}$$

To sum up, with formula (5.1) and (5.2), one can get the maximum likelihood estimate of θ, denoted by $\hat{\theta}$.

5.5 Item Information Function

A good item discriminates highly for some range of ability values, and therefore the size of slope of $P(\theta)$, measured by derivatives $(DP)(\theta)$, is a useful indicator. A closely related quantity that is used for item scored as right or wrong is the *item information function*, $I_a(\theta)$,

$$I_a(\theta) = \frac{[(DP_a)(\theta)]^2}{P_a(\theta)[1 - P_a(\theta)]}, \tag{5.3}$$

where P_a is the item response function for a fixed item.

The amount of information given by an item varies with ability level θ. The higher the curve, the more the information. If one information function is twice as high as another at some particular ability level, then it will take two items of the latter type to measure as well as one item of the former type at that ability level.

From the expression of the item information function (5.3), one can see there are some problems in estimating $I_a(\theta)$ directly:

- Function $I_a(\theta)$ is a ratio of $[(DP_a)(\theta)]^2$ over $P_a(\theta)[1 - P_a(\theta)]$. Producing good estimates of ratios has been a classically difficult problem in statistics.
- To estimate $(DP_a)(\theta)$ is also to estimate a ratio.
- Sampling variation in DP_a leads to bias in estimating $(DP_a)^2$.
- When $P_a(\theta)$ is near 0 or 1, $I_a(\theta)$ becomes unstable.

A good procedure must be sought which can minimize these problems in estimating $I_a(\theta)$.

A good procedure should have the following three properties: flexible, fast to compute and with stable derivatives. Parametric models for item and option characteristic functions are sure to give nice smooth derivatives, so that item information function also looks nice. But their shape rigidity leaves room for doubt about whether they capture all the important features of a curve, and fitting them has not turned out to be easy. Completely nonparametric approaches, such as the kernel smooth process of Ramsay (1991), are certainly flexible and fast, but produce very ugly and unstable derivatives, and hence are of limited value in estimating information functions. The procedure combining GLM method and nonparametric regression is one satisfying these three properties. And with the logit transform $h(\theta) = \ln\{P(\theta)/[1 - P(\theta)]\}$, the item information function can be estimated by

$$I_a(\theta) = P_a(\theta)[1 - P_a(\theta)](Dh)^2(\theta). \tag{5.4}$$

Now the real problem is how to give a good estimate of $h(\theta)$ and the corresponding $(Dh)(\theta)$.

5.6 Test Information Function

A maximum likelihood estimator $\hat{\theta}$ of a parameter θ is asymptotically normally distributed with mean θ_0 (the unknown true parameter) and variance

$$\mathrm{Var}(\hat{\theta}|\theta_0) = \frac{1}{\mathrm{E}[(\frac{d\ln L}{d\theta})^2_{\theta_0}]} = \frac{1}{I(\theta)}.$$

where L is the log-likelihood function, see Lord (1990). When the item response model is known, and by the local independence, one has,

$$I(\theta) = \frac{1}{\mathrm{Var}(\hat{\theta}|\theta_0)} = \sum_{a=1}^{N} \frac{[(DP_a)(\theta)]^2}{P_a(\theta)[1 - P_a(\theta)]}. \qquad (5.5)$$

That means that the information function for an unbiased estimator of ability is the reciprocal of the sampling variance of the estimator. Equation (5.5) is of such importance that it is given a special name and symbol. It is called the *test information function* and is denoted simply by $I(\theta)$. The importance of the test information function comes partly from the fact that it provides an (attainable) upper limit to the information that can be obtained from the test, no matter what method of scoring is used.

A very important feature of (5.5) is that the test information function consists entirely of independent and additive contributions from the items. The contribution of an item does not depend on what other items are included in the test. The contribution of a single item is $(DP_a)^2/[P_a(1 - P_a)]$. Because in the later chapters, only the item information function is considered, the notation $I(\theta)$ is also used as the item information function.

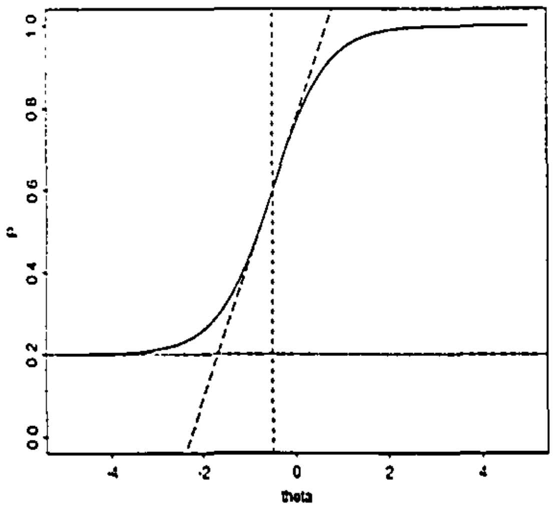


Figure 5.2: One 3PL model with $P(\theta) = 0.2 + 0.8[e^{1.7(\theta+0.5)}/(1 + e^{1.7(\theta+0.5)})]$, the parameters a, b and c are equal to 1, -0.5 and 0.2 separately.

5.7 The Three-Parameter Logistic Model

Whether calculating the log-likelihood or the item and test information function, we need to know the item response model $P(\theta)$. One of these models used very often in psychometrics is

$$P \equiv P(\theta) = c + (1 - c)\frac{e^{1.7a(\theta-b)}}{1 + e^{1.7a(\theta-b)}}.$$

which is called three-parameter logistic model (3PL). Figure 5.2 illustrates the characteristic of this model.

From Figure 5.2, one can see parameter c determines the guessing level, parameter b is the location point where the curve has its inflexion. This point is called the item difficulty. The more difficult the item, the further the curve is to the right. And the parameter a is proportional to the slope of the curve at the inflexion point. Thus parameter a represents the discrimination power of the item, the degree to which item response varies with ability levels.

The purpose of introduction of 3PL model is because in Chapter 7 the simulated data will be generated according to this model.

Chapter 6

A Psychometric Model for $h(\theta)$

6.1 A New Spline Model for $h(\theta)$

The method is used in this thesis to combine the generalized linear model and nonparametric regression to estimate the logit transformation $h(\theta)$ with the goal of obtaining a good estimate of the information function $I(\theta)$ as well as the ICC, $P(\theta)$.

Considering Figure 6.1, one can see that the main features of $P(\theta)$ which are generated by a 3PL model

$$P(\theta) = c + (1-c)\frac{e^{1.7a(\theta-b)}}{1+e^{1.7a(\theta-b)}}$$

are the following:

1. The curve has a lower asymptote of c, which is equal to guessing level, in this case, $c = 0.2$.

2. The function is monotone increasing with an upper asymptote of 1.

3. The point, $\theta = b$ is of intermediate difficult of the corresponding item, in this case, $b = 0$.

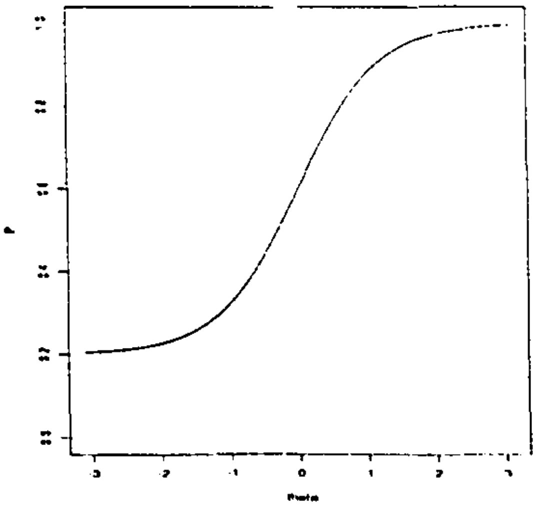


Figure 6.1: $P(\theta)$ generated by a 3PL model with $a = 1, b = 0, c = 0.2$.

4. Parameter, a, controls the slope of $P(\theta)$ at $\theta = b$. Here the slope is near $0.34 (= 0.415a(1 - c))$.

Figure 6.2 and Figure 6.3 show what $h(\theta)$ and $(Dh)(\theta)$ look like. It is also easy to see some properties of $h(\theta)$, such as that the curve of $h(\theta)$ is asymptotically linear in θ. For large negative values of θ the curve of $h(\theta)$ approaches the constant *logit(c)*, and the first derivative of $h(\theta)$ is asymptotically constant.

