Biostatistics (2013), **14**, 4, *pp.* 653–666
doi:10.1093/biostatistics/kxt008
Advance Access publication on April 9, 2013

Controlling the local false discovery rate in the adaptive Lasso

JOSHUA N. SAMPSON*, NILANJAN CHATTERJEE

Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd, EPS 8038, Rockville, MD 20852, USA

joshua.sampson@nih.gov

RAYMOND J. CARROLL

Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143, USA

SAMUEL MÜLLER

School of Mathematics and Statistic, University of Sydney, New South Wales 2006, Australia

Summary

The Lasso shrinkage procedure achieved its popularity, in part, by its tendency to shrink estimated coefficients to zero, and its ability to serve as a variable selection procedure. Using data-adaptive weights, the adaptive Lasso modified the original procedure to increase the penalty terms for those variables estimated to be less important by ordinary least squares. Although this modified procedure attained the oracle properties, the resulting models tend to include a large number of "false positives" in practice. Here, we adapt the concept of local false discovery rates (lFDRs) so that it applies to the sequence, λ_n, of smoothing parameters for the adaptive Lasso. We define the lFDR for a given λ_n to be the probability that the variable added to the model by decreasing λ_n to $\lambda_n - \delta$ is not associated with the outcome, where δ is a small value. We derive the relationship between the lFDR and λ_n, show lFDR $= 1$ for traditional smoothing parameters, and show how to select λ_n so as to achieve a desired lFDR. We compare the smoothing parameters chosen to achieve a specified lFDR and those chosen to achieve the oracle properties, as well as their resulting estimates for model coefficients, with both simulation and an example from a genetic study of prostate specific antigen.

Keywords: Adaptive Lasso; Local false discovery rate; Smoothing parameter; Variable selection.

1. Introduction

The Lasso procedure offers a means to fit a linear regression model when the number of parameters p is comparatively large (Tibshirani, 1996, 2011). The Lasso estimates coefficients by minimizing the residual sum of squares plus a penalty term. Let there be n subjects, let $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ be their outcomes, let $X_j = (X_{j1}, \ldots, X_{jn})^{\mathrm{T}}$ be their measurements for variable $j = 1, \ldots, p$, and let $\mathbf{X} = (X_1, \ldots, X_p)$. Then

*To whom correspondence should be addressed.

the estimated coefficients are

$$\hat{\beta}(\lambda_n) = \text{argmin} \left\{ \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2 + \sum_{j=1}^{p} \lambda_n |\beta_j| \right\}.$$

A major benefit of the L_1 penalty is that the Lasso also serves as a variable selection method, as a large proportion of $\hat{\beta}_j$ are reduced to 0 when λ_n is large.

The adaptive Lasso modifies the original version by adding a data-defined weight, $\hat{\omega}_{jn}$, to the penalty term (Zou, 2006). For our purposes, we consider only $\hat{\omega}_{jn} = 1/|\hat{\beta}_j^{\text{OLS}}|$, where $\hat{\beta}_j^{\text{OLS}}$ is the ordinary least squares estimate. The adaptive Lasso minimizes

$$\hat{\beta}(\lambda_n) = \text{argmin} \left\{ \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2 + \sum_{j=1}^{p} \lambda_n \hat{\omega}_{jn} |\beta_j| \right\}. \tag{1.1}$$

When $\lambda_n \to \infty$ and $\lambda_n/\sqrt{n} \to 0$, the adaptive Lasso is an oracle procedure (Cai and Sun, 2007; Fan and Li, 2001). Let the true relationship be described by the linear equation $E(Y|\mathbf{X}) = \beta_1 X_1 + \cdots + \beta_p X_p$ where only a strict subset of the β-coefficients are non-zero, this subset being $A = \{j : \beta_j \neq 0\}$. An oracle procedure is defined by having the following two properties:

- Consistent variable selection: $\text{pr}\{\hat{A}(\lambda_n) = A\} \to 1$ where $\hat{A}(\lambda) = \{j : \hat{\beta}_j(\lambda) \neq 0\}$ is the estimated set of influential variables.
- Asymptotic efficiency for $\beta_A = \{\beta_j : j \in A\} : \sqrt{n}\{\hat{\beta}_A(\lambda_n) - \beta_A\} \to$ Normal$(0, \Sigma)$, where Σ is the inverse of the information matrix when A is known.

In practice, with finite sample sizes, a sequence, λ_n, that satisfies the oracle requirements results in a model that includes a large number of false positives (i.e. the set $\{j : \beta_j = 0, j \in \hat{A}_n\}$ is large) (Martinez *and others*, 2010). In this manuscript, our three objectives are the following: (1) To demonstrate, mathematically, that choosing λ_n to meet the oracle properties will result in a high false positive rate for finite samples. (2) To quantify the probability that a variable selected into the model is a false positive. This probability can provide confidence that the included variable is independently associated with the outcome. (3) To show how to identify a sequence of smoothing parameters that controls the number of false positives, instead of achieving the oracle properties.

In order to measure and control the number of false positives, we introduce the concept of the local false discovery rate (lFDR) into the selection of λ_n (Efron *and others*, 2001; Efron and Tibshirani, 2002; Benjamini and Hochberg, 1995). Specifically, we define lFDR(λ_n) to be the probability that a variable added to the model is a false positive when the penalty term is incrementally lowered below λ_n. Our first goal is to derive the relationship between lFDR and λ_n. We then show that lFDR$(\lambda_n) \to 1$, an unusual choice for most problems, if λ_n satisfies the oracle requirements, thus explaining the observation that the adaptive Lasso results in a large number of false positives when the effect sizes are not too large. In more traditional problems, a value of 0.05 is often the targeted FDR or lFDR. Finally, we offer a parametric bootstrap method for selecting λ_n to achieve a desired lFDR which is similar to a step described by Hall *and others* (2009). Others have also noted this high false positive rate and proposed Bootstrapped and Bayesian versions of the Lasso for handling this problem (Bach, 2008; Hans, 2010; Park and Casella, 2005).

