School of Science

Department of Mathematics and Statistics

Statistical Analysis of Genomic Data: A New Model for Class Prediction and Inference

Zhenyu Jiang

This thesis is presented for the Degree of

Doctor of Philosophy

of

Curtin University

October 2011

Declaration

To the best of my knowledge and belief, this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature: Zhenyu Jiang

Date: 12 October 2011

Acknowledgments

I am extremely grateful to my supervisor, Professor Kok Lay Teo, for his helps and supports during my study. I also thank my co-supervisor, A/Professor Zudi Lu, and the departmental postgraduate coordinator, Professor Yong Hong Wu, for their supports and for their helpful advices. Finally, I sincerely thank all the students and staff in the Department of Mathematics and Statistics for their encouragement during my study at Curtin.

At the beginning of this project, I had many valuable discussions with Prof Luba Kalaydjieva for which I gratefully acknowledge. Futhermore, I wish to thank the Western Australian Institute for Medical Research and The University of Western Australia for providing me the datasets of WAFSS (Western Australian Family study of Schizophrenia).

Abstract

Genomics is a major scientific revolution in this century. High-throughput genomic data provides an opportunity for identifying genes and SNPs (single-nucleotide polymorphism) that are related to various clinical phenotypes. To deal with the sheer volume of genetic data being produced, it requires advanced methodological development in biostatistics that is lagging behind the technical capability to generate genomic data. SNPs have great importance in biomedical research for comparing regions of the genome between cohorts (such as case-control studies). Within a population, SNPs can be assigned a minor allele frequency, the lowest allele frequency at a locus that is observed in a particular population, and be recoded to binary datasets. Therefore, it is important to develop suitable statistical methods for SNPs analysis of genome alteration with the goal of contributing to the understanding of complex human diseases or traits such as mental health.

In this thesis, we develop new statistical methodologies for the analysis of schizophrenia genomic data from the WA Genetic Epidemiology Resource

(WAGER). The motivation is driven by the schizophrenia class prediction, (i.e. the prediction of individuals' disease status through their genotype and quantitative traits). In general, individual's disease status is a nominal variable, while genotypes can be converted into ordinal variables but are of high dimension. Note that the usual nonparametric regression that is developed for continuous variables cannot be applied here. There are some methodologies, such as the tree-based logistic Non-parametric Pathway-based Regression model (NPR) proposed by Wei and Li (2007)available in the literature. However, it is found that this model does not well adapt to the data set that we are analyzing. It is even worse than the (generalized) linear logistic regression model. Using logistic discrimination rule, together with adding quantitative traits, some important results have been obtained. However, some shortcomings remain. Firstly, the generalized linear logistic model has a high type I error rate for schizophrenia classification. Secondly, quantitative traits required for schizophrenia class prediction are performance assessments which demand several hours on-site participation by both assessor and assessee. These traits are generally quite difficult to reach even for a medium size sample. Meanwhile, though the laboratory analyzing cost is high, a person's genotype can be obtained by merely collecting a drop of blood.

Thus, two kinds of nonlinear models are proposed to capture the nonlinear effects in SNP datasets, which are categorical. The main contributions of this thesis are summarized as follows:

- Two kinds of nonlinear threshold index logistic regression models are proposed to capture the nonlinear effects by applying the idea of threshold models (Tong (1983, 1990)) which are parametric and therefore applicable to the categorical data.

  One of the proposed models, which is called the partially linear threshold index logistic regression (PL-TILoR) model, is given by

$$\log\left\{\frac{P(Y_i=1|\mathbf{X}_i)}{1-P(Y_i=1|\mathbf{X}_i)}\right\}=\boldsymbol{\alpha}^T\mathbf{X}_i+g(\boldsymbol{\beta}^T\mathbf{X}_i), \tag{0.1}$$

  where Y_i is the disease status of the ith person under case-control study, taking on values of 1 (case) or 0 (control), $\mathbf{X}_i$ is the vector of genotype variables, which is p-dimensional, and the superscript T stands for transpose of a vector or matrix. Here, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are p-dimensional unknown parameters with $\boldsymbol{\beta}$ being an index vector used for the reduction of dimension, satisfying $\|\boldsymbol{\beta}\|=1$ and $\boldsymbol{\alpha}^T\boldsymbol{\beta}=0$ for model identifiability, and g is, therefore, a one-dimensional nonlinear function, which is modelled as stepwise linear function through threshold effect (Tong, 1990), given below.

$$g(z)=(b_1z+b_2)I_{\{z\le c\}}+(b_3z+b_4)I_{\{z>c\}}, \tag{0.2}$$

  where b_i's and c are unknown parameters to be estimated and I_A is an

indicator function of the set A.

In practice, the first component in model (0.1) could also be nonlinear. In this case, model (0.1) becomes

$$\log\left\{\frac{P(Y_i=1|\mathbf{X}_i)}{1-P(Y_i=1|\mathbf{X}_i)}\right\} = g_1(\boldsymbol{\alpha}^T\mathbf{X}_i) + g_2(\boldsymbol{\beta}^T\mathbf{X}_i), \qquad (0.3)$$

where $\|\boldsymbol{\alpha}\| = 1$, $\|\boldsymbol{\beta}\| = 1$ and $\boldsymbol{\alpha}^T\boldsymbol{\beta} = 0$ for model identifiability, and g_1 and g_2 are two one-dimensional nonlinear functions which are modelled by stepwise linear functions through threshold effects as follows:

$$g_k(z) = (b_{k1}z + b_{k2})I_{\{z\le c_k\}} + (b_{k3}z + b_{k4})I_{\{z>c_k\}},\ k = 1, 2, \qquad (0.4)$$

where b_{ki}'s and c_k's are unknown parameters to be estimated. Thus, (0.3) and (0.4) form an additive threshold index logistic regression (A-TILoR) model.

- A maximum likelihood methodology is developed to estimate the unknown parameters in the PL-TILoR and A-TILoR models. Simulation studies have found that the proposed methodology works well for finite size samples.

- Empirical studies of the proposed models applied to the analysis of schizophrenia genomic data from the WA Genetic Epidemiology Resource (WAGER) have shown that A-TILoR model is very successful

in reducing the type I error rate in schizophrenia classification without even using quantitative traits. It outperforms the generalized linear logistic model that is widely used in the literature.

Contents

List of Figures

List of Tables

Chapter 1

Background: Literature and Problems

1.1 Introduction

Genomics is a major scientific revolution in this century. It is the study of all the genes of a cell or tissue at the DNA, mRNA or protein levels. High-throughput genomic data provides an opportunity for identifying pathways and genes that are related to various clinical phenotypes. To deal with the sheer volume of genetic data being produced, it requires advanced methodological development in biostatistics that is lagging behind the technical capability to generate genomic data.

In modern molecular biology and genetics, the genome is the entirety

of an organism's hereditary information. It is encoded either in DNA or, for many types of virus, in RNA. The genome includes both the genes and non-coding sequences of the DNA. A single-nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide (A, T, C, or G) differs between members of species. Within a population, SNPs can be assigned a minor allele frequency — a lesser allele frequency at a locus is observed in a particular population. In order to conduct quantitative analysis of SNP variables, SNP data must be first coded as 0-1-2 (or 0/1), depending on the number of the minor allele frequency (or have/have not the minor allele frequency), which is a kind of categorical data. In this study, we use the SNP variables with recoding format of 0-1-2, where 0-1-2 represents the number of the minor allele frequency on a locus. We will be particularly concerned with the analysis of schizophrenia genomic data.

In this research, we will focus on data mining of high-dimensional SNP data sets and developing new methodologies of classification. We will apply the statistical methodologies developed to the analysis of the schizophrenia SNP data, empirically investigating the performance of the new classification models. Before doing this, we first review some related background knowledge and literature.

1.2 Schizophrenia genomic data and classification literature review

Schizophrenia is a mental disorder characterized by a disintegration of thought processes and of emotional responsiveness. It is accompanied by significant social or occupational dysfunction. The onset of symptoms typically occurs in adolescence or young adulthood. With a lifetime risk of about 1%, over 50% of those affected develop chronic disabilities and nearly all experience a diminished quality of life. The current diagnosis procedure is based on observed behavior and the patient's reported experiences (ICD-10 and DSM-IV). Schizophrenia is one of the genetically complex disorders, with heritability at about 80% and likely multiple genes of small to moderate effect, as well as a host of enironmental influences. Notwithstanding availability of powerful techniques of genetic analsis, such as whole-genome association studies, the basic problem of "connecting phenotype with the genotype" in schizophrenia remains unresolved. The argument about whether schizophrenia is a single disease or a collection of pathogenetically distinct subtypes goes back to the inception of the diagnostic concept at the turn of the 20th century. E. Bleuler (1920) emphasized that "It is not a disease in the strict sense, but appears to be a group of disease. Therefore, we should speak of schizophrenias in the plural."

WAFSS (Western Australian Family Study of Schizophrenia) endorses the hypothesis that the syndrome of schizophrenia comprises several subtypes that could be delineated by objective endophenotype measurements of brain function and by exploring their genetic underpinnings. A core aim of the Western Australian Family Study of Schizophrenia (WAFSS), since its inception in 1996, is to address the problem of heterogeneity in Schizophrenia.

The WAFSS case-control study has been lasting more than a decade. The study is still continuing today. WAFSS study population includes 496 Western Australians of European descent. There are 325 members affected by schizophrenia (cases), and 171 population controls. The controls, recruited from a list of Red Cross blood donors or by random sampling from local telephone directories, were screened for psychopathology. For those if they or any of their first-degree relatives had been diagnosed with schizophrenia or bipolar affective disorder, they were excluded. Written informed consent was obtained from every participant. The study has complied with the ethics guidelines of the institutions involved. Genotyping was conducted on 23 selected genes according to neurological knowledge and research interests. A total of 1022 SNPs was found, which means that the data set volume reaches 1022*500. Through WAFSS's co-operations with the Welcome Trust in the UK to expand genotyping into whole genome, the number of genomic data will soon be doubled.

In 2005, WAFSS identifies a homogeneous familial subtype of the disease,

referred to as "cognitive deficit" (CD) subtype. Another markedly contrasting subtype, "cognitively spared" (CS), exhibits high performance on majority of cognitive tasks. WAFSS uses grade of membership (GoM) analysis (Woodbury et al. 1978; Manton et al. 1994) to analyze the test results, with control individuals providing the baseline data. GoM is a form of latent structure analysis, directed at defining a parsimonious number of latent groups or patterns of responses (representing, e.g., biological phenotypes) from complex data sets, and allowing individuals to resemble each group to varying degrees (rather than classifying them into mutually exclusive clusters, as done in standard latent class analysis). Using the individual-level GoM coefficients, they classify the schizophrenia into three subtypes: CD, CS and non-CD/CS. Although CD subtype is originally identified via phenotypes, further genotype studies like whole-genome scan and linkage analysis suggest this subtype characterizes a genetically distinct schizophrenia subtype that accounts for the linkage of schizophrenia to chromosome 6p25-24 region (see Hallmayer et al. 2005). In their study, CD subtype constitutes up to 50% of sample population.

An important part of the WAFSS project is statistical analysis of schizophrenia SNP dataset. WAFSS uses the generalized linear model-based method to examine each SNP for association with disease outcome at the genotype level. Risk is expressed as odds ratio (OR), which is a useful measure of association between some risk factor and disease (see Thomas, 2004, page 70),

with 95% confidence limit and associated p-value. The logistic regression is used to identify SNPs combinations within each gene to predict schizophrenia outcome and CD trait variation. Shortcoming of the individual gene analysis will be discussed in Section 3 of this Chapter.

A medical diagnosis is an attempt at classification. In statistics, classification methods can be dated back as early as 1930's. It was first applied in eugenics in 1935 by M. Barnard at the suggestion of R.A. Fisher. Fisher linear discriminant analysis (FLDA) is based on ratios of between-groups to within-groups sum of squares. This criterion is intuitively appealing. However, because genomic data has small number of sample but a fairly large number of genes, the matrices of between-groups and within-groups sum of squares may be quite unstable, leading to poor estimates of the corresponding population quantities. See Dudoit et al. 2002.

Other classical multivariate statistical discrimination approaches include the nearest neighbour classification and the Maximum likelihood discriminant rule. Both of them are known to perform well for the classification of tumours using gene expression data (Dudoit et. al. 2002). The nearest neighbour classification is a popular nonlinear classifier, which is developed for the measure of distance between observations. The K-nearest neighbour rule (Fix and Hodge (1951)) first finds K nearest neighbours of the unknown vector from the training vectors. Then, the unknown vector is assigned to the class which appears most frequently in the vectors identified in the previous step.

The maximum likelihood classifier (ML) predicts the class of an observation X such that the largest likelihood to X is obtained. If we assume that the conditional density for each class is multivariate Gaussian, then the ML discriminant rule reduces to the Quadratic discriminant analysis (QDA). Depending on the character of covariance matrices, there are two variants of the QDA——the Diagonal quadratic discriminant analysis (DQDA) and the Diagonal linear discriminant analysis (DLDA). Golub et al. (1999) propose a "weighted gene voting scheme" which turns out to be a variant of a special case of DLDA by Dudoit et al. (2002). It is worth mentioning that the work of Golub et al. (1999) is the first application of a discriminant rule to gene expression data, separating out two different but clinically indistinguishable types of leukaemia, ALL and AML. Modern machine learning algorithms provide sophisticated approaches to estimate the decision boundary or the distribution parameters. By treating the data mechanism as unknown, algorithmic modelling created another culture in the use of statistical modelling to reach conclusions from data (Breiman 2001). Recursive Partitioning is a statistical technique that forms the basis for two classes of nonparametric regression methods: classification and regression trees (CART) (see Breiman, Friedman etc. (1984)), and Multivariate Adaptive Regression Splines (MARS). CART consists of three main aspects to the tree construction: (a) selection of the splits, so that the data in each of the descendant subsets are "purer" than the data in the parent subset; (b) the decision to declare a node terminal, which

is done using cross-validation to "prune" the tree; and (c) the assignment of each terminal node to a class. Breiman (1996, 1998) finds that gains in accuracy could be obtained by aggregating predictors built from perturbed versions of the learning set. There are two main classes of methods for generating perturbed versions of the learning set: bagging (L. Breiman. 1996) and boosting (Freund et. al. 1997). They have gained popularity in building predictive models and identifying genes that are related to clinical phenotypes (Dettling et al. 2003; Li and Luan, 2005). Random forest (Breiman, 2001) consists of a large number of randomly constructed trees, each voting for a class. The forest is grown by perturbing the training set, growing a tree on the perturbed training set, perturbing the training set again, growing another tree, etc. It is a more accurate predictor than CART, but less interpretable. Other popular machine learning approaches include support vector machines (SVM) (Guyon et al. 2002; Mclachan et al., 2004), nearest shrunken centroids (NSC) (Tibshirani et al., 2002, 2003; Sharma et al., 2005) and neural networks (Khan et al. 2001). They are used in building predictive models, and studies often suggest lists of the genes likely to be involved in a disease. Medical diagnosis is a particularly fruitful area of application for statistical classification (Hand D.J. 1981). The methods have been applied to the assessment of the prognostic value of tests of lung function in miner with pneumoconiosis, to predicting Ischaemic heart disease, to predicting relapse in pulmonary tuberculosis sufferers, etc. With new high-

throughput genomic data now available, the discriminant analysis is useful in a new variety of settings. Also, class predictors can be constructed for known pathological categories and provide diagnostic confirmation or clarify unusual cases. Most importantly, the technique of class prediction can be applied to distinctions relating to future clinical outcome. For instance, Hedenfalk et al.(2001) compared gene expression profiles for two types of hereditary breast cancer (BRCA1 mutation and BRCA2 mutation) and found that distinctly different groups of genes are expressed by the two types, suggesting that a heritable mutation affects the gene expression profile of the cancer.

To dissect genetic predisposition to phenotypic traits, many studies assemble hundreds of SNPs in a panel of candidate genes presumably involved in regulating the underlying biologic mechanism. Involvement of multiple genes in each pathway suggests the importance of studying both main effects and interactions.

Wei and Li (2007) propose a novel nonparametric pathway-based regression model (NPR) for the analysis of genomic data. They assume that the phenotype is related to the total activity level across multiple pathways through an additive model,

$$F(x) = \sum_{k=1}^{K} F_k(x^{(k)}), \tag{1.1}$$

where $F_k(x^{(k)})$ can be interpreted as the activity level associated with the

kth pathway as determined by the genomic measurements of the p_k genes in this pathway. For a binary phenotype such as disease status or normal versus cancerous tissues, they assume a logistic model for Y,

$$P_r(Y = 1|x, Z) = \frac{\exp(2(F(x) + \gamma Z))}{1 + \exp(2(F(x) + \gamma Z))}, \tag{1.2}$$

where $Y = 1$ for diseased individual and $Y = -1$ for normal individual, while Z is the vector of other patient-specific covariates which is modelled parametrically with coefficient γ. To obtain an additive model in the form of (1.1), they propose to use regression trees as base learners, and a pathway-based gradient descent technique (Friedman, 2001) as the boosting algorithm. In such an NPR model, known biological pathways are treated as first-level regression units, which provide a nice biological interpretation of the resulting regression models. More importantly, using real breast cancer gene expression datasets, this model outperforms several other well-known classifiers including SVM, Logistic regression, Random Forest, Bagging, Nave Bayes and Neural Network.

In the WAFSS proposal of the year 2007, it was written: "We will use the nonparametric pathway regression approach of Wei and Li in addition to the methods described above. The method explicitly incorporates biological pathway information and combines both regression tree and boosting approaches to generate association models, with gene-specific and pathway-

specific importance scores." This will be discussed in Chapter 2.2.

1.3 Objectives of This Study

In this thesis, following the idea suggested in the analysis of functional modules proposed by WAFSS, we consider a new way of predicting schizophrenia outcome and CD trait variation. Although WAFSS's proposal and Wei and Li's approach are reasonable judging from the datasets they have analysed, their methods still suffer from some obvious shortcomings:

- Firstly, schizophrenia is a genetically complex disease, meaning that multiple genes have small to moderate effects on the determination of schizophrenia. Therefore, the individual gene analysis that is currently used in WAFSS study is not suitable for the classification of schizophrenia and CD trait variation.

- Secondly, to characterize the genetically complex features, Wei and Li (2007) propose to use the regression-tree-based nonparametric pathway regression (NPR) model. It has been examined using simulated SNP datasets and real gene expression datasets, but not real schizophrenia SNP datasets. Because of its ordinal/categorical characteristics, a SNP dataset is totally different from gene expression data sets, which are naturally continuous. Also, in their original paper, the performance of NPR model used in SNP selection was compared with other

machine learning methodologies only. We will show that Wei and Li's NPR model does not characterise well the genetically complex nature of schizophrenia SNP datasets. It performs even worse than the classical linear logistic regression model (see Chapter 2).

- Thirdly, linear logistic regression is a widely used approach to handling binary outcome datasets. Obviously, it cannot characterise the nonlinear features of genetical complexity. Based on logistic transformation, it is important to propose a new model that can capture the complex characters of SNP datasets. This is what we are pursuing.

