DEPARTMENT OF STATISTICS
University of Wisconsin
1300 University Ave.
Madison, WI 53706

TECHNICAL REPORT NO. 1131
October 28, 2006

LASSO-Patternsearch Algorithm with Application to Ophthalmology Data

Weiliang Shi[1]
Department of Statistics, University of Wisconsin, Madison WI

Grace Wahba[1]
Department of Statistics, Department of Computer Science and Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison WI

Stephen Wright[2]
Department of Computer Science, University of Wisconsin, Madison WI

Kristine Lee[3], Ronald Klein[4], Barbara Klein[4]
Department of Ophthalmology and Visual Science, University of Wisconsin, Madison WI

[1]Research supported in part by NIH Grant EY09946, NSF Grants DMS-0505636, DMS-0604572 and ONR Grant N0014-06-0095

[2]Research supported in part by NSF Grants SCI-0330538, DMS-0427689, CCF-0430504, CTS-0456694, CNS-0540147 and DOE Grant DE-FG02-04ER25627

[3]Research supported in part by NIH grants EY06594 and EY015286

[4]Research support in part by NIH grants EY06594, EY015286 and Research to Prevent Blindness Senior Investigator Awards

LASSO-Patternsearch Algorithm with Application to Ophthalmology Data

Weiliang Shi, Grace Wahba, Stephen Wright,
Kristine Lee, Ronald Klein and Barbara Klein
University of Wisconsin - Madison

October 30, 2006

Abstract

The LASSO-Patternsearch is proposed, as a two-stage procedure to identify clusters of multiple risk factors for outcomes of interest in large demographic studies, when the predictor variables are dichotomous or take on values in a small finite set. Many diseases are suspected of having multiple interacting risk factors acting in concert, and it is of much interest to uncover higher order interactions when they exist. The method is related to Zhang *et al*(2004) except that variable flexibility is sacrificed to allow entertaining models with high as well as low order interactions among multiple predictors. A LASSO is used to select important patterns, being applied conservatively to have a high rate of retention of true patterns, while allowing some noise. Then the patterns selected by the LASSO are tested in the framework of (parametric) generalized linear models to reduce the noise. Notably, the patterns are those that arise naturally from the log linear expansion of the multivariate Bernoulli density. Separate tuning procedures are proposed for the LASSO step and then the parametric step and a novel computational algorithm for the LASSO step is developed to handle the large number of unknowns in the problem. The method is applied to data from the Beaver Dam Eye Study and is shown to expose physiologically interesting interacting risk factors. In a study of progression of myopia in an older cohort, it is found in this group that the risk for smokers is reduced by taking vitamins, while the risk for non-smokers is independent of the "taking vitamins" variable, which is in agreement with the general result that smoking reduces the absorption of vitamins, and certain vitamins have been

associated with eye health. A second study involving the risk of pigmentary abnormalities demonstrated the counter-intuitive result that lower serum total cholesterol was part of a cluster of risk factors, agreeing with earlier studies using different analytic methods. Application to selection of SNP's related to a particular phenotype is noted briefly.

Contents

1 Introduction

We consider the problem which occurs in large demographic studies when there is some dichotomous outcome with a large number of risk factors that interact in complicated ways to induce elevated risk. Our goal is to search for important patterns of multiple risk factors among a very large number of possible patterns, with results that are easily interpretable.

In the examples motivating this study both dichotomous, polychotomous and continuous variables were available, but for the purposes of this work we assume that all variables are binary or that they have been dichotomized before the analyses considered here. As we will see, this allows the study of higher order interactions than would be possible with risk factors with more than two possible values.

There are many approaches that can model data with binary covariates and binary responses, see, for example [3] [4] [25] [26]. The method that we are proposing here takes account of correlations between clusters of predictors in a particularly transparent way, which is different than the above methods. The main core of the method is global, in that it deals with a very large number of patterns simultaneously, as opposed to tree-structured methods or greedy (sequential) algorithms which constitute much of the literature in this area. We will be using LASSO ideas on problems with many potential binary predictors and binary outcomes with many possible patterns of predictor variables to consider. For Gaussian data the LASSO was proposed in [27] as a variant of linear least squares ridge regression with many predictor variables. As proposed there, the LASSO minimized the residual sum of squares subject to a constraint that the sum of absolute values of the coefficients be less than some constant, say t. This is equivalent to minimizing the residual sum of squares plus a penalty which is some multiple λ (depending on t) of the sum of absolute values (l_1 penalty). It was demonstrated there that this approach tended to set many of the coefficients to zero, resulting in a sparse model, a property not generally obtaining with quadratic penalties. A similar idea was exploited in [5] to select a good subset of an overcomplete set of nonorthogonal

wavelet basis functions. The asymptotic behavior of LASSO type estimators was studied in [15], and [23] discussed computational procedures in the Gaussian context. More recently [7] discussed variants of the LASSO and methods for computing the LASSO for a continuous range of values of λ, in the Gaussian case. Variable selection properties of the LASSO were examined in [18] in some special cases, and many applications can be found on the web. In the context of nonparametric ANOVA decompositions [33] used an overcomplete set of basis functions obtained from a Spline ANOVA model, and used ℓ_1 penalties on the coefficients of main effects and low order interaction terms, in the spirit of [5]. The present paper uses some ideas from [33]. Other work in the context of nonparametric ANOVA decompositions have implemented ℓ_1 penalties along with quadratic (reproducing kernel square norm) penalties to take advantage of their sparsity-introducing properties, see for example [11] [17] [32].

In this paper, we present a two-stage method for logistic regression, beginning with a complete (in a sense to be defined) or otherwise overly large set of basis functions, embodying a large number of potential patterns. In the first stage the number of potential patterns is reduced, in the spirit of [5] and [33], by penalized likelihood with an l_1 penalty on the coefficients, the penalty parameter being chosen here in a conservative way for model selection. This is followed by a tuned parametric logistic regression in the second stage, to select out significant patterns. Finally, pseudo-significance tests can be carried out by scrambling the response data in various ways to see to what extent, if any the two stage method proposed will generate spurious patterns like the ones found. In constructing the patterns, for ease of interpretation we assume that it is known *a priori* in which direction each variable promotes risk, if it does. In practice this condition may be violated by a very small number of variables without losing interpretability. With this assumption the method is turning out to be extremely useful in teasing out and interpreting high order interactions in complex demographic data sets that are not readily found by more classical methods.

The rest of the article is organized as follows. In Section 2 we describe the first (LASSO) step of the LASSO-Patternsearch algorithm, including choosing the smoothing parameter by Generalized Approximate Cross Validation (GACV) and a discussion of an efficient algorithm for the LASSO step. Section 3 describes the second step of the LASSO-Patternsearch algorithm, along with a modification of the GACV for use in a sequential pattern deletion procedure. Section 4 presents two simulation examples, and Section 5 presents two real examples from Beaver Dam Eye Study. Section 6 notes some generalizations, and, finally, Section 7 gives a summary and conclusions. Appendix A relates the patterns to the log

linear expansion of the multivariate Bernoulli distribution, Appendix B derives a simplified expression for the GACV, Appendix C gives details of the specially designed code for the LASSO which is capable of handling a very large number of patterns simultaneously, and Appendix D gives a lemma related to erroneous risk directions.

2 The LASSO-Patternsearch Algorithm

2.1 The LASSO-Patternsearch Algorithm - Step 1

Considering n subjects, for which p variables are observed, we first reduce continuous variables to "high" or "low" in order to be able to examine *very many* variables and *their interactions* simultaneously. We will assume that for all or most of the the p variables, we know in which direction they are likely to affect the outcome or outcomes of interest, if at all. For some variables, for example smoking, it is clear for most endpoints in which direction the smoking variable is likely to be "bad" if it has any effect, and this is true of many but not all variables. For some continuous variables, for example systolic blood pressure, higher is generally "worse", but extremely low can also be "bad". For continuous variables, we need to initially assume the location of a cut point on one side of which the variable is believed to be "risky" ("high") and the other side "not risky" ("low"). For systolic blood pressure that might, for example, be 140mmHg. For an economic variable that might be something related to the current standard for poverty level. If the "risky" direction is known for most variables the results will be readily interpretable. Each subject thus has an attribute vector of p zeroes and ones, describing whether each of their p attributes is on one side or the other of the cutoff point. The LASSO-Patternsearch approach described below is able to deal with high order interactions and very large p. The data is $\{y_i, x(i), i = 1, \cdots, n\}$, where $x(i) = (x_1(i), x_2(i), \cdots, x_p(i))$ is the attribute vector for the ith subject, $x_j(i) \in \{0, 1\}$. Define the basis functions $B_{j_1 j_2 .. j_r}(x) = \prod_{\ell=1}^{r} x_{j_\ell}$, that is, $B_{j_1 j_2 .. j_r}(x) = 1$ if $x_{j_1}, ..., x_{j_r}$ are all 1's and 0 otherwise. We will call $B_{j_1 j_2 .. j_r}(x)$ an rth order pattern. Including all possibilities, we have a complete set of $N_B = 2^p - 1$ such patterns plus the constant function μ, spanning all possible patterns. See Appendix A for a discussion of how these patterns arise from the log linear parametrization of the multivariate Bernoulli distribution. If only patterns of order up to and including $r < p$ are used, then there will be $N_B = \sum_{\nu=1}^{r} \binom{p}{\nu}$ patterns. Letting

$p(x) = Prob\{y = 1|x\}$ and the logit $f(x) = \log[p(x)/(1-p(x))]$, we estimate f by minimizing

$$I_\lambda(y, f) = \mathcal{L}(y, f) + \lambda J(f) \tag{1}$$

where $\mathcal{L}(y, f)$ is $\frac{1}{n}$ times the negative log likelihood:

$$\mathcal{L}(y, f) = \frac{1}{n}\sum_{i=1}^{n}[-y_i f(x(i)) + \log(1 + e^{f(x(i))})] \tag{2}$$

with

$$f(x) = \mu + \sum_{\ell=1}^{N_B-1} c_\ell B_\ell(x), \tag{3}$$

where we are relabeling the $N_B - 1$ (non-constant) patterns from 1 to $N_B - 1$, and

$$J(f) = \sum_{\ell=1}^{N_B-1} |c_\ell|. \tag{4}$$

In Step 1 of the LASSO-Patternsearch we minimize (1) using the GACV criteria to choose λ. The next section describes the GACV and the kinds of results it can be expected to produce.