Our motivating example comes from a Genome-Wide Association Study (GWAS). Both the Lasso (Wu *and others*, 2009) and the adaptive Lasso (Kooperberg *and others*, 2010; Sun *and others*, 2010) have

become popular tools for GWAS because variable selection is an important step given that 100 000's of single nucleotide polymorphisms (SNPs) are available for testing. In our specific study, we focus on modeling the prostate specific antigen (PSA) level, a biomarker indicative of prostate cancer (Parikh *and others*, 2010).

The order of this paper is as follows. In Section 2, we introduce notation and review the adaptive Lasso. We then formalize our definition of the lFDR, derive the relationship between the lFDR and λ_n, and provide asymptotic theory. Finally, we describe our bootstrap approach for choosing λ_n. In Section 3 and supplementary material available at *Biostatistics* online, we evaluate the behavior of λ_n when selected by the lFDR through simulation and our motivating example. We conclude with a short discussion in Section 4.

2. Methods

2.1 *Notation*

We assume that there is a continuous outcome Y_i and its true value is defined by

$$Y_i = X_{i1}\beta_1 + \cdots + X_{ip}\beta_p + \epsilon_i, \tag{2.1}$$

where $\epsilon_i = \text{Normal}(0, \sigma^2)$. Further, we assume $n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} \to D$, where D is a positive-definite matrix. Recall that A is the set of covariates that are associated with a non-zero β, $A \equiv \{j : \beta_j \neq 0\}$, and $\beta_A \equiv \{\beta_j : j \in A\}$. We say that covariate j is *influential* if $j \in A$ or that it is *superfluous* if $j \notin A$. Without loss of generality, assume that $A = \{1, \ldots, p_0\}$, let $z = 1 - p_0/p$, and let D_{00} be the corresponding $p_0 \times p_0$ submatrix of D.

Let $\hat{\beta}(\lambda_n)$ be the parameter estimates produced by the adaptive Lasso,

$$\hat{\beta}(\lambda_n) = \text{argmin} \left\{ \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2 + \sum_{j=1}^{p} \lambda_n \hat{\omega}_{jn} |\beta_j| \right\},$$

where, for our purposes, $\hat{w}_{jn} = 1/|\hat{\beta}_j^{\text{OLS}}|$. The sequence λ_n is the set of smoothing parameters. We let $\hat{A}(\lambda_n)$ be the set of covariates predicted to have a non-zero β, so $\hat{A}(\lambda_n) \equiv \{j : \hat{\beta}_j(\lambda_n) \neq 0\}$.

Finally, we include the notation and definitions for the local FDR and related terms. We denote the probabilities, $P_{fp}(\lambda_n)$ and $P_{fn}(\lambda_n)$, that a variable will be a false positive and a false negative by

$$P_{fp}(\lambda_n) = \frac{1}{zp} \sum_{j \notin A} \text{pr}\{\hat{\beta}_j(\lambda_n) \neq 0\} \quad \text{and} \quad P_{fn}(\lambda_n) = \frac{1}{(1-z)p} \sum_{j \in A} \text{pr}\{\hat{\beta}_j(\lambda_n) = 0\}.$$

We define the lFDR by

$$\text{lFDR}(\lambda_n) = \frac{z\Delta_{fp}(\lambda_n)}{z\Delta_{fp}(\lambda_n) - (1-z)\Delta_{fn}(\lambda_n)}, \tag{2.2}$$

where

$$\Delta_{fp}(\lambda_n) = \left.\frac{\mathrm{d}P_{fp}}{\mathrm{d}\lambda}\right|_{\lambda=\lambda_n} \quad \text{and} \quad \Delta_{fn}(\lambda_n) = \left.\frac{\mathrm{d}P_{fn}}{\mathrm{d}\lambda}\right|_{\lambda=\lambda_n}.$$

By a Taylor series expansion, the expected difference in the number of false positives, $(1-z)\{P_{fp}(\lambda) - P_{fp}(\lambda - \delta)\}$, at λ and $\lambda - \delta$, is approximately $(1-z)\delta\Delta_{fp}(\lambda)$. Similarly, the expected difference in the number of false negatives and the total number of variables included in the model are $z\delta\Delta_{fp}(\lambda)$ and

$z\delta\Delta_{fp}(\lambda) + (1-z)\delta\Delta_{fn}(\lambda)$. Therefore, we define the lFDR by (2.2) as the probability that a variable added to our model will be superfluous, if added when the smoothing parameter is lowered below λ_n. Our definition of lFDR differs from that traditionally given for two reasons: (i) we interpret the lFDR from a frequentist point of view and (ii) we focus on the smoothing parameter λ_n instead of on a test statistic. The traditional definitions of lFDR and FDR have also been used for purposes of variable selection, usually by including only those variables with a q-value below a given threshold (Storey, 2002). However, such an approach would not be as appropriate for Lasso procedures, which try to avoid this post hoc selection. Note, an equivalent definition for FDR is available by replacing $\Delta(\lambda_n)$ with $P(\lambda_n)$ in (2.2).

2.2 *Prior results*

The adaptive Lasso has many theoretical properties. Here, we build on two previous results. Zou (2006) states the requirements needed for the adaptive Lasso to have the oracle properties.

THEOREM 1 Suppose that

$$\lambda_n/\sqrt{n} \to 0 \quad \text{and} \quad \lambda_n \to \infty. \tag{2.3}$$

Then the adaptive Lasso estimates must satisfy the following:

$$\text{Consistency in variable selection: } \lim_n \text{pr}\{\hat{A}_n = A\} = 1, \tag{2.4}$$

$$\text{Asymptotic Normality: } \sqrt{n}\{\hat{\beta}_A(\lambda_n) - \beta_A\} \to \text{Normal}(0, \sigma^2 \times D_{00}^{-1}). \tag{2.5}$$

If our focus is on variable selection, then a theorem identified by Pötscher and Schneider (2009) proves equally useful.