In this thesis, our main objectives are to investigate SNPs, leading to the determination of schizophrenia and the prediction of the possibility of a person who may develop schizophrenia or schizophrenia CD subtype purely according to his/her genotype information. More specifically, we are concerned with the following points:

- What is the performance of the NPR model on the schizophrenia SNP datasets? Does it characterise well the nonlinear complexity of the schizophrenia from the perspective of prediction? If not, could we develop more efficient and effective models for such an objective?

  We will show, in Chapter 2, that the performance of Wei and Li's NPR model is even worse than the linear logistic regression model on the schizophrenia SNP datasets. We will propose a new class of nonlin-

ear logistic regression models, called threshold index logistic regression (TILoR), in Section 3.2 of Chapter 3.

- How do we estimate the unknown parameters in the newly proposed models? This is a fundamentally important question in applications. A maximum likelihood estimation procedure will be suggested in Section 3.3 of Chapter 3. The performance of the estimators will be investigated via Monte Carlo simulation in Section 3.4 of Chapter 3.

- What benefits can be achieved by establishing a TILoR statistical approach to the schizophrenia SNP datasets? Empirical experiment will be carried out in Chapter 4, using the schizophrenia SNP data. Comparison with the linear logistic regression model will be addressed in Chapter 4.

We concluded with some suggestions for future research directions in Chapter 5.

1.4 Significance of This Study

To date, schizophrenia diagnosis procedure is based on observed behavior, patient and doctor SCAN interview, and the patient's reported experiences (ICD-10 and DSM-IV). Using the TILoR models we develop, we will show that cross-validation predictive accuracy rates of about 70% for gen-

eral schizophrenia and 80% for schizophrenia CD subtype can be achieved by utilising only the genotype information (see Chapter 4). Some significant points of our research are summarised as follows.

- In this thesis, we will extend the idea of threshold (auto)regression that was suggest by Tong (1983,1990) in nonlinear time series analysis to the nonlinear genomic analysis of SNP data that are of categorical nature. On this basis, we propose a new class of threshold index logistic regression (TILoR) models. Using this new framework of logistic regression with schizophrenia data sets, we find that the TILoR models can well capture the genetically complex features of the schizophrenia SNP datasets in terms of the cross-validation predictive accuracy rates.

- Our TILoR schizophrenia prediction is based on SNP genotype data alone, meaning that only a drop of blood taken from a case or control participant will be sufficient for genotyping used in our model. The final TILoR model involves about 40 SNPs on 12 genes, which dramatically reduces the costs of genotype and therefore, the costs of the prediction.

- The result of 70% accuracy of the cross-validation prediction with our TILoR models for general schizophrenia is quite close to the 80% broad heritability of schizophrenia, which, according to the experts' view from WAFSS, is an upper limit of the prediction accuracy using genotype data alone. Also, using our TILoR models, the specificity and sensi-

tivity for general schizophrenia prediction are 67.84% and 71.3%, respectively. Note that sensitivity and specificity are statistical measures of the performance of a binary classification test, and sensitivity measures the percentage of sick people who are correctly identified as having the condition while specificity measures the percentage of healthy people who are correctly identified as not having the condition (c.f., http://en.wikipedia.org/wiki/Sensitivity_and_specificity). As to newly discovered schizophrenia CD subtype, it is a more genetically significant subtype with severe cognitive deficit. Although there is no medical research findings about the broad heritability for CD subtype, it is expected that the figure would be higher than 80%. Our research findings corroborate this expectation. The results of our TILoR models for CD subtype are about 10% better than general schizophrenia: cross-validation predictive accuracy 81.93%, specificity 87.14% and sensitivity 76%. It should be emphasized that specificity and sensitivity accuracy figures are more important than the overall accuracy of prediction in medical practice. If a classification model have good results for both specificity and sensitivity, it will certainly have a good prediction accuracy rate. However, a not bad prediction accuracy rate can not guarantee acceptable specificity and sensitivity accuracy rates. All the comparisons from the perspective of these quantities show that our proposed TILoR models have made significant improvement over the

logistic linear regression and the NPR model of Wei and Li (2007) in the schizophrenia prediction based on SNPs from 23 genes.

Our new findings have the potential to become a part of medical diagnostic process. It should be noticed that medical diagnosis in psychiatry is problematic. Apart from the fact that there are differing theoretical views toward mental conditions, there are few lab tests available for various major disorders. Being a readily available and relatively low cost lab test in genotyping, our findings are outstandingly accurate.

Chapter 2

Analysing schizophrenia data: nonparametric pathway-based regression and linear logistic regression

2.1 Introduction

In this chapter, we first examine the performance of the analysis of the schizophrenia data sets that we introduced in Chapter 1 by applying the nonparametric pathway-based regression (NPR) model proposed by Wei and Li (2007) and the popular generalized linear (logistic) regression model. Al-

though NPR is a kind of tree-based nonlinear model, we will show that it cannot well capture the nonlinear features existed in the schizophrenia data sets. In fact, we will note that the popular generalized linear (logistic) regression model outperforms the NPR model in analysing the schizophrenia data sets.

2.2 Preliminary data handling

The research on statistical analysis of genomic data in this thesis focuses on a large comprehensive schizophrenia dataset. The whole dataset (WAFSS) was collected in WA over a period of more than 10 years, originally as a family study but later changed to a case-control study. A comprehensive examination of the role of the genetic variation of this dataset is a four-year project of WAFSS, which has already lab genotyped more than half a million genetic data, and through co-operations with the Welcome Trust in the UK, the genomic data is expected to be doubled.

After having access to the original dataset, it is required to "clean" and recode it before any further analysis can be carried out. This means that we need to eliminate the samples that are not genotyped and group subjects according to their phenotypes. After that, a sample consisting of 496 individuals (325 cases and 171 controls) is obtained. For each individual, this research focuses on 1022 single nucleotide polymorphisms (SNPs) data

in 23 genes. Then, we calculate the allele frequency of each SNP in the control population and decide which allele is the minor allele for each SNP in WAFSS population. We numerically recode genotypes into three categories, named 0, 1, 2, according to the number of minor allele frequency of a SNP for all the samples. Occasionally for some very few individuals, some SNPs can not be detected. In that case, we replace them with category 0. This replacing is basically reasonable for ease of the analysis of the data sets. The reasons are as follows. Firstly, 0 is the most probable status among the three SNP categories. Secondly, the ratio of the number of the missing values in the data sets analysed is very low. For example, the general schizophrenia data set analysed in Chapter 4 has only 11 missing values, the missing ratio of which is as low as 0.0005544355, and the CD subtype there has only 7 missing values with the missing ratio 0.0005556879.

We start our research from OR (odds ratio) χ^2 test for 1022 SNPs. Odds ratio is a useful measurement of association between risk factor and disease. In our case, the risk factor is the absence or presence of a minor allele (SNP recoded as zero or not zero accordingly). And the disease is schizophrenia or schizophrenia CD subtype. If the odds ratio of an individual SNP equal to one, that equivalents to the independence between the SNP and the disease. For details, the reader is referred to Pages 70–71 of Thomas (2004). A very significant OR P-value is a criterion that is widely used in Biology and Medical science. Biologists start from OR test then proceed to other Genetic

Epidemiology studies like Genome-wide association study. In this research, we start from the OR analysis then proceed to statistical model building. We make a preliminary selection of significant SNPs according to the p-value of the OR for schizophrenia and schizophrenia CD subtype. Then the selected SNPs are used in our proposed TILoR model and linear logistic regression model.

2.3 Analysis by nonparametric pathway-based regression model

Nonparametric pathway-based regression model (NPR model) proposed by Zhi Wei and Hongzhe Li in 2007 (Li et.al. 2007) is a novel approach that the WAFSS group is particularly interested in. The WAFSS group is attracted by the NPR method because it can explicitly incorporates biological pathway information. The model combines both regression tree and boosting approaches to generate association models, with gene-specific and pathway-specific importance scores.

NPR model assumes that phenotype is related to the total activity level across multiple pathways through an additive model,

$$F(x) = \sum_{k=1}^{K} F_k(x^{(k)}), \tag{2.1}$$

where $F_k(x^{(k)}) : R^{p_k} \mapsto R$ can be interpreted as the activity level associated with the kth pathway as determined by the genomic measurements of the p_k genes in this pathway. For a binary phenotype such as disease status (for example, normal versus cancerous tissues), they assume a logistic model for Y, i.e.,

$$P_r(Y = 1|x, Z) = \frac{\exp(2(F(x) + \gamma Z))}{1 + \exp(2(F(x) + \gamma Z))}, \tag{2.2}$$

where $Y = 1$ for diseased individual and $Y = -1$ for normal individual, Z is the vector of other patient-specific covariates which is modelled parametrically with coefficient γ. To obtain an additive model with the form of (2.1), Wei and Li (2007) propose to use the regression trees as base learners, and a pathway-based gradient descent technique (Friedman, 2001) as the boosting algorithm. The goal is to estimate the function $F : R^p \mapsto R$, minimizing an expected loss function $E[\ell(Y, F(X))]$, where $\ell(\cdot, \cdot)$ is a loss function and p is the dimension of X that is less than or equal to $\sum_{k=1}^{K} p_k$. Estimation of such an $F(\cdot)$ from data, $\{(y_i, x_i), i = 1, 2, \cdots, n\}$, can be done via a constrained minimization of the empirical loss

$$n^{-1} \sum_{i=1}^{n} \ell(y_i, F(x_i)), \tag{2.3}$$

by functional gradient descent, where $\ell(y_i, F(x_i))$ is the loss function for the ith observation (y_i, x_i). For binary disease status, the loss function is defined

as (c.f., Wei and Li, 2007)

$$L(y, F(x)) = \sum_{i=1}^{n} \ell(y_i, F(x_i)) = \sum_{i=1}^{n} \log(1 + \exp(-2y_i F(x_i))). \tag{2.4}$$

Pathways-based gradient descent boosting (GDB) procedure reads as follows:

1. Initialization: $F^{(0)}(x) = 0$, $F_k^{(0)}(x_k) = 0$, $k = 1, 2, \cdots, K$

   Repeat boosting steps for $m = 1$ to M, do:

2. calculating the gradients

$$\tilde{y}_i = 2y_i/(1 + \exp(2y_i F^{(m-1)}(x_i)));$$

3. Fitting trees to the gradient vector using $x^{(k)}$: let $h_k(x_i^{(k)}; \alpha)$ be the base learner procedure,

$$(\alpha^{(k)}, \beta^{(k)}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} [\tilde{y}_i - \beta h_k(x_i^{(k)}; \alpha)]^2, \quad k = 1, \cdots, K;$$

   let $k^* = \arg\min_k \sum_{i=1}^{n} [\tilde{y}_i - \beta^{(k)} h_k(x_i^{(k)}; \alpha^{(k)})]^2$.

4. Line search over ρ for the pathway k^* selected in Step 3,

$$\rho_m = \arg\min_{\rho} \sum_{i=1}^{n} L(y_i, F^{(m-1)}(x_i) + \rho h_k(x_i^{(k^*)}; \alpha^{(k^*)})).$$

5. Updating the function with ν being the learning rate,

$$F_{k^*}^{(m)}(x^{(k^*)}) = F_{k^*}^{(m-1)}(x^{(k^*)}) + \nu\rho_m h_{k^*}(x_i^{(k^*)}; \alpha^{(k^*)})),$$

$$F^{(m)}(x) = F^{(m-1)}(x) + F_{k^*}^{(m)}(x^{(k^*)}).$$

Wei and Li (2007) propose to apply the cross-validation on the error rates for the logistic model to determine the number of boosting steps M. After M is determined and the function $F(\cdot)$ (a linear combination of trees) is estimated, the issue of identifying important pathways and genes is assessed by calculating the importance scores.

First, for the influence of each gene, the importance score using the trees constructed on the kth pathway is:

$$\hat{I}_{lk} = \frac{1}{M_k} \sum_{m=1}^{M_k} I_l^2(T_{mk}), \tag{2.5}$$

where M_k is the number of times that the kth pathway was selected in Step 3 of the proposed pathway-based boosting algorithm, and T_{mk} is the mth tree built based on the kth pathway. Here $I_l^2(T)$, for a single tree T, is defined in the form (Breiman et al. 1984)

$$I_l^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2 I(v(t) = l),$$

as a measurement of relevance for each predictor variable X_l for the tree T with J nodes, where the sum is over the $J-1$ internal nodes, $I(\cdot)$ is an indicator function and $v(t)$ is the splitting variable associated with the tth node, and $\hat{i}_t^2$ is the empirical improvement in squared error risk as a result of a split at node t over the $J-1$ internal notes of the tree; see Wei and Li (2007, section 3.2) for details.

Second, for each pathway, the average of importance scores for genes selected within a pathway, Wei and Li (2007) assign it as the pathway importance score used as a measure of importance of the pathway to the phenotype. The most influential variable or pathway is given a score of 1, and the estimated importance scores of others are scaled accordingly.

Professor Luba Kalaydjieva from the Western Australian Institute for Medical research suggests the following biological pathway assumption for the WAFSS data set: Schizophrenia, specifically patients with a cognitive deficit, may result from impaired brain development (neuronal migration and synapse development) and/or impaired function of the adult synapse (plasticity) and that our network, containing a large number of genes encoding physically interacting proteins may be involved in schizophrenia pathogenesis - in terms of individual gene effects or multiple effects and interactions, together with three pathways showing the biological interactions between these genes. We rewrite these pathway information/assumptions in Table 2.1 below.

Table 2.1: Pathway assumptions for the WAFSS schizophrenia data set

	Synaptic organization	Plasticity	Neuronal migration	SNP number
ApoE	1	1	0	4
ApoER2	1	1	1	31
ATF4	1	0	0	5
ATF5	1	0	0	8
BDNF	1	1	0	10
CDK5	0	0	1	4
CITRON	1	0	0	30
DAB1	1	1	1	271
DCX	0	0	1	2
DISC1	1	1	1	125
DLG2	1	1	0	173
DLG4	1	1	0	9
FEZ1	1	0	0	13
LIS1	0	0	1	8
MAP1A	1	0	0	3
NUDE	0	0	1	10
NUDEL	0	0	1	6
PDE4B	0	1	0	77
RELN	1	1	1	170
VLDLR	1	1	1	29
Sum	14	10	10	988

As listed in Table 2.1, the names of the three pathways are synaptic organization, plasticity, and neuronal migration. If a gene is considered as part of some pathway, it will be assigned as 1, otherwise 0. The total SNP numbers involved in the three pathways are 988. From Li's website (http://www.cceb.upenn.edu/ hli/NPR), the R codes for implementing the NPR model can be downloaded.

We start from schizophrenia CD subtype. After customizing the R codes for the WAFSS schizophrenia CD subtype data and the Luba's pathways assumption, we obtain:

1. Pathway's importance ranking, given in Table 2.2.

Table 2.2: Pathway's importance ranking for the WAFSS schizophrenia CD subtype data set

Pathway	Relative Importance
Synaptic organization	1
Plasticity	0.942699
Neuronal migration	0.353067

From Table 2.2, we see that synaptic organization is the pathway that is most associated with schizophrenia CD subtype. According to Professor Luba's view (personal communication), the result obtained is reasonable and helpful in the understanding of the CD subtype pathways.

Table 2.3: SNP importance score of CD subtype for the WAFSS schizophrenia data set

Gene	SNP	RI score
DISC1	rs999710	1
DAB1	rs486706	0.777193
DLG2	rs790379	0.633021
DAB1	rs694060	0.467058
DISC1	rs2772122	0.418253
DAB1	rs852773	0.412278
DISC1	rs2806465	0.263882
DLG2	rs485199	0.255886
RELN	rs2283029	0.244975
RELN	rs2711881	0.228626
DAB1	rs534455	0.22041
DISC1	rs4658890	0.203394
BDNF	rs283531	0.199595
APOE	rs439401	0.14177
DLG4	rs17203281	0.126849
RELN	rs802787	0.122482
RELN	rs661575	0.116015
VLDLR	rs1454626	0.112308
ApoER2	rs3737983	0.104825

2. SNP importance score for CD subtype, given in Table 2.3.

   In Table 2.3, SNPs:rs999710, rs2806465, rs439401, rs17203281, rs1454626 and rs3737983 are cited by other medical research papers as related to various diseases (not necessary schizophrenia alone). We can see that NPR model does select some "hot" SNPs from a pool of 988 SNPs.

We then proceed from CD subtype to schizophrenia. Similar results have

Table 2.4: Important SNPs for schizophrenia selected by NPR model

Gene	SNP	Gene	SNP
APOE	rs439401	CITRON	rs2991515
APOER2	rs2297660	DISC1	rs999708
CITRON	rs1077451	DLG2	rs790379
CITRON	rs7960673	DLG4	rs1165023
CITRON	rs534455	RELN	rs1024805
CITRON	rs694060	RELN	rs1149612
CITRON	rs267647	VLDLR	rs1454626

been obtained. The synaptic organization pathway is the pathway most associated with schizophrenia and the plasticity pathway comes second. Important SNPs for schizophrenia selected by NPR model are listed in Table 2.4:

The analysis carried out using the NPR model as mentioned above does give biologists the pathway importance assessment and the SNP selection that they are interested in. However, the final estimate of the function $F(\cdot)$ in (2.1) is a linear combination of regression trees. It is very much like a "black box": we know the input and the output, but we do not know what the black box is. Biologists in WAFSS are not satisfied with the black box $F(\cdot)$.

Classification is the central topic of this thesis. It is closely linked to medical diagnosis. If we select the most-associated 12 SNPs (RI score from 1 to 0.2033 in Table 2.3) using the linear logistic model (i.e.$F(\cdot)$ in (2.2) is taken as linear) for CD subtype class prediction, the sensitivity is calculated

to be 58.55%, while specificity is 41.52%. A theoretical optimal prediction aims to achieve 100% sensitivity and 100% specificity. Therefore, using SNPs selected by NPR model in Table 2.3 as a group to explain schizophrenia is quite disappointing because specificity is even under 50%, while sensitivity is only slightly over 50%. The results for general schizophrenia prediction seem a bit better. Using important SNPs for schizophrenia selected by NPR model in Table 2.4, the sensitivity is 89.23% but specificity is 32.75%, which is still too low. See Table 2.5 for the details.

Although the NPR model successfully selects some important SNPs, the combination of these important SNPs used in disease status prediction is not satisfactory. How to select a group of SNPs that can give a reasonable prediction of schizophrenia risk remains unsolved.

2.4 Linear logistic discriminant rule based on preliminary analysis

As mentioned before, we have calculated the odds ratio for all 1022 SNPs using our R codes written at the preliminary analysis stage. Further research has been done based on reducing the number of variables by SNPs selection.

In the sense of Biology, if a SNP is selected in the regression model without meeting the criterion of the OR p-value being less than 0.05, then it is hardly convincing at all. For the schizophrenia and the CD subtype, 40 respective

SNPs are selected and they are used as regression variables.

