The Iterated Lasso for High-Dimensional Logistic Regression

By JIAN HUANG

Department of Statistics and Actuarial Science, 241 SH
University of Iowa, Iowa City, Iowa 52242, U.S.A.

SHUANGE MA

Division of Biostatistics, Department of Epidemiology and Public Health,
Yale University, New Haven, Connecticut 06520, U.S.A.

CUN-HUI ZHANG

Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, U.S.A.

The University of Iowa
Department of Statistics and Actuarial Science
Technical Report No. 392
November 5, 2008

Abstract. We consider an iterated Lasso approach for variable selection and estimation in sparse, high-dimensional logistic regression models. In this approach, we use the Lasso (Tibshirani 1996) to obtain an initial estimator and reduce the dimension of the model. We then use the Lasso as the initial estimator in the adaptive Lasso (Zou 2006) to obtain the final selection and estimation results. We provide conditions under which this two-step approach possesses asymptotic oracle selection and estimation properties. One important aspect of our results is that the total number of covariates can be larger than the sample size. Simulation studies indicate that the iterated Lasso has superior performance in variable selection relative to the standard Lasso. A data example is used to illustrate the proposed approach.

Key Words and Phrases. High-dimensional data; Oracle property; Penalized regression; Sparse models; Variable selection.

Short title. Iterated Lasso Logistic Regression

1 Introduction

The logistic regression is widely used in biomedical and epidemiological studies to identify risk factors associated with disease. Traditional studies usually involve a small number of potential risk factors, or covariates. Theoretical properties of the maximum likelihood estimator in the logistic regression models in low-dimensional settings are well established, and application of this model in such settings is facilitated by many widely available computer programs. In recent years, more and more high-dimensional genetic and genomic data are being generated using microarray technologies in studies that attempt to find genetic risk factors for disease and clinical outcomes. With such data, the number of covariates is larger than the sample size. The standard maximum likelihood method for logistic regression is not applicable.

Recently, there has been much work on the penalized methods for high-dimensional models, notably, considerable progress has been made in understanding the statistical properties of the Lasso (Tibshirani 1996) in both small p and large p settings, where p is the number of covariates. When p is fixed, Knight and Fu (2001) showed that, under appropriate conditions, the Lasso is consistent for estimating the regression parameter and its limiting distributions can have positive probability mass at 0 when the true value of the parameter is zero. Meinshausen and Bühlmann (2006) and Zhao and Yu (2006) showed that under a strong irrepresentable condition the Lasso is consistent for variable selection even when the number of variables p is as large as $\exp(n^a)$ for some $0 < a < 1$. Zhang and Huang (2008) studied the behavior of the Lasso regarding its selection properties if the irrepresentable condition is relaxed in linear regression models. They showed that under appropriate conditions on the sparseness of the model and the design matrix, the Lasso estimator is consistent in the ℓ_2 norm, and with high probability, all important variables are selected by the Lasso. In particular, under a sparse Riesz condition on the correlation of design variables, they showed that the Lasso selects a model of the right order of dimensionality, controls the bias of the selected model at a level determined by the contributions of small regression coefficients and threshold bias, and selects all coefficients of greater order than the bias of the selected model. An important aspect of the results of Zhang and Huang (2008) is that the logarithm of the number of variables can be of the same order as the sample size for certain random dependent designs. Related results have also been obtained by Meinshausen and Yu (2008). Van de Geer (2008) studied the Lasso in high-dimensional generalized linear models. However, her work focused on the prediction aspect of the Lasso, but did not address the question of variable selection. All the aforementioned work, except Van de Geer (2008), were carried out in the context of linear regression models.

While the Lasso has excellent properties in dimensional reduction and estimation, it is in general not selection consistent even in the low-dimensional case (Knight and Fu 2001; Leng, Lin and Wahba 2004). The adaptive Lasso proposed by Zou (2006) aims to improve the performance of the Lasso. The adaptive Lasso relies on an initial consistent estimator. When the number of covariates p is fixed, Zou (2006) proved that the adaptive Lasso has the asymptotic oracle property in linear regression. Huang, Ma and Zhang (2006) considered the case when $p \gg n$, where n is the sample size. They showed that the adaptive Lasso has the oracle property under an adaptive irrepresentable and other regularity conditions, provided a initial consistent estimator is available. This result allows $p = O(\exp(n^a))$ for some constant $0 < a < 1$, where a depends on the regularity conditions. Zou and Li (2008) studied a one-step approach in nonconcave penalized likelihood methods in models with fixed p. This approach is closely related to the adaptive Lasso.

We study the properties of a special form of the adaptive Lasso for variable selection and estimation in logistic regression in "large p, small n" settings. This approach was also suggested by Bühlmann and Meier (2008) in the context of linear regression. In this approach, we first use the Lasso to obtain an initial estimator and reduce the dimension of the model. We then use the Lasso estimates to form the weights in the ℓ_1 penalty in the adaptive Lasso in the second step to select the final set of variables. In the second step, variables with larger coefficients will receive smaller penalties, which leads to the selection consistency. Since the Lasso is used as the initial estimator in the adaptive Lasso, we call this approach the iterated Lasso to distinguish it from the adaptive Lasso that uses different initial estimators.

We show that, under appropriate conditions, the Lasso is consistent even when $p > n$, and the iterated Lasso possesses an oracle property in the sense of Fan and Li (2001), it correctly selects important covariates with high probability and estimates the nonzero coefficients with the same asymptotic distribution as though the true model were known in advance. Our results are among the first in its kind that establish the consistency of the Lasso and the oracle property of the iterated Lasso in logistic regression in sparse, high-dimensional settings. They provide theoretical justifications for using the Lasso to generate the initial estimates of the coefficients, and then use them in the adaptive Lasso for logistic regression. The computational cost of the iterated Lasso is about twice that of the Lasso. Therefore, it is also computationally feasible for high-dimensional logistic regression models.

The rest of the paper is organized as follows. The iterated Lasso estimator is defined in Section 2. The results on the Lasso and the iterated Lasso in high-dimensional logistic regression are described in Section 3. In Section 4, we use simulation to evaluate the iterated Lasso in logistic regression and demonstrate it on a real data example. Concluding remarks are given in Section 5. Proofs are provided in the Appendix.

2 The iterated Lasso for logistic regression

Suppose that $Y_1, \ldots, Y_n$ are independent binary response variables that take a value of 0 or 1, and that $x_1, \ldots, x_n$ are corresponding covariates with $x_i = (1, x_{i1}, \ldots, x_{ip})', 1 \leq i \leq n$.

Define $\pi(t) = e^t/(1+e^t)$. The logistic regression model assumes that

$$\mathrm{P}(Y_i = 1|x_i) = \pi(x_i'\beta) = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)},\ 1 \le i \le n,$$

where $\beta = (\alpha, \beta_1, \ldots, \beta_p)'$ is a is a $(p+1)$-dimensional vector of coefficients including the intercept. The negative log-likelihood function

$$\ell_n(\beta) = -\sum_{i=1}^n \Big\{ Y_i \log \pi(x_i'\beta) + (1-Y_i)\log[1-\pi(x_i'\beta)] \Big\}.$$

Let

$$L_1(\beta;\lambda_1) = \ell_n(\beta) + \lambda_1 \sum_{j=1}^p |\beta_j|,$$

where only the regression coefficients $\beta_1, \ldots, \beta_p$ are penalized, but not the intercept. The Lasso estimator

$$\widetilde{\beta}_n(\lambda_1) = \arg\min_\beta L_{n1}(\beta;\lambda_1). \tag{1}$$

where $\lambda_1 \ge 0$ is a penalty parameter.

