DEPARTMENT OF STATISTICS

North Carolina State University

2501 Founders Drive, Campus Box 8203

Raleigh, NC 27695-8203

Institute of Statistics Mimeo Series No. 2594

Variable Selection via Penalized Likelihood with Adaptive Penalty

Wenbin Lu and Hao Zhang

Department of Statistics, North Carolina State University, Raleigh, NC

Supported in part by National Science Foundation grants DMS-0504269 and DMS-0405913.

Variable Selection via Penalized Likelihood with Adaptive Penalty

Wenbin Lu and Hao Helen Zhang

North Carolina State University

SUMMARY. We study the variable selection problem in generalized linear models, and propose a penalized likelihood approach for joint variable selection and parameter estimation. Unlike many shrinkage methods such as the LASSO, the new method advocates different penalties for different coefficients: unimportant covariates receive larger penalties than important ones. Consequently, the coefficients of important variables can be efficiently estimated while those of unimportant ones are more shrunk to zeros. We show that the proposed penalized likelihood estimator has desired theoretical properties including root-n consistency and oracle properties. Another advantage of the new procedure lies in its easy implementation; one efficient algorithm is provided in the paper. We illustrate performance of the proposed procedure through numerous simulations, and then apply it to the Wisconsin Epidemiological Study of Diabetic Retinopathy.

KEY WORDS: Adaptive LASSO; Generalized linear models; LASSO; Oracle estimator; Variable selection.

1. Introduction

Generalized linear models (GLMs) include as special cases, linear regression, analysis of variance models, logistic and probit models for categorical responses, and log-linear models. Among all these models, a linear relationship between the response variable Y and the covariates $\mathbf{X} = (X_1, \cdots, X_p)'$ through some known link function is assumed. To be specific, the conditional expectation of Y given $\mathbf{X}$ is specified as

$$\mu = \mathrm{E}(Y|\mathbf{X}) = g(\boldsymbol{\beta}'\mathbf{X}), \tag{1}$$

where $g(\cdot)$ is a known link function and $\boldsymbol{\beta}$ is a p-dimensional vector of regression parameters. If g is the identity function, (1) reduces to the classical linear model. The standard approach to fitting GLMs is via maximum likelihood estimation for exponential families (McCullagh and Nelder, 1989). Given $\mathbf{X}$, assume Y follows a distribution from the exponential family with density

$$f_{Y|\mathbf{X}}(y;\theta,\phi) = \exp\left[\{\theta y - b(\theta)\}/a(\phi) + c(y,\phi)\right], \tag{2}$$

where a, b, c are known functions, $\theta = \boldsymbol{\beta}'\mathbf{X}$ is the canonical parameter, and ϕ is a dispersion parameter that may or may not be known. For Binomial, Poisson, and hypergeoemetric distribution, ϕ is fixed at unity. It is known that $\mu = \dot{b}(\theta)$ and $\mathrm{Var}(Y) = \ddot{b}(\theta)a(\phi)$. In practice, the number of covariates p can be large and some of X's are not relevant or predictive to the response variable Y, i.e. a subset of $\boldsymbol{\beta}$ are zeros. Variable selection is then needed to identify important variables, which often produces more interpretable models with better prediction power. For example, in epidemiological studies for accessing risk factors, a common question to ask is which risk factors should be included in the model. In this paper we consider the variable selection problem for GLMs, and propose a penalized likelihood method for automatic variable selection and efficient parameter estimation.

In statistics, there is a large body of literature on variable selection for classical linear models. Thorough reviews can be found in Hocking (1976), Linhart and Zucchini (1986), Rao and Wu (2001), and Miller (2002). Traditional approaches include forward selection, backward elimination, and best subset selection. Model selection criteria such as Mallow's C_p (Mallows, 1973), Akaike's Information Criterion (Akaike, 1973), and Bayesian Information Criterion (Schwartz, 1978) have been advocated. Recently a family of shrinkage methods are proposed such as the LASSO by Tibshirani (1996) and the SCAD by Fan and Li (2001). By imposing a penalty on the size of regression coefficients, these methods shrink small coefficients to exactly zeros and hence select variables and estimate coefficients simultaneously. Though both the LASSO and SCAD are attractive procedures, each of them has its own drawbacks. In particular, the LASSO solve a convex optimization problem and has shown good empirical performance in various contexts (Tibshirani, 1996). However, its solution does not satisfy oracle properties (Fan and Li, 2001). With regard to the SCAD estimator, it enjoys very nice theoretical properties, but its non-convex penalty form makes the implementation difficult.

In this paper, we propose a penalized likelihood approach by imposing an adaptive L_1 penalty on the regression coefficients in the GLMs. The new method is called the ALASSO, which combines the advantages of aforementioned methods and overcomes their drawbacks. We show that the ALASSO estimator possesses theoretical properties like root-n consistency and oracle properties, and its convex formulation makes the optimization convenient. The remainder of this article is organized as follows. Section 2 introduces the penalized likelihood approach equipped with the adaptive penalty. Section 3 establishes theoretical properties of the ALASSO estimator. In Section 4 we propose an efficient algorithm to compute the ALASSO. We also discuss the issue of parameter tuning and derive the sandwich-type formula for estimating the covariance matrix of the ALASSO estimator. Sections 5 is devoted to simulation studies and the application of the ALASSO to an epidemiological data set. Final remarks are given in Section 6. Major technical derivations and the detailed algorithm are contained in the Appendix.

2. Penalized Likelihood in Generalized Linear Models

Let $(\mathbf{x}_i, y_i), i = 1, \cdots, n$ be n independent and identically distributed samples from the distribution (2). The parameters in the GLM are often estimated by maximizing the log-likelihood function

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\{(\boldsymbol{\beta}'\mathbf{x}_i)y_i - b(\boldsymbol{\beta}'\mathbf{x}_i)\}/a(\phi) + c(y_i, \phi) \right]. \tag{3}$$

Without loss of generality, we assume $\mathbf{x}_i$'s are standardized, i.e. $\sum_{i=1}^{n} x_{ij} = 0, \sum_{i=1}^{n} x_{ij}^2 = 1$ for $j = 1, \cdots, p$. For simplicity, we also assume that the dispersion parameter ϕ is known, which is true for Binomial, Poisson, and Hypergeometric distributions. It is straightforward to generalize the proposed method to the case when ϕ is unknown. To select important variables, several authors have proposed to minimize the negative log-likelihood with certain shrinkage penalty (Fan and Li, 2001; Tibshirani, 1996)

$$-l_n(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} J(|\beta_j|).$$

Here $\lambda \geq 0$ is a tuning parameter which controls the amount of shrinkage: the larger the λ, the greater the amount of shrinkage. Tibshirani (1996) showed the LASSO estimator, using the L_1 penalty $J(|\beta_j|) = |\beta_j|$, can achieve smaller mean squared errors (MSE) than the ordinary least squares estimates. In addition, the L_1 penalty shrinks small coefficients to exactly zeros and produce sparse solutions; hence it does some kind of continuous subset selection.