We then use stepwise linear logistic regression via backward elimination (i.e. $F(\cdot)$ in (2.2) is taken as linear) procedure to further reduce the number of variables. In order to compare with the NPR model on CD subtype disease status prediction, we stop the variable selection when there are 12 variables (SNPs) being selected in the logistic model. Based on such a selected model, the sensitivity of the prediction is calculated, yielding 65.7%, while specificity is 73%. These represent a significant improvement over the NPR model, where the sensitivity is 58.55% and the specificity is 41.52%.

We further proceed to apply the stepwise logistic regression method on general schizophrenia. Again, we stop the variable selection when the number of variables is reduced to 12 variables. The sensitivity of the schizophrenia prediction obtained is 89.53% and the specificity is 35.67%. See Table 2.5 for the details.

Table 2.5: Comparison of class prediction by the NPR model with the linear logistic model(referred to GLM below) for the WAFSS schizophrenia data set

Data	Measure	NPR	GLM
CD subtype	Sensitivity	58.55%	65.7%
	Specificity	41.52%	73%
general schizophrenia	Sensitivity	89.23%	89.53%
	Specificity	32.75%	35.67%

In Table 2.5, we can see that the linear logistic discriminant rule(GLM) outperforms the NPR model on disease status prediction using SNP genotype. However, it is noted that the specificity of the linear logistic discriminant rule on general schizophrenia prediction is only 35.67%, even less than 50%. Therefore, it is still not a satisfactory predictor. This motivates us to propose a new model for class prediction. It is called the threshold index models and it will be discussed in the next chapter.

Chapter 3

Threshold index nonlinear logistic regression

3.1 Introduction

As indicated in Chapter 2, Wei and Li (2007)'s NPR model cannot characterise satisfactorily the genetically complex nature of schizophrenia SNP datasets. It performs even worse than the linear logistic regression model. Obviously, the linear logistic regression cannot characterise the nonlinear features of genetical complexity. Therefore, it is important to propose a new model that can capture the complex characters of SNP datasets.

In this chapter, a new class of threshold index logistic regression (TILoR) models, including partially linear and additive TILoR models, is proposed.

This model aims at capturing the nonlinear effects from SNP data. For SNP data, we shall determine which SNP should enter into the regime indices of the model and what nonlinear form should be used in the model through threshold approximation. This nonlinear model for SNP selection is generally verifiable and is also much easier to obtain than using the combination of the phenotype and genotype variables. This happens because the phenotype requires participants to answer many complicated and time-consuming questionnaires.

The new models will be introduced in Section 3.2, the maximum likelihood estimation for the proposed models is provided in Section 3.3, and the finite sample performance of the suggested estimators for the proposed models is studied in Section 3.4.

3.2 Threshold Index nonlinear logistic regression models

Logistic linear regression models are based on a fundamental assumption that

$$\log\left\{\frac{P(Y_i=1|\mathbf{X}_i)}{1-P(Y_i=1|\mathbf{X}_i)}\right\} = a + \mathbf{b}^T\mathbf{X}_i, \tag{3.1}$$

which is a linear function of the p regressor variables $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})^T$. It has been shown that this class of models is useful in modelling the main

effects of the gene expression data. However, it cannot capture the complex nonlinear interaction effects among the genes. General tree regression models belong to a class of nonlinear models which is used in the biostatistics literature (see, for example, Zhang and Singer, 1999) to characterise such nonlinear interactions. However, this class of models is too general to reveal specific relationships that are available. This is particularly so when the dimension, p, of $\mathbf{X}$ is very high.

Different from the linear models, many semiparametric nonlinear models have been proposed to approximate the nonlinear case when the regressor variables are continuous. See, for example, Fan and Gijbels (1996). As an example, let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, where $\mathbf{X}_2$ consists of continuous random variables. Thus, we consider a partially linear logistic regression model in the form of

$$\log\left\{\frac{P(Y_i = 1|\mathbf{X}_i)}{1 - P(Y_i = 1|\mathbf{X}_i)}\right\} = \tilde{\boldsymbol{\alpha}}^T \mathbf{X}_{1,i} + \tilde{g}(\mathbf{X}_{2,i}). \tag{3.2}$$

Partially linear regression is very popular in many applications; see e.g., Hardle, Liang and Gao (2000) and You and Zhou (2007). In model (3.2), $\mathbf{X}_1$ and $\mathbf{X}_2$, which are the sub-vectors of $\mathbf{X}$, are of dimensions p_1 and p_2, respectively, and $\tilde{\boldsymbol{\alpha}}$ is a parameter vector of dimension p_1, and $\tilde{g}$ is a p_2-dimensional nonlinear function which is estimated by some nonparametric methods. If the dimension p_2 is greater than 3, it may still suffer from the curse of dimensionality and some further semiparametric structure can be

applied. Fan *et al.* (1997) propose the generalised partially linear index models, which are in the form of the logistic case given below (see also Yi et al., 2009),

$$\log\left\{\frac{P(Y_i=1|\mathbf{X}_i)}{1-P(Y_i=1|\mathbf{X}_i)}\right\}=\tilde{\boldsymbol{\alpha}}^T\mathbf{X}_{1,i}+\tilde{g}(\tilde{\boldsymbol{\beta}}^T\mathbf{X}_{2,i}), \tag{3.3}$$

where $\tilde{\boldsymbol{\beta}}$ is an index vector of dimension p_2 satisfying $\|\tilde{\boldsymbol{\beta}}\|=1$ for model identifiability, and $\tilde{g}$ becomes a one-dimensional nonlinear function circumventing the curse of dimensionality. These models allow us to more effectively deal with nonlinear behaviour for the continuous $\mathbf{X}_2$. The variables taking discrete values are categorised into $\mathbf{X}_1$ modelled in the linear form. This means that for categorial regressor variables like SNP data, we can not apply the above models to capture the nonlinear interaction effects in the SNPs.

In this chapter, we therefore propose two kinds of nonlinear models, where the nonlinear effects for SNP data that are categorial are to be captured. In models (3.2) and (3.3) it is assumed in advance that which regressor variables will enter the models in the form of linear function and nonlinear function, respectively. In our models, variables which are to enter in the form of a nonlinear function so as to capture the nonlinear effects in categorial regressors are determined by data, and the nonlinear function in the stepwise linear form is also determined from the data. We will model the nonlinear effects by applying the idea of the threshold models of Tong (1990). One of the proposed models, which is referred to as partially linear threshold index

logistic regression (PL-TILoR) model, is as follows:

$$\log\left\{\frac{P(Y_i=1|\mathbf{X}_i)}{1-P(Y_i=1|\mathbf{X}_i)}\right\}=\boldsymbol{\alpha}^T\mathbf{X}_i+g(\boldsymbol{\beta}^T\mathbf{X}_i), \tag{3.4}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are p-dimensional unknown parameters with $\boldsymbol{\beta}$ being an index vector satisfying $\|\boldsymbol{\beta}\|=1$. Suppose that $\mathbf{X}_i$ is a continuous random vector. In this case, the model could be seen as a generalized version of the extended partially linear single-index model of Xia *et al.* (1999), with $\boldsymbol{\alpha}^T\boldsymbol{\beta}=0$ for model identifiability, where g is estimated nonparametrically. However, we note that the SNP data is categorical, i.e., $\mathbf{X}_i$ is not a continuous random vector, rather its components are all categorical (for details, see Chapter 4). Thus, the methodology developed by Xia *et al.* (1999) cannot be applied here. It is differently from the traditional tree regression for which it is directly applicable to the regressor vector $\mathbf{X}_i$ (c.f., Wei and Li 2007; Zhang and Singer 1999). However, the traditional tree regression may suffer from the curse of dimensionality and may perform even worse than the linear model as showed in Chapter 2. we propose to apply the idea of tree regression to the function g in (3.4), which is a one-dimensional nonlinear function. More specifically, we model g as a stepwise linear function through threshold effect (Tong, 1990) given by

$$g(z)=(b_1z+b_2)I_{\{z\le c\}}+(b_3z+b_4)I_{\{z>c\}}, \tag{3.5}$$

where b_i's and c are unknown parameters, which are to be estimated, and I_A

is an indicator function of the set A. Therefore, model (3.4) can be expressed as:

$$\log\left\{\frac{P(Y_i=1|\mathbf{X}_i)}{1-P(Y_i=1|\mathbf{X}_i)}\right\} = \boldsymbol{\alpha}^T\mathbf{X}_i + (b_1\boldsymbol{\beta}^T\mathbf{X}_i + b_2)I_{\{\boldsymbol{\beta}^T\mathbf{X}_i\le c\}} + (b_3\boldsymbol{\beta}^T\mathbf{X}_i + b_4)I_{\{\boldsymbol{\beta}^T\mathbf{X}_i> c\}}, \tag{3.6}$$

with $\|\boldsymbol{\beta}\| = 1$, $\boldsymbol{\alpha}^T\boldsymbol{\beta} = 0$, and the first non-zero component of $\boldsymbol{\beta}$ being positive, for model identifiability.

In practice, the first component could also be nonlinear. In this case, model (3.4) becomes

$$\log\left\{\frac{P(Y_i=1|\mathbf{X}_i)}{1-P(Y_i=1|\mathbf{X}_i)}\right\} = g_1(\boldsymbol{\alpha}^T\mathbf{X}_i) + g_2(\boldsymbol{\beta}^T\mathbf{X}_i), \tag{3.7}$$

where $\|\boldsymbol{\alpha}\| = 1$, $\|\boldsymbol{\beta}\| = 1$, $\boldsymbol{\alpha}^T\boldsymbol{\beta} = 0$, and the first non-zero components of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are positive, for model identifiability, and g_1 and g_2 are two one-dimensional nonlinear functions which are modelled by stepwise linear functions through threshold effects as follows:

$$g_k(z) = (b_{k1}z + b_{k2})I_{\{z\le c_k\}} + (b_{k3}z + b_{k4})I_{\{z> c_k\}},\ k = 1, 2, \tag{3.8}$$

where b_{ki}'s and c_k's are unknown parameters, which are to be estimated. Thus, (3.7) and (3.8) form an additive threshold index logistic regression

(A-TILoR) model

$$\begin{aligned}
&\log\left\{\frac{P(Y_i=1|\mathbf{X}_i)}{1-P(Y_i=1|\mathbf{X}_i)}\right\}\\
&=(b_{11}\boldsymbol{\alpha}^T\mathbf{X}_i+b_{12})I_{\{\boldsymbol{\alpha}^T\mathbf{X}_i\le c_1\}}+(b_{13}\boldsymbol{\alpha}^T\mathbf{X}_i+b_{14})I_{\{\boldsymbol{\alpha}^T\mathbf{X}_i>c_1\}}\\
&\quad+(b_{21}\boldsymbol{\beta}^T\mathbf{X}_i+b_{22})I_{\{\boldsymbol{\beta}^T\mathbf{X}_i\le c_2\}}+(b_{23}\boldsymbol{\beta}^T\mathbf{X}_i+b_{24})I_{\{\boldsymbol{\beta}^T\mathbf{X}_i>c_2\}}, \qquad (3.9)
\end{aligned}$$

with $\|\boldsymbol{\alpha}\|=1$, $\|\boldsymbol{\beta}\|=1$, $\boldsymbol{\alpha}^T\boldsymbol{\beta}=0$, and the first non-zero components of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ being positive. Obviously, the A-TILoR model (3.9) is more general than the PL-TILoR model (3.6).

We shall show in Chapter 4 that model (3.9) performs quite well in analysing schizophrenia data; it beats the linear logistic regression model, and hence the NPR model of Wei and Li (2007), as demonstrated in Chapter 2. An intuitive reason why the A-TILoR model can performs better than the NPR model for the WAFSS schizophrenia SNP datasets is as follows. The A-TILoR model appears similar to the tree-regression based NPR model in that both have stepwise constant or linear function over each of the partitioned regressor areas, and thus well capture the nonlinearity feature of real datasets. But if the dimension p of the covariate vector $\mathbf{X}$ is high (e.g., $p=40$ as considered in Chapter 4), then the NPR model suffers from severe curse of dimensionality due to its area partition based on each of the individual covariate, while our proposed A-TILoR model in the above does overcome this drawback of the NPR owing to its area partition based on the index

variables $\boldsymbol{\alpha}^T\mathbf{X}_i$ and $\boldsymbol{\beta}^T\mathbf{X}_i$ only, well capturing the nonlinearity feature of the WAFSS schizophrenia SNP datasets.

It may seem tempting to extend model (3.9) by adding more stepwise linear functions along other more directions of the covariate vector $\mathbf{X}$ of dimension p. If this is the case, then how to choose the number of the stepwise linear functions is important in practice, which can be solved by applying the Akaike (1973)'s information criterion (AIC), with the model of the smallest AIC value being chosen. However, further extension of model (3.9) will greatly increase the number of unknown parameters in the model for a large p, e.g., $p = 40$ as considered in Chapter 4. It leads to the rejection of the more extended model for the WAFSS schizophrenia SNP datasets by the Akaike (1973)'s information criterion (AIC). We therefore do not pursue such a straightforward extension here.

3.3 Maximum likelihood estimation

In this section, we propose a maximum likelihood methodology to estimate the unknown parameters in (3.6) and (3.9).

First of all, we look at the PL-TILoR model (3.6). Let

$$\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, b_1, b_2, b_3, b_4, \boldsymbol{\beta}^T, c)^T$$

and

$$\phi_i(\boldsymbol{\theta}) = \boldsymbol{\alpha}^T \mathbf{X}_i + (b_1 \boldsymbol{\beta}^T \mathbf{X}_i + b_2) I_{\{\boldsymbol{\beta}^T \mathbf{X}_i \leq c\}} + (b_3 \boldsymbol{\beta}^T \mathbf{X}_i + b_4) I_{\{\boldsymbol{\beta}^T \mathbf{X}_i > c\}}.$$

From model (3.6), it follows that

$$P(Y_i = 1 | \mathbf{X}_i) = \frac{\exp\{\phi_i(\boldsymbol{\theta})\}}{1 + \exp\{\phi_i(\boldsymbol{\theta})\}}. \tag{3.10}$$

Therefore, the likelihood can be expressed as:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} [P(Y_i = 1 | \mathbf{X}_i)]^{Y_i} [1 - P(Y_i = 1 | \mathbf{X}_i)]^{1 - Y_i} \\ &= \prod_{i=1}^{n} \left(\left[\frac{\exp\{\phi_i(\boldsymbol{\theta})\}}{1 + \exp\{\phi_i(\boldsymbol{\theta})\}} \right]^{Y_i} \left[\frac{1}{1 + \exp\{\phi_i(\boldsymbol{\theta})\}} \right]^{1 - Y_i} \right). \end{aligned} \tag{3.11}$$

So, the log-likelihood can be written as:

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \log L(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} Y_i \log \left[\frac{\exp\{\phi_i(\boldsymbol{\theta})\}}{1 + \exp\{\phi_i(\boldsymbol{\theta})\}} \right] \\ &\quad + \sum_{i=1}^{n} (1 - Y_i) \log \left[\frac{1}{1 + \exp\{\phi_i(\boldsymbol{\theta})\}} \right]. \end{aligned} \tag{3.12}$$

Maximizing the log-likelihood (3.12) with respect to

$$\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, b_1, b_2, b_3, b_4, \boldsymbol{\beta}^T, c)^T$$

subject to $\|\boldsymbol{\beta}\| = 1$ and $\boldsymbol{\alpha}^T\boldsymbol{\beta} = 0$ leads to the MLE (maximum likelihood estimator) $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$.

Similarly, we construct the MLE for the A-TILoR model (3.9). Let

$$\boldsymbol{\vartheta} = (b_{11}, b_{12}, b_{13}, b_{14}, , \boldsymbol{\alpha}^T, b_{21}, b_{22}, b_{23}, b_{24}, \boldsymbol{\beta}^T, c_1, c_2)^T$$

and

$$\begin{aligned}\varphi_i(\boldsymbol{\vartheta}) = (b_{11}\boldsymbol{\alpha}^T\mathbf{X}_i + b_{12})I_{\{\boldsymbol{\alpha}^T\mathbf{X}_i \le c_1\}} + (b_{13}\boldsymbol{\alpha}^T\mathbf{X}_i + b_{14})I_{\{\boldsymbol{\alpha}^T\mathbf{X}_i > c_1\}} \\ + (b_{21}\boldsymbol{\beta}^T\mathbf{X}_i + b_{22})I_{\{\boldsymbol{\beta}^T\mathbf{X}_i \le c_2\}} + (b_{23}\boldsymbol{\beta}^T\mathbf{X}_i + b_{24})I_{\{\boldsymbol{\beta}^T\mathbf{X}_i > c_2\}}.\end{aligned}$$

The log-likelihood can be expressed as:

$$\begin{aligned}&\ell_A(\boldsymbol{\vartheta}) = \log L_A(\boldsymbol{\vartheta}) \\ &= \sum_{i=1}^{n} Y_i \log\left[\frac{\exp\{\varphi_i(\boldsymbol{\vartheta})\}}{1+\exp\{\varphi_i(\boldsymbol{\vartheta})\}}\right] \\ &\qquad + \sum_{i=1}^{n}(1-Y_i)\log\left[\frac{1}{1+\exp\{\varphi_i(\boldsymbol{\vartheta})\}}\right].\end{aligned} \tag{3.13}$$

Maximizing the log-likelihood (3.13) with respect to

$$\boldsymbol{\vartheta} = (b_{11}, b_{12}, b_{13}, b_{14}, , \boldsymbol{\alpha}^T, b_{21}, b_{22}, b_{23}, b_{24}, \boldsymbol{\beta}^T, c_1, c_2)^T$$

subject to $\|\boldsymbol{\alpha}\| = 1$, $\|\boldsymbol{\beta}\| = 1$ and $\boldsymbol{\alpha}^T\boldsymbol{\beta} = 0$ leads to the MLE $\hat{\boldsymbol{\vartheta}}$ of $\boldsymbol{\vartheta}$.

Note that the log-likelihood (3.12) is not differentiable with respect to c

and $\boldsymbol{\alpha}$, and the log-likelihood (3.13) is not differentiable with respect to c_1 and c_2 as well as $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Therefore the widely used iteration procedure in optimization such as Newton-Raphson algorithm does not work here. We apply the downhill simplex method for the maximization of the log-likelihood (3.12) or (3.13), which does not require the multi-dimensional objective function of the optimisation to be differentiable; for details, the reader is referred to Press et al. (2002, page 413) on the method and code.

In our numerical experiments, we used the R version of the standard downhill simplex method, translated from the C code of Press et al. (2002, page 413). According to our experiences, this algorithm works rather stably in convergence with well specified $(D+1)$ initial values of the vector $\boldsymbol{\theta}$ or $\boldsymbol{\vartheta}$ (where D denotes for the dimension of $\boldsymbol{\theta}$ or $\boldsymbol{\vartheta}$), for which we need experimental tries to achieve a global maximum as done in using other optimisation algorithms.