THEOREM 2 Let $\mathbf{X}^{\mathrm{T}}\mathbf{X} = nI$, where I is the identity matrix. Then

$$\hat{\beta}_j(\lambda_n) = 0 \quad \text{if } |\hat{\beta}_j^{\text{OLS}}| \leqslant \sqrt{\lambda_n/(2n)},$$

$$\hat{\beta}_j(\lambda_n) = \hat{\beta}_j^{\text{OLS}}\left(1 - \frac{\lambda_n}{2n(\hat{\beta}_j^{\text{OLS}})^2}\right) \quad \text{if } |\hat{\beta}_j^{\text{OLS}}| > \sqrt{\lambda_n/(2n)}.$$

Because $\hat{\beta}_j^{\text{OLS}}$ is asymptotically normal with mean β_j, we immediately see

$$\lim_n \text{pr}\{\hat{\beta}_j(\lambda_n) = 0\} - \text{pr}\{\chi^2_{1,\gamma_j} \leqslant \lambda_n/(2\sigma^2)\} = 0, \tag{2.6}$$

where χ^2_{1,γ_j} follows a non-central χ^2 distribution with one degree of freedom and non-centrality parameter $\gamma_j = n\beta_j^2/\sigma^2$.

2.3 *Local false discovery rates*

When $\mathbf{X}$ is orthogonal, the total number of variables included in the model is monotonically non-decreasing as λ_n decreases. The lFDR is the proportion of added variables that are expected to be false positives. When $\mathbf{X}$ is orthogonal and $\sigma^2 = 1$, then

$$\text{lFDR}(\lambda_n) = \{1 + C(\lambda_n)\}^{-1}, \tag{2.7}$$

where $C(\lambda)$, which can be interpreted as the cost of removing a false positive, is

$$C(\lambda) = \frac{1-z}{z}\frac{\Delta_{fn}(\lambda)}{\Delta_{fp}(\lambda)} = \frac{1-z}{z^2 p}\sum_{j\in A}\frac{-f_{\chi^2_{1,\gamma_j}}(\lambda/2\sigma^2)}{f_{\chi^2_1}(\lambda/2\sigma^2)}$$
$$= \frac{1-z}{z^2 p}\sum_{j\in A}\frac{1}{2}\left[\exp\left(\sqrt{\frac{\lambda n\beta_j^2}{2}} - \frac{n\beta_j^2}{2}\right) + \exp\left(-\sqrt{\frac{\lambda n\beta_j^2}{2}} - \frac{n\beta_j^2}{2}\right)\right], \tag{2.8}$$

where $f_{\chi^2_{1,\gamma_j}}(\cdot)$ is the density for a χ^2 variable with non-centrality parameter $\gamma_j = n\beta_j^2/\sigma^2$.

Equations (2.7) and (2.8) allow us to choose λ_n to achieve a specific lFDR. For example, if, in addition to $\sigma^2 = 1$ and $\mathbf{X}$ being orthogonal, all $\beta_j = \beta$, then the lFDR will never exceed q if

$$\lambda_n = \frac{2}{n\beta^2}\left(\log\left(\frac{z}{1-z}\left[\frac{1-q}{q}\times\exp\left(\frac{n\beta^2}{2}\right) + \sqrt{\left(\frac{1-q}{q}\times\exp\left(\frac{n\beta^2}{2}\right)\right)^2 - \left(\frac{1-z}{z}\right)^2}\right]\right)\right)^2. \tag{2.9}$$

The sequence λ_n, when defined by (2.9), is independent of the number of variables p. Moreover, all properties discussed hold regardless of the size of β (e.g. β is constant or decreasing at a rate of $1/\sqrt{n}$). Therefore, although there is no λ_n that can attain the oracle property when β is decreasing at a rate of $1/\sqrt{n}$ (Pötscher and Schneider, 2009), the sequence defined by (2.9) would still attain the stated lFDR. As expected, we note that the lFDR decreases with increasing λ_n confirming that those variables added when λ_n is small are more likely to be false positives. We define λ_{qn} to satisfy $\text{lFDR}(\lambda_{qn}) = q$.

2.4 *Constant β*

The term $\exp(-\sqrt{\lambda n\beta_j^2/2} - n\beta_j^2/2)$ in (2.8) can be ignored when $n\beta_j^2$ is large. Specifically, when $\lambda_n n\beta_j^2 > 5.3\ \forall j$, the lFDR at a given value of λ_n can be approximated within 1% of its true value by

$$\text{lFDR}(\lambda_n) \approx \frac{1}{1 + ((1-z)/z^2 p)\sum_{j\in A}\frac{1}{2}[\exp(\sqrt{\lambda_n n\beta_j^2/2} - n\beta_j^2/2)]}. \tag{2.10}$$

Equation (2.10) shows more clearly that if we choose λ_n to achieve the oracle property (i.e. $\lambda_n/\sqrt{n} \to 0$), then we are choosing a λ_n that results in an lFDR $\to 1$. As an lFDR $= 1$ implies that all variables being added to the model are false positives, purposely choosing such a λ_n would seem counterintuitive. Therefore, even when λ_n can be chosen to achieve the oracle properties, it is unclear whether such a choice is desirable. An alternative approach would be to choose λ_n to ensure that lFDR $< q$. In the previous example, where $\sigma^2 = 1$, $\mathbf{X}$ is orthogonal, and $\beta_j = \beta$, we now see lFDR $< q$ if

$$\lambda_n = \frac{2}{n\beta^2}\left[\log\left(2\frac{z}{1-z}\frac{1-q}{q}\right) + \frac{n\beta^2}{2}\right]^2. \tag{2.11}$$

Purposely choosing a λ_n such that the lFDR $\to 0$ seems equally counterintuitive, limiting the reasonable choices for λ_n. If $\sigma^2 = 1$, $\mathbf{X}$ is orthogonal, and $\beta_j = \beta$, where β is a constant, we see that for the lFDR not to diverge to 0 or 1, $\lambda_n/n \to 0.5\beta^2$.

LEMMA 1 When $\beta_j = \beta\ \forall j$, β is constant, $\sigma^2 = 1$, X is orthogonal, and $t = 0.5\beta^2$, then

$$\frac{\lambda_n}{n} = t \Rightarrow \text{lFDR}(\lambda_n) \to \frac{2z}{1+z}, \tag{2.12}$$

$$\frac{\lambda_n}{n} \to t_1 > t \Rightarrow \text{lFDR}(\lambda_n) \to 0, \tag{2.13}$$

$$\frac{\lambda_n}{n} \to t_1 < t \Rightarrow \text{lFDR}(\lambda_n) \to 1, \tag{2.14}$$

where $0 \leqslant t_1 \leqslant \infty$.