However, the LASSO applies same penalty to all the coefficients, therefore a larger λ produces sparser solutions at the price of causing larger bias to nonzero coefficients. Ideally, different coefficients should be imposed with different penalties according to their relative importance. This motived our penalized likelihood estimation procedure with the weighted-L_1 penalty:

$$\min_{\boldsymbol{\beta}} -l_n(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|\tau_j, \tag{4}$$

where $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_p)'$ are non-negative leverage factors. We propose to choose τ's adaptively such that large penalties are used for unimportant covariates and small penalties for important ones. The choice of weights $\boldsymbol{\tau}$ in the ALASSO is crucial to assure good solutions. As shown in Section 3, the ALASSO estimator possesses oracle properties when τ_j's are chosen properly.

Denote the maximum likelihood estimator (MLE) of (3) as $\tilde{\boldsymbol{\beta}}$. Because $\tilde{\boldsymbol{\beta}}$ are consistent estimates (Nordberg, 1980), their sizes reflect the relative importance of covariates. In this paper, we choose $\tau_j^{-1} = |\tilde{\beta}_j|$ and solve

$$\min_{\boldsymbol{\beta}} -l_n(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|/|\tilde{\beta}_j|. \tag{5}$$

If $\tilde{\beta}_j = 0$, we set the solution $\hat{\beta}_j = 0$. Due to the consistency of $\tilde{\beta}_j$, the term $|\beta_j|/|\tilde{\beta}_j|$ converges to $I(\beta_j \neq 0)$ in probability as n goes to infinity. As a result, the adaptive penalty in (5) is closely related to the L_0 penalty $\sum_{j=1}^{p} I(|\beta_j| \neq 0)$, also known as the entropy penalty in wavelet literature (Donoho and Johnstone, 1998; Antoniadis and Fan, 2001). And the proposed penalized likelihood approach can be regarded as an automatic procedure to implement the best subset selection in some asymptotic sense. In general, the inverse of any consistent estimate of $\boldsymbol{\beta}$ can be used as weights. When the design matrix X is ill-posed due to collinearity among covariates, the ridge solution based on the L_2 penalty may be more appropriate for weights.

3. Theoretical Properties of ALASSO Estimators

In this section, we establish the asymptotic properties of the ALASSO estimator. Define

$$Q_n(\boldsymbol{\beta}) = l_n(\boldsymbol{\beta}) - n\lambda_n \sum_{j=1}^{p} |\beta_j|/|\tilde{\beta}_j|. \tag{6}$$

Denote the ALASSO solution of (6) by $\hat{\boldsymbol{\beta}}_n$, i.e. $\hat{\boldsymbol{\beta}}_n = \max_{\boldsymbol{\beta}} Q_n(\boldsymbol{\beta})$. Assume the true parameter $\boldsymbol{\beta}_0 = \{(\boldsymbol{\beta}_0^{(1)})', (\boldsymbol{\beta}_0^{(2)})'\}'$, where $\boldsymbol{\beta}_0^{(1)}$ consists of nonzero components and $\boldsymbol{\beta}_0^{(2)}$ consists of zero components. Write $\hat{\boldsymbol{\beta}}_n = (\hat{\beta}_{1n}, ..., \hat{\beta}_{pn}) = \{(\hat{\boldsymbol{\beta}}_n^{(1)})', (\hat{\boldsymbol{\beta}}_n^{(2)})'\}'$ accordingly. Assume the length of the true nonzero components $\boldsymbol{\beta}_0^{(1)}$ of $\boldsymbol{\beta}_0$ is d. Let $I(\boldsymbol{\beta}_0)$ be the Fisher information matrix based on the log likelihood and $I_1(\boldsymbol{\beta}_0^{(1)}) = I_{11}(\boldsymbol{\beta}_0^{(1)}, \mathbf{0})$, where $I_{11}(\boldsymbol{\beta}_0^{(1)}, \mathbf{0})$ is the first $d \times d$ submatrix of $I(\boldsymbol{\beta}_0)$ knowing $\boldsymbol{\beta}_0^{(2)} = \mathbf{0}$. Then under some regularity conditions of the density function $f_{Y|\mathbf{X}}(y; \theta, \phi)$, we have

Theorem 1 (Consistency): If $\sqrt{n}\lambda_n = O_p(1)$, then the ALASSO estimator satisfies $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$.

Theorem 2 (Oracle Property): Assume that $\sqrt{n}\lambda_n \to \lambda_0$ with $0 \leq \lambda_0 < \infty$ and $n\lambda_n \to \infty$, then under the conditions of Theorem 1, with probability tending to 1, the root-n consistent ALASSO estimator $\hat{\boldsymbol{\beta}}_n$ must satisfy:

(i) (Sparsity) $\hat{\boldsymbol{\beta}}_n^{(2)} = \mathbf{0}$;

(ii) (Asymptotic normality) $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{(1)} - \boldsymbol{\beta}_0^{(1)}) \to N\{-\lambda_0 I_1^{-1}(\boldsymbol{\beta}_0^{(1)})\mathbf{b}_1, I_1^{-1}(\boldsymbol{\beta}_0^{(1)})\}$ as n goes to infinity, where $\mathbf{b}_1 = (\text{sign}(\beta_{10})/|\beta_{10}|, \cdots, \text{sign}(\beta_{d0})/|\beta_{d0}|)'$ and $\boldsymbol{\beta}_0^{(1)} = (\beta_{10}, \cdots, \beta_{d0})'$.

A sketch of the proofs of Theorems 1 and 2 is given in Appendix A. From the proofs it is obvious that we only need the root-n consistency of $\tilde{\boldsymbol{\beta}}$. Therefore, any root-$n$ consistent estimates of $\boldsymbol{\beta}_0$ can be used for the adaptive weights $\boldsymbol{\tau}$ without changing the asymptotic properties of the ALASSO solution.

