Variable Selection and Model Building via Likelihood Basis Pursuit

Hao Helen Zhang, Grace Wahba, Yi Lin, Meta Voelker,
Michael Ferris, Ronald Klein, and Barbara Klein[1]
University of Wisconsin - Madison

Abstract This paper presents 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. 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 sets, its randomized version is proposed as well. A technique "slice modeling" is used to solve the optimization problem and makes the computation more efficient. LBP has great potential for a wide range of research and application areas such as medical studies, and in this paper we apply it to two large on-going epidemiologic studies: the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR) and the Beaver Dam Eye Study (BDES).
KEY WORDS: nonparametric variable selection; smoothing spline ANOVA; LASSO; generalized approximate cross validation; Monte Carlo bootstrap test; slice modeling.

1 Introduction

Variable selection, or dimension reduction, is fundamental to multivariate statistical model building. Not only does judicious variable selection improve the model's predictive ability, it generally provides a better understanding of the underlying concept that generates the data. Due to recent proliferation of large, high-dimensional databases, variable selection has become the focus of intensive research in several areas such as text processing, environmental sciences, and genomics, particularly gene expression array data which involves tens or hundreds of thousands of variables.

Traditional variable selection approaches such as stepwise selection and best subset selection are built in linear regression models, and the well-known criteria like Mallow's C_p, AIC and BIC are

[1]Hao Helen Zhang is now Assistant Professor, Department of Statistics, North Carolina State University. This work was supported in part by NSF grants DMS-0072292, DMS-0134987 and CCR-9972372, NIH grants EY09946 and EY03083, and AFOSR grant F49620-01-1-0040. The authors thank the editor, the associate editor, and the two referees for their constructive comments and suggestions that have led to significant improvement of this article.

often used to penalize the number of non-zero parameters. See Linhart & Zucchini (1986) for an introduction. To achieve better prediction and reduce the variances of estimators, many shrinkage estimation approaches have been proposed. 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 the ridge regression when $p = 2$. Due to the nature of the L_1 penalty, LASSO tends to shrink small coefficients to zero and hence gives concise models. It also exhibits the stability of ridge regression estimates. Fu (1998) made a thorough comparison between the bridge model and the 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. A related development is found in Bakin (1999). Gunn & Kandola (2002) proposed a structural modeling approach with sparse kernels. Recently Fan & Li (2001) suggested a non-concave penalized likelihood approach with the smoothly clipped absolute deviation (SCAD) penalty function, which resulted in an unbiased, sparse and continuous estimator. Our motivation of this study 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 method for variable selection in a nonparametric manner.

Smoothing spline analysis of variance (SS-ANOVA) provides a general framework for nonparametric multivariate function estimation and has been studied intensively for Gaussian data. Wahba, Wang, Gu, Klein & Klein (1995) gave a general setting for applying the SS-ANOVA model to exponential families. Gu (2002) provided a comprehensive review of the SS-ANOVA and some recent progress. In this work, we have developed a unified model which appropriately combines the SS-ANOVA model and basis pursuit for variable selection and model building. This article is organized as follows. Section 2 introduces the notation and illustrates the general structure of the likelihood basis pursuit (LBP) model. We focus on the main effects model and the two-factor interaction model. Then the models are generalized to incorporate categorical variables. Section 3 discusses the important issue of adaptively choosing regularization parameters. An extension of GACV proposed by Xiang & Wahba (1996) is derived as a tuning criterion. Section 4 proposes the measure of importance for the variables and, if desired, their interactions. A sequential Monte Carlo bootstrap test algorithm is developed to determine the selection threshold. Section 5 covers the numerical computation issue. Sections 6 through 8 present several simulation examples and the applications of LBP to two large epidemiologic studies. We carry out a data analysis for the four-year risk of progression of diabetic retinopathy in the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR) and for the five-year risk of mortality in the Beaver Dam Eye Study (BDES). The last

section contains some concluding remarks. Proofs are relegated to Appendix A and Appendix B.

2 Likelihood Basis Pursuit

2.1 Smoothing Spline ANOVA for Exponential Families

We are interested in estimating the dependence of Y on the covariates $\mathbf{X} = \big(X^{(1)}, ..., X^{(d)}\big)$. Typically $\mathbf{X}$ is in a high dimensional space $\mathcal{X} = \mathcal{X}^{(1)} \otimes \cdots \otimes \mathcal{X}^{(d)}$, where $\mathcal{X}^{(\alpha)}, \alpha = 1, ..., d$, is some measurable space and $\otimes$ denotes the tensor product operation. Conditioning on $\mathbf{x}$, assume Y is from an exponential family with the density of form $h(y|\mathbf{x}) = \exp[\{yf(\mathbf{x}) - b(f(\mathbf{x}))\}/a(\phi) + c(y, \phi)]$, where $a > 0$, b, and c are known functions, $f(\mathbf{x})$ is the parameter of interest dependent on $\mathbf{x}$, and ϕ is either known or a nuisance parameter independent of $\mathbf{x}$. We denote the observations of $Y_i|\mathbf{x}_i \sim h(y|\mathbf{x}_i)$, $i = 1, ..., n$, by the vector $\mathbf{y} = (y_1, ..., y_n)'$. The scaled conditional log-likelihood is

$$\mathcal{L}(\mathbf{y}, f) = \frac{1}{n}\sum_{i=1}^{n}\big[-l\{y_i, f(\mathbf{x}_i)\}\big] \equiv \frac{1}{n}\sum_{i=1}^{n}\big[-y_i f(\mathbf{x}_i) + b\{f(\mathbf{x}_i)\}\big]. \tag{2.1}$$

Though the methodology proposed here is general and valid for any exponential family, we use the Bernoulli case as our working example. In Bernoulli data, Y takes on values $\{0, 1\}$ with the conditional probability $p(\mathbf{x}) \equiv \text{prob}(Y = 1|\mathbf{X} = \mathbf{x})$. The logit function $f(\mathbf{x}) = \log\big(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\big)$, and the log-likelihood $l(y, f) = yf - \log(1 + e^f)$. Many parametric approaches, such as Tibshirani (1996), Fu (1998), and Fan & Li (2001), assume $f(\mathbf{x})$ to be a linear function of $\mathbf{x}$. Instead, we allow f to vary in a high-dimensional function space, which leads to a more flexible estimate for the target function. In this section and Section 2.2, we assume all the covariates are continuous. Later in Section 2.3, we take into account categorical variables. Similar to the classical analysis of variance (ANOVA), for any function $f(\mathbf{x}) = f(x^{(1)}, ..., x^{(d)})$ on a product domain $\mathcal{X}$, we can define its functional ANOVA decomposition as

$$f(\mathbf{x}) = b_0 + \sum_{\alpha=1}^{d} f_\alpha(x^{(\alpha)}) + \sum_{\alpha<\beta} f_{\alpha\beta}(x^{(\alpha)}, x^{(\beta)}) + \text{ all higher-order interactions}, \tag{2.2}$$

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. We are to estimate each f_α in an RKHS $\mathcal{H}^{(\alpha)}$, each $f_{\alpha\beta}$ in the tensor product space $\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}$, and so on. The full model space is then the dth-order tensor product space $\otimes_{\alpha=1}^{d}\mathcal{H}^{(\alpha)}$. Each functional component in the decomposition (2.2) falls in the corresponding subspace of $\otimes_{\alpha=1}^{d}\mathcal{H}^{(\alpha)}$.

For any continuous covariate $x^{(\alpha)}$, we scale it onto $[0,1]$ and choose $\mathcal{H}^{(\alpha)}$ to be the second-order Sobolev space $W_2[0,1]$, which is defined as $\{g : g(t), g'(t)$ are absolutely continuous, $g''(t) \in \mathcal{L}_2[0,1]\}$. When endowed with a certain inner product, $W_2[0,1]$ is an RKHS with the reproducing kernel (RK) $1 + K_0(s,t) + K_1(s,t)$. Here $K_0(s,t) = k_1(s)k_1(t)$ and $K_1(s,t) = k_2(s)k_2(t) - k_4(|s-t|)$, with $k_1(t) = t - \frac{1}{2}$, $k_2(t) = \frac{1}{2}(k_1^2(t) - \frac{1}{12})$, and $k_4(t) = \frac{1}{24}(k_1^4(t) - \frac{1}{2}k_1^2(t) + \frac{7}{240})$. Notice that $\mathcal{H}^{(\alpha)} = [1] \oplus \mathcal{H}_\pi^{(\alpha)} \oplus \mathcal{H}_s^{(\alpha)}$, with $[1]$ being the constant subspace, $\mathcal{H}_\pi^{(\alpha)}$ the "parametric" subspace generated by K_0 consisting of linear functions, and $\mathcal{H}_s^{(\alpha)}$ the "nonparametric" subspace generated by K_1 consisting of smooth functions. The reproducing kernel of $\otimes_{\alpha=1}^d \mathcal{H}^{(\alpha)}$ is

$$\prod_{\alpha=1}^{d} \big(1 + k_1(s^{(\alpha)})k_1(t^{(\alpha)}) + K_1(s^{(\alpha)}, t^{(\alpha)})\big). \tag{2.3}$$

Correspondingly, $\otimes_{\alpha=1}^d \mathcal{H}^{(\alpha)}$ can be decomposed into the tensor sum of parametric main effect subspaces, smooth main effect subspaces, two-factor interaction subspaces of three possible forms: parametric $\otimes$ parametric, smooth $\otimes$ parametric, and smooth $\otimes$ smooth, and similarly for three-factor or higher interaction subspaces. The RK's for these subspaces are given by the corresponding terms in the expansion of (2.3) and, the terms in (2.2) can be expanded in terms corresponding to those RK's in the expansion of (2.3). Therefore our model encompasses the linear model as a special case. See Wahba (1990) for more details. In various situations the ANOVA decomposition in (2.2) and in the expansion of (2.3) is truncated at some point. In this work we will consider truncation for the continuous variables no later than after the two-factor interactions. The remaining RK's will be used to construct an overcomplete set of basis functions for the likelihood basis pursuit, via details given below.

2.2 Likelihood Basis Pursuit

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 wavelet regression. In this paper we will apply basis pursuit to the negative log likelihood in the context of a dictionary based on the SS-ANOVA decomposition, and then select the important components using the multivariate function estimates. Let $\mathcal{H}$ be the model space after truncation. The variational problem for the likelihood basis pursuit (LBP) model is

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \big[-l(y_i, f(\mathbf{x}_i))\big] + J_\lambda(f). \tag{2.4}$$

$J_\lambda(f)$ denotes the l_1 norm of the coefficients of the basis functions in the representation of f. It is a generalization of the LASSO penalty in the context of nonparametric models. The l_1 penalty often produces coefficients that are exactly zero and therefore gives sparse solutions. This sparsity helps to distinguish important variables from unimportant ones easily and more effectively. See Tibshirani (1996) and Fan & Li (2001) for the comparison of the l_1 penalty with other forms of penalty. The regularization parameter λ balances the fitness in the likelihood and the penalty part.