3.4 Simulation studies: Finite sample performance

In order to examine the finite sample performance of the proposed maximum likelihood estimate of the unknown parameters in the A-TILOR model, we carry out some experiments via Monte Carlo simulations. In this section, we report the results with the setting of the model similar to that in our real

data analysis in the next chapter.

3.4.1 Simulation model

In real application of genomic data analysis, the dimension p of the regressor vector is quite large, and the regressor variables are categorical data representing different types of gene expression. To accommodate these scenarios, we consider the A-TILOR model, used for simulation, of the form (3.9) with $p = 39$. Let $\mathbf{X} = (X_1, X_2, \cdots, X_p)$, with $X_j \sim Binomial(2, q_j)$, and let $q_i = (1 + (j-1)/p)/2$, for $j = 1, 2, \cdots, p$. Assume that X_j's are linearly independent with each other. We also take the parameters in the model to

be equal to the estimated values for the CD data in Chapter 4, that is

$$
\begin{aligned}
b_1 =&(b_{11}b_{12}, b_{13}, b_{14}) = (0.1451343, -0.4199375, 2.1645830, 0.3385306),\\
c_1 =&0.1817811,\\
\alpha =&(0.160532828, 0.145342634, -0.105307522, 0.080778184, 0.213884518,\\
&-0.050812375, 0.330436703, -0.128131250, 0.129550296, -0.192552871,\\
&-0.026473093, -0.006149899, 0.068069191, -0.288425609, 0.009740996,\\
&0.123212627, -0.015222646, 0.170128772, -0.087436472, -0.143557690,\\
&0.046833668, -0.023903015, -0.215940820, -0.063833874, -0.234875792,\\
&0.173082853, 0.259160282, 0.072185675, 0.073419002, -0.444332340,\\
&-0.020977710, -0.149498653, 0.126362136, 0.108406749, -0.057101111,\\
&0.061013646, -0.123298587, 0.015689519, 0.244328756),\\
b_2 =&(b_{21}b_{22}, b_{23}, b_{24}) = (0.6700960, 1.0459350, -1.6852746, -0.7785389),\\
c_2 =&0.2594507,
\end{aligned}
$$

$$\begin{aligned}\beta =&(0.15020062, -0.13162367, -0.46794802, 0.01625887, 0.12274019,\\ &0.15892854, 0.13511445, 0.14823644, 0.24054483, 0.01113274,\\ &-0.29986520, -0.15894115, 0.02485366, -0.02145778, 0.28276706,\\ &-0.12458040, 0.13349062, -0.11010719, 0.15720584, -0.06474037,\\ &-0.04085526, -0.09366944, -0.08138479, 0.07599860, 0.17990750,\\ &0.03596593, 0.11245335, -0.09818820, 0.15916573, 0.17548150,\\ &0.16646456, 0.02726669, -0.20494686, 0.05136877, -0.05899783,\\ &-0.02131067, 0.13939863, -0.28623061, 0.13580588).\end{aligned}$$

We consider the simulated sample of size $n = 200$ and $n = 323$, respectively, where $n = 323$ is the sample size of the CD data set that we will analyse in Chapter 4. We first simulate the random vector $\mathbf{X}_i$ with its jth component $X_{i,j} \sim Bin(2, q_j)$. Then, we calculate $P(Y_i = 1|\mathbf{X}_i)$ according to (3.9). Thus, we simulate Y_i from the Bernoulli trial with probability equal to $P(Y_i = 1|\mathbf{X}_i)$. For each simulated sample, we apply the suggested maximum likelihood method to estimate the parameters. We repeat the simulation 100 times for each of the two cases with sample sizes $n = 200$ and $n = 323$.

3.4.2 Results

In Figures 3.1 and 3.2, the boxplots of the estimates of the parameters in g_1, α, g_2 and β based on 100 simulations are displayed for the cases with

sample sizes $n = 200$ and $n = 323$, respectively. In order to assess the precision of the estimate for each of the parameters, the absolute errors of the estimates of the parameters based on 100 simulations are depicted in boxplot in Figures 3.3 and 3.4 for the cases with sample sizes corresponding to those in Figures 3.1 and 3.2, respectively.

From these figures, we may conclude that

- Obviously, as the sample size increases, the absolute error of the estimate decreases. Comparing Figure 3.2 with Figure 3.1, we see that the boxplot becomes much narrower for each parameter in Figure 3.2 than those in Figure 3.1. This also clearly follows by comparing Figure 3.4 with Figure 3.3. It seems quite apparent that the sample size $n = 323$ used in Figure 3.2 and Figure 3.4 is acceptable for the model even with the regressor vector of the dimension $p = 39$.

- From all the figures, it is clear that both threshold parameters c_1 and c_2 can be estimated rather precisely, even for the case with sample size $n = 200$.

- It is also clear that the parameters of the vectors α and β can be estimated more precisely than the parameters b_{1j}'s and b_{2j}'s in g_1 and g_2, respectively. In Figures 3.1 and 3.3, the estimators of the parameters in g_1 and g_2 could be quite poor in some simulations. However, with $n = 323$, the accuracy of the estimators of these parameters are much

improved as seen in Figures 3.2 and 3.4.

Overall, the simulation results provide a strong support to our real schizophrenia data analysis to be carried out in Chapter 4, with the proposed A-TILOR model (3.9) applied for the case where the sample size $n = 323$ and the number of unknown parameters is up to $2 \times (p + 4) = 86$ when $p = 39$.

3.5 Bootstrapping method for estimating the standard deviations of the estimates

In many applications, we need to evaluate whether the estimated value of an unknown parameter is significantly away from zero or not, i.e., testing whether we can reject the null hypothesis that the estimated parameter is equal to zero. This requires the knowledge of the standard deviation of the estimator of each parameter.

One way to estimate the standard deviation of the estimator of each unknown parameter is through estimating the asymptotic variance of the estimator of the parameter, which can be established by following the argument of Chan (1993). However, asymptotic variance is based on the assumption that the sample size tends to infinity, which is difficult to apply in many situations in practice such as in our schizophrenia analysis. For example, in Chapter 4, we have up to 86 unknown parameters but we only have a sample with size roughly about 300–500. Clearly, the sample size may not

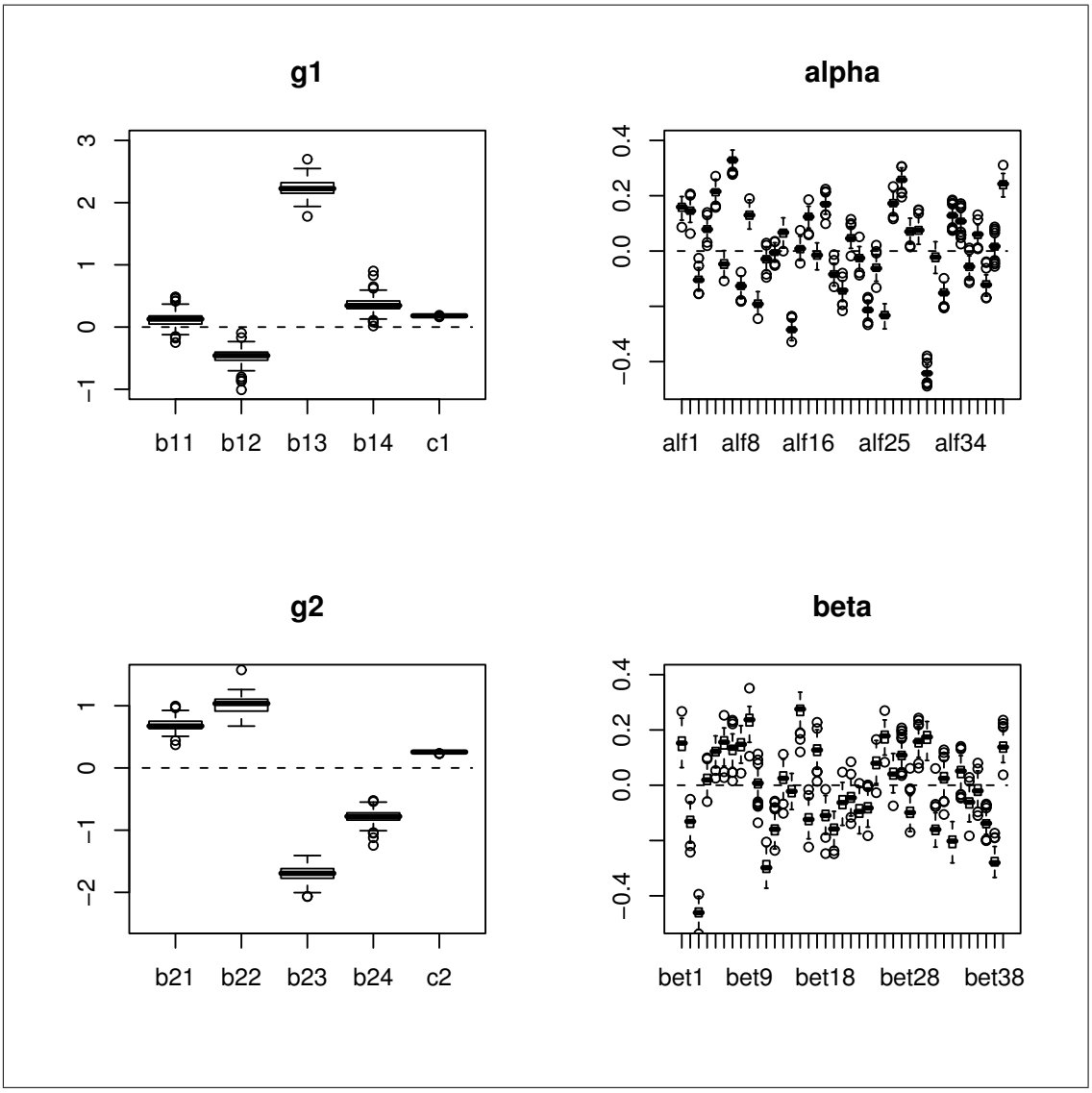

Figure 3.1: Boxplot of the estimates of the parameters in g_1, α, g_2 and β based on 100 simulations: $n = 200$.

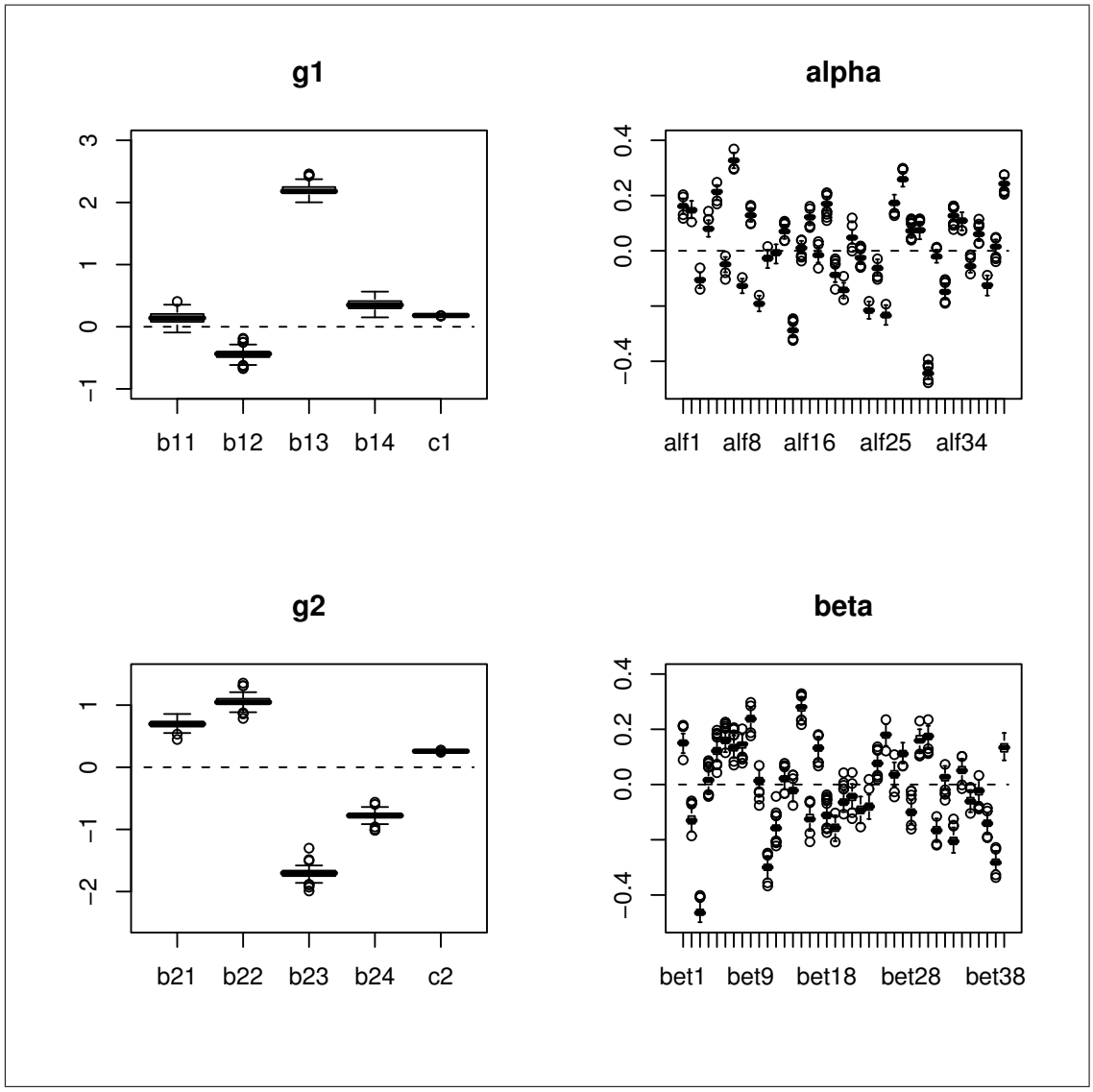

Figure 3.2: Boxplot of the estimates of the parameters in g_1, α, g_2 and β based on 100 simulations: $n = 323$.

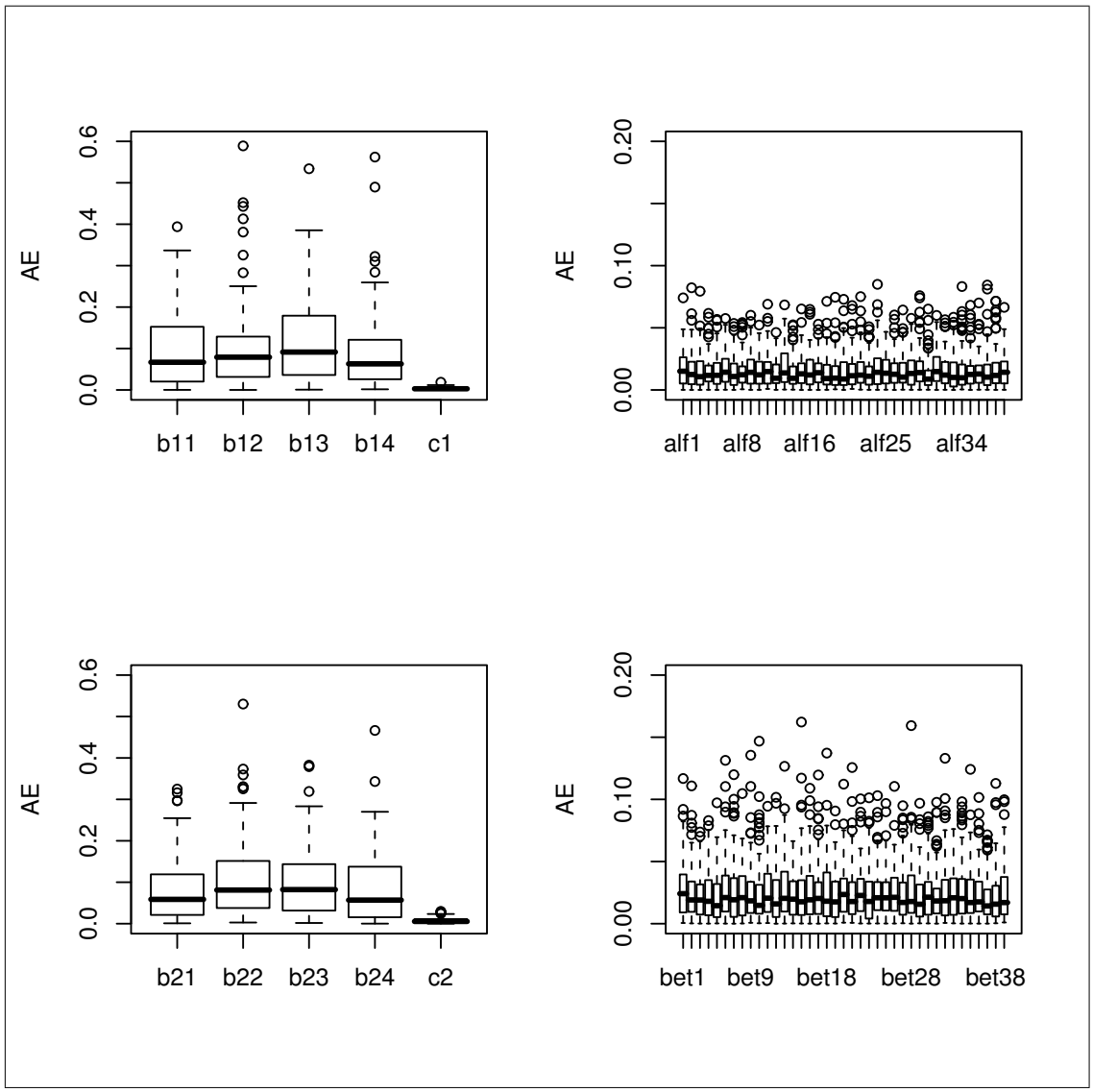

Figure 3.3: Boxplot of the absolute errors (AEs) of the estimates of the parameters in g_1, α, g_2 and β based on 100 simulations: $n = 200$.

