

DEPARTMENT OF STATISTICS

University of Wisconsin

1210 West Dayton St.

Madison, WI 53706

TECHNICAL REPORT NO. 1059r

March 14, 2003

## Variable Selection and Model Building via Likelihood Basis Pursuit

Hao Zhang    Grace Wahba    Yi Lin

Department of Statistics, University of Wisconsin, Madison WI

Meta Voelker    Michael Ferris

Department of Computer Sciences, University of Wisconsin, Madison WI

Ronald Klein    Barbara Klein

Department of Ophthalmology, University of Wisconsin, Madison WI

This paper is an invited revision and expansion of TR1059. Supported in part by National Science Foundation grants DMS-0072292, DMS-0134987 and CCR-9972372, National Institute of Health grants EY09946 and EY03083, and Air Force Office of Scientific Research grant F49620-01-1-0040.

# 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  
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 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 of application such as text processing, environmental sciences, and genomics, particularly gene expression array data, for which datasets with tens or hundreds of thousands of variables are available.

Traditional variable selection approaches such as stepwise selection and best subset selection are built in linear regression models, and the well-known criteria like Mallows’s  $C_p$ ,  $AIC$  and  $BIC$  are 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 data from exponential families. Gu (2002) provided a comprehensive review of the SS-ANOVA and some recent progress as well. 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 notations 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 details, especially the “slice modeling” technique. 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.

## 2 Likelihood Basis Pursuit

### 2.1 Smoothing Spline ANOVA for Exponential Families

Conditional on  $\mathbf{X}$ , suppose random variable  $Y$  is from the exponential family with the density form

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

where  $a, b$  and  $c$  are given.  $b(\cdot)$  is a strictly convex function on any bounded set and  $\phi$  are nuisance parameters. Denote the mean and variance of  $Y$  given  $\mathbf{X}$  by  $\mu$  and  $\sigma^2$  respectively. We are interested in the dependence of  $Y$  on the prediction covariates  $\mathbf{X} = (X^1, \dots, X^d)$ . Typically  $\mathbf{X}$  is in a high dimensional space  $\mathcal{X} = \mathcal{X}^1 \otimes \dots \otimes \mathcal{X}^d$ , where  $\mathcal{X}^\alpha, \alpha = 1, \dots, d$ , is some measurable space and  $\otimes$  denotes the tensor product operation. 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. In this paper, we focus on Bernoulli data, which is of particular interest due to its broad applications to risk estimation in scientific research such as medical studies.  $Y$  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  $f(\mathbf{x}) = \log\left(\frac{p(\mathbf{x})}{1 - p(\mathbf{x})}\right)$ . The density of  $Y$  has an expression (2.1) with  $a(\phi) = 1, b(f) = \log(1 + e^f)$ , and  $c(y, \phi) = 0$ . For  $n$  independent observations  $(\mathbf{x}_i, y_i), i = 1, \dots, n$ , the negative log likelihood is

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

Many parametric approaches such as those proposed by 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. Similar to the classical analysis of variance (ANOVA), for any function  $f(\mathbf{x}) = f(x^1, \dots, x^d)$  on a product domain  $\mathcal{X}$ , we can define its functional ANOVA decomposition by

$$f(\mathbf{x}) = b_0 + \sum_{\alpha=1}^d f_\alpha(x^\alpha) + \sum_{\alpha < \beta} f_{\alpha\beta}(x^\alpha, x^\beta) + \text{all higher-order interactions}, \quad (2.3)$$

where  $b_0$  is constant,  $f_\alpha$ '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”.

Firstly, we scale each covariate to the interval  $[0, 1]$  and construct a reproducing kernel Hilbert space (RKHS) on  $[0, 1]^d$  corresponding to the decomposition (2.3) following Wahba et al. (1995). Let  $\mathcal{H}^{(\alpha)}$ ,  $\alpha = 1, \dots, d$ , be the second-order Sobolev-Hilbert space on  $[0, 1]$ . Mathematically,  $\mathcal{H}^{(\alpha)} = \{g : g(x^\alpha), g'(x^\alpha) \text{ are absolutely continuous, } g''(x^\alpha) \in \mathcal{L}_2[0, 1]\}$ . When we endow  $\mathcal{H}^{(\alpha)}$  with the inner product  $(g_1, g_2) = (\int_0^1 g_1(t)dt)(\int_0^1 g_2(t)dt) + (\int_0^1 g_1'(t)dt)(\int_0^1 g_2'(t)dt) + \int_0^1 g_1''(t)g_2''(t)dt$ , the space  $\mathcal{H}^{(\alpha)}$  is an RKHS with kernel  $1 + k_1(s)k_1(t) + k_2(s)k_2(t) - k_4(|s - t|)$ , where  $k_1(t) = t - \frac{1}{2}$ ,  $k_2(t) = \frac{1}{2}(k_1^2(t) - \frac{1}{12})$ , and  $k_4(t) = \frac{1}{24}(k_1^4(t) - \frac{1}{2}k_1^2(t) + \frac{7}{240})$ . This is the special case of equation (10.2.4) in Wahba (1990). 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  $k_1(s)k_1(t) + k_2(s)k_2(t) - k_4(|s - t|)$ . Correspondingly we have the tensor product RKHS

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

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 in  $\otimes_{\alpha=1}^d \mathcal{H}^{(\alpha)}$ . To encompass the linear model as a special case of our model, we further decompose  $\mathcal{H}_1^{(\alpha)}$  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 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, three-factor interaction subspaces, and higher order interaction subspaces. The reproducing kernel of  $\otimes_{\alpha=1}^d \mathcal{H}^{(\alpha)}$  is

$$\prod_{\alpha=1}^d (1 + k_1(s^\alpha)k_1(t^\alpha) + K_1(s^\alpha, t^\alpha)). \quad (2.4)$$

Let  $\mathcal{H}$  be the model space after truncation. Then  $\mathcal{H}$  is a direct sum of  $Q$ , say, component subspaces. We denote each component subspace by  $\mathcal{H}_l$  and its reproducing kernel by  $R_l$ , which is one term in the expansion of (2.4). Then we have  $\mathcal{H} = \oplus_{l=1}^Q \mathcal{H}_l$ , and its kernel  $K = \sum_{l=1}^Q R_l$ .