For the usual smoothing spline modeling, the penalty J_λ is a quadratic norm or semi-norm in an RKHS. Kimeldorf & Wahba (1971) showed that the minimizer f_λ for the traditional smoothing spline model falls in a finite dimensional space, though the model space is of infinite dimension. For the penalized likelihood approach 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 span the model space and estimate the target function in that space. If all the data $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ are used to generate the bases, the resulting functional space demands intensive computation and the application is limited for large-scale problems. Thus we adopt the parsimonious bases approach used by Xiang & Wahba (1998), Ruppert & Carroll (2000), Lin, Wahba, Xiang, Gao, Klein & Klein (2000), and Yau et al. (2001). It has been shown by Gu & Kim (2001) 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 $\{\mathbf{x}_{1*}, ..., \mathbf{x}_{N*}\}$ from the original data and use them to generate basis functions which then span the model space $\mathcal{H}_*$. Notice that we are not wasting any data resource here, since all the data points are used for model fitting, though only a subset of them are selected to generate basis functions.

The issue of choosing N and the subsamples is important. In practice, we generally start with a reasonably large N. It is well known that "reasonably large" is not actually very large. See Lin et al. (2000). In principle, the subspace spanned by the chosen basis terms needs to be rich enough to provide a decent fit to the true curve. In this paper we use the simple random sampling technique to choose the subsamples. Alternatively, a cluster algorithm may be used, such as in Xiang & Wahba (1998) and Yau et al. (2001). Its basic idea involves two steps. First we group the data into N clusters which have maximum separation, by some good algorithm. 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.

2.2.1 Main Effects Model

The main effects model, also known as the additive model, is a sum of d functions of one variable. The function space is the tensor sum of constant and the main effect subspaces of $\otimes_{\alpha=1}^{d} \mathcal{H}^{(\alpha)}$.

Define $\mathcal{H}_* = [1] \oplus_{\alpha=1}^d \text{span}\big\{k_1(x^{(\alpha)}), K_1\big(x^{(\alpha)}, x_{j*}^{(\alpha)}\big), j = 1, ..., N\big\} \equiv [1] \oplus_{\alpha=1}^d \mathcal{H}_*^{(\alpha)}$, where $k_1(\cdot)$ and $K_1(\cdot,\cdot)$ are previously defined in Section 2.1. Then any component function $f_\alpha \in \mathcal{H}_*^{(\alpha)}$ has the representation $f_\alpha(x^{(\alpha)}) = b_\alpha k_1(x^{(\alpha)}) + \sum_{j=1}^N c_{\alpha,j} K_1\big(x^{(\alpha)}, x_{j*}^{(\alpha)}\big)$. The likelihood basis pursuit estimate $f \in \mathcal{H}_*$ is obtained by minimizing

$$\frac{1}{n}\sum_{i=1}^n \big[-l(y_i, f(\mathbf{x}_i))\big] + \lambda_\pi \sum_{\alpha=1}^d |b_\alpha| + \lambda_s \sum_{\alpha=1}^d \sum_{j=1}^N |c_{\alpha,j}|, \tag{2.5}$$

where $f(\mathbf{x}) = b_0 + \sum_{\alpha=1}^d f_\alpha(x^{(\alpha)})$. (λ_π, λ_s) are the regularization parameters. Here and in the sequel we have chosen to group terms of similar types (here "parametric" and "smooth") and to allow distinct λ's for different groups. Of course, we could choose to set $\lambda_\pi = \lambda_s$. Note that $\lambda_s = \infty$ yields a parametric model. Alternatively, we could choose different λ's for each coefficient if we decide to incorporate some prior knowledge of certain variables or their effects.

2.2.2 Two-factor Interaction Model

Two-factor interactions arise in many practical problems. See Hastie & Tibshirani (1990) Section 9.5.5. or Lin et al. (2000) Figures 9 and 10, for interpretable plots of two-factor interactions with continuous variables. In the LBP 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 and $\frac{d(d-1)}{2}$ parametric-parametric interaction terms The smooth part is the tensor sum of the subspaces generated by smooth main effect terms, parametric-smooth interaction terms, and smooth-smooth interaction terms. For each pair $\alpha \neq \beta$, the two-factor interaction subspace is $\mathcal{H}_*^{(\alpha\beta)} = \text{span}\{k_1(x^{(\alpha)})k_1(x^{(\beta)}), K_1\big(x^{(\alpha)}, x_{j*}^{(\alpha)}\big)k_1(x^{(\beta)})k_1(x_{j*}^{(\beta)}), K_1\big(x^{(\alpha)}, x_{j*}^{(\alpha)}\big)K_1\big(x^{(\beta)}, x_{j*}^{(\beta)}\big), j = 1, ...N\}$, and the interaction term $f_{\alpha\beta}(x^{(\alpha)}, x^{(\beta)})$ has the representation

$$f_{\alpha\beta} = b_{\alpha\beta}k_1(x^{(\alpha)})k_1(x^{(\beta)}) + \sum_{j=1}^N c_{\alpha\beta,j}^{\pi s} K_1\big(x^{(\alpha)}, x_{j*}^{(\alpha)}\big)k_1(x^{(\beta)})k_1(x_{j*}^{(\beta)}) + \sum_{j=1}^N c_{\alpha\beta,j}^{ss} K_1\big(x^{(\alpha)}, x_{j*}^{(\alpha)}\big)K_1\big(x^{(\beta)}, x_{j*}^{(\beta)}\big).$$

The whole function space is $\mathcal{H}_* \equiv [1] \oplus_{\alpha=1}^d \mathcal{H}_*^{(\alpha)} + \oplus_{\beta<\alpha} \mathcal{H}_*^{(\alpha\beta)}$. The LBP optimization problem is

$$\begin{aligned}\min_{f\in\mathcal{H}_*} \frac{1}{n}\sum_{i=1}^n \big[-l(y_i, f(\mathbf{x}_i))\big] + \lambda_\pi\Big(\sum_{\alpha=1}^d |b_\alpha|\Big) + \lambda_{\pi\pi}\Big(\sum_{\alpha<\beta} |b_{\alpha\beta}|\Big) \\ + \lambda_{\pi s}\Big(\sum_{\alpha\neq\beta}\sum_{j=1}^N |c_{\alpha\beta,j}^{\pi s}|\Big) + \lambda_s\Big(\sum_{\alpha=1}^d\sum_{j=1}^N |c_{\alpha,j}|\Big) + \lambda_{ss}\Big(\sum_{\alpha<\beta}\sum_{j=1}^N |c_{\alpha\beta,j}^{ss}|\Big).\end{aligned} \tag{2.6}$$

Note different penalties are allowed for the five different types of terms.

2.3 Incorporating Categorical Variables

In real applications, some covariates may be categorical such as sex, race, and smoking history in many medical studies. In previous sections, we proposed the main effects model (2.5) and the two-factor interaction model (2.6) for continuous variables only. Now we generalize these models to incorporate categorical variables, which are denoted by $\mathbf{Z} = (Z^{(1)}, ..., Z^{(r)})$. For simplicity, we assume each $Z^{(\gamma)}$ has two categories $\{T, F\}$ for $\gamma = 1, ..., r$. The following idea is easily extended to the situation when some variables have more than two categories. Define the mapping Φ_γ by

$$\begin{aligned} \Phi_\gamma(z^{(\gamma)}) &= \frac{1}{2} \quad \text{if} \quad z^{(\gamma)} = \mathrm{T} \\ &= -\frac{1}{2} \quad \text{if} \quad z^{(\gamma)} = \mathrm{F}. \end{aligned}$$

Generally 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.

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

$$\frac{1}{n}\sum_{i=1}^{n}[-l(y_i, f(\mathbf{x}_i, \mathbf{z}_i)] + \lambda_\pi(\sum_{\alpha=1}^{d}|b_\alpha| + \sum_{\gamma=1}^{r}|B_\gamma|) + \lambda_s\sum_{\alpha=1}^{d}\sum_{j=1}^{N}|c_{\alpha,j}|, \tag{2.7}$$

  where $f(\mathbf{x}, \mathbf{z}) = b_0 + \sum_{\alpha=1}^{d} b_\alpha k_1(x^{(\alpha)}) + \sum_{\gamma=1}^{r} B_\gamma \Phi_\gamma(z^{(\gamma)}) + \sum_{\alpha=1}^{d}\sum_{j=1}^{N} c_{\alpha,j} K_1\big(x^{(\alpha)}, x_{j*}^{(\alpha)}\big)$. For each γ, the function Φ_γ can be regarded as the main effect of the covariate $Z^{(\gamma)}$. Thus we choose to associate the coefficients $|B|$'s and $|b|$'s with the same parameter λ_π.

- Adding two-factor interactions with categorical variables to a model that already includes parametric and smooth terms adds a number of additional terms to the general model. Compared with (2.6), four new types of terms are involved when we take into account categorical variables. They are categorical main effects, categorical-categorical interactions, "parametric continuous"-categorical interactions, and "smooth continuous"-categorical interactions. The modified two-factor interaction model is: minimize

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}[-l(y_i, f(\mathbf{x}_i, \mathbf{z}_i))] + \lambda_\pi(\sum_{\alpha=1}^{d}|b_\alpha| + \sum_{\gamma=1}^{r}|B_\gamma|) + \lambda_{\pi\pi}(\sum_{\alpha<\beta}|b_{\alpha\beta}| + \sum_{\gamma<\theta}|B_{\gamma\theta}| + \sum_{\alpha=1}^{d}\sum_{\gamma=1}^{r}|P_{\alpha\gamma}|) \\ &\quad + \lambda_{\pi s}(\sum_{\alpha\neq\beta}\sum_{j=1}^{N}|c^{\pi s}_{\alpha\beta,j}| + \sum_{\alpha=1}^{d}\sum_{\gamma=1}^{r}\sum_{j=1}^{N}|c^{\pi s}_{\alpha\gamma,j}|) + \lambda_s(\sum_{\alpha=1}^{d}\sum_{j=1}^{N}|c_{\alpha,j}|) + \lambda_{ss}(\sum_{\alpha<\beta}\sum_{j=1}^{N}|c^{ss}_{\alpha\beta,j}|), \end{aligned} \tag{2.8}$$

where

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

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

3 Generalized Approximate Cross Validation (GACV)

Regularization parameter selection has been a very active research field. The ordinary cross validation (OCV) (Wahba & Wold (1975)), the generalized cross validation (GCV) (Craven & Wahba (1979)), and the generalized approximate cross validation (GACV) (Xiang & Wahba (1996)) are widely used in various contexts of smoothing spline models. We will derive the GACV to select λ's in the LBP models. Here we present the GACV score and its derivation only for the main effects model, and the extension to high-order interaction models is straightforward. 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 $\lambda = (\lambda_\pi, \lambda_{\pi\pi}, \lambda_{\pi s}, \lambda_s, \lambda_{ss})$ for the two-factor interaction model.