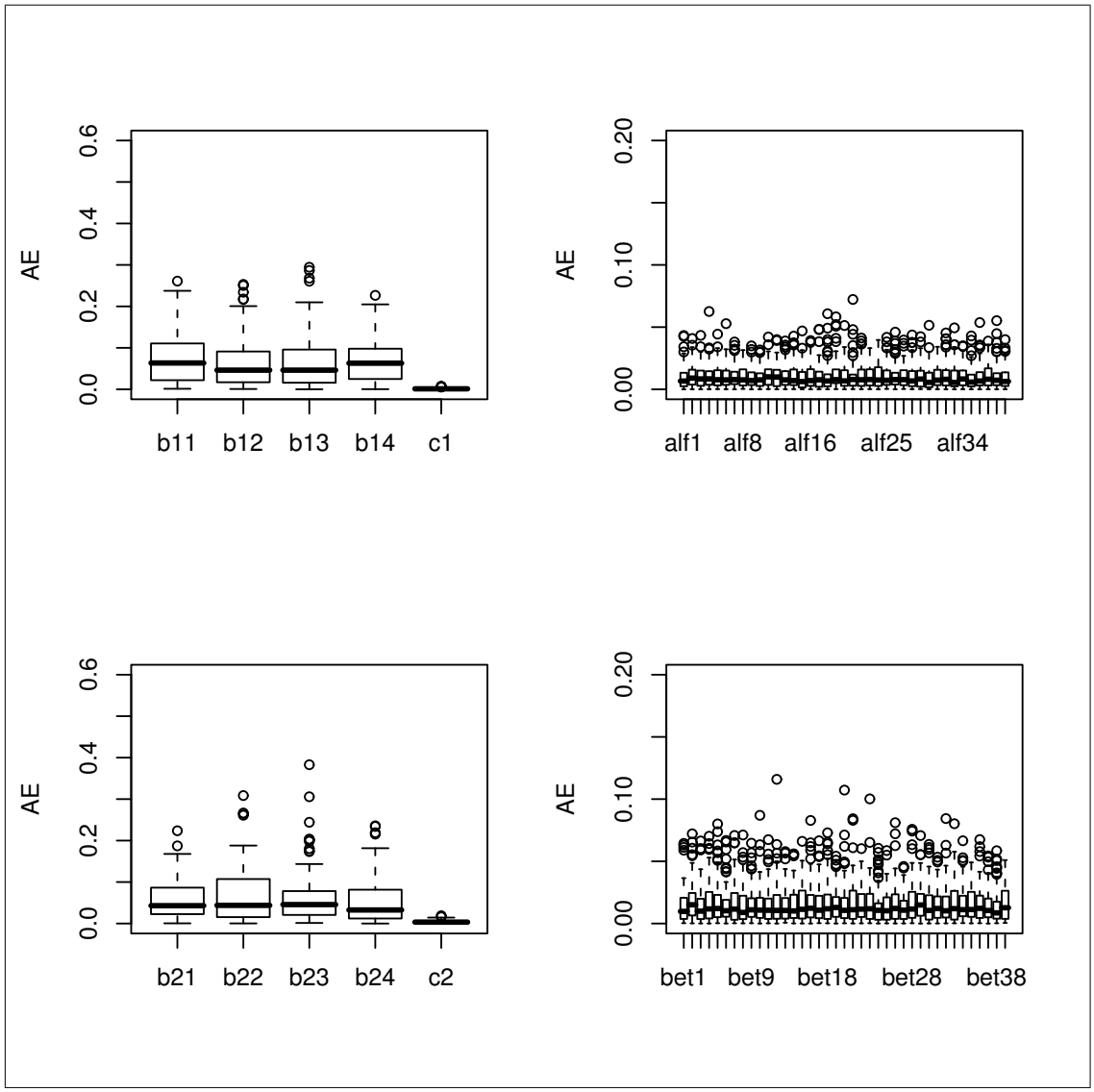

Figure 3.4: Boxplot of the absolute errors (AEs) of the estimates of the parameters in g_1, α, g_2 and β based on 100 simulations: $n = 323$.

be sufficiently large against the number of unknown parameters in such an application.

In this section, we therefore suggest an estimate of the standard deviation of the estimator of each parameter by proposing a bootstrap procedure to obtain a finite-sample based estimation of the standard deviation of the parameter estimate.

We only state the bootstrap procedure for model (3.9), which will be applied in the next chapter. For model (3.9), given the observations $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, we assume that the unknown parameters of the MLE (see Section 3.3) are denoted by

$$\hat{\boldsymbol{\vartheta}} = (\hat{b}_{11}, \hat{b}_{12}, \hat{b}_{13}, \hat{b}_{14}, , \hat{\boldsymbol{\alpha}}^T, \hat{b}_{21}, \hat{b}_{22}, \hat{b}_{23}, \hat{b}_{24}, \hat{\boldsymbol{\beta}}^T, \hat{c}_1, \hat{c}_2)^T.$$

Then, the bootstrap procedure reads as follows:

1) Generate a bootstrap sample of size n:

a) For the i-th observation $\mathbf{X}_i$*, calculate*

$$\begin{aligned}\hat{B}_i = &(\hat{b}_{11}\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i + \hat{b}_{12})I_{\{\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i \le \hat{c}_1\}} + (\hat{b}_{13}\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i + \hat{b}_{14})I_{\{\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i > \hat{c}_1\}} \\ &+ (\hat{b}_{21}\hat{\boldsymbol{\beta}}^T\mathbf{X}_i + \hat{b}_{22})I_{\{\hat{\boldsymbol{\beta}}^T\mathbf{X}_i \le \hat{c}_2\}} + (\hat{b}_{23}\hat{\boldsymbol{\beta}}^T\mathbf{X}_i + \hat{b}_{24})I_{\{\hat{\boldsymbol{\beta}}^T\mathbf{X}_i > \hat{c}_2\}},\end{aligned}$$

and

$$\hat{p}_i = \hat{P}(Y_i = 1|\mathbf{X}_i) = \frac{e^{\hat{B}_i}}{1 + e^{\hat{B}_i}}.$$

b) Generate the i-th bootstrap observation Y_i^* from a binomial distribution $Binorm(1, \hat{p}_i)$.

c) For $i = 1, 2, \cdots, n$ in Steps a) and b), a bootstrap sample of size n, $\{(\mathbf{X}_i, Y_i^*)\}_{i=1}^n$, is generated.

2) Obtain a bootstrap MLE of $\boldsymbol{\vartheta}$ using the bootstrap sample of size n, $\{(\mathbf{X}_i, Y_i^*)\}_{i=1}^n$:

The estimation is calculated by using the method provided in Section 3.3, where we use $\hat{\boldsymbol{\vartheta}}$ as the initial values of the parameters in the maximum likelihood procedure for the bootstrap sample $\{(\mathbf{X}_i, Y_i^*)\}_{i=1}^n$. We denote the unknown parameters of the bootstrap MLE by

$$\hat{\boldsymbol{\vartheta}}^* = (\hat{b}_{11}^*, \hat{b}_{12}^*, \hat{b}_{13}^*, \hat{b}_{14}^*, , \hat{\boldsymbol{\alpha}^*}^T, \hat{b}_{21}^*, \hat{b}_{22}^*, \hat{b}_{23}^*, \hat{b}_{24}^*, \hat{\boldsymbol{\beta}^*}^T, \hat{c}_1^*, \hat{c}_2^*)^T.$$

3) Repeat Steps 1) and 2) B times, where B is called the size of the bootstrap, which is usually quite large as required.

We denote the B bootstrap estimates of $\boldsymbol{\vartheta}$ by

$$\hat{\boldsymbol{\vartheta}}^{*(j)}, \ j = 1, \ 2, \ \cdots, \ B.$$

4) Calculate the bootstrap standard deviations of the MLE $\hat{\boldsymbol{\vartheta}}$:

The standard deviation of the k-th component of $\hat{\boldsymbol{\vartheta}}$, can be calculated

as

$$std(\hat{\vartheta}_k) = \sqrt{\frac{1}{B}\sum_{j=1}^{B}\left(\hat{\vartheta}_k^{*(j)} - \bar{\vartheta}_k^{*}\right)^2},$$

where $\hat{\vartheta}_k^{*(j)}$ *is the* k*-th component of* $\hat{\boldsymbol{\vartheta}}^{*(j)}$ *obtained in Step 3), and* $\bar{\vartheta}_k^{*} = \frac{1}{B}\sum_{j=1}^{B}\hat{\vartheta}_k^{*(j)}$.

The main burden of computation in the above bootstrap procedure lies in Step 2). Here the maximisation of the likelihood for each bootstrap sample by using the downhill simplex method, given at the end of Section 3.3, needs well specified $(D+1)$ initial values of the vector $\boldsymbol{\vartheta}$ (here D stands for the dimension of $\boldsymbol{\vartheta}$), which may require a bit time-consuming experimental tries in general if we have no information on the actual value of the vector $\boldsymbol{\vartheta}$. Luckily, in the bootstrap, a simple way to reduce this computation burden is to fully utilise the estimator $\hat{\boldsymbol{\vartheta}}$ because the bootstrap sample is generated based on this data-based estimator, and therefore we can well specify the $(D+1)$ initial values of the vector $\boldsymbol{\vartheta}$ in Step 2) by adding $(D+1)$ small randomly-generated (vector) values to $\hat{\boldsymbol{\vartheta}}$. In this way, the above bootstrap method works quite well in our experiences. We will use this method to estimate the standard deviation of the estimators of the model parameters in Chapter 4 (section 4.2).

Chapter 4

Analysing schizophrenia data with threshold index logistic regression model

4.1 Introduction

In Chapter 2, we have analysed the schizophrenia SNP datasets by using Wei and Li (2007)'s NPR models and the linear logistic regression models. It is found that the linear logistic regression models perform better than the NPR models. In this chapter, we are applying the model and methodology suggested in Chapter 3 to analysing the schizophrenia data. We will show that our proposed TILoR models will capture the nonlinearity feature of the

genetical complexity of the schizophrenia SNP datasets. Consequently, it will perform much better than the linear logistic regression models in terms of cross-validation prediction.

We will report the analysis for the datasets of the general schizophrenia and the CD subtype schizophrenia, respectively, in Section 4.2. The cross-validation results with 3 randomly selected folds for each dataset are reported in Sections 4.3.

4.2 Estimated models

We will examine the two datasets of schizophrenia SNPs by applying our proposed TILoR models. Note that the PL-TILoR is a special A-TILoR model. So our analysis of the schizophrenia SNP datasets will only be carried out with A-TILoR model.

4.2.1 General schizophrenia SNP dataset with A-TILoR model

We first look at the general schizophrenia SNP dataset, which includes 171 controls and 325 cases of different subtypes (CD subtype, CS subtype, non-

CD non-CS subtype), with A-TILoR model

$$\begin{aligned}
&\log\left\{\frac{P(Y_i=1|\mathbf{X}_i)}{1-P(Y_i=1|\mathbf{X}_i)}\right\}\\
&=(b_{11}\boldsymbol{\alpha}^T\mathbf{X}_i+b_{12})I_{\{\boldsymbol{\alpha}^T\mathbf{X}_i\le c_1\}}+(b_{13}\boldsymbol{\alpha}^T\mathbf{X}_i+b_{14})I_{\{\boldsymbol{\alpha}^T\mathbf{X}_i>c_1\}}\\
&\quad+(b_{21}\boldsymbol{\beta}^T\mathbf{X}_i+b_{22})I_{\{\boldsymbol{\beta}^T\mathbf{X}_i\le c_2\}}+(b_{23}\boldsymbol{\beta}^T\mathbf{X}_i+b_{24})I_{\{\boldsymbol{\beta}^T\mathbf{X}_i>c_2\}}, \qquad (4.1)
\end{aligned}$$

where the first components of α and β are nonnegative for identifiability. In Chapter 2, we applied the OR (odds ratio) principle to choose important SNPs, from which $p = 40$ SNPs are selected at the significance level (i.e, Type I error rate) of 5%:

$rs8074995$, $rs439401$, $rs10774517$, $rs7960673$, $rs6490272$, $rs534455$, $rs486706$, $rs694060$, $rs12128305$, $rs11207007$, $rs6687842$, $rs10047071$, $rs17424216$, $rs2991515$, $rs11581152$, $rs852787$, $rs9432024$, $rs11122357$, $rs877984$, $rs1400316$, $rs1399622$, $rs17507049$, $rs7121214$, $rs7928038$, $rs10501563$, $rs1940078$, $rs1943699$, $rs6592211$, $rs17203281$, $rs1615640$, $rs11220082$, $rs931671$, $rs17281921$, $rs1978198$, $rs2711881$, $rs2528865$, $rs10248053$, $rs2283029$, $rs1454626$, $rs1022307$.

So we use these SNPs as our regressors, denoted as:

$$\mathbf{X}_i=(X_{i,1},X_{i,2},\cdots,X_{i,40}).$$

The estimated coefficients in model (4.1) and their standard deviations (s.d.) calculated by bootstrap method (bootstrap sample size=100) are reported in table 4.1, table 4.2, table 4.3 and table 4.4.

Table 4.1: Estimated coefficients b1, b2 and their standard deviations calculated by bootstrap method in A-TILoR model for the WAFSS schizophrenia data set

b1=(b11,b12,b13,b14)	0.0274	0.4358	-2.7377	1.3744
s.d. (bootstrap)	0.0139	0.0873	0.0561	0.0547
b2=(b21,b22,b23,b24)	0.0260	-0.0748	2.4239	0.4685
s.d. (bootstrap)	0.0281	0.0875	0.0629	0.0553

From the results mentioned above for the fitted model, we can draw some interesting conclusions as follows:

- 1) In genetic analysis, it is a common sense to note that the individual SNPs make contributions through interactions. Our indices in the TILoR model confirm that the individual SNPs' contributions are made through such regime indices α and β. Let's look at the components of the index vectors α and β and their bootstrap standard deviations(table 4.2, table 4.3). Quite clearly, all the components of the index vectors α and β, except the coefficients of $X_{i,1}$ (SNP rs8074995) in α and that of $X_{i,27}$ (SNP rs1943699) in β, are significantly different from zero at the significance level (that is, the allowed Type I testing error rate) of both 5% and even 1%, or equivalently at the confidence level of both

95% and 99%, respectively. Hence all the 40 SNPs selected by the odds ratio test in Chapter 2 play some part in forming the regime indices α and β (although most of them may only make relatively slight contributions). As mentioned in Chapter 1, schizophrenia is a complex disorder. There are multiple susceptibility genes, each with small to modest effects that interact with each other and environmental factors to influence susceptibility for this disease. It is accepted that for each gene, more than one SNP shows association with schizophrenia, but rarely are data from individual SNPs highly significant(Harrison and Owen,2003). The combination of table 4.2 and table 4.3provides an explicit quantitative proof to this biological understanding of schizophrenia using proposed threshold index logistic regression model. We further notice that there are 8 components of α whose absolute values are greater than 0.2, while for β, there are 7 components of β whose absolute values are greater than 0.2. For the components of α and β whose absolute values are greater than 0.2, those components selected are listed in table 4.5. Obviously, it is clear from table 4.5 that the SNPs selected in these two indices are quite different, implying different sub-regimes may underlie the interaction mechanisms. These findings are of potential important, because the SNPs selected in table 4.5 could be more influential diagnostic indicators of schizophrenia.

- 2) On the thresholds(see table 4.4), the values $c_1 = -0.09512598$ and

$c_2 = 0.09161813$ appear near 0, but they are still very significant, as the confidence intervals c_1 and c_2 plus their three times standard deviations calculated by bootstrap method in table 4.4 still do not include 0.

- 3) We plotted the kernel density of the indices of $\alpha^T \mathbf{X}_i$'s and $\beta^T \mathbf{X}_i$'s, respectively, in Figure 4.1, with dashed lines for the thresholds c_1 and c_2. It follows that under the α-regime, there is a high empirical probability (90.32258%) that the values of the index variables are less than the threshold c_1, while under the β-regime, the empirical probability of the index variable less than the threshold c_2 is 66.33065%.

  By looking at the functions g_1 and g_2(Chapter 3, equation(3.8)), plotted in Figure 4.2, it is apparent that when the regime indices are lower than the corresponding thresholds, the impacts of the regimes are stable, but when indices are greater than the thresholds, the impacts are viably more significant. If combine with the fact the majority of the index variables are less than the two thresholds (with reference to Figure 4.1), it follows that the impacts of most of the index variables are small, only if the regime indices are greater than the corresponding thresholds then they will have significant impact, but that probability is relatively lower. Figure 4.2 also provides a visual exhibition of the nonlinear feature of schizophrenia SNP data sets.

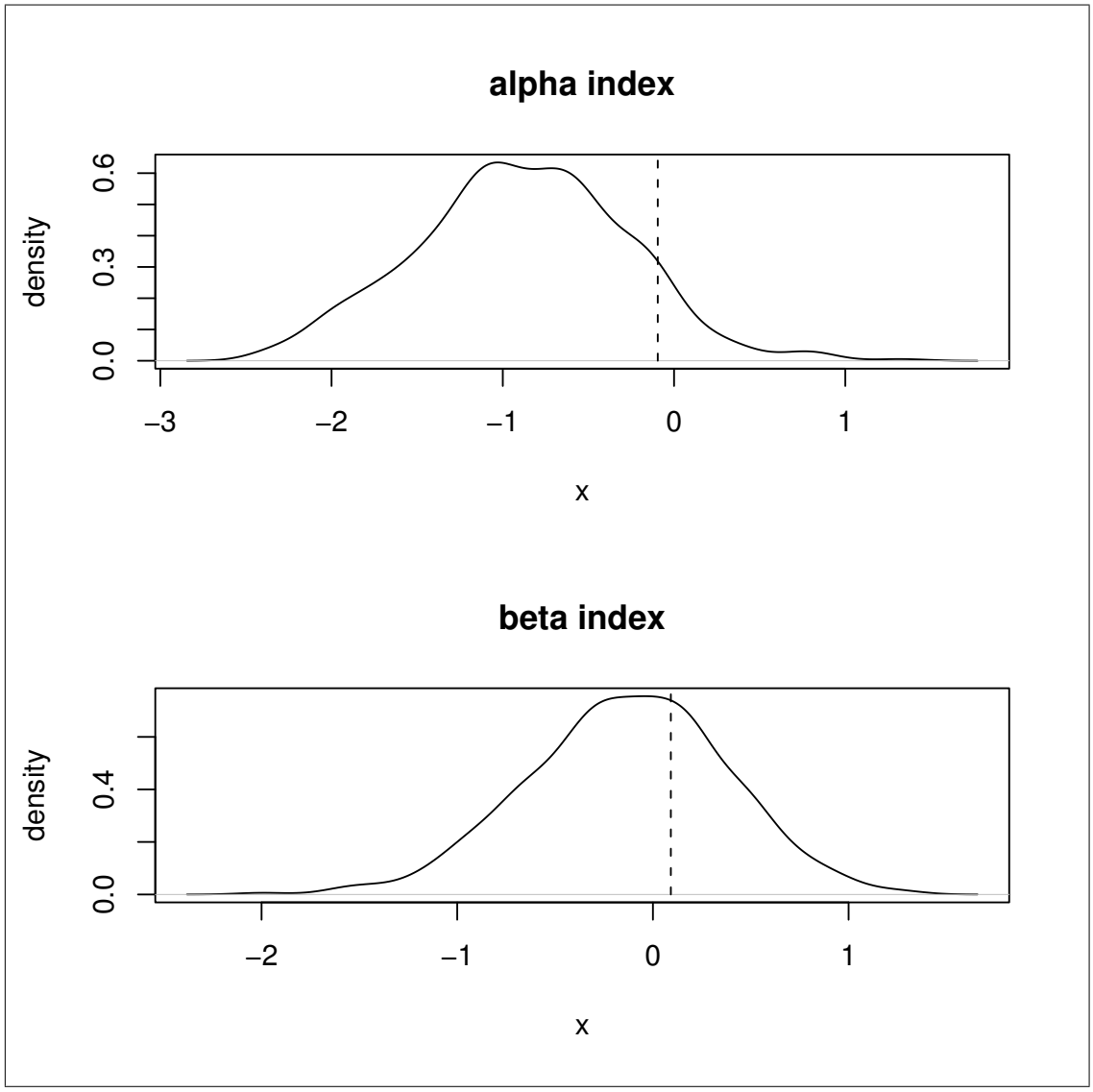

Figure 4.1: A-TILoR model for general schizophrenia: The kernel density for the indices of $\alpha^T \mathbf{X}_i$'s and $\beta^T \mathbf{X}_i$'s with dashed lines for the thresholds c_1 and c_2, respectively.