In order to approximate the special characteristics of $h(\theta)$, a basis for approximating $h(\theta)$ is chosen to be,

$$\{1, \theta, \ln(e^{a(\theta-b)} + 1)\} .$$

For any function f in the function space spanned by $\{1, \theta, \ln(e^{a(\theta-b)} + 1)\}$, we have

$$f(\theta) = c_1 + c_2\theta + c_3 \ln(e^{a(\theta-b)} + 1) \tag{6.1}$$

where a and b are taken as known or independently estimated. Functions $f(\theta)$

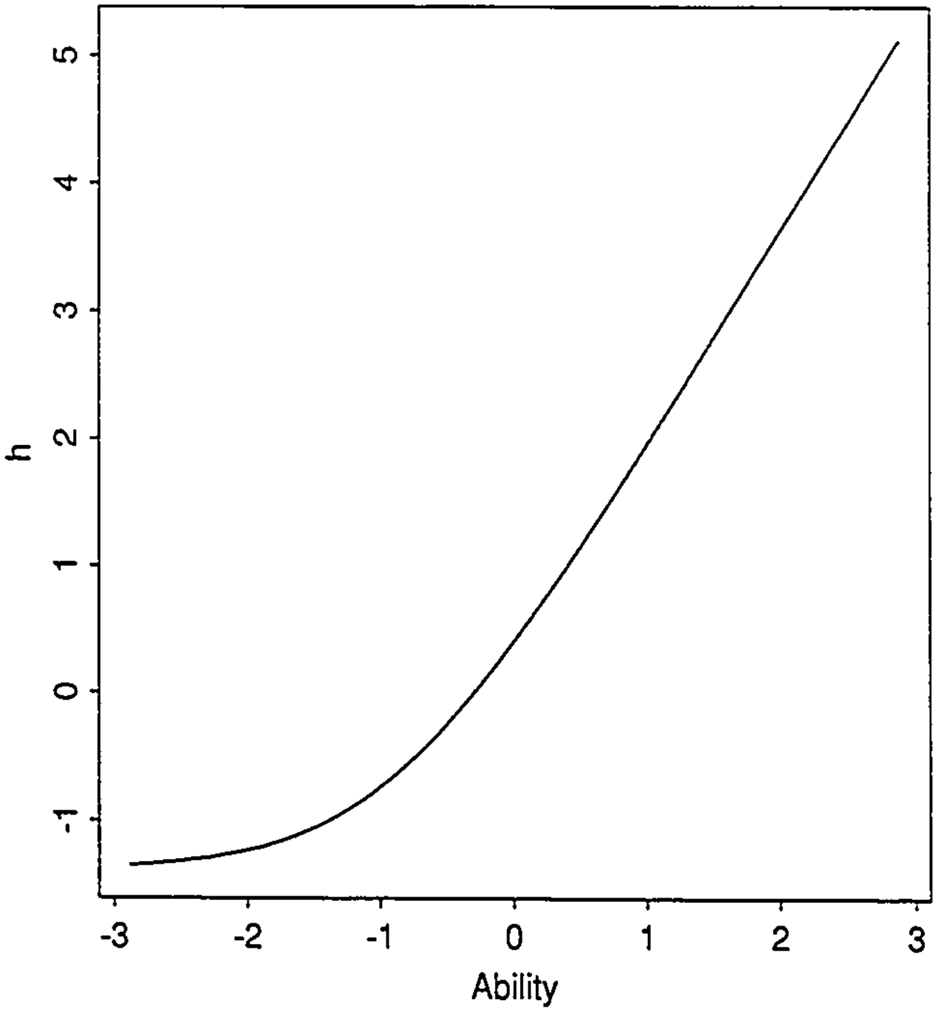


Figure 6.2: $h(\theta)$ of 3PL model with $a = 1, b = 0, c = 0.2$.

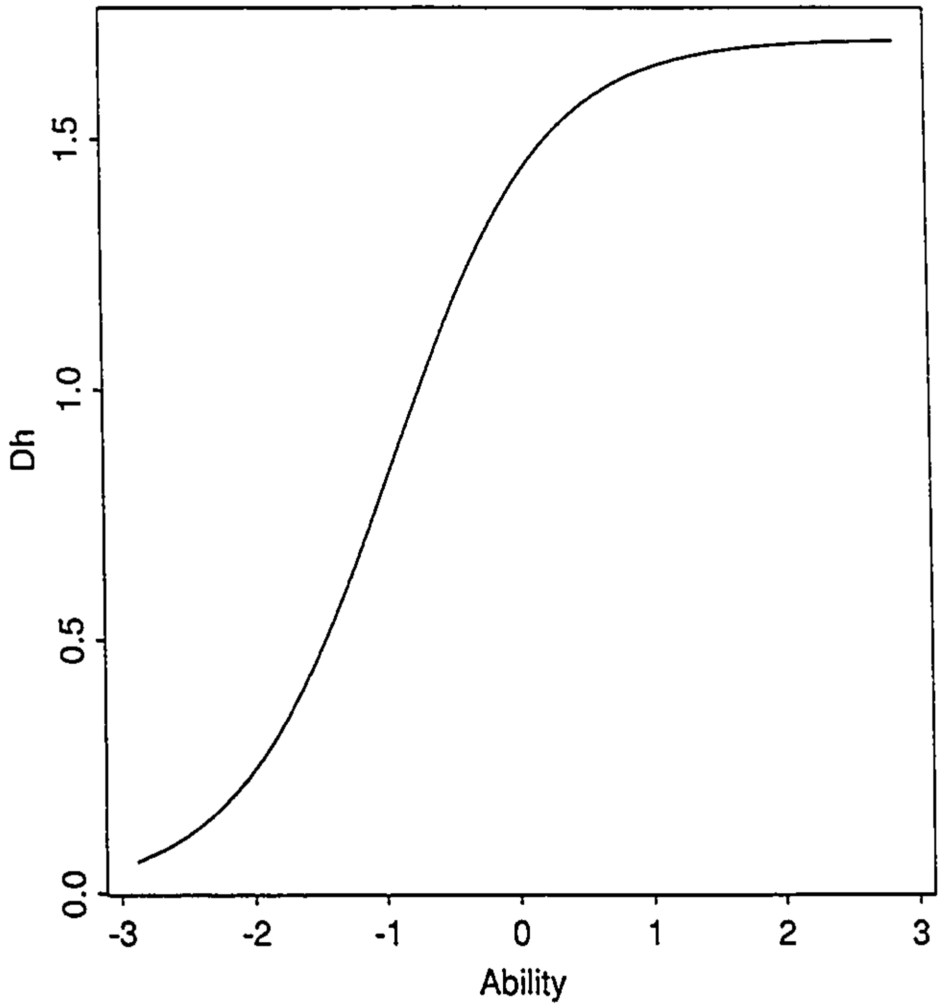


Figure 6.3: $(Dh)(\theta)$ of 3PL model with $a = 1, b = 0, c = 0.2$.

have the same properties as $h(\theta)$. For example, when θ has a large positive value, $\ln(e^{a(\theta-b)}+1)$ tends to equal θ. Therefore $f(\theta)$ is asymptotically linear in θ. And the first derivative of $f(\theta)$,

$$(Df)(\theta) = c_2 + c_3 a \frac{e^{a(\theta-b)}}{e^{a(\theta-b)}+1},$$

is asymptotically a constant near $c_2 + c_3 a$ for the large positive values of θ.

It would be naive to suppose that the function (6.1) is enough to describe $h(\theta)$. One will estimate $h(\theta)$ through a more generalized smooth function

$$h(\theta) = f(\theta) + r(\theta).$$

This first term f is the model component (6.1) and the second term r is a residual function which gives a more adequate approximation to the real curve. But one can only identify term r uniquely if it satisfies some suitable constraint that separates its contribution from that potentially provided by term f. Now let it satisfy the integral constraint,

$$(Br)(\theta) = \left\{ \int r(\theta)d\theta, \int \theta r(\theta)d\theta, \int \ln(e^{a(\theta-b)}+1)r(\theta)d\theta \right\}' = 0.$$

Define a differential operator L as

$$L = wD^2 - D^3$$

where

$$w = \frac{a(1-e^{a(\theta-b)})}{e^{a(\theta-b)}+1}.$$

Then $h(\theta)$ consists of two parts, $f(\theta)$ and $r(\theta)$, with $f(\theta) \in \ker L$ and $r(\theta) \in \ker B$. If let $H_0 = H_1 \oplus H_2$, with $H_1 = \ker L$ and $H_2 = \ker B$. Then H_0 is a Hilbert space for any h_1 and h_2 belonging to H_0 with inner product

$$\langle h_1, h_2 \rangle = \int h_1(\theta)d\theta \int h_2(\theta)d\theta$$

$$
\begin{aligned}
&+ \int \theta h_1(\theta)d\theta \int \theta h_2(\theta)d\theta \\
&+ \int \ln\left(\frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}}\right) h_1(\theta)d\theta \int \ln\left(\frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}}\right) h_2(\theta)d\theta \\
&+ \int (Lh_1)(\theta)(Lh_2)(\theta)d\theta. \qquad (6.2)
\end{aligned}
$$

6.2 Computing the Reproducing Kernels

With the same notation as in chapter 4.2. one can calculate the $k_1(\theta, t)$ and $k_2(\theta, t)$ for this specific model $h(\theta)$.