3.1 Generalized Approximate Cross Validation (GACV)

Let p be the "true" but unknown probability function and p_λ be its estimate associated with λ. Similarly, f and μ are respectively the true logit and mean functions, and f_λ and μ_λ are the corresponding estimates. Kullback-Leibler (KL) distance, also known as the relative entropy, is often used to measure the distance between two probability distributions. For Bernoulli data, we have $KL(p, p_\lambda) = E_{\mathbf{X}}\big[\frac{1}{2}\{\mu(f - f_\lambda) - (b(f) - b(f_\lambda))\}\big]$ with $b(f) = \log(1+e^f)$. Removing the quantity which does not depend on λ from the KL distance expression, we get the comparative KL (CKL)

distance $E_{\mathbf{X}}[-\mu f_\lambda + b(f_\lambda)]$. The ordinary leaving-out-one cross validation (CV) for CKL is

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} [-y_i f_\lambda^{[-i]}(\mathbf{x}_i) + b(f_\lambda(\mathbf{x}_i))], \tag{3.1}$$

where $f_\lambda^{[-i]}$ is the minimizer of the objective function with the ith data point omitted. CV is commonly used as a roughly unbiased estimate for CKL. See Xiang & Wahba (1996) and Gu (2002). Direct calculation of CV involves computing n leaving-out-one estimates, which is expensive and almost infeasible for large scale problems. Thus we will derive a second-order approximate cross validation (ACV) score to the CV. See Xiang & Wahba (1996), Lin et al. (2000), and Gao, Wahba, Klein & Klein (2001) for the ACV in traditional smoothing spline models. We firstly establish the leaving-out-one lemma for the LBP models. The proof of Lemma 1 is given in Appendix A.

Lemma 1: (Leaving-Out-One Lemma) Denote the objective function in the LBP model by

$$I_\lambda(f, \mathbf{y}) = \mathcal{L}(\mathbf{y}, f) + J_\lambda(f), \tag{3.2}$$

Let $f_\lambda^{[-i]}$ be the minimizer of $I_\lambda(f, \mathbf{y})$ with the ith observation omitted and $\mu_\lambda^{[-i]}(\cdot)$ be the mean function corresponding to $f_\lambda^{[-i]}(\cdot)$. For any $v \in R$, we define the vector $\mathbf{V} = (y_1, ..., y_{i-1}, v, y_{i+1}, \ldots, y_n)'$. Let $h_\lambda(i, v, \cdot)$ be the minimizer of $I_\lambda(f, \mathbf{V})$, then $h_\lambda(i, \mu_\lambda^{[-i]}(\mathbf{x}_i), \cdot) = f_\lambda^{[-i]}(\cdot)$.

Using Taylor series approximations and Lemma 1, we can derive (in Appendix B) the ACV score

$$ACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \big[- y_i f_\lambda(\mathbf{x}_i) + b(f_\lambda(\mathbf{x}_i))\big] + \frac{1}{n}\sum_{i=1}^{n} h_{ii}\frac{y_i(y_i - \mu_\lambda(\mathbf{x}_i))}{1 - \sigma_{\lambda i}^2 h_{ii}}, \tag{3.3}$$

where $\sigma_{\lambda i}^2 \equiv p_\lambda(\mathbf{x}_i)(1 - p_\lambda(\mathbf{x}_i))$ and h_{ii} is the ii-th entry of the matrix H defined in the equation (B.9). Let W be the diagonal matrix with $\sigma_{\lambda i}^2$ in the ii-th position. By replacing h_{ii} with $\frac{1}{n}\sum_{i=1}^n h_{ii} \equiv \frac{1}{n} tr(H)$ and replacing $1 - \sigma_{\lambda i}^2 h_{ii}$ with $\frac{1}{n} tr[I - (W^{1/2} H W^{1/2})]$ in (3.3), we obtain the generalized approximate cross validation (GACV) score

$$GACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \big[- y_i f_\lambda(\mathbf{x}_i) + b(f_\lambda(\mathbf{x}_i))\big] + \frac{tr(H)}{n}\frac{\sum_{i=1}^n y_i(y_i - \mu_\lambda(\mathbf{x}_i))}{tr[I - W^{1/2} H W^{1/2}]}. \tag{3.4}$$

3.2 Randomized GACV

Direct computation of (3.4) involves the inversion of a large scale matrix, whose size depends on the sample size n, basis size N, and dimension d. Large values of n, N, or d may make the computation expensive and produce unstable solutions. Thus the randomized GACV (ranGACV)

score is proposed as a computable proxy for GACV. Essentially, 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, (which has been exploited by numerous authors, see e.g. Girard (1998)).

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

Let $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$ be a zero mean random n-vector of independent components with variance σ_ϵ^2. Let $f_\lambda^{\mathbf{y}}$ and $f_\lambda^{\mathbf{y}+\epsilon}$ be respectively the minimizer of (2.5) using the original data $\mathbf{y}$ and the perturbed data $\mathbf{y} + \epsilon$. Using the derivation procedure in Lin et al. (2000), we get the ranGACV score for the LBP estimates

$$ranGACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\big[-y_i f_\lambda(\mathbf{x}_i) + b(f_\lambda(\mathbf{x}_i))\big] + \frac{\epsilon^T(f_\lambda^{\mathbf{y}+\epsilon} - f_\lambda^{\mathbf{y}})}{n}\frac{\sum_{i=1}^{n} y_i(y_i - \mu_\lambda(\mathbf{x}_i))}{\epsilon^T\epsilon - \epsilon^T W(f_\lambda^{\mathbf{y}+\epsilon} - f_\lambda^{\mathbf{y}})}. \quad (3.5)$$

In addition, two ideas help to reduce the variance of the second term in (3.5). (1) Choose ϵ as Bernoulli (0.5) taking values in $\{+\sigma_\epsilon, -\sigma_\epsilon\}$. This guarantees that the randomized trace estimate has the minimal variance given a fixed σ_ϵ^2. See Hutchinson (1989). (2) Generate U independent perturbations $\epsilon^{(u)}$, $u = 1, \ldots, U$, and compute U replicated ranGACVs. Their average is then used to compute the GACV estimate.

4 Selection Criteria for Main Effects and Two-Factor Interactions

4.1 The L_1 Importance Measure

After choosing the optimal $\hat{\lambda}$ by the GACV or ranGACV criteria, the LBP estimate $f_{\hat{\lambda}}$ is obtained by minimizing (2.5), (2.6), (2.7) or (2.8). How to measure the importance of a particular component in the fitted model is a key question. We propose using the functional L_1 norm as the importance measure. The sparsity in the solutions will help distinguish the significant terms from insignificant ones effectively, and thus improve the performance of our importance measure. In practice for each functional component, its L_1 norm is empirically calculated as the average of the function values evaluated at all the data points. For instance, $L_1(f_\alpha) = \frac{1}{n}\sum_{i=1}^{n}|f_\alpha(x_i^{(\alpha)})|$ and $L_1(f_{\alpha\beta}) = \frac{1}{n}\sum_{i=1}^{n}|f_{\alpha\beta}(x_i^{(\alpha)}, x_i^{(\beta)})|$. For the categorical variables in the model (2.7), the empirical L_1 norm of the main effect f_γ is $L_1(f_\gamma) = \frac{1}{n}\sum_{i=1}^{n}|B_\gamma\Phi_\gamma(z_i^{(\gamma)})|$ for $\gamma = 1, ..., r$. The norms of the interaction terms involved with categorical variables are defined similarly. The rank of the L_1 norm scores is used to rank the relative importance of functional components. For instance, the component with the largest L_1 norm is ranked as the most important one, and any component with zero or tiny L_1

norm is ranked as an unimportant one. We have also tried using the functional L_2 norm to rank the components, and this gave almost identical results in terms of the set of variables selected in numerous simulation studies (not reproduced here).

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 $\hat{f}_1, \ldots, \hat{f}_d$ and calculate their L_1 norms $L_1(\hat{f}_1), \ldots, L_1(\hat{f}_d)$. We will use a sequential procedure to select important terms. Denote the decreasingly ordered norms as $\hat{L}_{(1)}, \ldots, \hat{L}_{(d)}$ 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".

Now we develop a sequential Monte Carlo bootstrap test procedure to determine q. 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 $1 \leq \eta \leq d-1$, the null model is $f = \hat{b}_0 + \hat{f}_{(1)} + \cdots + \hat{f}_{(\eta)}$ and the hypotheses are $H_0 : L_{(\eta+1)} = 0$ 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. Conditional on the original covariates $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$, we generate $\{y_1^{*(\eta)}, ..., y_n^{*(\eta)}\}$ (responses 0 or 1) by using the null model $f = \hat{b}_0 + \hat{f}_{(1)} + \cdots + \hat{f}_{(\eta)}$ as the true logit function. Totally we sample T independent sets of data $(\mathbf{x}_1, y_{1,t}^{*(\eta)}), ..., (\mathbf{x}_n, y_{n,t}^{*(\eta)}), t = 1, ..., T$, from the null model f, fit the main effects model for each set, and compute $\hat{L}_t^{*(\eta+1)}, t = 1, ..., T$. If exactly k of the simulated $\hat{L}^{*(\eta+1)}$ values exceed $\hat{L}_{(\eta+1)}$ and none equals 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 $f = \hat{b}_0$. Test $H_0 : L_{(1)} = 0$ vs $H_1 : L_{(1)} > 0$. Generate T independent sets of data $(\mathbf{x}_1, y_{1,t}^{*(0)}), ..., (\mathbf{x}_n, y_{n,t}^{*(0)})$, $t = 1, ..., T$, 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 as any number slightly larger than $\hat{L}_{(1)}$.

Step 2: Let $\eta = \eta + 1$ and $f = \hat{b}_0 + \hat{f}_{(1)} + \cdots + \hat{f}_{(\eta)}$. Test $H_0 : L_{(\eta+1)} = 0$ vs $H_1 : L_{(\eta+1)} > 0$. Generate T independent sets of data $(\mathbf{x}_1, y_{1,t}^{*(\eta)}), \ldots, (\mathbf{x}_n, y_{n,t}^{*(\eta)})$ 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 $q = \hat{L}_{(d)}$.

The order of entry for sequential testing of the terms being entertained for the model is determined by the magnitude of the component L_1 norms. There are other reasonable ways to determine the order of entry and the particular strategy employed can affect the results, as is the case for any stepwise procedure. For the LBP approach, the relative ranking among the important terms usually does not affect the final component selection solution as long as the important terms are all ranked higher than the unimportant terms. Thus our procedure can usually distinguish between important and unimportant terms, as in most of our examples. When the distinction between important and unimportant terms is ambiguous with our method, the ambiguous terms can be recognized and further investigation will be needed on these terms.