If λ_n were chosen to achieve an lFDR strictly between 0 and 1, then only the first of the two oracle properties holds, $\lim_n \text{pr}(\hat{A}_n = A) = 1$ from (2.4). However, we claim that forgoing the second oracle property, in exchange for an lFDR between 0 and 1, is no loss. Although performing variable selection and fitting in a single step is convenient, it is unnecessary. Clearly, there is a two-step method that recovers the second oracle property. After using the adaptive Lasso with $\lambda_n/n \to 0.5\beta^2$ for variable selection, we can refit the model using OLS with only that subset of variables. This two-step procedure not only satisfies both oracle properties, but offers improved efficiency over the single-step procedure, reminding us that the oracle procedure is not an optimal procedure. Although an oracle procedure promises that $\hat{\beta}(\lambda_n) \to 0$ for all superfluous variables, it makes no claim as to the rate at which this occurs. Asymptotically, we can increase the rate at which $\text{pr}\{A^c \notin \hat{A}_n\} \to 1$ without decreasing the rate at which $\text{pr}\{A \in \hat{A}_n\} \to 1$. Returning to (2.6), this potential improvement is clear because, asymptotically, $\text{pr}\{\chi^2_{1,\gamma} \leqslant \lambda_n/(2\sigma^2)\}$, is unchanged by λ_n so long as $\lambda_n/n \to 0$.

2.5 *Empirical choice of* λ_n

In the idealized scenario, where $\mathbf{X}$ is orthogonal, $\beta_j = \beta\ \forall j \in A$, and both z and β are known, (2.9) can be used to choose a sequence λ_n to achieve a specified lFDR. If all values of $\{\beta_j : j \in A\}$ are not identical, then the solution to (2.8) would need to be obtained numerically. Although β and z are unknown, in practice, we could use an estimate of z and either an estimate of β or a lower bound for a biologically meaningful β. However, when (2.9) is evaluated with these estimates, the chosen λ_n tends to produce an lFDR above the desired value when $\mathbf{X}$ is not orthogonal. Therefore, we prefer a bootstrap approach similar to one of the steps discussed by Hall *and others* (2009). The algorithm is as follows. Let us first fit a simple model of Y on $\mathbf{X}$ to obtain estimates of β. In practice, as done in our simulations, we suggest identifying those β to have non-zero values by the adaptive Lasso with λ_{dn}, and then defining $\tilde{\beta}$ by the OLS estimates. Let us then denote the variance of the residuals from this model by $\tilde{\sigma}^2$. Next, set all components of $\tilde{\beta}$ below some threshold equal to zero. In practice, when $n > p$, we use $1/\sqrt{n}$ as this threshold. Then generate B sets of data, assuming the true model is $Y = X\tilde{\beta} + \tilde{\epsilon}$, where $\tilde{\epsilon} = \text{Normal}(0, \tilde{\sigma}^2)$. For each value of λ_n in a given set, we calculate the number of true, $N^b_{tp}(\lambda_n)$, and false, $N^b_{fp}(\lambda_n)$, positives added to the model between $\lambda_n - \delta$ and $\lambda_n + \delta$ where δ is an appropriately small number and the superscript b denotes the dataset. We can then estimate the lFDR for each λ_n by

$$\text{lFDR}_{\text{est}}(\lambda_n) = \frac{\sum_{b=1}^{B} N^b_{fp}(\lambda_n)}{\sum_{b=1}^{B} N^b_{fp}(\lambda_n) + \sum_{b=1}^{B} N^b_{tp}(\lambda_n)}, \tag{2.15}$$

and select the smoothing parameter that achieves a specified lFDR, q:

$$\lambda_{qn} = \text{argmin}\{|\text{lFDR}_{\text{est}}(\lambda_n) - q|\}.$$

For completeness, we define $\text{lFDR}_{\text{est}}(\infty) = 0$ and $\text{lFDR}_{\text{est}}(\lambda_n) = \max\{\text{lFDR}_{\text{est}}(\lambda) : \lambda > \lambda_n\}$ when $\sum_{b=1}^{B} N_{fp}^b + \sum_{b=1}^{B} N_{tp}^b = 0$. In practice, $B = 10$, but we base our estimates of lFDR_{est} on a monotonically smoothed version of $\text{lFDR}_{\text{est}}(\lambda_n)$.

For purposes of comparison, we consider the standard method for selecting λ_n to be cross-validation aimed at minimizing the prediction error of future estimates. Recall that standard 10-fold cross-validation starts by dividing the set S_n of n subjects into 10 mutually exclusive sets, $s_1 \cup s_2 \cup \cdots \cup s_{10} = S_n$, of roughly equal size. Let $\hat{\beta}_{jk}$, $1 \leqslant k \leqslant 10$ be the adaptive Lasso estimate for β_j based on those subjects not in s_k. Then

$$\hat{\lambda}_{dn} = \text{argmin} \left\{ \sum_k \sum_{i \in s_k} \left(Y_i - \sum_j X_{ij} \hat{\beta}_{jk}(\lambda) \right)^2 \right\}.$$

Also, $\hat{\lambda}_{dn}$ is an estimate of the deviance-optimized smoothing parameters:

$$\lambda_{dn} = \text{argmin} \left\{ E_{\mathbf{T}} \left[\left(Y_0 - \sum_j X_{0j} \hat{\beta}_j(\lambda) \right)^2 \right] \right\},$$

where $T = \{\overrightarrow{X}_1, Y_1, \ldots, \overrightarrow{X}_n, Y_n\}$ are the data input into the adaptive Lasso to obtain the estimates $\hat{\beta}$, $T_0 = \{\overrightarrow{X}_0, Y_0\}$ are the data from a new individual, and $\mathbf{T} = \{T, T_0\}$. When β is fixed and X is orthogonal, the smoothing parameters minimizing the deviance must satisfy the oracle properties.