## 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 the context of 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 from the multivariate function estimate. The variational problem for likelihood basis pursuit (LBP) model is

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n [-l(y_i, f(\mathbf{x}_i))] + J_\lambda(f). \quad (2.5)$$

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 in the solutions helps to distinguish important variables from unimportant ones easily and more effectively. 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  $\lambda$  balances the tradeoff 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  $f_\lambda$  for the traditional smoothing spline model falls in  $\text{span}\{K(\mathbf{x}_i, \cdot), i = 1, \dots, n\}$ , though the model space is of infinite dimensions. 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. When all the data  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  are included to generate the bases, the model space is  $\text{span}\{R_l(\mathbf{x}_i, \cdot), i = 1, \dots, n, l = 1, \dots, Q\}$ . This setup 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  $\{\mathbf{x}_{1*}, \dots, \mathbf{x}_{N*}\}$  from the whole data and use them to generate basis functions. Note the space  $\text{span}\{R_l(\mathbf{x}_{j*}, \cdot), j = 1, \dots, N\}$  is a subspace of  $\mathcal{H}_l$  for  $l = 1, \dots, Q$ . Then the tensor sum RKHS  $\mathcal{H}_*$  used for estimation is

$$\mathcal{H}_* = \bigoplus_{l=1}^Q \text{span}\{R_l(\mathbf{x}_{j*}, \cdot), j = 1, \dots, N\}. \quad (2.6)$$

Notice 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.

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 subsampling technique to choose the subsamples. 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.

### 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. By retaining only main effect subspaces in (2.6), we get the function space

$$\mathcal{H}_* = \bigoplus_{\alpha=1}^d \text{span}\{k_1(x^\alpha), K_1(x^\alpha, x_{j_*}^\alpha), j = 1, \dots, N\} \equiv \bigoplus_{\alpha=1}^d \mathcal{H}_*^{(\alpha)}.$$

Any element  $f_\alpha \in \mathcal{H}_*^{(\alpha)}$  has the representation  $f_\alpha(x^\alpha) = b_\alpha k_1(x^\alpha) + \sum_{j=1}^N c_{\alpha,j} K_1(x^\alpha, x_{j_*}^\alpha)$ , and the function estimate  $f$  is  $f(\mathbf{x}) = b_0 + \sum_{\alpha=1}^d b_\alpha k_1(x^\alpha) + \sum_{\alpha=1}^d \sum_{j=1}^N c_{\alpha,j} K_1(x^\alpha, x_{j_*}^\alpha)$ , where  $k_1(\cdot)$  and  $K_1(\cdot, \cdot)$  are defined in Section 2.1. The likelihood basis pursuit estimate of  $f$  is obtained by minimizing

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

where  $(\lambda_\pi, \lambda_s)$  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  $\lambda$ ’s for the different groups. By using different  $\lambda$ ’s here we allow the possibility of a parametric model. We could also set  $\lambda_\pi = \lambda_s$ .

### 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) page 1570-1600, 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  $d$  paramet-

ric main effect terms  $\{k_1(x^\alpha), \alpha = 1, \dots, d\}$  and  $\frac{d(d-1)}{2}$  parametric-parametric interaction terms  $\{k_1(x^\alpha)k_1(x^\beta), \alpha = 1, \dots, d, \beta < \alpha\}$ . 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. The function space used is  $\mathcal{H}_* \equiv \bigoplus_{\alpha=1}^d \mathcal{H}_*^{(\alpha)} + \bigoplus_{\beta < \alpha} \mathcal{H}_*^{(\alpha\beta)}$ . For each pair  $\alpha \neq \beta$ ,

$$\mathcal{H}_*^{(\alpha\beta)} = \text{span}\{ k_1(x^\alpha)k_1(x^\beta), K_1(x^\alpha, x_{j_*}^\alpha)k_1(x^\beta)k_1(x_{j_*}^\beta), K_1(x^\alpha, x_{j_*}^\alpha)K_1(x^\beta, x_{j_*}^\beta), j = 1, \dots, N\},$$

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

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

Different penalties are allowed for five 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 totally five tuning parameters  $\{\lambda_\pi, \lambda_{\pi\pi}, \lambda_{\pi s}, \lambda_s, \lambda_{ss}\}$  in the two-factor interaction model. The optimization problem is: minimize

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n [-l(y_i, f(\mathbf{x}_i))] + \lambda_\pi \left( \sum_{\alpha=1}^d |b_\alpha| \right) + \lambda_{\pi\pi} \left( \sum_{\alpha < \beta} |b_{\alpha\beta}| \right) \\ & + \lambda_{\pi s} \left( \sum_{\alpha \neq \beta} \sum_{j=1}^N |c_{\alpha\beta,j}^{\pi s}| \right) + \lambda_s \left( \sum_{\alpha=1}^d \sum_{j=1}^N |c_{\alpha,j}| \right) + \lambda_{ss} \left( \sum_{\alpha < \beta} \sum_{j=1}^N |c_{\alpha\beta,j}^{ss}| \right). \end{aligned} \quad (2.8)$$

### 2.3 Incorporating Categorical Variables

In real applications, some covariates may be categorical such as sex, race, smoking history, and marital status in many medical studies. In Section 2.1 and Section 2.2, the main effects model (2.7) and the two-factor interaction model (2.8) are proposed for continuous variables only. In this section we will generalize these models to incorporate categorical variables. Assume there are  $r$  categorical variables and denote them by a vector  $\mathbf{Z} = (Z^1, \dots, Z^r)$ . Usually each variable has several categories, and here we consider the simplest case that all  $Z$ 's are two-level categorical variables. Similar ideas are easily extended for variables having more than two categories. Assume  $Z^1$  takes two responses  $\{T, F\}$ , we define the mapping  $\Phi_1$ :

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



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

$$\frac{1}{n} \sum_{i=1}^n [-l(y_i, f(\mathbf{x}_i, \mathbf{z}_i))] + \lambda_\pi \left( \sum_{\alpha=1}^d |b_\alpha| + \sum_{\gamma=1}^r |B_\gamma| \right) + \lambda_s \sum_{\alpha=1}^d \sum_{j=1}^N |c_{\alpha,j}|, \quad (2.9)$$

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

- 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 (2.8), 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

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n [-l(y_i, f(\mathbf{x}_i, \mathbf{z}_i))] + \lambda_\pi \left( \sum_{\alpha=1}^d |b_\alpha| + \sum_{\gamma=1}^r |B_\gamma| \right) + \lambda_{\pi\pi} \left( \sum_{\alpha < \beta} |b_{\alpha\beta}| + \sum_{\gamma < \theta} |B_{\gamma\theta}| + \sum_{\alpha=1}^d \sum_{\gamma=1}^r |P_{\alpha\gamma}| \right) \\ & + \lambda_{\pi s} \left( \sum_{\alpha \neq \beta} \sum_{j=1}^N |c_{\alpha\beta,j}^{\pi s}| + \sum_{\alpha=1}^d \sum_{\gamma=1}^r \sum_{j=1}^N |c_{\alpha\gamma,j}^{\pi s}| \right) + \lambda_s \left( \sum_{\alpha=1}^d \sum_{j=1}^N |c_{\alpha,j}| \right) + \lambda_{ss} \left( \sum_{\alpha < \beta} \sum_{j=1}^N |c_{\alpha\beta,j}^{ss}| \right), \end{aligned} \quad (2.10)$$

where

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

We assign different regularization parameters for main effect terms, parametric-parametric in-

teraction 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 same parameter  $\lambda_{\pi\pi}$ , while the coefficients  $|c_{\alpha\gamma,j}^{\pi s}|$ 's are assigned with  $\lambda_{\pi s}$ .