5 Numerical Computation

Since the objective function in either (2.5) or (2.6) is not differentiable with respect to the coefficients b's and c's, some numerical methods for optimization fail to solve this kind of problem. We can replace the l_1 norms in the objective function by non-negative variables constrained linearly to be the corresponding absolute values using standard mathematical programming techniques, and then solve a series of programs with nonlinear objective functions and linear constraints. Many optimization methods can solve such problems. We use MINOS (Murtagh & Saunders (1983)) since it generally performs well with the linearly constrained models and returns consistent results.

Consider choosing the optimal λ by selecting the grid point for which ranGACV achieves a minimum. For each λ, to find $ranGACV$ the program (2.5), (2.6), (2.7), or (2.8) must be solved twice — once with $\mathbf{y}$ (the original problem) and once with $\mathbf{y} + \epsilon$ (the perturbed problem). This often results in hundreds or thousands of individual solves, depending upon the range for λ. In order to obtain solutions in a reasonable amount of time, we employ an efficient solving approach, namely slice modeling (see Ferris & Voelker (2000) and Ferris & Voelker (2001)). Slice modeling is an approach for solving a series of mathematical programs with the same structure but different data. Since for LBP we can consider the values $(\lambda, \mathbf{y})$ to be individual slices of data, as only these values change between solves, the program can be reduced to an example of nonlinear slice

modeling. By applying slice modeling ideas, namely maintaining common program structure and "core" data shared between solves and using previous solutions as starting points for late solves, we can improve efficiency and make the grid search feasible. We have developed efficient code for the main effect and two-way interaction LBP models. The code is easy to use and runs fast. For example, in one simulation example with $n = 1000$, $d = 10$, and λ fixed, the main effect model takes less than two seconds; and the two-way interaction model takes less than fifty seconds.

6 Simulation

6.1 Simulation 1: Main Effects Model

In this example, there are altogether $d = 10$ covariates: $X^{(1)}, ..., X^{(10)}$. They are taken to be uniformly distributed in $[0, 1]$ independently. The sample size $n = 1000$. We use the simple random sub-sampling technique to select $N = 50$ basis functions. The perturbation ϵ is distributed as Bernoulli(0.5) taking two values $\{+0.25, -0.25\}$. Four variables $X^{(1)}, X^{(3)}, X^{(6)}$ and $X^{(8)}$ are important, and the others are noise variables. The true conditional logit function is

$$f(\mathbf{x}) = \frac{4}{3}x^{(1)} + \pi sin(\pi x^{(3)}) + 8\left(x^{(6)}\right)^5 + \frac{2}{e-1}e^{x^{(8)}} - 5.$$

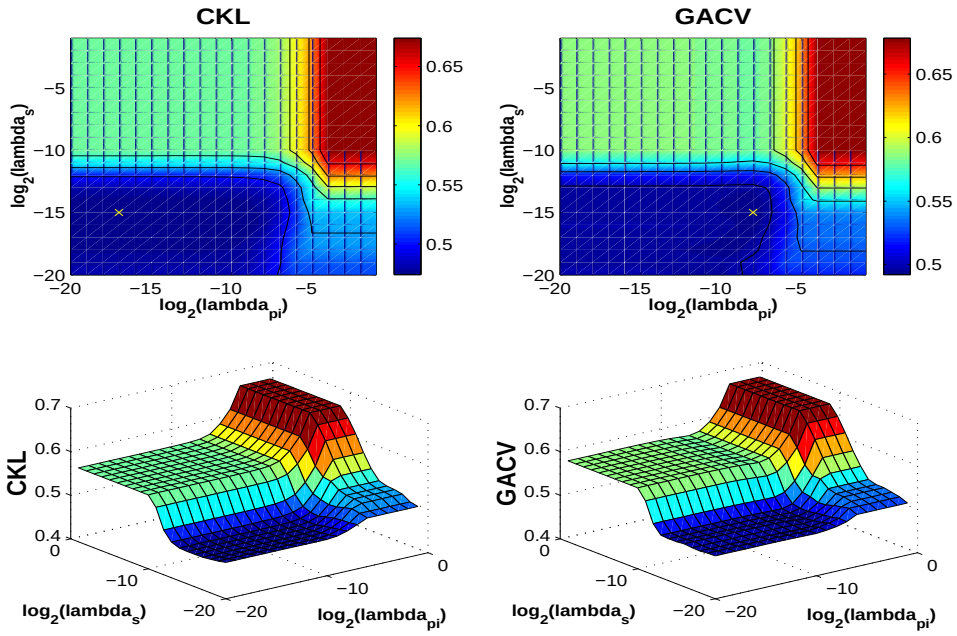

Figure 1: Contours and three-dimensional plots for CKL(λ) and GACV(λ).

We fit the main effects LBP model and search the parameters (λ_π, λ_s) globally. Since the true f is known, both CKL and ranGACV can be used for choosing the λ's. 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 $ranGACV(\lambda)$, where the white cross "x" indicates the location of the optimal regularization parameter. Here $\hat{\lambda}_{CKL} = (2^{-17}, 2^{-15})$ 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.

Using the optimal parameters we fit the main effects model and calculate the L_1 norm scores for the individual components $\hat{f}_1, \ldots, \hat{f}_{10}$. Figure 2 plots two sets of L_1 norm scores, obtained respectively using $\hat{\lambda}_{CKL}$ and $\hat{\lambda}_{ranGACV}$, in their decreasing order. The dashed line indicates the threshold 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.

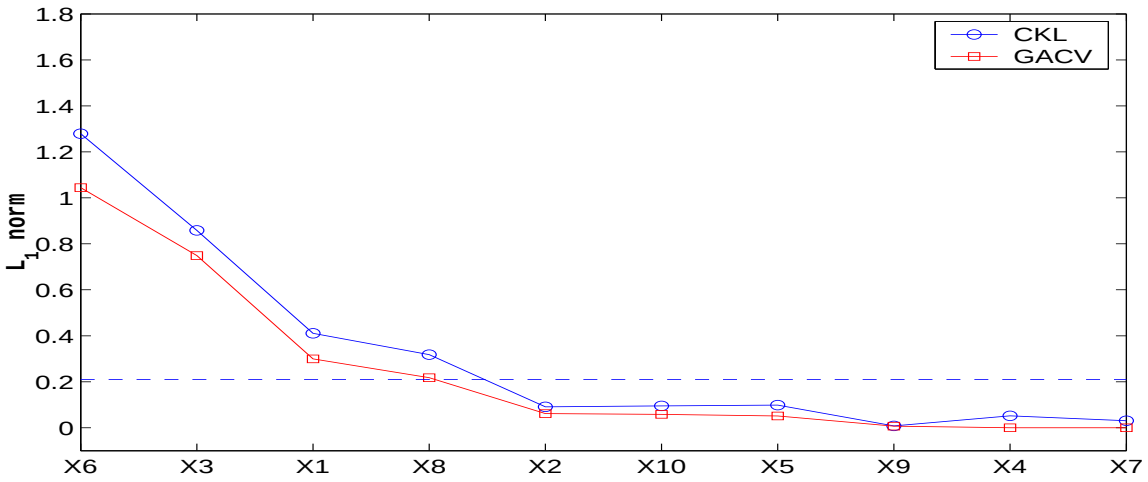

Figure 2: L_1 norm scores for the main effects model.

The procedure of the sequential bootstrap tests to determine 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 of Figure 3, grey circles denote the L_1 norms for the variables in the null model, and black circles denote the L_1 norms for those not in the null model. (In a color plot, grey circles are shown in green and black circles are shown in blue). Along the horizontal axis, the variable being tested for importance is bracketed by a pair of *. Our experiment shows that 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 the fifth component f_2 is accepted with the p-value $10/51 \doteq 0.20$. Thus f_6, f_3, f_1 and f_8 are selected as "important"

components and $q = L_{(4)} = 0.21$.

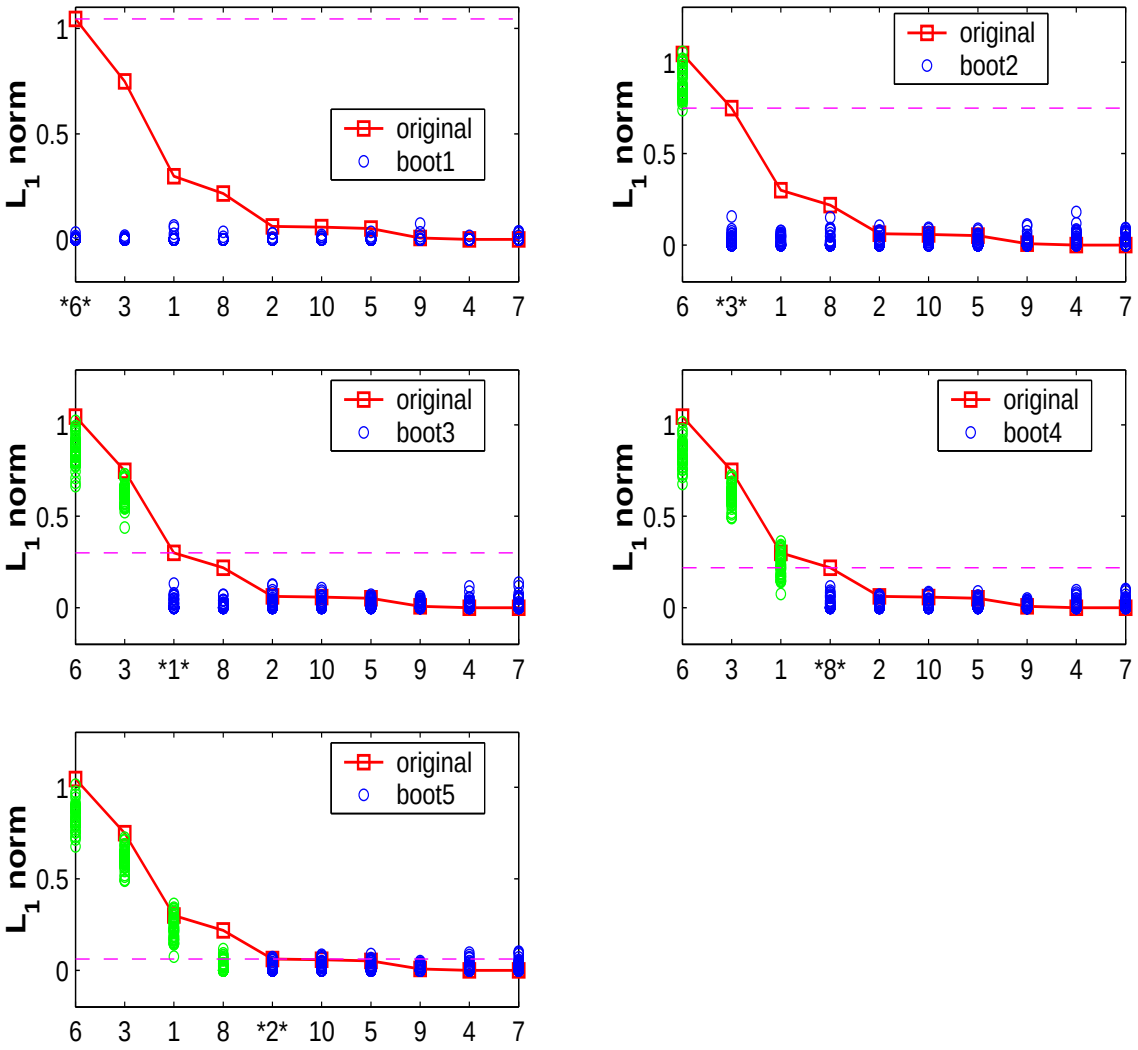

Figure 3: Monte Carlo bootstrap tests for Simulation 1.