Consider the adaptive Lasso criterion

$$L_2(\beta;\lambda_2) = \ell_n(\beta) + \lambda_2 \sum_{j=1}^p w_j|\beta_j|, \tag{2}$$

where w_j is the weight for β_j. Here $\lambda_2 \ge 0$ is the penalty level and will be selected independently of λ_1. In general, the weights w_j can be data-dependent. The basic requirement for the weights is that the w_j should be relatively large if the true value of $\beta_j = 0$ and w_{nj} should be relatively small if $\beta_j \neq 0$. In reality, the values of the regression coefficients are not known in advance. However, a good initial estimator will provide this information. A convenient way to determine w_j is to use an initial estimator. Here we use the Lasso estimator $\widetilde{\beta}_n$ and let

$$w_j = |\widetilde{\beta}_{nj}|^{-1},\ \ j = 1, \ldots, p. \tag{3}$$

The iterated Lasso estimator

$$\widehat{\beta}_n(\lambda_2) = \arg\min_\beta L_2(\beta, \lambda_2).$$

When $\widetilde{\beta}_{nj} = 0$, we define $w_j = \infty$. Then to minimize (2), we must have $\beta_j = 0$. Thus, if a variable is not selected by the Lasso, it will also not be selected by the adaptive Lasso. In Theorem 1 below, we show that the Lasso will select all the nonzero coefficients if they are not too small, in the sense that the the smallest nonzero coefficient does not converge to zero too fast. This means that the Lasso over selects with high probability. Therefore, the iterated Lasso can be regarded as a way to correct the problem of over selection by the Lasso.

3 Theoretical results

In this section, we study the theoretical properties of the iterated Lasso estimators in sparse, high-dimensional cases in which $p \geq n$. First, we then show that the Lasso estimator is consistent. We then show that the adaptive Lasso possesses an oracle property, provided that an initial consistent estimator is available. Together, these results imply that the iterated Lasso has the oracle property.

3.1 Consistency of the Lasso

The adaptive Lasso estimator is defined based on an initial estimator. In particular, for the adaptive Lasso to have the oracle property, a key requirement is that the initial estimator that is consistent at zero. In the low-dimensional settings, the maximum likelihood estimator for the logistic regression model is consistent under certain regularity conditions. In this case, we can use the maximum likelihood estimator as the initial estimators for the weights. However, when $p_n > n$, which is the case in many microarray gene expression studies, the maximum likelihood estimator is no longer feasible. In this section, we show that the logistic Lasso estimator is consistent under the sparse Riesz condition (SRC) formulated in Zhang and Huang (2008) and certain other conditions.

Let the true parameter value be $\beta_0 = (\beta_{00}, \beta_{01}, \ldots, \beta_{0p_n})'$. Here we write p_n to indicate that p is allowed to diverge with n. Most quantities and data objects in our discussion are functions of n, but this dependence on n is often made implicit, especially for n-vectors and matrices with n rows. We normalize the covariates so that

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1,\ j = 1, \ldots, p_n.$$

Let $x^j = (x_{1j}, \ldots, x_{nj})'$ be the j-th covariate vector and let $X = (x^0, x^1, \ldots, x^{p_n})$, where x^0 is a $n \times 1$ column vector of 1's. For any $A \subseteq \{1, \ldots, p_n\}$, let

$$X_A = (x^0, x^j, j \in A),\ C_A = X_A' X_A / n.$$

Define

$$c_{\min}(m) = \min_{|A|=m} \min_{\|\nu\|=1} \nu' C_A \nu,\ c_{\max}(m) = \max_{|A|=m} \max_{\|\nu\|=1} \nu' C_A \nu,$$

where $\|\cdot\|$ denote the ℓ_2 norm. The covariate matrix X satisfies the SRC with rank q and spectrum bounds $0 < c_* < c^* < \infty$ if

$$c_* \leq c_{\min}(q) \leq c_{\max}(q) \leq c^*, \quad \forall\ A \text{ with } |A| = q \text{ and } v \in \mathbb{R}^q. \tag{4}$$

Since $\|X_A \nu\|^2/n = \nu' C_A \nu$, all the eigenvalues of C_A are inside the interval $[c_*, c^*]$ under (4) when the size of A is no greater than q^*. Let ρ_n be the largest eigenvalue of $X'X/n$.

Denote the set of nonzero coefficients by $A^o = \{j : \beta_{0j} \neq 0, 1 \leq j \leq p_n\}$. Let $k_n = |A^o|$ be the number of nonzero coefficients and let $m_n = p_n - k_n$ be the number of zero coefficients. Let $b_{n2} = \max\{|\beta_{0j}| : j \in A^o\}$ be the absolute largest value of the nonzero coefficients.

Consider the following conditions.

(A1) (a) Bounded parameters: there exists a $0 < b_2 < \infty$ such that $b_{n2} < b_2$; (b) Bounded covariates: there exists a constant $M > 0$ such that $\max_{1\leq i\leq n} \max_{1\leq j\leq p_n} |x_{ij}| < M$.

(A2) (SRC) There exists a constant $M_1 > 0$, such that with $q_n \equiv M_1 n^2/\lambda_{n1}^2$, the covariate matrix X satisfies the SRC with rank q_n and spectrum bounds $\{c_*, c^*\}$.

Condition (A1a) requires that the regression coefficients do not diverge to infinity. Condition (A1b) requires that the covariates be bounded. This condition can probably be weakened but it facilitates the technical details in the proofs.

Under (A1), there exists a constant $0 < \epsilon_0 < 1/2$ such that

$$\epsilon_0 \leq \pi(x_i'\beta)(1 - \pi(x_i'\beta)) \leq 1 - \epsilon_0, 1 \leq i \leq n. \tag{5}$$

Let $\gamma = c_*^2\epsilon_0^2$. Define

$$h_n = \frac{q_n \log p_n}{\gamma n} + \frac{2\lambda_{n1}^2(|A^o| + 1)}{\gamma n^2}. \tag{6}$$

Theorem 1 *Suppose that for a finite constant $\rho > 0$, $\rho_n \leq \rho$ for all n sufficiently large.*

(i) Let $\widetilde{A} = \{j : \widetilde{\beta}_j \neq 0\}$. Then, with probability one,

$$|\widetilde{A}| \leq \rho n^2/\lambda_{n1}^2.$$

(ii) Additionally, suppose that (A1)-(A2) hold. If $h_n(d_n \log n)^{1/2} \to 0$, then

$$\|\widetilde{\beta}_n - \beta_0\|^2 = O_p(h_n).$$

Part (i) of this theorem provides an upper bound of the dimension of the model selected by the Lasso in terms of ρ, n and λ_{n1}. It says that the dimension of the model selected by the Lasso is bounded by $\rho n^2/\lambda_{n1}^2$. In particular, the number of nonzero estimates is inversely proportional to λ_{n1}^2. Part (ii) shows that the rate of convergence of the Lasso estimator $r_n = h_n^{-1/2}$.

3.2 An oracle property of the iterated Lasso

We study the property of the iterated Lasso by considering the general case of the adaptive Lasso. The following definition describes a key requirement on the initial estimator for the adaptive Lasso.