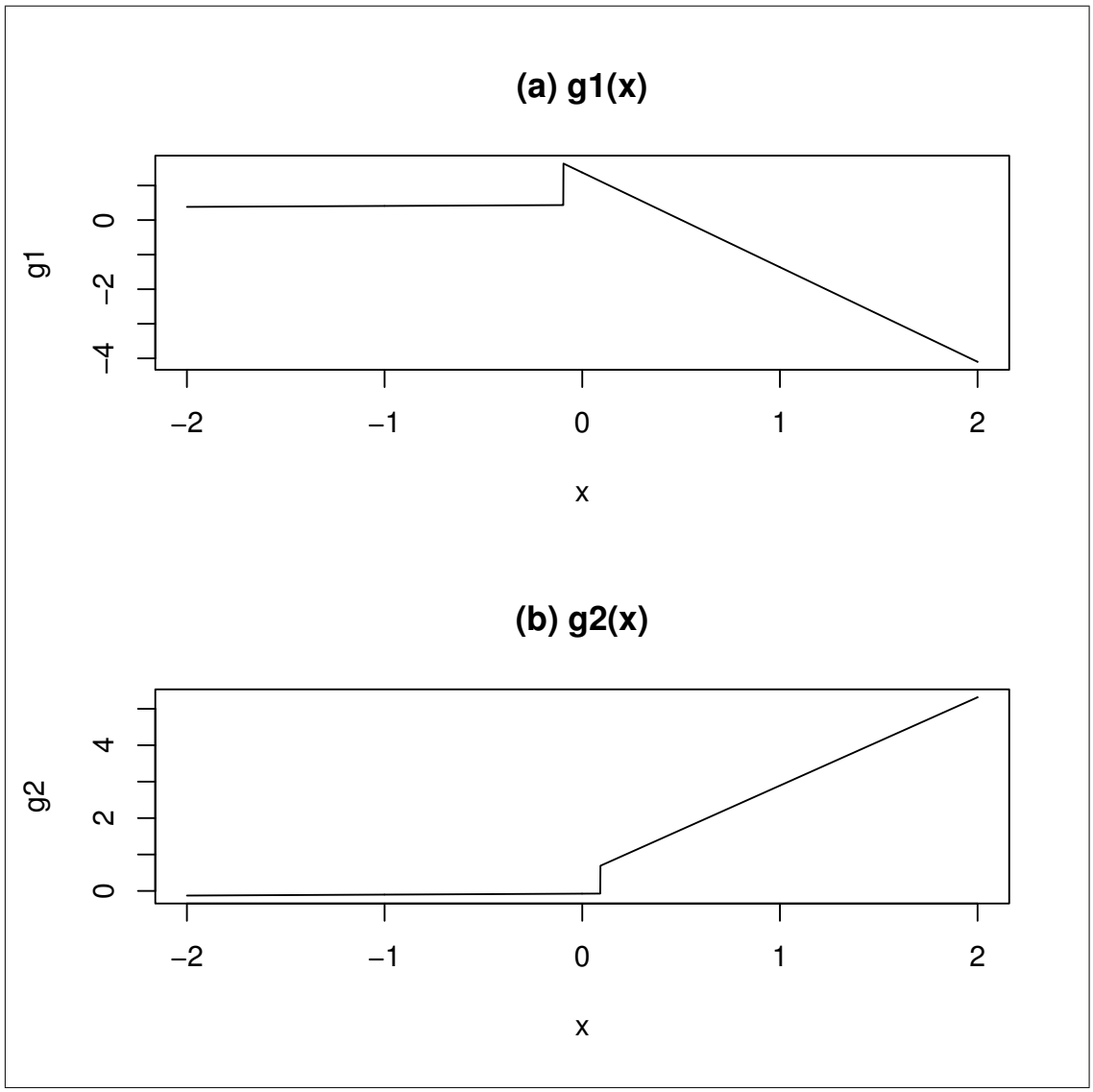

Figure 4.2: A-TILoR model for general schizophrenia: The plot of the functions g_1 and g_2, respectively.

4.2.2 CD subtype schizophrenia SNP dataset with A-TILoR model

We now look at the CD subtype schizophrenia SNP dataset, which includes 171 controls and 152 cases of CD subtype, with A-TILoR model (4.1). In Chapter 2, we apply the OR (odds ratio) principle to choose important SNPs, by which $p = 39$ SNPs are selected, for this dataset, at the significance level of 5%:

$rs220599$, $rs439401$, $rs738288$, $rs10889023$, $rs534455$, $rs486706$, $rs10789045$, $rs7548633$, $rs694060$, $rs595513$, $rs267638$, $rs11207007$, $rs6687842$, $rs1778032$, $rs2991515$, $rs17115797$, $rs3131726$, $rs7534106$, $rs2224823$, $rs11122324$, $rs7534681$, $rs9432024$, $rs11122357$, $rs967433$, $rs2772122$, $rs7533169$, $rs790369$, $rs2054023$, $rs4944481$, $rs17734854$, $rs1615640$, $rs11220082$, $rs931671$, $rs4542192$, $rs7412571$, $rs6664618$, $rs1572680$, $rs2283029$, $rs1454626$.

So here we use these SNPs as our regressors, denoted as

$$\mathbf{X}_i = (X_{i,1}, X_{i,2}, \cdots, X_{i,39}).$$

The estimated coefficients in A-TILoR model(4.1) are listed in table 4.6, table 4.7, table 4.8 and table 4.9.

From the above results for the fitted model, we can draw some interesting observations quite similar to general schizophrenia as follows:

- 1) For CD subtype, quite clearly, the coefficients of the indices α and β are nearly all significant at the significance level of either 5% or even 1%, except the coefficients of $X_{i,12}$ (SNP rs11207007) and $X_{i,15}$ (SNP rs2991515) in α and those of $X_{i,4}$ (SNP rs10889023) and $X_{i,10}$ (SNP rs595513) in β. Hence all the 39 SNPs selected by the odds ratio test in Chapter 2 play some part in forming the regime indices α and β (although most of them may only make relatively slight contributions). As mentioned in last section, schizophrenia is a complex disorder, and rarely are data from individual SNPs highly significant. If we look at the components of α and β whose absolute values are greater than 0.2, then the components selected are in table 4.10. Obviously, the main SNPs selected in these two indices for the CD subtype are quite different, implying different sub-regimes may underlie the interaction mechanisms.

- 2) On the thresholds for the CD subtype, in table 4.9, the values $c_1 = 0.1817811$ and $c_2 = 0.2594507$ are away from 0 and very significant. We plotted the kernel density of the indices of $\alpha^T \mathbf{X}_i$'s and

$\beta^T\mathbf{X}_i$'s, respectively, in Figure 4.3, with dashed lines for the thresholds c_1 and c_2. It follows that under the α-regime, the empirical probability of the values of the index variable less than the threshold c_1 is 65.94427%, while under the β-regime, there is a high empirical probability of 80.18576% that the values of the index variable is less than the threshold c_2.

- 3) By looking at the nonlinear functions g_1 and g_2 plotted in Figure 4.4, for the CD subtype, with reference to Figure 4.3, it is apparent that only when the regime indices are greater than the thresholds, the impacts of the regimes are more significant, but that chance seems relatively lower.

4.3 Cross-validation performance

In this section, we are going to examine the performance of our proposed model of threshold index logistic regression (denoted by A-TILoR, or for simplicity by TLR below) compared with the linear logistic regression model estimated through generalised linear model in R (GLM is referred to the linear logistic regression below). We will demonstrate that our proposed TLR method performs viably better than the GLM method with the analysis of the schizophrenia datasets. Note that we have shown in Chapter 2 that the GLM method performs better than Wei and Li's (2007) NPR method.

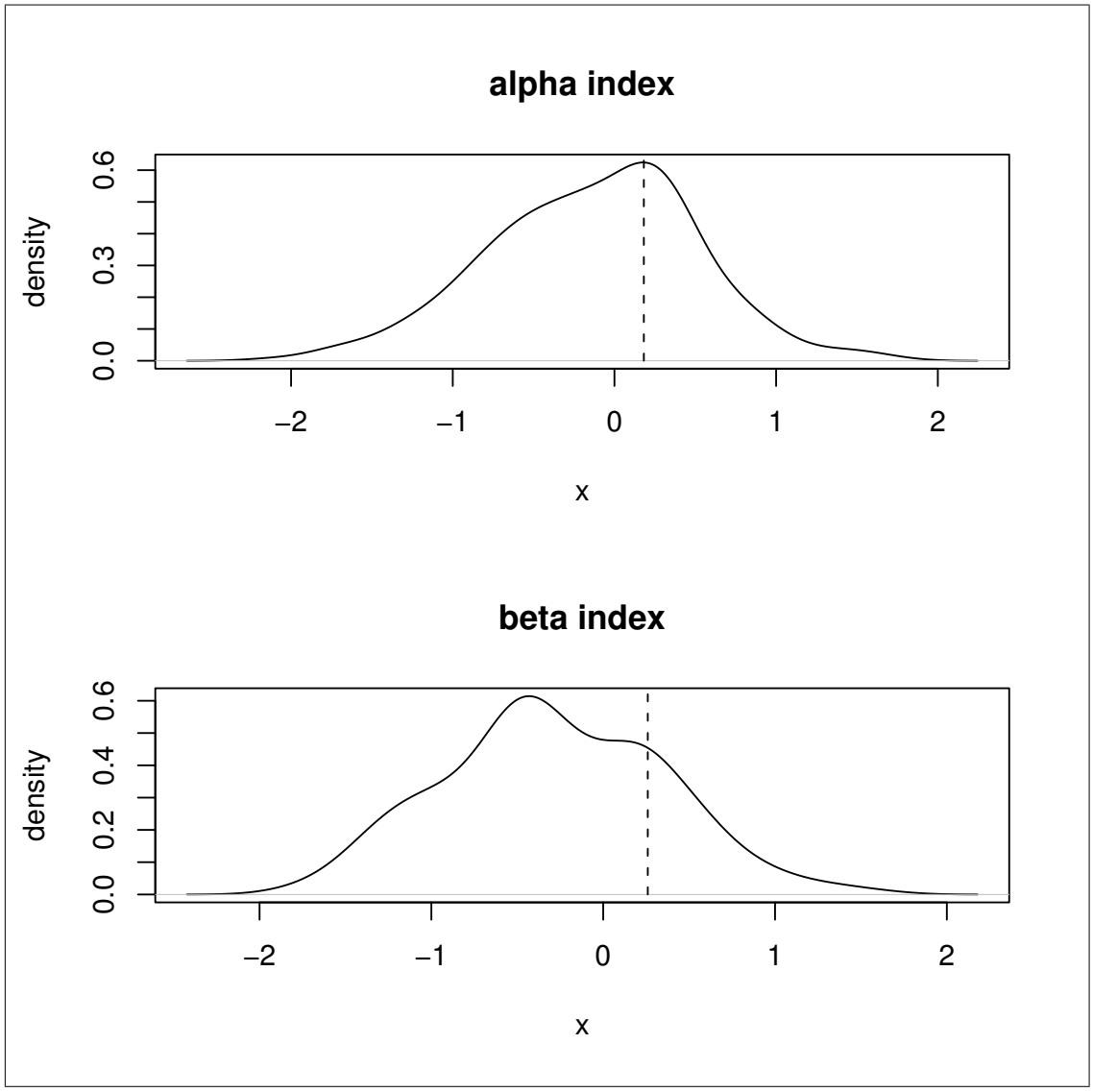

Figure 4.3: A-TILoR model for CD subtype schizophrenia: The kernel density for the indices of $\alpha^T \mathbf{X}_i$'s and $\beta^T \mathbf{X}_i$'s with dashed lines for the thresholds c_1 and c_2, respectively.

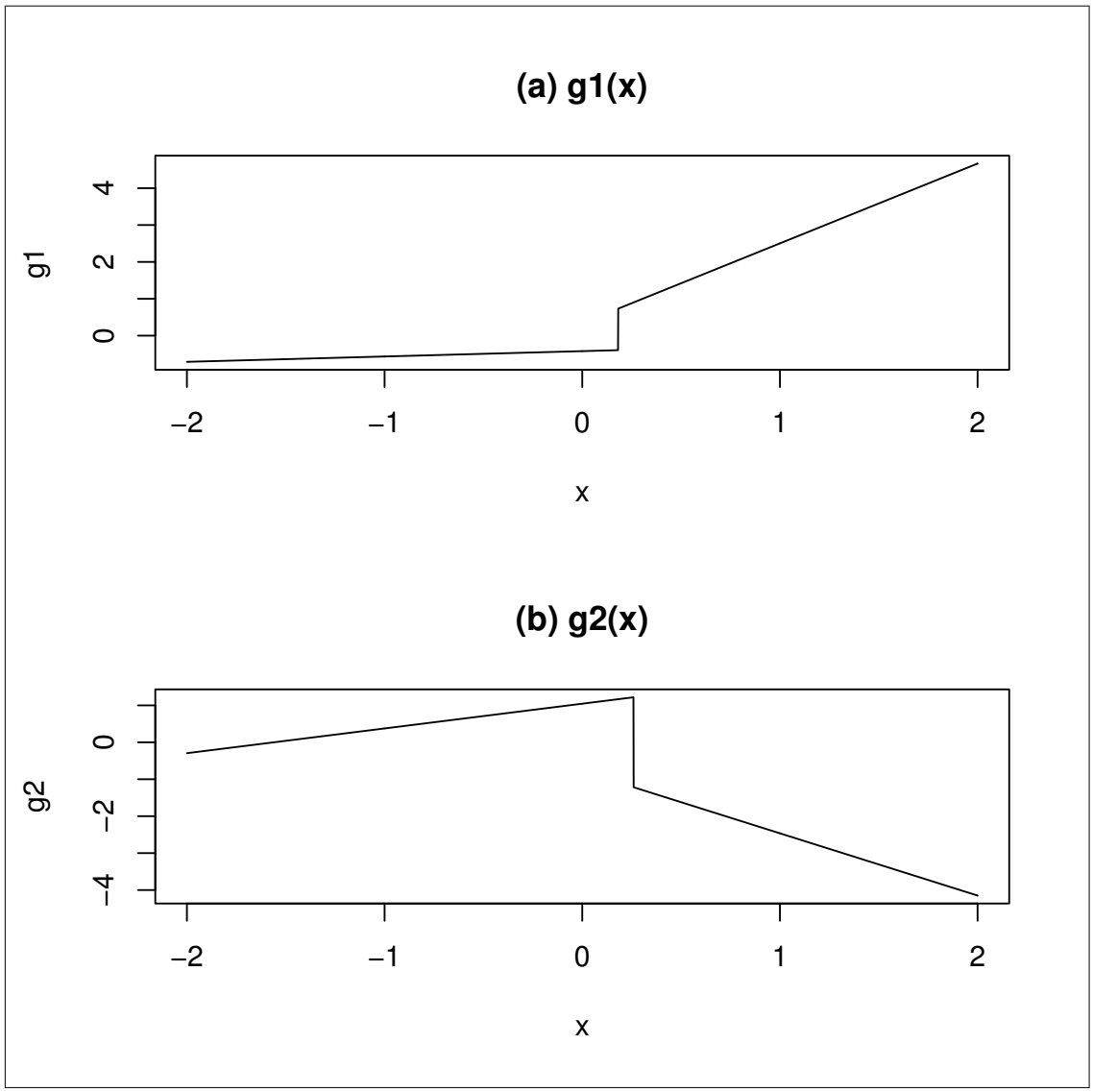

Figure 4.4: A-TILoR model for CD subtype schizophrenia: The plot of the functions g_1 and g_2, respectively.

We have carried out the comparison through cross-validation testing by partition the whole datasets of the general schizophrenia and the CD subtype, respectively, into some sub-folds.

It is known that the resubstitution estimate of predictive accuracy, derived by direct application of model predictions to the data from which the regression relationship is derived, gives, in general, an optimistic assessment. Because there is a mutual dependence between the model prediction and the data used to derive that prediction, an ideal is to assess the performance of the model on a new data set. The data that are used to develop the model from the training set, while the data on which predictions are tested form the test set. Cross-validation extends the training/test set approach. The data are divided into k sets (or folds), where k is typically in the range of 3 to 10. Each of the k sets becomes in turn the test set, with the remaining data forming the training set. The predictive accuracy assessments from the k folds are combined to give a measure of the predictive performance of the model. This may be done for several different measures of predictive performance. Here we use a 3-fold validation with special considerations based on the case-control character.

For the general schizophrenia data set (325 cases and 171 controls), we use a random number sampling system to divide the case data into three equal groups, and control data into three equal groups. Then we combine the case groups and the control group to form three folds. For each of the

three folds, it is set aside as the test data, with the remaining data making up the training data. In each time, there are 108 cases and 57 controls in the test set, and 217 cases and 114 controls in the training set. We do the same for the CD subtype data set (152 cases and 171 controls), where in each time, there are 50 cases and 57 controls in the test set, and 102 cases and 114 controls in the training set.

Schizophrenia's broad heritability is about 80%. Therefore, 80% is naturally the approximate upper limit of accuracy of models using genotypes only. In other words, without using other information such as phenotypes, whatever modelling technique applies, the accuracy rate is not supposed to be higher than 80%. If we consider 50% as a model-worthy lower limit accuracy, the interval (50%-80%) gives an idea what the accuracy rate will be in. Corresponding error rate will be between 20% to 50%. That gives us an idea about what to expect.

4.3.1 Comparison between GLM and TLR based on resubstitution estimates

Before we look at the performance through the cross-validation out-of-sample prediction, we first take a simple checking of the performance of GLM and TLR methods based on resubstitution estimates (i.e., in-sample prediction) for each fold.

For a given fold left for testing set, we also use the other two folds as

training set to estimate the models. We list the resubstitution estimates for training set in Tables 4.11 and 4.12 for the general schizophrenia and the CD subtype, respectively. Here note that the resubstitution for Fold 1 means the in-sample prediction of the other two folds (i.e., Folds 2 and 3) that are used as the training set in estimating the model. Also, note that in these tables and below, Type I error rate is the percentage of people wrongly classified as with schizophrenia among people without schizophrenia, while Type II error rate is the percentage of people wrongly classified as without schizophrenia among people with schizophrenia.

