DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 1042

October 3, 2001

Variable Selection via Basis Pursuit for Non-Gaussian Data

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 Ophthalmalogy, University of Wisconsin, Madison WI

This paper was prepared for the Proceedings of the 2001 ASA Joint Statistical Meetings (JSM 2001), Biometrics Section. Research partly supported by NSF Grants DMS-0072292 and CCR-9972372, the Wisconsin Alumni Research Foundation, NIH Grants EY09946 and EY03083 and AFOSR Grant F49620-01-1-0040. [1]

[1] An additional reference was added on October 15

Variable Selection via Basis Pursuit for Non-Gaussian Data

Hao Zhang [2] ,**Grace Wahba** [3] ,**Yi Lin** [4]
Meta Voelker [5] ,**Michael Ferris** [6]
Ronald Klein [7] ,**Barbara Klein** [8]
University of Wisconsin - Madison

Abstract

A simultaneous flexible variable selection procedure is proposed by applying a basis pursuit method to the likelihood function. The basis functions are chosen to be compatible with variable selection in the context of smoothing spline ANOVA models. Since it is a generalized LASSO-type method, it enjoys the favorable property of shrinking coefficients and gives interpretable results. We derive a Generalized Approximate Cross Validation function (GACV), an approximate leave-out-one cross validation function used to choose smoothing parameters. In order to apply the GACV function for a large data set situation, we propose a corresponding randomized GACV. A technique called 'slice modeling' is used to develop an efficient code. Our simulation study shows the effectiveness of the proposed approach in the Bernoulli case.

KEY WORDS: basis pursuit, spline ANOVA, LASSO, GACV, randomized GACV, slice modeling

1 Motivation

Variable selection and model building are very important in statistical data analysis. In many scientific areas, we are often faced with problems of selecting a subset of possible predictors. Stepwise regression procedures are traditional methods, such as forward selection, backward elimination and leaps and bounds methods. However they rely on the assumption of linear models or generalized linear models and always demand expensive computation. The LASSO proposed by Tibshirani (1996) minimizes the penalized ordinary least squares with L_1 penalty. A few penalized likelihood approaches were proposed by Fan & Li (1999) and Fu (1998) to select variables for linear regression models and generalized linear models. These methods are quite stable and produce interpretable results, except that the required assumptions of parametric forms make them limited in application. In addition, they need the underlying distribution for the variation to be at least approximately Gaussian to validate the statistical inferences and decisions. Gunn & Kandola (2001) proposed a structual modelling approach with sparse kernels. One motivation of this study is to develop a flexible and efficient variable selection method for non-Gaussian data.

In the setting of our method, we conduct a proper decomposition for the likelihood function by choosing the basis functions to be compatible with variable selection in the context of smoothing spline ANOVA (SS-ANOVA). Then we apply basis pursuit (BP) to find the optimal decomposition which minimizes the L_1 norm of the coefficients occurring in the representation. See Chen, Donoho & Saunders (1998) for more elaboration about BP. Our method generalizes the parametric version of LASSO to a more flexible nonparametric form, enjoys the favorable property of shrinking coefficients and gives interpretable results. Furthermore, it works for non-Gaussian data and is valid for most distributions. In this paper we focus on the Bernoulli case. Both the simulation results and examples show the effectiveness of our approach.

2 Introduction

In many demographic medical studies, the outcome Y is a binary variable that takes values 0 and 1. Suppose the distribution of Y depends on the explanatory variables $\{X^\alpha, \alpha = 1, \cdots, d\}$. Define $X = (X^1, X^2, \cdots, X^d)$. d can be very big. It is desirable to select the important covariates and exclude all extraneous variables that fit only sample-specific noise.