4. Computation Algorithm, Parameter Tuning and Variance Estimation

4.1 Computation Algorithm

For solving the standard LASSO, Tibshirani (1996) suggested an algorithm based on quadratic programming, and Fu (1998) proposed the shooting algorithm. Recently Efron et al. (2004) showed that, the whole solution path of the LASSO can be obtained by a modified Lars algorithm. In this section, we propose an iterative algorithm to solve the ALASSO estimator for fixed λ. We have modified the shooting algorithm (given in Appendix B) to handle the weighted L_1 penalty.

For Gaussian data, the problem (4) reduces to minimizing the penalized least squares problem

$$\sum_{i=1}^{n}(y_i - \boldsymbol{\beta}'\mathbf{x}_i)^2 + \lambda\sum_{j=1}^{p}|\beta_j|/|\tilde{\beta}_j|. \tag{7}$$

We get the ALASSO solution in two steps: firstly compute the ordinary least square estimate $\tilde{\boldsymbol{\beta}}$, secondly use the modified shooting algorithm to solve (7).

For non-normal distributions, we propose to use the Newton-Raphson procedure through the iterative least squares subject to the weighted L_1 penalty. Similar algorithms were used in McCullagh and Nelder (1989) for fitting standard GLMs. Define the gradient $\nabla l(\boldsymbol{\beta}) = -\partial l_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ and the Hessian matrix $\nabla^2 l(\boldsymbol{\beta}) = -\partial^2 l_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\boldsymbol{\beta}'$. Let X_* denote the Cholesky decomposition of $\nabla^2 l(\boldsymbol{\beta})$, i.e. $\nabla^2 l(\boldsymbol{\beta}) = X_*'X_*$, and set the pseudo response vector $\mathbf{y}_* = (X_*')^{-1}(\nabla^2 l(\boldsymbol{\beta})\boldsymbol{\beta} - \nabla l(\boldsymbol{\beta}))$. By the second-order Taylor expansion, $-l_n(\beta)$ can be approximated by the quadratic form $\frac{1}{2}(\mathbf{y}_* - X_*\boldsymbol{\beta})'(\mathbf{y}_* - X_*\boldsymbol{\beta})$, and in each iterative step we minimize

$$\frac{1}{2}(\mathbf{y}_* - X_*\boldsymbol{\beta})'(\mathbf{y}_* - X_*\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}|\beta_j|/|\tilde{\beta}_j|. \tag{8}$$

The problem (8) is different from the LASSO in that different penalties are imposed on β_j's. The following is the complete algorithm to compute the ALASSO estimator for fixed λ:

1. Solve $\tilde{\boldsymbol{\beta}}$ by minimizing the negative log likelihood $-l_n(\boldsymbol{\beta})$.
2. Initialization: $k = 1$ and $\boldsymbol{\beta}^{[k]} = (\beta_1^{[1]}, \cdots, \beta_p^{[1]})' = (0, \cdots, 0)'$.
3. Compute $\nabla l, \nabla^2 l, X_*, \mathbf{y}_*$ based on the current estimate $\boldsymbol{\beta}^{[k]}$.
4. Minimize (8) using the modified shooting algorithm. Denote the solution as $\boldsymbol{\beta}^{[k+1]}$.
5. Let $k = k + 1$. Return to step 3 until the convergence criterion meets.

4.2 Standard Errors

We use the conventional technique in the likelihood setting to approximate the covariance matrix of the ALASSO estimate with sandwich formula. As pointed out by Tibshirani (1996), the solution from L_1 penalty is nonlinear in $\mathbf{y}$ and it is difficult to obtain an accurate estimate of the standard error of $\hat{\boldsymbol{\beta}}$. Following the technique in Fan and Li (2001), we give a closed form estimate for the standard errors of the ALASSO estimate.

At convergence, we may approximate the L_1 penalty as $|\beta_j| = \frac{1}{2}|\beta_j^{[k]}| + \frac{1}{2|\beta_j^{[k]}|}\beta_j^2$, thus the optimization problem becomes

$$-l(\boldsymbol{\beta}^{[k]}) + \nabla l(\boldsymbol{\beta}^{[k]})'(\boldsymbol{\beta} - \boldsymbol{\beta}^{[k]}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{[k]})'\nabla^2 l(\boldsymbol{\beta}^{[k]})(\boldsymbol{\beta} - \boldsymbol{\beta}^{[k]}) + \lambda \sum_{j=1}^{p} \frac{\beta_j^2}{2|\beta_j^{[k]}||\tilde{\beta}_j|}.$$

Then the solution has the form

$$\boldsymbol{\beta}^{[k+1]} = \boldsymbol{\beta}^{[k]} - \left[\nabla^2 l(\boldsymbol{\beta}^{[k]}) + \lambda A(\boldsymbol{\beta}^{[k]})\right]^{-1} \left[\nabla l(\boldsymbol{\beta}^{[k]}) + \lambda b(\boldsymbol{\beta}^{[k]})\right],$$

where $A(\boldsymbol{\beta}) = \text{diag}\{\frac{1}{|\beta_1|\cdot|\tilde{\beta}_1|}, \cdots, \frac{1}{|\beta_p|\cdot|\tilde{\beta}_p|}\}$ and $b(\boldsymbol{\beta}) = \left(\text{sign}(|\beta_1|)/|\tilde{\beta}_1|, \cdots, \text{sign}(|\beta_p|)/|\tilde{\beta}_p|\right)'$. The sandwich formula for the covariance matrix of $\hat{\boldsymbol{\beta}}$ is given as

$$\left[\nabla^2 l(\hat{\boldsymbol{\beta}}) + \lambda A(\hat{\boldsymbol{\beta}})\right]^{-1} \widehat{\text{cov}}\left(\nabla l(\boldsymbol{\beta}_0) + \lambda b(\boldsymbol{\beta}_0)\right) \left[\nabla^2 l(\hat{\boldsymbol{\beta}}) + \lambda A(\hat{\boldsymbol{\beta}})\right]^{-1}.$$