Definition 1 *Let $b_{n1} = \min\{|\beta_{0j}| : j \in A^o\}$ the absolute smallest value of the nonzero coefficient. An estimator $\widetilde{\beta}_n$ is r_n-consistent at zero if there exists a constant $\xi_0 > 0$ such that* $\mathrm{P}(\min_{j\in A^o} |\widetilde{\beta}_{nj}| \geq \xi_0 \, b_{n1}) \to 0$ *and* $r_n \max_{j\notin A^o, j\geq 1} |\widetilde{\beta}_{nj}| = O_p(1)$, *where* $r_n \to \infty$ *as* $n \to \infty$.

In general, r_n-consistency at zero is different from the usual r_n-consistency. However, r_n-consistency at zero follows from the usual r_n-consistency if the smallest nonzero coefficient satisfies

$$b_{n1}/h_n^{1/2} \to \infty. \tag{7}$$

When the number of nonzero coefficients is fixed, this is always satisfied. In particular, under (7) and the conditions in Theorem 1, the Lasso estimator is r_n-consistent at zero with $r_n = h_n^{-1/2}$.

Consider the following assumption.

(A3) (a) The initial estimator $\widetilde{\beta}_n$ is r_n-consistent at zero; (b) The constants $\{k_n, m_n, b_{n1}, r_n, \lambda_{n2}\}$ satisfy

$$\frac{\sqrt{\log k_n}}{b_{n1}\sqrt{n}} + \frac{\sqrt{n \log m_n}}{\lambda_{n2}\, r_n} + \frac{\sqrt{k_n}\lambda_{n2}}{n b_{n1}} \to 0, \tag{8}$$

where we write λ_{n1} for λ_1 and λ_{n2} for λ_{n2} to indicate their dependence on n.

Condition (A3b) put restrictions on the numbers of covariates with zero and nonzero coefficients, the penalty parameter, and the smallest non-zero coefficient. It is worth noting that only the logarithm of m_n enters the equation. This implies that our results are applicable to models whose dimension is larger than n.

We now examine condition (8) in more detail. The first term in (8) requires that $b_{n1}/\sqrt{\log(k_n)/n} \to \infty$. In other words, for the adaptive Lasso to be able to distinguish nonzero coefficients from zero coefficients, we require that the nonzero coefficients be bigger than $\sqrt{\log(k_n)/n}$. If the number of nonzero coefficients is fixed, this is automatically satisfied. The second term requires that λ_{n2} cannot be too small and that the rate of convergence of the initial estimator at zero should not be two slow. For the given rate of convergence r_n of the initial estimator, the penalty level needed for consistent selection is related to the logarithm of the number of zero coefficients. The third them requires that λ_{n2} must be smaller than $nb_{n1}/\sqrt{k_n}$. This requirement is related to the bias introduced in the penalty term. For the bias due to penalty to be asymptotically negligible, more stringent condition is needed. See Theorem 3 below.

There are two special cases where (8) is especially simple: (1) When p_n is fixed, then b_{n1} is bounded away from zero. In this case, (8) is satisfied if $\lambda_{n2}/n \to 0$ and $\lambda_{n2} r_n/n^{1/2} \to \infty$. (2) The number of nonzero coefficients k_n is fixed, but p_n is allowed to diverge with n. Then b_{n1} is bounded away from zero. (A3b) is satisfied if $\lambda_{n2}/n \to 0$ and $\log p_n = o(1)(\lambda_{n2} r_n/n^{1/2})^2$.

Therefore, depending on r_n and λ_{2n}, the total number of covariates can be as large as $\exp(n^a)$ for some $0 < a < 1$.

For any vector $u = (u_1, u_2, \ldots)'$, define $\mathrm{sgn}(u) = (\mathrm{sgn}(u_1), \mathrm{sgn}(u_2), \ldots)'$, where $\mathrm{sgn}(u_1) = -1, 0$ or 1 if $u_1 < 0, = 0$ or 1. With a slight abuse of notation, we define $\mathrm{sgn}(\beta_{00}) = \mathrm{sgn}(\widehat{\beta}_{n0}) = 0$. That is, the signs of the intercept β_{00} and its estimate $\widehat{\beta}_{n0}$ are defined to be 0, although the intercept itself may not be zero.

Theorem 2 *Suppose that (A1)-(A3) hold and that the matrix C_{A^o} is strictly positive definite. Then*

$$\mathrm{P}\big(sgn(\widehat{\beta}_n) = sgn(\beta_0)\big) \to 1.$$

Thus under the conditions of Theorem 2, the adaptive Lasso can correctly select nonzero coefficients with probability converging to one.

Denote $\beta_{0A^0} = (\beta_{00}, \beta_{0j}, j \in A^o)'$, $\widehat{\beta}_{nA^0} = (\widehat{\beta}_{n0}, \widehat{\beta}_{nj}, j \in A^o)'$ and $x_{iA^o} = (1, x_{ij}, j \in A^o)'$. Let $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$, where $\epsilon_i = Y_i - \pi(x_i'\beta_0)$, $1 \le i \le n$, and let the diagonal matrix $D = \mathrm{diag}(d_1, \ldots, d_n)$, where $d_i = \pi(x_i'\beta_0)(1 - \pi(x_i'\beta_0)), 1 \le i \le n$. Denote $\Sigma_{A^o} = n^{-1}X_{A^o}'DX_{A^o}$.

Theorem 3 *Suppose that (A1) to (A3) hold. Let $s_n^2 = \sigma^2\alpha_n'\Sigma_{A^o}^{-1}\alpha_n$ for any $k_n \times 1$ vector α_n satisfying $\|\alpha_n\|_2 \le 1$. If $k_n^{1/2}\lambda_{n2}/n^{1/2} \to 0$,*

$$n^{1/2}s_n^{-1}\alpha_n'(\widehat{\beta}_{nA^o} - \beta_{0A^o}) = n^{-1/2}s_n^{-1}\sum_{i=1}^{n}\epsilon_i\alpha_n'\Sigma_{A^o}^{-1}x_{iA^o} + o_p(1) \to_D N(0,1),$$

where $o_p(1)$ is a term that converges to zero in probability uniformly with respect to α_n.

This theorem implies that the adaptive Lasso estimators of the non-zero parameters have the same asymptotic distribution that they would have if the zero coefficients were known in advance.

Under (7) and the conditions of Theorems 1 and 2, the iterated Lasso estimator is selection consistent. In addition, Theorem 3 implies that the iterated Lasso estimator is asymptotically efficient. Therefore, the iterated Lasso estimator has the asymptotic oracle property.

4 Computation and numerical studies

4.1 Computational Algorithm

The proposed approach involves computing the Lasso and adaptive Lasso estimates. One possibility is to use the modified LARS algorithm (Efron et al. 2004; Park and Hastie 2007).

As an alternative, we consider the gradient boosting approach (Kim and Kim 2004). First, we note that the Lasso estimator defined as

$$\widehat{\boldsymbol{\beta}} = \arg\min\{\ell_n(\beta) + \lambda \sum_j |\beta_j|\}$$

is equivalent to

$$\widehat{\beta} = \arg\min \ell_n(\beta)$$

subject to $\sum_j |\beta_j| \leq u$, with a one-to-one correspondence between λ and u. The above constraint estimate can be computed with the following iterative algorithm.

