DEPARTMENT OF STATISTICS

University of Wisconsin

1210 West Dayton St.

Madison, WI 53706

TECHNICAL REPORT NO. 1066

September 25, 2002

Nonparametric Variable Selection and Model Building Via Likelihood Basis Pursuit

Hao Helen Zhang [1]

[1]Presently at Department of Statistics, North Carolina State University, Raleigh NC 27617. This work was supported in part by NSF Grant DMS 0072292, NIH Grant EY09946 and NASA Grant NAG5-10273.

NONPARAMETRIC VARIABLE SELECTION AND MODEL BUILDING VIA LIKELIHOOD BASIS PURSUIT

By

Hao Zhang

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN – MADISON

2002

Abstract

We propose a nonparametric penalized likelihood approach for variable selection and model building, called likelihood basis pursuit (LBP). In the setting of a tensor product reproducing kernel Hilbert space, we decompose the log likelihood into the sum of different functional components such as main effects and interactions, with each component represented by appropriate basis functions. The basis functions are chosen to be compatible with variable selection and model building in the context of a smoothing spline ANOVA model. Basis pursuit is applied to obtain the optimal decomposition in terms of having the smallest l_1 norm on the coefficients. We use the functional L_1 norm to measure the importance of each component and determine the "threshold" value by a sequential Monte Carlo bootstrap test algorithm. As a generalized LASSO-type method, LBP produces shrinkage estimates for the coefficients, which greatly facilitates the variable selection process, and provides highly interpretable multivariate functional estimates at the same time. To choose the regularization parameters appearing in the LBP models, generalized approximate cross validation (GACV) is derived as a tuning criterion. To make GACV widely applicable to large data set situations, its randomized version is proposed as well. A technique "slice modeling" is used to solve the optimization problem and makes the computation more efficient.

Several simulation studies are conducted to show the performance of the proposed LBP method. LBP has great potential for a wide range of research and application areas such as medical studies. In this dissertation we apply it to two large on-going epidemiological studies: the Wisconsin Epidemiological Study of Diabetic Retinopathy (WESDR) and the Beaver Dam Eye Study (BDES).

Acknowledgements

I would like to express my deepest gratitude to my adviser Professor Grace Wahba. She gave me invaluable advise and full support in the course of my study and research. Her dedication to statistics, her care and patience to the students, and her great personality have been inspiring me very much. I shall always appreciate her guidance which led me into the wonderful world of statistics.

I am very grateful to Professor Yi Lin, whose timely help, stimulating suggestions and encouragement accompany me through the dissertation. I have benefited a great deal from the numerous discussions with him. I would also like to thank Professors Barbara Klein and Ronald Klein for their original data sets, constructive advise and valuable insights on the application parts of this dissertation; Meta Voelker and Professor Michael Ferris for their expertise in computer programming and the fruitful collaboration; Professors Bates Douglas, Wei-Yin Loh and Chunming Zhang for their helpful thoughts on my research and service in my preliminary and final examination committees.

My former and present colleagues from the Department of Statistics have helped me in various ways these years. I want to thank them for their great support and friendship. Especially I am obliged to Kin-Yee Chan, Yonghua

Chen, Yunfei Chen, Bin Cheng, Xiaoyin Fan, Chunyang Gai, Fangyu Gao, Peter Hoff, Yuan Ji, Hongyu Jiang, Yoonjung Lee, Yoonkyung Lee, Chenlei Leng, Liang Li, Ruojia Li, Xiwu Lin, Lei Shen, Park Soomin, Jiafeng Sun, Haonan Tan, Hansheng Wang, Xiehong Xie, Jun Yan, Ming Yuan, Zhilong Yuan and Wei Yue. My thanks also extends to my friends Bohui Chen, Qun Chen, Hongju Liu, Hao Wang and Jin Zhu.

Finally, I would like to give my special thanks to my husband Guang, my parents and sisters for their love and support.

This research is in part supported by NSF grant NSF Grant DMS 0072292, NIH Grant EY09946 and NASA Grant NAG5-10273.

List of Figures

Contents

Chapter 1

Introduction

1.1 Motivation

This dissertation aims at developing a new statistical methodology for variable selection and model building. Variable selection, or dimension reduction, is fundamental to multivariate statistical model building, model validation and model selection. Not only does judicious variable selection improve the model's predictive ability, it also results in highly interpretable models and provides a better understanding of the underlying concept that generates data. Sometimes it is essential to exclude "noisy" or "irrevalent" variables out of models. One among many scenarios where variable selection plays an important role could be an ophthalmic disease study. When faced with a large number of medical, demographic, ocular and other covariates collected from the study participants, the medical researchers or doctors always ask "Which are relevant predictors for incidence or progression of a specific eye disease? ". In recent years, variable selection has become the focus of intensive research in several areas of application, for which datasets with tens or hundreds of thousands of variables are available. These areas include text processing and genomics, particularly gene expression array data.

Traditionally, various variable selection approaches such as forward selection, backward selection, stepwise selection and best subset selection have been constructed in the frame of multivariate linear regression or logistic regression models, and the well-known criteria like Mallow's C_p, AIC and BIC are often used to penalize the number of non-zero parameters. See Linhart & Zucchini (1986) for an introduction of linear models. These are least squares methods and tend to exhibit numerical instability in the presence of collinearity. To achieve better prediction and reduce the variances of estimators, a number of shrinkage estimation approaches have been developed. The nonnegative garrotte by Brieman (1995) replaces the least squares criterion by constrained optimization criterion. Bridge regression was introduced by Frank & Friedman (1993) , which is a constrained least squares method subject to an L_p penalty with $p \geq 1$. Two special cases of bridge regression are: the LASSO proposed by Tibshirani (1996) when $p = 1$ and ridge regression when $p = 2$. Due to the nature of L_1 penalty, LASSO tends to shrink smaller coefficients to zero and hence gives concise models. It also shows the stability of ridge regression estimates. Fu (1998) made a thorough comparison between the bridge model and LASSO. Knight & Fu (2000) proved some asymptotic results for LASSO-type estimators. In the case of wavelet regression, this L_1 penalty approach is called "basis pursuit".

Chen, Donoho & Saunders (1998) discussed atomic decomposition by basis pursuit in some detail. Gunn & Kandola (2002) proposed a structural modeling algorithm with sparse kernels. Most recently, Fan & Li (2001) used a non-concave penalized likelihood approach with the smoothly clipped absolute deviation (SCAD) penalty function, which resulted in an unbiased, sparse and continuous estimator. The motivation of this dissertation is to provide a flexible nonparametric alternative to the parametric approaches for variable selection as well as model building. Yau, Kohn & Wood (2001) presented a Bayesian methodology of variable selection for high dimensional multinomial regression in a nonparametric manner.

1.2 Goal

Variable/Model selection approaches based on linear models usually lead to simple and interpretable models, except that one drawback commonly exists in them: the possible model mismatch. Since parametric methods rely heavily on model assumptions, an inadequate or incorrect assumption might produce the misleading results in real complex situations. In response to the lack of model flexibility, researchers often turn to nonparametric statistical modeling, which is desired due to its assumption-free and data-driven nature and great flexibility. Nonparametric approaches have proven to be very effective in a wide variety of statistical problems such as regression and classification. The principal focus of this dissertation is to develop a non-parametric approach for variable selection and model building.

The beautiful mathematical structure of Reproducing Kernel Hilbert Space (RKHS) theory makes it able to provide a rigorous and effective framework for discovering the implicit relationships that exist in data and providing accurate approximation of general multidimensional functions. See Aronszajn (1950) for a complete introduction on RKHS and its theoretical properties. Smoothing spline analysis of variance (SS-ANOVA), a kernel-based model developed on RKHS theory, is often used when statisticians need to investigate the interactions between variables. It has been widely used for smooth multivariate interpolation of arbitrarily scattered data in many areas and studied intensively for Gaussian data. Wahba, Wang, Gu, Klein & Klein (1995) gave a general setting for applying SS-ANOVA model to data from exponential families. Gu (2002) provided a comprehensive review of SS-ANOVA and some recent progress as well. In this disseration a unified model, which appropriately combines the SS-ANOVA model and basis pursuit, is developed for variable selection and model building. The constructed model is equipped with regularization parameters, which balance the tradeoff between the likelihood fit to data and selection of important functional components. This new approach has been inspired by the LASSO and is capable of both selecting important independent variables and providing accurate prediction for response variables.

1.3 Outline of Dissertation

This dissertation is organized as follows. Chapter 2 first introduces the notations used in this dissertation. It then reviews the basis setting of the smoothing spline ANOVA model for exponential families, in particular, for Bernoulli distribution. In the remainder of the chapter the general structure of the proposed likelihood basis pursuit (LBP) approach is illustrated, with emphasis on two types of models: the main effects model and the two-factor interaction model. And the models are generalized to incorporate categorical variables.

Chapter 3 addresses the important issue of how to choose regularization parameters for the LBP model. The generalized approximate cross validation (GACV) proposed by Xiang & Wahba (1996) for typical SS-ANOVA models is not directly applicable to the LBP model due to the appearance of L_1 penalty. Hence an extension of GACV is derived as an adaptive tuning criterion.

Chapter 4 suggests two measures of importance of the variables and, if desired, their interactions. The results related to two measures will be reported to compare the scores and ranks for all the variables. Central to the LBP model is the choice of a threshold, which distinguishes important variabes from non-important ones. A sequential Monte Carlo bootstrap test algorithm is developed to determine the threshold for variable selection.

Chapter 5 provides a comprehensive discussion as to numerical computation issue involved in the LBP model fitting. In this chapter, the proposed statistical model is treated as a challenging nonlinear optimization problem from the view point of computer programming. Then much of this discussion centers around the software MINOS used to solve the constrained optimization problem. In addition, we use the "slice modeling" technique to optimize the programming structure and improve the efficiency of the code. Ferris & Meta (2001) have addressed the advantage of deploying the slice technique under many circumstances.

Chapter 6 demonstrates the performance of the LBP model on synthetic data. Then the applications of the LBP model to real world datasets are presented. Chapter 7 applies the LBP model to the Wisconsin Epidemiological Study of Diabetic Retinopathy (WESDR) and carries out a variable selection analysis for the four-year risk of progression of diabetic retinopathy. Chapter 8 shows the application of the LBP on the Beaver Dam Eye Study (BDES) and selects important risk factors for the five-year risk of mortality. These data sets are used to illustrate how nonparametric variable selection can be used to model the data set and recover significant variables which might be omitted by some traditional parametric approaches.

Finally, Chapter 9 contains some concluding remarks and future research plan.

Chapter 2

Likelihood Basis Pursuit

In the following are stated the two problems under investigation. Given a dataset consisting of observations for independent variables and response variables, (1) how can important independent variables that are relevant to responses be selected? (2) how can interpretable and flexible models be built for the future response forecasts?

2.1 Exponential Family

Suppose conditional on $\mathbf{X}$, Y is a random variable from an exponential family with the density in the canonical form

$$h(y, f(\mathbf{x}), \phi) = \exp\{ \ [\ yf(\mathbf{x}) - b(f(\mathbf{x})) \ ]/a(\phi) + c(y, \phi) \ \}, \qquad (2.1)$$

where a, b and c are given. $b(\cdot)$ is a strictly convex function on any bounded set, $\mathbf{x}$ the vector of covariates, ϕ nuisance parameters and f an unknown regression function in $\mathbf{x}$. To allow a flexible estimation, we always assume f is a smooth function. Denote the mean and variance of Y by μ and σ^2.

We are interested in exploring the dependence of the variable Y on the predictor variables $\mathbf{X} = (X^1, \ldots, X^D)$, which are also known as explanatory variables. Typically $\mathbf{X}$ is in a high dimensional space $\mathcal{X} = \mathcal{X}^1 \otimes \cdots \otimes \mathcal{X}^D$, where $\mathcal{X}^\alpha,\ \alpha = 1, \ldots, D$, is some measurable space and $\otimes$ denotes the tensor product operation. Some of the covariates are continuous while others are categorical. In Section 2.3, all the components of $\mathbf{X}$ are assumed to be continuous; while in Section 2.4, we take into account categorical variables.

Bernoulli data is of particular interest because it has broad applications to risk estimation in scientific research such as medical studies. The main focus of this dissertation is on handling nonparametric binary regression with a large number of variables, possibly with interactions between variables. Suppose that Y takes on two values $\{0, 1\}$ with $p(\mathbf{x}) \equiv \text{prob}(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{e^{f(\mathbf{x})}}{1 + e^{f(\mathbf{x})}}$. f is the so-called "logit" function with $f(\mathbf{x}) = log(\frac{p(\mathbf{x})}{1 - p(\mathbf{x})})$. In the expression (2.1), $a(\phi) = 1, b(f) = log(1 + e^f), c(y, \phi) = 0$. For n independent observations $(\mathbf{x}_i, y_i),\ i = 1, \ldots, n$, the negative log likelihood is

$$\mathcal{L} = \sum_{i=1}^{n} [-y_i f(\mathbf{x}_i) + b(f(\mathbf{x}_i))\,] \equiv \sum_{i=1}^{n} [-l(y_i, f(\mathbf{x}_i))\,]. \tag{2.2}$$

2.2 Variable Selection Problem

The variable selection problem is most familiar in the linear regression context. Under normal assumption, if letting S be an index subset of $\{1, \ldots, D\}$, the variable selection problem is to select and fit a model of the form

$$Y = \sum_{j \in S} X^j \beta_j + \epsilon,$$

where β_j's are regression coefficients and $\epsilon \sim N(0, \sigma^2)$. In the generalized linear model setting, $f(\mathbf{X})$ is assumed to be of a linear from too. A wide variety of traditional methods are subset selection, stepwise selection, forward selection and so on. The criteria such as Mallow's C_p, Akaike Information Criterion (AIC) (Akaike 1974) and Schwarz's Bayesian Information Criteria (BIC) (Schwarz 1978) are widely used to gauge the number of variables included in the model. Under two situations the linear models may not produce stable estimates, when collinearity exists between covariates and small perturbation or change occurs to the data. Shrinkage estimation approaches mitigate the instability and improves the fitting. They include nonnegative garrotte by Brieman (1995) , bridge regression by Frank & Friedman (1993) and the LASSO by Tibshirani (1996). Some new methods related to shrinkage modeling are Gunn & Kandola (2002) and Fan & Li (2001). The linear models are attractive due to their tractability and simplicity. In addition, many problems of interest can be approximated well by linear methods.

However, the linear model is lack of flexibility which limits its application in complicated situations. A variety of approaches have been proposed to allow more flexibility than is inherent in simple linear/ parametric models. These nonparametric models include Classification and Regression Treed

(CART) and Multivariate Adaptive Regression Splines (MARS).

Generally speaking, variable selection is a special model selection problem, where each model under consideration corresponds to a distinct subset S. Selecting a model from a large class of plausible models is an important problem in statistics and machine learning. A distinguishing feature of this model selection problem is its enormous size. Even when we only take account in the additive linear models, with a moderate value of D, the model space has size 2^D. The computation can be prohibitively expensive.

2.3 Smoothing Spline ANOVA for Exponential Families

Rather assuming a simple form for the model as in linear regression, we allow f to vary in a high-dimensional function space, which leads to a more flexible estimate for the target function. For example, f is assumed as an elements of some (reproducing kernel Hilbert) space of smooth functions, and it is estimated by minimizing a penalized likelihood. Similar to the classical analysis of variance (ANOVA), for any function $f(\mathbf{x}) = f(x^1, \ldots, x^D)$ on a product domain $\mathcal{X}$, we define its functional ANOVA decomposition

$$
\begin{aligned}
f(\mathbf{x}) \;&=\; b_0 + \sum_{\alpha=1}^{D} f_\alpha(x^\alpha) \;+ \sum_{\alpha<\beta} f_{\alpha\beta}(x^\alpha, x^\beta) \\
&\quad + \;\text{all higher-order interactions,} \qquad (2.3)
\end{aligned}
$$

where b_0 is constant, f_α's are the main effects, and $f_{\alpha\beta}$'s are the two-factor interactions. The identifiability of the terms is assured by side conditions through averaging operators. In practice, the decomposition (2.3) is truncated somewhere to get different sub-models. Higher-order interaction terms are often excluded to make the model more "estimable" and "interpretable".

Without loss of generality, we assume each of the covariates is in the range $[0,1]$, or, we scale each covariate to the interval $[0,1]$. A reproducing kernel Hilbert space on $[0,1]^D$ corresponding to the decomposition (2.3) is constructed following Wahba et al. (1995) . Let $\mathcal{H}^{(\alpha)}$, $\alpha = 1,\dots,D$, be the second-order Sobolev-Hilbert space on $[0,1]$. Mathematically, $\mathcal{H}^{(\alpha)} = \{g : g(x^\alpha), g'(x^\alpha)$ are absolutely continuous, $g''(x^\alpha) \in \mathcal{L}_2[0,1]\}$. When we endow $\mathcal{H}^{(\alpha)}$ with the inner product

$$\begin{aligned}(g_1,g_2) = \quad & (\int_0^1 g_1(t)\ dt)(\int_0^1 g_2(t)\ dt) + (\int_0^1 g_1'(t)\ dt)(\int_0^1 g_2'(t)\ dt)\\ + \quad & \int_0^1 g_1''(t) g_2''(t)\ dt,\end{aligned}$$

$\mathcal{H}^{(\alpha)}$ is an RKHS with kernel $K(s,t) = 1 + k_1(s)k_1(t) + k_2(s)k_2(t) - k_4(|s-t|)$, where

$$\begin{aligned}k_1(t) &= t - \frac{1}{2}\\ k_2(t) &= \frac{1}{2}(k_1^2(t) - \frac{1}{12})\\ k_4(t) &= \frac{1}{24}(k_1^4(t) - \frac{1}{2}k_1^2(t) + \frac{7}{240}).\end{aligned}$$

This is the special case of equation (10.2.4) in Wahba (1990) when $m = 2$. Other kernels could be used to generate different kinds of RKHS, such as

Gaussian kernel and polynomial kernel. The forms of kernels are of central importance and they determine the class of functions from whom the solution is drawn and thus the accuracy of the estimation.

Next, we decompose $\mathcal{H}^{(\alpha)}$ into the direct sum of two orthogonal subspaces $\mathcal{H}^{(\alpha)} = \{1\} \oplus \mathcal{H}_1^{(\alpha)}$. Here $\{1\}$ is the "mean" space and $\mathcal{H}_1^{(\alpha)}$ is the "contrast" space generated by the kernel $k_1(s)k_1(t) + k_2(s)k_2(t) - k_4(|s-t|)$. Then the tensor product RKHS is

$$\otimes_{\alpha=1}^{D} \mathcal{H}^{(\alpha)} = [1] \oplus \sum_{\alpha=1}^{D} \mathcal{H}_1^{(\alpha)} \oplus \sum_{\alpha<\beta} [\mathcal{H}_1^{(\alpha)} \otimes \mathcal{H}_1^{(\beta)}] \oplus \cdots .$$

Each functional component in the decomposition (2.3) falls in the corresponding subspace of $\otimes_{\alpha=1}^{D} \mathcal{H}^{(\alpha)}$. Any truncation in the functional ANOVA decomposition corresponds to a truncation of subspaces of $\otimes_{\alpha=1}^{D} \mathcal{H}^{(\alpha)}$. To encompass the linear model as a special case of our model, we make a further orthogonal decomposition on $\mathcal{H}_1^{(\alpha)}$ by $\mathcal{H}_1^{(\alpha)} = \mathcal{H}_{1,\pi}^{(\alpha)} \oplus \mathcal{H}_{1,s}^{(\alpha)}$. $\mathcal{H}_{1,\pi}^{(\alpha)}$ is the "parametric" contrast generated by the kernel $k_1(s)k_1(t)$. $\mathcal{H}_{1,s}^{(\alpha)}$ is the "nonparametric" or "smooth" contrast, generated by the kernel $K_1(s,t) \equiv k_2(s)k_2(t) - k_4(|s-t|)$. Thus $\mathcal{H}_1^{(\alpha)} \otimes \mathcal{H}_1^{(\beta)}$ is a direct sum of four orthogonal subspaces:

$$\mathcal{H}_1^{(\alpha)} \otimes \mathcal{H}_1^{(\beta)} = [\mathcal{H}_{1,\pi}^{(\alpha)} \otimes \mathcal{H}_{1,\pi}^{(\beta)}] \oplus [\mathcal{H}_{1,\pi}^{(\alpha)} \otimes \mathcal{H}_{1,s}^{(\beta)}] \oplus [\mathcal{H}_{1,s}^{(\alpha)} \otimes \mathcal{H}_{1,\pi}^{(\beta)}] \oplus [\mathcal{H}_{1,s}^{(\alpha)} \otimes \mathcal{H}_{1,s}^{(\beta)}].$$

Continuing this way results in an orthogonal decomposition of $\otimes_{\alpha=1}^{D} \mathcal{H}^{(\alpha)}$ into tensor sums of products of finite dimensional parametric spaces, plus smooth main effect subspaces, plus two-factor interaction spaces of three possible forms: parametric $\otimes$ parametric, smooth $\otimes$ parametric and smooth

$\otimes$ smooth, plus three-factor and higher order interaction subspaces. The reproducing kernel of $\otimes_{\alpha=1}^{D} \mathcal{H}^{(\alpha)}$ is

$$\prod_{\alpha=1}^{D} (1 + k_1(s^\alpha)k_1(t^\alpha) + K_1(s^\alpha, t^\alpha)). \tag{2.6}$$

Let $\mathcal{H}$ be the model space after truncation. Then $\mathcal{H}$ is a direct sum of Q, say, component subspaces. Each component subspace is denoted as $\mathcal{H}_l$, and its reproducing kernel as R_l, for $l = 1, \ldots, Q$. Each R_l is one term in the expansion of (2.6). Then $\mathcal{H} = \oplus_{l=1}^{Q} \mathcal{H}_l$, and its kernel is $K = \sum_{l=1}^{Q} R_l$.