Define $D(\boldsymbol{\beta}) = \text{diag}\{\frac{I(\beta_1 \neq 0)}{\beta_1^2}, \cdots, \frac{I(\beta_p \neq 0)}{\beta_p^2}\}$. Straightforward calculations lead to the estimate

$$\begin{aligned}\widehat{\text{cov}}\left(\nabla l(\boldsymbol{\beta}_0) + \lambda b(\boldsymbol{\beta}_0)\right) &= \left[I + \lambda D(\hat{\boldsymbol{\beta}})\{\nabla^2 l(\hat{\boldsymbol{\beta}})\}^{-1}\right] \nabla^2 l(\hat{\boldsymbol{\beta}}) \left[I + \lambda D(\hat{\boldsymbol{\beta}})\{\nabla^2 l(\hat{\boldsymbol{\beta}})\}^{-1}\right] \\ &= \{\nabla^2 l(\hat{\boldsymbol{\beta}}) + \lambda D(\hat{\boldsymbol{\beta}})\}\{\nabla^2 l(\hat{\boldsymbol{\beta}})\}^{-1}\{\nabla^2 l(\hat{\boldsymbol{\beta}}) + \lambda D(\hat{\boldsymbol{\beta}})\}. \qquad (9)\end{aligned}$$

4.3 Parameter Tuning

The choice of λ in (4) is very important, since λ determines the tradeoff between model fit and model sparsity. In the literature, various criteria have been proposed for model selection; the well-known criteria include Mallow's C_p (Mallows, 1973, 1995), AIC (Akaike, 1973), BIC (Schwarz, 1978), and generalized cross validation (GCV) (Wahba, 1990). Among them, the GCV score was used to choose λ for the LASSO by Tibshirani (1996) and for the SCAD by Fan and Li (2001). In this work, we consider both GCV and BIC for parameter tuning. As we have shown in Section 4.2, the ALASSO solution $\hat{\boldsymbol{\beta}}$ can be regarded as a ridge solution to (8), i.e. $\hat{\boldsymbol{\beta}} = \{\nabla^2 l(\hat{\boldsymbol{\beta}}) + \lambda A(\hat{\boldsymbol{\beta}})\}^{-1} X_*' \mathbf{y}_*$. Therefore the number of effective parameters in the ALASSO estimator can be approximated by $p(\lambda) = \text{tr}[\{\nabla^2 l(\hat{\boldsymbol{\beta}}) + \lambda A(\hat{\boldsymbol{\beta}})\}^{-1}\nabla^2 l(\hat{\boldsymbol{\beta}})]$. The GCV score is given by

$$\text{GCV}(\lambda) = \frac{-l_n(\hat{\boldsymbol{\beta}})}{[n\{1 - p(\lambda)/n\}^2]}. \qquad (10)$$

Alternatively, the BIC criteria is given by

$$\text{BIC}(\lambda) = -2l_n(\hat{\boldsymbol{\beta}}) + \log(n) \cdot r,$$

where r is the number of nonzero coefficients in $\hat{\boldsymbol{\beta}}$. Performance of GCV and BIC are compared through numerical studies in Section 5.

5. Numerical Studies

5.1 Simulation Study

We demonstrate the performance of the proposed penalized likelihood method via simulations, and compare it with the LASSO estimator and the standard maximum likelihood estimator (MLE). To measure the prediction error of any fitted model, we compute the model error (ME)

$$\text{ME}(\hat{\boldsymbol{\beta}}) = E[g(\hat{\boldsymbol{\beta}}'\mathbf{X}) - g(\boldsymbol{\beta}'\mathbf{X})]^2.$$

For the linear model $Y = \boldsymbol{\beta}'\mathbf{X} + \epsilon$, we have $\text{ME}(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'E(\mathbf{X}\mathbf{X}')(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. For the logistic model, we have $\text{ME}(\hat{\boldsymbol{\beta}}) = E\left[\exp(\hat{\boldsymbol{\beta}}'\mathbf{X})/\{1+\exp(\hat{\boldsymbol{\beta}}'\mathbf{X})\} - \exp(\boldsymbol{\beta}'\mathbf{X})/\{1+\exp(\boldsymbol{\beta}'\mathbf{X})\}\right]^2$. The ME is estimated via 1000 Monte Carlo simulations.

Two regression settings are considered: (i) data with Normal error; (ii) binary data. In each setting, 100 data sets are generated and we report the average performance of the fitted models. For overall model fitting, we compute the relative model error (RME) for each procedure, i.e., the ratio of its model error ratio against that of the MLE. For parameter estimation, we report the mean of $\hat{\boldsymbol{\beta}}$ and its standard error over 100 simulations. To evaluate performance in variable selection for each procedure, we report the average selected model size, the number of "correct" zero coefficients which present truly unimportant variables, and the "incorrect" zeros which present important variables erroneously left-out by the procedure. All simulations are done with R codes.

Example 1: Linear Model for Normal Data

We simulate 100 data sets consisting of n observations from the linear model

$$Y = \boldsymbol{\beta}'\mathbf{X} + \epsilon, \quad \epsilon \sim N(0,1), \quad \boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)'.$$

There are eight covariates $\mathbf{X} = (X_1, ..., X_8)'$, which are marginally standard normal and the correlation between X_j and X_k is $\rho^{|j-k|}$ for $j \neq k$ with $\rho = 0.5$. This example was also used in Tibshirani (1996) and Fan and Li (2001). The sample size is chosen as $n = 50$ and $n = 100$. Table 1 summarizes the average model errors and variable selection results for various procedures. For the LASSO and ALASSO, λ is tuned with either GCV or BIC. Here $BIC(\lambda) = (Y - \hat{\boldsymbol{\beta}}_\lambda'\mathbf{X})'(Y - \hat{\boldsymbol{\beta}}_\lambda'\mathbf{X})/\hat{\sigma}^2 + \log(n) \cdot r$, where r is the number of nonzeros in $\hat{\boldsymbol{\beta}}_\lambda$ and the variance

estimator $\hat{\sigma}^2$ is obtained from the least square residuals, i.e. $\hat{\sigma}^2 = (Y - \tilde{\boldsymbol{\beta}}'\mathbf{X})'(Y - \tilde{\boldsymbol{\beta}}'\mathbf{X})/(n-p)$ with $\tilde{\boldsymbol{\beta}}$ the least square estimator.