2.2 Generalized Approximate Cross Validation (GACV)

The tuning parameter λ in (1) balances the trade-off between data fitting and the sparsity of the model. The bigger λ is, the sparser the model is, and vice versa. The choice of λ is generally a crucial part of penalized likelihood methods and machine learning techniques like the Support Vector Machine. For the smoothing spline models with Gaussian data, [29] proposed the ordinary cross validation (OCV). The generalized cross validation (GCV), derived from OCV was suggested by [6][10], and theoretical properties were obtained in [19]. For smoothing spline models with Bernoulli data and quadratic penalty functionals, [31] derived Generalized Approximate Cross Validation (GACV) from an OCV estimate following the method used to obtain GCV. In [33] GACV was extended to the case of Bernoulli data with l_1 penalties. We will use the GACV of [33], however, in the present context the expression for it reduces to a much simpler form, so we shall provide a simplified derivation.

The GACV score derivation begins with a leaving out one likelihood to minimize an estimate of the comparative Kullback-Leibler distance (CKL) between the true and estimated

model distributions. The ordinary leave out one cross validation score for CKL is

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[-y_i f_\lambda^{[-i]}(x(i)) + \log(1 + e^{f_\lambda(x(i))})], \tag{5}$$

where f_λ is the minimizer of the objective function (1), and $f_\lambda^{[-i]}$ is the minimizer of (1) with the ith data point left out. In Appendix B, a series of approximations and an averaging step are described, which results in the much easier to compute $GACV(\lambda)$ score:

$$GACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[-y_i f_{\lambda i} + \log(1 + e^{f_{\lambda i}})] + \frac{1}{n}(\frac{1}{n}trH)\frac{\sum_{i=1}^{n} y_i(y_i - p_{\lambda i})}{(1 - \frac{1}{n}N_{B_0})}, \tag{6}$$

here $H = B^*(B^{*\prime}WB^*)^{-1}B^{*\prime}$, where W is the $n \times n$ diagonal matrix with iith element the estimated variance at $x(i)$ and B^* is the $n \times N_{B_0}$ design matrix for the N_{B_0} non-zero c_ℓ in the model, see Appendix B.

2.3 GACV as a conservative pattern selector

The GACV criteria, being a predictive criteria rather than a direct model selection criteria (e. g. say, like BIC) is conservative in selecting basis functions (patterns) in that it has a high probability of including true patterns but at the expense of including noise patterns. We consider this a desirable feature, since in exploratory analysis, the cost of losing a true pattern is greater than that of including a noise pattern, since confirmatory analysis can take place later. We illustrate this with a small simulation.

In this example, we generate 7 independent variables from Bernoulli(0.5). The sample size is 1000 and the true logit function contains two main effects, one size-two pattern and one size-three pattern:

$$f(x) = -2 + 1.5B_1(x) + 1.5B_7(x) + 1.5B_{23}(x) + 2B_{456}(x).$$

The experiment was repeated 100 times. Table 1 shows the appearance frequency and estimate of the patterns in the model selected by Step 1 of the LASSO-Patternsearch. The LASSO never misses any of the patterns in the 100 runs. However, the average model size (the number of patterns) is 13.9 and the standard deviation is 4.32. This model size is much bigger than the true model size, 4. Figure 1 is the histogram of model sizes. It can be seen that the smallest model in the 100 runs still has 8 patterns.

Table 1 also presents the mean estimates of the parameters (numbers in the parenthesis are standard deviations). Clearly they are much smaller than the truth. The l_1 penalty is

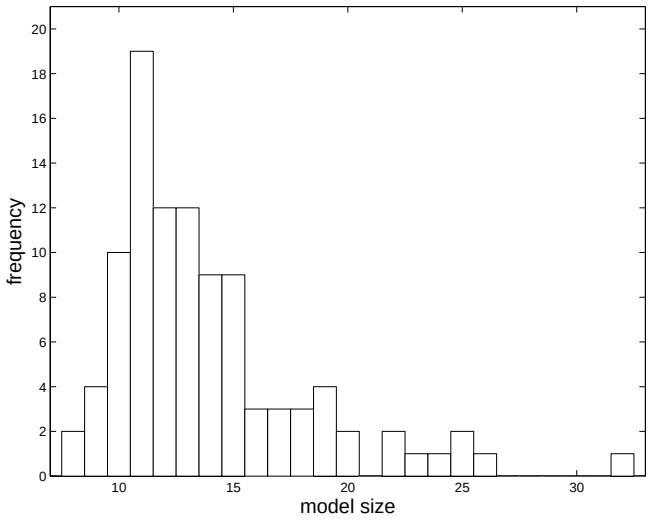

Figure 1: The histogram of model sizes given by LASSO-Patternsearch Step 1

trying to shrink them, but the excessive number of noise patterns is no doubt contributing to this.

	Constant	B_1	B_7	B_{23}	B_{456}
frequency	100	100	100	100	100
estimate	-1.68 (0.172)	1.14 (0.197)	1.13 (0.215)	1.15 (0.213)	1.41 (0.348)
true coef	-2	1.5	1.5	1.5	2

Table 1: The appearance frequency and estimate of patterns in the model by Step 1

2.4 Computation

From a mathematical point of view, this optimization problem (1) is the same as the likelihood basis pursuit (LBP) algorithm in [33], but with different basis functions. The solution can easily be computed via a general constrained nonlinear minimization code such as MATLAB's `fmincon` on a desktop, for a range of values of λ, provided n and N_B are not too large. However, for extremely large data sets with a more than a few attributes p (and therefore a large number N_B of possible basis functions), the problem becomes much more difficult to solve computationally with general optimization software, and algorithms that

exploit the structure of the problem are needed. We use an algorithm that formulates the problem as nonlinear optimization with nonnegativity constraints and makes use of the fact that first and second partial derivatives of the function in (1) with respect to the coefficients c_ℓ are easy to compute analytically once the function has been evaluated at these values of c_ℓ. It is not economical to compute the full Hessian (the matrix of second partial derivatives), so the algorithm computes only the second derivatives of the log likelihood function $\mathcal{L}$ with respect to those coefficients c_ℓ that appear to be nonzero at the solution. (Generally, just a small fraction of these $N_B - 1$ coefficients are nonzero at the solution.) It applies a gradient projection approach, using the partial Hessian to compute a Newton-like step in the apparently-nonzero coefficient. Rapid convergence is obtained once the correct nonzero coefficients are identified correctly.

The algorithm is particularly well suited to solving the problem (1) for a number of different values of λ in succession; the solution for one value of λ provides an excellent starting point for the minimization with a nearby value of λ.

Further details of this approach can be found in Appendix C.

3 The LASSO-Patternsearch Algorithm - Step 2

In Step 2 of LASSO-Patternsearch algorithm, the N_{B_0} patterns surviving Step 1 are entered into a linear logistic regression model, (`glmfit` in MATLAB is used here). Pattern coefficients and p-values for the significance level of each coefficient are obtained. With $n \sim 1000$ and the pattern sizes obtained in Figure 1, there was no difficulty executing the code. The simplest procedure is just to choose a cutoff p-value arbitrarily or consistent with subjective prior information concerning the scientific problem at hand, and to delete all patterns with a p-value over the cutoff. Here the p-value cutoff is "tuned" as follows: The patterns are rank ordered by their p-values, from high to low, as $1, 2, \cdots, N_{B_0}$. A tuning score is computed for the model with all N_{B_0} patterns. Then the rank 1 pattern is deleted and the model refitted with the remaining $N_{B_0} - 1$ patterns and the tuning score computed. Then the rank 2 pattern is deleted, and so forth, until there are no patterns left. A final model is chosen from the pattern set with the best tuning score. It is possible for the rank orders to change as patterns are removed, however, for simplicity the initial rank ordering is maintained, and it is believed that the results would not, in general be different. Note that all subset selection is not being done, since this will introduce an overly large number of degrees of freedom into

the process. If a tuning set is available, then it can be used to create the tuning score.

It is easy to derive a GACV score which could be used as the tuning score for each model. The Leave-Out-One Lemma still holds, the proof is simple, and the derivation of the GACV score follows that in Appendix B. The only difference is that the likelihood function is now smooth with respect to the parameters so the robust assumption that appears in Appendix B is not needed. All columns of the design matrix will be in the GACV score. That's not a problem because the matrix is much smaller than that in Step 1. Let B_s be the design matrix here, then the GACV score for logistic regression is the same as (6) with $H = B_s(B_s'WB_s)^{-1}B_s'$. However, it was found that using the GACV score was still conservative in problems at the scale (e.g. n, N_{B_0}, signal strength) of the data sets of interest here, retaining too many noise patterns, both in simulations and in retrospective examination of analyses (to be described) of observational data. Thus, a weight was inserted in front of the second term, the so called "optimism" part of the GACV score. In the spirit of the variable selection criteria BIC the weight $\log n/2$ was chosen, to obtain the BIC-type GACV (BGACV):

$$BGACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[-y_i f_{\lambda i} + \log(1+e^{f_{\lambda i}})] + \frac{\log n}{2}\frac{1}{n}(\frac{1}{n}trH)\frac{\sum_{i=1}^{n} y_i(y_i - p_{\lambda i})}{(1-\frac{1}{n}N_{B_0})}. \qquad (7)$$

The BGACV scores are compared for each model in the ordered sequence of models obtained above, and the model with the smallest BGACV score is taken as the final model.