In addition to selecting 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 fitted using $\hat{\lambda}_{ranGACV}$. In each panel, the solid line denotes the true curve and the dashed line denotes the corresponding estimate. In general, the fitted main effects model provides a reasonably good estimate for each important component. Altogether we generated 20 datasets and fitted the main effects model for each dataset with regularization parameters tuned separately. Throughout all these 20 runs, variables $X^{(1)}, X^{(3)}, X^{(6)}$ and $X^{(8)}$ are always the four top-ranked variables. The results and figures shown above are based on the first dataset.

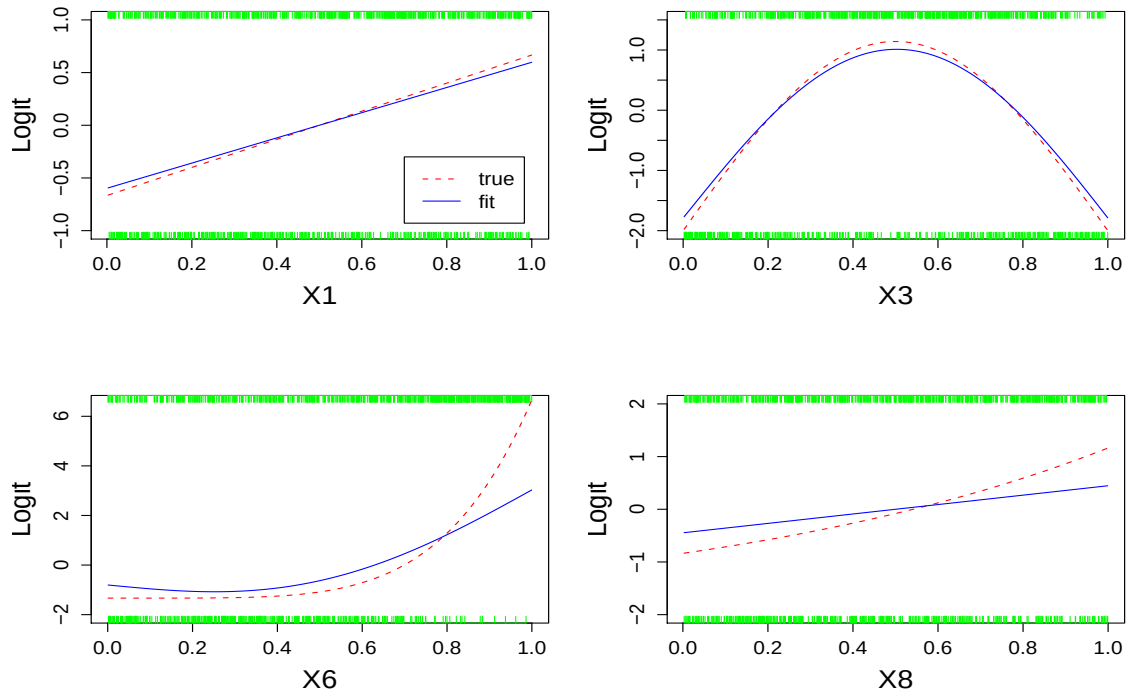

Figure 4: True and estimated univariate logit component.

6.2 Simulation 2: Two-factor Interaction Model

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 their interaction. The true logit function f is

$$f(\mathbf{x}) = 4x^{(1)} + \pi sin\left(\pi x^{(1)}\right) + 6x^{(2)} - 8\left(x^{(2)}\right)^3 + \cos\left(2\pi(x^{(1)} - x^{(2)})\right) - 4.$$

We choose $n = 1000, N = 50$, and use the same perturbation ϵ as in the previous example. There are five tuning parameters $(\lambda_\pi, \lambda_{\pi\pi}, \lambda_s, \lambda_{\pi s}, \lambda_{ss})$ in the two-factor interaction model. In practice, extra constraints may be added on the parameters for different needs. Here we penalize all the two-factor interaction terms equally by setting $\lambda_{\pi\pi} = \lambda_{\pi s} = \lambda_{ss}$. The optimal parameters are $\hat{\lambda}_{CKL} = (2^{-10}, 2^{-10}, 2^{-15}, 2^{-10}, 2^{-10})$ and $\hat{\lambda}_{ranGACV} = (2^{-20}, 2^{-20}, 2^{-18}, 2^{-20}, 2^{-20})$. The ranked L_1 norm scores are plotted in Figure 5. The dashed line is the threshold q chosen by the Monte Carlo bootstrap test procedure. The LBP two-factor interaction model, using either $\hat{\lambda}_{CKL}$ or $\hat{\lambda}_{GACV}$, selects all the important effects $X^{(1)}, X^{(2)}$, and their interaction effect correctly.

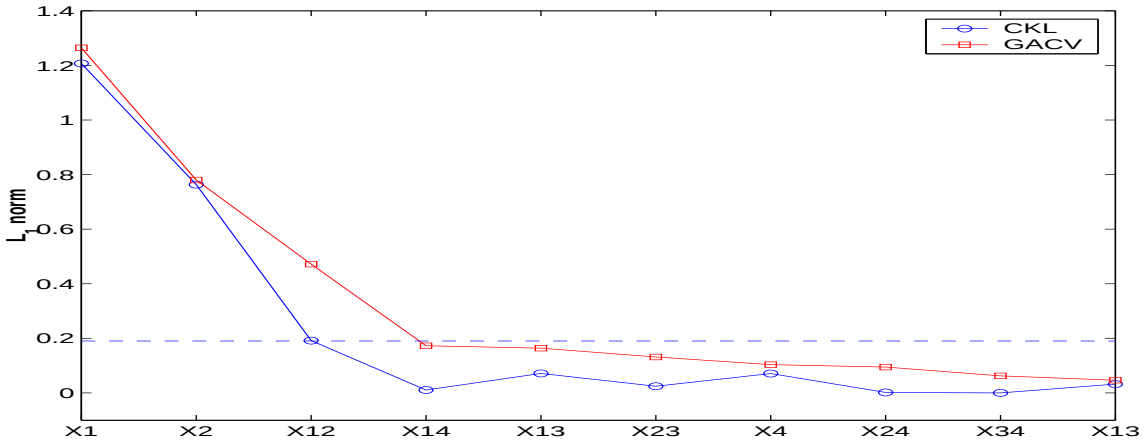

Figure 5: L_1 norm scores for the two-factor interaction model.

There is a strong interaction effect between variable $X^{(1)}$ and $X^{(2)}$, which is shown clearly by the cross section plots in Figure 6. Solid lines are the cross sections of the true logit function $f\left(x^{(1)}, x^{(2)}\right)$ at distinct values $x^{(1)} = 0.2, 0.5, 0.8$, while the dashed lines are their corresponding estimates given by the LBP model. The parameters are tuned by the ranGACV criterion.

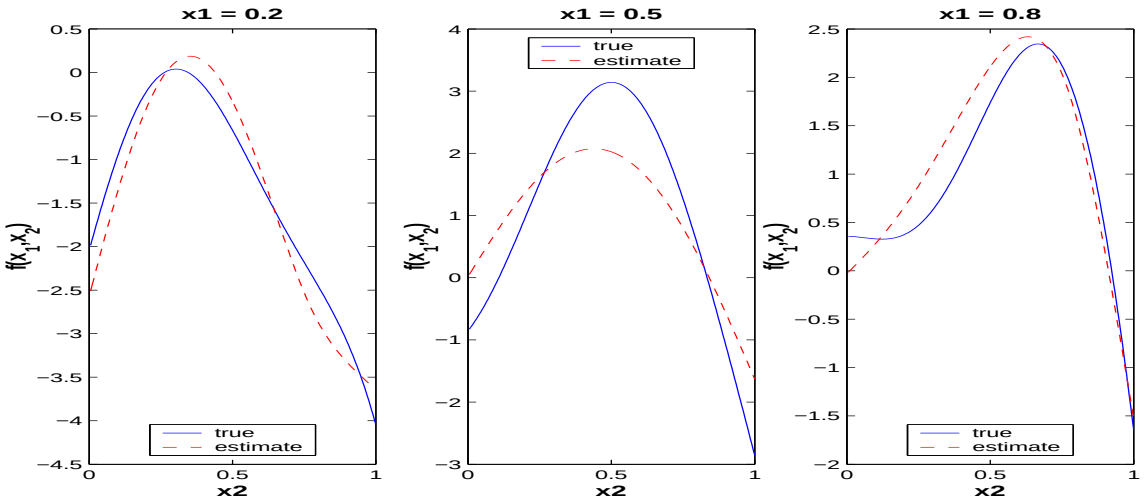

Figure 6: Cross section plots for the two-factor interaction model.

6.3 Simulation 3: Main Effects Model Incorporating Categorical Variables

In this example, among the 12 covariates $X^{(1)}, ..., X^{(10)}$ are continuous and $Z^{(1)}, Z^{(2)}$ are categorical. The continuous variables are uniformly distributed in $[0, 1]$ and the categorical variables are

Bernoulli(0.5) with values $\{0, 1\}$. The true logit function is

$$f(\mathbf{x}) = \frac{4}{3}x^{(1)} + \pi sin\big(\pi x^{(3)}\big) + 8\big(x^{(6)}\big)^5 + \frac{2}{e-1}e^{x^{(8)}} + 4z^{(1)} - 7.$$

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 in the previous examples. The main effects model incorporating categorical variables in (2.7) is fitted. Figure 7 plots the ranked L_1 norm scores for all the covariates. The LBP main effects models using $\hat{\lambda}_{CKL}$ and $\hat{\lambda}_{GACV}$ both select the important continuous and categorical variables correctly.