2.6 *High-dimensional adaptive Lasso: $p > n$*

As defined in (1.1), the weights in the adaptive Lasso are $1/\hat{\beta}^{\text{OLS}}$. However, when $p > n$, the weights must substitute a different estimate of β in place of $\hat{\beta}^{\text{OLS}}$. Two possible substitutes that have been studied include $\hat{\beta}^{\text{sep}}$, the estimates obtained by fitting separate models for each variable (Huang *and others*, 2008), and $\hat{\beta}^{L}$, the estimates from a regular Lasso procedure (Zhou *and others*, 2009). The properties of the latter estimates, $\hat{\beta}^{L}$, with $\lambda_n = \sqrt{24 \log(p)/n}$, have been studied and demonstrated to have useful qualities (Zhou *and others*, 2009). In practice, however, we found that $1/\hat{\beta}^{\text{sep}}$ performed better, and chose to use those weights in our simulations. For defining $\tilde{\beta}$, we cannot use $1/\sqrt{n}$ as our cutoff threshold. Instead, we first perform the adaptive Lasso on our data and count the number of coefficients estimated to be non-zero. We then find the threshold, such that by setting all $\tilde{\beta}$ below that threshold to 0 and simulating data, the adaptive Lasso on the simulated data estimates a similar number of non-zero coefficients.

3. Results

3.1 *Simulation design: comparing λ_{qn} and λ_{dn}*

Our first goal is to offer an example comparing the magnitude and performance of λ_{dn} and λ_{qn}. As with all simulations here, our objective is not to describe the performance of the estimates $\hat{\lambda}_{dn}$ and $\hat{\lambda}_{qn}$, but to calculate, describe, and compare the true values of λ_{dn} and λ_{qn}. We assume that the covariate matrix $\mathbf{X}$ is orthogonal and that the outcome Y can be described by linear regression, (2.1), with $\beta_j = 0.15$ if $\beta_j \in A$ and $\sigma^2 = 1$. For these examples, we fixed the number of covariates $p = 50$, but let the size of A vary, $z \in \{0.5, 0.7, 0.9\}$. As described below, we used simulation to calculate λ_{dn} and λ_{qn}, their corresponding lFDR and the proportion of variables that were misclassified, err_{MC}, for a sequence of samples between $n = 200$ and $n = 2000$.

Our second goal is to show that results are essentially unchanged when we vary p. For efficiency, we calculated λ_{dn}, λ_{qn}, lFDR, and err_{MC} at only $n = 1000$ for $p \in \{100, 200, 500\}$, maintaining all of the other assumptions.

Our third goal is to examine whether λ_{qn}, calculated assuming that $\mathbf{X}$ is orthogonal, was appropriate when there was dependence. Specifically, we repeated the abbreviated analyses assuming that the covariance structure of $(X_{i1}, X_{i2}, \ldots, X_{ip})$ is block diagonal. Correlation ρ within a block was constant, $\rho \in \{0.3, 0.6\}$, each block contained the same number of influential variables (or possibly no influential variables if there were more blocks than influential variables), and each block contained the same number of total variables. Variables were divided into 2, 5, or 10 groups.

For any combination of n, p, z, and covariance structure, we estimate the values of λ_{dn}, λ_{qn}, lFDR, and err_{MC} by simulating 200 000 values of $\mathbf{X}$ and Y. For each simulation, at a specified set of λ ranging from 0.01 to 100, we calculate the residual deviance and err_{MC}. Furthermore, for the same set of λ, we count the number of true and false positives added to the model when the smoothing parameter was between $\max(0, \lambda - 0.5)$ and $\min(\lambda + 0.5, 100)$. Then, for each λ, we average the number of true and false positives added, deviance, and err_{MC} over all 200 000 datasets to obtain estimates of each desired value. The lFDR was estimated by the ratio between the average number of false positives added, compared with the total number of variables added to the model. For each combination of n, p, z, and covariance structure discussed, we simulated a new set of 200 000 simulations. To generate a dataset $\mathbf{X}$, we assumed that $\{X_{i1}, \ldots, X_{ip}\}$ followed a normal distribution with mean 0 and specified covariance matrix. When $\mathbf{X}$ was assumed orthogonal, we used the resulting principal components. All datasets were standardized, so each variable had mean 0 and variance 1.

3.2 *Simulation results: comparing λ_{qn} and λ_{dn}*

First, consider the example when $\mathbf{X}$ is orthogonal, $p = 50$, and $z = 0.7$. Figure 1(a) shows that λ_{qn} increases linearly with the number of subjects and that the slope is approximately $0.5\beta^2$ as (2.11) suggests, except when $\lambda_n n\beta^2$ is small. The same equation promises that the lFDR-selected λ_n's, for values q_1 and q_2, differ by approximately $2[\log((1 - q_1)/q_1) - \log((1 - q_2)/q_2)]$. As the deviance-optimized λ_n's achieve the oracle properties when X is orthogonal, they must be increasing at a rate less than $\sqrt{n}$, and therefore, the representative black line in Figure 1(a) is significantly below those illustrating the smoothing parameters chosen to achieve the specified values of the lFDR.

The advantage to choosing a sequence λ_n that increases linearly with the number of the subjects is that the proportion of misclassified variables converges to 0 much quicker. Figure 1(b) shows that when there are 1000 individuals in the study and 35 out of 50 of the SNPs are superfluous, on average, 12% of the variables are misclassified with the deviance-optimized parameters, whereas less than 2.1% are misclassified when using lFDR-selected parameters. The relationship between lFDR and percentage misclassified is not monotone, as it depends on z. Here, setting $q = 0.5$ minimized the proportion misclassified. Figure 2 shows that when λ_n minimizes deviance, the cost of reducing false positives is very low, or equivalently, the lFDR is high, so there is great benefit in increasing λ_n. In terms of identifying A exactly, with 1000 individuals, the probability that there is at least one misclassified variable, $\text{pr}\{\hat{A}_n \neq A\}$, exceeds 0.999 when using deviance-optimized smoothing parameters, whereas that probability is less than 0.64 when using lFDR-selected parameters.