The following is a summary of the LASSO-Patternsearch algorithm:

1. Solve (1) and choose λ by GACV. Keep the patterns with non-zero coefficients.

2. Put the patterns with non-zero coefficients from Step 1 into a logistic regression model, generate the sequence of reduced models, evaluate the BGACV score, and select the model with the smallest BGACV score.

For simulated data, the results can be compared with the simulated model. For observational data, a post-analysis step involving scrambling the data will be used to validate the results.

4 Simulation Studies

In this section we study the empirical performances of the LASSO-Patternsearch algorithm through two simulated examples. The first example continues with simulated data of Section

2.3. There the input variables were generated independently. In the second example below the input variables are generated with a correlation structure. In each experiment setting, we generate 100 data sets for fitting and tuning. We summarize the model size, the frequency of each pattern appearing in the logit function and the estimates of coefficients of the patterns over the 100 runs.

4.1 Example 1: Independent Covariates Data

We use the same example as the one in Section 2.3. We had 7 independent variables generated from Bernoulli(0.5) and the true logit function was $f(x) = -2 + 1.5B_1(x) + 1.5B_7(x) + 1.5B_{23}(x) + 2B_{456}(x)$. We reported the results from Step 1 of LASSO-Patternsearch and showed why we need the second step. Here we are going to see the performance of the whole algorithm.

	Constant	B_1	B_7	B_{23}	B_{456}
frequency	100	99	100	100	98
estimate	-1.99(0.193)	1.48(0.217)	1.51(0.193)	1.51(0.219)	2.05(0.314)
true coef	-2	1.5	1.5	1.5	2

Table 2: The appearance frequency and estimate of patterns in the final model

Table 2 shows the appearance frequency and estimated coefficients of the patterns in the final model. We only missed B_1 once and B_{456} twice in 100 runs. Both B_7 and B_{23} were selected in all 100 runs. The third row presents the mean and standard deviation of estimated coefficients of the patterns. Clearly the estimates are much closer to the truth than those in Table 1. Figure 2 Left panel shows a histogram of the final model sizes. For over 70% of the time, the right size, 4, was obtained. Sizes 5 and 6 occurred most of the other times. Very few had a model size greater than 6. No model sizes less than 4 were obtained.

To see what would happen if we made a mistake in determining which direction is bad for the risk factors, we recoded x_3 by switching 0 to 1 and 1 to 0, and ran the algorithm again on 100 sets of simulated data.

In this case, the true logit function becomes

$$f(x) = -2 + 1.5B_1(x) + 1.5B_7(x) + 1.5B_2(x) - 1.5B_{23}(x) + 2B_{456}(x).$$

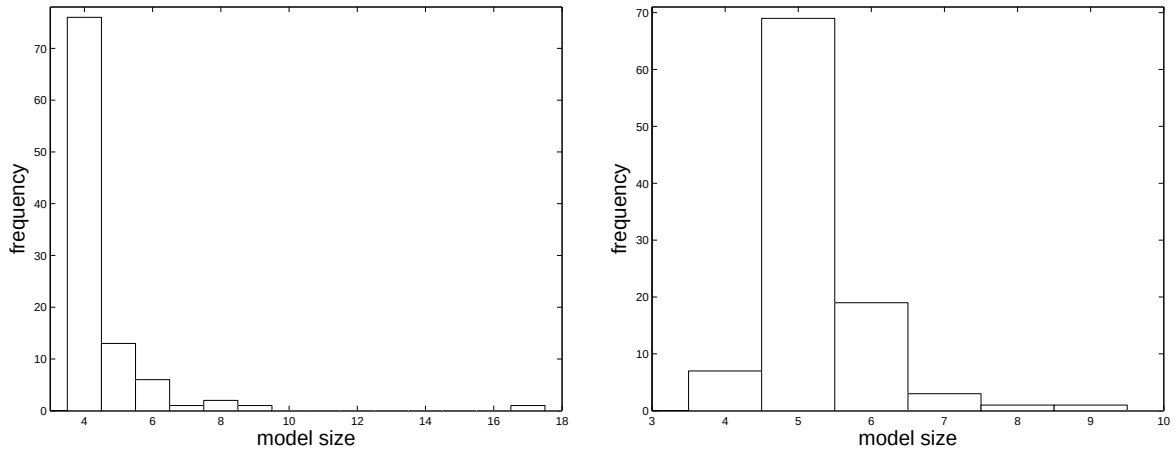

Figure 2: Histogram of sizes of the final models. Left panel: Simulation Example 1. Right panel: Simulation Example 2.

92 out of 100 times, B_2 shows up with an estimate close to 1.5 and B_{23} shows up with an estimate close to -1.5. Therefore, we can tell that the direction of x_3 is wrong. This is a general phenomenon, coding in the wrong direction can only increase the number of patterns and include negative coefficients, and too many coding "errors" will make the results difficult to interpret. See Appendix D.

4.2 Example 2: Correlated Covariates Data

All the covariates in the last example are independent. In the real world however, risk factors within a pattern are frequently correlated to various degrees. In this example, we intentionally introduce some correlation among the risk variables and increase the number of risk factors to 8 and the sample size to 2000. Only the first five variables are correlated and the first seven are related to the risk. x_1 is generated from a Bernoulli distribution with a success rate of 0.5. x_2 and x_3 are conditionally independent given x_1. When x_1 is 1, both of them follow Bernoulli(0.7). When x_1 is 0, both of them follow Bernoulli(0.5). x_4 and x_5 are conditionally independent given x_2 and x_3. When both x_2 and x_3 are 1, x_4 and x_5 follow Bernoulli(0.7). When only one of x_2 and x_3 is 1, x_4 and x_5 follow Bernoulli(0.5). When none of x_2 and x_3 is 1, x_4 and x_5 follow Bernoulli(0.3). x_6, x_7 and x_8 are independently generated from Bernoulli(0.5). They are also independent of the first five variables. The true logit

function f is

$$f(x) = -2 + 1.5B_5(x) + 1.5B_{23}(x) + 1.5B_{34}(x) + 2B_{123}(x) + 2B_{4567}(x)$$

The simulation is run 100 times and the result is shown in Table 3.

	constant	B_5	B_{23}	B_{34}	B_{123}	B_{4567}
frequency	100	99	98	100	99	87
estimate	-2.01(0.12)	1.50(0.15)	1.54(0.21)	1.49(0.18)	2.02(0.36)	2.16(0.44)
true coef	-2	1.5	1.5	1.5	2	2

Table 3: The appearance frequency and estimate of patterns in simulation example 2

All terms showed up almost every time except the size-four pattern I_{4567}. It appeared 87 times. It is very difficult to get high appearance frequency for high order patterns because often times, very few observations have these patterns. For example, suppose the four covariates are independent (they are not in the above example) and their occurrence rates are p_1, p_2, p_3 and p_4. We can only expect to see $np_1p_2p_3p_4$ subjects with this pattern, where n is the sample size. This feature, on the other hand, tells us that the variables in high order patterns are probably correlated.

Figure 2 Right panel shows the histogram of model sizes in the 100 runs. Again, the correct model size was found around 70% of the time. Unlike simulation example 1, there were a few cases where the model size is smaller than the true model size. This is caused by the low appearance frequency of B_{4567}.

5 The Beaver Dam Eye Study

The Beaver Dam Eye Study (BDES) is an ongoing population-based study of age-related ocular disorders. The purpose of the study is to collect information on the prevalence and incidence of age-related cataract, macular degeneration, and diabetic retinopathy. Between 1987 and 1988, a private census identified 5924 people aged 43 through 84 years in Beaver Dam, WI. 4926 of these people participated the baseline exam (bd1) between 1988 and 1900. Five (bd2), ten (bd3) and fifteen (bd4) year follow-up data have been collected and there have been several hundred publications on this data. A detailed description of the study can

be found in [14]. We will focus on two types of eye diseases here: the progression of myopia and the pigmentary abnormality for women.

5.1 Progression of Myopia

Myopia, or nearsightedness, is one of the most prevalent world-wide eye conditions. Myopia occurs when the eyeball is slightly longer than usual from front to back for a given level of refractive power of the cornea and lens and people with myopia find it hard to see objects at a distance. Approximately one-third of the population experience this eye problem and in some countries like Singapore, more than 70% of the population have myopia upon completing college. It is believed that myopia is related to various environmental risk factors as well as genetic factors. The ten-year change of refraction for the BDES population was summarized in [16]. We will study the five-year progression of myopia on an older cohort aged 60 through 69 years. We focus on a small age group since the change of refraction differs for different age groups.

Based on [16] and some preliminary analysis we tried on this data, we choose seven risk factors: *sex, inc, jomyop, catct, pky, asa* and *vtm*. Descriptions and binary cut points are found in Table 4.

code	variable	unit	higher risk
sex	sex		Male
inc	income	\$1000	< 30
jomyop	juvenile myopia	yes/no	< 21
catct	cataract	severity 1-5	4-5
pky	packyear	pack per day×years smoked	>30
asa	aspirin	yes/no	not taking
vtm	vitamin	yes/no	not taking

Table 4: The variables in the progression of myopia example. The forth column shows which direction is risky. The definition of juvenile myopia is myopic before the age of 21.

For most of these variables, we know which direction is bad. For example, male gender is a risk factor for most diseases and smoking is never good. The binary cut points are somewhat subjective here. Regarding *pky*, a pack a day for 30 ears, for example, is a fairly

substantial smoking history. *catct* has five levels of severity and we cut it at the third level. Aspirin *(asa)* and vitamin supplements *(vtm)* are commonly taken to maintain good health so we treat not taking them as risk factors. y is assigned 1 if a person's refraction has changed more than -0.75 diopters from the baseline exam to the five-year follow-up and 0 otherwise. There are 1374 participants in this age group at the baseline examination, of which 952 have measurements of refraction at the baseline and the five-year follow-up. Among the 952 people, 76 have missing values in the covariates. We assume that the missing values are missing at random for both response and covariates, although this assumption is not necessarily valid. However the examination of the missingness and possible imputation are beyond the scope of this study. Our final data consists of 876 subjects without any missing values in the seven risk factors.