### 3 Generalized Approximate Cross Validation (GACV)

The  $\lambda$ 's in the LBP models are called regularization parameters, tuning parameters, or smoothing parameters in the context of smoothing models. Regularization parameter selection has been a very active research field, appearing in various contexts of penalized likelihood methods and other nonparametric methods. For the smoothing splines with Gaussian data, ordinary cross validation (OCV) was originally proposed by Wahba & Wold (1975). Craven & Wahba (1979) suggested the generalized cross validation (GCV) which has been widely used since then. Later for the smoothing splines with non-Gaussian data, Xiang & Wahba (1996) proposed the generalized approximate cross validation (GACV) as an extension of GCV. We will derive the GACV to select the  $\lambda$ 's in the likelihood basis pursuit models. With an abuse of notation, we use  $\lambda$  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  $p_\lambda$  be its estimate associated with  $\lambda$ . Similarly,  $f$  and  $\mu$  are respectively the true logit and mean functions, and  $f_\lambda$  and  $\mu_\lambda$  are the corresponding estimates. We focus on the main effects model, since the ideas are easily extended to two-factor interaction model and more complicated models. The objective function in the main effects model (2.7) is expressed as

$$I_\lambda(f, \mathbf{y}) \equiv \sum_{i=1}^n [-l(y_i, f(\mathbf{x}_i))] + J_\lambda(f), \quad (3.1)$$

where the penalty function  $J_\lambda(f) = \lambda_\pi \sum_{\alpha=1}^d |b_\alpha| + \lambda_s \sum_{\alpha=1}^d \sum_{j=1}^N |c_{\alpha,j}|$ . 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  $KL(p, p_\lambda) = E_{\mathbf{x}}[\frac{1}{2}\{\mu(f-f_\lambda) - (b(f) - b(f_\lambda))\}]$ . By removing the quantity which does not depend on  $\lambda$  from the KL distance expression, the comparative Kullback-Leibler distance (CKL) is obtained

$$CKL(p, p_\lambda) = E_{\mathbf{x}}[-\mu f_\lambda + b(f_\lambda)]. \quad (3.2)$$

The ordinary leaving-out-one cross validation function CV for CKL is

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n [-y_i f_{\lambda}^{[-i]}(\mathbf{x}_i) + b(f_{\lambda}(\mathbf{x}_i))], \quad (3.3)$$

where  $f_{\lambda}^{[-i]}$  is the minimizer of (3.1) with the  $i$ -th data point omitted. In other words,  $f_{\lambda}^{[-i]}$  minimizes the objective function  $\sum_{j \neq i} [-l(y_j, f(\mathbf{x}_j))] + J_{\lambda}(f)$ . CV is commonly used as a roughly unbiased estimate for CKL in (3.2), see, for example, 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. The following leaving-out-one lemma will be used to derive an approximate cross validation (ACV) score.

**Lemma 1: (Leaving-out-one lemma for LBP)** Let  $-l(y_i, f(\mathbf{x}_i)) = -y_i f(\mathbf{x}_i) + b(f(\mathbf{x}_i))$  and  $I_{\lambda}(f, \mathbf{y}) = \sum_{j=1}^n [-l(y_j, f(\mathbf{x}_j))] + J_{\lambda}(f)$ .  $J_{\lambda}(f)$  is the  $l_1$  norm of the coefficients in the decomposition of  $f$ . Let  $\mu_{\lambda}^{[-i]}(\cdot)$  be the mean function corresponding to  $f_{\lambda}^{[-i]}(\cdot)$ . Suppose  $h_{\lambda}(i, v, \cdot)$  is the minimizer of  $I_{\lambda}(f, \mathbf{V})$ , where  $\mathbf{V} = (y_1, \dots, y_{i-1}, v, y_{i+1}, \dots, y_n)$ , then

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

The proof of Lemma 1 is given in Appendix A. Using Taylor series approximations and Lemma 1, we can derive the ACV which is a second-order approximation to CV. The argument is similar to Xiang & Wahba (1996), Lin et al. (2000), and Gao, Wahba, Klein & Klein (2001). However unlike in any of these papers just mentioned, the  $l_1$  norm in the LBP model is non-differentiable, which caused some difficulty in the derivation of ACV. We only present the expression of ACV here and its derivation can be found in Zhang (2002).

$$ACV(\lambda) = \frac{1}{n} \sum_{i=1}^n [-y_i f_{\lambda}(\mathbf{x}_i) + b(f_{\lambda}(\mathbf{x}_i))] + \frac{1}{n} \sum_{i=1}^n h_{ii} \frac{y_i(y_i - \mu_{\lambda}(\mathbf{x}_i))}{1 - \sigma_{\lambda_i}^2 h_{ii}}, \quad (3.4)$$

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 a matrix  $H$  defined in Zhang (2002),

$$h_{ii} \approx \frac{f_{\lambda}(\mathbf{x}_i) - f_{\lambda}^{[-i]}(\mathbf{x}_i)}{y_i - \mu_{\lambda}^{[-i]}(\mathbf{x}_i)}.$$

Let  $W$  be the  $n \times n$  diagonal matrix with  $\sigma_{\lambda_i}^2$  in the  $ii$ -th position. By replacing  $h_{ii}$  with  $\frac{1}{n} \sum_{i=1}^n h_{ii} \equiv \frac{1}{n} \text{tr}(H)$  and replacing  $1 - \sigma_{\lambda_i}^2 h_{ii}$  with  $\frac{1}{n} \text{tr}[I - (W^{1/2} H W^{1/2})]$ , we obtain the generalized approximate

cross validation (GACV) score

$$GACV(\lambda) = \frac{1}{n} \sum_{i=1}^n [-y_i f_\lambda(\mathbf{x}_i) + b(f_\lambda(\mathbf{x}_i))] + \frac{tr(H)}{n} \frac{\sum_{i=1}^n y_i (y_i - \mu_\lambda(\mathbf{x}_i))}{tr[I - W^{1/2} H W^{1/2}]} \quad (3.5)$$

### 3.2 Randomized GACV

Direct computation of (3.5) 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  $tr[I - \frac{1}{2}(W^{1/2} H W^{1/2})]$  based on the following theorem, (which has been exploited by numerous authors, see e.g. Girard (1998)).