2.4 Likelihood Basis Pursuit

To avoid model overfitting we place some constraints on the coefficients in the decomposition, which leads to a penalized likelihood approach. Basis pursuit (BP) is a principle for decomposing a signal into an optimal superposition of dictionary elements, where "optimal" means having the smallest l_1 norm of the coefficients among all such decompositions. Chen et al. (1998) illustrated atomic decomposition by basis pursuit in the context of wavelet regression. In this dissertation likelihood basis pursuit (LBP) is proposed as a nonparametric variable selection and model building approach. Essentially we will apply basis pursuit to the negative log likelihood in the context of a dictionary based on an SS-ANOVA decomposition, and then select the important components from the multivariate function estimate. The variational problem for LBP model is

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \left[-l(y_i, f(\mathbf{x}_i))\right] + J_\lambda(f). \tag{2.7}$$

Here $J_\lambda(f)$ denotes the l_1 norm of the coefficients in the decomposition of f. It is a generalized version of LASSO penalty for nonparametric models. The l_1 penalty often produces coefficients that are exactly zero, therefore, gives sparse solutions. The sparsity of the LBP solutions enhances ability of the method to select important variables from a large set. The comparison of the l_1 penalty with other forms of penalty can be found in Tibshirani (1996) and Fan & Li (2001) . The regularization parameter λ balances the trade-off between minimizing the negative log likelihood function and the penalty part.

For the usual smoothing spline modeling, the penalty $J_\lambda(f)$ is a quadratic norm or seminorm in an RKHS. Kimeldorf & Wahba (1971) showed that the minimizer f_λ for the smoothing spline model falls in span$\{K(\mathbf{x}_i, \cdot),\ i = 1, \ldots, n\}$, though the model space is of infinite dimensions. For penalized likelihood with a non-quadratic penalty like the l_1 penalty, it is very hard to obtain analytic solutions. In light of the results for the quadratic penalty situation, we propose using a sufficiently large number of basis functions to approximate the target function. This assures that the proposed method can handle a large number of main effects and interactions because each functional component is represented as a linear combination of basis terms instead of a smoothing spline. When including all the n data points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ to generate the basis functions, we use span$\{R_l(\mathbf{x}_i, \cdot),\ i = 1, \ldots, n, l = 1, \ldots, Q\}$ as the approximating function space.

This setup demands intensive computation and the application is limited for large-scale problems (when n is big). Thus we adopt the parsimonious bases approach used by Xiang & Wahba (1998) , Ruppert & Carroll (2000) , Lin, Wahba, Xiang, Gao, Klein & Klein (2000) , Lin et al. (2000) and Yau et al. (2001) . It has been shown that the number of basis terms can be much smaller than n without degrading the performance of the estimation. For $N \leq n$, we subsample N points from the whole data and denote them as $\{\mathbf{x}_{1*}, \ldots, \mathbf{x}_{N*}\}$. We use these subsamples to generate basis functions and the tensor sum RKHS $\mathcal{H}_*$ as the approximation function space,

$$\mathcal{H}_* = \oplus_{l=1}^{Q} \text{ span}\{R_l(\mathbf{x}_{j*}, \cdot), j = 1, \ldots, N\}. \tag{2.8}$$

Notice that the space span$\{R_l(\mathbf{x}_{j*}, \cdot), j = 1, \ldots, N\}$ is a subspace of $\mathcal{H}_l$ for $l = 1, \ldots, Q$.

The issue of choosing N and the subsamples is important. In principle, the subspace spanned by the chosen basis terms needs to be rich enough to provide a decent fit to the true curve. Note that we are not wasting any data resource here, since all the data points are involved in fitting the model, though only a subset of them are selected for generating basis functions. We apply the simple random subsampling technique to choose the subsamples in this paper. Alternatively, a cluster algorithm may be used, such as in Xiang & Wahba (1998) and Yau et al. (2001) . The basic idea is to first group the data into N clusters which have maximum separation by some good algorithm, and then within each cluster one data point is randomly chosen as a representative to be included in the base pool. This scheme usually provides well-separated subsamples.

In practice, we always truncate the functional ANOVA decomposition in (2.3) to get various models. Two popular truncated models focused in this dissertation are the main effects model and the two-factor interaction model. The proposed approach can be easily extended to multi-factor interaction models.

2.4.1 Main Effects Model

The main effects model, also known as the additive model, is a sum of D functions of one variable. By retaining only main effect component spaces in (2.8), we use the following function space

$$\mathcal{H}_* = \oplus_{\alpha=1}^{D} \operatorname{span}\{k_1(x^\alpha), K_1(x^\alpha, x_{j*}^\alpha), j = 1, \ldots, N\} \equiv \oplus_{\alpha=1}^{D} \mathcal{H}_*^{(\alpha)}. \quad (2.9)$$

Any element $f_\alpha \in \mathcal{H}_*^{(\alpha)}$ has the representation

$$f_\alpha(x^\alpha) = b_\alpha k_1(x^\alpha) + \sum_{j=1}^{N} c_{\alpha,j} K_1(x^\alpha, x_{j*}^\alpha), \quad (2.10)$$

and the function estimate f is

$$f(\mathbf{x}) = b_0 + \sum_{\alpha=1}^{D} b_\alpha k_1(x^\alpha) + \sum_{\alpha=1}^{D} \sum_{j=1}^{N} c_{\alpha,j} K_1(x^\alpha, x_{j*}^\alpha), \quad (2.11)$$

where $k_1(\cdot)$ and $K_1(\cdot,\cdot)$ are defined in Section 2.1. The likelihood basis pursuit estimate f is obtained by minimizing

$$\frac{1}{n}\sum_{i=1}^{n} \left[-l(y_i, f(\mathbf{x}_i))\right] + \lambda_\pi \sum_{\alpha=1}^{D} |b_\alpha| + \lambda_s \sum_{\alpha=1}^{D}\sum_{j=1}^{N} |c_{\alpha,j}|, \tag{2.12}$$

where (λ_π, λ_s) are the regularization parameters. It is possible to allow different λ's for different variables and/or to put constraints on the λ's.

2.4.2 Two-factor Interaction Model

The two-factor interaction space consists of the "parametric" part and the "smooth" part. The parametric part is generated by D parametric main effect terms $\{k_1(x^\alpha),\ \alpha = 1,\ldots,D\}$ and $\frac{D(D-1)}{2}$ parametric-parametric interaction terms $\{k_1(x^\alpha)k_1(x^\beta), \alpha = 1,\ldots,D, \beta < \alpha\}$. The smooth term is the tensor sum of the spaces generated by smooth main effect terms, parametric-smooth interaction terms and smooth-smooth interaction terms. The function space used is

$$\mathcal{H}_* \equiv \ \oplus_{\alpha=1}^{D}\ \mathcal{H}_*^{(\alpha)} \ + \ \oplus_{\beta<\alpha}\ \mathcal{H}_*^{(\alpha\beta)}. \tag{2.13}$$

For each pair $\alpha \neq \beta$,

$$\begin{aligned}\mathcal{H}_*^{(\alpha\beta)} = \text{span}\ \ \{\ \ & k_1(x^\alpha)k_1(x^\beta),\ K_1(x^\alpha, x_{j*}^\alpha)k_1(x^\beta)k_1(x_{j*}^\beta), \\ & K_1(x^\alpha, x_{j*}^\alpha)K_1(x^\beta, x_{j*}^\beta),\ j = 1,\ldots,N\ \ \},\end{aligned} \tag{2.14}$$

and the interaction term $f_{\alpha\beta}(x^\alpha, x^\beta)$ has the representation

$$\begin{aligned} f_{\alpha\beta}(x^\alpha, x^\beta) &= b_{\alpha\beta} k_1(x^\alpha) k_1(x^\beta) + \sum_{j=1}^{N} c^{\pi s}_{\alpha\beta,j} K_1(x^\alpha, x^\alpha_{j*}) k_1(x^\beta) k_1(x^\beta_{j*}) \\ &+ \sum_{j=1}^{N} c^{ss}_{\alpha\beta,j} K_1(x^\alpha, x^\alpha_{j*}) K_1(x^\beta, x^\beta_{j*}). \end{aligned}$$

Different penalties are allowed for different types of terms: parametric main effect terms, parametric-parametric interaction terms, smooth main effect terms, parametric-smooth interaction terms and smooth-smooth interaction terms. Therefore there are five tuning parameters $\{\lambda_\pi, \lambda_{\pi\pi}, \lambda_s, \lambda_{\pi s}, \lambda_{ss}\}$ in the two-factor interaction model. The optimization problem is: minimize

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^{n} [-l(y_i, f(\mathbf{x}_i))\ ] + \lambda_\pi (\sum_{\alpha=1}^{D} |b_\alpha|) + \lambda_{\pi\pi} (\sum_{\alpha<\beta} |b_{\alpha\beta}|) \\ &+ \lambda_s (\sum_{\alpha=1}^{D} \sum_{j=1}^{N} |c_{\alpha,j}|) + \lambda_{ss} (\sum_{\alpha<\beta} \sum_{j=1}^{N} |c^{ss}_{\alpha\beta,j}|) + \lambda_{\pi s} (\sum_{\alpha\neq\beta} \sum_{j=1}^{N} |c^{\pi s}_{\alpha\beta,j}|). \end{aligned} \tag{2.15}$$

2.4.3 Incorporating Categorical Variables

In real applications, some of the covariates may be categorical. For example, in many medical studies, sex, race, smoking history and marital status all take discrete values. Similar to regression analysis, categorical variables require special attention because, unlike continuous variables, they cannot by entered into the LBP model just as they are. Thus categorical variables need to be recoded into a series of variables which can then be entered into the fitting model. In Section 2.1 and 2.2, the main effects model (2.12) and the two-factor interaction model (2.15) are proposed for continuous variables only. In this section we generalize these models to incorporate categorical variables.

Assume there are r categorical variables and denote them by a vector $\mathbf{Z} = (Z^1, \ldots, Z^r)$. Usually each variable has several categories. There are a variety of coding systems that can be used when coding categorical variables such as simple coding, forward difference coding and backward difference coding and helmert coding. Ideally, we should choose a coding system that reflects the comparisons that we want to make. For example,For example, to compare each level to the next higher level, we may use "forward difference" coding. To compare each level to the mean of the subsequent levels of the variable, "Helmert" coding is one choice. By deliberately choosing a coding system, we can obtain comparisons that are most meaningful for testing your hypotheses. Here we will use "Helmert" coding system to incorporate the categorical variables. Assume Z^1 takes two responses $\{1, 2\}$, a mapping Φ_1 corresponding to Helmert coding is defined as:

$$
\begin{aligned}
\Phi_1(z^1) &= \frac{1}{2} \quad \text{if} \;\; z^1 = 1 \\
&= -\frac{1}{2} \quad \text{if} \;\; z^1 = 2.
\end{aligned}
$$

The mapping is chosen to make the range of categorical variables comparable with that of continuous variables. For any variable with $C > 2$ categories, $C-1$ contrasts are needed. One set of mappings which correspond to Helmert coding can be defined as

$$[\Phi_1, \ldots, \Phi_{C-1}] = \frac{1}{2} \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\frac{1}{C-1} & 1 & 0 & \cdots & 0 \\ -\frac{1}{C-1} & -\frac{1}{C-2} & 1 & \cdots & 0 \\ & \cdots & & \cdots & \\ -\frac{1}{C-1} & -\frac{1}{C-2} & -\frac{1}{C-3} & \cdots & -1 \end{bmatrix}.$$

Here we derive the models for the simplest case: all Z's are two-level categorical. Similar ideas are easily extended for variables having more than two categories.

- The main effects model which incorporates the categorical variables is: minimize

$$\frac{1}{n}\sum_{i=1}^{n}[-l(y_i, f(\mathbf{x}_i, \mathbf{z}_i)] + \lambda_\pi(\sum_{\alpha=1}^{D}|b_\alpha| + \sum_{\gamma=1}^{r}|B_\gamma|) + \lambda_s\sum_{\alpha=1}^{D}\sum_{j=1}^{N}|c_{\alpha,j}| \tag{2.16}$$

  subject to

$$f(\mathbf{x}, \mathbf{z}) = b_0 + \sum_{\alpha=1}^{D} b_\alpha k_1(x^\alpha) + \sum_{\gamma=1}^{r} B_\gamma \Phi_\gamma(z^\gamma) + \sum_{\alpha=1}^{D}\sum_{j=1}^{N} c_{\alpha,j} K_1(x^\alpha, x^\alpha_{j*}).$$

  For $\gamma = 1, \ldots, r$, the function Φ_γ is actually the main effect of Z^γ. Thus we assign the same parameter λ_π to the coefficients $|B|$'s as to the coefficients $|b|$'s.

- The two-factor interaction model which incorporates categorical variables is much more complicated than in the continuous case. Compared with the expression in (2.15), four new types of terms are taken into account: categorical main effects, categorical-categorical interactions, "parametric continuous"-categorical interactions and "smooth continuous"-categorical interactions. The modified two-factor interaction model is: minimize

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[-l(y_i, f(\mathbf{x}_i,\mathbf{z}_i))\right] + \lambda_\pi\Big(\sum_{\alpha=1}^{D}|b_\alpha| + \sum_{\gamma=1}^{r}|B_\gamma|\Big)\\
&\quad +\lambda_{\pi\pi}\Big(\sum_{\alpha<\beta}|b_{\alpha\beta}| + \sum_{\gamma<\theta}|B_{\gamma\theta}| + \sum_{\alpha=1}^{D}\sum_{\gamma=1}^{r}|P_{\alpha\gamma}|\Big)\\
&\quad +\lambda_{\pi s}\Big(\sum_{\alpha\neq\beta}\sum_{j=1}^{N}|c^{\pi s}_{\alpha\beta,j}| + \sum_{\alpha=1}^{D}\sum_{\gamma=1}^{r}\sum_{j=1}^{N}|c^{\pi s}_{\alpha\gamma,j}|\Big)\\
&\quad +\lambda_s\Big(\sum_{\alpha=1}^{D}\sum_{j=1}^{N}|c_{\alpha,j}|\Big) + \lambda_{ss}\Big(\sum_{\alpha<\beta}\sum_{j=1}^{N}|c^{ss}_{\alpha\beta,j}|\Big) \qquad (2.18)
\end{aligned}$$

subject to

$$\begin{aligned}
f(\mathbf{x},\mathbf{z}) &= b_0 + \sum_{\alpha=1}^{D} b_\alpha k_1(x^\alpha) + \sum_{\gamma=1}^{r} B_\gamma \Phi_\gamma(z^\gamma) + \sum_{\alpha<\beta} b_{\alpha\beta} k_1(x^\alpha)k_1(x^\beta)\\
&+ \sum_{\gamma<\theta} B_{\gamma\theta}\Phi_\gamma(z^\gamma)\Phi_\theta(z^\theta) + \sum_{\alpha=1}^{D}\sum_{\gamma=1}^{r} P_{\alpha\gamma}k_1(x^\alpha)\Phi_\gamma(z^\gamma)\\
&+ \sum_{\alpha\neq\beta}\sum_{j=1}^{N} c^{\pi s}_{\alpha\beta,j} K_1(x^\alpha, x^\alpha_{j*})k_1(x^\beta)k_1(x^\beta_{j*})\\
&+ \sum_{\alpha=1}^{D}\sum_{\gamma=1}^{r}\sum_{j=1}^{N} c^{\pi s}_{\alpha\gamma,j} K_1(x^\alpha, x^\alpha_{j*})\Phi_\gamma(z^\gamma)\\
&+ \sum_{\alpha=1}^{D}\sum_{j=1}^{N} c_{\alpha,j}K_1(x^\alpha, x^\alpha_{j*}) + \sum_{\alpha<\beta}\sum_{j=1}^{N} c^{ss}_{\alpha\beta,j}K_1(x^\alpha, x^\alpha_{j*})K_1(x^\beta, x^\beta_{j*}).
\end{aligned}$$

We assign different regularization parameters for main effect terms, parametric-parametric interaction terms, parametric-smooth interaction terms and smooth-smooth interaction terms. Thus the coefficients $|B_{\gamma\theta}|$'s and $|P_{\alpha\gamma}|$'s are associated with the parameter $\lambda_{\pi\pi}$ and the coefficients $|c^{\pi s}_{\alpha\gamma,j}|$'s with $\lambda_{\pi s}$.

There are other ways of grouping terms or assigning parameters based on the needs. For example, if we do not want to penalize the linear terms in the model, we can set λ_{π} to be zero. Or, in case all the two-factor interactions are treated equally, we may let $\lambda_{\pi\pi} = \lambda_{\pi s} = \lambda_{ss}$. In practice, the researchers' prior knowledge, experience or preference may help to choose the way of setting up the parameters in the model subjectively.

Chapter 3

Adaptive Choice of the Regularization Parameters

The λ's in the LBP models are called regularization parameters, tuning parameters or smoothing parameters in the context of smoothing models. Different values of λ give different models. When λ is small, the function estimate tends to interpolate the data and has a small bias but large variance. As λ increases, the solution becomes more sparse due to the larger penalty, thus the estimate has a small variance but large bias. λ controls the tradeoff between the goodness-of-fit and the sparsity of the solution. A proper choice of λ is desired.

Regularization parameter selection has been a very active research field, appearing in various contexts of penalized likelihood methods and other nonparametric methods. An automated data-driven approach for parameter selection is highly desirable. For the smoothing splines with Gaussian data, ordinary cross validation (OCV) was originally proposed by Wahba & Wold (1975) . Craven & Wahba (1979) proposed the generalized cross validation

(GCV) which has been widely used. Later for the smoothing splines with non-Gaussian data, Xiang & Wahba (1996) proposed the generalized approximate cross validation (GACV) as an extension of GCV.

In this chapter, we derive the GACV as a tuning criterion of λ for the LBP models. We focus on the main effects model, and the ideas are easily applied to the two-factor interaction model and more complicated models. With an abuse of notation, we use λ to represent the collective set of tuning parameters. In particular, $\lambda = (\lambda_\pi, \lambda_s)$ for the main effects model and $\lambda = (\lambda_\pi, \lambda_{\pi\pi}, \lambda_s, \lambda_{\pi s}, \lambda_{ss})$ for the two-factor interaction model.

3.1 Comparative Kullback-Liebler Distance

Let us consider the random pair $(Y, \mathbf{X})$, where $\mathbf{X} = (X_1, \ldots, X_D)$ is the covariate vector. We are interested in estimating the conditional probability $\text{prob}(Y = y|\mathbf{X})$. Let $p(y|\mathbf{X})$ be the "true" but unknown conditional probability function and $p_\lambda(y|\mathbf{X})$ be its estimate associated with λ. Respectively $f(y|\mathbf{X})$ and $f_\lambda(y|\mathbf{X})$ are the true logit and its estimate. Let $\mu(\mathbf{X}) = E(Y|\mathbf{X})$, $\sigma^2(\mathbf{X}) = Var(Y|\mathbf{X})$ and $\mu_\lambda(\mathbf{X}), \sigma^2_\lambda(\mathbf{X})$ be their estimates.

Kullback-Liebler distance, also known as the relative entropy, is often used to measure the distance between two probability distributions. Conditioning on the value of $\mathbf{X}$, the Kullback-Liebler distance between the true probability function $p(y|\mathbf{X})$ and its estimate $p_\lambda(y|\mathbf{X})$ is defined as:

$$KL(p, p_\lambda|\mathbf{X}) \equiv E_p \; log(\frac{p(y|\mathbf{X})}{p_\lambda(y|\mathbf{X})}), \tag{3.1}$$

where E_p denotes the expectation under the true conditional probability $p(y|\mathbf{X})$. Note the Kullback-Liebler distance is not a true metric because it is not symmetric and does not satisfy the triangle inequality. However we do have $KL(p, p_\lambda) \geq 0$ with equality if and only if $p = p_\lambda$.

For Bernoulli outcomes,

$$\begin{aligned} KL(p, p_\lambda(y|\mathbf{X}) &= \frac{1}{2}[\ \dot{b}(f(\mathbf{X}))(f(\mathbf{X}) - f_\lambda(\mathbf{X})) - (b(f(\mathbf{X})) - b(f_\lambda(\mathbf{X})))\ ] \\ &= \frac{1}{2}[\ \mu(\mathbf{X})(f(\mathbf{X}) - f_\lambda(\mathbf{X})) - (b(f(\mathbf{X})) - b(f_\lambda(\mathbf{X})))\ ]. \end{aligned}$$

The comparative Kullback-Liebler distance is defined by

$$\begin{aligned} CKL(p, p_\lambda|\mathbf{X}) &\equiv KL(p, p_\lambda|\mathbf{X}) - E_p \; log(p(y|\mathbf{X})) \\ &= -E_p \; log(p_\lambda(y|\mathbf{X})). \end{aligned} \tag{3.2}$$

The *CKL* distance differs from the *KL* distance in (3.1) by a quantity which does not depend on the estimator, or the λ. The *CKL* distance can be regarded as the expected negative log-likelihood based on the estimated density function. To minimize the *CKL* distance is equivalent to maximize the expected log-likelihood for the future observations. The objective function desired to be minimized should be the expectation of $CKL(p, p\lambda|\mathbf{X})$ with respect to $\mathbf{X}$

$$E(CKL(p, p_\lambda|\mathbf{X})) = -E(E_p \; log(p_\lambda(y|\mathbf{X})|\mathbf{X}))).$$

For Bernoulli data, we have

$$
\begin{aligned}
CKL(p, p_\lambda|\mathbf{X}) &= -\mu(\mathbf{X}) f_\lambda(\mathbf{X}) + b(f_\lambda(\mathbf{X})) \\
E(CKL(p, p_\lambda|\mathbf{X})) &= E(-\mu(\mathbf{X}) f_\lambda(\mathbf{X}) + b(f_\lambda(\mathbf{X}))|\mathbf{X}\ ). \qquad (3.4)
\end{aligned}
$$

Suppose there are n pairs of observations $(y_i, \mathbf{x}_i), i = 1, \ldots, n$. For the functions $f(\mathbf{x})$ and $\mu(\mathbf{x})$, we define $f_i = f(\mathbf{x}_i)$ and $\mu_i = \mu(\mathbf{x}_i)$ for $i = 1, \ldots, n$. Correspondingly let $f_{\lambda i} = f_\lambda(\mathbf{x}_i)$ and $\mu_{\lambda i} = \mu_\lambda(\mathbf{x}_i)$. Then a consistent estimate of the quantity in (3.4) is

$$
\begin{aligned}
CKL(p, p_\lambda) &= \frac{1}{n} \sum_{i=1}^{n} [\ -\mu_i f_{\lambda i} + b(f_{\lambda i})\ ] \\
&= \frac{1}{n} \sum_{i=1}^{n} [\ -\mu_i f_{\lambda i} + log(1 + e^{f_{\lambda i}})\ ]. \qquad (3.5)
\end{aligned}
$$

Since the mean function μ is usually unknown, the expression in (3.5) is not directly computable. One approximating method is to replace μ_i with y_i for $i = 1, \ldots, n$, and calculate

$$
OBS(\lambda) = \frac{1}{n} \sum_{i=1}^{n} [\ -y_i f_{\lambda i} + b(f_{\lambda i})\ ],
$$

which is the observed negative log-likelihood. However, because y_i and $f_{\lambda i},\ i = 1, \ldots, n,$ are usually positively correlated, the $OBS(\lambda)$ tends to underestimate the CKL. To correct this bias, we use the leave-out-one cross validation

$$
CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} [\ -y_i {f_{\lambda i}}^{[-i]} + b(f_{\lambda i})\ ], \qquad (3.7)
$$

where ${f_{\lambda i}}^{[-i]}$ is the estimate of f with the ith data point omitted. Since ${f_{\lambda i}}^{[-i]}$ only depends on the observations other than y_i, it is independent of y_i and expected to be close to $f_{\lambda i}$ for a large n. In other words, $E(y_i {f_{\lambda i}}^{[-i]}) = E(y_i)E({f_{\lambda i}}^{[-i]}) \approx \mu_i E(f_{\lambda i})$. Therefore $CV(\lambda)$ can be expected to be at least roughly unbiased for the quantity in (3.5) and a computable proxy for the CKL distance.