Table 5 lists all the patterns selected by Step 1 of the LASSO-Patternsearch. There are 12 patterns plus the constant. We can easily handle 13 variables in a linear logistic regression, with the sample size here. Figure 3 shows the results of Step 2 of the LASSO-Patternsearch. Each point represents a pattern, in the order as they appeared in Table 5. The red or green circle is the estimated coefficient of that pattern and the bar is the 96.92% (selected by BGACV) confidence interval. The red ones are significant (confidence interval doesn't contain zero) and the green ones are not (confidence interval contains zero). Therefore, the model after Step 2 consists of four basis functions, 1 for a main effect and 3 additional patterns. They are patterns 2 (*catct*), 8 (*pky vtm*), 12 (*sex inc jomyop asa*) and 13 (*sex inc catct asa*).

Pattern		Estimate	Pattern		Estimate
1	constant	-3.4435	2	catct	2.6987
3	asa	0.8336	4	vtm	0.1128
5	inc pky	-0.3034	6	catct asa	0.0089
7	catct vtm	-0.7967	8	pky vtm	1.2418
9	sex asa vtm	-0.6961	10	inc catct pky	0.4199
11	inc pky vtm	0.4848	12	sex inc jomyop asa	2.1071
13	sex inc catct asa	1.4860			

Table 5: Patterns selected at Step 1 in the progression of myopia data.

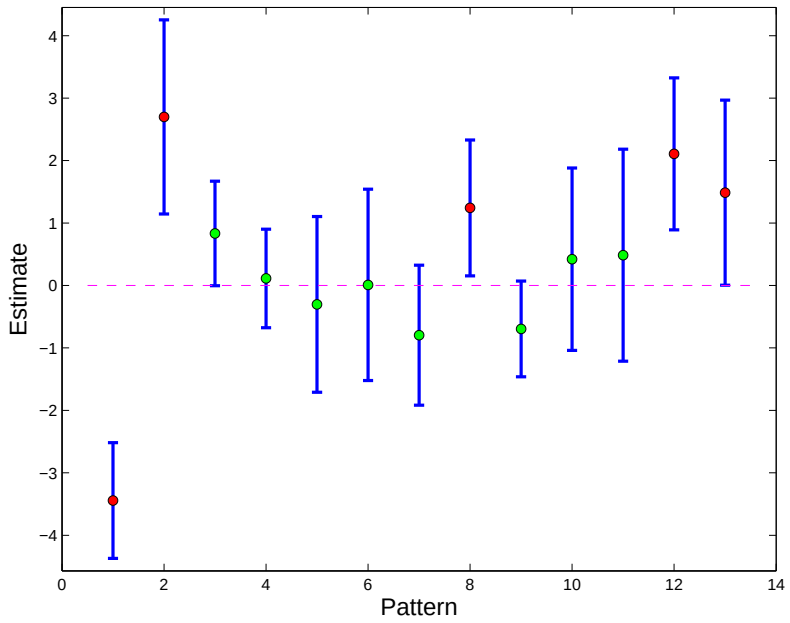

Figure 3: The patterns that survived Step 1 of the Lasso-Patternsearch and the confidence intervals at Step 2 of the Lasso-Patternsearch in the progression of myopia data

The (refitted) model is $f(catct, pky, vtm, sex, inc, jomyop, asa) = -3.29 + 2.42 * cact + 1.18 * pky * vtm + 1.84 * sex * inc * jomyop * asa + 1.08 * sex * inc * cat * asa$.

The pattern (*pky vtm*) catches our attention because the pattern effect is very strong and both variables are controllable. This model tells us that the distribution of y, progression of myopia conditional on $pky = 1$ depends on *catct*, as well as *vtm* and higher order interactions, but progression of myopia conditional on $pky = 0$ is independent of *vtm*. This interesting effect can easily be seen by going back to a table of the original *catct, pky* and *vtm* data (Table 6). The denominators in the risk column are the number of persons with the given pattern and the numerator are the number of those with $y = 1$. The first two rows list the heavy smokers with cataract. Heavy smokers who take vitamins have a smaller risk of having progression of myopia. The third and forth rows list the heavy smokers without cataract. Again, taking vitamins is protective. The first four rows suggest that taking vitamins in heavy smokers is associated with a reduced risk of getting more myopic. The last four rows list all non-heavy smokers. Apparently taking vitamins does not similarly reduce the risk of becoming more myopic in this population. Actually, it is commonly known that smoking significantly decreases the serum and tissue vitamin level, especially Vitamin C and Vitamin

E, for example [8]. Our data suggest a possible reduction in the progression of myopia in persons who smoke who take vitamins. However, our data are observational and subject to uncontrolled confounding. A randomized controlled clinical trial would provide the best evidence of any effect of vitamins on progression of myopia in smokers.

catct	pky	not taking vitamins	risk
1	1	1	$17/23 = 0.7391$
1	1	0	$7/14 = 0.5000$
0	1	1	$22/137 = 0.1606$
0	1	0	$2/49 = 0.0408$
1	0	1	$18/51 = 0.3529$
1	0	0	$19/36 = 0.5278$
0	0	1	$22/363 = 0.0606$
0	0	0	$13/203 = 0.0640$

Table 6: The raw data for cataract, smoking and not taking vitamins.

5.2 Scrambling the Myopia Data

To further investigate the significance of the patterns found by our algorithm, we randomly scramble the response data while keeping the attribute data fixed and apply the entire LASSO-Patternsearch algorithm to see how frequently false patterns are selected. The procedure is repeated 600 times and Figure 4 shows the $log_{10}p$-values of patterns of different sizes that are extracted from the scrambled data. These p-values are computed in the final logistic regression model. The responses have been decoupled from the predictors, but some noise patterns still show up occasionally. The four plots from left to right and top to bottom are for patterns of size 1, 2, 3 and 4, respectively. The red line in the first plot is the $log_{10}p$-value of *catct* in the original model. Clearly it is much smaller than all the noise patterns. The red line in the second plot corresponds to the pattern (*pky vtm*). Again it is much smaller than the noise. The lower red line in the bottom right plot corresponds to the pattern (*sex inc jomyop asa*) and it is almost the same as the smallest p-value of the noise patterns. The upper red line of the last plot is the p-value of the pattern (*sex inc catct asa*). It is weaker than all the eleven noise patterns of size four which were generated. The

significance level of this pattern in the final model was .031 (= 1-.969), (the last pattern to be included). Comparing (11+1)/600) = .02 we get a rough agreement between the false pattern rate for this pattern from the simulation and the final model, although it could be argued that a higher noise rate could be expected to appear in the scrambling due to the two step "mining".

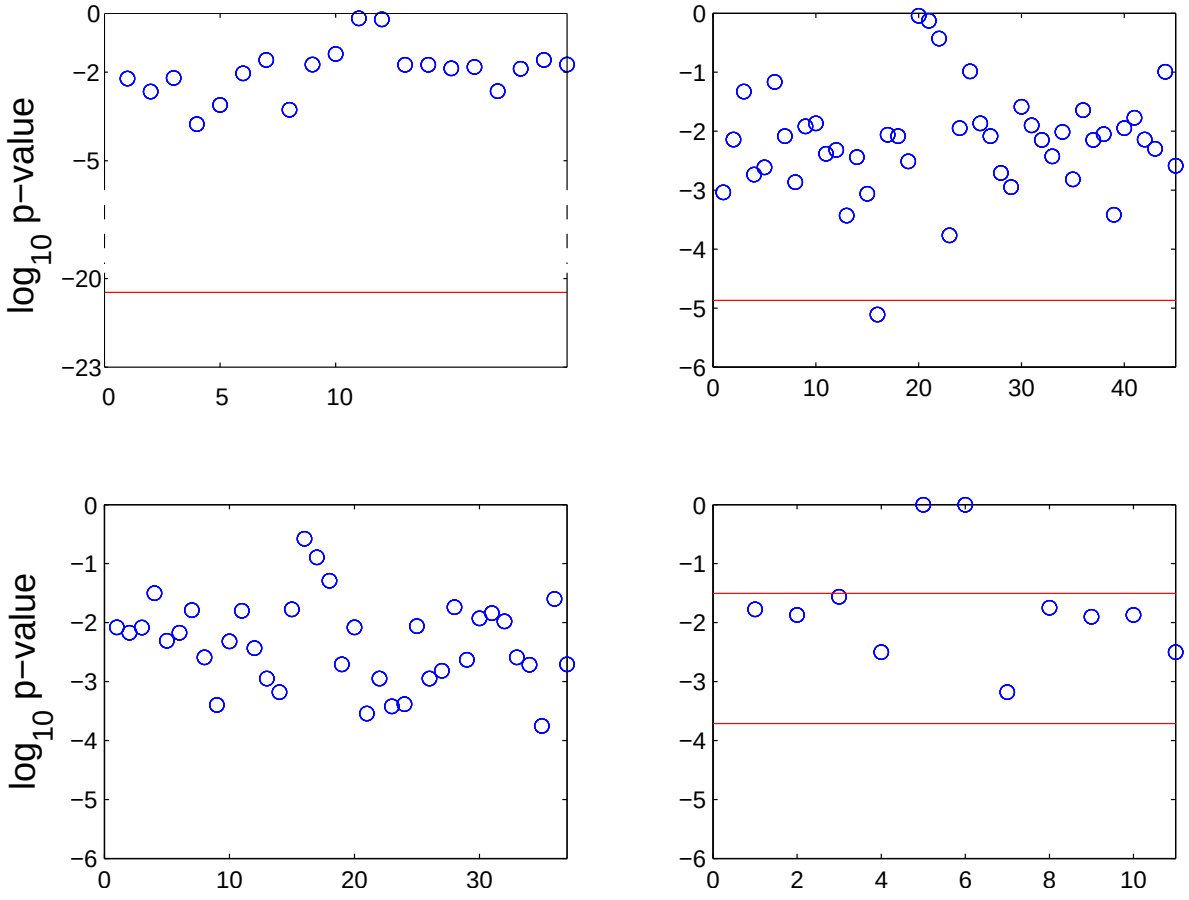

Figure 4: Scatter plots of $log_{10}p$-values of noise patterns surviving the LASSO-Patternsearch in the completely reassigned data, 600 runs. Left to right, top to bottom: patterns of size 1, 2, 3 and 4.