Suppose we are given a training set consisting of n examples $(x_1, y_1), \cdots, (x_n, y_n)$. Independent observations y_i, for $i = 1, \cdots, n$, have a Bernoulli distribution, with parameter $p(x_i) = \text{prob}(y_i = 1|x_i)$. Then the negative log likelihood function is $-l(y_i, f(x_i)) = -y_i f(x_i) + b(f(x_i))$, where $f(x) = logit(p(x)), b(f) =$

[2]Corresponding author address: Hao Zhang, Department of Statistics, University of Wisconsin, 1210 W. Dayton St., Madison, WI 53706. Supported by NSF under Grant DMS-0072292, and NIH under Grant EY09946
[3]Supported by NSF under Grant DMS-0072292, and NIH under Grant EY09946
[4]Supported by the Wisconsin Alumni Research Foundation
[5]Supported by NSF under Grant CCR-9972372 and AFOSR under Grant F49620-01-1-0040
[6]Supported by NSF under Grant CCR-9972372 and AFOSR under Grant F49620-01-1-0040
[7]Supported by NIH under Grant EYO3083
[8]Supported by NIH under Grant EYO3083

$log(1+e^f)$. The mean of y_i is $\mu_i = p_i$ and the variance $\sigma_i^2 = p_i(1-p_i)$.

This paper is organized as follows. Firstly, we introduce the SS-ANOVA decomposition, which is then combined with basis pursuit to formulate a general model. Two specific models proposed are the main effects model and the two-factor interaction model. Section 4 describes the important issue of adaptively choosing smoothing parameters. The criteria for selecting variables and numerical computation issues are discussed in Section 5-7. In the end, we will show some simulation results and a detailed analysis of a real set of data.

3 Basis Pursuit via SS-ANOVA

3.1 SS-ANOVA Decomposition

A functional SS-ANOVA decomposition of a function $f = f(x^1, \cdots, x^d)$ in an RKHS $\mathcal{H}$ is

$$\begin{aligned} f(x) &= b_0 + \sum_{\alpha=1}^{d} f_\alpha(x^\alpha) + \sum_{\alpha<\beta} f_{\alpha\beta}(x^\alpha, x^\beta) \\ &+ \text{all higher-order interactions}, \end{aligned} \tag{3.1}$$

where b_0 is constant, f_α's are the main effects, and $f_{\alpha\beta}$'s are the two-factor interactions. It is assumed that $f_\alpha \in \mathcal{H}^{(\alpha)}$, where $\mathcal{H}^{(\alpha)}$ is an RKHS generated by some specified kernel, and $f_{\alpha\beta} \in \mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}$ and so on. Here $\otimes$ denotes the tensor product operation. Under this assumption, $\mathcal{H}$ is constructed as

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

Furthermore, we decompose $\mathcal{H}^{(\alpha)}$ into a parametric part and a smooth part, denoted as $\mathcal{H}^{(\alpha)} = \mathcal{H}_\pi^{(\alpha)} \oplus \mathcal{H}_s^{(\alpha)}$. Here $\mathcal{H}_\pi^{(\alpha)}$ is finite dimensional (the "parametric" part) and $\mathcal{H}_s^{(\alpha)}$ is the ortho-complement of $\mathcal{H}_\pi^{(\alpha)}$ in $\mathcal{H}^{(\alpha)}$ (the "smooth part"). Now $\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}$ is a direct sum of four orthogonal subspaces: $[\mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_\pi^{(\beta)}] \oplus [\mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_s^{(\beta)}] \oplus [\mathcal{H}_s^{(\alpha)} \otimes \mathcal{H}_\pi^{(\beta)}] \oplus [\mathcal{H}_s^{(\alpha)} \otimes \mathcal{H}_s^{(\beta)}]$. Continuing this way results in an orthogonal decomposition of $\mathcal{H}$ into sums of products of finite dimensional spaces, plus "smooth" main effects subspaces, plus two-factor interaction spaces of three possible forms: parametric $\otimes$ smooth, smooth $\otimes$ parametric and smooth $\otimes$ smooth, plus three-factor and higher order interaction subspaces. See Wahba, Wang, Gu, Klein & Klein (1995) and Gao, Wahba, Klein & Klein (2001). Linear models and additive models are special cases in the setup we use.