Now $m = 3$, and $u = \{u_1, u_2, u_3\}'$ is equal to $u(\theta) = \{1, \theta, \ln(e^\theta + 1)\}'$ respectively, i.e. it is assumed for convenience that $a = 1$ and $b = 0$.

First, look at how to compute the reproducing kernel $k_1(\theta, t)$. Recall the formula of how to calculate $k_1(\theta, t)$ in Chapter 4,

$$
\begin{aligned}
k_1(\theta, t) &= u(\theta)'\mathbf{U}_1^{-1}u(t) \\
&= (1, \theta, \ln(e^\theta + 1))\mathbf{U}_1^{-1} \begin{pmatrix} 1 \\ \theta \\ \ln(e^\theta + 1) \end{pmatrix}
\end{aligned}
$$

where $\mathbf{U}_1$ is a 3×3 matrix with $u_{ij} = \langle u_i, u_j \rangle$. And

$$
\begin{aligned}
\langle u_i, u_j \rangle &= \left(\int u_i(\theta)d\theta, \int \theta u_i(\theta)d\theta, \int \ln(e^\theta + 1)u_i(\theta)d\theta\right) \\
&\times \begin{pmatrix} \int u_j(\theta)d\theta \\ \int \theta u_j(\theta)d\theta \\ \int \ln(e^\theta + 1)u_j(\theta)d\theta \end{pmatrix}.
\end{aligned}
$$

Therefore three special one-dimension numerical integral were computed:

1. $\int \ln(e^\theta + 1)d\theta$,

2. $\int \theta \ln(e^\theta + 1)d\theta$,

3. $\int \ln^2(e^\theta + 1)d\theta$.

Using the Romberg quadrature algorithm (see Press, Teukolsky, Vetterling and Flannery 1993), the approximate values of $\mathbf{U}_1$ for $T = 3$ are given as follows

$$\mathbf{U}_1 = \begin{pmatrix} 72.56 & 54.42 & 102.84 \\ 54.42 & 405.00 & 261.07 \\ 102.84 & 261.07 & 238.73 \end{pmatrix}.$$

Therefore

$$\begin{aligned} k_1(\theta, t) &= (1, \theta, \ln(e^\theta + 1)) \begin{pmatrix} 72.56 & 54.42 & 102.84 \\ 54.42 & 405.00 & 261.07 \\ 102.84 & 261.07 & 238.73 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ t \\ \ln(e^t + 1) \end{pmatrix} \\ &= (1, \theta, \ln(e^\theta + 1)) \begin{pmatrix} 14.29 & 6.94 & -13.75 \\ 6.94 & 3.38 & -6.69 \\ -13.75 & -6.69 & 13.24 \end{pmatrix} \begin{pmatrix} 1 \\ t \\ \ln(e^t + 1) \end{pmatrix} \\ &= 14.29 + 6.94t - 13.75\ln(e^t + 1) + \theta(6.94 + 3.38t - 6.69\ln(e^t + 1)) \\ &\quad + \ln(e^\theta + 1)(-13.75 - 6.69t + 13.24\ln(e^t + 1)). \end{aligned}$$

One needs only compute the values of $k_2(\theta, t)$ on some fixed points. Since

$$k_2(\theta, t) = \int g(\theta, w) g(t, w) dw$$

and

$$g(\theta, w) = g_0(\theta, w) - u(\theta)' \mathbf{B}^{*-1} b_0(w),$$

computing $k_2(\theta, t)$ involves computing two-dimensional numerical integral. To summarize, the following three steps are required to compute the reproducing kernel $k_2(\theta, t)$:

1. compute the initial value constraint Green's function $g_0(\theta, w)$,

2. compute the integral constraint Green's function $g(\theta, w)$,

3. compute the reproducing kernel $k_2(\theta, t)$.

6.2.1 Computing the Green's Function $g_0(\theta, w)$ for the Initial Value Constraint

The following steps show how to calculate the Green's function with the initial value constraint:

- Calcvlate the Wronskian matrix W^*:

$$W^*(w) = \begin{pmatrix} 1 & 0 & 0 \\ w & 1 & 0 \\ \ln(e^w + 1) & \frac{e^w}{e^w+1} & \frac{e^w}{(e^w+1)^2} \end{pmatrix}.$$

- Calculate the inverse of the Wronskian matrix W^{*-1}:

$$W^{-1}(w) = \begin{pmatrix} 1 & 0 & 0 \\ -w & 1 & 0 \\ w(e^w + 1) - \frac{(e^w+1)^2 \ln(e^w+1)}{e^w} & -(e^w + 1) & \frac{(e^w+1)^2}{e^w} \end{pmatrix}.$$

  Therefore the adjoint function is written as

$$u^*(w)' = \left(w(e^w + 1) - \frac{(e^w + 1)^2 \ln(e^w + 1)}{e^w}, -(e^w + 1), \frac{(e^w + 1)^2}{e^w} \right).$$

- Calculate the $g_0(\theta, w)$, the Green's function with the initial value constraint:

$$g_0(\theta, w) = u(\theta)' u^*(w), \quad 0 \le w \le \theta.$$

  So the initial value Green's function is written as,

$$g_0(\theta, w) = \begin{cases} w(e^w + 1) - \frac{(e^w+1)^2 \ln(e^w+1)}{e^w} & \\ -\theta(e^w + 1) + \ln(e^\theta + 1)\frac{(e^w+1)^2}{e^w}, & 0 \le w \le \theta \\ -w(e^w + 1) + \frac{(e^w+1)^2 \ln(e^w+1)}{e^w} & \\ +\theta(e^w + 1) - \ln(e^\theta + 1)\frac{(e^w+1)^2}{e^w}, & \theta \le w \le 0 \\ 0 & \text{otherwise} \end{cases}$$

- To check the initial value Green's function $g_0(\theta, w)$: one uses the adjoint operator L^* of the operator L and apply it to the initial value Green's function. One has,

$$L^* = w_0 I + w_1 D + w_2 D^2 + D^3,$$

where,

$$w_0 = \frac{2e^w(e^w+1)}{(e^w+1)^3},$$
$$w_1 = -\frac{4e^w}{(e^w+1)^2},$$
$$w_2 = \frac{1-e^w}{e^w+1}.$$

When we apply it to the initial value Green's function, the result is showing,

$$L_w^* g(\theta, w) = 0, \quad \text{when} \ \ \theta \neq w.$$

This satisfies the basic property of the initial value Green's function.

6.2.2 Computing the Green's Function $g(\theta, w)$ for the Integral Constraint

To compute the Green's function with the integral constraint, one needs to use the Romberg quadrature algorithm again. Denote

$$H(\theta, w) = \begin{cases} 1, & \theta w \geq 0 \\ -1, & \theta w < 0 \end{cases}$$

Rewrite $g_0(\theta, w)$ as

$$g_0(\theta, w) = \text{sgn}(\theta) H(\theta, w) u'(\theta) u^*(w)$$

where $u(\theta)' = (1, \theta, \ln(e^\theta + 1))$, and

$$u^*(w)' = \left(w(e^w+1) - \frac{(e^w+1)^2 \ln(e^w+1)}{e^w} - (e^w+1), \frac{(e^w+1)^2}{e^w} \right).$$

Let $Z(\theta, w) = \text{sgn}(\theta) H(\theta, w)$, so

$$g_0(\theta, w) = Z(\theta, w) u'(\theta) u^*(w).$$

Define order 3 matrix functions

$$\mathbf{U}(w) = \int_w^T u(\theta) u'(\theta) d\theta, \ \ w \geq 0,$$

$$\mathbf{V}(w) = -\int_{-T}^{-w} u(\theta)u'(\theta)d\theta, \ w \geq 0.$$

The approximate values of $\mathbf{B}^* = \ddot{B}u'$ for $T = 3$ are

$$\mathbf{B}^* = \begin{pmatrix} 6 & 0 & 6.0468 \\ 0 & 18 & 9.0000 \\ 6.0468 & 9.0000 & 11.0077 \end{pmatrix}.$$