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

Let  $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$  be a zero mean random  $n$ -vector of independent components with variance  $\sigma_\epsilon^2$ . Let  $f_\lambda^{\mathbf{y}}$  and  $f_\lambda^{\mathbf{y}+\epsilon}$  respectively be the minimizer of (2.7) using the original data  $\mathbf{y} \equiv (y_1, \dots, y_n)'$  and the perturbed data  $\mathbf{y} + \epsilon$ . Then the ranGACV score is given by

$$ranGACV(\lambda) = \frac{1}{n} \sum_{i=1}^n [-y_i f_\lambda(\mathbf{x}_i) + b(f_\lambda(\mathbf{x}_i))] + \frac{\epsilon^T (f_\lambda^{\mathbf{y}+\epsilon} - f_\lambda^{\mathbf{y}})}{n} \frac{\sum_{i=1}^n y_i (y_i - \mu_\lambda(\mathbf{x}_i))}{\epsilon^T \epsilon - \epsilon^T W (f_\lambda^{\mathbf{y}+\epsilon} - f_\lambda^{\mathbf{y}})} \quad (3.6)$$

Its derivation is given in Lin et al. (2000). In addition, two facts help to reduce the variance of the second term in (3.6). (1) It is shown in Hutchinson (1989) that given the variance  $\sigma_\epsilon^2$ , when each component of  $\epsilon$  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  $U$  independent perturbations  $\epsilon^{(u)}$ ,  $u = 1, \dots, U$ , and compute  $U$ -replicate ranGACVs. Their average has a smaller variance.

## 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 (2.7), (2.8), (2.9) or (2.10). How to measure the importance of a particular component in the fitted model is a key question. We consider the main effects and, possibly, the two factor interactions as the model components of interest, and propose using the functional  $L_1$  norm as the importance measure.

In the fitting step, the advantage of using  $l_1$  regularization is that it produces sparse solutions, by pushing the coefficients of unimportant components towards zero more quickly than the traditional quadratic regularization. Some terms in the ANOVA decomposition will automatically become zero. This sparsity will distinguish the significant terms from insignificant ones more effectively, and thus improve the performance the importance measure. 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 (2.7), the empirical  $L_1$  norms of the main effect  $f_\alpha$  and the two-factor interaction  $f_{\alpha\beta}$ ,  $\alpha = 1, \dots, d$ ,  $\beta < \alpha$ , are

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

For the categorical variables in the model (2.9), the empirical  $L_1$  norm of the main effect  $f_\gamma$  is  $L_1(f_\gamma) = \frac{1}{n} \sum_{i=1}^n |B_\gamma \Phi_\gamma(z_i^\gamma)|$  for  $\gamma = 1, \dots, r$ . The norms of the interaction terms involved with categorical variables are defined similarly. The rank of the  $L_1$  norm scores indicates the relative importance of various effect terms. For instance, the component with the largest  $L_1$  norm is the most important, and any variable with zero or tiny  $L_1$  norm is unimportant. The functional  $L_2$  norm gave almost identical results 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  $\hat{f}_1, \dots, \hat{f}_d$  and calculate their  $L_1$  norms  $L_1(\hat{f}_1), \dots, L_1(\hat{f}_d)$ . We will use a sequential procedure to select important terms. Denote the decreasingly ordered norms as  $\hat{L}_{(1)}, \dots, \hat{L}_{(d)}$  and the corresponding components  $\hat{f}_{(1)}, \dots, \hat{f}_{(d)}$ . 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  $\eta$  ( $0 \leq \eta \leq d - 1$ ) 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  $f = \hat{b}_0 + \hat{f}_{(1)} + \dots + \hat{f}_{(\eta)}$  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  $\alpha$ . If the null distribution of  $\hat{L}_{(\eta+1)}$  were known, we could get the critical value  $\alpha$ -percentile and make a decision of rejection or acceptance. In practice the exact  $\alpha$ -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  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ , we generate  $\{y_1^{*(\eta)}, \dots, y_n^{*(\eta)}\}$  (responses 0 or 1) by using the null model  $f = \hat{b}_0 + \hat{f}_{(1)} + \dots + \hat{f}_{(\eta)}$  as the true logit function. Totally we sample  $T$  independent sets of data  $(\mathbf{x}_1, y_{1,t}^{*(\eta)}), \dots, (\mathbf{x}_n, y_{n,t}^{*(\eta)})$ ,  $t = 1, \dots, T$ , from the null model  $f$ , fit the main effects model for each set, and compute  $\hat{L}_t^{*(\eta+1)}$ ,  $t = 1, \dots, T$ . 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  $f = \hat{b}_0$ . We test  $H_0 : L_{(1)} = 0$  vs  $H_1 : L_{(1)} > 0$ . Generate  $T$  independent sets of data  $(\mathbf{x}_1, y_{1,t}^{*(0)}), \dots, (\mathbf{x}_n, y_{n,t}^{*(0)})$ ,  $t = 1, \dots, T$ , 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  $f = \hat{b}_0 + \hat{f}_{(1)} + \dots + \hat{f}_{(\eta)}$ . We test  $H_0 : L_{(\eta+1)} = 0$  vs  $H_1 : L_{(\eta+1)} > 0$ . Generate  $T$  independent sets of data  $(\mathbf{x}_1, y_{1,t}^{*(\eta)}), \dots, (\mathbf{x}_n, y_{n,t}^{*(\eta)})$  based on  $f$ , fit the main effects model and compute the Monte Carlo  $p$ -value  $p_\eta$ . 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  $q = \hat{L}_{(d)}$ .

## 5 Numerical Computation

Since the objective function in either (2.7) or (2.8) is not differentiable with respect to the coefficients  $b$  and  $c$ , some numerical methods for optimization fail to solve this kind of problem. By introducing proper constraints, we can change this problem into minimizing a nonlinear smooth and convex

function with polyhedral constraints. Many methods can be used for such problems; we choose to employ MINOS (see Murtagh & Saunders (1983)) as the underlying non-linear solver. MINOS performs well with the linearly constrained models and returns consistent results.

For every value of  $\lambda$ , program (2.7), (2.8), (2.9), or (2.10) must be solved twice — once with  $y$  (the original problem) and once with  $y + \epsilon$  (the perturbed problem). This often results in hundreds or thousands of individual solves, depending upon the range for  $\lambda$ . 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 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 for the LBP models are examples where non-linear slice modeling is useful. The  $l_1$  norms in the objective function can be replaced by non-negative variables constrained linearly to be the corresponding absolute values using standard mathematical programming techniques. Then we have a series of programs with non-linear objective functions and linear constraints. These programs only vary in the objective functions (in the  $\lambda$  values and/or the  $y$  values). Slice modeling improves efficiency of the programs by removing the necessity of regenerating the constraints for each solve and allowing previous solutions to be used for starting values.

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  $\lambda$ . 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”. Once we have solutions for the original and perturbed problems at a particular  $\lambda$ , `ranGACV` can be calculated. This suggests solving the original and perturbed problems together for each  $\lambda$ . However, the slice

modeling approach suggests the opposite: because fewer changes in the solution take place moving from one  $\lambda$  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.