5.3 Pigmentary Abnormality in Women

Pigmentary abnormality is an early sign of age-related macular degeneration and an important cause of visual loss in people 65 years of age or older. It is defined by the presence of

retinal depigmentation or increased retinal pigmentation in association with retinal drusen. Based on previous work, age is a strong risk factor for pigmentary abnormality and age-related macular degeneration. In [13] the association between cardiovascular disease and its risk factors and the incidence of age-related macular degeneration was examined. It was shown in [12] that hormone replacement therapy has a weak protective effect while [24] and [22] said a history of heavy alcohol consumption has a deleterious effect for pigmentary abnormality. The association of pigmentary abnormality with six attributes was studied by the "univariate" penalized logistic regression in [21]. The same data was studied by [9] in the setting of penalized multivariate logistic regression. The six variables used there were: current usage of hormone replacement therapy, history of heavy drinking, body mass index, age, systolic blood pressure and serum total cholesterol.

Similar to the progression of myopia data, we only chose people aged 60 through 69 years as age is a special risk factor. The seven variables we use and their binary cut points are listed in Table 7. One interesting variable here is *chol*. We all know that high serum total cholesterol is associated with higher risk of cardiovascular disease morbidity and mortality, but both [21] and [9] found a protecting effect of serum total cholesterol for pigmentary abnormality. Therefore, we define lower serum total cholesterol as higher risk accordingly. High systolic blood pressure and high body mass index increased the risk and current usage of hormone replacement therapy was seen to reduce the risk in [9], where serum total cholesterol, systolic blood pressure and body mass index were treated as continuous variables.

All the risk factors are the measurements at the baseline examination and the response is the status of pigmentary abnormality at the five-year follow-up. There are 2762 women at the baseline exam, of whom 749 are between the age of 60 and 69 years. After deleting the missing values, there remains 561 people. The incidence rate of pigmentary abnormality is $82/561 = 14.6\%$. The model fitted by LASSO-Patternsearch is:

$$f(x) = -2.1793 + 0.6238 * sys + 0.9640 * bmi * hormone * chol. \tag{8}$$

Two terms showed up in the final model: the main effect of *sys* and a size-three pattern (*bmi hormone chol*). Remembering that the risk direction of *chol* is coded as *lower* serum total cholesterol and the risk direction of *hormone* is coded as *no* current usage of hormone this result agrees with [21] and [9], although less information is being used here since the data has been dichotomized. *hist* doesn't show up partly because there are very few (30) women with the history of heavy drinking in this age group.

code	variable	unit	higher risk
bmi	body mass index	kg/cm	>30
nsummer	time outside in summer	day	>3/4
pky	packyear	pack per day×year smoked	> 30
hormone	current usage of hormone	yes/no	no
hist	history of heavy drinking	yes/no	yes
chol	serum total cholesterol	mg/dL	≤240
sys	systolic blood pressure	mmHg	>140

Table 7: The variables in the pigmentary abnormality example. The forth column shows which direction is risky.

We would like to know which direction is bad for the risk factors before we apply the LASSO-Patternsearch. In the case of pigmentary abnormality, we know higher serum total cholesterol is protective from previous studies. However, such information is often not known prior to the beginning of a study. As shown in simulation example 1, LASSO-Patternsearch is able to tell if the sign of one of the variables is wrong. If we assume higher serum total cholesterol is bad, we will get the following model:

$$f(x) = -2.1602 + 0.6359 * sys + 0.939 * bmi * hormone - 1.0484 * bmi * hormone * chol. \quad (9)$$

The constant and the coefficient of *sys* are very close to those in (8). The remaining two patterns have coefficients whose absolute values are close to that of the pattern (*bmi hormone chol*) in (8). This implies that we are very likely to have the sign of *chol* wrong because we will get (8) back if we do reverse the coding of *chol*, giving a simpler model without negative coefficients.

5.4 Scrambling the Pigmentary Abnormality Data

To examine the significance of the two terms in (8), we follow the method in the progression of myopia example. We permuted the responses completely and ran the LASSO-Patternsearch on the scrambled data. The whole procedure was repeated 600 times and the result is shown in Figure 5. These two terms are not as strong as *catct* and (*pky vtm*) in the previous example as some of the noise patterns have smaller p-values. However, only 8 noise patterns

are stronger than *sys* out of 600 runs. Obviously, *sys* is a very strong term and should be kept in the model. The same argument applies for (*bmi hormone chol*).

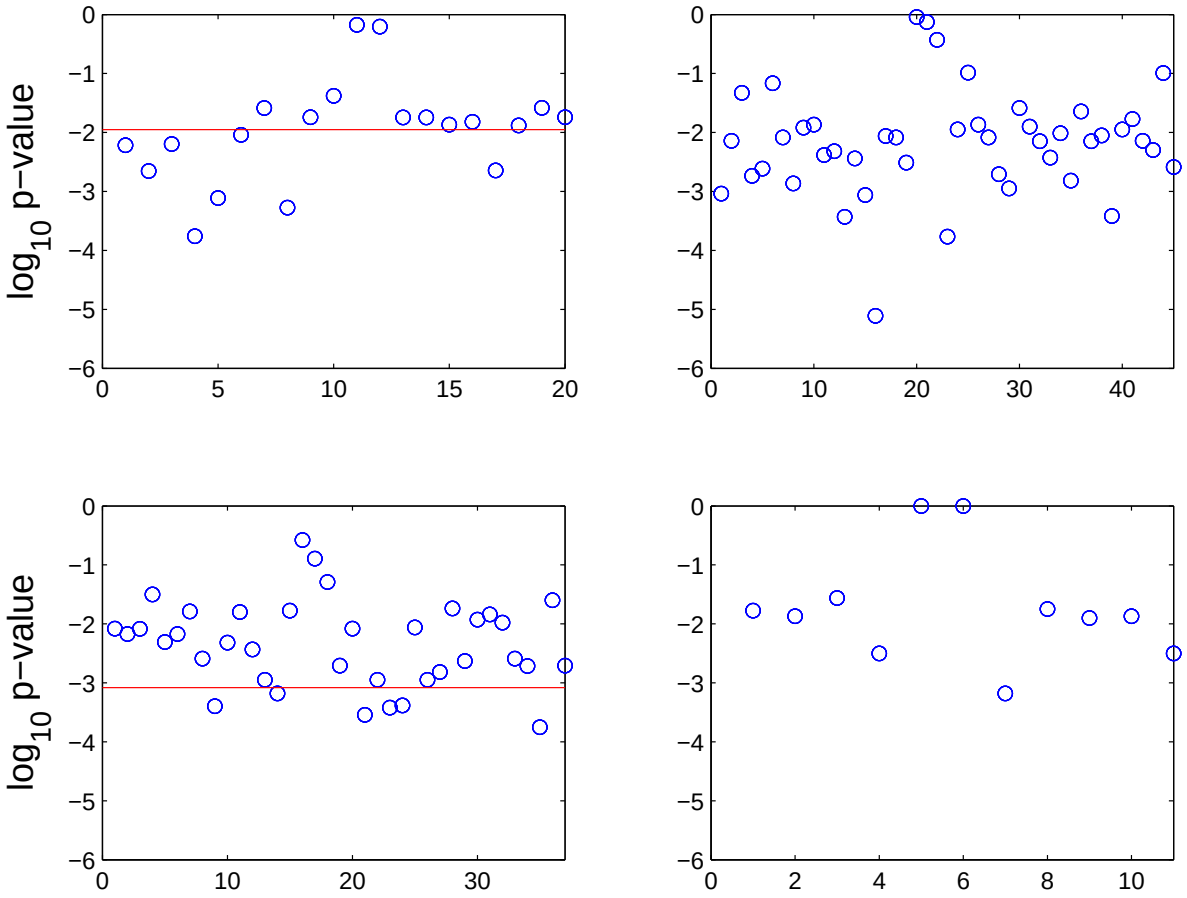

Figure 5: The scatter plots of log_{10}p-values of noise patterns surviving the LASSO-Patternsearch in the completely permuted data, 600 runs. Left to right, top to bottom: patterns of size 1, 2, 3 and 4.

6 Discussion

In any problem where there are a large number of highly interacting predictor variables that are or can be reduced to dichotomous values, the LASSO-Patternsearch can be profitably used. If the "risky" direction (with respect to the outcome of interest) is known from previous studies for all or almost all of the variables, the results are readily interpretable. If the risky

direction is coded correctly for all of the variables, the fitted model can be expected to be sparser than that for any other coding. The transparent two step procedure can, if desired, accommodate subjective prior information by modifying the estimated tuning parameters at either step. The method can be used as a preprocessor when there are very many continuous variables in contention, to reduce the number of variables for more detailed nonparametric analysis.