Define also

$$\mathbf{X}(w) = \mathbf{B}^{*-1}\mathbf{U}(w),$$

$$\mathbf{Y}(w) = \mathbf{B}^{*-1}\mathbf{V}(w).$$

Following the formula (4.5), one gets the Green's function of integral constraint condition

$$g(\theta, t) = \begin{cases} Z(\theta, w)u'(\theta)u^*(w) - u'(\theta)\mathbf{X}(w)u^*(w), & 0 \leq w \leq T, \\ Z(\theta, w)u'(\theta)u^*(w) - u'(\theta)\mathbf{Y}(w)u^*(w), & -T \leq w \leq 0. \end{cases}$$

By dividing $\theta - w$ plane into four regions, one can get a clearer expression of $g(\theta, w)$:

1. $0 \leq w \leq \theta \leq T$,

$$g(\theta, w) = u'(\theta)[\mathbf{I} - \mathbf{X}(w)]u^*(w),$$

2. $-T \leq \theta \leq 0 \leq w$ and $0 \leq \theta \leq w \leq T$,

$$g(\theta, w) = -u'(\theta)\mathbf{X}(w)u^*(w),$$

3. $-T \leq \theta \leq w \leq 0$,

$$g(\theta, w) = -u'(\theta)[\mathbf{I} + \mathbf{Y}(w)]u^*(w),$$

4. $-T \leq w \leq \theta \leq 0$ and $-T \leq w \leq 0 \leq \theta$,

$$g(\theta, w) = -u'(\theta)\mathbf{Y}(w)u^*(w).$$

6.2.3 Computing the Reproducing Kernel $k_2(\theta, t)$

Let $\mathbf{G}$ be the $l \times l$ matrix of values of $g(\theta, w)$ for l estimated values of θ and w. Denote $k_2(\theta, t)$ for these fixed points as a $l \times l$ matrix $\mathbf{K}$, the Simpson rule was used for the numerical calculation. Then

$$\mathbf{K} = \mathbf{G}'\mathbf{W}\mathbf{G},$$

where $\mathbf{W}$ is a diagonal weighted matrix by the Simpson rule,

$$\mathbf{W} = \text{diag}\left(\frac{1}{3}, \frac{4}{3}, \frac{2}{3}, \frac{4}{3}, \cdots, \frac{4}{3}, \frac{2}{3}, \frac{4}{3}, \frac{1}{3}\right).$$

6.3 Computing the Operators Both in H_1 and H_2

The next step after computing the reproducing kernels is to compute the d, c, coefficients of the two components in H_1 and H_2.

Depending on the analysis in Chapter 4.3, one can understand how spline smoothing works in this problem. But in practical computation, this method can not be used directly because it involves computing the inverse of a big matrix. Following are the details of computation using special matrix decompositions which make computation stable and efficient.

Since weight matrix $\mathbf{W}$ is a diagonal matrix and all the diagonal elements are positive, let $\mathbf{W} = \mathbf{W}^{\frac{1}{2}}\mathbf{W}^{\frac{1}{2}}$, i.e.

$$\mathbf{W}^{\frac{1}{2}} = \text{diag}\left(w_{11}^{\frac{1}{2}}, w_{22}^{\frac{1}{2}}, \cdots, w_{ll}^{\frac{1}{2}}\right),$$

so

$$Q_\lambda = \frac{1}{n}(\mathbf{W}^{\frac{1}{2}}h - \mathbf{W}^{\frac{1}{2}}\mathbf{T}d - \mathbf{W}^{\frac{1}{2}}\mathbf{K}c)'(\mathbf{W}^{\frac{1}{2}}h - \mathbf{W}^{\frac{1}{2}}\mathbf{T}d - \mathbf{W}^{\frac{1}{2}}\mathbf{K}c) + \lambda c'\mathbf{K}c.$$

Applying QR decomposition to $\mathbf{W}^{\frac{1}{2}}\mathbf{T}$, get,

$$\mathbf{W}^{\frac{1}{2}}\mathbf{T} = \mathbf{QR}$$

where $\mathbf{Q}'\mathbf{Q} = \mathbf{QQ}' = \mathbf{I}$ and $\mathbf{R}$ is an upper-triangular matrix. For $\mathbf{W}^{\frac{1}{2}}\mathbf{T}$ is a $n \times 3$ matrix, so

$$\mathbf{W}^{\frac{1}{2}}\mathbf{T} = (\mathbf{Q}_1, \mathbf{Q}_2)\begin{pmatrix} \mathbf{R} \\ 0 \end{pmatrix}$$

with $\mathbf{Q}_1 \in R^{n3}$, $\mathbf{Q}_2 \in R^{n(n-3)}$ and $\mathbf{R} \in R^{33}$. Obviously,

$$\mathbf{Q}_2'(\mathbf{W}^{\frac{1}{2}}\mathbf{T}) = 0.$$

Using $\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}$ and $\mathbf{R}$ instead of $\mathbf{W}^{\frac{1}{2}}\mathbf{T}$ in Q_λ, Q_λ becomes

$$\begin{aligned} Q_\lambda(h, \tilde{h}|w) &= \frac{1}{n}[\mathbf{Q}'(\mathbf{W}^{\frac{1}{2}}h - \mathbf{W}^{\frac{1}{2}}\mathbf{T}d - \mathbf{W}^{\frac{1}{2}}\mathbf{K}c)]' \\ &\times [\mathbf{Q}'(\mathbf{W}^{\frac{1}{2}}h - \mathbf{W}^{\frac{1}{2}}\mathbf{T}d - \mathbf{W}^{\frac{1}{2}}\mathbf{K}c)] + \lambda c'\mathbf{K}c \\ &= \frac{1}{n}(\mathbf{Q}_1'\mathbf{W}^{\frac{1}{2}}h + \mathbf{Q}_2'\mathbf{W}^{\frac{1}{2}}h - \mathbf{R}d - \mathbf{Q}'\mathbf{W}^{\frac{1}{2}}\mathbf{K}c)' \\ &\times (Q_1'\mathbf{W}^{\frac{1}{2}}h + \mathbf{Q}_2'\mathbf{W}^{\frac{1}{2}}h - \mathbf{R}d - \mathbf{Q}'\mathbf{W}^{\frac{1}{2}}\mathbf{K}c)'. \end{aligned}$$

define

$$Z_1 = \mathbf{Q}_1'\mathbf{W}^{\frac{1}{2}}h,$$

$$Z_2 = \mathbf{Q}_2'\mathbf{W}^{\frac{1}{2}}h,$$

then

$$Z_1 - \mathbf{R}d - \mathbf{Q}_1'\mathbf{W}^{\frac{1}{2}}\mathbf{K}c$$

and

$$Z_2 - \mathbf{Q}_2'\mathbf{W}^{\frac{1}{2}}\mathbf{K}c$$

are orthonormal, and thus Q_λ can be written as

$$
\begin{aligned}
Q_\lambda(h, \tilde{h}|w) &= \frac{1}{n}\|Z_1 + Z_2 - \mathbf{R}d - \mathbf{Q}_1'\mathbf{W}^{\frac{1}{2}}\mathbf{K}c - \mathbf{Q}_2'\mathbf{W}^{\frac{1}{2}}\mathbf{K}c\|^2 \\
&\quad + \lambda c'\mathbf{K}c \\
&= \frac{1}{n}\|Z_1 - \mathbf{R}d - \mathbf{Q}_1'\mathbf{W}^{\frac{1}{2}}\mathbf{K}c\|^2 + \frac{1}{n}\|Z_2 - \mathbf{Q}_2'\mathbf{W}^{\frac{1}{2}}\mathbf{K}c\|^2 \\
&\quad + \lambda c'\mathbf{K}c.
\end{aligned}
$$

Since $\mathbf{T}'c = 0$, c can be defined as

$$c = \mathbf{W}^{\frac{1}{2}}\mathbf{Q}_2 b,$$

i.e.