## 6 Simulation

### 6.1 Simulation 1: Main Effects Model

In this example, there are altogether  $d = 10$  covariates:  $X_1, \dots, X_{10}$ . 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  $\epsilon$  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

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

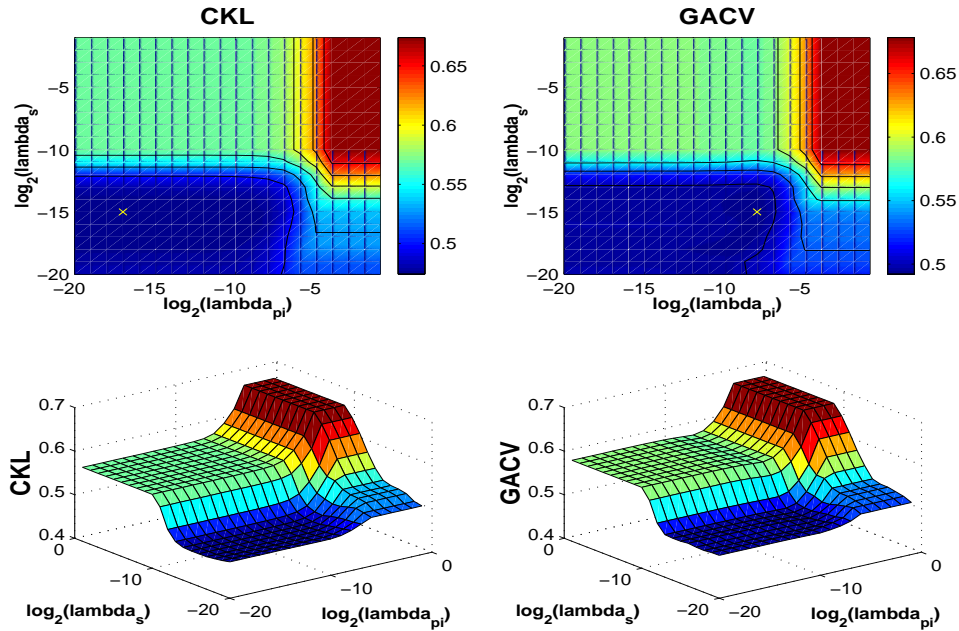


Figure 1: Contours and three-dimensional plots for  $CKL(\lambda)$  and  $GACV(\lambda)$ .



We fit the main effects LBP model and search the parameters  $(\lambda_\pi, \lambda_s)$  globally. Since the true  $f$  is known, both CKL and ranGACV can be used for choosing the  $\lambda$ 's. Figure 1 depicts the values of  $CKL(\lambda)$  and  $ranGACV(\lambda)$  as functions of  $(\lambda_\pi, \lambda_s)$  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  $\hat{f}_1, \dots, \hat{f}_{10}$ . 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.

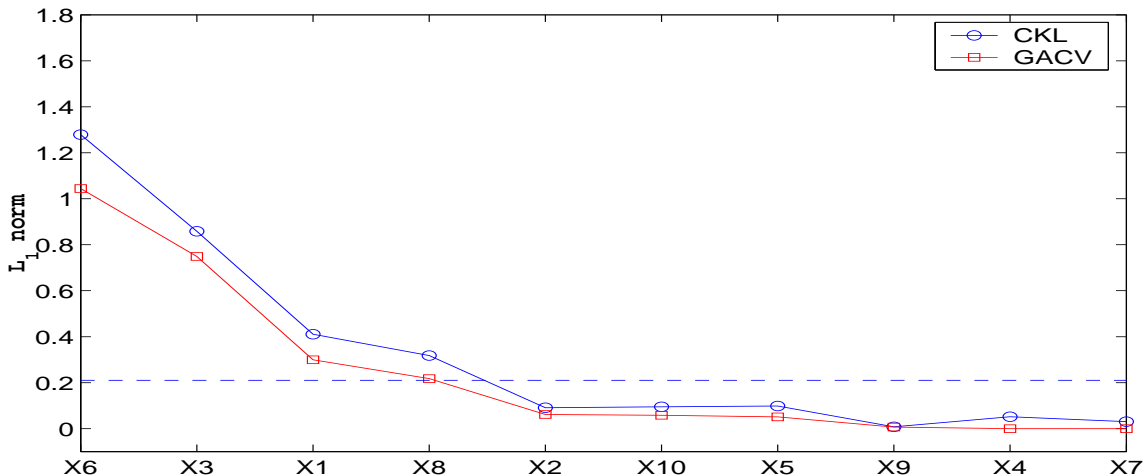


Figure 2:  $L_1$  norm scores for the main effects model.

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  $H_1 : L_{(\eta)} > 0, \eta = 1, \dots, 10$ . In each plot, along the horizontal axis the variable being tested for importance is bracketed by a pair of \*. Light color (green in a colored plot) is used for the variables which are in the null model, and dark color (blue 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 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  $q = L_{(4)} = 0.21$ .