Overall the ALASSO outperforms the LASSO and MLE in terms of variable selection and model prediction. We also find that BIC works better than GCV for both LASSO and ALASSO. In particular, the ALASSO combined with BIC tuning gives the best performance for both $n = 50$ and $n = 100$. For example, in the case $n = 50$, the ALASSO with BIC selects important covariates most accurately (average model size: true 3, ALASSO+BIC 3.05, LASSO+BIC 4.08, ALASSO+GCV 4.27, LASSO+GCV 5.95), and achieves the smallest model error (RME: ALASSO+BIC 44%, ALASSO+GCV 56%, LASSO+GCV 72%, LASSO+BIC 77%).

(Insert Table 1)

In Table 2, we show the frequency of each variable being selected by the LASSO and ALASSO over 100 simulations. Both procedures never miss important variables in all the settings. Furthermore, the ALASSO tends to choose unimportant variables with a much lower frequency than the LASSO. Again, the ALASSO with BIC gives outstanding performance by avoiding the inclusion of "wrong" variables in the model. For example, when $n = 50$, the ALASSO with BIC selects five noise variables in no more than 2 out of 100 runs.

(Insert Table 2 here)

In Table 3, we demonstrate the point estimation performance of different procedures. To test the accuracy of the proposed standard error formula, we compute both the MC standard error (SD) and the Monte Carlo average of estimated standard error (SE) calculated using the formula (9) in Section 4.2. For the LASSO solutions, we use the formula in Tibshirani (1996) to compute their standard errors. The estimated standard errors of all the procedures are close to the sample standard errors. Not surprisingly, the estimation performance becomes better when the sample size increases.

(Insert Table 3 here)

Example 2: Logistic Model for Binary Data

We simulate 100 data sets consisting of n observations from the logistic model

$$P[Y = 1|\mathbf{X}] = \exp(\boldsymbol{\beta}'\mathbf{X})/\{1 + \exp(\boldsymbol{\beta}'\mathbf{X})\}, \quad \text{with} \quad \boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)'.$$

The eight covariates $\mathbf{X} = (X_1, ..., X_8)'$ are the same as before. This model was also considered in Fan and Li (2001). Two sample sizes are considered: $n = 200$ and 300. Table 4 summarizes the average model errors and variable selection results for three methods, with parameter tuned with GCV and BIC. Table 5 shows the frequency of each variable being selected by LASSO and ALASSO over 100 simulations. Similar to the results in Example 1, the ALASSO performs

better than the LASSO in all settings. The ALASSO with the BIC performs the best among all methods. The BIC proves to be more effective than GCV as a tuning score also in this context. Table 6 shows the point estimation results of various procedures.

(Insert Tables 4-6 here)

5.2 Application to Wisconsin Epidemiological Study of Diabetic Retinopathy

We apply the proposed methodology to the data from the Wisconsin Epidemiological Study of Diabetic Retinopathy (WESDR), an ongoing epidemiological study of a cohort of patients receiving their medical care southern Wisconsin. The baseline examination was conducted in 1980-82, and four, ten, fourteen, and twenty year followups have been carried out. Details about the study can be found in Klein et al. (1984, 1989, 1998). The data set consists of 648 observations. The binary response variable Y is 1 for those patients who experienced four-year progression of retinopathy and 0 otherwise. This data was analyzed by nonparametric smoothing spline method in Wahba et al. (1995) and Zhang et al. (2004). Following their setup, we consider 14 potential risk factors. The continuous covariates are *dur, gly, bmi, sys, ret, pulse, ins, sch, iop*, and categorical covariates are *smk, sex, asp, famdb, mar*. As reported in Wahba et al. (1995) and Zhang et al. (2004), after many laborious parametric and nonparametric regression analyses of small groups of variable at a time, it would be reasonable to assume that three variables: *dur, gly, bmi* are the "true" risk factors. Table 7 summarizes the estimated coefficients by the MLE, ALASSO, and the LASSO methods, with the associated standard errors. Tuning with the BIC leads to a more compact model than with the GCV. When tuned with the BIC, the LASSO selects four variables *gly, bmi, sch, sex* while the ALASSO selects only two of them *gly, bmi*. Based on the nonparametric model fitting result in Zhang et al. (2004), the effect of *dur* has a nonparametric hilly shape which can not approximated well by a straight line; this might explain why *dur* is missed by the procedures based on linear models.

(Insert Table 7 here)

6. Discussion

We propose the penalized likelihood method with an adaptive L_1 penalty for automatic variable selection in generalized linear models. Based on our numerical results, the ALASSO gives better performance than the LASSO in terms of selecting variables more correctly and achieving smaller model errors. Theoretical properties of the ALASSO estimator such as root-n consistency and oracle property are shown. We also find that the BIC chooses the tuning parameter more effectively than GCV in the context. It is possible to extend this work to other models like Cox's proportional hazards models (Zhang and Lu, 2006).

Acknowledgment

The work of the first author was supported in part by NSF grant DMS-0504269, and that of the second author was supported in part by NSF grant DMS-0405913.

References

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* **60**, 255-265.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelets approximations. *Journal of the American Statistical Association* **96**, 939-963.

Donoho, D. and Johnstone, I. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90**, 1200-1224.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least Angle Regression. *Annals of Statistics* **32**, 407-451.

Fan, J. and Li, R. (2001). Variable selection via penalized likelihood. *Journal of American Statistical Association* **99**, 1348-1360.

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

Hocking, R. (1976). The analysis and selection of variables in linear regression. *Biometrics* **32**, 1-49.

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

Mallows, C.L. (1973). Some comments on C_p. *Technometrics* **15**, 661-675.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd edition. Chapman & Hall: New York.

Miller, A. (2002). *Subset Selection in Regression*. CRC Press, Chapman & Hall.

Nordberg, L. (1980). Asymptotic normality of maximum likelihood estimators based on independent, unequally distributed observations in exponential family models. *Scandinavian Journal of Statistics* **7**, 27-32.

Rao, C.R. and Wu, Y. (2001). On model selection (with discussion). In *Institute of Mathematical Statistical Lecture Notes - Monograph Series*, P. Lahiri (ed.) **38**, 1-64.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461-464.

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

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

Wahba, G., Wang, Y., Gu, C., Klein, R. and Klein, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Annals of Statistics* **23**, 1865-1895.

Zhang, H. H. and Lu, W. (2006). Adaptive-LASSO for Cox's proportional hazards model. *Submitted*.

Zhang, H. H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R. and Klein, B. (2004). Variable Selection and Model Building via Likelihood Basis Pursuit. *Journal of American Statistical Association* **99**, 659-672.