1. Initialize $\widehat{\beta} = 0$ and $s = 0$.
2. With the current estimate of $\widehat{\beta}$, compute $\psi(\beta) = \partial\ell_n(\beta)/\partial\beta$. Denote the kth component of ψ as ψ_k.
3. Find k^* that minimizes $\min(\psi_k(\beta), -\psi_k(\beta))$. If $\psi_{k^*}(\beta) = 0$, then stop the iteration.
4. Otherwise denote $r = -\text{sign}(\psi_{k^*}(\beta))$. Find $\hat{\pi} = \text{argmin}_{\pi \in [0,1]} \ell_n((1-\pi)\beta + \pi \times u \times r\eta^{k^*})$, where η^{k^*} has the k^*th element equals to 1 and the rest equal to 0.
5. Let $\beta_k = (1-\hat{\pi})\beta_k$ for $k \neq k^*$ and $\beta_{k^*} = (1-\hat{\pi})\beta_{k^*} + ru\hat{\pi}$. Let $s = s+1$.
6. Repeat steps 2–5 until convergence or a fixed number of iterations S has been reached. The β at convergence is the Lasso estimate.

The gradient boosting algorithm is less sensitive to the number of covariates, and can be more efficient than alternatives (Kim and Kim 2004). The adaptive Lasso estimator can also be computed using this algorithm by a simple rescaling of the covariates.

The proposed approach involves tuning parameters u_1 and u_2 corresponding to λ_1 and λ_2, respectively, which determines the sparsity of the estimates. In this study, we use V-fold cross validation to select u_1 and u_2.

4.2 Simulation Study

We conduct simulation studies to assess the finite sample performance of the iterated Lasso. We are interested in comparing performance of the proposed iterated Lasso method with the standard Lasso.

In the simulations, we generate n iid samples, with $n/2$ having $Y = 1$ (cases) and the rest $Y = 0$ (controls), and $p = 500$ covariates for each sample. Covariates for cases and controls are generated as multivariate normal distributed, with pair-wise correlation coefficient between the i^{th} and j^{th} covariates $\rho^{|i-j|}$. We set the marginal variances for all covariates equal to 1. The means for all covariates of controls are 0. The means for the first $k = 20$ covariates of cases are μ, and the rest 0. The simulated scenario corresponds to the logistic regression model, with the first $k = 20$ covariates having nonzero coefficients. We consider the following simulation parameters: $n = 100$ or 200, $\rho = 0$, (no correlation) 0.3 (weak correlation) and 0.5 (strong correlation), and $\mu = 1.0$ (strong signal) and 0.5 (weak signal). A total of 12 models are generated by considering different combinations of simulation parameters. We simulate 100 datasets under each simulation scenario.

We use V-fold cross validation to select the optimal tunings with $V = 5$. Our theoretical investigations show that the requirements for λ_1 and λ_2 are different. Therefore, the penalty parameters will be determined independently via 5-fold cross validation.

In Table 1, we show the medians (of 100 replicates) of the number of selected covariates and number of true positives using the Lasso and iterated Lasso approaches. We see that under all simulated scenarios, the iterated Lasso has smaller false positive rates. It selects a model that is closer to the true model and that it has smaller false positive rates and has similar or only slightly larger false negative rates.

4.3 Analysis of Breast Cancer Study

Breast cancer is the second leading cause of deaths from cancer among women in the United States. Despite major progresses in breast cancer treatment, the ability to predict the metastatic behavior of tumor remains limited. The Breast Cancer study was first reported in van't Veer et al. (2002). 97 lymph node-negative breast cancer patients 55 years old or younger participated in this study. Among them, 46 developed distant metastases within 5 years (metastatic outcome coded as 1) and 51 remained metastases free for at least 5 years (metastatic outcome coded as 0). The dataset is publicly available at *http://www.rii.com/publications/2002/vantveer.html.*

We first pre-process gene expression data as follows: (1) Remove genes with more than 30% missing measurements. (2) Fill in missing gene expression measurements with median values across samples. (3) Normalize gene expressions to have zero means and unit variances. (4) Compute the simple correlation coefficients of gene expressions with the binary outcome. (5) Select the 500 genes with the largest absolute values of correlation coefficients.

We analyze the breast cancer data using both the Lasso and the proposed iterated approach. Optimal tunings are selecting using the 5-fold cross validation. Since with practical data, it is not clear which are the true positives, we use the leave-one-out cross validation (LOOCV) to evaluate the predictive power of the two approaches. If an approach can properly select genes with predictive power, then the LOOCV prediction error should be small.

We show the estimation results in Table 2. With the Lasso, 42 genes are selected. With the iterated Lasso, only 22 genes are selected. Estimates (if nonzero) under both approaches have the same signs, which suggests similar biological conclusions. However the estimates can be different. With the Lasso, 14 subjects cannot be properly predicted. With the proposed approach, 14 subjects cannot be properly predicted. So the same predictive performance is achieved with both approaches. However, the findings from the iterated Lasso provide a more focused set of genes for further investigation of their biological functions.

5 Concluding remarks

A key requirement for the iterative Lasso to possess the oracle property is that the initial estimator is consistent and does not miss important variables with high probability. In low-dimensional settings, finding an initial consistent estimator is relatively easy and can be achieved by many well established approaches such as the maximum likelihood method. However, in high-dimensional settings, finding an initial consistent estimator is difficult. Under the conditions stated in Theorem 1, the Lasso is shown to be consistent and selects all the important variables as long as their corresponding coefficients are not too small. Thus the Lasso can be used as the initial estimator in the adaptive Lasso to achieve asymptotic oracle efficiency. Our simulation results show that the iterated Lasso performs well in terms variable selection. Therefore, the iterated Lasso is a useful approach for variable selection and estimation in sparse, high-dimensional logistic regression.

Our theoretical results on the Lasso (Theorem 1) depend on the form of the logistic likelihood. However, the results on the adaptive Lasso (Theorems 2 and 3) do not require logistic model assumption, provided that a consistent initial estimator is available. It is clear that the iterated Lasso can also be applied to general regression problems, including other generalized linear model, Cox regression and robust regression. Further work is needed to verify that similar theoretical results still hold generally in these problems.

6 Appendix: Proofs

In this section, we prove the results stated in Section 3. For simplicity, we will drop the subscript n from certain quantities in many instances. For example, we will simply write p for p_n, λ_1 for λ_{n1} etc. We first prove Theorems 2 and 3, then we prove Theorem 1.

6.1 Proof of Theorems 2 and 3

Denote the log-likelihood function by $\ell(y,\theta) = y\log\pi(\theta) + (1-y)\log(1-\pi(\theta))$. Simple calculation shows that the first and second derivatives of ℓ with respect to θ are $\dot{\ell}(y,\theta) = -y + \pi(\theta)$ and $\ddot{\ell}(y,\theta) = \pi(\theta)(1-\pi(\theta))$.