Table 1 shows that the large difference between λ_{dn} and λ_{qn} remains for $p > 50$, and, in fact, both λ_{dn} and λ_{qn} appear to be essentially independent of p when $\mathbf{X}$ is orthogonal. When the covariates are correlated, compared with when they are independent, λ_{qn} tends to be larger, as more stringent penalty terms are needed to exclude null variables that are correlated with influential variables. Increasing ρ or block size magnifies this effect. Therefore, in practice, we suggest choosing λ_{qn} by the bootstrapping

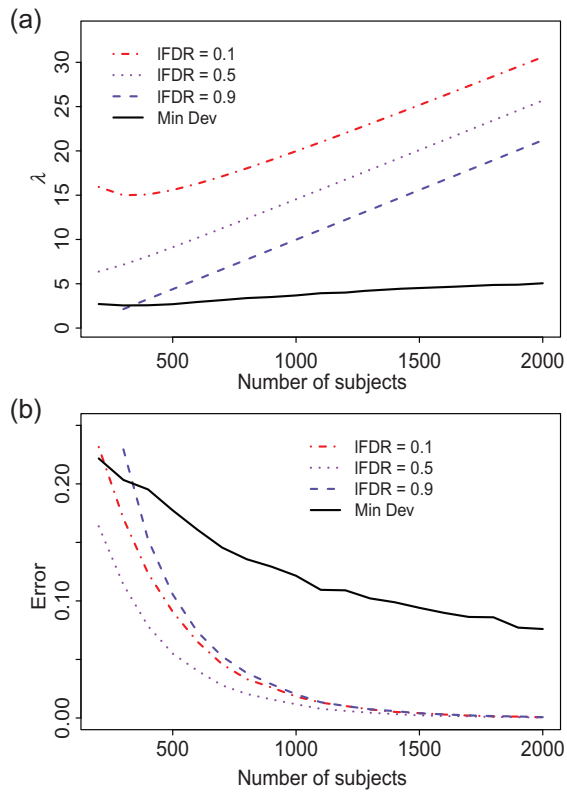

Fig. 1. (a) The sequence of λ_n chosen to minimize the deviance (solid black line) or chosen to achieve a specified lFDR (broken lines) increases with the number of subjects in the study. For these simulations, $z = 0.7$, $p = 50$, $\beta = 0.15$ for associated variables. (b) The average proportion of variables that are misclassified (error, y-axis), or the number of false positives and false negatives, quickly drops to 0 when λ_n is chosen to achieve a specified lFDR, but remains above 0 for deviance-optimized smoothing parameters.

method described in Section 2.5. Table 1 also demonstrates the obvious result that as the proportion z of null variables increases, λ_{qn} must also increase.

3.3 *Simulation design: evaluating the performance of adaptive Lasso with* λ_{qn}

Our next goal is to evaluate the performance of the adaptive Lasso when using λ_{qn}, estimated by our bootstrap approach described in Sections 2.5 and 2.6. This method selects a set of variables that should satisfy the specified lFDR criteria. For comparison, we consider a more traditional method for selecting variables targeting the same criteria. This method, implemented by the R function FDRtool (Strimmer, 2008), inputs the p-values calculated from models including each variable individually. In brief, the method decomposes the overall distribution of p-values into two distributions, representing the p-values from the null and influential variables. Given these two distributions, the traditional method first estimates the p-value thresholds that would result in the specified lFDR and then selects all variables meeting the appropriate threshold.

We consider the lFDR, lFDR_{AL}, resulting from using the bootstrap version of the adaptive Lasso and the rates lFDR_{TR} resulting from the traditional method. We compare these observed rates to the targeted

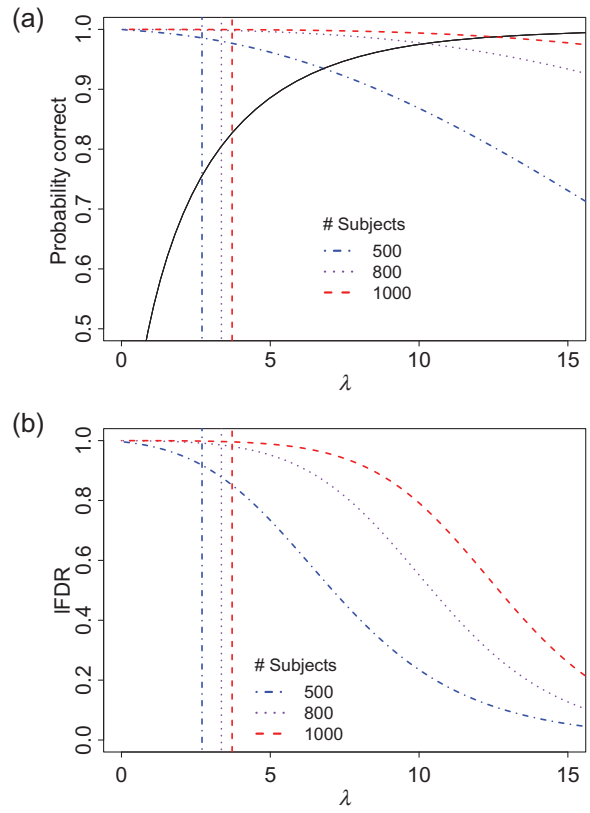

Fig. 2. (a) The solid black line shows the probability, $\text{pr}(j \notin \hat{A}_n | \beta_j = 0)$, that a null variable is excluded from $\hat{A}_n$ when $\mathbf{X}$ is orthogonal. The top curved dashed line (red) shows the probability, $\text{pr}(j \in \hat{A}_n | \beta_j = 0.15)$, that a non-null variable is included in $\hat{A}_n$ when $\mathbf{X}$ is orthogonal, $z = 0.7$, and $n = 1000$. The vertical dashed line (red) farthest to the right indicates λ_{dn}. The other pairs of broken lines show the equivalent values when $n = 500$ and $n = 800$. (b) The local FDR, lFDR, is illustrated as a function of λ for the three scenarios above.

values: $q \in \{0.1, 0.5, 0.9\}$. These comparisons are performed in two types of datasets. When $n > p$, settings are similar to those in Section 3.1: $n = 1000$, $p = 500$, $z = 0.9$, $\beta_j = 0.1$ if $\beta_j \in A$ and $\sigma^2 = 1$. In order for the traditional methods to produce rates below 1, we reduce the number of correlated variables per block to 5. Again $\rho \in \{0.0, 0.3, 0.6\}$. When $p > n$, specifically $n = 1000$ and $p = 5000$, we increase z to 0.96 and include 10 variables per correlated block. To achieve $q = 0.1$ when $p > n$, we further increase z to 0.99 and β_j to 0.35. We provide an extended set of simulations, exploring other correlation structures and effect distributions, in supplementary material available at *Biostatistics* online.