From tables 4.11 and 4.12, we clearly see that our TLR method performs quite stable for both the general schizophrenia and the CD subtype schizophrenia in the in-sample prediction. For the Type II error rate, GLM appears better than the TLR. However notice that the average Type I error rate of the GLM for the general schizophrenia is as high as 50.58%, which is obviously unacceptable from the model estimation perspective. For the CD subtype, the Type II error rates for these two methods are quite close, while our TLR method has much lower Type I error rate than the GLM. Overall, our TLR method is acceptable for both the general schizophrenia and the CD subtype schizophrenia in the in-sample prediction.

4.3.2 Comparison between TLR and GLM based on cross-validation prediction

Now we turn to the comparison between the TLR and the GLM based on the cross-validation out-of-sample prediction, which is more important in applications.

In tables 4.13 and 4.14, we report the comparison between the GLM and the TLR from the predictive accuracy and the Type I and Type II error rates for the general schizophrenia. In tables 4.15 and 4.16, these comparisons are made for the CD subtype schizophrenia.

Comparing the above tables, we may summarise:

- 1. From the predictive accuracy perspective, the TLR obviously performs better than the GLM for general schizophrenia and CD subtype in both Tables 4.13 and 4.15. For the general schizophrenia, the cross validation predictive accuracy for the TLR and the GLM is 70.10% and 66.26%, respectively. Schizophrenia's broad heritability is about 80%. In our case, whatever modelling technique applies, the accuracy rate for general schizophrenia is not supposed to be higher than 80%. If we consider 50% as a model-worthy lower limit accuracy, it seems that both the GLM and the TLR are acceptable in tables 4.13 and TLM performs better.

  Schizophrenia CD subtype is proposed by WAFSS in 2005. WAFSS's

research has found that Schizophrenia can be divided into three subgroups based on seven phenotype trait tests: general cognitive ability, sustained attention, executive function, verbal memory, speed of information processing, neurobehavioral features and personality factors. These seven tests represent seven domains of the neurocognitive function. The three schizophrenia subgroups are: CD (cognitive deficit), CS (cognitive spared), and non-CD/CS. The group CD subtype accounts for up to 45% of the schizophrenia population and has the most severe symptom of mental disorders among all schizophrenia groups. The genetic basis of CD subtype is still under investigation. WAFSS group suggested that this subtype has a distinct genetic basis(Hallmayer, J.F. etc. 2005). For the CD subtype, the TLR has the predictive accuracy of 81.93% vs. that of 61.68% for the GLM in table 4.15. Because the TLR cross validation predictive accuracy is over the 80% upper limit of general schizophrenia predictive accuracy, this does suggest CD subtype is genetically more significant.

- 2. For the prediction of general schizophrenia, from the perspective of the Type I and Type II error rates, the problem with the GLM is that it has a too ideal type II error (12.44% for resubstitution error and 19.75% cross-validation error) but far too worse type I error (50.58% for resubstitution error and 60.23% cross-validation error) in table 4.11 and table 4.14. The bad performance on type II error has made GLM itself

unsuitable to be used as a practical model for general schizophrenia. In contrast, in same tables, using TLR, both the type I error (33.33% for resubstitution and 32.16% for cross-validation error) and type II error (27.8% for resubstitution and 28.70% for cross-validation error) are stable and close to the 20% ideal lower error rate limit, which makes it an eligible and nice predictor for schizophrenia classification.

- 3.For the prediction of CD subtype, from the prespective of the type I and type error, further than the TLR is a bit better than the GLM in the performance of resubstitution in table 4.12, the TLR has much better Type I and Type II error rates than the GLM in the performance of cross validation in table 4.16. The TLR has the cross validation type I error of 12.86% vs that of 32.75% for the GLM. For the type II error, the cross validation error rate for the TLR and the GLM is 24% and 44.67%, respectively.

- 4. The threshold index nonlinear logistic regression model has much superior prediction ability to the logistic model regarding specificity and sensitivity for both general schizophrenia and CD subtype. Specificity and sensitivity are two indexes that biologists are particularly interested in. Specificity measures the percentage of healthy people who are correctly identified as not having the condition while sensitivity measures the percentage of sick people who are correctly identified

as having the condition. In table 4.17, for general schizophrenia, the TLR has a balanced cross validation results of sensitivity 71.3% and specificity 67.84%, vs. that of 80.25% and 39.77% for the GLM. The GLM specificity result of 39.77% makes itself unacceptable as a modelling technique for general schizophrenia. In chapter two, the poor performance of GLM on specificity prediction for general schizophrenia is the motivation drives us to propose a new model. Now the proposed TLR model not only has successfully solved the specificity problem for general schizophrenia, but also has much superior results for CD subtype. TLR is about 20% better in both sensitivity and specificity results compare to GLM for CD subtype.

In table 2.5, the linear logistic model regressors is selected from regressors in table 4.17 until the regressor number reaches the same number as in NPR model using step by step method. Also, the prediction of sensitivity and specificity are calculated based on resubstitution method. Therefore, it is not surprising that the GLM results in table 2.5 are slightly better than that in table 4.17. But even so, they are still much worse than the TLR results.

Table 4.2: Estimated coefficients α , β and their standard deviations calculated by bootstrap method in A-TILoR model for the WAFSS schizophrenia data set (1)

SNP	α	α s.d(bootstrap)	β	β s.d(bootstrap)
$X_{i,1}$(rs8074995)	0.0058	0.0042	0.1393	0.0050
$X_{i,2}$(rs439401)	0.3166	0.0052	0.1727	0.0051
$X_{i,3}$(rs10774517)	-0.0797	0.0041	-0.1082	0.0044
$X_{i,4}$(rs7960673)	-0.0161	0.0043	-0.0541	0.0044
$X_{i,5}$(rs6490272)	0.0004	0.0048	0.1058	0.0042
$X_{i,6}$(rs534455)	0.1194	0.0042	0.1804	0.0047
$X_{i,7}$(rs486706)	-0.0343	0.0055	0.0503	0.0047
$X_{i,8}$(rs694060)	-0.0905	0.0047	0.0630	0.0042
$X_{i,9}$(rs12128305)	-0.1112	0.0042	0.0810	0.0048
$X_{i,10}$(rs11207007)	0.1359	0.0036	-0.0288	0.0054
$X_{i,11}$(rs6687842)	-0.0203	0.0040	-0.0993	0.0050
$X_{i,12}$(rs10047071)	-0.0531	0.0051	-0.2190	0.0054
$X_{i,13}$(rs17424216)	-0.2258	0.0059	0.0227	0.0040
$X_{i,14}$(rs2991515)	-0.0350	0.0047	0.0800	0.0048
$X_{i,15}$(rs11581152)	0.1220	0.0051	0.0241	0.0042
$X_{i,16}$(rs852787)	-0.1378	0.0060	0.0976	0.0039
$X_{i,17}$(rs9432024)	-0.2081	0.0056	0.1916	0.0043
$X_{i,18}$(rs11122357)	0.0368	0.0056	-0.3270	0.0042
$X_{i,19}$(rs877984)	0.1109	0.0046	0.0651	0.0046
$X_{i,20}$(rs1400316)	-0.0826	0.0050	-0.2375	0.0049

Table 4.3: Estimated coefficients α , β and their standard deviations calculated by bootstrap method in A-TILoR model for the WAFSS schizophrenia data set (2)

SNP	α	α s.d(bootstrap)	β	β s.d(bootstrap)
$X_{i,21}$(rs1399622)	-0.0473	0.0036	-0.0544	0.0052
$X_{i,22}$(rs17507049)	-0.2784	0.0050	0.1464	0.0046
$X_{i,23}$(rs7121214)	0.1016	0.0052	-0.0622	0.0049
$X_{i,24}$(rs7928038)	0.1064	0.0045	-0.3077	0.0042
$X_{i,25}$(rs10501563)	-0.1824	0.0050	-0.1235	0.0048
$X_{i,26}$(rs1940078)	-0.0405	0.0060	-0.4768	0.0041
$X_{i,27}$(rs1943699)	0.2444	0.0049	-0.0024	0.0051
$X_{i,28}$(rs6592211)	-0.1094	0.0048	-0.1192	0.0035
$X_{i,29}$(rs17203281)	-0.5100	0.0053	-0.0415	0.0050
$X_{i,30}$(rs1615640)	-0.1139	0.0047	0.0162	0.0047
$X_{i,31}$(rs11220082)	-0.0795	0.0050	-0.1194	0.0047
$X_{i,32}$(rs931671)	-0.2502	0.0048	0.1427	0.0048
$X_{i,33}$(rs17281921)	0.0332	0.0043	-0.0568	0.0039
$X_{i,34}$(rs1978198)	0.0342	0.0055	0.2519	0.0048
$X_{i,35}$(rs2711881)	-0.0555	0.0048	-0.1884	0.0045
$X_{i,36}$(rs2528865)	0.0770	0.0051	-0.0204	0.0051
$X_{i,37}$(rs10248053)	-0.1033	0.0056	0.1180	0.0058
$X_{i,38}$(rs2283029)	0.0845	0.0049	0.2366	0.0050
$X_{i,39}$(rs1454626)	-0.3238	0.0045	0.0979	0.0043
$X_{i,40}$(rs1022307)	0.0410	0.0043	-0.0302	0.0047

Table 4.4: Estimated coefficients c1, c2 and their standard deviations calculated by bootstrap method in A-TILoR model for the WAFSS schizophrenia data set

c1	c1 s.d(bootstrap)	c2	c2 s.d(bootstrap)
-0.0951	0.0005	0.09162	0.0004

Table 4.5: A-TILoR model for general schizophrenia: The components of α and β whose absolute values are greater than 0.2.

Component of $\mathbf{X}_i$	(Gene:SNP)	Component of α
$X_{i,2}$	(APOE:rs439401)	0.3166165734
$X_{i,13}$	(DAB:rs17424216)	-0.2257977475
$X_{i,17}$	(DISC1:rs9432024)	-0.2080899519
$X_{i,22}$	(DLG2:rs17507049)	-0.2784461531
$X_{i,27}$	(DLG2:rs1943699)	0.2443957136
$X_{i,29}$	(DLG4:rs17203281)	-0.5098565580
$X_{i,32}$	(NUDEL:rs931671)	-0.2502394725
$X_{i,39}$	(VLDLR:rs1454626)	-0.3237843665
Component of $\mathbf{X}_i$	(Gene:SNP)	Component of β
$X_{i,12}$	(DAB:rs10047071)	-0.219001605
$X_{i,18}$	(DISC1:rs11122357)	-0.326981794
$X_{i,20}$	(DLG2:rs1400316)	-0.237457527
$X_{i,24}$	(DLG2:rs7928038)	-0.307667557
$X_{i,26}$	(DLG2:rs1940078)	-0.476838576
$X_{i,34}$	(RELN:rs1978198)	0.251949608
$X_{i,38}$	(RELN:rs2283029)	0.236604681

Table 4.6: Estimated coefficients b1, b2 and their standard deviations calculated by bootstrap method in A-TILoR model for the WAFSS schizophrenia CD subtype data set

b1=(b11,b12,b13,b14)	0.1451	-0.4199	2.1646	0.3385
s.d. (bootstrap)	0.0485	0.0403	0.0470	0.0395
b2=(b21,b22,b23,b24)	0.6701	1.0459	-1.6853	-0.7785
s.d. (bootstrap)	0.0555	0.0445	0.0435	0.0329

Table 4.7: Estimated coefficients α , β and their standard deviations calculated by bootstrap method in A-TILoR model for the WAFSS schizophrenia CD subtype data set (1)

SNP	α	α s.d(bootstrap)	β	β s.d(bootstrap)
$X_{i,1}$(rs220599)	0.1605	0.0067	0.1502	0.0103
$X_{i,2}$(rs439401)	0.1453	0.0080	-0.1316	0.0097
$X_{i,3}$(rs738288)	-0.1053	0.0064	-0.4679	0.0089
$X_{i,4}$(rs10889023)	0.0808	0.0061	0.0163	0.0107
$X_{i,5}$(rs534455)	0.2139	0.0067	0.1227	0.0096
$X_{i,6}$(rs486706)	-0.0508	0.0089	0.1589	0.0123
$X_{i,7}$(rs10789045)	0.3304	0.0068	0.1351	0.0111
$X_{i,8}$(rs7548633)	-0.1281	0.0073	0.1482	0.0100
$X_{i,9}$(rs694060)	0.1296	0.0073	0.2405	0.0115
$X_{i,10}$(rs595513)	-0.1926	0.0067	0.0111	0.0117
$X_{i,11}$(rs267638)	-0.0265	0.0072	-0.2999	0.0104
$X_{i,12}$(rs11207007)	-0.0061	0.0071	-0.1589	0.0102
$X_{i,13}$(rs6687842)	0.0681	0.0078	0.0249	0.0093
$X_{i,14}$(rs1778032)	-0.2884	0.0070	-0.0215	0.0109
$X_{i,15}$(rs2991515)	0.0097	0.0061	0.2828	0.0111
$X_{i,16}$(rs17115797)	0.1232	0.0065	-0.1246	0.0095
$X_{i,17}$(rs3131726)	-0.0152	0.0076	0.1335	0.0103
$X_{i,18}$(rs7534106)	0.1701	0.0069	-0.1101	0.0091
$X_{i,19}$(rs2224823)	-0.0874	0.0071	-0.1572	0.0109
$X_{i,20}$(rs11122324)	-0.1436	0.0069	-0.0647	0.0117

Table 4.8: Estimated coefficients α , β and their standard deviations calculated by bootstrap method in A-TILoR model for the WAFSS schizophrenia CD subtype data set (2)

SNP	α	α s.d(bootstrap)	β	β s.d(bootstrap)
$X_{i,21}$(rs7534681)	0.0468	0.0070	-0.0409	0.0099
$X_{i,22}$(rs9432024)	-0.0239	0.0067	-0.0937	0.0099
$X_{i,23}$(rs11122357)	-0.2159	0.0075	-0.0814	0.0095
$X_{i,24}$(rs967433)	-0.0638	0.0071	0.0760	0.0086
$X_{i,25}$(rs2772122)	-0.2349	0.0074	0.1799	0.0099
$X_{i,26}$(rs7533169)	0.1731	0.0073	0.0360	0.0104
$X_{i,27}$(rs790369)	0.2592	0.0074	0.1125	0.0106
$X_{i,28}$(rs2054023)	0.07219	0.0067	-0.0982	0.0102
$X_{i,29}$(rs4944481)	0.0734	0.0068	0.1592	0.0106
$X_{i,30}$(rs17734854)	-0.4443	0.0056	0.1755	0.0100
$X_{i,31}$(rs1615640)	-0.0210	0.0076	-0.1665	0.0095
$X_{i,32}$(rs11220082)	-0.1495	0.0067	0.0273	0.0102
$X_{i,33}$(rs931671)	0.1264	0.0072	-0.2049	0.0099
$X_{i,34}$(rs4542192)	0.1084	0.0057	0.0514	0.0091
$X_{i,35}$(rs7412571)	-0.0571	0.0075	-0.0590	0.0101
$X_{i,36}$(rs6664618)	0.0610	0.0074	-0.0213	0.0089
$X_{i,37}$(rs1572680)	-0.1233	0.0068	-0.1394	0.0096
$X_{i,38}$(rs2283029)	0.0157	0.0074	-0.2862	0.0095
$X_{i,39}$(rs1454626)	0.2443	0.0073	0.1358	0.0105

Table 4.9: Estimated coefficients c1, c2 and their standard deviations calculated by bootstrap method in A-TILoR model for the WAFSS schizophrenia CD subtype data set

c1	c1 s.d(bootstrap)	c2	c2 s.d(bootstrap)
0.1818	0.0014	0.2595	0.0028

Table 4.10: A-TILoR model for CD subtype schizophrenia: The components of α and β whose absolute values are greater than 0.2.

Component of $\mathbf{X}_i$	(Gene:SNP)	Component of α
$X_{i,5}$	(DAB:rs534455)	0.213884518
$X_{i,7}$	(DAB:rs10789045)	0.330436703
$X_{i,14}$	(DAB:rs1778032)	-0.288425609
$X_{i,23}$	(DISC1:rs11122357)	-0.215940820
$X_{i,25}$	(DISC1:rs2772122)	-0.234875792
$X_{i,27}$	(DLG2:rs790369)	0.259160282
$X_{i,30}$	(DLG2:rs17734854)	-0.444332340
$X_{i,39}$	(VLDLR:rs1454626)	0.244328756
Component of $\mathbf{X}_i$	(Gene:SNP)	Component of β
$X_{i,3}$	(ATF4:rs738288)	-0.46794802
$X_{i,9}$	(DAB:rs694060)	0.24054483
$X_{i,11}$	(DAB:rs267638)	-0.29986520
$X_{i,15}$	(DAB:rs2991515)	0.28276706
$X_{i,33}$	(NUDEL:rs931671)	-0.20494686
$X_{i,38}$	(RELN:rs2283029)	-0.28623061

Table 4.11: Comparison between GLM and TLR for the general schizophrenia data set: Resubstitution Type I and Type II error rates.

	Type	Fold1	Fold2	Fold3	Average
GLM	I	52.63%	45.61%	53.50%	50.58%
	II	11.05%	13.36%	12.90%	12.44%
TLR	I	34.21%	24.56%	41.23 %	33.33%
	II	24.88%	28.11%	30.41 %	27.8%

Table 4.12: Comparison between GLM and TLR for the CD subtype schizophrenia: Resubstitution Type I and Type II error rates.

	Type	Fold1	Fold2	Fold3	Average
GLM	I	21.05%	18.42%	17.54%	19%
	II	24.51%	21.57%	22.55%	22.87%
TLR	I	16.67%	12.28%	7.9 %	12.28%
	II	24.51%	19.6%	27.45 %	23.85%

Table 4.13: Comparison between GLM and TLR for the general schizophrenia: Cross-validation estimate of schizophrenia predictive accuracy

	Fold1	Fold2	Fold3	Average
GLM	66.67%	66.67%	65.45 %	66.26%
TLR	69.69%	66.67%	73.94 %	70.10%

Table 4.14: Comparison between GLM and TLR for the general schizophrenia: Cross-validation Type I and Type II error rates.

	Type	Fold1	Fold2	Fold3	Average
GLM	I	52.63%	57.89%	70.17%	60.23%
	II	23.14%	20.37%	15.74%	19.75%
TLR	I	38.59%	36.84%	21.05 %	32.16%
	II	25.92%	31.48%	28.70 %	28.70%

Table 4.15: Comparison between GLM and TLR for the CD subtype of schizophrenia: Cross-validation estimate of predictive accuracy.

	Fold1	Fold2	Fold3	Average
GLM	59.81%	63.55%	61.68 %	61.68%
TLR	85.05%	80.37%	80.37 %	81.93%

Table 4.16: Comparison between GLM and TLR for the CD subtype of schizophrenia: Cross-validation Type I and Type II error rates.

	Type	Fold1	Fold2	Fold3	Average
GLM	I	36.84%	26.32%	35.09%	32.75%
	II	44%	48%	42%	44.67%
TLR	I	15.79%	17.54%	5.26 %	12.86%
	II	14%	22%	36 %	24%

Table 4.17: Comparison betwen TLR and GLM for the general schizophrenia and the CD subtype: Cross-validation prediction of specificity and sensitivity.