3.2 Leaving-out-one Lemma for LBP

For any fixed λ, direct calculation of $CV(\lambda)$ involves computing n leaving-out-one estimates $\{f_{\lambda i}^{[-i]}, i = 1, \ldots, n\}$, which is expensive and almost infeasible for large-scale problems. Thus it is desired to derive the approximate $CV(\lambda)$. The following leaving-out-one lemma is important for deriving a second-order approximation to $CV(\lambda)$. Xiang (1996) generalized the leaving-out-one lemma of Craven & Wahba (1979) for the smoothing spline models. But in his lemma, the penalty part $J(\lambda)$ is a quadratic penalty functional instead of the non-differentiable l_1 penalty function as in the LBP model. Thus we have to generalize the theorem of the quardratic case.

For Bernoulli data, we define $-l(y_i, f(\mathbf{x}_i)) = -y_i f(\mathbf{x}_i) + b(f(\mathbf{x}_i))$ for $i = 1, \ldots, n$. The objective function in the LBP model is expressed as

$$I_\lambda(f, y) = \sum_{i=1}^{n} \left[-l(y_i, f(\mathbf{x}_i)) \right] + J_\lambda(f), \tag{3.8}$$

where the penalty function $J_\lambda(f)$ is the l_1 norm of the coefficients in the decompositon of f.

Lemma 1: (Leaving-out-one Lemma for LBP)

Denote $I_\lambda(f, y) = \sum_{j=1}^{n} \, [-l(y_j, f(\mathbf{x}_j)) \,] + J_\lambda(f)$. Let $\mu_\lambda^{[-i]}(\cdot)$ be the mean functions corresponding to $f_\lambda^{[-i]}(\cdot)$. Define $\mathbf{V} = (y_1, \ldots, y_{i-1}, v, y_{i+1}, \ldots, y_n)$. Suppose $h_\lambda(i, v, \cdot)$ is the minimizer of $I_\lambda(f, \mathbf{V})$, then

$$h_\lambda(i, \mu_\lambda^{[-i]}(\mathbf{x}_i), \cdot) = f_\lambda^{[-i]}(\cdot).$$

Proof:

For $i = 1, \ldots, n$, we have

$$-l(\mu_\lambda^{[-i]}(\mathbf{x}_i), \tau) = -\mu_\lambda^{[-i]}(\mathbf{x}_i)\tau + b(\tau),$$

and $f_\lambda{}^{[-i]}$ minimizes the objective function

$$\sum_{j \neq i} \, [-l(y_j, f(\mathbf{x}_j)) \,] + J_\lambda(f). \tag{3.10}$$

Since

$$\frac{\partial(-l(\mu_\lambda^{[-i]}(\mathbf{x}_i), \tau))}{\partial \tau} = -\mu_\lambda^{[-i]}(\mathbf{x}_i) + b'(\tau)$$

and

$$\frac{\partial^2(-l(\mu_\lambda^{[-i]}(\mathbf{x}_i, \tau))}{\partial^2 \tau} = b''(\tau) > 0,$$

we see that $-l(\mu_\lambda^{[-i]}(\mathbf{x}_i), \tau)$ achieves its unique minimum at $\hat{\tau}$ that satisfies $b'(\tau) = \mu_\lambda^{[-i]}(\mathbf{x}_i)$. So $\hat{\tau} = f_\lambda^{[-i]}(\mathbf{x}_i)$. Then for any f, we have

$$-l(\mu_\lambda^{[-i]}(\mathbf{x}_i), f_\lambda^{[-i]}(\mathbf{x}_i)) \leq -l(\mu_\lambda^{[-i]}(\mathbf{x}_i), f(\mathbf{x}_i)). \tag{3.13}$$

Define $\mathbf{y}^{-i} = (y_1, \ldots, y_{i-1}, \mu_\lambda^{[-i]}(\mathbf{x}_i), y_{i+1}, \ldots, y_n)$. For any f,

$$\begin{aligned} I_\lambda(f, \mathbf{y}^{-i}) &= -l(\mu_\lambda^{[-i]}(\mathbf{x}_i), f(\mathbf{x}_i) + \sum_{j \neq i} [-l(y_j, f(\mathbf{x}_j))\ ] + J_\lambda(f) \\ &\geq -l(\mu_\lambda^{[-i]}(\mathbf{x}_i), f_\lambda^{[-i]}(\mathbf{x}_i)) + \sum_{j \neq i} [-l(y_j, f(\mathbf{x}_j))\ ] + J_\lambda(f) \\ &\geq -l(\mu_\lambda^{[-i]}(x_i), f_\lambda^{[-i]}(\mathbf{x}_i)) + \sum_{j \neq i} [-l(y_j, f_\lambda^{[-i]}(\mathbf{x}_j))\ ] + J_\lambda(f_\lambda^{[-i]}). \end{aligned}$$

The first inequality comes from (3.13). The second inequality is due to the fact that $f_\lambda^{[-i]}(\cdot)$ is the minimizer of (3.10). Thus we have

$$h_\lambda(i, \mu_\lambda^{[-i]}(\mathbf{x}_i), \cdot) = f_\lambda^{[-i]}(\cdot). \qquad \blacksquare$$

This lemma says if we replace the ith leaving out observation y_i by $\mu_\lambda^{[-i]}(x_i)$, the minimizer of I_λ with respect to f is $f_\lambda^{[-i]}(\cdot)$. This nice result will play an important role in deriving the approximate cross validation score.

3.3 Generalized Approximate Cross Validation

In this section, we derive the generalized approximate cross validation ($GACV$) for the LBP models to choose the regularization parameter λ adaptively.

Firstly, we reformulate $CV(\lambda)$ defined in (3.7) as follows.

$$
\begin{aligned}
CV(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}\left[\ -y_i {f_{\lambda i}}^{[-i]} + b(f_{\lambda i})\ \right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\ (-y_i f_{\lambda i} + b(f_{\lambda i})) + y_i(f_{\lambda i} - {f_{\lambda i}}^{[-i]})\ \right] \\
&= OBS(\lambda) + \frac{1}{n}\sum_{i=1}^{n}\ y_i(f_{\lambda i} - {f_{\lambda i}}^{[-i]}) \\
&= OBS(\lambda) + \frac{1}{n}\sum_{i=1}^{n}\ y_i\ \frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda_i}^{[-i]}}\ (y_i - \mu_{\lambda_i}^{[-i]}) \\
&= OBS(\lambda) + \frac{1}{n}\sum_{i=1}^{n}\ y_i\ \frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda_i}^{[-i]}}\ \frac{y_i - \mu_{\lambda i}}{\frac{y_i - \mu_{\lambda i}}{y_i - \mu_{\lambda i}^{[-i]}}} \\
&= OBS(\lambda) + \frac{1}{n}\sum_{i=1}^{n}\ y_i\ \frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda_i}^{[-i]}}\ \frac{y_i - \mu_{\lambda i}}{1 - \frac{\mu_{\lambda i} - \mu_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda i}^{[-i]}}}. \qquad (3.15)
\end{aligned}
$$

We denote the quantities in the second term of (3.15) by

$$
G_1 \equiv \frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda_i}^{[-i]}}, \qquad (3.16)
$$

$$
G_2 \equiv \frac{\mu_{\lambda i} - \mu_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda i}^{[-i]}}. \qquad (3.17)
$$

Note $\mu_{\lambda i} = \dot{b}(f_{\lambda i})$. Using an approximation to the finite difference expression in (3.16), we find an approximate linear relation between G_1 and G_2:

$$
\begin{aligned}
G_2 &= \frac{\mu_{\lambda i} - \mu_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda i}^{[-i]}} \\
&= \frac{\dot{b}(f_{\lambda i}) - \dot{b}(f_{\lambda i}^{[-i]})}{y_i - \mu_{\lambda i}^{[-i]}} \\
&\approx \ddot{b}(f_{\lambda i})\ \frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda_i}^{[-i]}} \\
&= \ddot{b}(f_{\lambda i}) \cdot G_1. \qquad (3.18)
\end{aligned}
$$

For Bernoulli case, $\ddot{b}(f_{\lambda i}) = \sigma^2_{\lambda i} = p_\lambda(\mathbf{x}_i)\ (1 - p_\lambda(\mathbf{x}_i))$. In the equation (3.15) substituting G_2 by (3.18), we get

$$\begin{aligned} CV(\lambda) &\approx OBS(\lambda) + \frac{1}{n}\sum_{i=1}^{n} y_i\ G_1 \cdot \frac{y_i - \mu_{\lambda i}}{1 - \ddot{b}(f_{\lambda i}) \cdot G_1} && (3.19) \\ &= OBS(\lambda) + \frac{1}{n}\sum_{i=1}^{n} \frac{y_i(y_i - \mu_{\lambda i})}{\frac{1}{G_1} - \ddot{b}(f_{\lambda i})}. && (3.20) \end{aligned}$$

Now define a perturbation vector of length n

$$\epsilon_0 = (0, \ldots, \mu^{[-i]}_{\lambda i} - y_i, \ldots, 0)^T, \qquad (3.21)$$

and $\varepsilon_0 = \mu^{[-i]}_{\lambda i} - y_i$. Lemma 1 shows that $f^{[-i]}_{\lambda}$ is the minimizer of (3.8) with y_i being replaced by $\mu^{[-i]}_{\lambda i}$. Using this fact, we may express the expression G_1 as

$$\begin{aligned} G_1 &= \frac{f_{\lambda i} - f^{[-i]}_{\lambda i}}{y_i - \mu^{[-i]}_{\lambda_i}} \\ &= \frac{f^{[-i]}_{\lambda i} - f_{\lambda i}}{\varepsilon_0}. && (3.22) \end{aligned}$$

3.3.1 Some Notations

Let $m = ND$. The superscript t of a matrix denotes the transpose of the matrix. For independent observations $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$, define the vectors

$$\mathbf{y} = (\ y_1, \ldots, y_n\ )^t,$$
$$\mathbf{f} = (\ f_1, \ldots, f_n\ )^t.$$

The coefficients in the LBP main effects model (2.12) are denoted by

$$\mathbf{d} = (\ b_0, b_1, \ldots, b_D\ )^t,$$

$$\mathbf{c} = (\ c_{1,1}, \ldots, c_{D,N}\ )^t = (\ c_1, \ldots, c_m\ )^t.$$

Furthermore, we define the matrices

$$K_\alpha = \left(\ K(x_i^\alpha, x_{j*}^\alpha)\ \right)_{n\times N} \quad \text{for } \alpha = 1, \ldots, D,$$

$$K = (\ K_1, \ldots, K_D\ )_{n\times m}$$

$$T = \begin{bmatrix} 1 & k_1(x_1^1) & \cdots & k_1(x_1^d) \\ & \cdots & & \cdots \\ 1 & k_1(x_n^1) & \cdots & k_1(x_n^d) \end{bmatrix}.$$

Then the quantities in (2.10) and (2.11) can be expressed in terms of the matrices T and K:

$$\begin{aligned} f_i &= \sum_{\alpha=1}^{D+1} T_{i\alpha} d_\alpha + \sum_{j=1}^{m} K_{ij} c_j, \quad \text{for} \quad i = 1, \ldots, n, \\ \mathbf{f} &= T\,\mathbf{d} + K\,\mathbf{c}. \end{aligned} \tag{3.23}$$

Since $\mathbf{f}$ is linear in $\mathbf{d}$ and $\mathbf{c}$, we can express the objective function (3.8) in terms of $\mathbf{d}$ and $\mathbf{c}$

$$\begin{aligned} I_\lambda(\ \mathbf{d}, \mathbf{c}, \mathbf{y}) &= \frac{1}{n} \sum_{i=1}^{n} [-y_i\ (\ \sum_{\alpha=1}^{D+1} T_{i\alpha} d_\alpha + \sum_{j=1}^{m} K_{ij} c_j\ ) \\ &\quad + log(\ 1 + \exp\ (\ \sum_{\alpha=1}^{D+1} T_{i\alpha} d_\alpha + \sum_{j=1}^{m} K_{ij} c_j)\ )\ ] \\ &\quad + \lambda_d \sum_{\alpha=2}^{D+1} |d_\alpha| + \lambda_c \sum_{j=1}^{m} |c_j| \end{aligned} \tag{3.24}$$

Fix λ, let $(\mathbf{d}_\lambda^\mathbf{y}, \mathbf{c}_\lambda^\mathbf{y}) = (d_1, d_2, \ldots, d_{D+1}, c_1, \ldots, c_m)^t$ be the minimizer of (3.24), with the original data vector $\mathbf{y}$ being used for fitting the model. It is well known that the l_1 penalty in (3.24) tends to produce sparse solutions, or, many components of $\mathbf{d}$ and $\mathbf{c}$ will be exactly zero in the solution.

3.3.2 Robustness Assumption

Without loss of generality, we let only the first s components of $\mathbf{d}_\lambda^\mathbf{y}$ and the first r components of $\mathbf{c}_\lambda^\mathbf{y}$ are nonzero elements

$$\begin{aligned}
\mathbf{d}_\lambda^\mathbf{y} &= (\underbrace{d_1,\ d_2, \ldots,\ d_s}_{\neq 0},\ \underbrace{d_{s+1},\ d_{s+2}, \ldots,\ d_{D+1}}_{=0})^t, \\
\mathbf{c}_\lambda^\mathbf{y} &= (\underbrace{c_1,\ \ldots,\ c_r}_{\neq 0},\ \underbrace{c_{r+1},\ c_{r+2}, \ldots,\ c_m}_{=0})^t. \qquad (3.25)
\end{aligned}$$

For any random vector ε, let $(\mathbf{d}_\lambda^{\mathbf{y}+\varepsilon}, \mathbf{c}_\lambda^{\mathbf{y}+\varepsilon}) = (d_1^*, d_2^*, \ldots, d_{D+1}^*, c_1^*, \ldots, , c_m^*)^t$ be the minimizer of $I_\lambda(\mathbf{d}, \mathbf{c}, \mathbf{y}+\varepsilon)$, where $\mathbf{y}+\varepsilon$ is the perturbed data. Many components of $\mathbf{d}_\lambda^{\mathbf{y}+\varepsilon}$ and $\mathbf{c}_\lambda^{\mathbf{y}+\varepsilon}$ are zeros due to the shrinkage property of the l_1 penalty.

We assume the zeros in the solutions are robust against small perturbation in data. In other words, when the perturbation ε is small enough, all zero components will stay at zero. Intuitively speaking, if the zero components are very sensitive to a small perturbation, then it is extremely hard to get so many zeros in the solution. When the objective function has a simple mathematical form such as a quadratic function, we can prove the robustness property mathematically. For a complicated objective function like in (3.15), some numerical simulations can be used to show the robustness of zero components.

With this assumption, for any small ε, the perturbed solution vector $(\mathbf{d}_\lambda^{\mathbf{y}+\varepsilon}, \mathbf{c}_\lambda^{\mathbf{y}+\varepsilon})$ has the form

$$\begin{aligned} \mathbf{d}_\lambda^{\mathbf{y}+\varepsilon} &= (\underbrace{d_1^*,\ d_2^*, \ldots,\ d_s^*}_{\neq 0},\ \underbrace{d_{s+1}^*,\ d_{s+2}^*, \ldots,\ d_{D+1}^*}_{=0})^t, \\ \mathbf{c}_\lambda^{\mathbf{y}+\varepsilon} &= (\underbrace{c_1^*,\ \ldots,\ c_r^*}_{\neq 0},\ \underbrace{c_{r+1}^*,\ c_{r+2}^*, \ldots,\ c_m^*}_{=0})^t. \end{aligned} \tag{3.26}$$

By comparing the expressions in (3.25) and (3.26), we have

$$\begin{aligned} d_{s+j} &= d_{s+j}^* = 0, \quad \text{for} \quad j = 1, \ldots, D+1-s \\ c_{r+j} &= c_{r+j}^* = 0, \quad \text{for} \quad j = 1, \ldots, m-r. \end{aligned} \tag{3.27}$$

For convenience, we denote the nonzero segments of $(\mathbf{d}, \mathbf{c})$ as

$$\begin{aligned} \tilde{\mathbf{d}} &\equiv (d_1,\ \ldots, d_s)^t, \\ \tilde{\mathbf{d}}^* &\equiv (d_1^*,\ \ldots, d_s^*)^t, \\ \tilde{\mathbf{c}} &\equiv (c_1,\ \ldots, c_r)^t, \\ \tilde{\mathbf{c}}^* &\equiv (c_1^*,\ \ldots, c_r^*)^t. \end{aligned}$$

3.3.3 Derivation of GACV

Denote the submatrix consisting of the first r columns of K as $K_{n\times r}$ and the submatrix consisting of the first s columns of T as $T_{n\times s}$. Using the expressions in (3.23) and (3.27), we get

$$\begin{aligned} \mathbf{f}_\lambda^{\mathbf{y}+\epsilon} - \mathbf{f}_\lambda^{\mathbf{y}} &= T_{n\times(D+1)}\,(\mathbf{d}_\lambda^{\mathbf{y}+\epsilon} - \mathbf{d}_\lambda^{\mathbf{y}}) + K_{n\times m}\,(\mathbf{c}_\lambda^{\mathbf{y}+\epsilon} - \mathbf{c}_\lambda^{\mathbf{y}}) \\ &= [T_{n\times s}, K_{n\times r}] \begin{bmatrix} \tilde{\mathbf{d}} - \tilde{\mathbf{d}}^* \\ \tilde{\mathbf{c}} - \tilde{\mathbf{c}}^* \end{bmatrix} \end{aligned} \tag{3.28}$$

Since $I_\lambda(\mathbf{d}, \mathbf{c}, \mathbf{y})$ is linear in $\mathbf{y}$, it is not hard to see that the minimizer of (3.24) is continuous in y_i, $i = 1, \ldots, n$, irrespective of whether or not realizations of y_i are continuous. Furthermore, though $I_\lambda(\mathbf{d}, \mathbf{c}, \mathbf{y})$ is not differentiable at zero, its marginal derivatives at $(\tilde{\mathbf{d}}, \tilde{\mathbf{c}})$ exist

$$\begin{bmatrix} \dfrac{\partial I_\lambda}{\partial \tilde{\mathbf{d}}} \\ \dfrac{\partial I_\lambda}{\partial \tilde{\mathbf{c}}} \end{bmatrix}_{\left(\mathbf{d}_\lambda^{\mathbf{y}+\epsilon}, \mathbf{c}_\lambda^{\mathbf{y}+\epsilon}, \mathbf{y}+\epsilon\right)} = 0,$$

$$\begin{bmatrix} \dfrac{\partial I_\lambda}{\partial \tilde{\mathbf{d}}} \\ \dfrac{\partial I_\lambda}{\partial \tilde{\mathbf{c}}} \end{bmatrix}_{\left(\mathbf{d}_\lambda^{\mathbf{y}}, \mathbf{c}_\lambda^{\mathbf{y}}, \mathbf{y}\right)} = 0. \tag{3.30}$$