Many generalizations are available. Two classes of models where the LASSO-Patternsearch approach can be expected to be useful are the multicategory end points model in [20][28], where an estimate is desired of the probability of being in class k when there are K possible outcomes; another is the multiple correlated endpoints model in [9]. In this latter model, the correlation structure of the multiple endpoints can be of interest. Another generalization allows the coefficients c_ℓ to depend on other variables, however, the penalty functional must involve ℓ_1 penalties if it is desired to have a convex optimization problem with good sparsity properties with respect to the patterns. The LASSO-Patternsearch approach may be used when some of the predictor variables are trichotomous (or higher), see [30] for a simple example, but of course this will generate increasingly more basis functions. In work in preparation the LASSO-Patternsearch has been applied to SNP's in DNA sequences to separate cases and controls, with data vectors with thousands of SNP's which may take on three levels of interest. There a preprocessing step is carried out to reduce the number of SNP's in the model, by examining many main effects and two-factor interaction models. The algorithm of Appendix C was able to handle 4000 basis functions involving 12 variables and 4000 samples, taking about 2.5 hours on a 3.4 GHz machine. Problems the size of the BDES data were solved in a few minutes.

7 Summary and Conclusions

The LASSO-Patternsearch combines together several known ideas in a novel way, using a new purpose built computational algorithm. We have examined its properties by analysis of observational data and simulation studies at a scale similar to the observational data. The results are verified in the simulation studies by examination of the generated "truth", and in the observational data, by examining the false pattern generation rate by scrambling the data, by selective examination of the observational data directly, and by comparison with previous analyses of the data, with excellent results. The novel computational algorithm

allows the examination of a very large number of patterns, and, hence, high order interactions. We believe the LASSO-Patternsearch will be an important addition to the toolkit of statistical data analysts.

Acknowledgments

Thanks to David Callan for Appendix D. Research supported by NIH Grants EY09946 (WS, GW) NSF Grants DMS-0505636, DMS-0604572 (WS, GW) ONR Grant N0014-06-0095 (WS,GW) NIH Grants EY06594 and EY015286 (BEK, RK, KL) Research to Prevent Blindness Senior Investigator Awards (RK, BEK) NSF Grants SCI-0330538, DMS-0427689, CCF-0430504, CTS-0456694, CNS-0540147, DOE Grant DE-FG02-04ER25627 (SW)

Appendices

A The Multivariate Bernoulli Distribution and Pattern Basis Functions

Let $(x_0, x_1, \cdots, x_p)$ be a $p+1$ dimensional vector of possibly correlated Bernoulli random variables. The most general form of the joint density is

$$p(x_0, x_1, \cdots, x_p) = p(0,0,\cdots,0)^{\pi_{j=0}^p (1-x_j)} p(1,0,\cdots,0)^{x_0 \pi_{j=1}^p (1-x_j)} \cdots p(1,1,\cdots,1)^{\pi_{j=0}^p x_j}. \tag{10}$$

Taking logs gives

$$\begin{aligned} \log p(x_0, x_1, \cdots x_p) &= \pi_{j=0}^p (1-x_j) \log p(0,0,\cdots,0) \\ &+ x_0 \pi_{j=1}^p (1-x_j) \log p(1,0,\cdots,0) \\ &+ \cdots \\ &+ \pi_{j=0}^p x_j \log p(1,1,\cdots,1) \end{aligned}$$

and, following [30] collecting terms gives

$$\log p(x_0, x_1, \cdots x_p) = \phi + \sum_{j=0}^{p} \mu_j x_j + \sum_{0 \le j < k} \mu_{jk} x_j x_k$$

$$
\begin{aligned}
&+ \sum_{0\le j<k<\ell} \mu_{jk\ell} x_j x_k x_\ell \\
&+ \cdots \\
&+ \mu_{0,1,\cdots,p} x_0 x_1 x_2 \cdots x_p.
\end{aligned}
$$

The μ terms involve various cross product ratios that can be interpreted in various contexts, and the non-zero μ's define a graphical model, see [30]. It is now easy to see what the log odds ratio (logit) $f(x_1, \cdots, x_p)$ for variable x_0 given $(x_1, \cdots, x_p)$ is:

$$
f(x_1, \cdots x_p) = \log p(1, x_1, \cdots, x_p) - \log p(0, , x_1, \cdots, x_p), \tag{11}
$$

and is easily seen to be

$$
f(x_1, \cdots, x_p) = \mu_0 + \sum_{j=1} \mu_{0j} x_j + \sum_{1<j<k} \mu_{0jk} x_j x_k + \cdots + \mu_{0,1,\cdots,p} x_1 \cdots x_p. \tag{12}
$$

which is exactly of the form of Eqn (3) with $N_B = 2^p - 1$.

B The GACV Score

We denote the estimated logit function by $f_\lambda(\cdot)$ and define $f_{\lambda i} = f_\lambda(x(i))$ and $p_{\lambda i} = \frac{e^{f_{\lambda i}}}{1+e^{f_{\lambda i}}}$ for $i = 1, \cdots, n$. Now define

$$
OBS(\lambda) = \frac{1}{n} \sum_{i=1}^{n} [-y_i f_{\lambda i} + \log(1 + e^{f_{\lambda i}})]. \tag{13}
$$

From (2.1) in [31], the leave-one-out CV is

$$
\begin{aligned}
CV(\lambda) &= \frac{1}{n} \sum_{i=1}^{n} [-y_i f_{\lambda i}^{[-i]} + \log(1 + e^{f_{\lambda i}})] \\
&= OBS(\lambda) + \frac{1}{n} \sum_{i=1}^{n} y_i (f_{\lambda i} - f_{\lambda i}^{[-i]}) \\
&= OBS(\lambda) + \frac{1}{n} \sum_{i=1}^{n} y_i \frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - p_{\lambda i}^{[-i]}} \times \frac{y_i - p_{\lambda i}}{1 - (p_{\lambda i} - p_{\lambda i}^{[-i]})/(y_i - p_{\lambda i}^{[-i]})}.
\end{aligned} \tag{14}
$$

Following [31], [33] approximate $(p_{\lambda i} - p_{\lambda i}^{[-i]})/(y_i - p_{\lambda i}^{[-i]})$ by Taylor expansion.

$$
\begin{aligned}
\frac{p_{\lambda i} - p_{\lambda i}^{[-i]}}{y_i - p_{\lambda i}^{[-i]}} &\approx \dot{p}(f_{\lambda i}) \frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - p_{\lambda i}^{[-i]}} \\
&= \frac{e^{f_{\lambda i}}}{(1 + e^{f_{\lambda i}})^2} \frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - p_{\lambda i}^{[-i]}}.
\end{aligned} \tag{15}
$$

Denote

$$G_i \equiv \frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - p_{\lambda i}^{[-i]}}; \tag{16}$$

then from (14) and (15), we get

$$CV(\lambda) \approx OBS(\lambda) + \frac{1}{n}\sum_{i=1}^{n} G_i \frac{y_i(y_i - p_{\lambda i})}{1 - \frac{e^{f_{\lambda i}}}{(1+e^{f_{\lambda i}})^2} G_i}. \tag{17}$$

Denote the objective function in (1) - (4) by $I_\lambda(y, c)$, let $B_{ij} = B_j(x(i))$ be the entries of the design matrix B, and, for ease of notation denote $\mu = c_{N_B}$. Then the objective function can be written

$$I_\lambda(y, c) = \frac{1}{n}\sum_{i=1}^{n}[-y_i \sum_{j=1}^{N_B} c_j B_{ij} + log(1 + e^{(\sum_{j=1}^{N_B} c_j B_{ij})})] + \lambda \sum_{j=1}^{N_B - 1} |c_j|. \tag{18}$$

Denote the minimizer of (18) by $\hat{c}$. We know that the l_1 penalty produces sparse solutions. Without loss of generality, we assume that the first s components of $\hat{c}$ are nonzero. When there is a small perturbation ϵ on the response, we denote the minimizer of $I_\lambda(c, y + \epsilon)$ by $\tilde{c}$. The 0's in the solutions are robust against a small perturbation in the response. That is, when ϵ is small enough, the 0 elements will stay at 0. This can be seen by looking at the the KKT conditions when minimizing (18). Therefore, the first s components of $\tilde{c}$ are nonzero and the rest are zero. For simplicity, we denote the first s components of c by c^* and the first s columns of the design matrix B by B^*. Then let f_λ^y be the column vector with i entry $f_\lambda(x(i))$ based on data y, and let $f_\lambda^{y+\epsilon}$ be the same column vector based on data $y + \epsilon$.

$$f_\lambda^{y+\epsilon} - f_\lambda^y = B(\tilde{c} - \hat{c}) = B^*(\tilde{c}^* - \hat{c}^*). \tag{19}$$

Now we take the first-order Taylor expansion of $\frac{\partial I_\lambda}{\partial c^*}$:

$$\left[\frac{\partial I_\lambda}{\partial c^*}\right]_{(\tilde{c}, y+\epsilon)} \approx \left[\frac{\partial I_\lambda}{\partial c^*}\right]_{(\hat{c}, y)} + \left[\frac{\partial^2 I_\lambda}{\partial c^* \partial c^{*\prime}}\right]_{(\hat{c}, y)} (\tilde{c}^* - \hat{c}^*) + \left[\frac{\partial^2 I_\lambda}{\partial c^* \partial y^\prime}\right]_{(\hat{c}, y)} (y + \epsilon - y). \tag{20}$$

Define

$$U \equiv n\left[\frac{\partial^2 I_\lambda}{\partial c^* \partial c^{*\prime}}\right]_{(\hat{c}, y)} = B^{*\prime} \mathrm{diag}([\frac{e^{f_{\lambda 1}}}{(1+e^{f_{\lambda 1}})^2}, \cdots, \frac{e^{f_{\lambda n}}}{(1+e^{f_{\lambda n}})^2}]) B^* = B*^\prime W B*, \quad \text{say,}$$

and

$$V \equiv -n\left[\frac{\partial^2 I_\lambda}{\partial c^* \partial y^\prime}\right]_{(\hat{c}, y)} = B^{*\prime}.$$