3.2 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 coefficients among all such decompositions. See Chen et al. (1998) for more details. We propose to pursue the logit function $f \in \mathcal{H}$ in the context of a dictionary based on the SS-ANOVA decomposition with L_1 penalty and call it the likelihood basis pursuit estimate. The variational problem is

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}[-l(y_i, f(x_i)] + J_\lambda(f). \tag{3.2}$$

Here $J_\lambda(f) = |f|$, denoting the L_1 norm of the coefficients of its decomposition. λ is the smoothing parameter to balance the likelihood fit and the penalty. Our approach to flexible basis pursuit/LASSO methods is to use a reproducing kernel arising in the usual penalized likelihood estimation problem to generate basis functions, and then apply the L_1 norm to their coefficients. Any positive-definite function may play the role of a reproducing kernel. In this work, we let $X^\alpha \in [0,1]$ for $\alpha = 1, \cdots, d$, and we use the spline kernel given by equation (10.2.4) with $m = 2$ in Wahba (1990). Denote it by K, and define $k_1(t) = t - \frac{1}{2}$. The so-called representer of evaluation at t in $\mathcal{H}_s^{(\alpha)}$ is defined as $K_t(\cdot) = K(t, \cdot)$.

In the usual penalized likelihood problem with $J_\lambda(f)$ a squared norm or seminorm in an RKHS, the minimizer f_λ has a finite representation in terms of representers and a basis for the parametric part (see Kimeldorf & Wahba (1971)). Then f_α in (3.1) has the form

$$f_\alpha(x^\alpha) = b_\alpha k_1(x^\alpha) + \sum_{j=1}^{n} c_{\alpha,j} K(x^\alpha, x_j^\alpha) \tag{3.3}$$

3.3 Subsets of basis functions

To reduce the computational requirements we will choose a subset of the implied basis functions. They can be chosen either randomly or using a clustering scheme. We denote by N the size of the subset. Next, choose a subset, $\{x_{i_1}, \cdots, x_{i_N}\}$ of the observed explanatory vectors. This subset remains fixed. Then for each $\alpha = 1, \cdots, d$, define

$$\mathcal{H}_*^{(\alpha)} = \text{span}\{1, k_1(\cdot), K_{x_{i_1}^\alpha}(\cdot), \cdots, K_{x_{i_N}^\alpha}(\cdot)\}$$

We solve the minimization problem for each $f_\alpha \in H_*^{(\alpha)}$ instead of in $H^{(\alpha)}$. Thus the solution has the form

$$f_\alpha(x^\alpha) = b_\alpha k_1(x^\alpha) + \sum_{j=1}^{N} c_{\alpha,j} K(x^\alpha, x_{i_j}^\alpha). \tag{3.4}$$

In the usual penalized likelihood setting, the penalized likelihood estimate with $N < n$ is known to be a good approximation to the estimate with the full set of representers. See Xiang & Wahba (1997), Gao et al. (2001) and Lin, Wahba, Xiang, Gao, Klein & Klein (2000). For notational convenience we relabel the observed explanatory variables so that the subset selected above are labeled $\{x_1, \cdots, x_N\}$ in the sections to follow.

3.3.1 Main Effects Model

The main effects spline, also known as the additive spline, is a function of d variables, actually a sum of d functions of one variable. By adopting the expression in (3.4), the main effects model is of the form

$$f(x) = b_0 + \sum_{\alpha=1}^{d} b_\alpha k_1(x^\alpha) + \sum_{\alpha=1}^{d}\sum_{j=1}^{N} c_{\alpha,j} K(x^\alpha, x_j^\alpha).$$

Besides the spline kernel used here, we can also use Gaussian kernel or polynomial kernel. The likelihood basis pursuit estimate f is obtained by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}[-l(y_i, f_i)] + \lambda_\pi \sum_{\alpha=1}^{d}|b_\alpha| + \lambda_s \sum_{\alpha=1}^{d}\sum_{j=1}^{N}|c_{\alpha,j}|, \tag{3.5}$$

where (λ_π, λ_s) are the tuning parameters. It is reasonable to assign different penalties to the parametric terms and the smoothing terms.