$$b = \mathbf{Q}_2'\mathbf{W}^{-\frac{1}{2}}c.$$

Therefore

$$
\begin{aligned}
Q_\lambda &= \frac{1}{n}\|Z_1 - \mathbf{R}d - \mathbf{Q}_1'\mathbf{W}^{\frac{1}{2}}\mathbf{K}\mathbf{W}^{\frac{1}{2}}\mathbf{Q}_2 b\|^2 + \\
&\quad \frac{1}{n}\|Z_2 - (\mathbf{Q}_2'\mathbf{K}\mathbf{W}^{\frac{1}{2}}\mathbf{Q}_2)b\|^2 + \lambda b'(\mathbf{Q}_2'\mathbf{W}^{\frac{1}{2}}\mathbf{K}\mathbf{W}^{\frac{1}{2}}\mathbf{Q}_2)b \\
&= \frac{1}{n}\|Z_1 - \mathbf{R}d - \mathbf{Q}_1'\mathbf{W}^{\frac{1}{2}}\mathbf{K}\mathbf{W}^{\frac{1}{2}}\mathbf{Q}_2 b\|^2 + \frac{1}{n}\|Z_2 - \mathbf{J}b\|^2 + \lambda b'\mathbf{J}b,
\end{aligned}
$$

where $\mathbf{J} = \mathbf{Q}_2'\mathbf{W}^{\frac{1}{2}}\mathbf{K}\mathbf{W}^{\frac{1}{2}}\mathbf{Q}_2$. Recall $d = \mathbf{T}^-(\mathbf{I} - \mathbf{K}\mathbf{M}_\lambda^{-1}\mathbf{W}\mathbf{Q})h$, using the decompositions above, d becomes

$$\mathbf{R}d = Z_1 - \mathbf{Q}_1\mathbf{W}^{\frac{1}{2}}\mathbf{K}\mathbf{W}^{\frac{1}{2}}\mathbf{Q}_2 b.$$

Because $\mathbf{R}$ is an upper-triangular matrix, d can be solved by back substitution. Also because both d and c depend on b, if b is found, the problem can be completely solved. To compute b will apply the eigendecomposition of $\mathbf{J}$ be $\mathbf{J} = \mathbf{U}\mathbf{D}^2\mathbf{U}'$. Since the value of d makes the first part of Q_λ equal to zero, so

$$Q_\lambda = \frac{1}{n}\|Z_2 - \mathbf{U}\mathbf{D}^2\mathbf{U}'b\|^2 + \lambda b'\mathbf{U}\mathbf{D}^2\mathbf{U}'b.$$

The problem of finding b is changed to find a equivalently. Use the similar method as before, let

$$D_b Q_\lambda = 0.$$

Solving this equation will give the solution

$$b = \mathbf{U}(\mathbf{D}^2 + n\lambda\mathbf{I})^{-1}\mathbf{U}'Z_2.$$

6.4 Smoothing $P(\theta)$, $(Dh)(\theta)$ and $I(\theta)$

Once a good smoothed function $h(\theta)$ is attained, the *anti-logit* transformation gives an estimate of $P(\theta)$. To get a smoothed $(Dh)(\theta)$, one can use the similar method as getting the smoothed $h(\theta)$. The difference is to use the $(Dk_1)_\theta(\theta, t)$ and $(Dk_2)_\theta(\theta, t)$ instead of $k_1(\theta, t)$ and $k_2(\theta, t)$.

Once smoothed $P(\theta)$ and $(Dh)(\theta)$ are available, using formula (5.4), the smoothed estimate of $I(\theta)$ can be easily computed.

6.5 Summary of the Algorithm

If one has the data in the form $\{\theta_i, y_i\}, i = 1, \ldots, n$, with $P(\theta)$ to be estimated, it is not necessary to compute the evaluated values of $P(\theta_i)$ for each point θ_i. For most purposes, estimating $P(\theta)$ at $\theta_\xi, \xi = 1, \ldots, l$, l is much smaller than n, is sufficient. The evaluation points θ_ξ can be equally spaced: because calculating the reproducing kernels does not involve the original data, one needs to compute $k_2(\theta, t)$ and $(Dk_2)_\theta(\theta, t)$ only once. The steps of this algorithm are as follows:

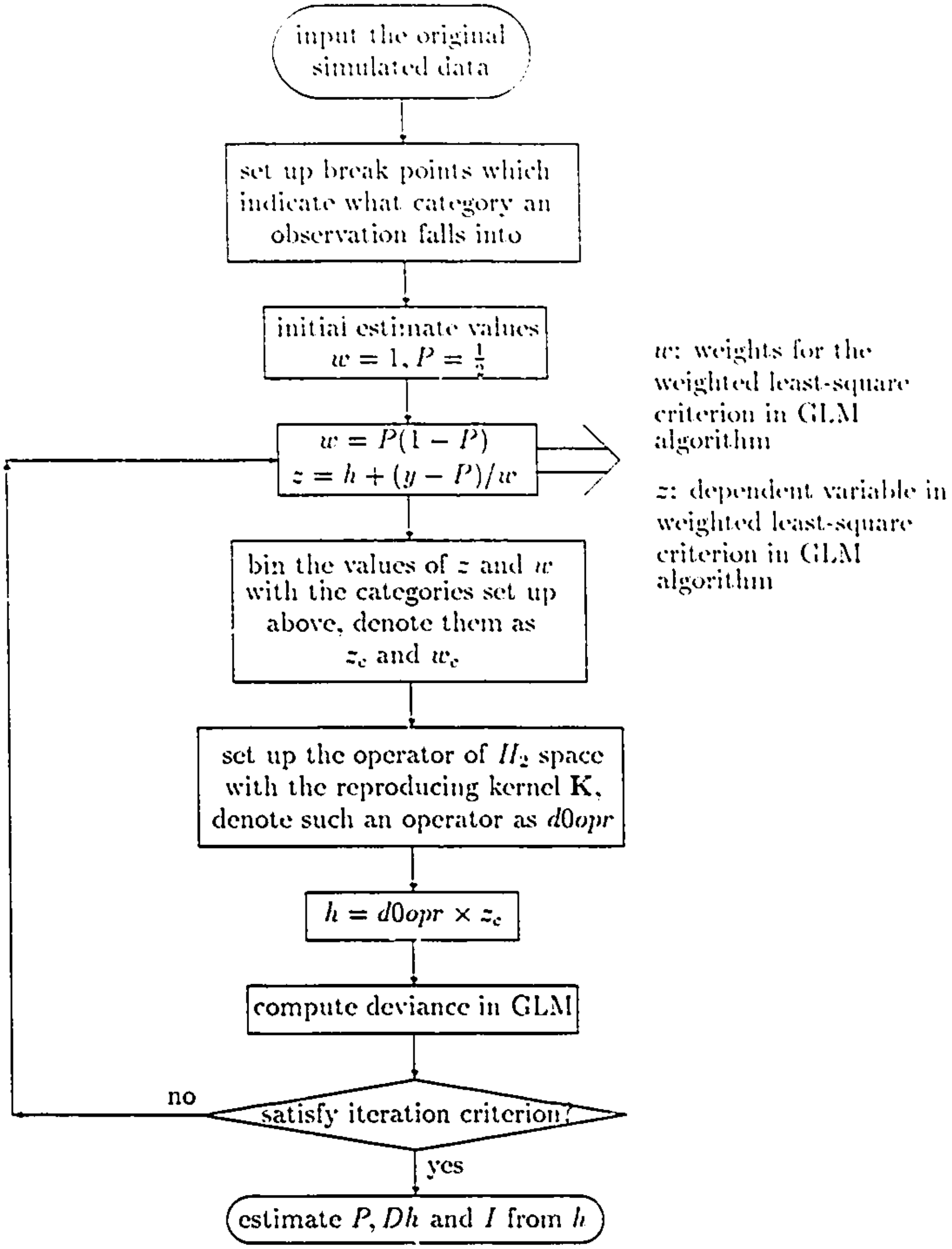
1. Compute $k_2(\theta, t)$ and $(Dk_2)_\theta(\theta, t)$ for points $\theta_\xi, \xi = 1, \ldots, l$. Denote them by notations **K** and **DK** respectively.

2. Set up the operator for Dh in H_2 space using **DK**.

3. With the reproducing kernel **K**, using the GLM algorithm, first set up the operator of h in H_2 space in each iteration, and at last get smoothed $h(\theta)$.

4. With the operator of Dh in H_2 space and the estimated $h(\theta)$, compute the smoothed $(Dh)(\theta)$ and further $P(\theta)$ and $I(\theta)$.