Using the first order Taylor series approximation, we expand $[\,\dfrac{\partial I_\lambda}{\partial \tilde{\mathbf{d}}}, \dfrac{\partial I_\lambda}{\partial \tilde{\mathbf{c}}}\,]^t_{(\mathbf{d}_\lambda^{\mathbf{y}+\epsilon}, \mathbf{c}_\lambda^{\mathbf{y}+\epsilon}, \mathbf{y}+\epsilon)}$ at $(\mathbf{d}_\lambda^{\mathbf{y}}, \mathbf{c}_\lambda^{\mathbf{y}}, \mathbf{y})$ and get

$$\begin{aligned} \begin{bmatrix} \dfrac{\partial I_\lambda}{\partial \tilde{\mathbf{d}}} \\ \dfrac{\partial I_\lambda}{\partial \tilde{\mathbf{c}}} \end{bmatrix}_{\left(\mathbf{d}_\lambda^{\mathbf{y}+\epsilon}, \mathbf{c}_\lambda^{\mathbf{y}+\epsilon}, \mathbf{y}+\epsilon\right)} &\approx \begin{bmatrix} \dfrac{\partial I_\lambda}{\partial \tilde{\mathbf{d}}} \\ \dfrac{\partial I_\lambda}{\partial \tilde{\mathbf{c}}} \end{bmatrix}_{\left(\mathbf{d}_\lambda^{\mathbf{y}}, \mathbf{c}_\lambda^{\mathbf{y}}, \mathbf{y}\right)} + \begin{bmatrix} \dfrac{\partial^2 I_\lambda}{\partial \tilde{\mathbf{d}} \partial \tilde{\mathbf{d}}^T} & \dfrac{\partial^2 I_\lambda}{\partial \tilde{\mathbf{d}} \partial \tilde{\mathbf{c}}^T} \\ \dfrac{\partial^2 I_\lambda}{\partial \tilde{\mathbf{c}} \partial \tilde{\mathbf{d}}^T} & \dfrac{\partial^2 I_\lambda}{\partial \tilde{\mathbf{c}} \partial \tilde{\mathbf{c}}^T} \end{bmatrix}_{(d',c',y')} \begin{bmatrix} \tilde{\mathbf{d}}^* - \tilde{\mathbf{d}} \\ \tilde{\mathbf{c}}^* - \tilde{\mathbf{c}} \end{bmatrix} \\ &+ \begin{bmatrix} \dfrac{\partial^2 I_\lambda}{\partial \tilde{\mathbf{d}} \partial y^T} \\ \dfrac{\partial^2 I_\lambda}{\partial \tilde{\mathbf{c}} \partial y^T} \end{bmatrix}_{(d',c',y')} (y + \epsilon - y)\,, \end{aligned}$$

where $(\mathbf{d}', \mathbf{c}', \mathbf{y}')$ is an intermediate vector. The equations in (3.30) imply that

$$\begin{bmatrix} \frac{\partial^2 I_\lambda}{\partial \tilde{\mathbf{d}} \partial \tilde{\mathbf{d}}^T} & \frac{\partial^2 I_\lambda}{\partial \tilde{\mathbf{d}} \partial \tilde{\mathbf{c}}^T} \\ \frac{\partial^2 I_\lambda}{\partial \tilde{\mathbf{c}} \partial \tilde{\mathbf{d}}^T} & \frac{\partial^2 I_\lambda}{\partial \tilde{\mathbf{c}} \partial \tilde{\mathbf{c}}^T} \end{bmatrix}_{(d',c',y')} \begin{pmatrix} \tilde{\mathbf{d}}^* - \tilde{\mathbf{d}} \\ \tilde{\mathbf{c}}^* - \tilde{\mathbf{c}} \end{pmatrix} + \begin{pmatrix} \frac{\partial^2 I_\lambda}{\partial \tilde{\mathbf{d}} \partial y^T} \\ \frac{\partial^2 I_\lambda}{\partial \tilde{\mathbf{c}} \partial y^T} \end{pmatrix}_{(\mathbf{d}',\mathbf{c}',\mathbf{y}')} (\mathbf{y} + \epsilon - \mathbf{y}) = 0.$$

Define the diagonal matrix

$$W(\mathbf{f}) = \ diag\ [\ \sigma_1^2, \ldots, \sigma_n^2\ ], \tag{3.32}$$

and

$$\begin{aligned} U_{(s+r)\times(s+r)} \quad &\equiv \quad \begin{pmatrix} \frac{\partial^2 I_\lambda}{\partial \tilde{\mathbf{d}} \partial \tilde{\mathbf{d}}^T} & \frac{\partial^2 I_\lambda}{\partial \tilde{\mathbf{d}} \partial \tilde{\mathbf{c}}^T} \\ \frac{\partial^2 I_\lambda}{\partial \tilde{\mathbf{c}} \partial \tilde{\mathbf{d}}^T} & \frac{\partial^2 I_\lambda}{\partial \tilde{\mathbf{c}} \partial \tilde{\mathbf{c}}^T} \end{pmatrix} \\ &= \quad \begin{pmatrix} T_{n\times s}^T \\ K_{n\times r}^T \end{pmatrix} W\left(T_{n\times s}, K_{n\times r}\right), \end{aligned} \tag{3.33}$$

$$-V_{r\times n} \equiv \begin{pmatrix} \frac{\partial^2 I_\lambda}{\partial \tilde{\mathbf{d}} \partial \mathbf{y}^T} \\ \frac{\partial^2 I_\lambda}{\partial \tilde{\mathbf{c}} \partial \mathbf{y}^T} \end{pmatrix}. \tag{3.34}$$

Then we have

$$U_{(\mathbf{d}',\mathbf{c}',\mathbf{y}')} \begin{pmatrix} \tilde{\mathbf{d}}^* - \tilde{\mathbf{d}} \\ \tilde{\mathbf{c}}^* - \tilde{\mathbf{c}} \end{pmatrix} = V_{(\mathbf{d}',\mathbf{c}',\mathbf{y}')}\ \epsilon.$$

When ϵ is small, $(\mathbf{d}', \mathbf{c}', \mathbf{y}')$ is very close to $(\mathbf{d}, \mathbf{c}, \mathbf{y})$, which gives the approximation

$$U_{(\mathbf{d},\mathbf{c},\mathbf{y})} \begin{pmatrix} \tilde{\mathbf{d}}^* - \tilde{\mathbf{d}} \\ \tilde{\mathbf{c}}^* - \tilde{\mathbf{c}} \end{pmatrix} \approx V_{(\mathbf{d},\mathbf{c},\mathbf{y})}\ \epsilon.$$

Now we show that U is a full-rank matrix with $rank(U) = s + r$. Note

$$\begin{aligned} U &= [\ \begin{pmatrix} T^T_{n\times s} \\ K^T_{n\times r} \end{pmatrix} W^{\frac{1}{2}}\ ]\ [\ W^{\frac{1}{2}}\,(T_{n\times s}, K_{n\times r})\ \ ] \\ &= [\ W^{\frac{1}{2}} \begin{pmatrix} T^T_{n\times s} \\ K^T_{n\times r} \end{pmatrix}]^T\ [\ W^{\frac{1}{2}} \begin{pmatrix} T^T_{n\times s} \\ K^T_{n\times r} \end{pmatrix}\ ], \end{aligned} \tag{3.37}$$

and the robustness assumption implies that $(T_{n\times s}, K_{n\times r})$ has a full column rank $s + r$. Since $W^{\frac{1}{2}}$ has full rank n, the matrix in the first bracket in (3.37) has a full column rank $s + r$; so does the prodcuct matrix obtained by multiplying its transpose. Thus U is invertible. Using the expression in (3.28), we estimate the quantity $\mathbf{f}^{\mathbf{y}+\epsilon}_\lambda - \mathbf{f}^{\mathbf{y}}_\lambda$ by

$$\mathbf{f}^{\mathbf{y}+\epsilon}_\lambda - \mathbf{f}^{\mathbf{y}}_\lambda\ \approx (T_{n\times s}, K_{n\times r})\ \ U^{-1}\ V\ \epsilon\ \equiv H\epsilon, \tag{3.38}$$

where

$$H_{n\times n}\ \equiv\ \ (T_{n\times s}, K_{n\times r})\ \ U^{-1}\ V. \tag{3.39}$$

Denote the diagonal element of H by h_{ii}. Using the special perturbation vector ε_0 defined in (3.21), we finally get the estimate for G_1

$$G_1 = \frac{f_{\lambda i} - f^{[-i]}_{\lambda i}}{y_i - \mu^{[-i]}_{\lambda_i}} \approx h_{ii}.$$

Then the approximate cross validation $ACV(\lambda)$ is given by

$$\begin{aligned} CV(\lambda) &\approx ACV(\lambda) \\ &= \frac{1}{n}\sum_{i=1}^{n}\ [-y_i f_{\lambda i} + b(f_{\lambda i})\ ]\ +\ \frac{1}{n}\sum_{i=1}^{n}\ \frac{y_i(y_i-\mu_{\lambda i})}{\frac{1}{h_{ii}} - \ddot{b}(f_{\lambda i})} \\ &= \frac{1}{n}\sum_{i=1}^{n}\ [-y_i f_{\lambda i} + b(f_{\lambda i})\ ]\ +\ \frac{1}{n}\sum_{i=1}^{n}\ h_{ii}\ \frac{y_i(y_i-\mu_{\lambda i})}{1-\ddot{b}(f_{\lambda i})\ h_{ii}} \\ &= \frac{1}{n}\sum_{i=1}^{n}\ [-y_i f_{\lambda i} + b(f_{\lambda i})\ ]\ +\ \frac{1}{n}\sum_{i=1}^{n}\ h_{ii}\ \frac{y_i(y_i-\mu_{\lambda i})}{1-\sigma^2_{\lambda i}\ h_{ii}}. \end{aligned} \quad (3.41)$$

By replacing h_{ii} with $\frac{1}{n}\sum_{i=1}^{n} h_{ii} \equiv \frac{1}{n}tr(H)$ and replacing $1-\sigma^2_{\lambda i}\ h_{ii}$ with the quantity $\frac{1}{n}tr[I-(W^{1/2}HW^{1/2})]$, we obtain the generalized approximate cross validation (GACV)

$$GACV = \frac{1}{n}\sum_{i=1}^{n}\ [-y_i f_{\lambda i} + b(f_{\lambda i})\ ] + \frac{tr(H)}{n}\frac{\sum_{i=1}^{n}\ y_i(y_i-\mu_{\lambda i})}{tr\ [\ I\ -\ W^{1/2}HW^{1/2}\ ]}. \quad (3.42)$$

3.4 Randomized GACV

Direct computation of (3.42) involves the inversion of a large-scale matrix, whose size depends on the sample size n, basis size N and dimension d. Large N, n or d may make the computation expensive and produce unstable solutions. Thus the randomized GACV (ranGACV) is proposed as a computable proxy for GACV. We use the randomized trace estimates for $tr(H)$ and $tr[I\ -\ \frac{1}{2}(W^{1/2}HW^{1/2})]$ based on the following theorem:

If A is any square matrix and ϵ is a zero mean random n-vector with independent components with variance σ^2_ϵ, then $\frac{1}{\sigma^2_\epsilon}E\epsilon^T A\epsilon = tr(A)$.

Let $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$ be a zero mean random n-vector of independent components with variance σ_ϵ^2. Let $\mathbf{f}_\lambda^{\mathbf{y}}$ and $\mathbf{f}_\lambda^{\mathbf{y}+\epsilon}$ respectively be the minimizer of (2.12) fitted with the original data vector $\mathbf{y}$ and the perturbed data vector $\mathbf{y}+\epsilon$. Then $ranGACV(\lambda)$ is given by

$$ranGACV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} [-y_i f_\lambda(\mathbf{x}_i) + b(f_\lambda(\mathbf{x}_i))] \tag{3.43}$$

$$+ \frac{\epsilon^T(\mathbf{f}_\lambda^{\mathbf{y}+\epsilon} - \mathbf{f}_\lambda^{\mathbf{y}})}{n} \frac{\sum_{i=1}^{n} y_i(y_i - \mu_{\lambda i})}{\epsilon^T\epsilon - \epsilon^T W(\mathbf{f}_\lambda^{\mathbf{y}+\epsilon} - \mathbf{f}_\lambda^{\mathbf{y}})}. \tag{3.44}$$

Its derivation is given in Lin et al. (2000) . In addition, two facts may help to reduce the variance of the second term in (3.44). (1) It is shown in Hutchinson (1989) that given the variance σ_ϵ^2, when each component of ϵ has a Bernoulli(0.5) distribution taking values $\{+\sigma_\epsilon, -\sigma_\epsilon\}$, the randomized trace estimate for the trace of a matrix has the minimal variance. Thus the perturbation based on Bernoulli distribution is suggested. (2) Generate Z independent perturbations $\epsilon^{(z)}$, $z = 1, \dots, Z$, and compute Z-replicate ranGACVs. Their average has a smaller variance.

Chapter 4

Selection Criteria for Main Effects and Two-Factor Interactions

4.1 The L_1 Importance Measure

After choosing $\hat{\lambda}$ by the GACV or ranGACV criteria, the LBP estimate $f_{\hat{\lambda}}$ is obtained by minimizing (2.12), (2.15), (2.16) or (2.18). How to measure the importance of a particular component of the fitted model is a key question. We will consider the main effects and, possibly, the two factor interactions as the model components of interest.

We propose using the functional L_1 norm as the importance measure. This measure is especially useful when the optimization problem has sparse solutions, which is exactly the case of LBP. In practice, we calculate the empirical L_1 norm for each functional component, which is the average of the function values evaluated at all the data points.

- For the continuous variables in the model (2.12), the empirical L_1 norms of the main effect f_α and the two-factor interaction $f_{\alpha\beta},\ \alpha = 1, \ldots, D$, $\beta < \alpha$, are

$$
\begin{aligned}
L_1(f_\alpha) &= \frac{1}{n}\sum_{i=1}^{n} |f_\alpha(x_i^\alpha)| = \frac{1}{n}\sum_{i=1}^{n} |b_\alpha k_1(x_i^\alpha) + \sum_{j=1}^{N} c_{\alpha,j} K_1(x_i^\alpha, x_{j*}^\alpha)| \\
L_1(f_{\alpha\beta}) &= \frac{1}{n}\sum_{i=1}^{n} |f_{\alpha\beta}(x_i^\alpha, x_i^\beta)| \\
&= \frac{1}{n}\sum_{i=1}^{n} |b_{\alpha\beta} k_1(x_i^\alpha) k_1(x_i^\beta) + \sum_{j=1}^{N} c_{\alpha\beta,j}^{\pi s} K_1(x_i^\alpha, x_{j*}^\alpha) k_1(x_i^\beta) k_1(x_{j*}^\beta) \\
&\quad + \sum_{j=1}^{N} c_{\beta\alpha,j}^{\pi s} K_1(x_i^\beta, x_{j*}^\beta) k_1(x_i^\alpha) k_1(x_{j*}^\alpha) \\
&\quad + \sum_{j=1}^{N} c_{\alpha\beta,j}^{ss} K_1(x^\alpha, x_{j*}^\alpha) K_1(x^\beta, x_{j*}^\beta)|.
\end{aligned}
$$

- For the categorical variables in the model (2.16), the empirical L_1 norm of the main effect f_γ, $\gamma = 1, \ldots, r$, is:

$$
L_1(f_\gamma) = \frac{1}{n}\sum_{i=1}^{n} \mid B_\gamma \Phi_\gamma(z_i^\gamma) \mid = \frac{1}{n} |B_\gamma| \sum_{i=1}^{n} \mid \Phi_\gamma(z_i^\gamma) \mid
$$

  and the empirical L_1 norms for the interactions between categorical variables are defined similarly.

The rank of the L_1 norm scores indicates the relative importance of all the main effect terms and the interaction terms. For instance, the component with the largest L_1 norm is the most important, and any variable with near zero L_1 norm might be unimportant. An alternative measure is based on the functional L_2 norm. The empirical L_2 norms of the main effect f_α and the two-factor interaction $f_{\alpha\beta},\ \alpha = 1, \ldots, D,\ \beta < \alpha$, are

$$\begin{aligned}
L_2(f_\alpha) &= [\ \frac{1}{n}\sum_{i=1}^{n} (f_\alpha(x_i^\alpha))^2\ ]^{\frac{1}{2}} \\
&= [\ \frac{1}{n}\sum_{i=1}^{n} (\ b_\alpha k_1(x_i^\alpha) + \sum_{j=1}^{n} c_{\alpha,j} K(x_i^\alpha, x_{j*}^\alpha)\ )^2\ ]^{\frac{1}{2}} \\
L_2(f_{\alpha\beta}) &= [\ \frac{1}{n}\sum_{i=1}^{n} (\ f_{\alpha\beta}(x_i^\alpha, x_i^\beta))^2\ ]^{\frac{1}{2}} \\
L_2(f_\gamma) &= |B_\gamma|\ [\ \frac{1}{n}\sum_{i=1}^{n} (\ \Phi_\gamma(z_i^\gamma))^2\ ]^{\frac{1}{2}}
\end{aligned}$$

L_2 norm works equally well as L_1 norm in our simulation studies. But we omit further discussion of it.

4.2 Choosing the Threshold

We focus on the main effects model in this section. Using the chosen parameter $\hat{\lambda}$, we obtain the estimated main effect components $\hat{f}_1, \cdots, \hat{f}_D$ and calculate their L_1 norms $L_1(\hat{f}_1), \ldots L_1(\hat{f}_D)$. Denote the decreasingly ordered norms as $\hat{L}_{(1)}, \cdots, \hat{L}_{(D)}$ and the corresponding components $\hat{f}_{(1)}, \ldots \hat{f}_{(D)}$. A universal threshold value is needed to differentiate the important components from unimportant ones. Call the threshold q. Only variables with their L_1 norms greater than or equal to q are “important”.

Essentially we will test the variables’ importance one by one in their L_1 norm rank order. If one variable passes the test (hence “important”), it enters the null model for testing the next variable; otherwise the procedure stops. After the first η $(0 \leq \eta \leq D-1)$ variables enter the model, it is a one-sided hypothesis testing problem to decide whether the next component $\hat{f}_{(\eta+1)}$ is important or not. When $\eta = 0$, the null model f is the constant, say, $f = \hat{b}_0$, and the hypotheses are $H_0 : L_{(1)} = 0$ vs $H_1 : L_{(1)} > 0$. When $\eta \geq 1$, the null model is $f = \hat{b}_0 + \hat{f}_{(1)} + \cdots + \hat{f}_{(\eta)}$ and the hypotheses $H_0 : L_{(\eta+1)} = 0$ vs $H_1 : L_{(\eta+1)} > 0$. Let the desired one-sided test level be α. If the null distribution of $\hat{L}_{(\eta+1)}$ were known, we could get the critical value α-percentile and make a decision of rejection or acceptance. In practice the exact α-percentile is difficult or impossible to calculate. However the Monte Carlo bootstrap test provides a convenient approximation to the full test.

Now we develop a sequential Monte Carlo bootstrap test procedure to determine q. Conditional on the original covariates $\{\mathbf{x}_1, \ldots \mathbf{x}_n\}$, we sample T independent sets of data $(\mathbf{x}_1, y_{1,t}^{*(\eta)}), \ldots (\mathbf{x}_n, y_{n,t}^{*(\eta)})$, $t = 1, \ldots T$ from the null model $f = \hat{b}_0 + \hat{f}_{(1)} + \cdots + \hat{f}_{(\eta)}$. We fit the main effects model for each set and compute $\hat{L}_t^{*(\eta+1)}$, $t = 1, \cdots, T$. Under the null hypothesis, these B values are equally likely values of $\hat{L}_{(\eta+1)}$, since $L_{(\eta+1)}, L_1^{*(\eta+1)}, \cdots, L_T^{*(\eta+1)}$ are independently and identically distributed. If exactly k of the simulated $\hat{L}^{*(\eta+1)}$ values exceed $\hat{L}_{(\eta+1)}$ and none equal it, the Monte Carlo p-value is $\frac{k+1}{T+1}$. See Davison & Hinkley (1997) for an introduction on Monte Carlo bootstrap test.

Sequential Monte Carlo Bootstrap Tests Algorithm:

Step 1: Let $\eta = 0$ and $f = \hat{b}_0$. We test $H_0 : L_{(1)} = 0$ vs $H_1 : L_{(1)} > 0$. Generate T independent sets of data $(\mathbf{x}_1, y_{1,t}^{*(0)}), \cdots, (\mathbf{x}_n, y_{n,t}^{*(0)}),\ t = 1, \cdots, T$ from $f = \hat{b}_0$. Fit the LBP main effects model and compute the Monte Carlo p-value p_0. If $p_0 < \alpha$, go to step 2; otherwise stop and define $q > \hat{L}_{(1)}$.

Step 2: Let $\eta = \eta{+}1$ and $f = \hat{b}_0 + \hat{f}_{(1)} + \cdots + \hat{f}_{(\eta)}$. We test $H_0 : L_{(\eta+1)} = 0$ vs $H_1 : L_{(\eta+1)} > 0$. Generate T independent sets of data $(\mathbf{x}_1, y_{1,t}^{*(\eta)}), \cdots, (\mathbf{x}_n, y_{n,t}^{*(\eta)})$ based on f, fit the main effects model and compute the Monte Carlo p-value p_η. If $p_\eta < \alpha$ and $\eta < D - 1$, repeat step 2; and if $p_\eta < \alpha$ and $\eta = D - 1$, go to step 3; otherwise stop and define $q = \hat{L}_{(\eta)}$.

Step 3: Stop the procedure and define $q = \hat{L}_{(D)}$.

Chapter 5

Numerical Computation

5.1 Nonlinear Optimization Programming

The objective function in the LBP model either (2.12) or (2.15) consists of two parts: the negative log likelihood part and the l_1 norm of the coefficients. The log likelihood part is a nonlinear, convex and differentiable function of the unknowns (d, c). However, the l_1 penalty part is non-differentiable at the origin. When the objective function is not differentiable, many methods for optimization involved in computing derivatives, such as Newton-Ralphson method, fail to solve the problem. Fan & Li (2001) proposed using a locally quadratic function to approximate the penalty function and applying a modified Newton-Ralphson algorithm. The way we handle this situation is instead of solving the original problem, we will change the problem into its equivalent form but in better condition and solve the new problem.

By introducing proper constraints, we change this problem into minimizing a nonlinear and convex objective function with polyhedral constraints. For the optimization problem in (2.12), we introduce the artificial variables

ξ_α and $\zeta_{\alpha,j}$ for $\alpha = 1, \ldots D, j = 1, \ldots N$ and get the new problem:

$$\begin{aligned}
\min_{d,c,\xi,\zeta} \quad \frac{1}{n} \quad & \sum_{i=1}^{n} \{ -y_i (b_0 + \sum_{\alpha=1}^{D} \; b_\alpha k_1(x_i^\alpha) + \sum_{\alpha=1}^{D} \sum_{j=1}^{N} \; c_{\alpha,j} K_1(x_i^\alpha, x_{j*}^\alpha)) \\
& + log \; [\; 1 + \exp(b_0 + \sum_{\alpha=1}^{D} \; b_\alpha k_1(x_i^\alpha) + \sum_{\alpha=1}^{D} \sum_{j=1}^{N} \; c_{\alpha,j} K_1(x_i^\alpha, x_{j*}^\alpha)) \;] \} \\
& + \lambda_\pi \sum_{\alpha=1}^{D} \; \xi_\alpha \; + \lambda_s \sum_{\alpha=1}^{D} \sum_{j=1}^{N} \; \zeta_{\alpha,j}, \qquad (5.1)
\end{aligned}$$

subject to

$$\begin{cases}
\xi_\alpha \geq b_\alpha & \alpha = 1, \ldots D \\
\xi_\alpha \geq -b_\alpha & \alpha = 1, \ldots D \\
\zeta_{\alpha,j} \geq c_{\alpha,j} & \alpha = 1, \ldots D, \; j = 1, \ldots N \\
\zeta_{\alpha,j} \geq -c_{\alpha,j} & \alpha = 1, \ldots D, \; j = 1, \ldots N
\end{cases}$$

The unknowns are (d, c, ξ, ζ). Thus instead of handling the original problem, we solve its equivalent problem which minimizes a differentiable objective function under linear constraints. Many tools can solve this optimization problem, such as MATLAB, GAMS, MINOS and etc. We tried various solvers and found that MINOS outperformed others for the LBP model.