Proof of Theorem 2. By the Karush-Kunh-Tucker conditions, $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)'$ is the unique adaptive Lasso estimator if and only if

$$\begin{cases} \sum_{i=1}^n \dot{\ell}\big(Y_i, x_i'\widehat{\beta}\big)x_{ij} = \lambda_2 w_j \mathrm{sgn}(\widehat{\beta}_j), & j = 0 \text{ or } \widehat{\beta}_j \neq 0, \\ \big|\sum_{i=1}^n \dot{\ell}\big(Y_i, x_i'\widehat{\beta}\big)x_{ij}\big| \leq \lambda_2 w_{nj} & \widehat{\beta}_j = 0, j \geq 1. \end{cases} \tag{9}$$

Let $\beta_{0A^o} = (\beta_{00}, \beta_{0j}, j \in A^o)$, $\widehat{\beta}_{A^o} = (\widehat{\beta}_0, \widehat{\beta}_j, j \in A^o)$ and $x_{iA^o} = (1, x_{ij}, j \in A^o)'$. We write $\mathrm{sgn}(\widehat{\beta}_{A^o}) = \mathrm{sgn}(\beta_{0A^o})$ to mean that the signs are equal component wise with the convention that $\mathrm{sgn}(\widehat{\beta}_0) = \mathrm{sgn}(\beta_{00}) \equiv 0$. If $\mathrm{sgn}(\widehat{\beta}_{A^o}) = \mathrm{sgn}(\beta_{A^o})$, then (9) holds for $\widehat{\beta} = (\widehat{\beta}_{A^o}', \mathbf{0}')'$. Thus, since $x_i'\widehat{\beta} = x_{iA^o}'\widehat{\beta}_{A^o}$ for this $\widehat{\beta}$ and x^j are linearly independent,

$$\mathrm{sgn}(\widehat{\beta}) = \mathrm{sgn}(\beta_0) \quad \text{if} \quad \begin{cases} \mathrm{sgn}(\widehat{\beta}_j) = \mathrm{sgn}(\beta_{0j}), j \in A^o, \\ \Big|\sum_{i=1}^n \dot{\ell}\big(Y_i, x_{iA^o}'\widehat{\beta}_{A^o}\big)x_{ij}\Big| \leq \lambda_2 w_j, \ \forall j \notin A^o, j \geq 1, \\ \sum_{i=1}^n \dot{\ell}\big(Y_i, x_{iA^o}'\widehat{\beta}_{A^o}\big) = 0. \end{cases} \tag{10}$$

where $\widehat{\beta}_{A^o}$ is the solution to

$$\sum_{i=1}^n \dot{\ell}\big(Y_i, x_{iA^o}'\widehat{\beta}_{A^o}\big)x_{ij} = -\lambda_2 w_j \,\mathrm{sgn}(\beta_{0j}),\ j = 0 \text{ and } j \in A^o.$$

To prove the theorem, it suffices to prove (10). We can write (10) as

$$\mathrm{sgn}(\widehat{\beta}) = \mathrm{sgn}(\beta_0) \quad \text{if} \quad \begin{cases} \mathrm{sgn}(\beta_{0j})(\beta_{0j} - \widehat{\beta}_j) < |\beta_{0j}|, \forall j \in A^o \\ \big|\sum_{i=1}^n \dot{\ell}\big(Y_i, x_{iA^o}'\widehat{\beta}_{A^o}\big)x_{ij}\big| \leq \lambda_2 w_j, \ \forall j \notin A^o, j \geq 1, \\ \sum_{i=1}^n \dot{\ell}\big(Y_i, x_{iA^o}'\widehat{\beta}_{A^o}\big) = 0. \end{cases} \tag{11}$$

Define

$$Q_{n1}(\beta_{A^o}) = \frac{1}{2}(\beta_{A^o} - \beta_{0A^o})'X_1'DX_1(\beta_{A^o} - \beta_{0A^o}) - \epsilon'X_1(\beta_{A^o} - \beta_{0A^o}) + \lambda_2 \sum_{j \in A^o} w_{nj}\mathrm{sgn}(\beta_{0j})\beta_j,$$

where $X_1 \equiv X_{A^o} = (x^0, x^j, j \in A^o)$ and recall $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$. Let $\beta_{A^o}^* = \arg\min Q_{n1}(\beta_{A^o})$. Denote $\xi_1 = (0, w_{nj}\mathrm{sgn}(\beta_{0j}), j \in A^o)'$. Then

$$\beta_{A^o}^* - \beta_{0A^o} = (X_1DX_1)^{-1}(X_1'\epsilon - \lambda_2\xi_1). \tag{12}$$

Thus we have

$$\|\beta^*_{A^o} - \beta_{0A^o}\|^2 = O_p\left(\frac{k_n}{n} + \frac{\lambda_2^2 k_n}{n^2 b_{n1}^2}\right). \tag{13}$$

Following Huang, Ma, and Zhang (2007), it can be shown that, with probability converging to one,

$$\begin{cases} |\beta_{0j} - \beta^*_j| < 0.5|\beta_{0j}|, & \forall j \in A^o \\ \left|\sum_{i=1}^n [\epsilon_i - d_i x'_{iA^o}(\beta^*_{A^o} - \beta_{A^o})]x_{ij}\right| \le \frac{1}{4}\lambda_2 w_j, & \forall j \notin A^o, j \ge 1. \end{cases} \tag{14}$$

Since $\epsilon_i = \dot{\ell}(Y_i, x'_{iA^o}\beta_{0A^o})$, by the Taylor expansion, we have

$$\left|\sum_{i=1}^n \dot{\ell}(Y_i, x_{iA^o}\beta^*_{A^o})x_{ij} - \sum_{i=1}^n [\epsilon_i + d_i x'_{iA^o}(\beta^*_{A^o} - \beta_{0A^o})]x_{ij}\right| \le M_1 c^* \|\beta^*_{A^o} - \beta_{0A^o}\|_2^2.$$

This inequality, (14) and (13) imply that, with high probability,

$$\begin{cases} |\beta_{0j} - \beta^*_j| < 0.5|\beta_{0j}|, & \forall j \in A^o \\ \left|\sum_{i=1}^n \dot{\ell}(Y_i, x'_{iA^o}\beta^*_{A^o})x_{ij}\right| \le 0.5\lambda_2 w_j, & \forall j \notin A^o. \end{cases}$$

Therefore, by (11), to prove (10), it suffices to show that with probability converging to 1,

$$\begin{cases} |\widehat{\beta}_{A^o} - \beta^*_{A^o}| < 0.5|\beta_{0j}|, & \forall j \in A^o \\ \frac{1}{n}\left|\sum_{i=1}^n \dot{\ell}(Y_i, x'_{iA^o}\widehat{\beta}_{A^o})x_{ij} - \sum_{i=1}^n \dot{\ell}(Y_i, x'_{iA^o}\beta^*_{A^o})x_{ij}\right| \le 0.5\lambda_2 w_j, & \forall j \notin A^o. \end{cases} \tag{15}$$

By Lemma 4 below, we can show that

$$\|\widehat{\beta}_{A^o} - \beta^*_{A^o}\|^2 = o_p\Big(\frac{k_n}{n} + \frac{\lambda_2^2 k_n}{n^2 b_{n1}^2}\Big). \tag{16}$$

By (A1b), we have

$$\begin{aligned} \left|\sum_{i=1}^n [\dot{\ell}(Y_i, x'_{iA^o}\widehat{\beta}_{A^o}) - \dot{\ell}(Y_i, x'_{iA^o}\beta^*_{A^o})]x_{ij}\right| &\le O(1)\sum_{i=1}^n |x'_{iA^o}(\widehat{\beta}_{A^o} - \beta^*_{A^o})||x_{ij}| \\ &\le O(1)M_1 \sum_{i=1}^n \|x_{iA^o}\|\|\widehat{\beta}_{A^o} - \beta^*_{A^o}\| \\ &\le O(1)nM_1^2 k_n^{1/2}\|\widehat{\beta}_{A^o} - \beta^*_{A^o}\|. \end{aligned} \tag{17}$$

Therefore, (15) follows from (16), (17), and (A3b). □

Proof of Theorem 3. Theorem 3 follows from (12), (13), (16), the assumption that $\sqrt{k_n}\lambda_2/\sqrt{n} \to 0$ and the Lindeberg-Feller central limit theorem.