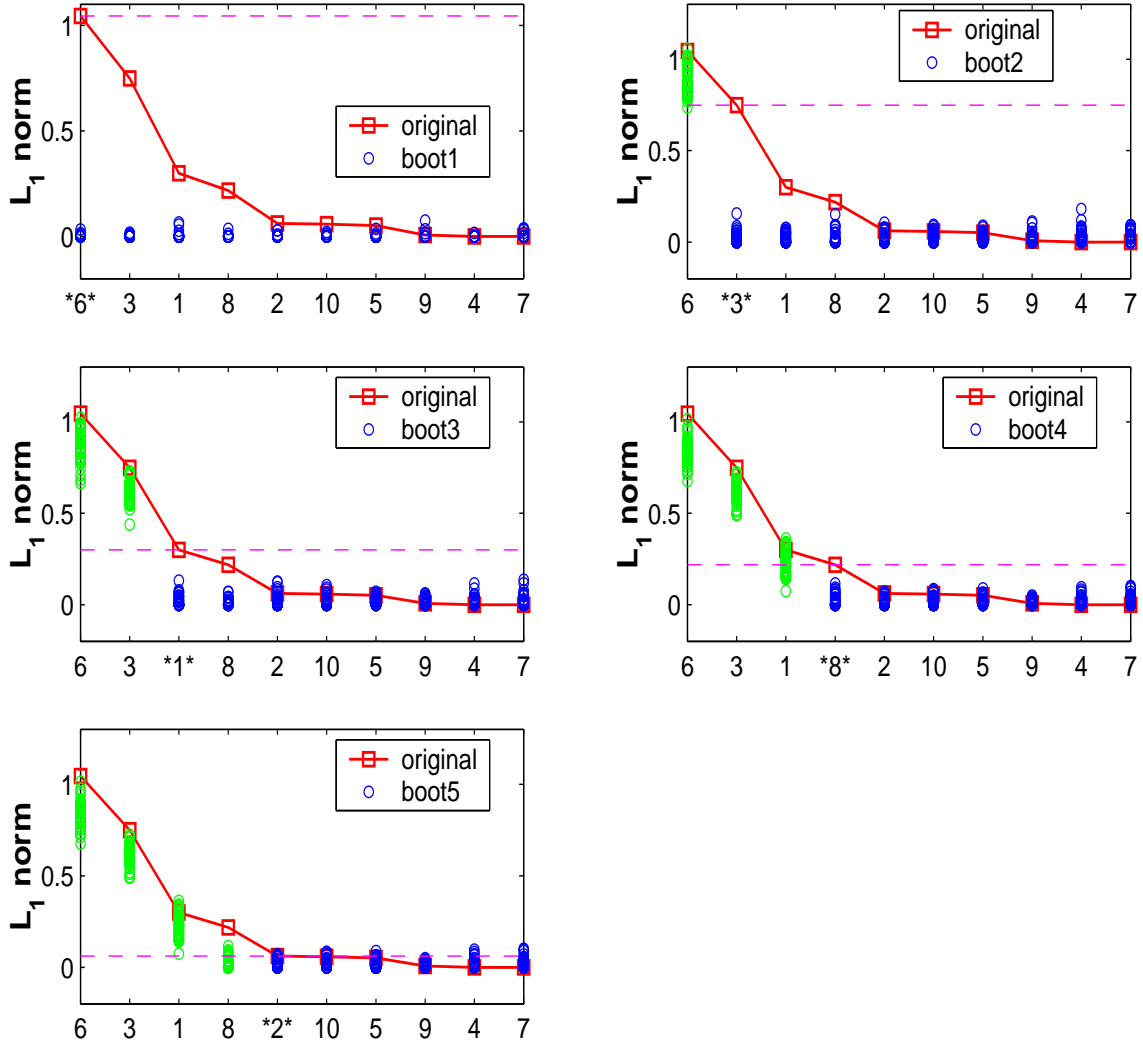


Figure 3: Monte Carlo bootstrap tests for Simulation 1.

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 the 20 runs, variables  $X_1, X_3, X_6$  and  $X_8$  are the four top-ranked variables. The results shown above are based on the first dataset.

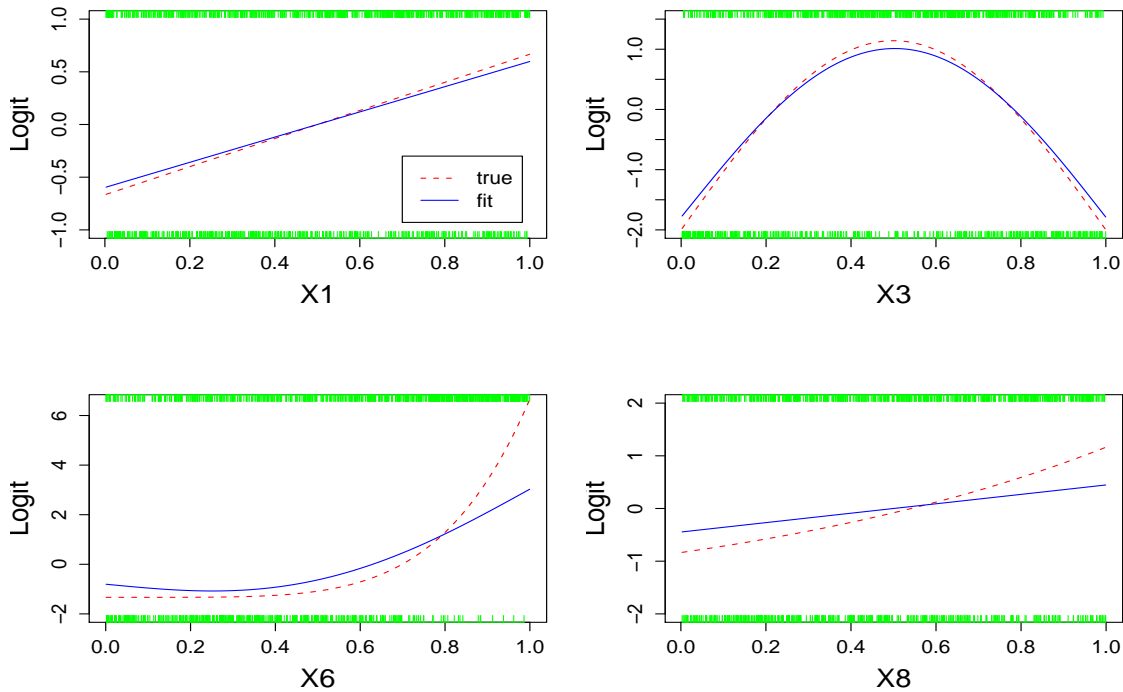


Figure 4: True and estimated univariate logit component.

## 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

$$f(x) = 4x_1 + \pi \sin(\pi x_1) + 6x_2 - 8x_2^3 + \cos(2\pi(x_1 - x_2)) - 4.$$

We choose  $n = 1000$ ,  $N = 50$ , and use the same perturbation  $\epsilon$  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.

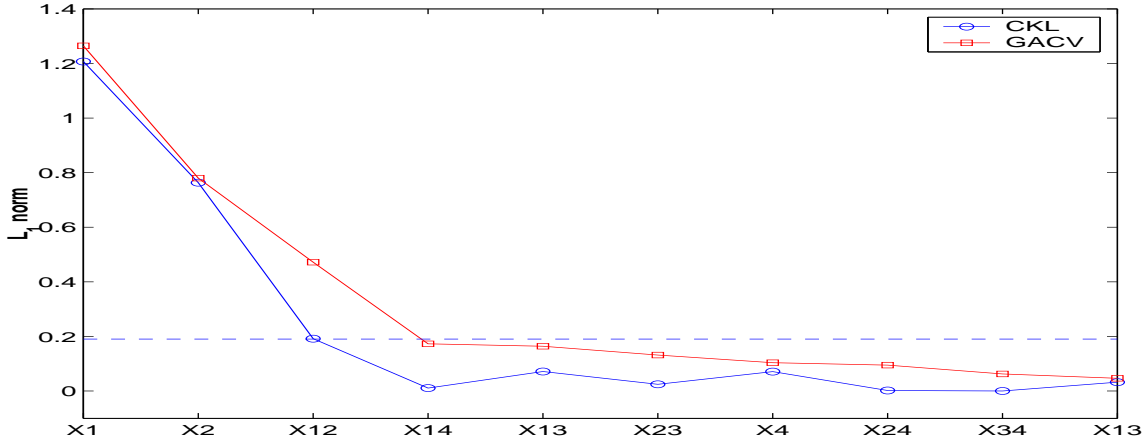


Figure 5:  $L_1$  norm scores for the two-factor interaction model.

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  $f(x_1, x_2)$  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 ranGACV criterion.