5.2 MINOS Software

We use MINOS (Murtagh & Saunders 1983) as the underlying non-linear solver. MINOS is a software package for solving large-scale optimization problems (linear and nonlinear programs). It can also process large numbers of nonlinear constraints. The nonlinear functions should be smooth but need not be convex. For linear programs, MINOS uses a sparse implementation of the primal simplex method. For nonlinear objective functions (and linear constraints), it uses a reduced-gradient method with quasi-Newton approximations to the reduced Hessian. For problems with nonlinear constraints, MINOS uses a sparse SLC algorithm (a projected Lagrangian method, related to Robinson's method) and solves a sequence of subproblems in which the constraints are linearized and the objective is an augmented Lagrangian (involving all nonlinear functions). Convergence is rapid near a solution. MINOS makes use of nonlinear function and gradient values. The solution obtained will be a local optimum (which may or may not be a global optimum).

MINOS is especially effective for linear programs and for problems with a nonlinear objective function and sparse linear constraints (e.g., quadratic programs), which is exactly the situation of the LBP model. Thus we find that MINOS performs well with the linearly constrained models and returned consistent results. Under MINOS, non-linear programs are specified in three pieces: the linear portion, the non-linear objective function, and the non-linear constraints. Originally, MINOS required the linear portion of the program to be specified by an MPS file; later versions of MINOS include the subroutine `minoss`, that reads the linear portion from parameters. Using `minoss`, we are able to specify and store the linear portion of the programs internally, eliminating the need to write a new MPS file every time we change

λ. Besides saving time in accessing files, it enables us to hold the program structure and common data constant throughout all solves. Since the only changes to the program occur in the objective function, we are able to utilize solutions from one problem as feasible starting points for the next problem. In addition, we maintain certain internal data structures from one problem to the next, generating faster solution times by the so-called "hot-start".

5.3 Slice Modeling

For every value of λ, program (2.12) or (2.15) must be solved twice — once with y (the original problem) and once with $y + \varepsilon$ (the perturbed problem). This often results in hundreds or thousands of individual solves, depending upon the range for λ. So, in order to obtain solutions in a reasonable amount of time, we need to employ an efficient solving approach, namely slice modeling. See Ferris & Voelker (2000) and Ferris & Voelker (2001) .

Slice modeling is the name that we have given to an approach for solving a series of mathematical programs with the same structure but different data. The name comes from the idea that individual models within the series can be defined by selecting a particular "slice" of data. Under slice modeling, the common program structure is held constant, as well as any "core" data which is shared between programs. The individual programs are then defined simply as data modifications of one another. Further, solutions to slice models solved earlier can be used as starting points for later solves in order to speed up the individual solves. Doing so provides a starting point that has a good chance of being near a solution.

Programs (2.12) and (2.15) are examples of non-linear slice modeling. The L_1 norms can be replaced by non-negative variables constrained linearly to be the corresponding absolute values using standard mathematical programming techniques. After doing so, we have a series of programs with non-linear objective functions and linear constraints. These programs only vary in the objective functions (in the λ values and/or the y values). By applying slice modeling ideas to these programs, we can improve efficiency. Slice modeling removes the necessity of regenerating the constraints for each solve, and also allows previous solutions to be used for starting values.

Once we have solutions for the original and perturbed problems at a particular λ, $ranGACV$ can be calculated. This suggests the approach of solving the original and perturbed problems together for each λ. However, the slice modeling approach suggests the opposite: because fewer changes in the solution take place moving from one λ to another while maintaining the problem type (original or perturbed), previous solutions will have greater impact on future solves if the sequence of original and perturbed solves are separated. Such separation requires extra storage: we must store solution values. However, these solution values require significantly smaller memory than the problem specification, allowing this approach to achieve a significant time improvement. The code is very efficient and easy to use.

Chapter 6

Simulation Study

6.1 Simulation 1: Main Effects Model

6.1.1 Example 1

In this example, there are altogether $D = 10$ covariates: $X_1, \ldots X_{10}$. They are taken to be uniformly distributed in $[0, 1]$ independently. The sample size $n = 1000$. We use the simple random subsampling technique to select $N = 50$ basis functions. The perturbation ϵ is distributed as Bernoulli(0.5) taking two values $\{+0.25, -0.25\}$. The true conditional logit function is

$$f(x) = \frac{4}{3}x_1 + \pi sin(\pi x_3) + 8x_6^5 + \frac{2}{(e-1)}e^{x_8} - 5 \tag{6.1}$$

Four variables X_1,X_3, X_6 and X_8 are important, and the others are noise variables. We fit the main effects LBP model and search the parameters (λ_π, λ_s) globally. Since the true f is known, both $CKL(\lambda)$ and $ranGACV(\lambda)$ are available for choosing the λ's.

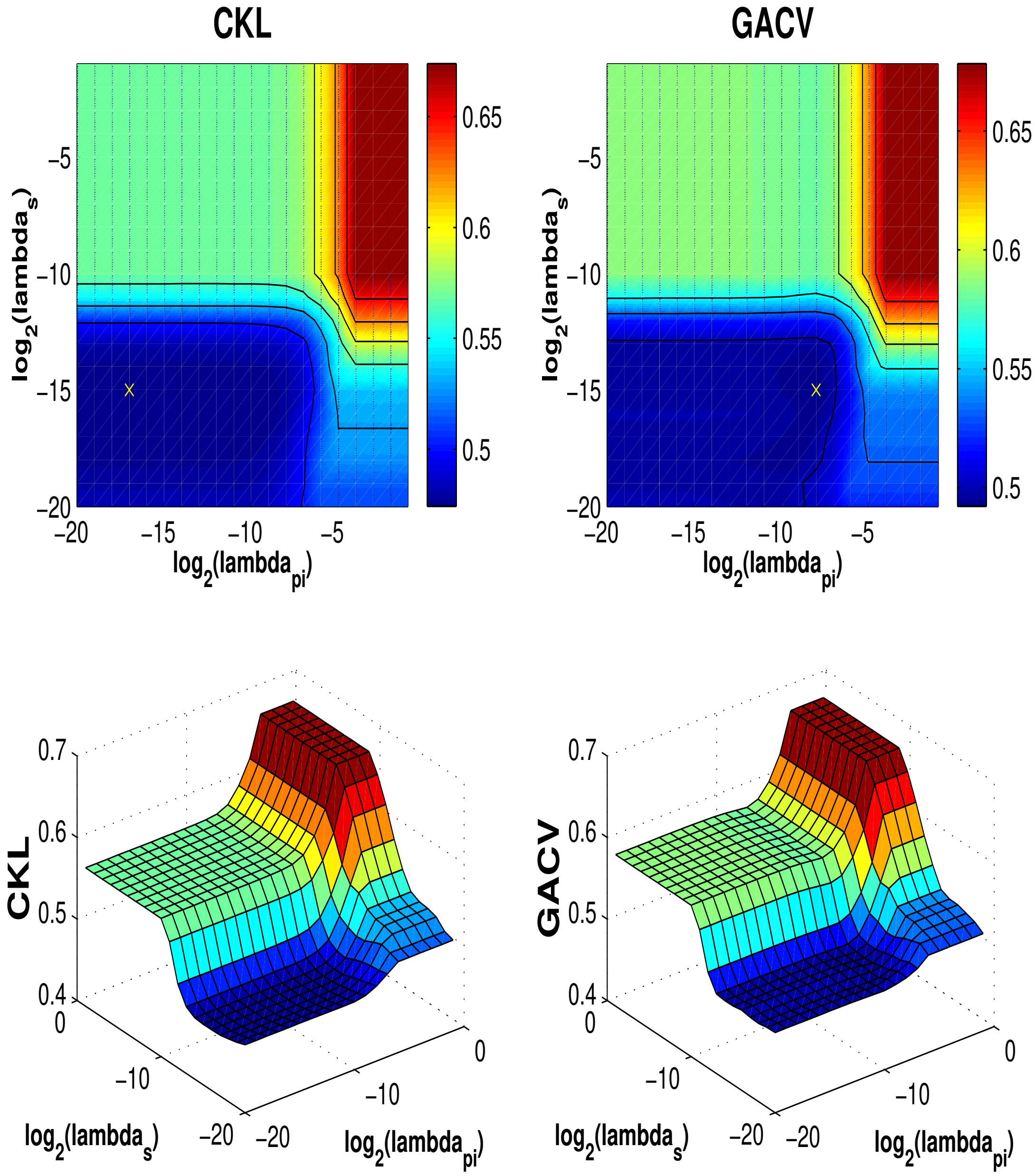

Figure 1: Contours and three-dimensional plots for CKL(λ) and GACV(λ).

Figure 1 depicts the values of $CKL(\lambda)$ and $ranGACV(\lambda)$ as functions of (λ_π, λ_s) within the region of interest $[2^{-20}, 2^{-1}] \times [2^{-20}, 2^{-1}]$. In the top row, are the contours for $CKL(\lambda)$ and $ranGACV(\lambda)$, with the white cross "x" denoting the location of the optimal regularization parameter, $\hat{\lambda}_{CKL} = (2^{-17}, 2^{-15})$ and $\hat{\lambda}_{ranGACV} = (2^{-8}, 2^{-15})$. The bottom row shows their three-dimensional plots. In general $ranGACV(\lambda)$ approximates $CKL(\lambda)$ quite well globally.

In simulation studies, since the true model generating data is known, it is possible to compute $CKL(\lambda)$, which is the true optimization criterion for choosing the smoothing parameter. However, in real life examples when the true model is not available, the LBP model will relies on $ranGACV(\lambda)$ for tuning parameters. Thus it is vital for $ranGACV(\lambda)$ to be a "good" approximate for $CKL(\lambda)$. By "good" we mean that two functions should have similar variation patterns in the parameter space and achieve their optimal values at same locations. It is not necessary for the two functions to have same range in terms of their functions values.

After choosing the optimal parameters $\hat{\lambda}$, we fit the main effects LBP model. The direct solutions obtained are the estimates for all the coefficients. With a little extra work, it is not hard to get the estimates for all the functional components in the model. In order to select important variables or functional components, we compute their importance measure and make a thorough comparison. Here, both L_1 and L_2 norm scores are calculated for the individual component $\hat{f}_1, \ldots \hat{f}_{10}$.

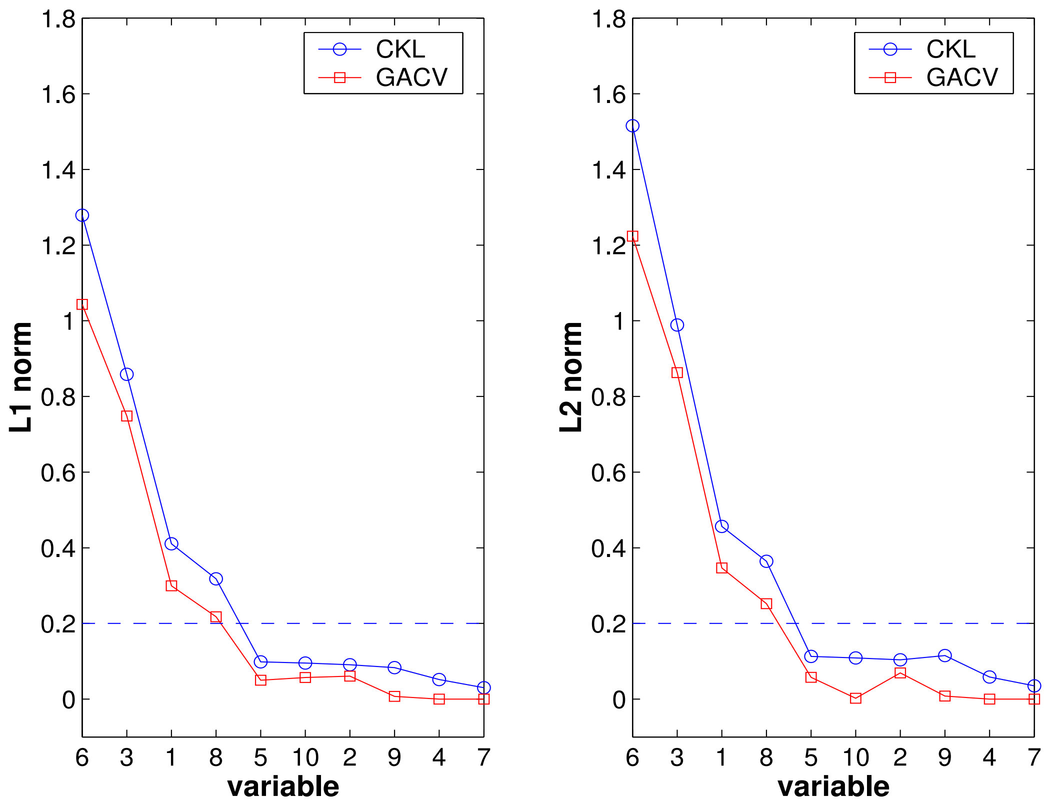

Figure 2: L_1 and L_2 norm scores for the main effects model (Example 1).

Figure 2 plots two sets of the L_1 and L_2 norm scores for the 10 variables, each set obtained respectively using $\hat{\lambda}_{CKL}$ and $\hat{\lambda}_{ranGACV}$, in decreasing orders. Simple observation tells us that all the lines drop dramatically at the beginning and then level off gradually. If a selection is made by using the "eyeball" method, we may either pick up the two variables corresponding to the highest two scores, or the four variables corresponding to the highest four scores.

Different decision-makers may give different lists of important variables. To make an objective decision, we need a cut-off value decided by data-driven or model-driven method. In this dissertation the proposed sequential Monte Carlo bootstrap test algorithm is used to compute the threshold value, which will help us differentiate important variables out of the rest.

Let us focus on the left-hand sided plot first. The dashed line indicates the threshold $q = 0.21$, chosen by the proposed sequential Monte Carlo bootstrap test algorithm. By using this threshold, variables X_6, X_3, X_1, X_8 are selected as "important" variables correctly either by CKL or by $GACV$. In the right-hand sided plot are shown the L_2 norm scores for each of the variables. Using the same threshold q, the same list of variables is selected.

The procedure of the bootstrap tests to determine the threshold q is depicted in Figure 3. We fit the main effects model using $\hat{\lambda}_{ranGACV}$ and sequentially test the hypotheses $H_0 : L_{(\eta)} = 0$ vs $H_1 : L_{(\eta)} > 0,\ \eta = 1, \ldots 10$. In each plot, the variable being tested for importance is bracketed by a pair of *. Light color (green color in a colored plot) is used for the variables which are in the null model, and dark color (blue color in a colored plot) for those not being tested yet. The null hypotheses of the first four tests are all rejected at level $\alpha = 0.05$ based on their Monte Carlo p-value $1/51 \doteq 0.02$. However, the null hypothesis for next component f_5 is accepted with the p-value $10/51 \doteq 0.20$. Thus f_3, f_1, f_6 and f_8 are selected as "important" components and $q = L_{(4)} = 0.34$.

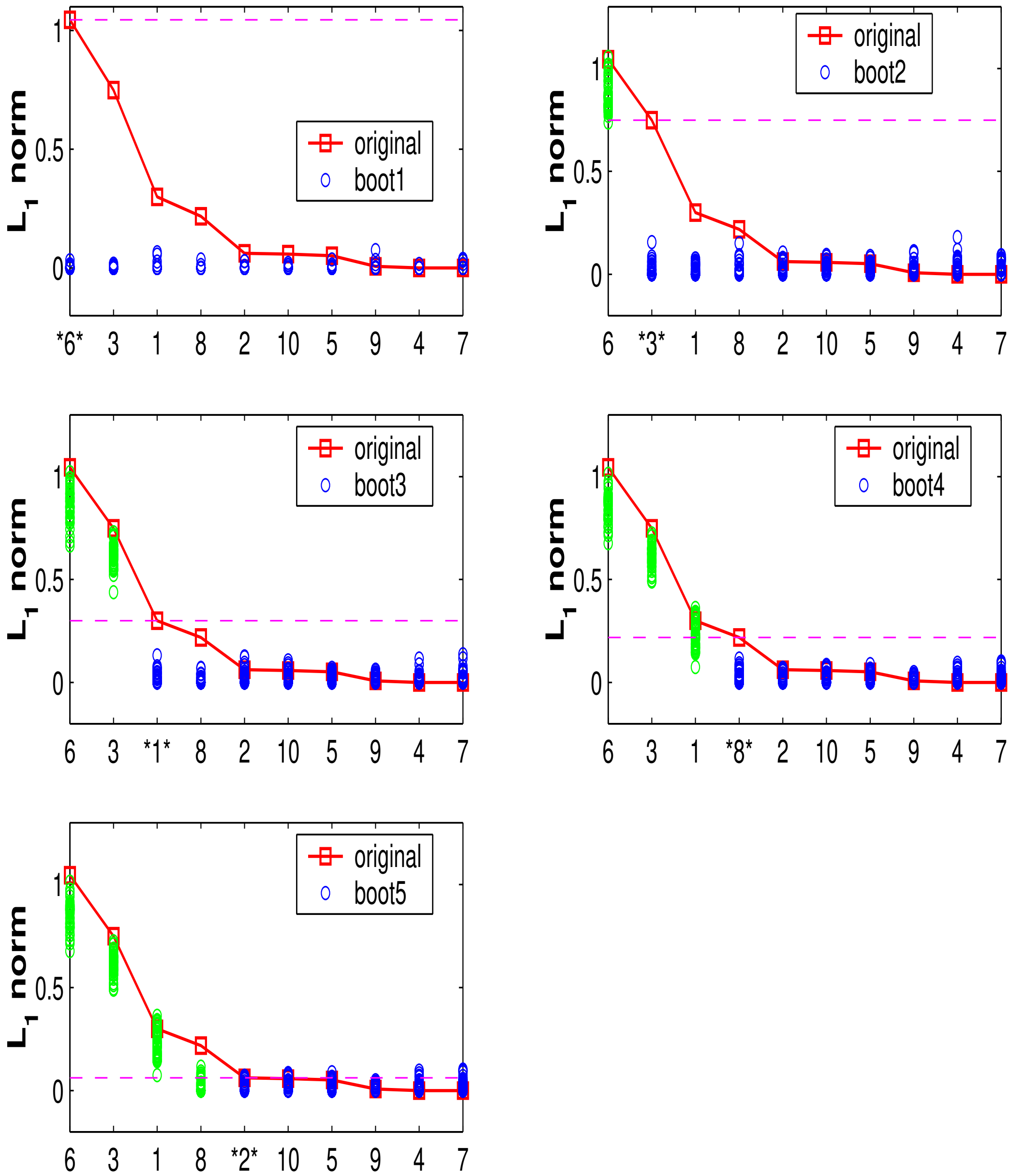

Figure 3: Monte Carlo bootstrap tests (Example 1).

In addition to select important variables, LBP also produces functional estimates for the individual components in the model. Figure 4 plots the true main effects f_1, f_3, f_6 and f_8 and their estimates obtained using $\hat{\lambda}_{ranGACV}$.

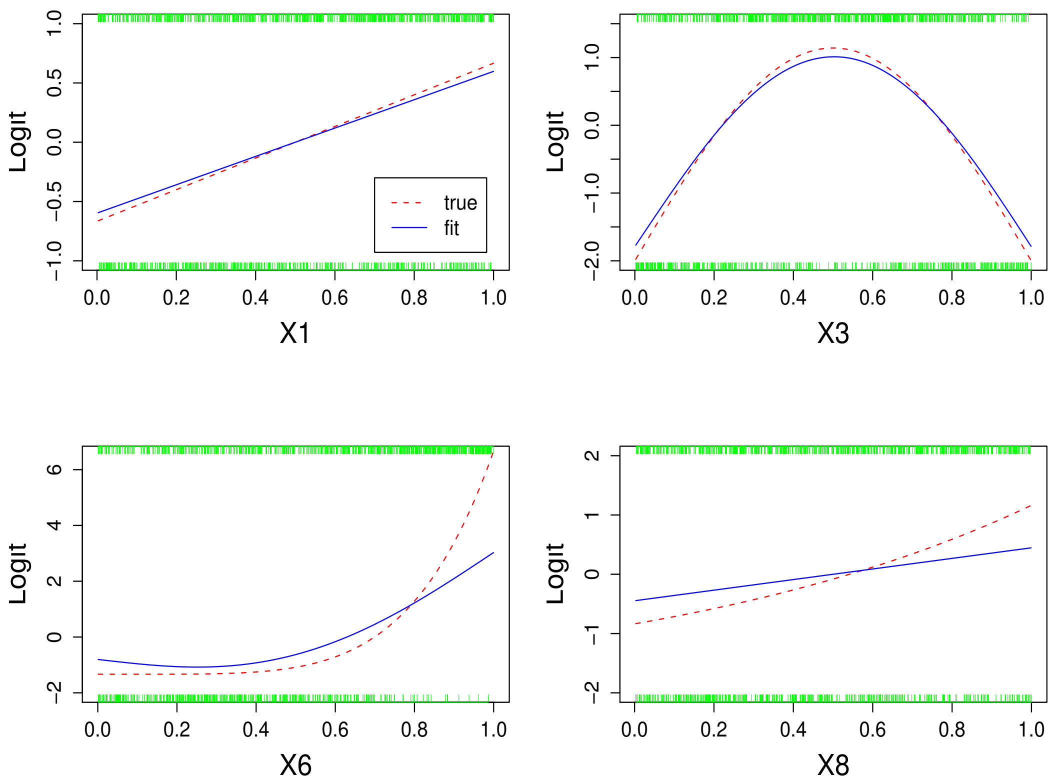

Figure 4: True and estimated univariate logit component.

The mathematical expressions of the main effects are: $f_1 = \frac{4}{3}x_1$, $f_3 = \pi sin(\pi x_3)$, $f_6 = 8x_6^5$, $f_8 = \frac{2}{e-1}e^{x_8}$. In each panel, the solid line is the true curve and the dotted line is the corresponding estimate. In general, the fitted main effects model provides a reasonably good estimate for each important component.

Altogether we generated 20 datasets from the true model and fitted the LBP model for each of the datasets. Figure 5 depicts the L_1 norm scores for the fitted models for all the datasets. The dashed line corresponds to the data set used above, which is the first data set we generated. The dotted lines are for the other datasets. In all the runs, variables X_1, X_3, X_6 and X_8 are ranked at the top, and X_6 always has the largest L_1 norm score, X_3 the second largest score.