By the first-order conditions, the left-hand side and the first term of the right-hand side of (20) are zero. So we have

$$U(\tilde{c}^* - \hat{c}^*) \approx V\epsilon. \tag{21}$$

Combine (19) and (21) we have $f_\lambda^{y+\epsilon} - f_\lambda^y \approx H\epsilon$, where

$$H \equiv B^* U^{-1} V \equiv B^* U^{-1} B^{*\prime}. \tag{22}$$

Now consider the following perturbation $\epsilon_0 = (0, \cdots, p_{\lambda i}^{[-i]} - y_i, \cdots, 0)'$; then $f_\lambda^{y+\epsilon_0} - f_\lambda^y \approx \epsilon_{0i} H_{.i}$, where $\epsilon_{0i} = p_{\lambda i}^{[-i]} - y_i$ and $H_{.i}$ is the ith row of H. By the Leave-Out-One Lemma (stated below), $f_\lambda{}^{[-i]} = f_\lambda{}^{y+\epsilon_0}$. Therefore

$$G_i = \frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - p_{\lambda i}^{[-i]}} = \frac{f_{\lambda i}^{y+\epsilon_0} - f_{\lambda i}}{\epsilon_{0i}} \approx h_{ii} \tag{23}$$

where h_{ii} is the iith entry of H. From (17), the ACV score is

$$ACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[-y_i f_{\lambda i} + \log(1 + e^{f_{\lambda i}})] + \frac{1}{n}\sum_{i=1}^{n} h_{ii}\frac{y_i(y_i - p_{\lambda i})}{(1 - w_{ii}h_{ii})} \tag{24}$$

where we have used $w_{ii} = \frac{e^{f_{\lambda i}}}{(1+e^{f_{\lambda i}})^2}$, (the estimated variance at $x(i)$). The $GACV$ score is obtained from the ACV score in (24) by replacing h_{ii} by $\frac{1}{n}tr(H)$ and $w_{ii}h_{ii}$ by $\frac{1}{n}tr(WH)$. It is not hard to see that $tr(WH) = trW^{1/2}HW^{1/2} = s \equiv N_{B_0}$, the number of basis functions in the model, giving

$$GACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[-y_i f_{\lambda i} + \log(1 + e^{f_{\lambda i}})] + \frac{1}{n}(\frac{1}{n}trH)\frac{\sum_{i=1}^{n} y_i(y_i - p_{\lambda i})}{(1 - \frac{1}{n}N_{B_0})}. \tag{25}$$

Equation (25) has the same form as Equation (12) in [33], except that the ingredients of H are much simpler in this problem.

Lemma B.1 *(Leave-Out-One Lemma)*
Let the objective function $I_\lambda(y, f)$ be defined as before. Let $f_\lambda^{[-i]}$ be the minimizer of $I_\lambda(y, f)$ with the ith observation omitted and let $p_\lambda^{[-i]}$ be the corresponding probability. For any real number ν, we define the vector $z = (y_1, \cdots, y_{i-1}, \nu, y_{i+1}, \cdots, y_n)'$. Let $h_\lambda(i, \nu, \cdot)$ be the minimizer of $I_\lambda(z, f)$; then $h_\lambda(i, p_\lambda^{[-i]}, \cdot) = f_\lambda^{[-i]}(\cdot)$.

The proof of lemma $B.1$ is quite simple and very similar to the proof of the Leave-Out-One-Lemma in [33] so we will omit it here.

C Minimizing the Penalized Log Likelihood Function

The function (1) is not differentiable with respect to the coefficients $\{c_\ell\}$ in the expansion (4), so most software for large-scale continuous optimization cannot be used to minimize it directly. We reformulate this problem as a smooth nonlinear minimization problem with bound constraints by "splitting" the variables c_ℓ, $\ell = 1, 2, \ldots, N_B - 1$ into positive and negative parts, that is, by introducing nonnegative variables γ_ℓ and η_ℓ such that

$$c_\ell = \gamma_\ell - \eta_\ell, \quad \gamma_\ell \geq 0, \quad \eta_\ell \geq 0, \quad \ell = 1, 2, \ldots, N_B - 1. \tag{26}$$

With this change of variables, we can make the substitution $|c_\ell| = \gamma_\ell + \eta_\ell$ in (4). (There is no need to split the variable $c_{N_B} = \mu$, since its absolute value does not appear in (4).) Denoting $B_{i\ell} = B_\ell(x(i))$, and replacing the arguments (y, f) of I_λ by the coefficients (γ, η, μ), we have

$$\begin{aligned} I_\lambda(\gamma, \eta, \mu) &= \frac{1}{n} \sum_{i=1}^{n} \left[-y_i \left(\mu + \sum_{\ell=1}^{N_B-1} (\gamma_\ell - \eta_\ell) B_{i\ell} \right) + \log \left(1 + \exp \left(\mu + \sum_{\ell=1}^{N_B-1} (\gamma_\ell - \eta_\ell) B_{i\ell} \right) \right) \right] \\ &\quad + \lambda \sum_{\ell=1}^{N_B-1} (\gamma_\ell + \eta_\ell) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[-y_i \left(\mu + \sum_{\ell=1}^{N_B-1} (\gamma_\ell - \eta_\ell) B_{i\ell} \right) + \log \left(1 + F_i(B; \gamma - \eta, \mu) \right) \right] \\ &\quad + \lambda \sum_{\ell=1}^{N_B-1} (\gamma_\ell + \eta_\ell), \end{aligned}$$

where we have defined

$$F_i(B; c, \mu) := \exp \left(\mu + \sum_{\ell=1}^{N_B-1} c_\ell B_{i\ell} \right)$$

Our problem of minimizing (1) is therefore equivalent to the following bound-constrained optimization problem:

$$\min_{\gamma,\eta,\mu} I_\lambda(\gamma, \eta, \mu) \ \text{ s.t. } \ \gamma \geq 0, \ \eta \geq 0. \tag{27}$$

It is easy to show that one of γ_ℓ and η_ℓ must be zero for each $\ell = 1, 2, \ldots, N_B - 1$. First-order optimality conditions are

$$0 = \frac{\partial I_\lambda}{\partial \mu}, \quad 0 \leq \frac{\partial I_\lambda}{\partial \gamma_\ell} \ \perp \ \gamma_\ell \geq 0, \quad 0 \leq \frac{\partial I_\lambda}{\partial \eta_\ell} \ \perp \ \eta_\ell \geq 0, \quad \ell = 1 \cdots, N_B - 1, \tag{28}$$

where the symbol $\perp$ denotes complementarity (that is, $a \perp b$ for two scalars a and b if $ab = 0$.) It is easy to verify that

$$\frac{\partial I_\lambda}{\partial \gamma_\ell} = \frac{1}{n} \sum_{i=1}^{n} \left[-y_i + \frac{F_i(B; \gamma - \eta)}{1 + F_i(B; \gamma - \eta)} \right] B_{i\ell} + \lambda,$$

$$\frac{\partial I_\lambda}{\partial \eta_\ell} = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \frac{F_i(B;\gamma-\eta)}{1+F_i(B;\gamma-\eta)}\right]B_{i\ell} + \lambda,$$

so that

$$\frac{\partial I_\lambda}{\partial \gamma_\ell} + \frac{\partial I_\lambda}{\partial \eta_\ell} = 2\lambda > 0, \quad \ell = 1, \cdots, N_B - 1,$$

for all $\lambda > 0$. Hence, at least one of $\partial I_\lambda/\partial\gamma_\ell$ and $\partial I_\lambda/\partial\eta_\ell$ is positive at the solution, so from (28), at least one of γ_ℓ and η_ℓ is zero at optimality. It is easy to see that I_λ is a convex function, so that any variable (μ, γ, η) satisfying (28) is a solution of (27). We recover the solution $f(x)$ by substituting the solution of (27) into (26) and then into (3).

Generally, only a small fraction of the variables γ_ℓ, η_ℓ, $\ell = 1, 2, \ldots, N_B - 1$ are nonzero at the solution, and this fraction tends to decrease as the parameter λ increases.

To solve (27), we use the gradient projection approach, which has proved effective for large-scale minimization problems with bound constraints. In its most basic form, this approach searches along the negative gradient direction, projecting onto the feasible region at each step and choosing the step length to ensure that the objective function is decreased at each iteration. (When bound constraints are not present, this approach reduces to the familiar steepest-descent approach for unconstrained minimization.) Since second derivatives for the function I_λ are relatively inexpensive to compute, we use an enhanced version of this approach known as *two-metric* gradient projection (See Bertsekas [2, 1]). In this variant, we compute at each iteration k an estimate of the *active set* $\mathcal{A}_k$, consisting of the set of unknowns that are apparently at their bound at the solution. (The active set excludes all variables that are not bounded, such as μ in the problem above.) Second partial derivatives are computed with respect to all variables *not* contained in $\mathcal{A}_k$, and a Newton-like step is computed in the subspace of these variables. (When no bounds are present, the two-metric approach reduces to Newton's method for unconstrained minimization.) We use a backtracking line search in conjunction with projection; if a particular steplength α_k is found not to decrease the function I_λ, we multiply α_k by a constant in the range $(0, 1)$ (typically $1/2$) and try again.

Often, the sequence of active-set estimates $\mathcal{A}_k$ converges to the true active set $\mathcal{A}^*$ in relatively few iterations, and convergence in the subspace of inactive variables is rapid because of the use of second-derivative information in this subspace.

We give further details of the algorithm in terms of the following general bound-constrained problem:

$$\min T(z) \text{ s.t. } z \geq 0, \tag{29}$$

where T is a smooth convex function of m variables. (The method is easily adapted to the problem (27) above in which not all variables have a lower bound.) Given any constant $\epsilon > 0$ and a feasible z, we can compute

$$\rho_\epsilon(z) = \epsilon^{-1} \left\| z - (z - \epsilon \nabla T(z))_+ \right\|_\infty$$

where for any vector y, we use y_+ to denote the vector whose components are $\max(y_i, 0)$, $i = 1, 2, \ldots, m$. Note that z^* satisfies the first-order conditions for optimality if and only if $\rho_\epsilon(z^*) = 0$. By convexity of T, any such z^* is a minimizer.