3.3.2 Two-factor Interaction Model

The two-factor interaction spline is more complicated than the additive model. The parametric space consists of d parametric main effect subspaces with basis functions $\{k_1(x^\alpha), \alpha = 1, \cdots, d\}$, and the parametric-parametric interaction subspaces with basis as $\{k_1(x^\alpha)k_1(x^\beta)\}$ for all $\alpha < \beta$. The smoothing part consists of d smooth main effect subspaces, $d(d-1)$ subspaces of k_1-K interactions, and $d(d-1)/2$ subspaces of K-K interactions. Thus for any $\alpha < \beta$, the interaction term $f_{\alpha\beta}(x^\alpha, x^\beta)$ in the approximate solution f has the form

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

Different penalties are given to different types of subspaces. We have five tuning parameters: λ_π, $\lambda_{\pi\pi}$, λ_s, $\lambda_{\pi s}$ and λ_{ss}. The optimization problem is: minimize

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}[-l(y_i, f_i)] + \lambda_\pi \sum_{\alpha=1}^{d}|b_\alpha| + \lambda_{\pi\pi}\sum_{\alpha<\beta}|b_{\alpha,\beta}| \\ &+\lambda_{\pi s}\Big(\sum_{\beta>\alpha}\sum_{j=1}^{N}|c^{\pi s}_{\beta\alpha,j}| + \sum_{\alpha<\beta}\sum_{j=1}^{N}|c^{\pi s}_{\alpha\beta,j}|\Big) \\ &+\lambda_s \sum_{\alpha=1}^{d}\sum_{j=1}^{N}|c^{s}_{\alpha,j}| + \lambda_{ss}\sum_{\alpha<\beta}\sum_{j=1}^{N}|c^{ss}_{\alpha\beta,j}| \end{aligned} \tag{3.6}$$

4 GACV

With an abuse of notation, we use λ to represent the collective set of tuning parameters. Since λ controls the tradeoff between the likelihood fit to the training data and the sparsity of coefficients for f_λ, it is important to find a good value of λ. We propose using Generalized Approximate Cross Validation (GACV) to choose λ. It is a proxy for the Comparative Kullback-Liebler (CKL) from the estimate to the true distribution. The derivation of GACV (to appear) is similar to that of GACV for smoothing splines in Xiang & Wahba (1996), Lin et al. (2000) and Gao et al. (2001). In practice, it is expensive to compute GACV directly due to matrix trace computing. We may produce randomized estimates of traces without doing any explicit calculation. If we put a small disturbance $\epsilon = (\epsilon_1, \cdots, \epsilon_n)'$ on $y = (y_1, \cdots, y_n)$, we get a new pseudo data set $y + \epsilon$. Let f_λ^y and $f_\lambda^{y+\epsilon}$ be respectively the estimates with respect to data y and $y + \epsilon$. Let W be the $n \times n$ diagonal matrix with $\sigma_i^2 = p_i(1 - p_i)$ in the ii-th position. The randomized GACV is given by

$$\begin{aligned} ranGACV(\lambda) = \; & \frac{1}{n}\sum_{i=1}^{n}[-y_i f_{\lambda i} + b(f_{\lambda i})] \\ &+ \frac{\epsilon^T(f_\lambda^{y+\epsilon} - f_\lambda^y)}{n}\frac{\sum_{i=1}^{n} y_i(y_i - \mu_{\lambda i})}{\epsilon^T\epsilon - \epsilon^T W(f_\lambda^{y+\epsilon} - f_\lambda^y)}. \end{aligned}$$

Refer to Lin et al. (2000) for its theoretical derivation.