The following tries to illustrate the third step of this algorithm:


input the original simulated data
set up break points which indicate what category an observation falls into
initial estimate values $w = 1, P = \frac{1}{2}$
$w = P(1 - P)$
$z = h + (y - P)/w$
bin the values of z and w with the categories set up above, denote them as z_c and w_c
set up the operator of H_2 space with the reproducing kernel **K**, denote such an operator as $d0opr$
$h = d0opr \times z_c$
compute deviance in GLM
satisfy iteration criterion?
no
yes
estimate P, Dh and I from h
w: weights for the weighted least-square criterion in GLM algorithm
z: dependent variable in weighted least-square criterion in GLM algorithm

Chapter 7

Simulation, Discussion and Conclusion

7.1 Illustration with Simulated Data

The purpose of this tiny simulation is to illustrate the method described in the previous chapters. To illustrate the simulation, three methods were used:

1. GLM with a basis $\{1, \theta, \ln(e^{\theta} + 1)\}$.

2. GLM combining the spline smoothing with the basis $\{1, \theta\}$, or cubic spline smoothing.

3. GLM combining spline smoothing with a special basis $\{1, \theta, \ln(e^{\theta} + 1)\}$. This will be called psychometric spline smoothing.

The first method is very simple: use the linear regression to approximate $h(\theta)$ according to the basis $\{1, \theta, \ln(e^{\theta} + 1)\}$, then use the GLM method to estimate $P(\theta), (Dh)(\theta)$ and $I(\theta)$. The second combines GLM method with polynomial spline with a cubic polynomial basis. The third one combines GLM with the spline smoothing with a special basis $\{1, \theta, \ln(e^{\theta} + 1)\}$.

The simulated binary data were carried out for a number of values of θ typical of many applications, namely $n = 500$. The actual values of θ were the corresponding n quartiles of the standard normal distribution. Conditional on these θ's, the function $P(\theta), h(\theta), (Dh)(\theta)$ and $I(\theta)$ to be estimated were generated from a three-parameter logistic function with parameter values, $a = 1, b = 0, c = 0.2$. For evaluated points, 21 values with equal spacing were used in each simulated sample, $\theta_\xi = -3(0.3)3$. These evaluated points are reasonable because there are few points beyond interval $[-3, 3]$ with the standard normal distribution.

The number of simulated samples analyzed were 100. Within each simulated sample, binary values $y_i \in \{0, 1\}$ were generated as follows: generate 500 u_i's randomly from the uniform distribution, let $y_i = 0$ if $u_i \leq P(\theta_i)$, otherwise $y_i = 1$ if $\mu_i > P(\theta_i)$.

Because both cubic polynomial spline and psychometric spline involve choosing the smoothing parameter λ, and the first method does not, so the comparison between the second and the third method is considered first.

The smoothing parameter λ, was given values of $0.001, 0.01, 0.1, 1, 10$ and 100. For each parameter λ, the average of mean square errors of each simulated sample for $P(\theta), h(\theta), (Dh)(\theta)$ and $I(\theta)$ were computed.

In each graph of Figure 7.1, the x-axis takes values of $\ln_{10}(\lambda)$, and y-axis takes values of the corresponding mean square errors. The solid line connects the points by psychometric splines, the dotted line is by cubic polynomial splines. It is clear that for each value of λ, the mean square errors for $P(\theta), h(\theta), (Dh)(\theta)$ and $I(\theta)$ by psychometric splines are always smaller than the cubic polynomial splines. And the optimal λ values of Dh and I are greater than those for P and h in both situations. That means that estimating Dh and I requires more smoothing than P and h.

There is another way to compare psychometric splines with cubic polynomial splines. Compare the mean square errors of both across the simulated samples for different ability level θ_i's.

Figure 7.2 and Figure 7.3 give the mean square errors of both cubic polynomial splines and psychometric splines for different values of λ with different ranges of θ. The range of θ for Figure 7.2 is from -3 to 3, and for Figure 7.3 is from -1.5 to 1.5. In both figures, the solid line shows the values of mean square errors of the psychometric splines and the dotted line shows the values of cubic polynomial splines. In Figure 7.2, one can see that cubic polynomial splines are much more unstable than psychometric splines, especially on boundaries, but it is difficult to see the difference between them when θ varies from -1.5 to 1.5. In Figure 7.3, it is clear that whatever the value of smoothing parameter λ is, the mean square errors of psychometric splines are almost always less than cubic polynomial splines. Recall the properties of $h(\theta)$: $h(\theta)$ is asymptotically linear in θ when θ tends to infinity. The psychometric splines with the basis $\{1, \theta, \ln(e^{\theta}+1)\}$ satisfies this property, but not the cubic polynomial splines. Therefore, the result confirms the value of using a basis for H_1, which does have this property.

Once the optimal smoothing parameter, λ, is chosen, the comparison between the first method and the psychometric splines can be made.

In Figure 7.4, the solid line represents the values generated from the 3PL model, the dashed line stands for the values of the linear approximation by GLM method, and the dotted line is from the estimated values by psychometric splines, using $\lambda = 0.01$ for estimating h and $\lambda = 0.1$ for estimating Dh. Comparing these three lines, one can see that the linearly approximated values catch almost all of the characteristics of functions $h(\theta)$ and $P(\theta)$, especially when θ has a larger positive value. It illustrates again that the basis, $\{1, \theta, \ln(e^{\theta} + 1)\}$, satisfies the properties of $h(\theta)$. The estimated values by psychometric splines are closer than the linearly approximated values to the original data because the spline method is more flexible. Although the estimated Dh doesn't fit very well, good estimated values of $I(\theta)$ were still found.

7.2 Discussion and Conclusion

The purpose of this thesis is to propose a method combining parametric models and nonparametric regression together. It was hoped that by controlling the smoothing parameter λ, one can make the smoothed values as close as possible to the parametric models, and on the other hand, one can also leave more room for the random component. For a very simple simulation described above, it seems to work well.

Comparing the method used in this thesis with the parametric model and kernel smoothing, the psychometric spline is easier to compute and more flexible than the parametric model.

However, this method needs improving. To go further, first, the method of computation should be improved, so that one can improve the accuracy of the computation. Also a greater range of simulations should be carried out to test this method.