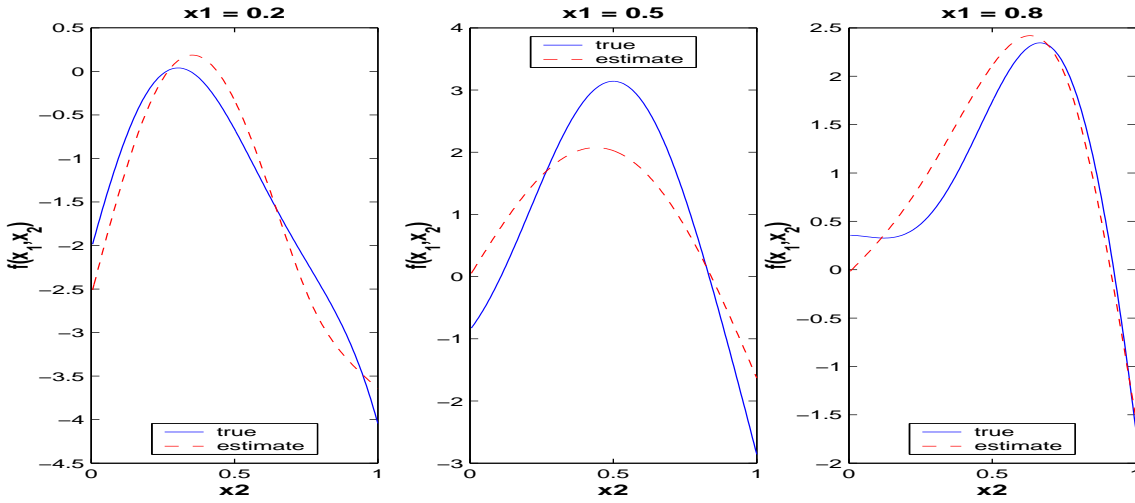


Figure 6: Cross section plots for the two-factor interaction model.

### 6.3 Simulation 3: Main Effects Model Incorporating Categorical Variables

In this example, among the 12 covariates  $X_1, \dots, X_{10}$  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

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

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  $\epsilon$  as in the previous examples. The main effects model incorporating categorical variables in (2.9) 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.

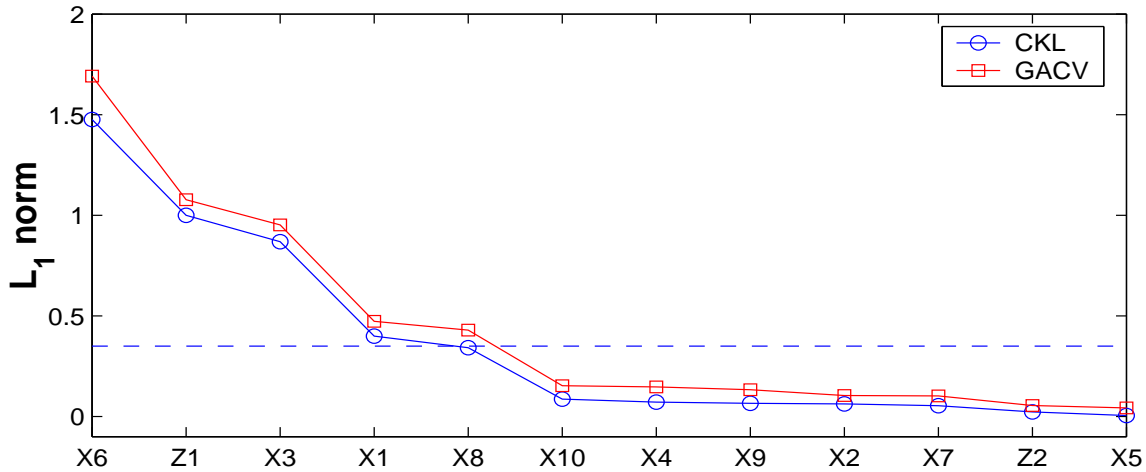


Figure 7:  $L_1$  norm scores for the main effects model incorporating categorical variables.

## 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. Diabetic retinopathy, a complication of diabetes, can lead to severe decrease in vision and blindness. Non-proliferative 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. 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, members of which 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 were 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 B. Since the true  $f$  is not known for real data, only ranGACV is available

for tuning  $\lambda$ . 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.

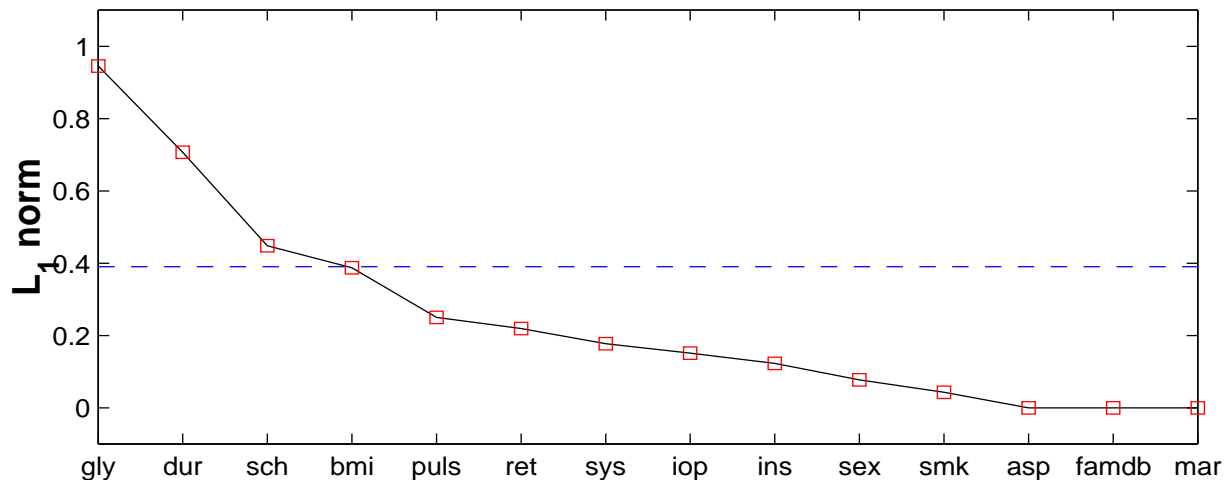


Figure 8:  $L_1$  norm scores for the WESDR main effects model.

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 component given by the LBP main effects model 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). When we fit the 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 9 exhibits a hilly shape, which indicates that a quadratic function might fit the curve better than a linear function. 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 (2.8) 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.