Data	Measure	TLR	GLM
CD subtype	Sensitivity	76%	55.33%
	Specificity	87.14%	67.25%
general schizophrenia	Sensitivity	71.3%	80.25%
	Specificity	67.84%	39.77%

Chapter 5

Conclusions and outlook

In this thesis, we have proposed a new class of threshold index logistic regression(TILoR) models, including partially linear and additive TILoR models. We have provided a maximum likelihood methodology to estimate the unknown parameters and studied the finite sample performance of the suggested estimators for the TILoR models. Empirical study by applying TILoR model to schizophrenia SNP data has found that our TILoR model outperforms linear logistic model and Nonparametric pathway-based regression model(NPR model) proposed by Zhi Wei and Hongzhe Li in 2007. In fact, TILoR model is the only model that is practically worthy among these three models. TILoR model has achieved the cross-validation predictive accuracy rates of about 70% for the general schizophrenia and 80% for the schizophrenia CD subtype by utilising only the genotype information in Chapter 4. In summary,

- We have extended the idea of threshold (auto)regression that was suggest by Tong (1983,1990) in nonlinear time series analysis to the nonlinear genomic analysis of SNP data which are categorical, and proposed a new class of threshold index logistic regression (TILoR) models. Based on this new framework of logistic regression with schizophrenia data sets, we have demonstrated that the TILoR models can well catch the genetically complex features of the schizophrenia SNP datasets in terms of the cross-validation predictive accuracy rates, specificity and sensitivity.

- The result of 70% accuracy of the cross-validation prediction with our TILoR models for general schizophrenia is quite close to the 80% broad heritability of schizophrenia, which, according to the experts' view from WAFSS(Western Australian Family Study of Schizophrenia), is an upper limit of prediction accuracy using genotype data alone. Using our TILoR models, the specificity for general schizophrenia prediction is 67.84%, sensitivity 71.3%. There are two possible reasons to explain the 10% prediction error gap. First, our empirical study is based on WAFSS schizophrenia data. The WAFSS group selected 23 genes which they assumed are most relevant to schizophrenia and genotyped them. It is possible that there are other schizophrenia genes that was not included in the WAFSS data. Second, like any classification model, prediction error is inevitable for TILoR model. It is not clear so far

among these two reasons, which one plays a more important role. To investigate the data reason, SNP data on the whole genome will be an ideal resource. Actually, through cooperation with the Welcome Trust, WAFSS is working through this direction. To investigate the model reason, we will put TILoR model into other empirical tests, which we leave that for future work.

- Schizophrenia CD subtype accounts for about 40% of the schizophrenia population and has the most severe symptom of mental disorders among all schizophrenia subgroups. Patients belong to this group have endured greatest life difficulties because of cognitive deficiency. Therefore, any breakthrough on this subtype will be particularly meaningful. Using TILoR model, for CD subtype, the cross-validation prediction accuracy rate is 81.93%, specificity 87.14% and sensitivity 76%. Our findings support the WAFSS view that CD subtype is genetically more significant and suggest that the broad heritability of schizophrenia may well above 80%.

- Schizophrenia is a complex disorder which means there are multiple susceptibility genes, each with small to modest effects that interact with each other and environmental factors to influence susceptibility for this disease. For each gene, more than one SNP shows association with schizophrenia, but rarely are individual SNPs highly significant.

The detailed analysis of our TILoR model results in Chapter 4 given a mathematical illustration of these biological understandings. Using TILoR model, two groups (15 and 14) SNPs have been selected for schizophrenia and CD subtype respectively, according to their components of α and β whose absolute values are greater. Among these two groups of SNPs, three SNPs: rs439401(APOE), rs17203281(DLG4), rs1454626(VLDLR) have already been widely reported associated with schizophrenia, cardiovascular risk factors and other diseases by medical studies around the world. Apart from these three "hot" SNPs, there are another three SNPs:rs11122357(DISC1), rs2283029(RELN), rs931671(NUDEL). They have been selected in both the schizophrenia group and the CD subtype group in our study, which suggests they are somehow special.These three genes(DISC1, RELN, NUDEL) have long been considered "schizophrenia" genes. The neurological mechanism study of these three polymorphisms is left to neurologists.

- Our TILoR schizophrenia prediction is based on the SNP genotype data alone, meaning that only a drop of blood taken from a participant will be sufficient for genotying. The final TILoR model involves about 40 SNPs on 12 genes, which dramatically reduces the cost of genotype and therefore, the cost of the prediction. It has the potential to becoming a part of medical diagnostic process. The medical diagnosis in psychiatry is problematic. Apart from the fact that there are differ-

ing theoretical views toward mental conditions, there are few lab tests available for various disorders. Being a readily available and relatively low cost lab test in genotyping sense, our findings appear promising by its accuracy. In particular, for children coming from a schizophrenia family, our findings can provide a disease risk reference to their life style chosen. For example, late adolescence and early adulthood are peak periods for the onset of schizophrenia. At this stage, avoiding environmental disadvantageous influences will be a sensible and rational way to better manage disease risk.

We have studied a rather promising new class of TILOR models for genomic data. Obviously, along the direction, there are lots of work worth doing. For example, we can extend the idea of the TILOR models to propose various new models for the categorial data with nonlinearity complexity and curse of dimensionality taken into account. Through the canonical link function, a threshold varying-coefficient logistic regression model could be suggested as:

$$\log\left\{\frac{P(Y_i=1|\mathbf{X}_i)}{1-P(Y_i=1|\mathbf{X}_i)}\right\} = a(U_i) + \mathbf{b}(U_i)^T\mathbf{X}_i. \tag{5.1}$$

Here, X is still a $p-$ dimensional covariate, and U_i a covariate of scalar.

A multi-threshold logistic regression model could be expressed as:

$$\log\left\{\frac{P(Y_i=1|\mathbf{X}_i)}{1-P(Y_i=1|\mathbf{X}_i)}\right\}=\sum_{k=1}^{K} I_{\{c_{k-1}\le\boldsymbol{\beta}^T\mathbf{X}_i<c_k\}}\{a_k+\mathbf{b}_k^T\mathbf{X}_i\}, \tag{5.2}$$

where $-\infty = c_0 < c_1 < c_2 \cdots < c_{K-1} < c_K = \infty$ are the thresholds, $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_p)^T\in\mathcal{R}^p$ with $\|\boldsymbol{\beta}\|=(\beta_1^2+\cdots+\beta_p^2)^{1/2}=1$ for identifiability. Also models (5.2) can be expressed as:

$$\log\left\{\frac{P(Y_i=1|\mathbf{X}_i)}{1-P(Y_i=1|\mathbf{X}_i)}\right\}=a(\boldsymbol{\beta}^T\mathbf{X}_i)+\mathbf{b}(\boldsymbol{\beta}^T\mathbf{X}_i)^T\mathbf{X}_i, \tag{5.3}$$

with

$$a(\boldsymbol{\beta}^T\mathbf{X}_i)=\sum_{k=1}^{K} I_{\{c_{k-1}\le\boldsymbol{\beta}^T\mathbf{X}_i<c_k\}}a_k$$

and

$$\mathbf{b}(\boldsymbol{\beta}^T\mathbf{X}_i)=\sum_{k=1}^{K} I_{\{c_{k-1}\le\boldsymbol{\beta}^T\mathbf{X}_i<c_k\}}\mathbf{b}_k,$$

of looks similar to a kind of generalised adaptive varying-coefficient models proposed by Fan, Yao and Cai (2003); see also Lu, Tjøstheim (2007) and Lu *et al.* (2009). But they are essentially different. In Fan, Yao and Cai (2003), $\mathbf{X}_i$ consists of continuous random variables, with nonparametric functions a and $\mathbf{b}$ in the form specified in (5.3). However, in the SNP data, $\mathbf{X}_i$ is a random vector of discrete categorial variables, and the nonparametric methodologies developed in Fan, Yao and Cai (2003) cannot apply here. Note that model (5.2) is parametric threshold logistic regression models, which could be well

applied to the SNP data. Model (5.2) could be called a class of adaptive threshold nonlinear logistic regression models. Other extension could also be made. In addition, the above models could be extended with genotyping errors taken into account (c.f., Zou, et al., 2008). We here only list a few for an appreciation. We leave these for future work.

Bibliography

[1] Akaike, H (1973), Information theory and an extension of the maximum likelihood principle. In second international Symposium in Information Theory. (B.N. Petroc and F. Caski, ects.) Akademiai Kiado, Budapest, PP.276-281.

[2] Amaratunga D. and Cabrera J. 2004, Exploration and analysis of DNA microarray and protein array data. New York, Wiley.

[3] Barnard, M.(1935) The secular variations of skull in four series of Egyptian skulls. Annals of Eugenics, 6:352-371.

[4] Braga-Neto, U.M. and Dougherty, E. R. (2004) Is cross-validation valid for small-sample microarray classification? Bioinformatics, 20, 374-380.

[5] Breiman, L. (1996) Bagging predictors. Machine Learning, 24: 123-140.

[6] Breiman, L. (1998) Arcing classifiers. Annals of Statistics, 26: 801-824.

[7] Breiman, L. (2001) Statistical modelling: the two cultures (with discussion). Statistical Science 16,199-231.

[8] Breiman, L. (2001) Random forests. Machine Learning Wadsworth. 45, 5-32.

[9] Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A.(1984) Classification and regression trees. Belmont, Wadsworth International Group.

[10] Chan, K. S. (1993) Consistency and Limiting Distribution of the Least Squares Estimator of a Threshold Autoregressive Model. The Annals of Statistics, 21, 520–533.

[11] Clark D, Skrobot OA, Adebiyi I, Susce MT, de Leon J, Blakemore AF, Arranz MJ.(2009)Apolipoprotein-E gene variants associated with cardiovascular risk factors in antipsychotic recipients.Eur Psychiatry. Oct;24(7):456-63.

[12] Crawford DC, Nord AS, Badzioch MD, Ranchalis J, McKinstry LA, Ahearn M, Bertucci C, Shephard C, Wong M, Rieder MJ, Schellenberg GD, Nickerson DA, Heagerty PJ, Wijsman EM, Jarvik GP.(2008),A common VLDLR polymorphism interacts with APOE genotype in the prediction of carotid artery disease risk. J Lipid Res. Mar;49(3):588-96.

[13] Dettling, M. and Buhlmann, P. (2003), Boosting for tumor classification with gene expression data. Bioinformatics 19, 1061-1069.

[14] Dudoit, S., Fridlyand,J. and Speed, T.P.(2002) Comparison of discrimination methods for the classification of tumors using gene expression data. JASA, March, vol 97, 457:77-87.

[15] Dudoit, S., Y.H. Yang, M.C. Callow and T. P. Speed (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Stat. Sinica.,12,111-140.

[16] Fan, J., Yao, Q. and Cai, Z. (2003) Adaptive varying-coefficient linear models. J. R. Statist. Soc. B, 65, 57–80.

[17] Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7:179-188.

[18] Fix, E. and J. Hodges.(1951) Discriminatory analysis. Nonparametric discrimination: consistency properties. Tech. Report 4, USAF School of Aviation Medicine, Randolph Field, Texas.

[19] Freund, Y. and R. E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55:119-139.

[20] Friedman, J.H. (1989). Regularized discriminant analysis, JASA, 84,165-175.

[21] Friedman J. (2001). Greedy function approximation:a gradient boosting machine. Annals of Statistics 29, 1189-1232.

[22] Furey, T. S., Cristianini, N., Duffy, N.,Bednarski, D.W., Schummer, M. and Haussler D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16, 906-914.

[23] Gao, J. (2007). Nonlinear time series: semiparametric and nonparametric methods. Chapman & Hall/CRC, Boca Raton.

[24] Golub, T.R., D.K. Slonim, P. Tamayo, C. Huard, M. Gasenbeek, J.P. Mesirov, H. Coller, M.L.Loh, J.R.Downing, M.A. Caligiuri, C.D. Bloomfield, and E. S. Lander(1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286:531-537.

[25] Good P. (1994) Permutation Tests: a practical Guide to Resampling Methods for Testing Hypotheses. Springer-verlag, New York.

[26] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002) Gene selection for cancer classification using support vector machines. Machine Learning, 46, 389-422.

[27] Hallmayer, J.F., Kalaydjieva, L., Badcock, J., Dragovic, M., Howell, S., Michie, P.T., Rock, D., Vile, D., illiams, R., Corder, E.H., Hollingsworth, K., and Jablensky, A. (2005) Genetic Evidence for a Dis-

tinct Subtype of Schizophrenia Characterized by Pervasive Cognitive Deficit. *Am. J. Hum. Genet.*, **77**, 468–476.

[28] Harrison P.J., Owen M.J.(2003) Genes for schizophrenia? Recent findings and their pathophysiological implications. The Lancet, 361:417-419.

[29] Hand, D.J. (1981) Discrimination and classification, Wiley Series in Probability and Mathematical Statistics.

[30] Hand, D.J. (1997) Construction and Assessment of Classification Rules. New York, Wiley.

[31] Härdle, W., Liang, H., and Gao, J. T. (2000). Partially Linear Models. Springer Phisica-Verlag.

[32] Hastie, T., R. Tibshirani, and J. Friedman (2001). Elements of Statistical Learning: Data Mining, Inference and Prediction. New York, Springer Verlag

[33] Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P. and Gusterson, B. (2001) Gene-expression profiles in hereditary breast cancer. New Eng. J. Med.,344,539-548.

[34] Kettenring J. R. (2006) The practice of cluster analysis. Journal of Classification 23:3-30.

[35] Khan, J., Wei, J.S.,Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., and Berthold, F., Schwab, M., Antonescu, C.R., Peterson C. and Meltzer, P.S. (2001) Classification and diagnostic prediction of cancers, using gene expression profiling and artificial neural networks. Nature Med., 7, 673-679.

[36] Kring SI, Brummett BH, Barefoot J, Garrett ME, Ashley-Koch AE, Boyle SH, Siegler IC, Srensen TI, Williams RB(2010) Impact of psychological stress on the associations between apolipoprotein E variants and metabolic traits: findings in an American sample of caregivers and controls, Psychosom Med., Jun;72(5):427-33.

[37] Li, H. and Luan, Y. (2005). Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. Bioinformatics 21, 2403-2409.

[38] Lu, Z., Steinskog, D.J., Tjøstheim, D. and Yao, Q. (2009) Adaptively varying coefficient spatial-temporal models, *J. Roy. Statist. Soc. Ser. B.* **71**, 859–880.

[39] Lu, Z., Tjøstheim, D. and Yao, Q. (2007). Adaptive Varying-Coefficient Linear Models for Stochastic Processes: Asymptotic Theory. *Statistica Sinica*, **17**, 177-197 .

[40] Manton KG, Woodbury MA and Tolley DH (1994) Statistical applications using fuzzy sets. John Wiley, New York

[41] Mclachlan, G.J. (1992) Discriminant Analysis and Statistical Pattern Recognition. Wiley, New York.

[42] Mclachlan, G.J., Do, K., and Ambroise, C. (2004) Analyzing Microarray Gene Expression Data. Wiley, Hoboken, NJ, USA.

[43] Molinaro, A. M., Simon, R. and Pfeiffer, R.M. (2005) Prediction error estimation: a comparison of resampling methods. Bioinformatics, 21, 3301-3307.

[44] Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (2002). Numerical Recipes in C++: The Art of Scientific Computing. Second Edition. Cambridge University Press, Cambridge.

[45] Schena M, Shalon D, Davis RW and Brown PO (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray'. Science 270:467-470.

[46] Schwarz, G. (1978) Estimating the dimension of a model. The Annals of Statistics 6, 461-464.

[47] Praveen Sharma, Sahni, N.S., Tibshirani, R., Skaane, P., Urdal, P., Berghagen, H., Jensen, M., Kristiansen, L., Moen, C., Pradeep Sharma,

Zaka, A., Arnes, J., Sauer, T., Akslen, L.A., Schlichting, E., Borresen-Dale, A. and Lonneborg A. (2005) Early detection of breast cancer based on gene-expression patterns in peripheral blood cells. Breast Cancer Res., 7, R634-R644.

[48] Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken c entroids of gene expression. PNAS, 99, 6567-6572.

[49] Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2003) Class prediction by nearest shrunken centriods, with applications to DNA microarrays. Stat. Sci., 18, 104-117.

[50] Thomas, D.C. (2004) *Statistical Methods in Genetic Epidemiology*. Oxford University Press, Oxford.

[51] Tong, H. 1983. *Threshold Models in Non-linear Time Series Analysis*. Springer-Verlag, New York.

[52] Tong, H. (1990). *Nonlinear time series: a dynamical system approach*. Oxford University Press, Oxford.

[53] Tsai SJ, Hong CJ, Cheng CY, Liao DL, Liou YJ.(2007) Association study of polymorphisms in post-synaptic density protein 95 (PSD-95) with schizophrenia. J Neural Transm. 114(4):423-6.

[54] Vant Veer L., Dai, H., Vijver, M.J., He, Y.D., Hart, A.M., Mao, M., Peterse, H.L., Kooy, K.V., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R. and Friend, S.H. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Letters to nature, Jan., 415:530-536.

[55] Vapnik, V. (1998). Statistical Learning Theory. Chichester, GB:John Wiley & Sons.

[56] Wei, Z. and Li, H. (2007) Nonparametric pathway-based regression models for analysis of genomic data. Biostatistics, **8**, 265–284.

[57] Wood I.A., Visscher, P.M. and Mengersen, K.L. (2007) Classification based upon gene expression data: bias and precision of error rates. Bioinformatics, vol. 23, 1363-1370.

[58] Woodbury MA, Clive J and Garson A Jr (1978) Mathematical typology: a grade of membership technique for obtaining disease definition. *Comput. Biomed. Res.*, **11**, 277–298.

[59] Xia, Y., Tong, H. and Li, W. K. (1999) On Extended Partially Linear Single-Index Models. *Biometrika*, **86**, 831–842.

[60] Yi, G.Y., He, W.Q., and Liang, H. (2009) "Analysis of correlated binary data under partially linear single-index logistic models". *Journal of Multivariate Analysis*, 100(2), 278–290.

[61] You, J. and Zhou, X. (2009) Partially linear models and polynomial spline approximation for the analysis of unbalanced panel data, Journal of Statistical Planning and Inference, 139(3), 679–695.

[62] Zhang, H. and Singer, B. (1999) *Recursive Partitioning in the Health Sciences*. Springer, New York.

[63] Zhang M.Q.(2000) Discriminant analysis and its application in DNA sequence motif recognition. Briefings in Bioinformatics. Vol. 1 No 4. 331-342. November.

[64] Zhu, X. (2005) Semi-supervised learning literature survey. Computer Sciences Technical Report 1530, University of Wisconsin, available at http://www.cs.wisc.edu/ jerryzhu/pub/ssl_survey.pdf

[65] Zuo Y, Zou G, Wang J, Zhao H and Liang H (2008). Optimal two-stage design for case-control association analysis incorporating genotyping errors. Annals of Human Genetics, 72(3), 375-387.

Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.