Appendix A: *Proofs of Theorems 1 and 2*

The same regularity conditions (A)-(C) used in Fan and Li (2001) are also assumed here.

Proof of Theorem 1

Following the similar steps of Fan and Li (2001), we want to show that

$$P\left\{\sup_{\boldsymbol{\beta}\in\partial B_n(C)} Q_n(\boldsymbol{\beta}) < Q_n(\boldsymbol{\beta}_0)\right\} \geq 1-\epsilon, \tag{A.1}$$

where $\partial B_n(C)$ is the boundary of the C-ball $B_n(C) \equiv \{\boldsymbol{\beta} : \boldsymbol{\beta} = \boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}, ||\mathbf{u}|| \leq C\}$. Recall that p is the dimension of $\boldsymbol{\beta}$, and d is the number of true nonzero components in $\boldsymbol{\beta}_0$. By the second-order Taylor expansion of the log likelihood and simple calculations, we have

$$\begin{aligned}
D_n(\mathbf{u}) &\equiv \frac{1}{n}\{Q_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) - Q_n(\boldsymbol{\beta}_0)\} \\
&= \frac{1}{n}\{l_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) - l_n(\boldsymbol{\beta}_0)\} - \lambda_n \sum_{j=1}^{p}\left\{\frac{|\beta_{j0} + n^{-1/2}u_j|}{|\tilde{\beta}_j|} - \frac{|\beta_{j0}|}{|\tilde{\beta}_j|}\right\} \\
&\leq \frac{1}{n}\{l_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) - l_n(\boldsymbol{\beta}_0)\} - \lambda_n \sum_{j=1}^{d}(|\beta_{j0} + n^{-1/2}u_j| - |\beta_{j0}|)/|\tilde{\beta}_j| \\
&\leq \frac{1}{n}\{l_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) - l_n(\boldsymbol{\beta}_0)\} + n^{-1/2}\lambda_n \sum_{j=1}^{d}|u_j|/|\tilde{\beta}_j| \\
&= -\frac{1}{2n}\mathbf{u}'\{I(\boldsymbol{\beta}_0) + o_p(1)\}\mathbf{u} + \frac{1}{n}O_p(1)\sum_{j=1}^{p}|u_j| + \frac{1}{\sqrt{n}}\lambda_n \sum_{j=1}^{d}|u_j|/|\tilde{\beta}_j|,
\end{aligned} \tag{A.2}$$

where $\mathbf{u} = (u_1, \cdots, u_d)'$. Using the fact that $\tilde{\boldsymbol{\beta}}$ satisfies $||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| = O_p(n^{-1/2})$, we have, for $1 \le j \le d$,

$$\frac{1}{|\tilde{\beta}_j|} = \frac{1}{|\beta_{j0}|} - \frac{\text{sign}(\beta_{j0})}{\beta_{j0}^2}(\tilde{\beta}_j - \beta_{j0}) + o_p(|\tilde{\beta}_j - \beta_{j0}|) = \frac{1}{|\beta_{j0}|} + \frac{O_p(1)}{\sqrt{n}}.$$

In addition, since $\sqrt{n}\lambda_n = O_p(1)$, we have

$$\begin{aligned}\frac{1}{\sqrt{n}}\lambda_n \sum_{j=1}^{d} |u_j|/|\tilde{\beta}_j| &= \frac{1}{\sqrt{n}}\lambda_n \sum_{j=1}^{d} \left\{ \frac{|u_j|}{|\beta_{j0}|} + \frac{|u_j|}{\sqrt{n}} O_p(1) \right\} \\ &\le Cn^{-1/2}\lambda_n O_p(1) = Cn^{-1}(\sqrt{n}\lambda_n)O_p(1) = Cn^{-1}O_p(1).\end{aligned}$$

Therefore in (A.2), by choosing a sufficiently large C, the first term is of the order $C^2 n^{-1}$. The second and third terms are of the order Cn^{-1}, which are dominated by the first term. Therefore (A.1) holds and it completes the proof.

Proof of Theorem 2

To show the sparsity: $\hat{\boldsymbol{\beta}}_n^{(2)} = \mathbf{0}$, it is sufficient to show that for any sequence $\boldsymbol{\beta}_1$ satisfying that $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0^{(1)}\| = O_p(n^{-1/2})$ and any constant C,

$$Q_n(\boldsymbol{\beta}_1, \mathbf{0}) = \max_{\|\boldsymbol{\beta}_2\| \le Cn^{-1/2}} Q_n(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2). \tag{A.3}$$

To prove (A.3), it is sufficient to show that with probability tending to 1, for any $\boldsymbol{\beta}_1$ satisfying that $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0^{(1)}\| = O_p(n^{-1/2})$, $\partial Q(\boldsymbol{\beta})/\partial\beta_j$ and β_j have different signs for $\beta_j \in (-Cn^{-1/2}, Cn^{-1/2})$ with $j = d+1, \cdots, p$. We have

$$\begin{aligned}\frac{\partial Q_n(\boldsymbol{\beta})}{\partial \beta_j} &= \frac{\partial l_n(\boldsymbol{\beta})}{\partial \beta_j} - n\lambda_n \frac{\text{sign}(\beta_j)}{|\tilde{\beta}_j|} = O_p(n^{1/2}) - (n\lambda_n)n^{1/2}\frac{\text{sign}(\beta_j)}{|n^{1/2}\tilde{\beta}_j|} \\ &= n^{1/2}\left\{ O_p(1) - n\lambda_n \frac{\text{sign}(\beta_j)}{|O_p(1)|} \right\},\end{aligned} \tag{A.4}$$

since $n^{1/2}(\tilde{\beta}_j - 0) = O_p(1)$. Then as $n\lambda_n \to \infty$, the sign of $\frac{\partial Q_n(\beta_j)}{\partial \beta_j}$ in (A.4) is completely determined by the sign of β_j when n is large, and they always have different signs.