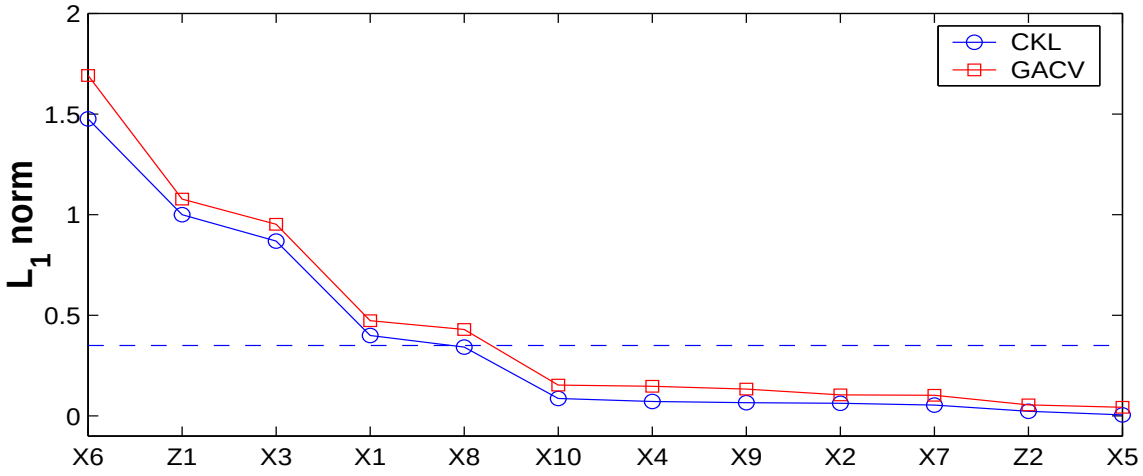

Figure 7: L_1 norm scores for the main effects model incorporating categorical variables.

7 Wisconsin Epidemiologic Study of Diabetic Retinopathy

The Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR) is an ongoing epidemiologic study of a cohort of patients receiving their medical care southern Wisconsin. 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 in 1980-82, and of those, 891 participated in the first follow-up examination. Details about the study can be found in Klein, Klein, Moss, Davis & DeMets (1984*a*), Klein, Klein, Moss, Davis & DeMets (1984*b*), Klein, Klein, Moss, Davis & DeMets (1989), Klein, Klein, Moss & Cruickshanks (1998) and elsewhere. A large number of medical, demographic, ocular, and other covariates were recorded in each examination.

A multilevel retinopathy score is assigned to each eye based on its stereoscopic color fundus photographs. This data set has been extensively analyzed using a variety of statistical methods, such as Craig, Fryback, Klein & Klein (1999), Kim (1995) and others.

In this section we examine the relation of a large number of possible risk factors at baseline to the four year progression of diabetic retinopathy. 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. Wahba et al. (1995) examined risk factors for progression of diabetic retinopathy on a subset of the younger onset group, whose members had no or non-proliferative retinopathy at 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 further in Appendix B). 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 dataset 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 prediction variables available was made, to reduce the set for examination to 14 potential risk factors. The continuous covariates are *dur, gly, bmi, sys, ret, pulse, ins, sch, iop*, and categorical covariates are *smk, sex, asp, famdb, mar.* The full names are in Appendix C. Since the true f is not known for real data, only ranGACV is available for tuning λ. Figure 8 plots the L_1 norm scores of the individual functional components in the fitted LBP main effects model. The dashed line indicates the threshold $q = 0.39$, which is chosen by the sequential bootstrap tests.

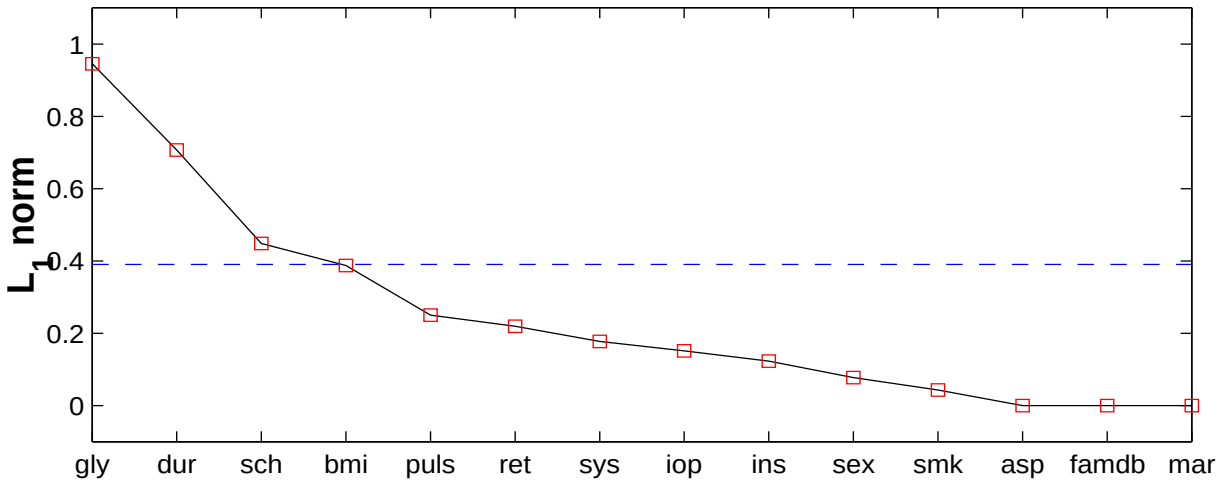

Figure 8: L_1 norm scores for the WESDR main effects model.

We note that the LBP picks out three most important variables *gly, dur*, and *bmi*, that appeared in Wahba et al. (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. The sequential Monte Carlo bootstrap 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 9 plots the estimated logit for *dur*. 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). The linear logistic regression model (using the function *glm* in *R* package) shows that *dur* is not significant at level $\alpha = 0.05$. The curve in Figure 9 exhibits a hilly shape, which means that a quadratic function fits the curve better than a linear function. We refit the linear logistic model by intentionally including dur^2, the term 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 (2.6) 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.

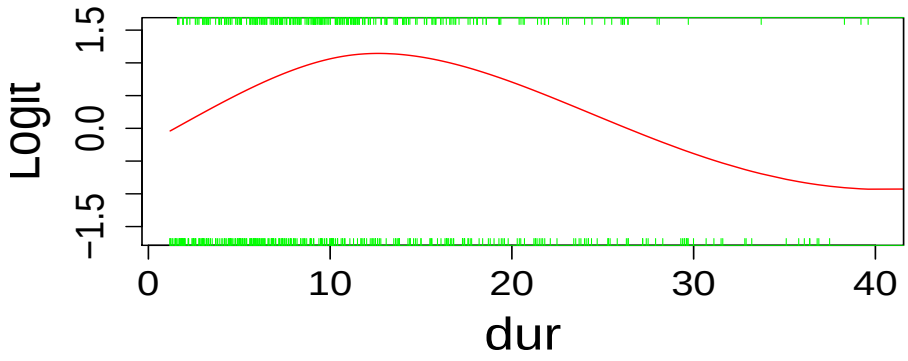

Figure 9: Estimated logit component for *dur*.

8 Beaver Dam Eye Study

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, 5924 eligible people (age 43-84 years) 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 & Chappell (2001).

We are interested in the relation between the five-year mortality occurrence 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 and the rest of the population differs. We consider 10 continuous and 8 categorical covariates, whose detailed information is given in Appendix D. The abbreviated names of continuous covariates are *pky, sch, inc, bmi, glu, cal, chl, hgb, sys, age*, and those of categorical covariates are *cv, sex, hair, hist, nout, mar, sum, vtm*. We deliberately take into account some "noisy" variables in the analysis, such as *hair, nout,* and *sum*, which are not directly related to mortality in general. Their inclusion is to show the performance of the LBP approach and they are not expected to be picked out eventually by the model. Y is assigned 1 if a person participated in the baseline examination and died prior to the start of the

first 5-year follow-up; Y is assigned 0 otherwise. There are 4422 non-diabetic study participants 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 335 subjects are not included in our analysis. This assumption is not necessarily valid, since age, blood pressure, body mass index, cholesterol, sex, smoking and 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.

The main effects model incorporating categorical variables in (2.7) is fitted. The sequential Monte Carlo bootstrap tests for 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 10 plots the L_1 norm scores for all the potential risk factors. Using the threshold (dashed line) 0.25 chosen by the sequential bootstrap test procedure, the LBP model identifies six important risk factors: *age, hgb, pky, sex, sys, cv* for the five-year mortality.

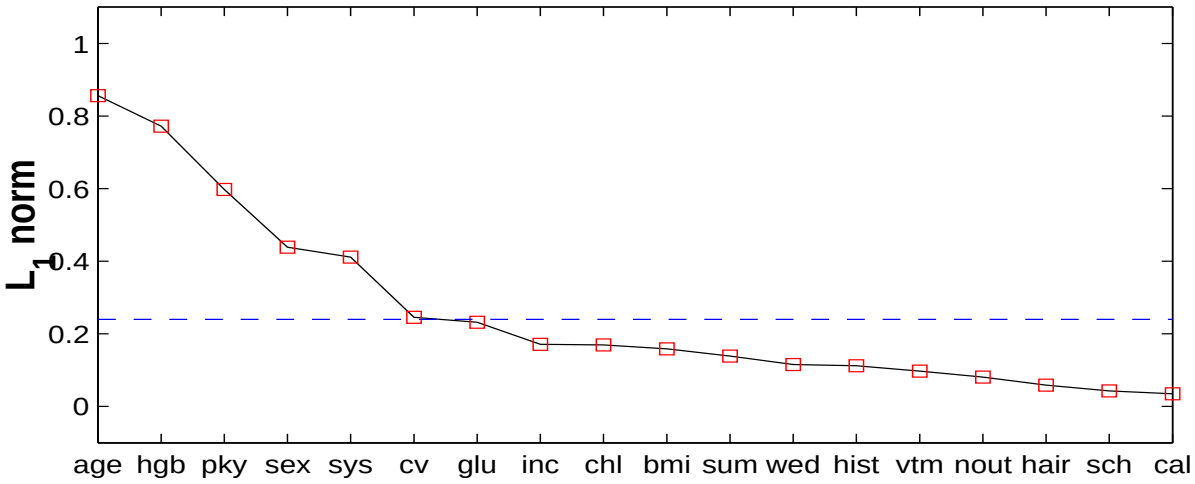

Figure 10: L_1 norm scores for the BDES main effects model

Compared with the LBP model, the linear logistic model with stepwise selection using *AIC* criterion, implemented by the function `glm` in *R* package, misses the variable *sys* but selects three more variables: *inc, bmi* and *sum*. Figure 11 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 linear 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.