Given iterate z^k with $z^k \geq 0$, we estimate the active set $\mathcal{A}_k$ as follows:

$$\mathcal{A}_k = \{i = 1, 2, \ldots, m \mid z_i^k \leq \min(\eta, \tau \rho_\epsilon(z^k)) \text{ and } \partial T(z^k)/\partial z_i > 0\}, \tag{30}$$

where η and τ are positive constants, with η small. Essentially the indices in $\mathcal{A}_k$ are those at or near zero for which the partial derivative indicates that the optimum may actually be at zero. We define the "inactive set" $\mathcal{I}_k$ to be the complement of the active set, that is,

$$\mathcal{I}_k = \{1, 2, \ldots, m\} \setminus \mathcal{A}_k.$$

We compute the rows and columns of the Hessian corresponding to $\mathcal{I}_k$, denoting the partial Hessian by $\nabla^2_{\mathcal{I}_k \mathcal{I}_k} T(z^k)$, and compute a scaling factor σ_k to be the average of the diagonal elements of this matrix. The component of the search direction $p^k_{\mathcal{I}_k}$ in the complement of the active set is computed by solving the following system:

$$(\nabla^2_{\mathcal{I}_k \mathcal{I}_k} T(z^k) + \delta_k I) p^k_{\mathcal{I}_k} = -\nabla_{\mathcal{I}_k} T(z^k),$$

where $\delta_k = \min(\sigma_k, \bar{\tau} \rho_\epsilon(z^k))$ (for some constant $\bar{\tau}$) is a damping factor that goes to zero as z^k approaches the solution. If δ_k were set to zero, $p_{\mathcal{I}_k}$ would be the Newton step for the subspace defined by $\mathcal{I}_k$; the use of a damping parameter ensures that the step is well defined even when the partial Hessian is singular or nearly singular, as happens with our problems.

The search direction in the components $\mathcal{A}_k$ is simply a multiple of the negative gradient direction, that is, $p^k_{\mathcal{A}_k} = -\sigma_k^{-1} \nabla_{\mathcal{A}_k} T(z^k)$.

We can now outline the complete algorithm for (29).

Algorithm C.1

given *initial point* z^0*, positive constants* ϵ*,* η*,* τ*,* $\bar{\tau}$*, and* `tol`*;*

for $k = 0, 1, 2, \cdots$

Evaluate $\nabla T(z^k)$ *and compute* $\rho_\epsilon(z^k)$*;*

if $\rho_\epsilon(z^k) <$ `tol`

stop *with approximate solution* z^k*;*

Evaluate $\mathcal{A}_k$ *and* $\mathcal{I}_k$ *and partial Hessian* $\nabla^2_{\mathcal{I}_k\mathcal{I}_k} T(z^k)$*;*

Compute search direction p^k *as described above;*

Set initial guess of steplength α *to* 1*;*

while $T((z^k + \alpha p^k)_+) > T(z^k)$

$\alpha \leftarrow \alpha/2$*;*

Set $z^{k+1} \leftarrow (z^k + \alpha p^k)_+$*;*

end(for)

In our implementation, we used the values $\epsilon = .01$, $\eta = .05$, $\tau = 10$, $\bar{\tau} = 10$.

We usually want to solve the problem for a decreasing sequence of values for the parameter λ. In this case, the computed solution for one value of λ can be used as a starting point for the next problem in the sequence. The number of iterations of the algorithm above is usually much fewer for each value of λ than if it were started "cold."

D Effect of Coding Flips

Proposition: Let $f(x) = \mu + \sum c_{j_1 j_2 .. j_r} B_{j_1 j_2 .. j_r}(x)$ with all $c_{j_1 j_2 .. j_r}$ which appear in the sum strictly positive. If $x_j \to 1 - x_j$ for $j \in$ some subset of $\{1, 2, ..., p\}$ such that at least one x_j appears in f, then the resulting representation has at least one negative coefficient and at least as many terms as f. This follows from the

Lemma: Let $g_k(x)$ be the function obtained from f by transforming $x_j \to 1 - x_j, 1 \leq j \leq k$. Then the coefficient of $B_{j_1 j_2 .. j_r}(x)$ in $g_k(x)$ is

$$(-1)^{|\{j_1,..,j_r\} \cap \{1,...,k\}|} \sum_{\{j_1,\cdots,j_r\} \subseteq T \subseteq \{j_1,\cdots,j_r\} \cup \{1,\cdots k,\}} c_T$$

where $|\cdot|$ means number of entries.

References

[1] D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods.* Academic Press, New York, 1982.

[2] D. P. Bertsekas. Projected Newton methods for optimization problems with simple constraints. *SIAM Journal on Control*, 20:221–246, 1982.

[3] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[4] K.-Y. Chan and W.-Y. Loh. Lotus: An algorithm for building accurate and comprehensible logistic regression trees. *Journal of Computational and Graphical Statistics*, 13:826–852, 2004.

[5] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20:33–61, 1998.

[6] P. Craven and G. Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31:377–403, 1979.

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32:407–499, 2004.

[8] P. Galan, F. Viteri, and S. Bertrais et al. Serum concentrations of beta-carotene, vitamins C and E, zinc and selenium are influenced by sex, age, diet, smoking status, alcohol consumption and corpulence in a general French adult population. *Eur. J. Clin. Nutr.*, 59:1181–90, 2005.

[9] F. Gao, G. Wahba, R. Klein, and B. Klein. Smoothing spline ANOVA for multivariate Bernoulli observations, with applications to ophthalmology data, with discussion. *J. Amer. Statist. Assoc.*, 96:127–160, 2001.

[10] G.H. Golub, M. Heath, and G. Wahba. Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–224, 1979.

[11] S. Gunn and J. Kandola. Structural modelling with sparse kernels. *Machine Learning*, 48:137–163, 2002.

[12] B. E. K. Klein, R. Klein, S. Jensen, and L. Ritter. Are sex hormones associated with age-related maculopathy in women? The Beaver Dam Eye Study. *Tr. Am. Ophth. Soc.*, 92:289–297, 1994.

[13] R. Klein, B.Klein, and S. Jensen. The relation of cardiovascular disease and its risk factors to the 5-year incidence of age-related maculopathy: the Beaver Dam Eye Study. *Ophthalmology.*, 104:1804–1812, 1997.

[14] R. Klein, B. E. K. Klein, K. Linton, and D. DeMets. The Beaver Dam eye study: Visual acuity. *Ophthalmology*, 98:1310–1315, 1991.

[15] K. Knight and W. Fu. Asymptotics for LASSO-type estimators. *Ann. Statist*, 28:1356–1378, 2000.

[16] K. Lee, B. Klein, R. Klein, and T. Wong. Changes in refraction over 10 years in an adult population: the Beaver Dam Eye Study. *Investigative ophthalmology visual science*, 43:2566–71, 2002.

[17] Y. Lee, Y. Kim, S. Lee, and J.-Y. Koo. Structured multicategory support vector machine with ANOVA decomposition. Technical Report 743rr, Department of Statistics, The Ohio State University, Colombus OH, 2005. to apear, Biometrika.

[18] C. Leng, Y. Lin, and G. Wahba. A note on the LASSO and related procedures in model selection. *Statistica Sinica*, 16:xx, 2006.

[19] K. C. Li. Asymptotic optimality of C_L and generalized cross validation in ridge regression with application to spline smoothing. *Ann. Statist.*, 14:1101–1112, 1986.

[20] X. Lin. Smoothing spline analysis of variance for polychotomous response data. Technical Report 1003, PhD thesis, Department of Statistics, University of Wisconsin, Madison WI, 1998. Available via G. Wahba's website.

[21] X. Lin, G. Wahba, D. Xiang, F. Gao, R. Klein, and B. Klein. Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV. *Ann. Statist.*, 28:1570–1600, 2000.

[22] S. Moss, R. Klein, B. Klein, S. Jensen, and S. Meuer. Alcohol consumption and the 5-year incidence of age-related maculopathy: the Beaver Dam Eye Study. *Ophthalmology*, 105:789–794, 1998.

[23] M. Osborne, B. Presnell, and B. Turlach. On the LASSO and its dual. *J. Comp. Graph. Stat.*, 9:319–337, 2000.

[24] L. Ritter, R. Klein, B. Klein, J. Mares-Perlman, and S. Jensen. Alcohol use and age-related maculopathy in the Beaver Dam Eye Study. *Am J Ophthalmol*, 120:190–196, 1995.

[25] I. Ruczinski, C. Kooperberg, and M. Leblanc. Logic regression. *Journal of Computational and Graphical Statistics*, 12:475–511, 2003.

[26] D. Steinberg and N.S. Cardell. The hybrid cart-logit modelin classification and data mining. *Eighth Annual Advanced Research Techniques Forum*, 1998.

[27] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc, B*, 58:267–288, 1996.

[28] G. Wahba. Soft and hard classification by reproducing kernel Hilbert space methods. *Proceedings of the National Academy of Sciences*, 99:16524–16530, 2002.

[29] G. Wahba and S. Wold. A completely automatic French curve. *Commun. Stat.*, 4:1–17, 1975.

[30] J. Whittaker. *Graphical Models in Applied Mathematical Multivariate Statistics*. Wiley, 1990.

[31] D. Xiang and G. Wahba. A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statistica Sinica*, 6:675–692, 1996.

[32] H. Zhang and Y. Lin. Component selection and smoothing for nonparametric regression in exponential families. *Statistica Sinica*, 16:1021–1042, 2006.

[33] H. Zhang, G. Wahba, Y. Lin, M. Voelker, M. Ferris, R. Klein, and B. Klein. Variable selection and model building via likelihood basis pursuit. *J. Amer. Statist. Assoc.*, 99:659–672, 2004.