6.2 Proof of Theorem 1

Proof of Theorem 3, part (i). By the KKT, a necessary and sufficient condition for $\widetilde{\beta} = (\widetilde{\beta}_0, \widetilde{\beta}_1, \ldots, \widetilde{\beta}_p)'$ to be the Lasso solution is

$$\begin{cases} \sum_{i=1}^n [Y_i - \pi(x_i'\widetilde{\beta})] = 0, & \\ \sum_{i=1}^n [Y_i - \pi(x_i'\widetilde{\beta})]x_{ij} = \lambda_1 \operatorname{sgn}(\widetilde{\beta}_j), & \widetilde{\beta}_j \neq 0, j \geq 1, \\ \left| \sum_{i=1}^n [Y_i - \pi(x_i'\widetilde{\beta})]x_{ij} \right| \leq \lambda_1, & \widetilde{\beta}_j = 0, j \geq 1. \end{cases} \tag{18}$$

Let $u = (Y_1 - \pi(x_1'\widetilde{\beta}), \ldots, Y_n - \pi(x_n'\widetilde{\beta}))'$. By (18), $\|X_{A_1}'u\|_2^2 = |A_1|\lambda_1^2$. Thus $|A_1|\lambda_1^2 = u'X_{A_1}X_{A_1}'u \leq nc_{\max}(|A_1|)\|u\|^2$. Since $\|u\|_2^2 \leq n$ with probability one, it follows that $|A_1| \leq c_{\max}(|A_1|)n^2/\lambda_1^2$. This completes the proof of part (i).

Let $\epsilon_i = Y_i - \pi(x_i'\beta_0)$ and $d_i = \pi(x_i'\beta_0)[1 - \pi(x_i'\beta_0)], 1 \leq i \leq n$. By (5), Taylor expansion gives

$$\ell_n(\beta) - \ell_n(\beta_0) \;=\; \sum_{i=1}^n \left\{ \frac{1}{2} d_i [x_i'(\beta - \beta_0)]^2 - \varepsilon_i x_i'(\beta - \beta_0) \right\} + R_n(\beta),$$

where

$$|R_n(\beta)| \leq \frac{1}{24} \sum_{i=1}^n [x_i'(\beta - \beta_0)]^3. \tag{19}$$

Define

$$Q_n(\beta) = \frac{1}{2} \sum_{i=1}^n d_i [x_i'(\beta - \beta_0)]^2 - \sum_{i=1}^n \varepsilon_i x_i'(\beta - \beta_0) + \lambda_1 \sum_{j=1}^p |\beta_j|.$$

Let $\beta_n^* = \arg\min Q_n(\beta)$.

Lemma 1 *Let $A^* = \{j : \beta_{nj}^* \neq 0\}$. With probability 1,*

$$|A^*| \leq M_2 c_{\max}(|A^*|) \frac{n^2}{\lambda_1^2} \leq M_2 c_{\max}(\min\{n, p_n\}) \frac{n^2}{\lambda_1^2},$$

where $M_2 > 0$ is a finite constant.

This lemma can be proved the same way as Theorem 1, part (i).

Lemma 2 *Under conditions of Theorem 1,*

$$\|\beta_n^* - \beta_0\|_2^2 \leq \frac{2\lambda_1^2 |A^o|}{n^2 c_*^2 \gamma^2} + O_p\left(\frac{\log p_n}{n} \frac{q_n}{c_*^2 \gamma^2} \right).$$

Proof. Write

$$Q_n(\beta) = \frac{1}{2}(\beta - \beta_0)' X' D X (\beta - \beta_0) - \epsilon' X (\beta - \beta_0) + \lambda_1 \sum_{j=1}^p |\beta_j|.$$

By the definition of β_n^*, $Q_n(\beta_n^*) \leq Q_n(\beta_0)$. This means

$$\frac{1}{2}(\beta_n^* - \beta_0)'X'DX(\beta_n^* - \beta_0) - \epsilon' X(\beta_n^* - \beta_0) + \lambda_1 \|\beta_n^*\|_1 \leq \lambda_1 \|\beta_0\|_1,$$

where $\|\cdot\|_1$ denotes the ℓ_1 norm. Thus

$$\frac{1}{2}(\beta_n^* - \beta_0)'X'DX(\beta_n^* - \beta_0) - \epsilon' X(\beta_n^* - \beta_0) \leq \lambda_1 \sum_{j \in A^o} |\beta_{nj}^* - \beta_{0j}|. \tag{20}$$

Let $B = A^o \cup A^* = \{j : \beta_{0j} \neq 0 \text{ or } \beta_{nj}^* \neq 0\}$, $X_B = (x^0, x^j, j \in B)$, $\beta_{nB}^* = (\beta_0^*, \beta_{nj}^*, j \in B)$, and $\beta_{0B} = (\beta_0, \beta_{0j}, j \in B)$. Denote $v = D^{-1/2}\varepsilon$ and $\eta_B = D^{1/2}X_B(\beta_{nB}^* - \beta_{0B})$. Since $A^* \in B$, the Cauchy-Schwarz inequality implies,

$$\sum_{j \in A^o} |\beta_{nj}^* - \beta_{0j}| \leq \sqrt{|A^o|}\, \|\beta_{nA^*}^* - \beta_{0A^*}\|_2 \leq \sqrt{|A^o|}\, \|\beta_{nB}^* - \beta_{0B}\|_2. \tag{21}$$

By (20) and (21),

$$\|\eta_{nB}\|^2 - 2v'\eta_B \leq \lambda_1 2\sqrt{|A^o|}\, \|\beta_{nB}^* - \beta_{0B}\|_2. \tag{22}$$

Let v^* be the projection of v to the span of $D^{1/2}X_B$, i.e., $v^* = D^{1/2}X_B(X_B'DX_B)^{-1}X_B'D^{1/2}v$. By the Cauchy-Schwarz inequality,

$$2|v'\eta_B| \leq 2\|v^*\|_2 \cdot \|\eta_B\|_2 \leq 2\|v^*\|_2^2 + \frac{1}{2}\|\eta_B\|_2^2. \tag{23}$$

Combining (22) and (23), $\|\eta_B\|_2^2 \leq 4\|v^*\|_2^2 + 2\lambda_1\sqrt{|A^o|} \cdot \|\beta_{nB}^* - \beta_{0B}\|_2$. Since $\|\eta_B\|_2^2 \geq nc_*\epsilon_0\|\beta_{nB}^* - \beta_{0B}\|_2^2$ and $2ab \leq a^2 + b^2$,

$$nc_*\epsilon_0\|\beta_{nB}^* - \beta_{0B}\|_2^2 \leq 4\|v^*\|_2^2 + \frac{(2\lambda_1\sqrt{|A^o|})^2}{2nc_*\epsilon_0} + \frac{1}{2}nc_{n*}\epsilon_0\|\beta_{nB}^* - \beta_{0B}\|_2^2.$$