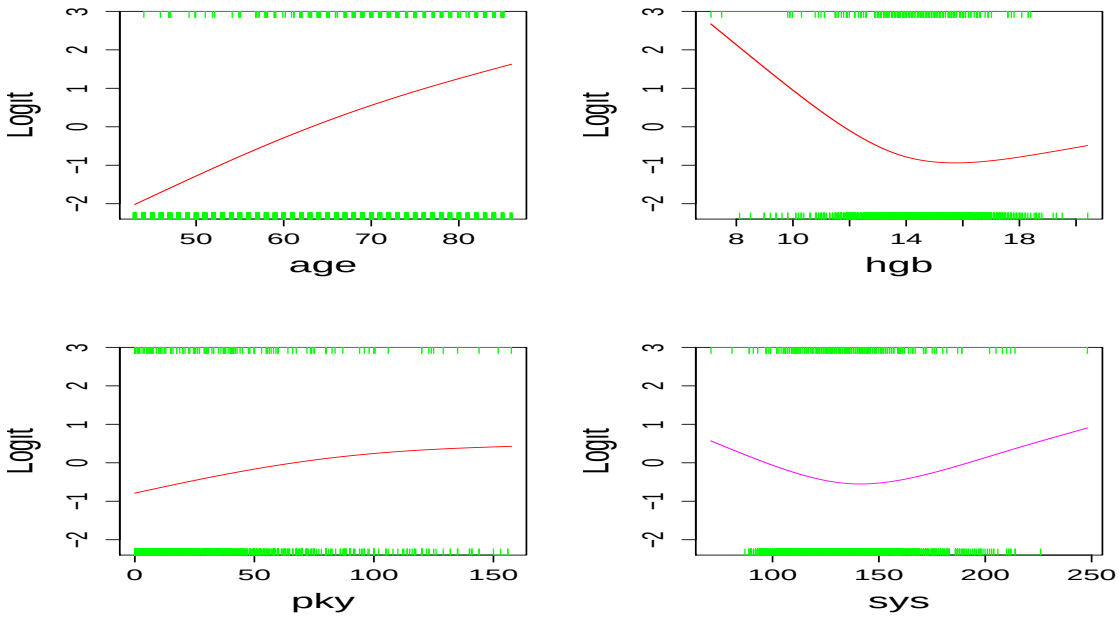

Figure 11: Estimated univariate logit components for important variables.

9 Discussion

We propose the likelihood basis pursuit (LBP) approach for variable selection in high dimensional nonparametric model building. In the spirit of LASSO, LBP produces 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 paper our continuous variables have all been on subsets of the real line, it is clear that other continuous domains are possible. We have used LBP in the context of the Bernoulli distribution, but it can be extended to other exponential distributions as well, of course to Gaussian data. We expect that larger numbers of variables than that considered here may be handled, and we expect that there will be many other scientific applications of the method. We plan to provide freeware for public use.

We believe that this method is 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. In practice, it can be an efficient tool for examining complex data sets to identify and prioritize variables (and, possibly, interactions) for further study. Based only on the variables or interactions identified by the LBP, one can build more traditional parametric or penalized likelihood models, for which confidence intervals and theoretical properties are known.

Appendix A

Proof of Lemma 1

For $i = 1, ..., n$, we have $-l(\mu_\lambda^{[-i]}(\mathbf{x}_i), \tau) = -\mu_\lambda^{[-i]}(\mathbf{x}_i)\tau + b(\tau)$. Let $f_\lambda{}^{[-i]}$ be the minimizer of the objective function

$$\frac{1}{n}\sum_{j\neq i}\big[\,-l(y_j, f(\mathbf{x}_j))\big] + J_\lambda(f). \tag{A.1}$$

Since

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

and

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

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

$$-l(\mu_\lambda^{[-i]}(\mathbf{x}_i), f_\lambda^{[-i]}(\mathbf{x}_i)) \leq -l(\mu_\lambda^{[-i]}(\mathbf{x}_i), f(\mathbf{x}_i)). \tag{A.2}$$

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

$$\begin{aligned}
I_\lambda(f, \mathbf{y}^{-i}) &= \frac{1}{n}\Big\{\,-l(\mu_\lambda^{[-i]}(\mathbf{x}_i), f(\mathbf{x}_i) + \sum_{j\neq i}\,[-l(y_j, f(\mathbf{x}_j))]\,\Big\} + J_\lambda(f) \\
&\geq \frac{1}{n}\Big\{\,-l(\mu_\lambda^{[-i]}(\mathbf{x}_i), f_\lambda^{[-i]}(\mathbf{x}_i)) + \sum_{j\neq i}\,[-l(y_j, f(\mathbf{x}_j))]\,\Big\} + J_\lambda(f) \\
&\geq \frac{1}{n}\Big\{\,-l(\mu_\lambda^{[-i]}(x_i), f_\lambda^{[-i]}(\mathbf{x}_i)) + \sum_{j\neq i}\,[-l(y_j, f_\lambda^{[-i]}(\mathbf{x}_j))]\,\Big\} + J_\lambda(f_\lambda^{[-i]}).
\end{aligned}$$

The first inequality comes from $(A.2)$. The second inequality is due to the fact that $f_\lambda^{[-i]}(\cdot)$ is the minimizer of $(A.1)$. Thus we have $h_\lambda(i, \mu_\lambda^{[-i]}(\mathbf{x}_i), \cdot) = f_\lambda^{[-i]}(\cdot)$.

Appendix B

We derive the approximate cross validation (ACV) in (3.3) for the main effects model. The derivation for two- or higher-order interaction models is similar. Write $\mu(\mathbf{x}) = \mathrm{E}(Y|\mathbf{X} = \mathbf{x})$ and $\sigma^2(\mathbf{x}) = \mathrm{Var}(Y|\mathbf{X} = \mathbf{x})$. For the functional estimate f, we define $f_i = f(\mathbf{x}_i)$, $\mu_i = \mu(\mathbf{x}_i)$ for $i = 1, ..., n$. To emphasize an estimate is associated with the parameter λ, we also use $f_{\lambda i} = f_\lambda(\mathbf{x}_i)$ and $\mu_{\lambda i} = \mu_\lambda(\mathbf{x}_i)$. Now define

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

This is the observed CKL, which, since y_i and $f_{\lambda i}$ are usually positively correlated, tends to underestimate the CKL. To correct this, we will use the leave-out-one cross validation in (3.1)

$$\begin{aligned} CV(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}\big[\,-y_i f_{\lambda i}{}^{[-i]} + b(f_{\lambda i})\big] \\ &= OBS(\lambda) + \frac{1}{n}\sum_{i=1}^{n} y_i(f_{\lambda i} - f_{\lambda i}{}^{[-i]}) \\ &= OBS(\lambda) + \frac{1}{n}\sum_{i=1}^{n} y_i \frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda i}^{[-i]}}\,\frac{y_i - \mu_{\lambda i}}{1 - \dfrac{\mu_{\lambda i} - \mu_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda i}^{[-i]}}}. \end{aligned} \tag{B.1}$$

For exponential families, we have $\mu_{\lambda i} = \dot{b}(f_{\lambda i})$ and $\sigma^2_{\lambda i} = \ddot{b}(f_{\lambda i})$. If we approximate the finite difference by the functional derivative, we get

$$\frac{\mu_{\lambda i} - \mu_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda i}^{[-i]}} = \frac{\dot{b}(f_{\lambda i}) - \dot{b}(f_{\lambda i}^{[-i]})}{y_i - \mu_{\lambda i}^{[-i]}} \approx \ddot{b}(f_{\lambda i})\;\frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda i}^{[-i]}} = \sigma^2_{\lambda i}\;\frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda i}^{[-i]}}. \tag{B.2}$$

Denote

$$G_i \equiv \frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda i}^{[-i]}}, \tag{B.3}$$

from $(B.2)$ and $(B.1)$, we get

$$CV(\lambda) \approx OBS(\lambda) + \frac{1}{n}\sum_{i=1}^{n}\; G_i\;\frac{y_i(y_i - \mu_{\lambda i})}{1 - G_i\sigma^2_{\lambda i}}. \tag{B.4}$$

Now we proceed to approximate G_i. Consider the main effects model with

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

Let $m = Nd$. Define the vectors $\mathbf{b} = (b_0, b_1, ..., b_d)'$ and $\mathbf{c} = (c_{1,1}, ..., c_{d,N})' = (c_1, ..., c_m)'$. For $\alpha = 1, ...d$, define the $n \times N$ matrix $R_\alpha = \Big(K_1(x_i^{(\alpha)}, x_{j*}^{(\alpha)})\Big)$. Let $R = [R_1, ..., R_d]$ and

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

Define $\mathbf{f} = (f_1, ..., f_n)'$, then $\mathbf{f} = T\mathbf{b} + R\mathbf{c}$. The objective function in (3.2) can be expressed as

$$I_\lambda(\mathbf{b}, \mathbf{c}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}\big[-y_i(\sum_{\alpha=1}^{d+1} T_{i\alpha}b_\alpha + \sum_{j=1}^{m} R_{ij}c_j) + b\big(\sum_{\alpha=1}^{d+1} T_{i\alpha}b_\alpha + \sum_{j=1}^{m} R_{ij}c_j\big)\big] + \lambda_d\sum_{\alpha=2}^{d+1}|b_\alpha| + \lambda_s\sum_{j=1}^{m}|c_j| \tag{B.5}$$

With the observed response $\mathbf{y}$, we denote the minimizer of (B.5) by $(\hat{\mathbf{b}}, \hat{\mathbf{c}})$. When a small perturbation ε is imposed on the response, denote the minimizer of $I_\lambda(\mathbf{b}, \mathbf{c}, \mathbf{y} + \varepsilon)$ by $(\tilde{\mathbf{b}}, \tilde{\mathbf{c}})$. Then we have $\mathbf{f}_\lambda^{\mathbf{y}} = T\hat{\mathbf{b}} + R\hat{\mathbf{c}}$ and $\mathbf{f}_\lambda^{\mathbf{y}+\varepsilon} = T\tilde{\mathbf{b}} + R\tilde{\mathbf{c}}$. The l_1 penalty in (B.5) tends to produce sparse solutions, that is, many components of $\hat{\mathbf{b}}$ and $\hat{\mathbf{c}}$ are exactly zero in the solution. For simplicity of explanation, we assume the first s components of $\hat{\mathbf{b}}$ and the first r components of $\hat{\mathbf{c}}$ are nonzero, i.e.

$$\hat{\mathbf{b}} = (\underbrace{\hat{b}_1, ..., \hat{b}_s}_{\neq 0}, \underbrace{\hat{b}_{s+1}, \ldots, \hat{b}_{d+1}}_{=0})', \qquad \hat{\mathbf{c}} = (\underbrace{\hat{c}_1, ..., \hat{c}_r}_{\neq 0}, \underbrace{\hat{c}_{r+1}, ..., \hat{c}_m}_{=0})'.$$

The general case is similar but notationally more complicated.