To show the asymptotic normality of $\hat{\boldsymbol{\beta}}_n^{(1)}$, define $s_n(\boldsymbol{\beta}) = \partial l_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ and $\bigtriangledown s_n(\boldsymbol{\beta}) = \partial s_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}'$. Let $s_n^{(1)}(\boldsymbol{\beta})$ be the first s elements of $s_n(\boldsymbol{\beta})$ and $\hat{I}_n^{(11)}(\boldsymbol{\beta})$ be the first $s \times s$ submatrix of $\bigtriangledown s_n(\boldsymbol{\beta})$. Then

$$\begin{aligned}0 &= \frac{\partial Q_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1}\Big|_{\boldsymbol{\beta}=\{(\hat{\boldsymbol{\beta}}_n^{(1)})', \mathbf{0}'\}'} \\ &= \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1|}\Big|_{\boldsymbol{\beta}=\{(\hat{\boldsymbol{\beta}}_n^{(1)})', \mathbf{0}'\}'} - n\lambda_n \left(\frac{\text{sign}(\hat{\beta}_{1n})}{\tilde{\beta}_1}, \cdots, \frac{\text{sign}(\hat{\beta}_{dn})}{\tilde{\beta}_d} \right)' \\ &= s_n^{(1)}(\boldsymbol{\beta}_0) - \hat{I}_n^{(11)}(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}}_n^{(1)} - \boldsymbol{\beta}_0^{(1)}) - n\lambda_n \left(\frac{\text{sign}(\beta_{10})}{\tilde{\beta}_1}, \cdots, \frac{\text{sign}(\beta_{d0})}{\tilde{\beta}_d} \right)',\end{aligned}$$

where $\boldsymbol{\beta}^*$ is between $\hat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_0$. The last equation is implied by $\text{sign}(\hat{\beta}_{jn}) = \text{sign}(\beta_{j0})$ when n is large, since $\hat{\boldsymbol{\beta}}_n$ is a root-n consistent estimator of $\boldsymbol{\beta}_0$. Since $s_n^{(1)}(\boldsymbol{\beta}_0)/\sqrt{n} \to N\{\mathbf{0}, I_1(\boldsymbol{\beta}_0^{(1)})\}$ in distribution, $\hat{I}_n^{(11)}(\boldsymbol{\beta}^*)/n \to I_1(\boldsymbol{\beta}_0^{(1)})$ in probability, $\sqrt{n}\lambda_n \to \lambda_0$ and $\tilde{\beta}_j \to \beta_{j0} \neq 0$, for $1 \leq j \leq d$, as $n \to \infty$, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{(1)} - \boldsymbol{\beta}_0^{(1)}) = I_1^{-1}(\boldsymbol{\beta}_{10}) \left\{ \frac{1}{\sqrt{n}} s_n^{(1)}(\boldsymbol{\beta}_0) - \lambda_0 \mathbf{b}_1 \right\} + o_p(1).$$

Therefore, by Slutsky's Lemma,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{(1)} - \boldsymbol{\beta}_0^{(1)}) \to N\{-\lambda_0 I_1^{-1}(\boldsymbol{\beta}_0^{(1)})\mathbf{b}_1, I_1^{-1}(\boldsymbol{\beta}_0^{(1)})\}$$

in distribution as $n \to \infty$.

Appendix B: *Modified Shooting Algorithm for ALASSO*

A modified shooting algorithm taking into account the weighted L_1 penalty in the ALASSO is proposed for solving (8). Define $S(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \boldsymbol{\beta}'\mathbf{x}_i)^2$, $\dot{S}_j(\boldsymbol{\beta}) = \partial S(\boldsymbol{\beta})/\partial\beta_j$, $j = 1, \cdots, p$, and denote $\boldsymbol{\beta}$ by $(\beta_j, \boldsymbol{\beta}^{-j})'$, where $\boldsymbol{\beta}^{-j}$ is the $(p-1)$-dimensional vector consisting of the β_i's other than β_j. The modified shooting algorithm is then given as follows

(i) Start with $\hat{\boldsymbol{\beta}}_0 = \tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \cdots, \tilde{\beta}_p)'$ and let $\lambda_j = \lambda/|\tilde{\beta}_j|$ for $j = 1, \cdots, p$.

(ii) At step m, for each $j = 1, \cdots, p$, let $S_0 = \dot{S}_j(0, \hat{\boldsymbol{\beta}}_{m-1}^{-j})$ and set

$$\hat{\beta}_j = \begin{cases} \frac{\lambda_j - S_0}{2(\mathbf{x}^j)'\mathbf{x}^j} & \text{if} \quad S_0 > \lambda_j \\ \frac{-\lambda_j - S_0}{2(\mathbf{x}^j)'\mathbf{x}^j} & \text{if} \quad S_0 < -\lambda_j \\ 0 & \text{if} \quad |S_0| \leq \lambda_j, \end{cases}$$

where $\mathbf{x}^j = (x_{1j}, \cdots, x_{nj})'$. Form a new estimator $\hat{\boldsymbol{\beta}}_m = (\hat{\beta}_1, \cdots, \hat{\beta}_p)'$ after updating all the $\hat{\beta}_j$'s.

(iii) Repeat (ii) until $\hat{\boldsymbol{\beta}}_m$ converges.

Table 1: Model fitting and variable selection results for Normal example.

n	Model	RME (%)	Correct Zero (5)	Incorrect Zero (0)	Model Size (3)
50	MLE	100	0	0	8
	LASSO (BIC)	77	3.92	0	4.08
	ALASSO (BIC)	**44**	**4.95**	**0**	**3.05**
	LASSO (GCV)	72	2.06	0	5.94
	ALASSO (GCV)	56	3.73	0	4.27
100	MLE	100	0	0	8
	LASSO (BIC)	85	3.91	0	4.09
	ALASSO (BIC)	**44**	**4.89**	**0**	**3.11**
	LASSO (GCV)	67	2.1	0	5.9
	ALASSO (GCV)	47	4.07	0	3.93

Table 2: Variable selection frequencies for Normal example

n	Tuning	Method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
50	BIC	LASSO	100	100	27	27	100	23	18	13
		ALASSO	**100**	**100**	**2**	**1**	**100**	**0**	**1**	**1**
	GCV	LASSO	100	100	60	50	100	60	56	63
		ALASSO	100	100	22	26	100	22	27	30
100	BIC	LASSO	100	100	25	28	100	22	13	21
		ALASSO	**100**	**100**	**1**	**1**	**100**	**3**	**2**	**4**
	GCV	LASSO	100	100	53	64	100	50	60	57
		ALASSO	100	100	23	20	100	16	19	15

Table 3: Point estimation for Normal example; true $(\beta_1, \beta_2, \beta_5) = (3, 1.5, 2)$.