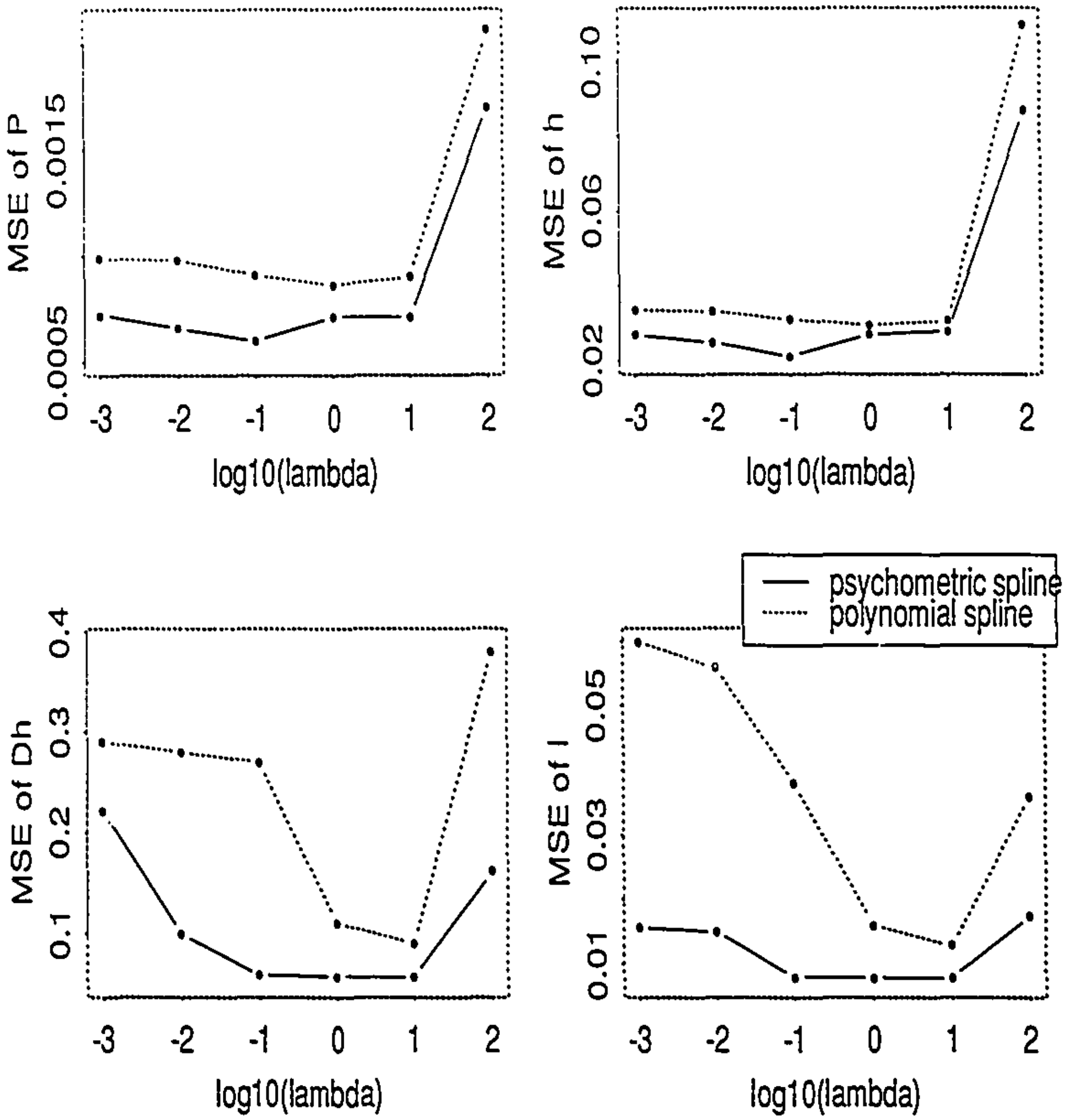


Figure 7.1: Comparison between polynomial splines and psychometric splines by the mean square errors of $P(\theta), h(\theta), (Dh)(\theta)$ and $I(\theta)$ within simulated sample for different values of λ.

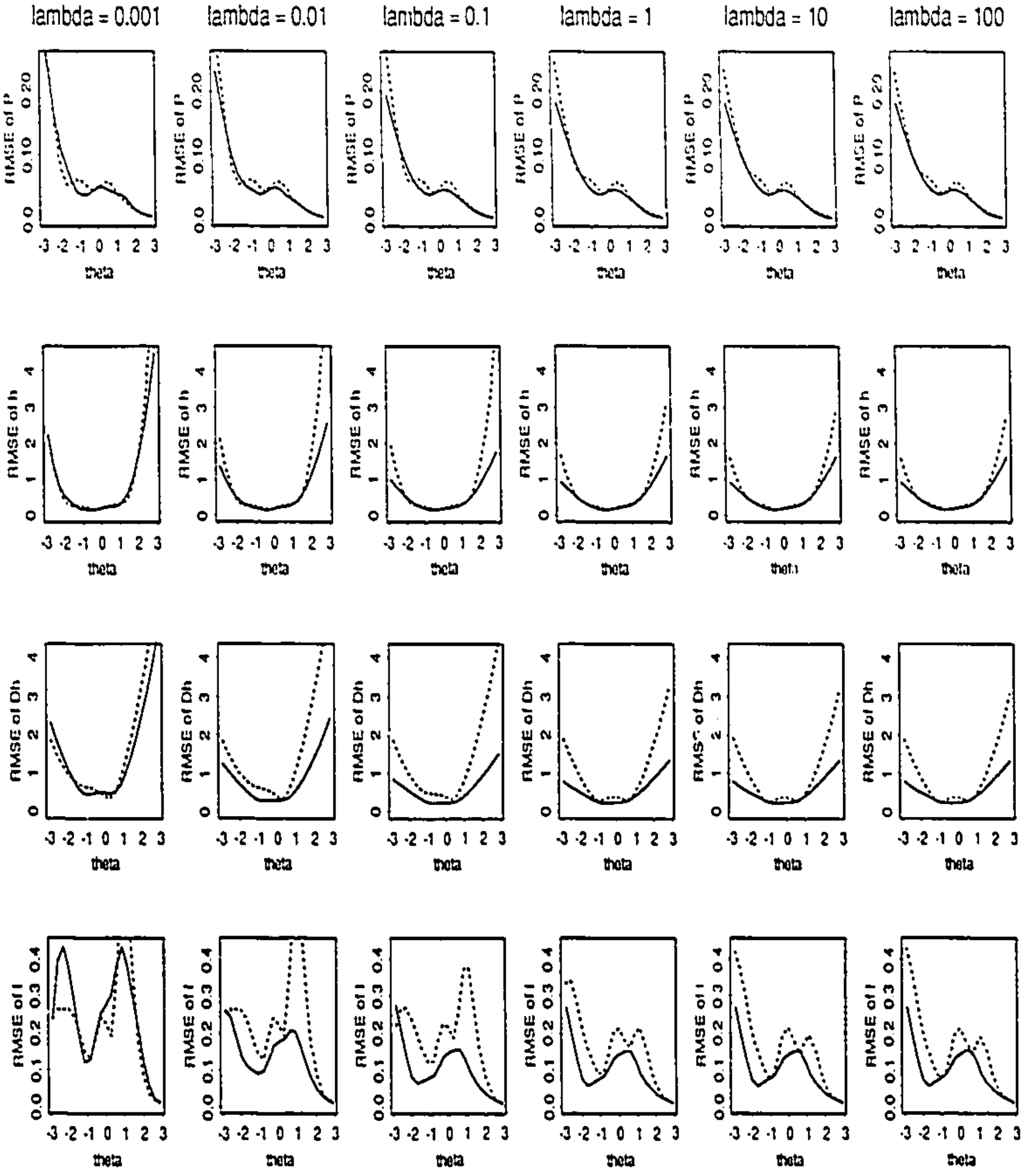


Figure 7.2: Comparison between polynomial splines and psychometric splines by mean square errors of $P(\theta), h(\theta), (Dh)(\theta)$ and $I(\theta)$ across simulated samples, and for different values of λ with the range of θ from -3 to 3. The solid line shows the result for the psychometric splines and the dotted line for the cubic splines