For each combination of parameters, we generated 1000 datasets and then averaged the resulting lFDR_{AL} and lFDR_{TR} across all 1000 datasets. For each dataset, we defined the lFDR to be 0 if the last variable selected was influential, 1 otherwise.

3.4 *Simulation results: evaluating the performance of adaptive Lasso with* λ_{qn}

The bootstrap approach proposed in Sections 2.5 and 2.6 selected values of λ_{qn} that, when applied to the full dataset, resulted in lFDR values similar to the targeted value. In the example where $n > p$ and $\rho = 0$,

Table 1. *The smoothing parameters designed to achieve* lFDR = 0.5 *are larger than those designed to minimize deviance*

				Low correlation			High correlation		
		Independent		10	5	2	10	5	2
p	z	λ_{dn}	λ_{qn}		λ_{qn}			λ_{qn}	
100	0.9	6.31	18.52	18.72	20.32	27.33	22.32	27.63	30.23
100	0.7	3.7	15.02	16.22	18.32	18.32	14.92	17.12	15.22
200	0.9	6.21	17.92	20.72	23.63	28.33	26.23	28.13	39.44
200	0.7	3.7	14.82	17.02	18.72	19.82	15.02	14.92	15.42
500	0.9	6.21	18.12	28.73	34.84	45.15	38.44	50.56	51.66
500	0.7	3.7	14.22	18.62	18.92	19.52	15.42	15.42	16.02

The first two columns list the number of variables p and the proportion z that are unassociated with the outcome for each scenario. The next two columns list the value of the smoothing parameter that minimizes deviance λ_{dn} and the smoothing parameter λ_{qn} that achieves an lFDR = 0.5. The remaining columns show how λ_{qn} increases with the extent of correlation among the variables. In each case all covariates are divided into 2, 5, or 10 groups, and the correlation is set to either 0.3 (low correlation) or 0.6 (high correlation). The total number of subjects was fixed at 1000.

Table 2. *A comparison between the newly proposed bootstrap* (B) *method for obtaining a specified lFDR with the traditional* (T) *approach*

	$\rho = 0.0$		$\rho = 0.3$		$\rho = 0.6$	
Target	B	T	B	T	B	T
$n > p$						
0.1	0.059	0.088	0.083	0.232	0.101	0.688
0.5	0.477	0.435	0.505	0.737	0.52	0.919
0.9	0.893	0.804	0.905	0.94	0.909	0.967
$p > n$						
0.1	0.008	0.075	0.019	0.298	0.105	0.904
0.5	0.466	0.464	0.592	0.679	0.639	0.905
0.9	0.751	0.864	0.861	0.953	0.894	0.978

The table entries list the observed, or true, lFDR for different targeted values (rows) and for different correlation structures (columns). Covariates were divided into either 100 independent blocks ($n > p$) or 500 independent blocks ($p > n$), with constant correlation of 0.0, 0.3, and 0.6. The top set of rows show results when $n = 1000$ and $p = 500$, whereas the bottom set of rows show results when $n = 1000$ and $p = 5000$. When non-zero, $\beta_j = 0.10$.

the observed lFDR was 0.06, 0.48, and 0.89 when λ_{qn} was chosen to achieve lFDR = 0.1, 0.5, and 0.9. When targeting lFDR = 0.1, our method achieved a lower lFDR, and therefore our chosen λ_q was larger than desired. This inflated λ_q arises, in part, from a tendency to select too few non-zero $\tilde{\beta}$ in our bootstrap models. Table 2 and results in supplementary material available at *Biostatistics* online show that the lFDR estimates were only minimally altered by changing the correlation structure or when considering the $p > n$.

The traditional approach, based on estimating the p-value distribution of the null and influential variables, performed poorly when there was high correlation between variables (Table 2). When there was high correlation, models with only a single variable assigned low p-values to those null variables associated with influential variables. This resulted in more variables achieving the lFDR threshold, but a higher proportion were false positives. With $n > p$, $\rho = 0.6$, and a targeted lFDR = 0.5, the observed lFDR = 0.9.

3.5 *Application*

In the United States, prostate cancer is the most commonly diagnosed non-cutaneous cancer in men, with approximately 200 000 new diagnoses each year. Because levels of PSA are elevated in the presence of prostate cancer, it is commonly used as a biomarker for early detection. Unfortunately, the specificity of tests based on PSA is often very low, as many healthy individuals also have high levels. Specificity could be greatly improved by a method that can identify individuals with naturally high levels. To this end, there have been large GWASs searching for genetic markers associated with the PSA level. The Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial, or PLCO, which recorded PSA levels, genotyped 2200 healthy men using an Illumina genotyping platform containing more than 500 000 SNPs (Andriole *and others*, 2009). We focus on a subset of 530 SNPs in and around the *KLK3* gene (Parikh *and others*, 2010).

Let $\mathbf{X}$ be the 2200×530 matrix containing the genotypes for the study population at these 530 SNPs. Genotypes are coded as 0, 1, or 2, indicating the number of minor alleles at that SNP. Let Y be the log-transformed PSA levels. Then we regressed Y on $\mathbf{X}$ using a linear model with the adaptive Lasso procedure. We repeated this analysis with the two values of λ: $\hat{\lambda}_d$ and $\hat{\lambda}_q$, where $\hat{\lambda}_q$ was estimated by the previously defined bootstrap procedure. Unfortunately, the truth is unknown, and therefore we can only use this example to illustrate their relative performance. The estimated values of the smoothing parameter were $\hat{\lambda}_d = 14.2$ and $\hat{\lambda}_q = 58.0$, with $q = 0.5$. As expected, $\hat{\lambda}_d$ is significantly smaller than that value estimated to achieve $\text{lFDR} = 0.5$. As a consequence, $\hat{\lambda}_d$ allowed 17 SNPs to have non-zero coefficients, whereas $\hat{\lambda}_q$ allowed only 1 SNP (Table S2 of supplementary material available at *Biostatistics* online). Although we cannot be certain that either model is correct, it seems doubtful that 17 SNPs in that region are directly associated with PSA levels. To estimate β corresponding to rs2735839 using the two-stage approach, first selecting variables with $\hat{\lambda}_q$ and then estimating β using OLS, we calculate $\hat{\beta} = 0.21$ from a model containing only rs2735839.