5 Variable Selection Criteria

With the optimal λ_{opt} chosen by GACV, we get the basis pursuit estimate $f_{\lambda_{opt}}$ by minimizing (3.5) or (3.6). How to decide which variables are important and which are not is a key question. Since the L_1

penalty is used for the likelihood basis pursuit estimate, we propose to use the rank of empirical L_1 norm of functions $f_\alpha(x^\alpha)$ and $f_{\alpha\beta}(x^\alpha, x^\beta)$ to decide importance of the variables. Empirical L_2-norm for functions can also be used. Simulation results show they work equally well. The norms are defined as follows:

$$\begin{aligned} L_1(f_\alpha) &= \frac{1}{n}\sum_{i=1}^{n} \mid f_\alpha(x_i^\alpha) \mid \\ L_1(f_{\alpha\beta}) &= \frac{1}{n}\sum_{i=1}^{n} \mid f_{\alpha\beta}(x_i^\alpha, x_i^\beta) \mid \\ L_2(f_\alpha) &= [\ \frac{1}{n}\sum_{i=1}^{n} (f_\alpha(x_i^\alpha))^2\ ]^{\frac{1}{2}} \\ L_2(f_{\alpha\beta}) &= [\ \frac{1}{n}\sum_{i=1}^{n} (f_{\alpha\beta}(x_i^\alpha, x_i^\beta))^2\ ]^{\frac{1}{2}}. \end{aligned}$$

6 Choosing Basis Functions

Since $\mathcal{H}_*^{(\alpha)}$ is a subspace of $\mathcal{H}^{(\alpha)}$, how to choose a subset of basis functions is important. In general, we should make the subspace spanned by the subset of basis functions rich enough to provide a decent fit to the true curve. Note that we are not wasting any data resource here, since all the data points are used for the model fitting, though only part of them are used to generate basis functions. With fixed size of the subset, the complexity of the code will not be heavily affected by the sample size. In our simulations and the example, we choose the subset size $N = 5\%n$ and very good results are obtained. In order to improve the accuracy of the estimation, we can use better sampling methods to choose basis points. For example, cluster analysis sampling may replace simple random sampling.

7 Numerical Computation

Since the objective functions in both (3.5) and (3.6) are not differentiable with respect to b and c, many numerical methods for unconstrained 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 which is solved using Murtagh & Saunders (1983). For each λ, problem (3.5) or (3.6) must be solved twice — once with y (the original problem) and once with $y + \varepsilon$ (the perturbed problem). This often results in hundreds or thousands of individual solves. We employ an efficient solution approach, namely slice modeling proposed by Ferris & Voelker (2000)

We found that MINOS performed very well with the linearly constrained models and returned consistent results. Once we have solutions for the original and perturbed problems at a particular λ, randomized GACV can be calculated. This suggests the approach of solving two problems together for each λ. However, the slice modeling approach suggests the opposite: because fewer changes in the solution take place moving from one λ to another while maintaining the problem type (original or perturbed), previous solutions will have greater impact on future solves if the sequences 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 easily implemented.

8 Simulation

8.1 Additive Model

There are $d = 10$ covariates, $X_1, \cdots, X_{10}$. The true logit function is

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

Among the ten variables, four variables X_1,X_3,X_6 and X_8 are important, and the rest are just noise variables. The sample size $n = 1000$. The basis size $N = 50$. Since we know the true probability function, we are able to compute GACV and CKL as well. CKL gives the optimal $\lambda = (2^{-15}, 2^{-17})$, and GACV gives $(2^{-8}, 2^{-15})$. Figure 1 gives the ranked scores of L_1 and L_2 norms respectively given by CKL and GACV. They both successfully pick out the important variables if we use a proper threshold, say 0.15.

8.2 Two-factor Interaction Model

There are 4 variables and important effects are X_1,X_2 and $X_1 * X_2$. The true logit function is

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

Figure 2 gives the ranked scores of L_1 and L_2 norms respectively given by CKL and GACV. They both successfully pick out the important terms, which are the 1 and 2 main effects and the $1, 2$ interaction.