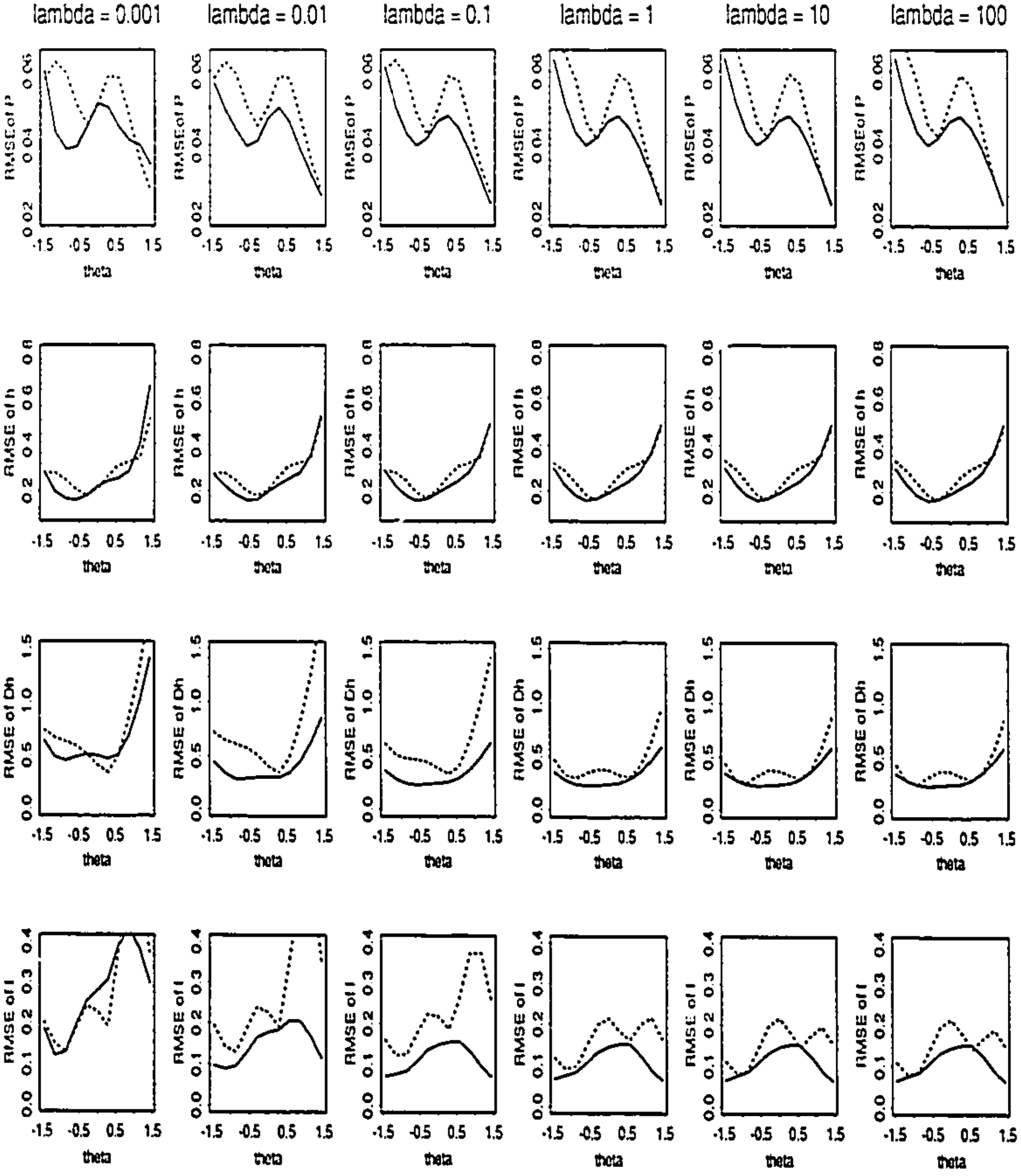


Figure 7.3: Comparison between polynomial splines and psychometric splines by mean square errors of $P(\theta)$, $h(\theta)$, $(Dh)(\theta)$ and $I(\theta)$ across simulated samples, and for different values of λ with the range of θ from -1.5 to 1.5. The solid line shows the result for the psychometric splines and the dotted line for the cubic splines

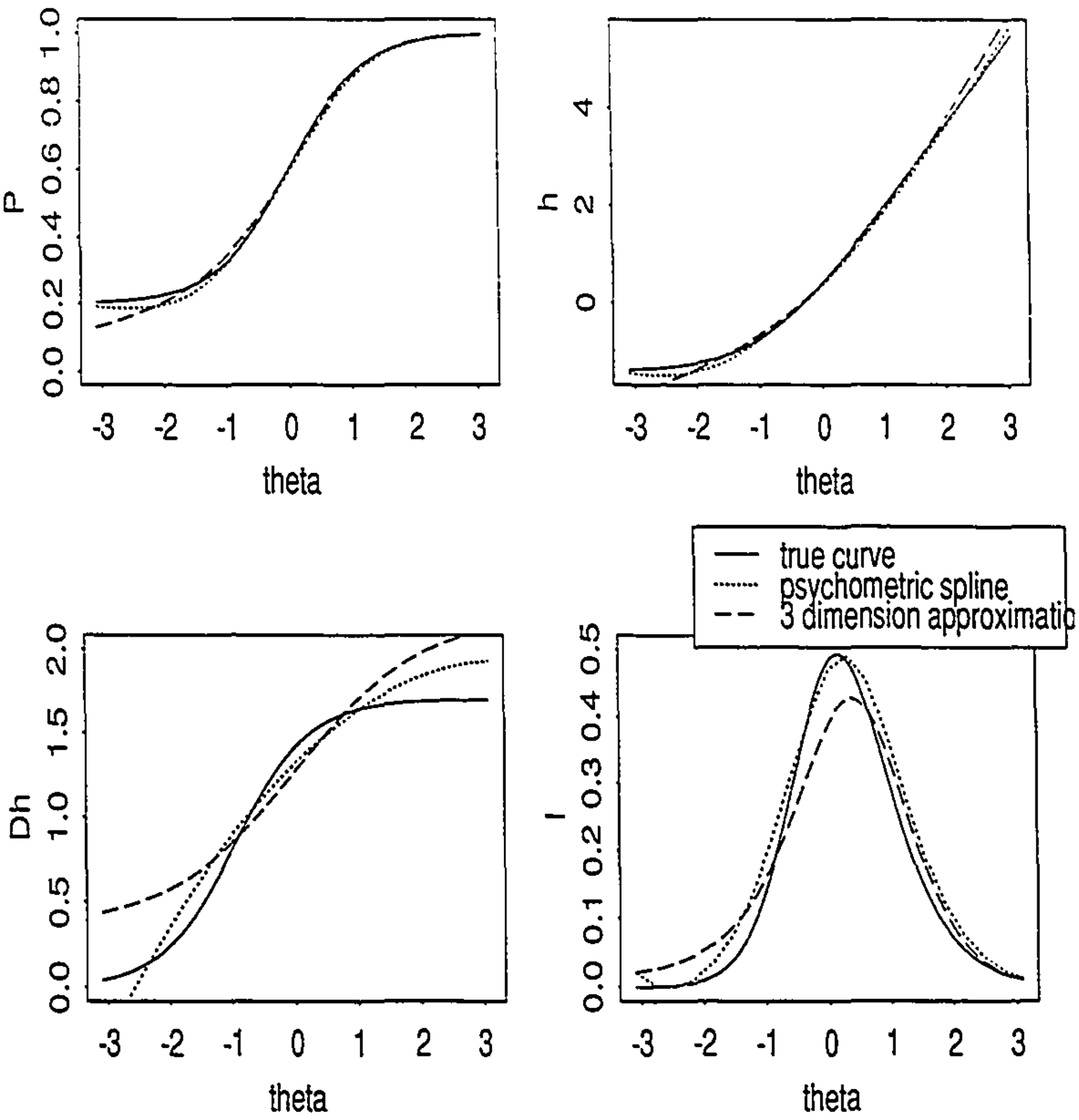


Figure 7.4: Comparison between the psychometric splines and GLM method with the basis $\{1, \theta, \ln(e^{\theta}+1)\}$ for one simulated sample. P and h were estimated with $\lambda = 0.01$, and Dh and I were estimated with $\lambda = 0.1$. The solid line shows the true curve, the dotted line shows the estimate by the psychometric splines, and the dashed line shows the estimate by the 3 dimension linear regresion.

Bibliography

[1] Adams, R.A. (1975) *Sobolev spaces*, New York: Academic Press.

[2] Anderson J.A. and Blair V. (1982) Penalized maximum likelihood estimation in logistic regression and discrimination, *Biometrika*, **69**, 123-136.

[3] Aubin Jean-Pierre (1979) *Applied functional analysis*, New York: Wiley.

[4] Cox D.R. and Snell E.J. (1989) *Analysis of binary data*, second edition, New York; London: Chapman and Hall.

[5] Craven P. and Wahba G. (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation, *Numerische Mathematik*, **31**, 377-403.

[6] Dalzell C.J. and Ramsay J.O. (1993) Computing reproducing kernels with arbitrary boundary constraints, *Society for Industrial and Applied Mathematics*, **14**, 511-515.

[7] Good I.J. and Gaskins R.A. (1971) Nonparametric roughness penalties for probability densities, *Biometrika*, **58**, 225-277.

[8] Gu C. and Qiu C. (1993) Smoothing spline density estimation: theory, *The Annals of Statistics*, **21**, 217-334.

[9] Gu C. (1993) A dimensionless automatic algorithm, *Journal of the American Statistical Association*, **88**, 495-504.

[10] Lord E.M. (1980) *Application of item response theory to practical testing problem*, Hillsdale, New Jersey: Lawrence Erlbaum Associates.

[11] McCullagh P. and Nelder J.A. (1989) *Generalized linear models*, second edition, London: New York: Chapman and Hall.

[12] Nelder J.A. and Wedderburn R.W.M., (1972) Generalized linear models, *Journal of the Royal Statistical Society, Series A*, **135**, 370-384.

[13] Press W.H., Teukolsky S.A., Vetterling W.T. and Flannery B.P. (1992) *Numerical recipes in Fortran*, second edition, Cambridge: New York: Cambridge University Press.

[14] O'Sullivan F., Yandell B.S. and Raynor W.J. (1986) Automatic smoothing of regression functions in generalized linear models, *Journal of the American Statistical Association*, **81**, 96-103.

[15] Ramsay J.O. and Dalzell C.J. (1991) Some tools for functional data analysis (with discussion), *Journal of the Royal Statistical Society, Series B*, **53**, 539-572.

[16] Ramsay J.O. (1991/93) *Testgraf Manual*, McGill University: unpublished manuscript.

[17] Silverman B.W. (1978) Density ratios, empirical likelihood and cot death, *Applied Statistics*, **27**, 26-33.

[18] Wahba G. and Wold S. (1975) A completely automatic French curve: fitting spline functions by cross-validation, *Communications in Statistics*, **4**, 1-17.

[19] Wahba G. (1990) *Spline models for observational data*, Philadelphia: Society of Industrial and Applied Mathematics.