4. Discussion

The adaptive Lasso has become a popular model-building procedure because it shrinks a subset of coefficients to zero, thereby simultaneously performing variable selection and simplifying model interpretation. Although, asymptotically, using the traditional smoothing parameters promises that the adaptive Lasso will achieve consistent variable selection, their use often leaves a large number of false positives in the model when sample size is finite.

The lFDR is usually a form of post-processing, in that we would first perform a statistical procedure to attach a p-value to an estimate of each parameter and then determine the probability that the true value of a parameter with that p-value is the null value. We have adapted the lFDR framework to select smoothing parameters in the adaptive Lasso. Instead of defining an lFDR for a specific p-value, we define it for a specific value of the tuning constant λ. The framework offers an alternative means for selecting the smoothing parameters. When chosen to achieve a specified value of the lFDR, the adaptive Lasso procedure promises both asymptotically consistent variable selection and better control of the false positive rate for finite samples.

By itself, a single-step, adaptive Lasso procedure using λ_q, the lFDR-selected smoothing parameter, does not achieve the oracle properties. If one believed that the optimal, or best, estimator had to have these properties, then a combined variable selection and model fitting procedure with λ_q would not be a viable option. However, we do not consider the absence of the second oracle property to be a deterrent to using λ_q. First, the oracle properties can be regained by a two-step procedure that adds a separate model fitting step, where OLS is applied only to those variables retained by the initial adaptive Lasso. Although the convenience of a one-step procedure is sacrificed, the final estimate would still have the stated properties. Second, the first oracle property, consistent variable selection, is not a statement of optimality. That

property makes no claims on the rate at which $\text{pr}\{\hat{A}_n = A\} \to 1$. In some sense, the rate of our two-step procedure is faster than the rate of the single-step procedure. Therefore, there is a benefit to our method, even if it cannot be measured by a characteristic as coarse as the oracle properties.

We chose to select the smoothing parameters to achieve a desired lFDR, instead of FDR, because we wanted to judge each variable on its own merits, and not the merits of all selected variables. As discussed previously (Efron *and others*, 2001; Efron and Tibshirani, 2002), if one fit a model with 1000 variables and aimed to achieve an FDR of 0.1, then if the first 90 variables selected were guaranteed to be non-null, the next 10 would be included regardless of the evidence. Note also that in addition to providing examples when $\text{lFDR} = 0.1$, we offered examples with an lFDR as high as 0.9, a larger value than that generally used. For the adaptive Lasso procedure, where standard practice has been to choose $\text{lFDR} = 1$ and there is often the desire not to omit any non-null variables, aiming for larger lFDR values may be preferred for the Lasso procedure.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the NIH, Bethesda, Md. (http://biowulf.nih.gov). *Conflict of Interest*: None declared.

Funding

Sampson's and Chatterjee's research was supported by the Intramural Research Program of the NCI. Chatterjee's research was supported by a gene-environment initiative grant from the NHLBI (RO1-HL091172-01). Müller's research was supported by a grant from the Australian Research Council (DP110101998). Carroll's research was supported by a grant from the National Cancer Institute (R37-CA057030). Carroll was also supported by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST).

References

Andriole, G. L., Grubb, R. L., Buys, S. S., Chia, D., Church, T. R., Fouad, M. N., Gelmann, E. P., Kvale, P. A., Reding, D. J., Weissfeld, J. L. *and others*. (2009). Mortality results from a randomized prostate-cancer screening trial. *New England Journal of Medicine* **360**, 1310–1319.

Bach, F. R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML)*, Helsinki, Finland.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57**, 289–300.

Cai, T. and Sun, W. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association* **102**, 901–912.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* **23**, 70–86.

EFRON, B., TIBSHIRANI, R., STOREY, J. D. AND TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151–1160.

FAN, J. AND LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

HALL, P., LEE, E. R. AND PARK, B. (2009). Bootstrap-based penalty choice for the lasso achieving oracle performance. *Statistica Sinica* **19**, 449–471.

HANS, C. (2010). Model uncertainty and variable selection in Bayesian Lasso regression. *Statistics and Computing* **20**, 221–229.

HUANG, J., MA, S. AND ZHANG, C.-H. (2008). Adaptive lasso for sparse high dimensional regression models. *Statistica Sinica* **18**, 1603–1618.

KOOPERBERG, C., LEBLANC, M. AND OBENCHAIN, V. (2010). Risk prediction using genome-wide association studies. *Genetic Epidemiology* **34**, 643–652.

MARTINEZ, J. G., CARROLL, R. J., MULLER, S., SAMPSON, J. N. AND CHATTERJEE, N. (2010). A note on the effect on power of score tests via dimension reduction by penalized regression under the null. *The International Journal of Biostatistics* **6**, Article 12.

PARIKH, H., DENG, Z., YEAGER, M., BOLAND, J., MATTHEWS, C., JIA, J., COLLINS, I., WHITE, A., BURDETT, L., HUTCHINSON, A. *and others.* (2010). A comprehensive resequence analysis of the KLK15-KLK3-KLK2 locus on chromosome 19q13.33. *Human Genetics* **127**, 91–99.

PARK, T. AND CASELLA, G. (2005). The Bayesian Lasso. *Technical Report*.

PÖTSCHER, B. M. AND SCHNEIDER, U. (2009). On the distribution of the adaptive lasso estimator. *Journal of Statistical Planning and Inference* **139**, 2775–2790.

STOREY, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64**, 479–498.

STRIMMER, K. (2008). fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics* **24**, 1461–1462.

SUN, W., IBRAHIM, J. G. AND ZOU, F. (2010). Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. *Genetics* **185**, 349–359.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B* **58**, 267–288.

TIBSHIRANI, R. (2011). Regression shrinkage and selection via the Lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 273–282.

WU, T. T., CHEN, Y. F., HASTIE, T., SOBEL, E. AND LANGE, K. (2009). Genome-wide association analysis by Lasso penalized logistic regression. *Bioinformatics* **25**, 714–721.

ZHOU, S., VAN DE GEER, S. AND BUHLMANN, P. (2009). Adaptive lasso for high dimensional regression and gaussian graphical modeling. ArXiv:0903.2515.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

[*Received December 18, 2011; revised February 7, 2013; accepted for publication February 19, 2013*]