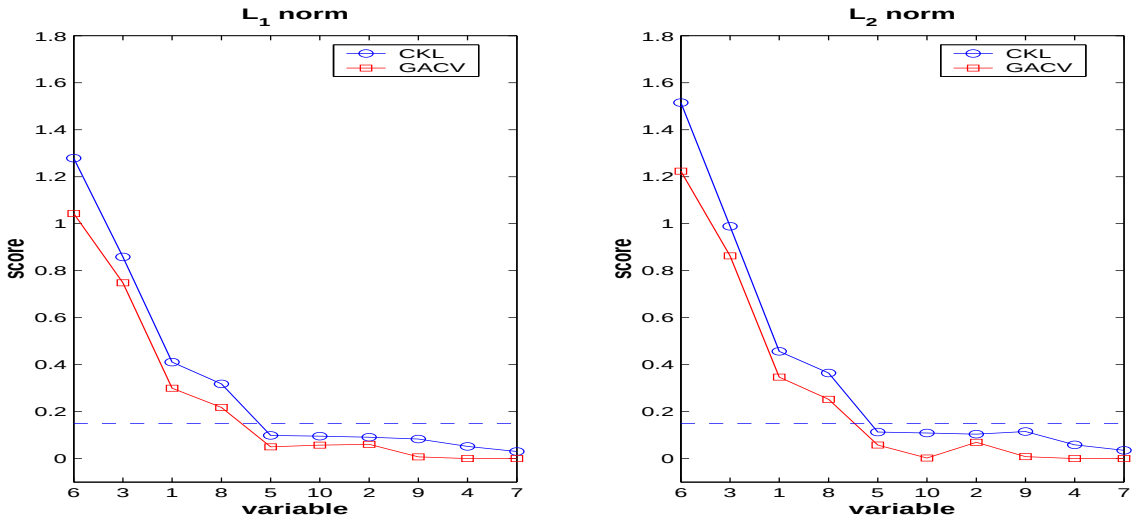

Figure 1: L-scores given by CKL and GACV fits (Additive model)

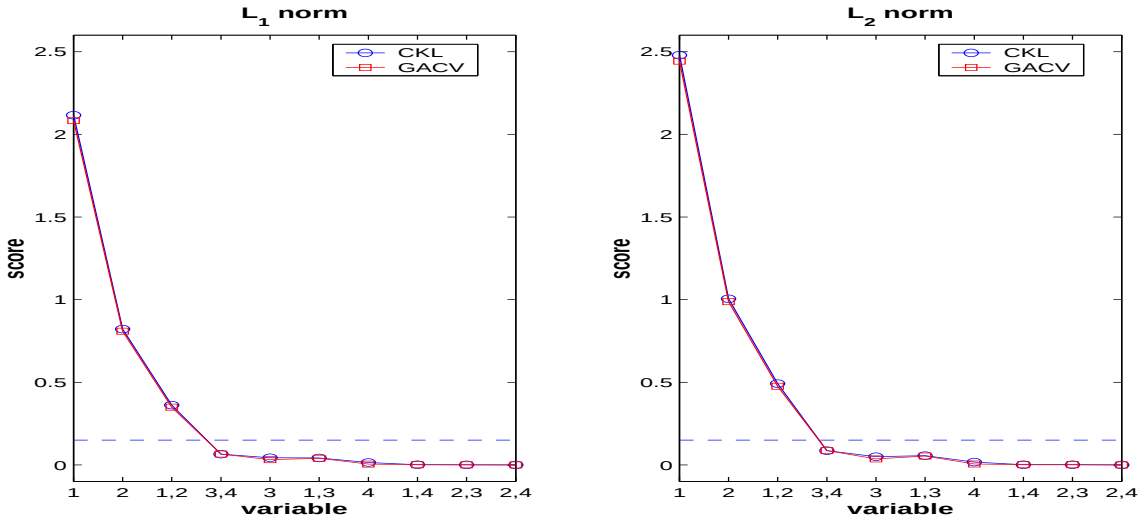

Figure 2: L-scores given by CKL and GACV fits (Two-factor interaction model)