It follows that

$$\|\beta_{nB}^* - \beta_{0B}\|_2^2 \leq \frac{8\|v^*\|_2^2}{nc_*\epsilon_0} + \frac{4\lambda_1^2|A^o|}{n^2c_*^2\epsilon_0^2}.$$

Now $\|v^*\|_2^2 = \|(X_B'DX_B)^{-1/2}X_B'D^{1/2}\epsilon\|_2^2 \leq \|X_B\epsilon\|_2^2/(4nc_*\epsilon_0)$. We have

$$\max_{A:|A|\leq q_n} \|X_A'\varepsilon\|_2^2 = \max_{A:|A|\leq q_n} \sum_{j \in A} |\varepsilon' x^j|^2 \leq q_n \max_{1\leq j\leq p} |\varepsilon' x^j|^2.$$

Since $\|n^{-1/2}x^j\|_2^2 = 1, 1 \leq j \leq p$, by the maximal inequality for subgaussian random variables, $\max_{1\leq j\leq p}|\varepsilon' x^j|^2 = n\max_{1\leq j\leq p}|n^{-1/2}\varepsilon' x^j|^2 = O_p(n\log p_n)$. Therefore,

$$\|\beta_{nB}^* - \beta_{0B}\|_2^2 \leq O(1)\frac{\log(p_n)}{nc_{n*}^2\epsilon_0^2} + \frac{4\lambda_1^2|A^o|}{n^2c_{n*}^2\epsilon_0^2}.$$

The lemma follows. □

For any subset $A \subset \{1, \ldots, p_n\}$ and $\beta_A = (\beta_0, \beta_j : j \in A)$, write $Q_n(\beta_A) = Q_n((\beta_A', 0_{p-|A|}')')$, where $0_{p-|A|}$ is a $p - |A|$ dimensional vector of zeros. In $Q_n(\beta_A)$, the columns of X corresponding to the zeros in $(\beta_A', 0_{p-|A|}')'$ are dropped from the expression. Similarly define $L_n(\beta_A)$.

Let $A = A^o \cup \widetilde{A} \cup A^* = \{j : \beta_{0j} \neq 0 \text{ or } \widetilde{\beta}_{nj} \neq 0 \text{ or } \beta_{nj}^* \neq 0\}$. Since $\widetilde{\beta} = (\widetilde{\beta}_{nA}', 0_{p-|A|}')'$ and $\beta_n^* = (\beta_{nA}^{*'}, 0_{p-|A|}')'$ minimize $L_n(\beta)$ and $Q_n(\beta)$, respectively, we have

$$\widetilde{\beta}_{nA} = \arg\min L_n(\beta_A, \lambda_1) \quad \text{and} \quad \beta_{nA}^* = \arg\min Q_n(\beta_A, \lambda_1).$$

Lemma 3 *Define* $\Lambda_n(\delta, A) = \inf_{\|\beta_A - \beta_{nA}^*\|_2 = \delta} Q_n(\beta_A; \lambda_1) - Q_n(\beta_{nA}^*; \lambda_1)$. *Under the conditions of Theorem 1,*

$$\Lambda_n(\delta, A) \geq nc_*\epsilon_0\delta^2. \tag{24}$$

Proof of Lemma 3. Define $\nu(\beta) = (0, \nu(\beta_j), j \in A)'$ where $\nu(\beta_j)$ is a subdifferential of $|\beta_j|$. That is, $\nu(\beta_j) = 1$ if $\beta_j > 0$, $\nu(\beta_j) = -1$ if $\beta_j < 0$, and $-1 \leq \nu(\beta_j) \leq 1$ if $\beta_j = 0$. Let $\Sigma_A = n^{-1}X_A'DX_A$. Since $Q_n(\beta_A)$ is minimized at β_{nA}^*, there is a subdifferential of Q_n at β_{nA}^* that satisfies $dQ_n(\beta_{nA}^*; \lambda_1) = n\Sigma_A(\beta_{nA}^* - \beta_{0A}) + X_A'\epsilon + \lambda_1\nu(\beta_{nA}^*) = 0$. By the property of a subdifferential, $\|\beta_A\|_1 - \|\beta_{nA}^*\|_1 \geq \sum_{j \in A} \nu(\beta_{nj}^*)(\beta_j - \beta_{nj}^*)$. Therefore,

$$Q_n(\beta_A; \lambda_1) - Q_n(\beta_{nA}^*; \lambda_1) \geq n(\beta_A - \beta_{nA}^*)'\Sigma_{nA}(\beta_A - \beta_{nA}^*) + n(\beta_A - \beta_{nA}^*)'dQ_n(\beta_{nA}^*).$$

It follows that $Q_n(\beta_A; \lambda_1) - Q_n(\beta_{nA}^*; \lambda_1) \geq n(\beta_A - \beta_{nA}^*)'\Sigma_{nA}(\beta_A - \beta_{nA}^*)$. The lemma follows from (A2). □

The following lemma from Hjort and Pollard (1993) concerning convex minimization will be useful.

Lemma 4 *(Hjort and Pollard 1993) Suppose* $A_n(s)$ *is a convex function in* $\mathbb{R}^p$ *and is approximated by* $B_n(\mathbf{s})$. *Let* $\mathbf{a}_n = \arg\min A_n(\mathbf{s})$ *and assume that* B_n *has a unique argmin* $\mathbf{b}_n$. *Then for each* $\delta > 0$, $\mathrm{P}\big(\|\mathbf{a}_n - \mathbf{b}_n\| \geq \delta\big) \leq \mathrm{P}\big(\Delta_n(\delta) \geq \Lambda_n(\delta)/2\big)$, *where*

$$\Delta_n(\delta) = \sup_{\|\mathbf{s} - \mathbf{b}_n\| \leq \delta} |A_n(\mathbf{s}) - B_n(\mathbf{s})| \quad \textit{and} \quad \Lambda_n(\delta) = \inf_{\|\mathbf{s} - \mathbf{b}_n\| = \delta} B_n(\mathbf{s}) - B_n(\mathbf{b}_n).$$

Here $\|\cdot\|$ *can be any norm in* $\mathbb{R}^p$.

Proof of Theorem 1, part (ii). For any $\delta > 0$, define $\Delta_n(\delta, A) = \sup_{\|\beta_A - \beta_{nA}^*\|_2 \leq \delta} |R_n(\beta_A)|$. Let $\Lambda_n(\delta, A)$ be as defined in Lemma 3. By Proposition 4,

$$\mathrm{P}\big(|\widetilde{\beta}_{nA} - \beta_{nA}^*\|_2 \geq \delta\big) \leq \mathrm{P}\big(\Delta_n(\delta, A) \geq \Lambda_n(\delta, A)/2\big).$$

By Lemma 3, $\Lambda_n(\delta, A) \geq nc_*\epsilon_0\delta^2$. Therefore,

$$\mathrm{P}\big(|\widetilde{\beta}_{nA} - \beta_{nA}^*\|_2 \geq \delta\big) \leq \mathrm{P}\big(\Delta_n(\delta, A) \geq nc_*\gamma\delta^2/2\big). \tag{25}$$