n	Method	$\hat{\beta}_1$ (SD, SE)	$\hat{\beta}_2$ (SD, SE)	$\hat{\beta}_5$ (SD, SE)
50	MLE	2.979 (0.200, 0.158)	1.505 (0.219, 0.176)	1.991 (0.212, 0.175)
	LASSO (BIC)	2.903 (0.212,0.159)	1.411 (0.164, 0.166)	1.845 (0.177, 0.144)
	ALASSO (BIC)	2.996 (0.208, 0.161)	1.472 (0.174, 0.166)	1.977 (0.163,0.136)
	LASSO (GCV)	2.932 (0.203, 0.154)	1.442(0.200, 0.159)	1.908 (0.199, 0.145)
	ALASSO (GCV)	2.984 (0.203, 0.158)	1.475 (0.199, 0.164)	1.990 (0.188, 0.145)
100	MLE	2.994 (0.126, 0.115)	1.505 (0.132, 0.126)	2.005 (0.134, 0.123)
	LASSO (BIC)	2.961 (0.119, 0.114)	1.398(0.133, 0.116)	1.883 (0.120, 0.103)
	ALASSO (BIC)	3.026(0.116, 0.115)	1.452 (0.123, 0.115)	1.984 (0.119, 0.098)
	LASSO (GCV)	2.962 (0.124, 0.114)	1.466(0.121, 0.116)	1.944 (0.124, 0.108)
	ALASSO (GCV)	2.998(0.123, 0.115)	1.492 (0.120, 0.117)	1.998 (0.118, 0.104)

Table 4: Model fitting and variable selection results for Binary example.

n	Model	RME (%)	Correct Zero (5)	Incorrect Zero (0)	Model Size (3)
200	MLE	100	0	0	8
	LASSO (BIC)	76	3.15	0	4.85
	ALASSO (BIC)	**45**	**4.58**	**0**	**3.42**
	LASSO (GCV)	79	0.94	0	7.06
	ALASSO (GCV)	70	2.17	0	5.83
300	MLE	100	0	0	8
	LASSO (BIC)	78	3.45	0	4.55
	ALASSO (BIC)	**45**	**4.75**	**0**	**3.25**
	LASSO (GCV)	80	0.99	0	7.01
	ALASSO (GCV)	71	2.55	0	5.45

Table 5: Variable selection frequencies for binary example

n	Tuning	Method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
200	BIC	LASSO	100	100	40	40	100	35	28	42
		ALASSO	**100**	**100**	**6**	**10**	**100**	**12**	**8**	**6**
	GCV	LASSO	100	100	79	86	100	80	77	84
		ALASSO	100	100	54	60	100	56	53	60
300	BIC	LASSO	100	100	25	28	100	22	13	21
		ALASSO	**100**	**100**	**1**	**1**	**100**	**3**	**2**	**4**
	GCV	LASSO	100	100	83	79	100	76	79	84
		ALASSO	100	100	52	48	100	51	54	40

Table 6: Point estimation for binary example; true $(\beta_1, \beta_2, \beta_5) = (3, 1.5, 2)$.

n	Method	$\hat{\beta}_1$ (SD, SE)	$\hat{\beta}_2$ (SD, SE)	$\hat{\beta}_5$ (SD, SE)
200	MLE	3.360 (0.630, 0.606)	1.628 (0.404,0.435)	2.150 (0.347, 0.258)
	LASSO (BIC)	2.413 (0.569,0.356)	1.106 (0.347, 0.258)	1.449 (0.436, 0.268)
	ALASSO (BIC)	2.765 (0.500, 0.477)	1.239 (0.326, 0.333)	1.720 (0.412, 0.350)
	LASSO (GCV)	3.012 (0.589,0.493)	1.427 (0.372, 0.353)	1.888 (0.512,0.380)
	ALASSO (GCV)	3.174 (0.592, 0.563)	1.506 (0.380, 0.397)	2.014 (0.506, 0.438)
300	MLE	3.230 (0.471, 0.466)	1.614 (0.361, 0.342)	2.172 (0.431, 0.388)
	LASSO (BIC)	2.421 (0.430, 0.290)	1.152(0.268, 0.212)	1.527 (0.326, 0.221)
	ALASSO (BIC)	2.849(0.394, 0.399)	1.354 (0.307, 0.279)	1.865 (0.313, 0.297)
	LASSO (GCV)	2.945 (0.458, 0.392)	1.450(0.346, 0.284)	1.943 (0.415, 0.308)
	ALASSO (GCV)	3.120 (0.458, 0.445)	1.541 (0.351, 0.318)	2.086 (0.400, 0.352)

Table 7: Estimated coefficients and standard errors for WESDR data.

		GCV		BIC	
Covariate	MLE	LASSO	ALASSO	LASSO	ALASSO
intercept	-7.429 (1.16)	-7.056 (0.687)	-7.575 (0.948)	-5.713 (0.480)	-5.876 (0.796)
dur	-0.014 (0.014)	-0.008 (0.005)	-0.006 (0.013)	0 (-)	0 (-)
gly	0.537 (0.055)	0.486 (0.046)	0.520 (0.052)	0.421 (0.038)	0.472 (0.050)
bmi	0.068 (0.025)	0.051 (0.017)	0.056 (0.023)	0.032 (0.010)	0.024 (0.023)
sys	-0.003 (0.007)	0 (-)	0 (-)	0 (-)	0 (-)
ret	-0.001 (0.009)	0 (-)	0 (-)	0 (-)	0 (-)
pulse	0.006 (0.014)	0 (-)	0 (-)	0 (-)	0 (-)
ins	-0.224 (0.334)	0 (-)	0 (-)	0 (-)	0 (-)
sch	0.054 (0.030)	0.036 (0.017)	0.037 (0.037)	0.018 (0.007)	0 (-)
iop	-0.030 (0.027)	-0.005 (0.006)	-0.005 (0.026)	0 (-)	0 (-)
smk	0.297 (0.207)	0.177 (0.109)	0.156 (0.197)	0 (-)	0 (-)
sex	0.328 (0.191)	0.196 (0.110)	0.213 (0.181)	0.019 (0.011)	0 (-)
asp	0.266 (0.385)	0.009 (0.013)	0 (-)	0 (-)	0 (-)
famdb	-0.088 (0.181)	0 (-)	0 (-)	0 (-)	0 (-)
mar	-0.111 (0.223)	0 (-)	0 (-)	0 (-)	0 (-)