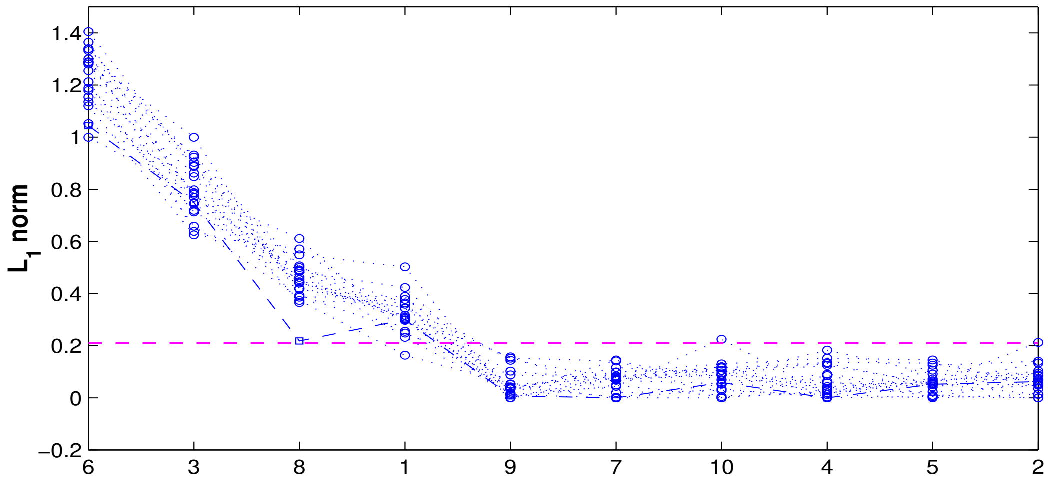

Figure 5: L_1 norm scores of 20 simulated data sets (Example 1)

We observe that in 17 runs, X_8 has a higher L_1 norm score than X_1, while the other 3 runs give X_1 higher score. The theoretical L_1 norms for the variables can be computed: $L_1(f_6) = 1.55, L_1(f_3) = 0.84, L_1(f_8) = 0.48, L_1(f_1) = 0.33$ and zeros for the rest variables. In theory, X_8 has a larger L_1 score than

X_1, however their difference is not big. Thus the order of their L_1 norm estimates could be switched due to randomness in the simulated data. However we should note that this randomness in the data does not prevent the procedure from differentiating important variables out of all the candidate variables. In the following is drawn the L_1 norm plot for the second data set, which belongs to one of the 17 runs that rank X_8 higher than X_1.

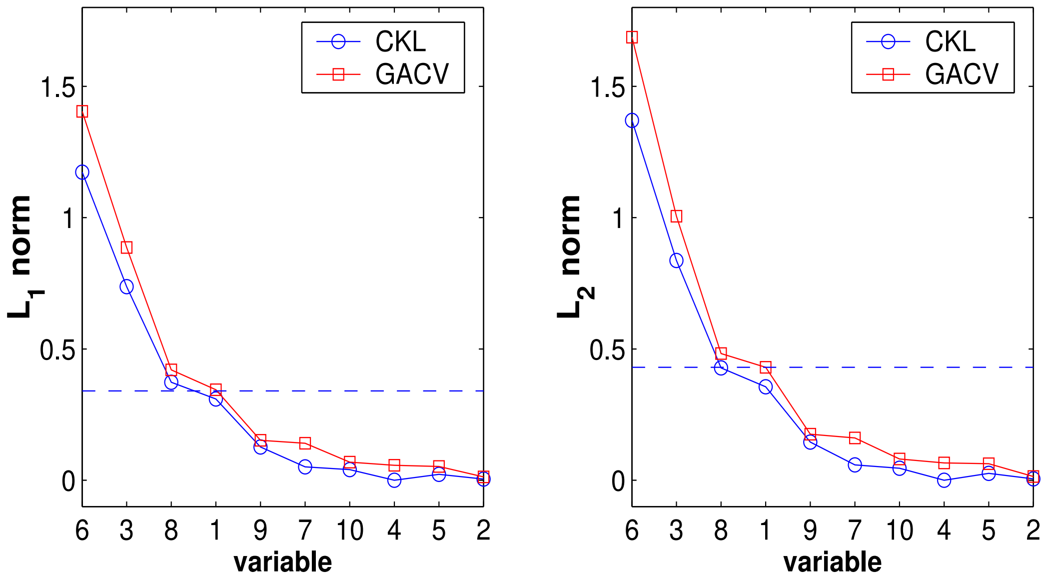

Figure 6: L_1 and L_2 scores of the second simulated dataset (Example 1)

6.1.2 Example 2

The only difference between Example 2 and Example 1 is that in Example 2 the covariates $X_1, \ldots, X_{10}$ have the truncated Normal distribution $N(0.5, 0.3)$ on $[0, 1]$. They are independently distributed. The sample size $n = 1000$ and we select $N = 50$ basis functions. The true logit function and perturbation ϵ are both same as in Example 1.

Firstly, we fit the main effects model using the original data and tune parameters by both $CKL(\lambda)$ and $GACV(\lambda)$. The optimal regularization parameters are $\hat{\lambda}_{CKL} = (2^{-9}, 2^{-15})$ and $\hat{\lambda}_{ranGACV} = (2^{-13}, 2^{-20})$. Figure 7 plots two sets of the L_1 and L_2 norm scores for the 10 variables, each set obtained respectively using $\hat{\lambda}_{CKL}$ and $\hat{\lambda}_{ranGACV}$, in decreasing orders. It shows that variables X_6, X_3, X_8 and X_1 have the largest scores.

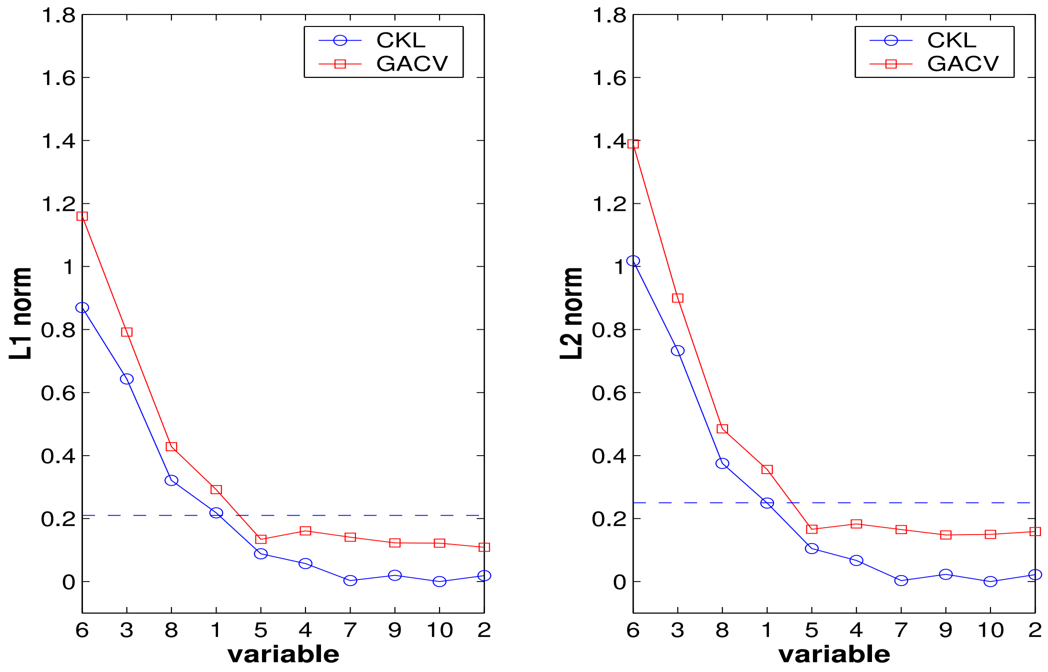

Figure 7: L_1 and L_2 norm scores using the original data (Example 2)

Secondly, we transform each of the covariates by using its true distribution function, thus all the transformed covariates are uniformly distributed. Then we fit the main effects model using the transformed data and tune parameters. The optimal regularization parameters are $\hat{\lambda}_{CKL} = (2^{-10}, 2^{-14})$ and $\hat{\lambda}_{ranGACV} = (2^{-12}, 2^{-20})$. Figure 8 plots two sets of the L_1 and L_2 norm scores for the 10 variables. Again X_6, X_3, X_8 and X_1 are ranked at the top. Also we notice that the optimal parameters chosen using the transformed data are quite close to those chosen using the original data.

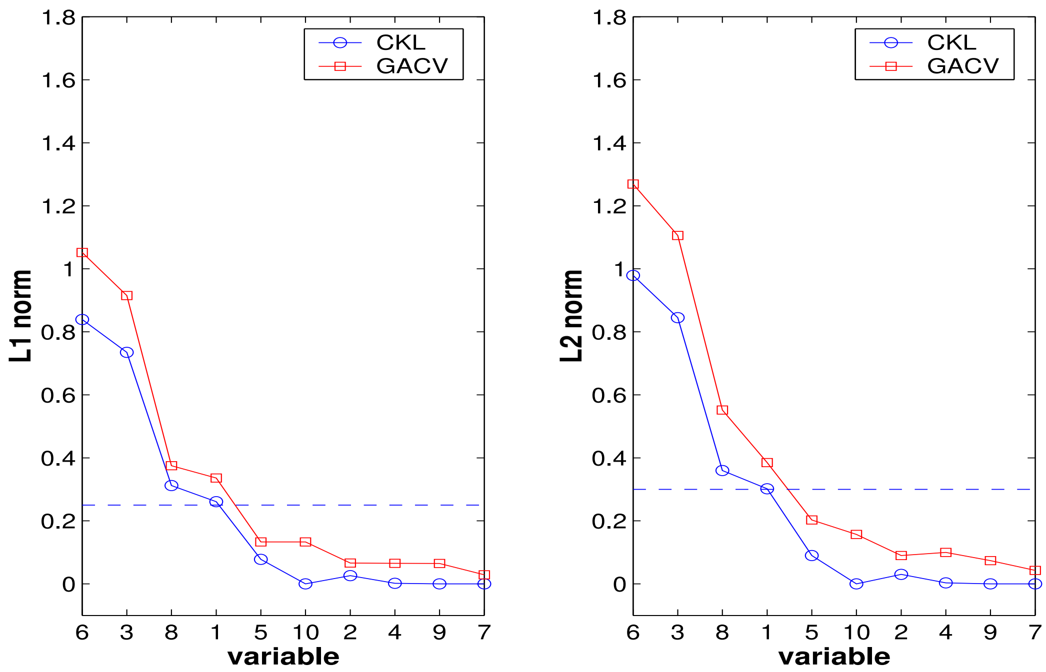

Figure 8: L_1 and L_2 norm scores using the transformed data (Example 2)

In this example, it shows that there are two ways to apply the LBP model to the data with Normally distributed covariates. One way is to handle the original data directly, and the other way is to first transform the data to uniform distribution and then apply the LBP model. Both of them select the important variables successfully in this example.

6.2 Simulation 2: Two-factor Interaction Model

In this example, there are $d = 4$ continuous covariates, independently and uniformly distributed in $[0, 1]$. The true model is a two-factor interaction model, and the important effects are X_1, X_2 and $X_1 * X_2$. We generate $n = 1000$ samples and choose $N = 50$ basis functions. The distribution of the perturbation ϵ is the same as in Simulation 1. The true f is

$$f(x) = 4x_1 + \pi sin(\pi x_1) + 6x_2 - 8x_2^3 + 8x_1 * x_2 - 6$$

There are five tuning parameters $(\lambda_\pi, \lambda_{\pi\pi}, \lambda_s, \lambda_{\pi s}, \lambda_{ss})$ for the two-factor interaction model. In practice, extra constraints may be added on the parameters for different needs. In this example we force all the two-factor interaction terms to have the same penalty parameter, or equivalently, we set $\lambda_{\pi\pi} = \lambda_{\pi s} = \lambda_{ss}$. The optimal parameters obtained are $\hat{\lambda}_{CKL} = (2^{-7}, 2^{-6}, 2^{-8}, 2^{-6}, 2^{-6})$ and $\hat{\lambda}_{GACV} = (2^{-8}, 2^{-6}, 2^{-8}, 2^{-6}, 2^{-6})$. Two sets of optimal parameters are pretty close. The ranked L_1 scores are plotted in Figure 9. . The dashed line in the plot denotes the threshold q. This experiment shows that the LBP two-factor interaction model, fitted using either $\hat{\lambda}_{CKL}$ or $\hat{\lambda}_{GACV}$, selects all the important effects X_1, X_2 and $X_1 * X_2$ correctly.

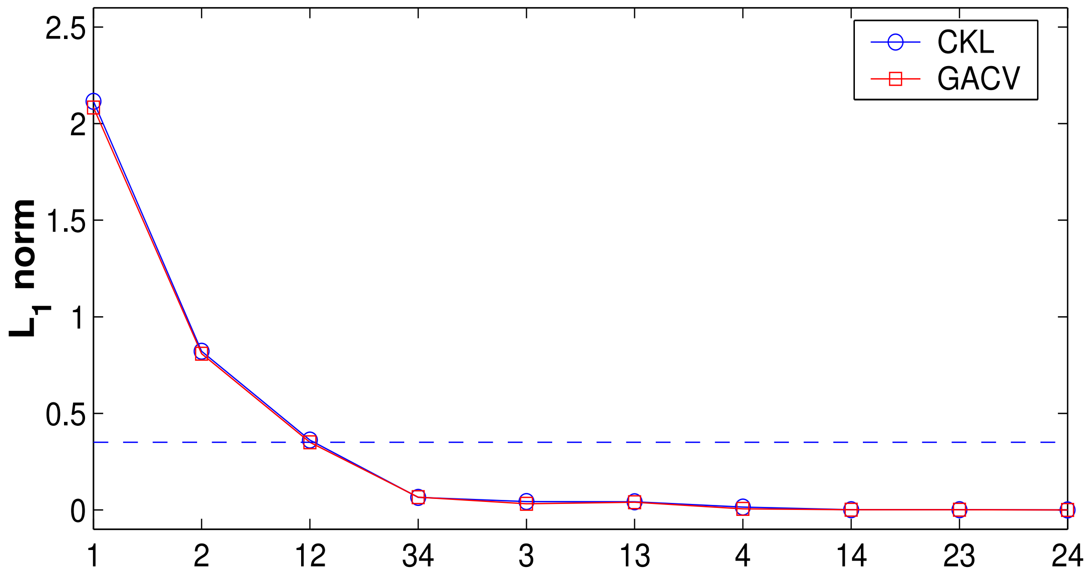

Figure 9: L_1 norm scores for the two-factor interaction model.

6.3 Simulation 3: Incorporating Categorical Variables

In this example, there are both continuous covariates $X_1, \ldots X_{10}$ and categorical covariates Z_1, Z_2. The continuous variables are uniformly distributed in $[0, 1]$ and the categorical variables are Bernoulli(0.5) distributed with values $\{0, 1\}$. The true logit function is

$$f(x) = \frac{4}{3}x_1 + \pi sin(\pi x_3) + 8x_6^5 + \frac{2}{(e-1)}e^{x_8} + 4z_1 - 7.$$

The important main effects are X_1, X_3, X_6, X_8, Z_1. Sample size $n = 1000$ and basis size $N = 50$. We use the same perturbation ϵ as before. The model in (2.16) is fitted. The ranked L_1 norm scores are plotted in Figure 10. When using the threshold chosen by the bootstrap test procedure (denoted by the dashed line), the LBP models using $\hat{\lambda}_{CKL}$ and $\hat{\lambda}_{GACV}$ both select the important variables correctly.

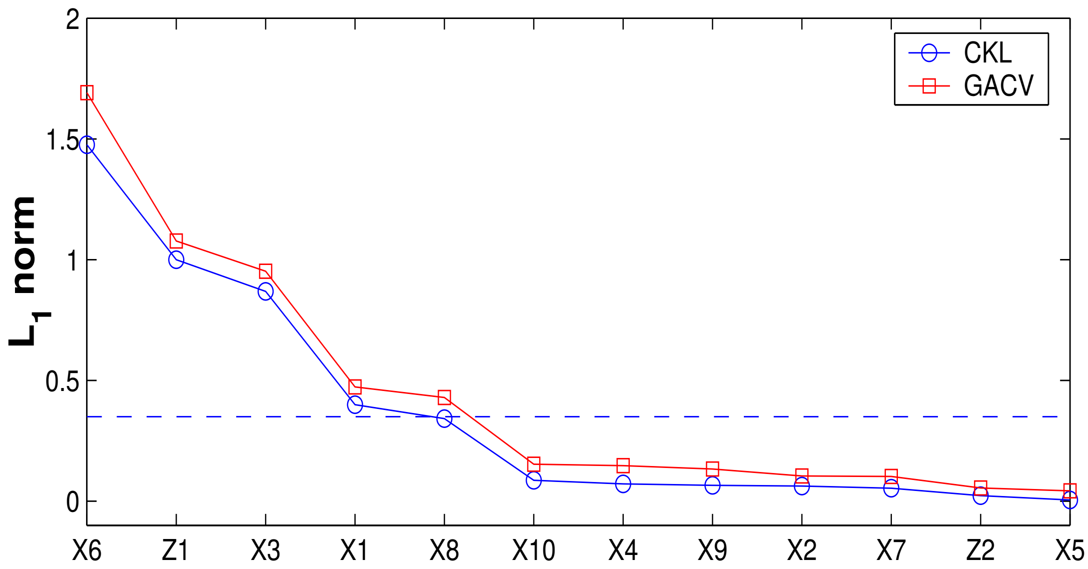

Figure 10: L_1 norm scores for the model incorporating categorical variables.

Chapter 7

Wisconsin Epidemiological Study of Diabetic Retinopathy

7.1 Introduction

The Wisconsin Epidemiological Study of Diabetic Retinopathy (WESDR) is an ongoing epidemiological study of a cohort of patients receiving their medical care in an 11-county area in southern Wisconsin. Diabetic retinopathy, a complication of diabetes can lead to severe decrease in vision and blindness. Nonproliferative retinopathy is an early, usually asymptomatic manifestation which often progresses to proliferative retinopathy which is associated with high risk of loss of vision. It is usually a bilateral condition (both eyes usually affected).

The baseline examination was conducted in 1980-82, and four, ten, fourteen and twenty year followups have been carried out. Details about the study can be found in Klein, Klein, Moss, Davis & DeMets (1984a) , Klein, Klein, Moss, Davis & DeMets (1984b) , Klein, Klein, Moss, Davis & DeMets

(1989) , Klein, Klein, Moss & Cruickshanks (1998) and elsewhere. All younger onset diabetic persons (defined as less than 30 years of age at diagnosis and taking insulin) and a probability sample of older onset persons receiving primary medical care in an 11-county area of southwestern Wisconsin in 1979-1980 were invited to participate. Among 1210 identified younger onset patients, 996 agreed to participate in the baseline examination, and of those, 891 participated in the first follow-up examination. A large number of medical, demographic, ocular and other covariates were recorded in each examination. In particular, stereoscopic color fundus photographs of each eye were graded in a masked fashion using the modified Airline House classification system, and multilevel retinopathy score is assigned to each eye. The severity scale for retinopathy is an ordinal scale.

In this chapter, we examine the relation of a large number of possible risk factors at baseline to the four year progression of diabetic retinopathy. This data set has been extensively analyzed using a variety of statistical methods, such as Craig, Fryback, Klein & Klein (1999) , Kim (1995) and others. Wahba et al.(1995) examined risk factors for progression of diabetic retinopathy on a subset of the younger onset group, members of which had no or non-proliferative retinopathy at baseline. Each person's retinopathy score was defined as the score for the worse eye, and four year progression of retinopathy was defined as occurring if the retinopathy score degraded two levels from baseline.

669 persons were in that data set. A model of the risk of progression of diabetic retinopathy in this population was built using a Smoothing Spline ANOVA model (which has a quadratic penalty functional), using the predictor variables glycosylated hemoglobin (*gly*), duration of diabetes (*dur*) and body mass index (*bmi*). (These variables are described later in this section). That study began with these variables and two other (not independent) variables, age at baseline and age at diagnosis, and these latter two were eliminated at the start. Although it was not discussed in Wahba et al.(1995) , we report here that that study began with a large number (perhaps about 20) of potential risk factors, which was reduced to *gly, dur* and *bmi* as being likely the most important, after many extended and laborious parametric and nonparametric regression analyses of small groups of variables at a time, and by linear logistic regression, by the authors and others. At that time it was recognized that a (smoothly) nonparametric model selection method which could rapidly flag important variables in a data set with many candidate variables was much to be desired. For the purposes of the present study, we make the reasonable assumption that *gly, dur* and *bmi* are the 'truth' (that is, the most important risk factors in the analyzed population)- and thus we are presented with a unique opportunity to examine the behavior of the LBP method in a real data set where, arguably, the truth is known, by giving it many variables in this data set and comparing the results to Wahba et al. (1995) .

Minor corrections and updatings of that data set have been made, (but are not believed to affect the conclusions), and we have 648 persons in the updated data set used here. Some preliminary winnowing of the many potential predictor variables available were made,to reduce the set for examination to 14 potential risk factors. The variables are described as follows.

- Continuous Covariates:

Variable	Name	Description
X_1:	*dur*	duration of diabetes at the time of baseline examination, years
X_2:	*gly*	glycosylated hemoglobin, a measure of hyperglycemia, %
X_3:	*bmi*	body mass index, kg/m^2
X_4:	*sys*	systolic blood pressure, $mmHg$
X_5:	*ret*	retinopathy level
X_6:	*pulse*	pulse rate, count for 30 seconds
X_7:	*ins*	insulin dose, kg/day
X_8:	*sch*	years of school completed
X_9:	*iop*	intra-ocular pressure, $mmHg$

- Categorical Covariates:

Variable	Name	Description	Definition
Z_1:	*smk*	smoking status	0 = no, 1 = any
Z_2:	*sex*	gender	0 = female, 1 = male
Z_3:	*asp*	use of ≥ 1 aspirin for ≥ 3 months while diabetic	0 = no, 1 = yes
Z_4:	*fambd*	family history of diabetes	0 = none, 1 = yes
Z_5:	*mar*	marital status	0 = no, 1 = yes/ever

7.2 Analysis of Four-year Risk of Progression of Diabetic Retinopathy

7.2.1 Covariates Selection

Since the true f is not known in real data analysis, $CKL(\lambda)$ is not computable. Thus we use only $ranGACV(\lambda)$ for tuning the λ. The Monte Carlo sequential bootstrap tests are presented in Figure 11. Along the x-axis, the covariates are coded as *2=gly, 1=dur, 8=sch, 3=bmi, 6=pulse, 5=ret, 4=sys, 9=iop, 7=ins, b=sex, a=smk,c=asp, d=famdb, e=mar*, and are listed in decreasing order of their L_1 norm scores. The tests for *gly, dur, sch, bmi* all have p-value $1/51 \doteq 0.02$, thus these four covariates are selected as important risk factors at the significance level $\alpha = 0.05$.