By Lemma 2, $\|\beta^*_{nA} - \beta_0\| = O_p(h_n)$. When $\|\beta_A - \beta^*_{nA}\|_2 \leq \delta$, we have $\|\beta_A - \beta_0\|_2 \leq \delta + \|\beta^*_{nA} - \beta_0\|_2$. Thus $\Delta_n(\delta) \leq \sup\big\{|R_n(\beta_A)| : \|\beta_A - \beta_0\|_2 \leq \delta + \|\beta^*_{nA} - \beta_0\|\big\}$. When $\|\beta_A - \beta_0\|_2 \leq \delta + \|\beta^*_{nA} - \beta_0\|_2$, by (19), $|R_n(\beta_A)| \leq O(1)\sum_{i=1}^n [(\beta_A - \beta_0)' x_{iA} x'_{iA}(\beta_A - \beta_0)]^{3/2}$. By condition (A1), $\max_{1\leq i\leq n}\|x_{iA}\| \leq q_n^{1/2} M_1$. This in turn implies

$$\Delta_n(\delta, A) \leq nc_*[\delta + O_p(h_n)]^3 q_n^{1/2} M_1,$$

where h_n is defined in (6). By (25), it follows that

$$\mathrm{P}\big(\|\widetilde{\beta}_{nA} - \beta^*_{nA}\|_2 \geq \delta\big) \leq \mathrm{P}\big(nc_*[\delta + O_p(h_n)]^3 q_n^{1/2} M_1 \geq c_*\delta^2/2\big). \tag{26}$$

For any δ that satisfies $\delta = o(h_n)$ and $(\delta + h_n)^3 q_n^{1/2}/\delta^2 \to 0$, the right-hand side of (26) goes to zero and $\widetilde{\beta}_{nA} - \beta^*_{nA} = o_p(h_n)$. Therefore, $\widetilde{\beta}_{nA} - \beta_0 = \widetilde{\beta}_{nA} - \beta^*_{nA} + \beta^*_{nA} - \beta_0 = O_p(h_n)$. In particular, we can choose $\delta_n = h_n/\log^{1/4} n$. Then the required condition on h_n and q_n is

$$\frac{h_n^3 q_n^{1/2}}{(h_n/\sqrt{\log n})^2} = h_n(q_n \log n)^{1/2} \to 0.$$

This completes the proof of Theorem 1. □

References

[1] Bühlmann, P. and Meier, L. (2007). Discussion of "One-step sparse estimates in nonconcave penalized likelihood models" by H. Zou and R. Li. *Ann. Statist.* **36**, 1534-1541.

[2] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with Discussion). *Ann. Statist.* **32**, 407-499.

[3] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties.*J. Am. Statist. Assoc.* **96**, 1348–1360.

[4] Hjort, N. L. and Pollard, D. (1993). Asymptotics for minimisers of convex processes. Preprint, Department of Statistics, Yale University.

[5] Huang, J., Ma, S. and Zhang, C.-H. (2006). Adaptive LASSO for sparse high-dimensional regression models. Technical Report 374, Department of Statistics and Actuarial Science, University of Iowa. To appear in *Statist. Sinica.*

[6] Kim, Y and Kim, J (2004). Gradient LASSO for feature selection. *Proceedings of the 21st International Conference on Machine Learning.* ACM International Conference Proceeding Series, **69**. Association for Computing Machinery. New York.

[7] Knight, K. and Fu, W. J. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28**, 1356-1378.

[8] Leng, C., Lin, Y., and Wahba, G. (2004). A Note on the LASSO and related procedures in model selection. *Statist. Sinica* **16**, 12731284.

[9] Meinshausen, N. and Buhlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34**, 1436-1462.

[10] Meinshausen, N. and Yu, B. (2008). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, to appear.

[11] Park, M. and Hastie, T. (2007). L1 regularization path for generalized linear models. *J. Roy. Statist. Soc. Ser. B* **69** 659-677.

[12] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267-288.

[13] van de Geer, S. (2006). High-dimensional generalized linear models and the Lasso. *Ann. Statist.* **2**, 614-645.

[14] Zhao, P. and Yu, B. (2006). On model selection consistency of LASSO. *J. Machine Learning Res.* **7**, 2541-2567.

[15] Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **4**, 1567-1594.

[16] Zhang, H. H. & Lu, W. B. (2007). Adaptive-Lasso for Cox's proportional hazards model. *Biometrika* **94**, 691-703.

[17] Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418-1429.

[18] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with Discussion). *Ann. Statist.* **36**, 1509-1533.

Table 1: Simulation study. Nonzero: median number of nonzero estimates. True: median of true positives. Numbers in "()" are inter-quartile ranges. The number of nonzero coefficients in the generating model is 20.

			Lasso		Proposed	
n	μ	ρ	Nonzero	True	Nonzero	True
100	1	0	34 (30,38)	15 (14,17)	24 (20,27)	14 (13,16)
200	1	0	46 (39,50)	19 (18,20)	33 (30,36)	18 (18,19)
100	1	0.3	35 (31,39)	14 (13,16)	24 (21,27)	13 (12,15)
200	1	0.3	45 (42,50)	17 (17,18)	33 (30,36)	17 (16,18)
100	1	0.5	35 (30,39)	13 (12,15)	24 (20,26)	13 (12,14)
200	1	0.5	48 (43,51)	16 (16,17)	31 (26,33)	16 (15,17)
100	0.5	0	42 (39,45)	17 (17,18)	29 (27,31)	16 (14,17)
200	0.5	0	60 (53,68)	20 (19,20)	40 (35,43)	19 (18,19)
100	0.5	0.3	42 (36,45)	16 (14,17)	29 (25,31)	14 (12,15)
200	0.5	0.3	69 (58,75)	18 (17,19)	40 (35,45)	17 (17,18)
100	0.5	0.5	41 (34,47)	13 (11,14)	25 (20,28)	11 (10,13)
200	0.5	0.5	51 (45,54)	16 (15,17)	35 (32,38)	15 (13,16)

Table 2: Analysis of Breast Cancer Data: Systematic names of identified genes; Estimates using Lasso and the proposed approach.

Systematics Name	Lasso	Proposed	Systematics Name	Lasso	Proposed
NM_003009	0.049	0	Contig64940_RC	-0.200	-0.320
NM_000936	-0.021	0	Contig20866_RC	0.035	0
AB033113	-0.265	-0.103	NM_004962	0.636	1.880
Contig52994_RC	-0.225	0	NM_006465	0.180	0
NM_002494	0.038	0.046	NM_007235	0.068	0.039
AL133018	-0.200	-0.125	NM_006574	0.186	0.180
U89715	0.019	0	D86979	0.513	0.089
Contig44278_RC	-0.196	0	Contig35655_RC	0.620	1.078
Contig38907_RC	0.322	0	NM_014713	-0.025	-0.027
NM_012104	0.027	0	Contig33778_RC	0.020	0.003
NM_002749	-0.051	-0.040	Contig22842_RC	-0.042	0
NM_002789	-0.150	-0.073	AB037823	-0.477	-0.029
NM_020307	-0.028	-0.009	X61070	0.449	0
AB040924	0.336	0	NM_015846	-0.021	0
AB032989	0.146	0	NM_018010	0.083	0.059
Contig60950	-0.230	0	NM_000076	-0.135	0
Contig31018_RC	-0.026	0	Contig40751_RC	0.255	0.345
Contig15795_RC	-0.190	-0.177	Contig59553	-0.072	-0.027
NM_003885	0.257	0	Contig13678_RC	0.050	0
Contig719_RC	0.087	0.025	Contig26520_RC	-0.060	-0.043
NM_020682	0.307	0.040	AL157449	0.060	0