8.3 Wisconsin Epidemiological Study of Diabetic Retinopathy (WESDR)

WESDR is an ongoing epidemiological study of a cohort of patients receiving their medical care in an 11-county area in Wisconsin, who were first examined in 1980-82, then again in 1984-86 and 1990-92. Detailed descriptions are in Klein, Klein, Moss, Davis & DeMets (1989). Here we focus on the younger onset group (sample size 668) of the baseline examination. Progression Y was defined to be 1 for a participant if at the second followup exam, the retinopathy level degraded two scales from the baseline. Seven variables are candidates. X_1 = duration of diabetes at baseline examination; X_2 = glycosylated hemoglobin; X_3 = body mass index; X_4 = systolic blood pressure; X_5 = retinopathy level; X_6 = smoking; X_7 = age at baseline examination. The correlation between AGE and DUR is as high as 0.76.

In real examples, only GACV is used to tune parameters. First, the additive model is fitted excluding AGE. Figure 3 shows that three covariates DUR, GLY and BMI are selected. When including AGE, both AGE and DUR were selected. These results are in general agreement with the penalized likelihood models in Wahba et al. (1995), where the final model was selected in a much more labor intensive method, al-

though there are some differences. This example also shows another big advantage of our approach. It is still valid even when high correlation exists among two covariates (AGE and DUR here), where stepwise methods usually fail in selecting out both. Further discussion will appear elsewhere (in preparation).

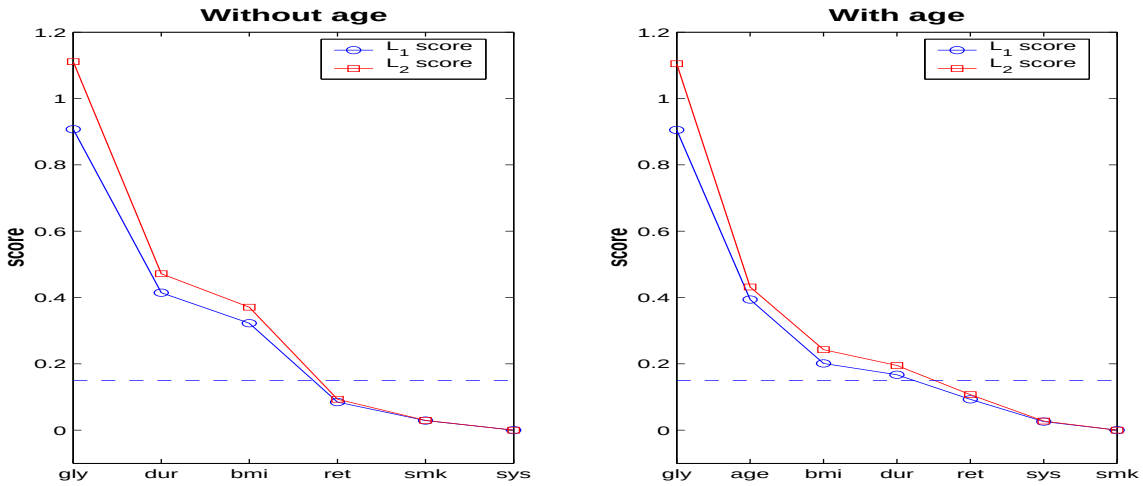

Figure 3: L-scores given by GACV for WESDR data

References

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

Fan, J. & Li, R. Z. (1999), 'Variable selection via penalized likelihood'. to appear in Journal of American Statistical Association, 2001.

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

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 American Statistical Association* **96**, 127–160.

Gunn, S. R. & Kandola, J. S. (2001), 'Structural modelling with sparse kernels'. to appear in Machine Learning.

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

Klein, R., Klein, B., Moss, S. E., Davis, M. D. & DeMets, D. L. (1989), 'The WESDR.IX. Four year incidence and progression of diabetic retinopathy when age at diagnosis is less than 30 years', *Archives of Ophthalmology* **107**, 237–243.

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.

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

Tibshirani, R. J. (1996), 'Regression shrinkage and selection via the lasso', *Journal of 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 WESDR', *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. (1997), Approximate smoothing spline methods for large data sets in the binary case, Technical Report 982, Department of Statistics, University of Wisconsin, Madison, WI.