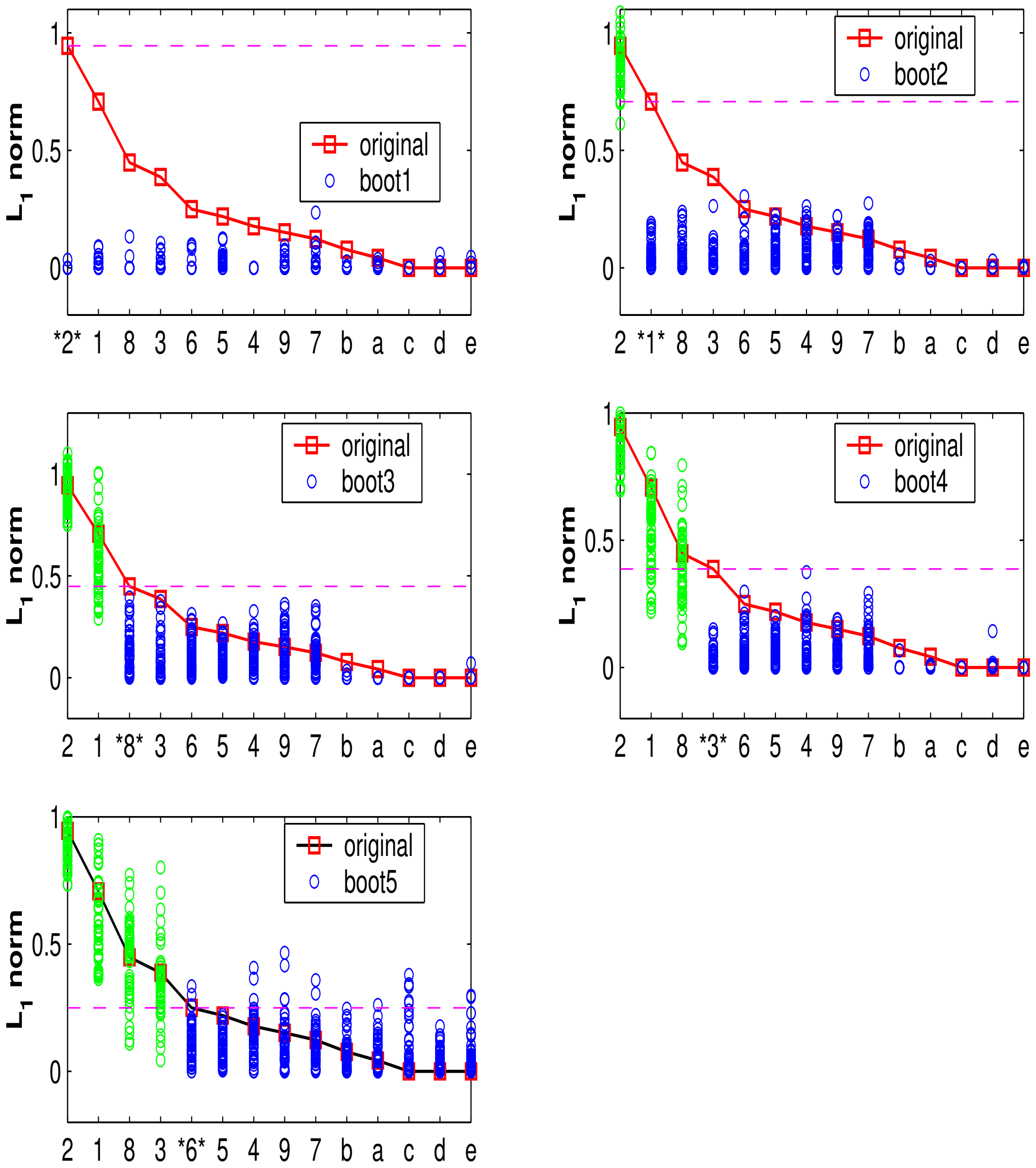

Figure 11: Monte Carlo bootstrap tests for the WESDR main effects model.

Figure 12 plots the L_1 norm scores of the individual functional components. The dotted line indicates the threshold, $q = 0.39$, chosen by the bootstrap tests. We note that the LBP picks out the three most important variables *gly, dur*, and *bmi*, that appeared in Wahba, Wang, Gu, Klein & Klein (1995) . The LBP also chose *sch* (highest year of school/college completed). This variable frequently shows up in demographic studies, when one looks for it, because it is likely a proxy for other variables that are related to disease, e.g. lifestyle or quality of medical care. It did show up in preliminary studies in Wahba et al. (1995) (not reported there) but was not included, because it was not considered a direct cause of disease itself.

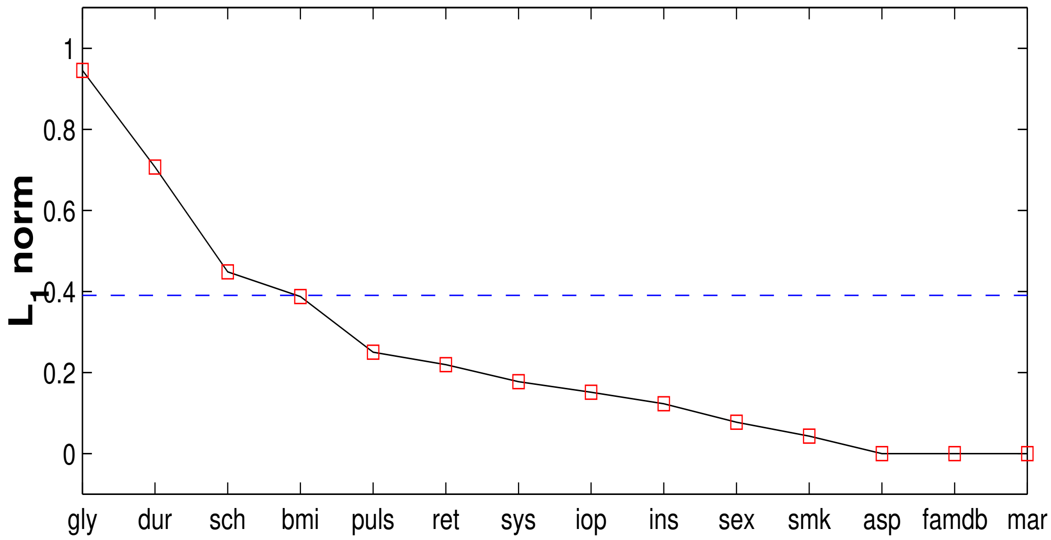

Figure 12: L_1 norm scores for the WESDR main effects model.

Figure 13 plots the estimated logit component for *dur* given by the LBP main effects model. It shows that the risk of progression of diabetic retinopathy increases up to a duration of about 15 years, before decreasing thereafter, which generally agrees with the analysis in Wahba et al.(1995)

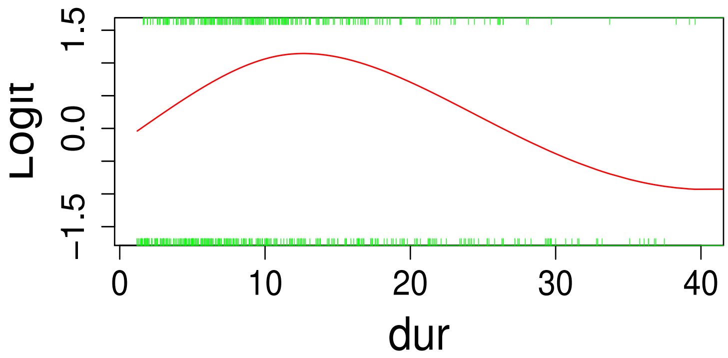

Figure 13: Estimated logit component for *dur*.

When we fit a linear logistic regression model using the function *glm* in *R* package, the linear coefficient for *dur* is not significant at level $\alpha = 0.05$. The curve in Figure 13 exhibits a hilly shape, which indicates that a quadratic function might fit the curve better. We refit the linear logistic model by intentionally including dur^2, the hypothesis test for dur^2 is significant with p-value 0.02. This fact confirms the discovery of the LBP, and shows that LBP can be a valid screening tool to help us decide the appropriate functional form for the individual covariate.

When fitting the two-factor interaction model in (2.15) with the constraints $\lambda_{\pi\pi} = \lambda_{\pi s} = \lambda_{ss}$, the *dur-bmi* interaction in Wahba et al.(1995) was not found here. We note that the interaction terms tend to be washed out if there are only a few interactions. However further exploratory analysis may be carried out by rearranging the constraints and/or varying the tuning parameters subjectively.

It is noted that the solution to the optimization problem is very sparse. In this example, we observed that approximately 90% of the coefficients are zeros in the solution.

7.2.2 Correlated Covariates Selection

In this data set, the correlation between AGE and DUR is as high as 0.76. Classical variable selection methods, such as stepwise method, usually have difficulty in handling the collinearity situation. In this section, we conduct a second analysis on the WESDR data set to show the performance of the LBP model when collinearity exists. We only take into account six continuous covariates: *age, dur, gly, bmi, sys* and *ret*. In order to make comparisons, we fit the LBP main effects model twice, once with *age* included and once with *age* excluded. The L_1 norm scores for the two models are shown in Figure 14.

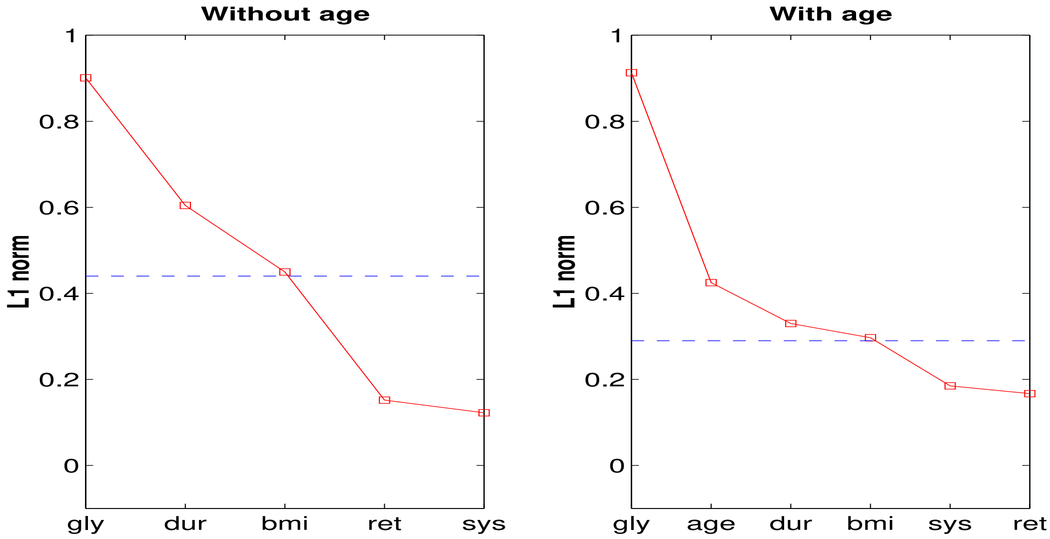

Figure 14: L_1 norm scores for the WESDR study: (left) without *age*, (right) with *age*.

In the left-hand sided plot is the fitting result of the LBP model by excluding *age*. Monte Carlo bootstrap tests decide the threshold $q = 0.44$. The covariates *gly, dur* and *bmi* are selected as important variables at level $\alpha = 0.05$ with the Monte Carlo p-values, respectively, $1/51, 1/51 \doteq 0.05, 2/51 \doteq 0.04$. In the second experiment, we include all the six variables. The threshold obtained is $q = 0.30$. As shown in the right-hand sided plot of Figure 14, four important covariates are selected: *gly, age, dur* and *bmi*. If we fit the linear logistic model by the function *glm* in *R* package, using the stepwise selection with *AIC* criterion, the covariate *dur* is missed. In this example it shows that the LBP model is still valid when high correlation exists among the covariates.

Chapter 8

Beaver Dam Eye Study

8.1 Introduction of BDES

The Beaver Dam Eye Study (BDES) is an ongoing population-based study of age-related ocular disorders. It aims at collecting information related to the prevalence, incidence and severity of age-related cataract, macular degeneration and diabetic retinopathy. Between 1987 and 1988, 5925 eligible people (age 43-84) were identified in Beaver Dam, WI. and of those, 4926(83.1%) participated in the baseline exam. Five and ten year followup data have been collected and results are being reported. Many variables of various kinds are collected, including mortality between baseline and the followups. A detailed description of the study is given by Klein, Klein, Linton & DeMets (1991). . Recent reports include Klein, Klein, Lee & Cruickshanks (2001). .

We are interested in the relation between five-year mortality for the non-diabetic study participants and possible risk factors at baseline. We focus on the non-diabetic participants since the pattern of risk factors for people with diabetes differs from that of the rest of the population. We consider

10 continuous and 8 categorical covariates, whose detailed information is given in the following tables. Y is assigned 1 if a patient participated in the baseline examination and died prior to the start of the first 5-year follow-up; Y is assigned 0 otherwise.

- Continuous Covariates:

Variable	Name	Description
X_1:	*pky*	pack years smoked, packs per day/20)∗years smoked
X_2:	*sch*	highest year of school/college completed, years
X_3:	*inc*	total household personal income, thousands/month
X_4:	*bmi*	body mass index, kg/m^2
X_5:	*glu*	glucose (serum), mg/dL
X_6:	*cal*	calcium (serum),mg/dL
X_7:	*chl*	cholesterol (serum), mg/dL
X_8:	*hgb*	hemoglobin (blood), g/dL
X_9:	*sys*	systolic blood pressure, $mmHg$
X_{10}:	*age*	age at examination, years

- Categorical Covariates:

Variable	Name	Description	Definition
Z_1:	cv	cardiovascular disease history	$0 =$ no, $1 =$ yes
Z_2:	sex	gender	$0 =$ female, $1 =$ male
Z_3:	$hair$	hair color	$0 =$ blond/red, $1 =$ brown/black
Z_4:	$hist$	heavy drinking	$0 =$ never, $1 =$ past/currently
Z_5:	$nout$	winter leisure time	$0 =$ indoors,$1 =$ outdoors
Z_6:	mar	marital status	$0 =$ no, $1 =$ yes/ever
Z_7:	sum	day spent outdoors in summer	$0 =< 1/4$ day, $1 => 1/4$ day
Z_8:	vtm	vitamin use	$0 =$ no, $1 =$yes

There are 4422 non-diabetic study participant in the baseline examination, and 395 of them have missing data in the covariates. For the purpose of this study we assume the missing data are missing at random, thus these 395 subjects are not included in our analysis. This assumption is not necessarily valid, age, blood pressure, body mass index, cholesterol, sex, smoking, hemoglobin may well affect the missingness, but a further examination of the missingness is beyond the scope of the present study. In addition, we exclude another 10 participants who have either outlier values $pky > 158$ or very abnormal records $bmi > 58$ or $hgb < 6$. Thus we report an analysis of the remaining 4017 non-diabetic participants from the baseline population.

8.2 Analysis of Five-Year Risk of Mortality

The goal of this analysis is to select important risk factors for the five-year mortality event. A full and correct discovery of risk factors can increase our knowledge of the existing outcome pattern and help to reduce the mortality rate of the participants in the future study. Thus the variable selection is significantly meaningful for medical research.

8.2.1 Including All Variables

There are 18 covariates listed above. Though some prior knowledge or experience of the doctors can be used to narrow down the number of variables for consideration, we deliberately take into account some "noisy" variables in the analysis. The variables *hair, nout* and *sum* are not directly related to mortality in general, and their inclusion is to show the performance of the proposed approach. These variables are not expected to be picked out eventually by the model. The main effects model incorporating categorical variables in (2.16) is fitted. The sequential Monte Carlo bootstrap tests are shown as follows.

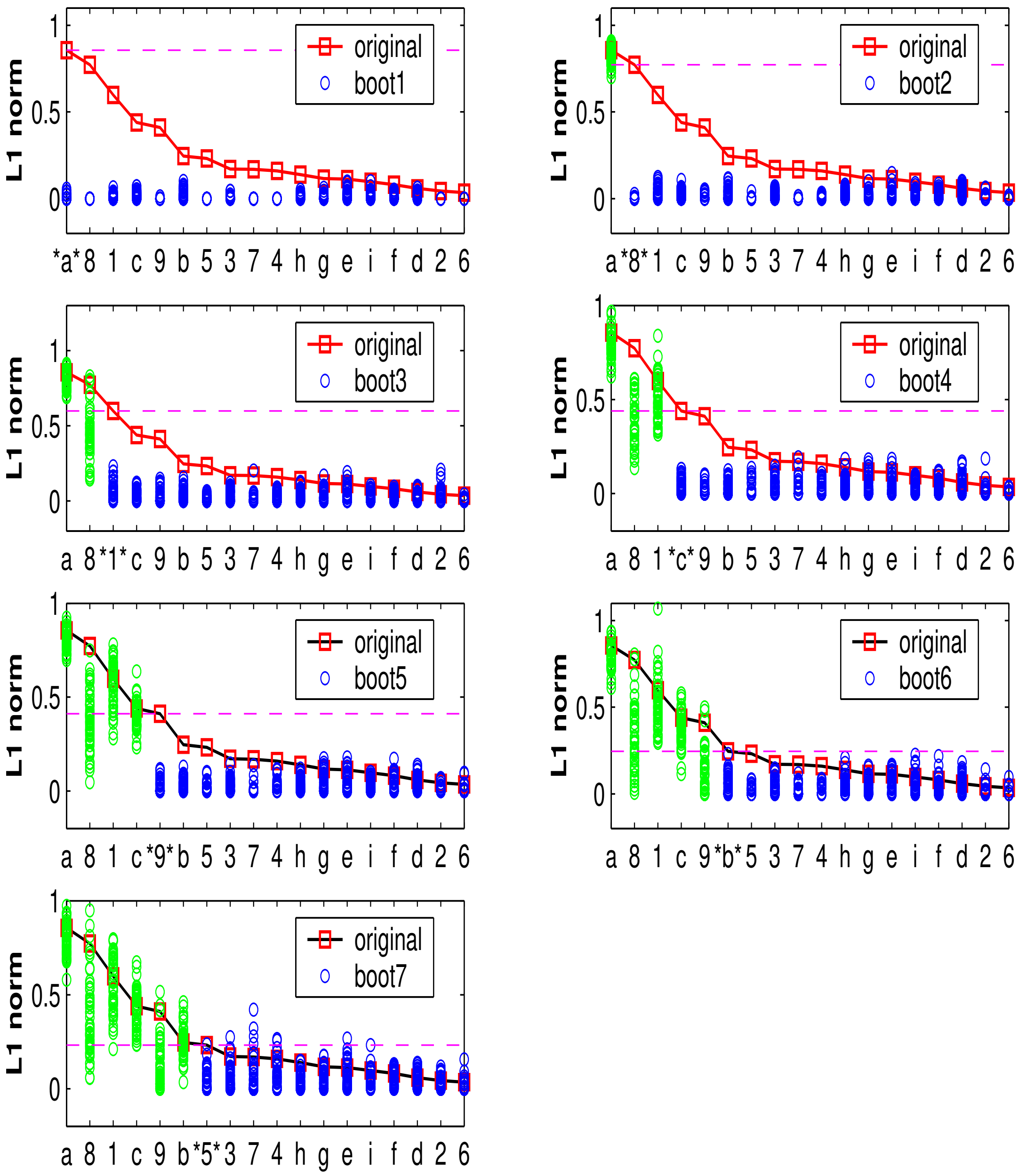

Figure 15: Monte Carlo bootstrap tests for the BDES (all variables included)

In Figure 15 Along the x-axis, the covariates are coded as *0=age, 8=hgb, 1=pky, b=sex, 9=sys, a=cv, 5=glu, 3=inc, 7=chl, 4=bmi, g=sum, f=mar, d=hist, h=vtm, e=nout, c=hair, 2=sch, 6=cal*, and are listed in decreasing order of their L_1 norm scores. The tests for the first six covariates: *age, hgb, pky, sex, sys, cv* all have Monte Carlo p-values $1/51 \doteq 0.02$; while the test for *glu* is not significant with p-value $9/51 = 0.18$. The threshold is chosen as $q = L_{(6)} = 0.25$.

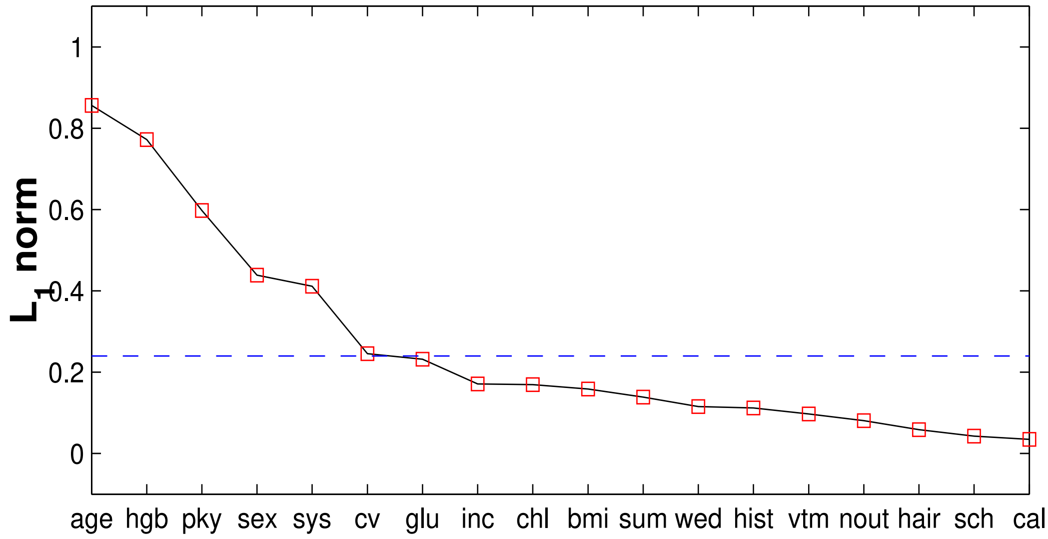

Figure 16: L_1 norm scores for the BDES (all variables included).