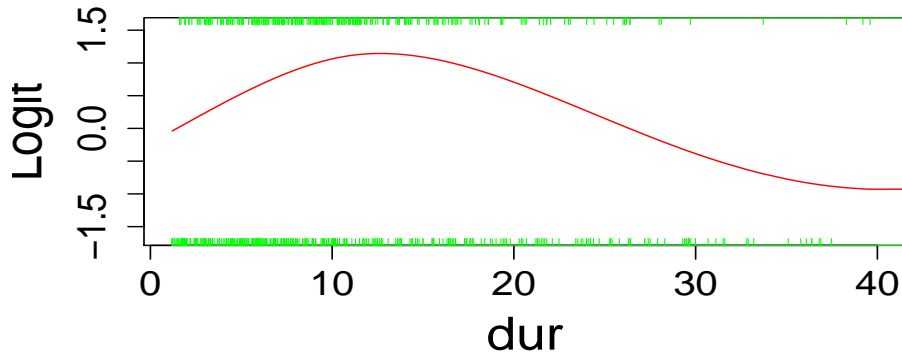


Figure 9: Estimated logit component for *dur*.

## 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 C. 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 (2.9) 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.

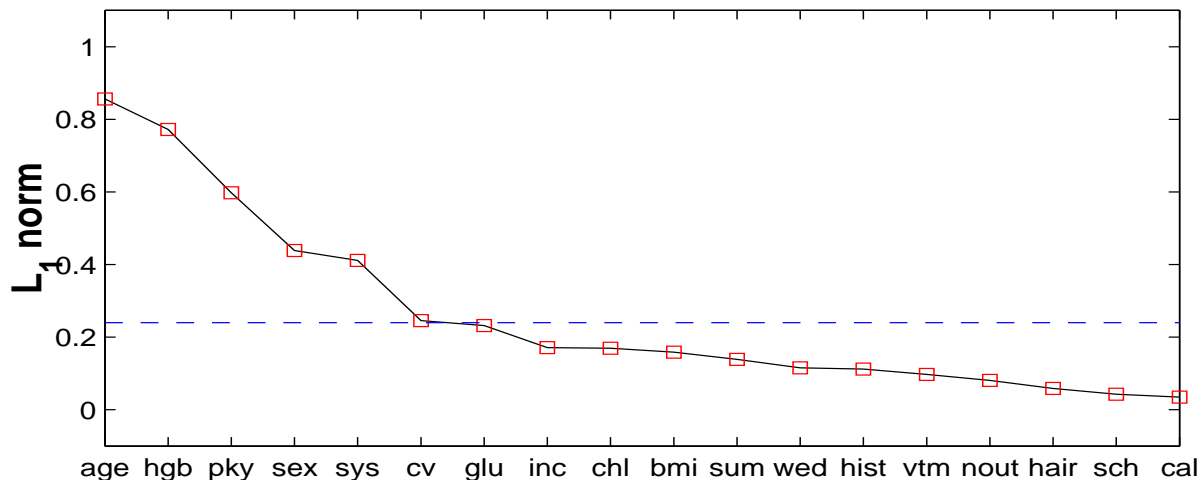


Figure 10:  $L_1$  norm scores for the BDES main effects model

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  $sys^2$ .

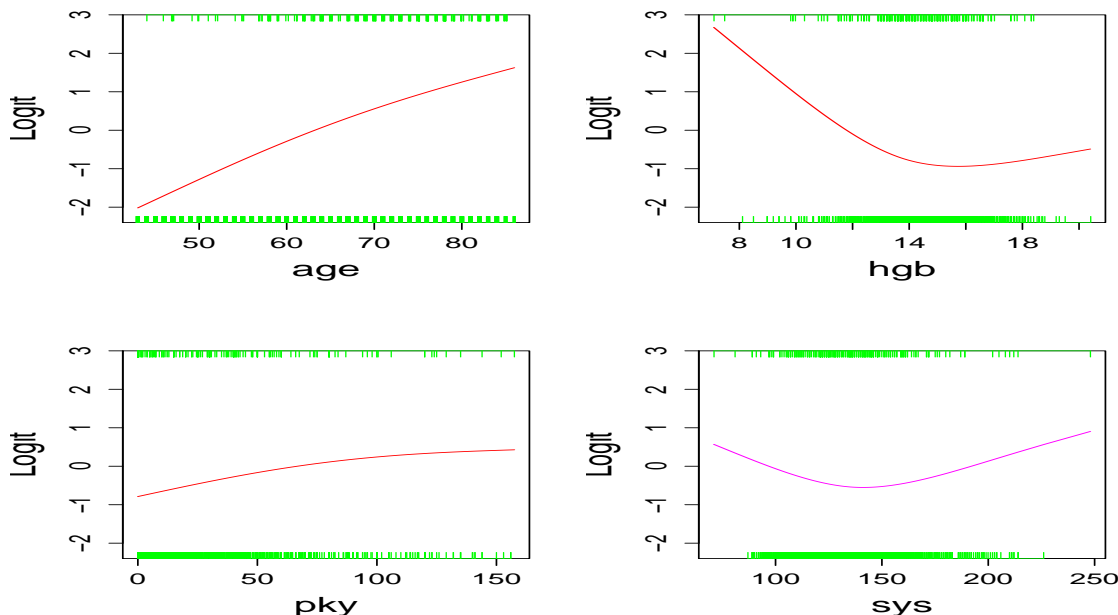


Figure 11: Estimated univariate logit components for important variables.

## 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. 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.

## Appendix A

### Proof of Lemma 1

For  $i = 1, \dots, n$ , 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

$$\sum_{j \neq i} [-l(y_j, f(\mathbf{x}_j))] + J_\lambda(f). \quad (9.1)$$

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  $b'(\hat{\tau}) = \mu_\lambda^{[-i]}(\mathbf{x}_i)$ . So  $\hat{\tau} = f_\lambda^{[-i]}(\mathbf{x}_i)$ . Then for any  $f$ , we have

$$-l(\mu_\lambda^{[-i]}(\mathbf{x}_i), f_\lambda^{[-i]}(\mathbf{x}_i)) \leq -l(\mu_\lambda^{[-i]}(\mathbf{x}_i), f(\mathbf{x}_i)). \quad (9.2)$$

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

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

The first inequality comes from (9.2). The second inequality is due to the fact that  $f_\lambda^{[-i]}(\cdot)$  is the minimizer of (9.1). Thus we have  $h_\lambda(i, \mu_\lambda^{[-i]}(\mathbf{x}_i), \cdot) = f_\lambda^{[-i]}(\cdot)$ .

## Appendix B

### Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR)

- Continuous covariates:

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

- Categorical covariates:

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

## Appendix C

### Beaver Dam Eye Study (BDES)

- Continuous covariates:

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

- Categorical covariates:

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

## References

- Bakin, S. (1999), ‘Adaptive regression and model selection in data mining problems’. Ph.D. thesis, Australian 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 Bayesian 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. Stadler, 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. (1984a), ‘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. (1984b), ‘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.
- 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., 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.
- Wahba, G. & Wold, S. (1975), ‘A completely automatic French curve’, *Commun. Statist.* **4**, 1–17.
- 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.