The zeros in the solutions are robust against a small perturbation in data. That is, when the magnitude of perturbation ε is small enough, the zero elements will stay at zero. This can be seen by transforming the optimization problem (B.5) into a nonlinear programming problem with a differentiable convex objective function and linear constraints, and considering the Karush-Kuhn-Tucker (KKT) conditions (see, for example, Mangasarian (1969)). Thus it is reasonable to assume that the solution of (B.5) with the perturbed data $\mathbf{y} + \varepsilon$ has the form

$$\tilde{\mathbf{b}} = (\underbrace{\tilde{b}_1, ..., \tilde{b}_s}_{\neq 0}, \underbrace{\tilde{b}_{s+1}, ..., \tilde{b}_{d+1}}_{=0})', \qquad \tilde{\mathbf{c}} = (\underbrace{\tilde{c}_1, ..., \tilde{c}_r}_{\neq 0}, \underbrace{\tilde{c}_{r+1}, ..., \tilde{c}_m}_{=0})'.$$

For convenience, we denote the first s components of $\mathbf{b}$ as $\mathbf{b}^*$ and the first r components of $\mathbf{c}$ as $\mathbf{c}^*$. Correspondingly, let T^* be the sub-matrix containing the first s columns of T and R^* be the sub-matrix consisting of the first r columns of R. Then

$$\mathbf{f}_\lambda^{\mathbf{y}+\varepsilon} - \mathbf{f}_\lambda^{\mathbf{y}} = T\ (\tilde{\mathbf{b}} - \hat{\mathbf{b}}) + R\ (\tilde{\mathbf{c}} - \hat{\mathbf{c}}) = \begin{bmatrix} T^* & R^* \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{b}}^* - \hat{\mathbf{b}}^* \\ \tilde{\mathbf{c}}^* - \hat{\mathbf{c}}^* \end{bmatrix}. \tag{B.6}$$

Now

$$[\frac{\partial I_\lambda}{\partial \mathbf{b}^*}, \frac{\partial I_\lambda}{\partial \mathbf{c}^*}]'_{(\tilde{\mathbf{b}},\tilde{\mathbf{c}},\mathbf{y}+\varepsilon)} = \mathbf{0}, \qquad [\frac{\partial I_\lambda}{\partial \mathbf{b}^*}, \frac{\partial I_\lambda}{\partial \mathbf{c}^*}]'_{(\hat{\mathbf{b}},\hat{\mathbf{c}},\mathbf{y})} = \mathbf{0}. \tag{B.7}$$

The first-order Taylor approximation of $[\frac{\partial I_\lambda}{\partial \mathbf{b}^*}, \frac{\partial I_\lambda}{\partial \mathbf{c}^*}]'_{(\tilde{\mathbf{b}},\tilde{\mathbf{c}},\mathbf{y}+\varepsilon)}$ at $(\hat{\mathbf{b}}, \hat{\mathbf{c}}, \mathbf{y})$ gives

$$\begin{aligned} \begin{bmatrix} \frac{\partial I_\lambda}{\partial \mathbf{b}^*} \frac{\partial I_\lambda}{\partial \mathbf{c}^*} \end{bmatrix}_{(\tilde{\mathbf{b}},\tilde{\mathbf{c}},\mathbf{y}+\varepsilon)} \approx & \begin{bmatrix} \frac{\partial I_\lambda}{\partial \mathbf{b}^*} \\ \frac{\partial I_\lambda}{\partial \mathbf{c}^*} \end{bmatrix}_{(\hat{\mathbf{b}},\hat{\mathbf{c}},\mathbf{y})} + \begin{bmatrix} \frac{\partial^2 I_\lambda}{\partial \mathbf{b}^* \partial \mathbf{b}^{*\prime}} & \frac{\partial^2 I_\lambda}{\partial \mathbf{b}^* \partial \mathbf{c}^{*\prime}} \\ \frac{\partial^2 I_\lambda}{\partial \mathbf{c}^* \partial \mathbf{b}^{*\prime}} & \frac{\partial^2 I_\lambda}{\partial \mathbf{c}^* \partial \mathbf{c}^{*\prime}} \end{bmatrix}_{(\hat{\mathbf{b}},\hat{\mathbf{c}},\mathbf{y})} \begin{bmatrix} \tilde{\mathbf{b}}^* - \hat{\mathbf{b}}^* \\ \tilde{\mathbf{c}}^* - \hat{\mathbf{c}}^* \end{bmatrix} \\ & + \begin{bmatrix} \frac{\partial^2 I_\lambda}{\partial \mathbf{b}^* \partial \mathbf{y}'} \\ \frac{\partial^2 I_\lambda}{\partial \mathbf{c}^* \partial \mathbf{y}'} \end{bmatrix}_{(\hat{\mathbf{b}},\hat{\mathbf{c}},\mathbf{y})} (\mathbf{y} + \varepsilon - \mathbf{y}), \end{aligned} \tag{B.8}$$

when the magnitude of ε is small. Define

$$U \equiv \begin{bmatrix} \frac{\partial^2 I_\lambda}{\partial \mathbf{b}^* \partial \mathbf{b}^{*\prime}} & \frac{\partial^2 I_\lambda}{\partial \mathbf{b}^* \partial \mathbf{c}^{*\prime}} \\ \frac{\partial^2 I_\lambda}{\partial \mathbf{c}^* \partial \mathbf{b}^{*\prime}} & \frac{\partial^2 I_\lambda}{\partial \mathbf{c}^* \partial \mathbf{c}^{*\prime}} \end{bmatrix}_{(\hat{\mathbf{b}},\hat{\mathbf{c}},\mathbf{y})}, \qquad V \equiv - \begin{bmatrix} \frac{\partial^2 I_\lambda}{\partial \mathbf{b}^* \partial \mathbf{y}'} \\ \frac{\partial^2 I_\lambda}{\partial \mathbf{c}^* \partial \mathbf{y}'} \end{bmatrix}_{(\hat{\mathbf{b}},\hat{\mathbf{c}},\mathbf{y})}.$$

From (B.7) and (B.8), we have

$$U \begin{bmatrix} \tilde{\mathbf{b}}^* - \hat{\mathbf{b}}^* \\ \tilde{\mathbf{c}}^* - \hat{\mathbf{c}}^* \end{bmatrix} \approx V\epsilon.$$

Then (B.6) gives us $\mathbf{f}_\lambda^{\mathbf{y}+\varepsilon} - \mathbf{f}_\lambda^{\mathbf{y}} \approx H\varepsilon$, where

$$H \equiv \begin{bmatrix} T^* & R^* \end{bmatrix} U^{-1} V. \tag{B.9}$$

Now consider a special form of perturbation $\varepsilon_0 = (0, ..., \mu_{\lambda i}^{[-i]} - y_i, ..., 0)'$, then $\mathbf{f}_\lambda^{\mathbf{y}+\varepsilon_0} - \mathbf{f}_\lambda^{\mathbf{y}} \approx \varepsilon_{0i} H_{\cdot i}$, where $\epsilon_{0i} = \mu_{\lambda i}^{[-i]} - y_i$. Lemma 1 in Section 3 shows that $f_\lambda^{[-i]}$ is the minimizer of $I(f, \mathbf{y} + \varepsilon_0)$.

Therefore G_i in (B.3) is

$$G_i = \frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda i}^{[-i]}} = \frac{f_{\lambda i}^{[-i]} - f_{\lambda i}}{\varepsilon_{0i}} \approx h_{ii}.$$

From $(B.4)$ an approximate cross validation score is then

$$ACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\big[-y_i f_{\lambda i} + b(f_{\lambda i})\big] + \frac{1}{n}\sum_{i=1}^{n} h_{ii}\frac{y_i(y_i - \mu_{\lambda i})}{1 - \sigma_{\lambda i}^2 h_{ii}}. \tag{B.10}$$

Appendix C

Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR)

- Continuous covariates:

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*)	intraocular pressure, $mmHg$

- Categorical covariates:

Z_1:	(*smk*)	smoking status	(0 = no, 1 = any)
Z_2:	(*sex*)	gender	(0 = female, 1 = male)
Z_3:	(*asp*)	use of at least one aspirin for at least three months while diabetic	(0 = no, 1 = yes)
Z_4:	(*famdb*)	family history of diabetes	(0 = none, 1 = yes)
Z_5:	(*mar*)	marital status	(0 = no, 1 = yes/ever)

Appendix D

Beaver Dam Eye Study (BDES)

- Continuous covariates:

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:

Z_1:	(cv)	history of cardiovascular disease	(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:	$(hist)$	history of 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)	part of day spent outdoors in summer	(0 =< 1/4 day, 1 => 1/4 day)
Z_8:	(vtm)	vitamin use	(0 = no, 1 =yes)

References

Bakin, S. (1999), 'Adaptive regression and model selection in data mining problems'. Ph.D. thesis, Australin National University, Canberra ACT 0200, Australia.

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.

Donoho, D. L. & Johostone, I. M. (1994), 'Ideal spatial adaptation by wavelet shrinkage', *Biometrika* **81**, 425–455.

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

Ferris, M. C. & Voelker, M. M. (2000), 'Slice models in general purpose modeling systems', *Optimization Methods and Software.* forthcoming in 2002.

Ferris, M. C. & Voelker, M. M. (2001), Slice models in GAMS, *in* P. Chamoni, R. Leisten, A. Martin, J. Minnemann & H. Stadtler, eds, 'Operations Research Proceedings', Springer-Verlag, pp. 239–246.

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.

Gao, F., Wahba, G., Klein, R. & Klein, B. (2001), 'Smoothing spline ANOVA for multivariate Bernoulli observations, with application to ophthalmology data', *Journal of the American Statistical Association* **96**, 127–160.

Girard, D. (1998), 'Asymptotic comparison of (partial) cross-validation, GCV and randomized GCV in nonparametric regression', *The Annals of Statistics* **26**, 315–334.

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

Gu, C. & Kim, Y. J. (2001), 'Penalized likelihood regression: General formulation and efficient approximation'. To appear in *Canadian Journal of Statistics*.

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

Hastie, T. J. & Tibshirani, R. J. (1990), *Generalized additive models*, Chapman and Hall.

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 over a 10-year period. 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 less than 30 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 Wisconsin Epidemiologic Study of Diabetic Retinopathy. 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.

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

Mangasarian, O. (1969), *Nonlinear Programmin*, McGraw-Hill, New York.

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 the Royal Statistical Society, B* **58**, 267–288.

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 Wisconsin Epidemiological Study of Diabetic Retinopathy', *The Annals of Statistics* **23**, 1865–1895.

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', *Proceedings of 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*.

Zhang, H. H. (2002), 'Nonparametric variable selection and model building via likelihood basis pursuit'. Ph.D. thesis, Department of Statistics, University of Wisconsin, Madison.