Figure 16 plots the L_1 norm scores for all the potential risk factors. Using the threshold (dashed line) $q = 0.25$ chosen by the bootstrap test procedure, the LBP model identifies six important variables: *age, hgb, pky, sex, sys, cv* for the five-year mortality. It is noticed that the L_1 curve decreases very quickly for the first few variables, then levels off at a slower speed.

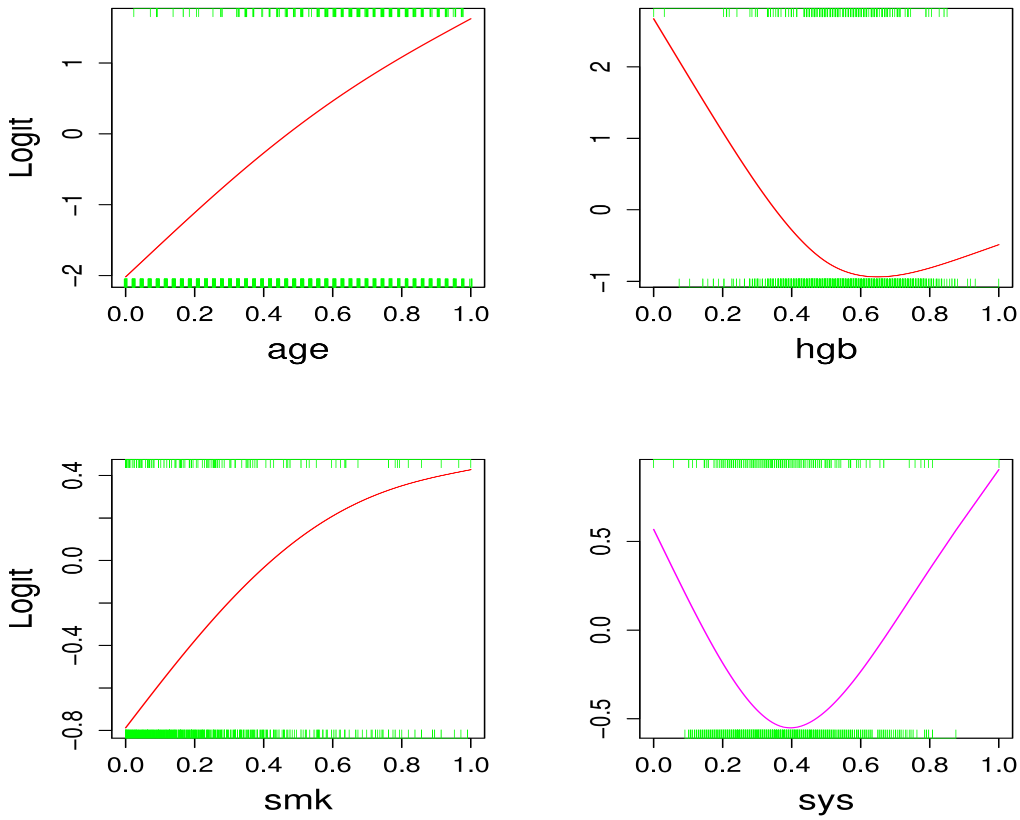

Figure 17: Estimated univariate logit component for important variables.

Compared with the LBP model, the linear logistic model with stepwise selection using *AIC* criterion, implemented with the function `glm` in *R* package, misses the variable *sys* but selects three more variables: *inc, bmi* and *sum.* Figure 15 depicts the estimated univariate logit components for the important continuous variables selected by the LBP model. All the curves can be approximated reasonably well by linear models except *sys*, whose functional form exhibits a quadratic shape. This explains why *sys* is not selected by the logistic model. When we refit the logistic regression model by including sys^2 in the model, the stepwise selection picked out both *sys* and sys^2.

8.2.2 Excluding Noisy Variables

Sometimes with help of the prior knowledge or experience of the experts, we know some variables are either irrelevant or unimportant to the event of interest before doing our analysis. For example, in this data set, doctors might doubt that *hair, nout, sum* are directly related to mortality and question whether to take into account all the variables. Thus we refitted the LBP model by excluding these three variables. The bootstrap tests are shown in Figure 18.

The null hypotheses of the first five tests are all rejected at level $\alpha = 0.05$ based on their Monte Carlo p-value $1/51 \doteq 0.02$. The sixth test has the p-value $3/51 \doteq 0.06$. If using the significance level 0.1, the LBP model selects six important variables: *age, hgb, pky, sys, sex, cv*. If using the level 0.05, *cv* is significant by the margin. This list of important risk factors is the same as the analysis above by including all the variables.

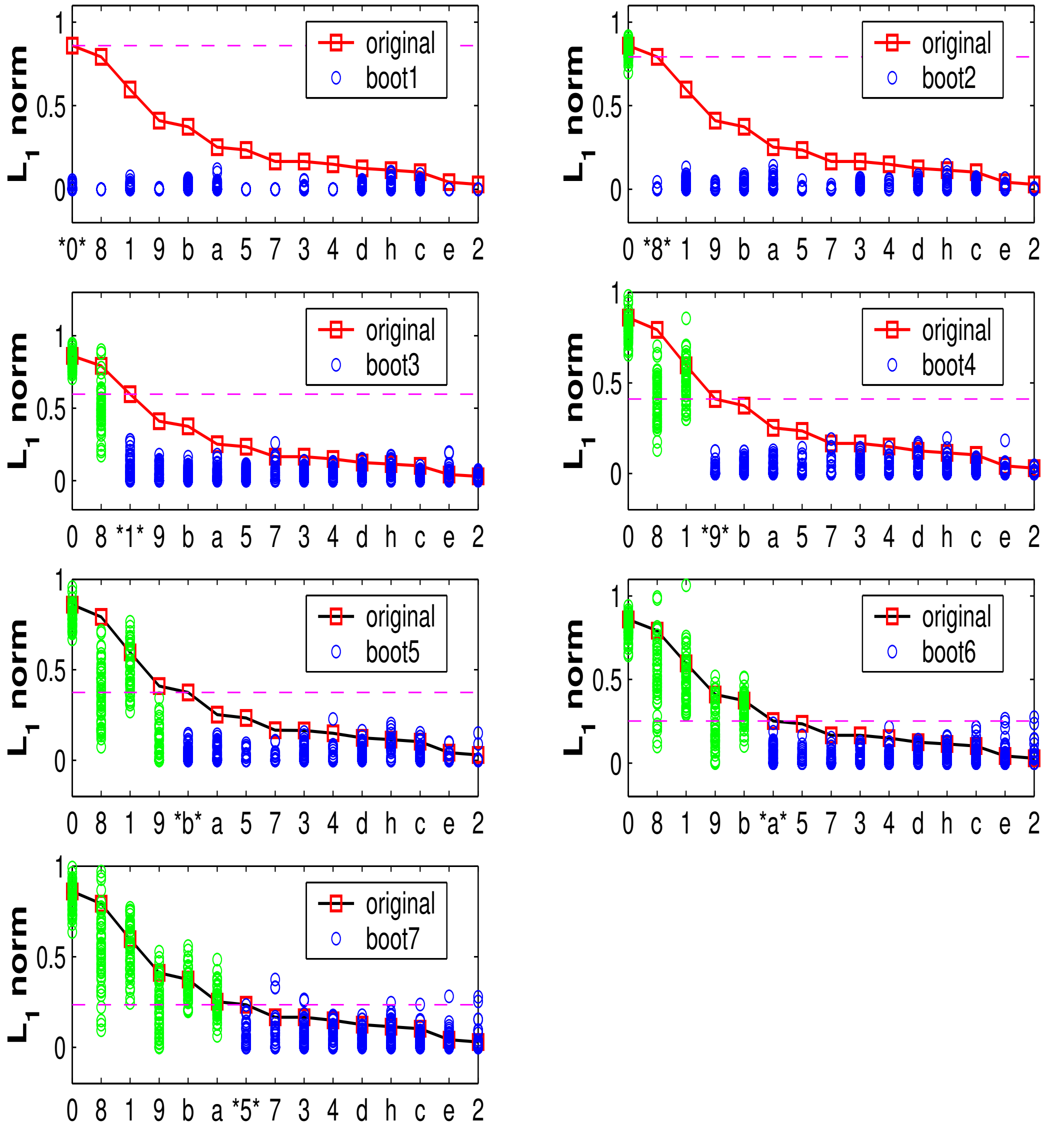

Figure 18: Monte Carlo bootstrap for the BDES (excluding "noisy" variables).

Figure 19 plots the L_1 norm scores for the model fitted without including *hair, nout, sum.* The threshold $q = 0.25$ chosen by the bootstrap test procedure is denoted by the dashed line. We note the rank of variables is a little bit different from Figure 16. Two pairs of variables switch their orders: *sex* and *sys*, *inc* and *chol*. Since the L_1 scores for the two variables in each pair are kind of lose, the order switch could be due to randomness in the data.

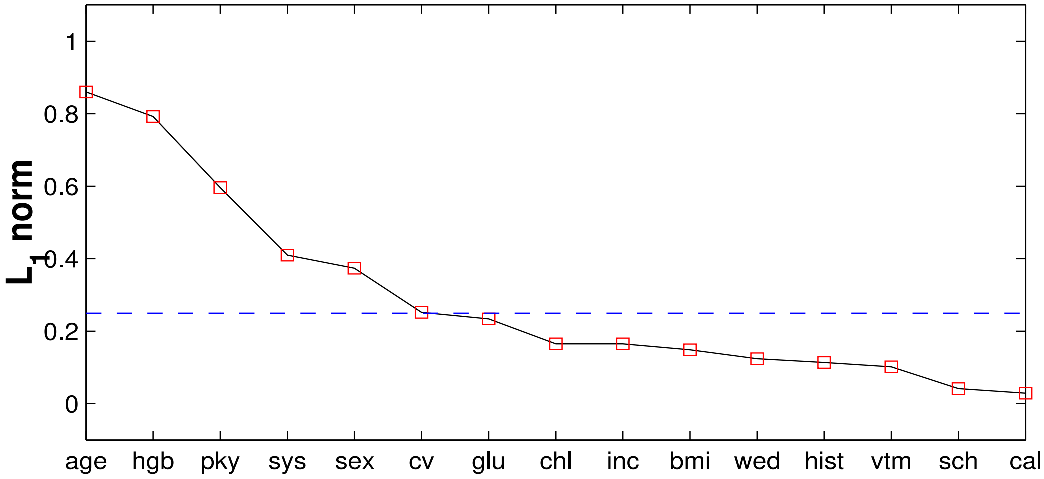

Figure 19: L_1 norm scores for the BDES (excluding “noisy” variables).

Chapter 9

Conclusion

9.1 Summary

The aim of this dissertation has been to explore the construction of data driven models that both fit data well and select important independent variables which are relevant in predicting the dependent variables. Nonparametric variable selection approaches are in demand because of their predictive accuracy and great flexibility. This dissertation has developed the likelihood basis pursuit (LBP) approach based on smoothing spline ANOVA model. In the spirit of LASSO, LBP produces the shrinkage functional estimates by imposing the l_1 penalty on the coefficients of the basis functions. Using the proposed measure of importance for the functional components, LBP selects important variables effectively and the results are highly interpretable.

LBP can handle continuous variables and categorical variables simultaneously. Although in this dissertation the continuous variables have all been on subsets of the real line, it is clear that other continuous domains are possible. LBP is fully developed for the problem of nonparametric binary regression in this dissertation, but it is applicable to any of the other exponential distributions as well, of course to Gaussian data.

This method is believed to be a useful addition to the toolbox of the data analyst. It provides a way to examine the possible effects of a large number of variables in a nonparametric manner, complimentary to standard parametric models in its ability to find nonparametric terms that may be missed by parametric methods. It has an advantage over quadratically penalized likelihood methods when it is desired to examine a large number of variables or terms simultaneously inasmuch as the l_1 penalties result in sparse solutions. It can be an efficient tool for examining complex data sets to identify and prioritize variables (and, possibly, interactions) for further study, and for building more traditional parametric or penalized likelihood models, for which confidence intervals and theoretical properties are known, based only on the variables or interactions identified by the LBP.

9.2 Future Work

9.2.1 Classification Problem

Consider a set of training data samples in which each sample is labelled as belonging to one of several pre-specified "classes". Data classification then involves the assignment of newly observated data samples to the classes on the basis of statistical "models" built for each of the classes. The goal is to choose the model which minimises the classification error. There are two basic types of classifier, namely (1) those which attempt to minimise the error rate without regard to density estimation and (2) those which use density estimates to derive a classification. The former method gives only the class assignment, while the latter method also give the likelihood of a sample belonging to each class. This means that the former methods, despite often giving good classification accuracy, is not recommended for use when accountability is essential, e.g. medical image analysis, or when ranked probabilities are required, e.g. speech recognition.

In this dissertation the LBP is applied Bernoulli data analysis, which can be cast as a classification problem if our interest focuses on minimizing the classification error and predicting the future response (label) rather than estimating the conditional probabilities. Similarly, the multinomial regression problem can be cast as a multi-category classification problem. With some proper modifications in the model setting, such as the loss function and parameters tuning criteria, LBP can also be a promising classification approach to any two-category or multi-category classification problem.

Compared with other classification methods, the advantage of LBP is to select the important variables/features which play important roles in deciding the classification boundary. Removing the redundant variables in the model potentially will decreases the misclassification rate as well. The classification results can be highly interpretable to the practitioner. Thus the application of the LBP to some real classification problems with its comparison to modern classification methods is one future direction.

9.2.2 Support Vector Machines

Support Vector Machines (SVMs), a class of large-margin classifiers developed by Vapnik (1995) , have become quite popular due to many attractive properties and successful performance in many areas. For any given classification problem there exists a fundamental limit to the classification accuracy achievable, and this minimum error rate is called the "Bayes error" for that problem. Lin (2002) showed that the solutions of SVMs directly targets the Bayes decision rule. One important feature of the SVMs is the solution to the optimization problem is quite sparse, and the decision boundary only relies on a small number of important examples, which are usually called "support vectors". However the classical version of SVMs does not bear the feature of selecting important variables. Or in other words, SVMs has a power of selecting important examples/observations rather than variables.

By endowing the SVMs with certain types of L_1 penalties as in the LBP methods, we may develop the so-called "variable selection Support Vector Machines", or "feature selection Support Vector Machines", which will definitely enhance the power of SVMs. The combination of SVMs and LBP will select both important covariates (columns in the data) and important examples (rows in the data) and produce flexible model estimates, which are the most desired features for any model selection methodology and data mining tool. There is a significant potential in the application.

9.2.3 Microarray Gene Expression Data

Another interesting problem I am eager to work on is gene selection for the microarray gene expression data. One problem of interest is to classify n cell lines in cancer group and non-cancer group correctly based on expression levels of p genes. For a typical gene data set, always millions of genes are assayed on a smaller number of samples. This "large p, small n" data format has been a challenge to both statisticians and genetists.

In theory, LBP as a methodology developed in the framework of reproducing kernel Hilbert space (RKHS) has the potential to handle the situation where the number of covariates exceeds the number of observations. In addition to make a prediction or a classification, in many times the biologists are more interested in selecting just a few important genes which adequately explain the dependent variable ("tumor type" here), among many candidate genes. As a nonparametric variable selection approach, the LBP could be a potential tool of gene selections for microarray gene expression study. Given the vast quantity and special background of gene expression data, collaborations with genetists will be needed to evaluate the performance of the LBP approach.

Bibliography

Aronszajn, N. (1950), 'Theory of reproducing kernels', *Trans. Am. Math. Soc* **68**, 337–404.

Brieman, L. (1995), 'Better subset selection using the nonnegative garrotte', *Technometrics* **37**, 373–384.

Chen, S., Donoho, D. & Saunders, M. (1998), 'Atomic decomposition by basis pursuit', *SIAM J. Sci. Comput.* **20**, 33–61.

Craig, B. A., Fryback, D. G., Klein, R. & Klein, B. (1999), 'A Bayesiuan approach to modeling the natural history of a chronic condition from observations with intervention', *Statistics in Medicine* **18**, 1355–1371.

Craven, P. & Wahba, G. (1979), 'Smoothing noise data with spline functions: estimating the correct degree of smoothing by the method of generalized cross validation', *Numerische Mathematik* **31**, 377–403.

Davison, A. C. & Hinkley, D. V. (1997), *Bootstrap methods and their application*, Cambridge.

Fan, J. & Li, R. Z. (2001), 'Variable selection via penalized likelihood', *Journal of American Statistical Association* **96**, 1348–1360.

Ferris, M. C. & Voelker, M. M. (2000), Slice models in general purpose modeling systems, Technical Report 00-10, Data Mining Institute, Computer Sciences Department, University of Wisconsin.

Ferris, M. C. & Voelker, M. M. (2001), Slice models in GAMS, Technical Report 01-07, Data Mining Institute, Computer Sciences Department, University of Wisconsin. To appear in the Proceeding of OR 2001.

Frank, I. E. & Friedman, J. H. (1993), 'A statistical view of some chemometrics regression tools', *Technometrics* **35**, 109–148.

Fu, W. J. (1998), 'Penalized regression: the bridge versus the lasso', *Journal of Computational and Graphical Statistics* **7**, 397–416.

Gu, C. (2002), *Smoothing spline ANOVA models*, Springer-Verlag.

Gunn, S. R. & Kandola, J. S. (2002), 'Structural modelling with sparse kernels', *Machine Learning* **48**, 115–136.

Hutchinson, M. (1989), 'A stochastic estimator for the trace of the influence matrix for Laplacian smoothing splines', *Commun. Statist.-Simula.* **18**, 1059–1076.

Kim, K. (1995), 'A bivariate cumulative probit regression model for ordered categorical data', *Statistics in Medicine* **14**, 1341–1352.

Kimeldorf, G. & Wahba, G. (1971), 'Some results on Tchebycheffian spline functions', *Journal of Math. Anal. Applic.* **33**, 82–95.

Klein, R., Klein, B., Lee, K., Cruickshanks, K. & Chappell, R. (2001), 'Changes in visual acuity in a population overr a 10-year period. The Beaver Dam Eye Study', *Ophthalmology* **108**, 1757–1766.

Klein, R., Klein, B., Linton, K. & DeMets, D. L. (1991), 'The Beaver Dam eye study: Visual acuity', *Ophthalmology* **98**, 1310–1315.

Klein, R., Klein, B., Moss, S. & Cruickshanks, K. (1998), 'The Wisconsin Epidemiologic Study of diabetic retinopathy. XVII. the 14-year incidence and progression of diabetic retinopathy and associated risk factors in type 1 diabetes', *Ophthalmology* **105**, 1801–1815.

Klein, R., Klein, B., Moss, S. E., Davis, M. D. & DeMets, D. L. (1984*a*), 'The Wisconsin Epidemiologic Study of Diabetic Retinopathy. II. Prevalence and risk of diabetes when age at diagnosis is 30 or more years', *Archives of Ophthalmology* **102**, 520–526.

Klein, R., Klein, B., Moss, S. E., Davis, M. D. & DeMets, D. L. (1984*b*), 'The Wisconsin Epidemiologic Study of Diabetic Retinopathy. III. Prevalence and risk of diabetes when age at diagnosis is 30 or more years', *Archives of Ophthalmology* **102**, 527–532.

Klein, R., Klein, B., Moss, S. E., Davis, M. D. & DeMets, D. L. (1989), 'The

WESDR.IX. Four year incidence and progression of diabetic retinopathy when age at diagnosis is less than 30 years', *Archives of Ophthalmology* **107**, 237–243.

Knight, K. & Fu, W. J. (2000), 'Asymptotics for Lasso-type estimators', *The Annals of Statistics* **28**, 1356–1378.

Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R. & Klein, B. (2000), 'Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV', *The Annals of Statistics* **28**, 1570–1600.

Lin, Y. (2002), 'Support vector machines and the Bayes rule in classification', *Data mining and knowledge discovery* **6**, 259–275.

Linhart, H. & Zucchini, W. (1986), *Model Selection*, New York: Wiley.

Murtagh, B. A. & Saunders, M. A. (1983), Minos 5.5 user's guide, Technical Report SOL 83-20R, OR Dept., Stanford University.

Ruppert, D. & Carroll, R. J. (2000), 'Spatially-adaptive penalties for spline fitting', *Australian and New Zealand Journal of Statistics* **45**, 205–223.

Tibshirani, R. J. (1996), 'Regression shrinkage and selection via the lasso', *Journal of Royal Statistical Society, B* **58**, 267–288.

Vapnik, V. N. (1995), *The Nature of Statistical Learning Theory*, Springer, New York.

Wahba, G. (1990), *Spline Models for Observational Data*, Vol. 59, SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics.

Wahba, G. & Wold, S. (1975), 'A completely automatic French curve', *Commun. Statist.* **4**, 1–17.

Wahba, G., Wang, Y., Gu, C., Klein, R. & Klein, B. (1995), 'Smoothing spline ANOVA for exponential families, with application to the WESDR', *The Annals of Statistics* **23**, 1865–1895.

Xiang, D. (1996), Model fitting and testing for non-Gaussian data with a large data set, Technical Report 957, Department of Statistics, University of Wisconsin, Madison, WI. Ph.D. thesis.

Xiang, D. & Wahba, G. (1996), 'A generalized approximate cross validation for smoothing splines with non-Gaussian data', *Statistica Sinica* **6**, 675–692.

Xiang, D. & Wahba, G. (1998), Approximate smoothing spline methods for large data sets in the binary case, *in* 'Proc. of the 1997 ASA Joint Statistical Meetings, Biometrics Section', pp. 94–98.

Yau, P., Kohn, R. & Wood, S. (2001), 'Bayesian variable selection and model averaging in high dimensional multinomial nonparametric regression'. To appear in *Journal of Computational and Graphical Statistics*.