Iterative Selection Using Orthogonal Regression Techniques

Bradley Turnbull[1], Subhashis Ghosal[1]* and Hao Helen Zhang[2]

[1]*Department of Statistics, North Carolina State University, Raleigh, NC, USA*

[2]*Department of Mathematics, University of Arizona, Tucson, AZ*

Received 15 October 2013; revised 1 November 2013; accepted 2 November 2013
DOI:10.1002/sam.11212
Published online in Wiley Online Library (wileyonlinelibrary.com).

Abstract: High dimensional data are nowadays encountered in various branches of science. Variable selection techniques play a key role in analyzing high dimensional data. Generally two approaches for variable selection in the high dimensional data setting are considered—forward selection methods and penalization methods. In the former, variables are introduced in the model one at a time depending on their ability to explain variation and the procedure is terminated at some stage following some stopping rule. In penalization techniques such as the least absolute selection and shrinkage operator (LASSO), as optimization procedure is carried out with an added carefully chosen penalty function, so that the solutions have a sparse structure. Recently, the idea of penalized forward selection has been introduced. The motivation comes from the fact that the penalization techniques like the LASSO give rise to closed form expressions when used in one dimension, just like the least square estimator. Hence one can repeat such a procedure in a forward selection setting until it converges. The resulting procedure selects sparser models than comparable methods without compromising on predictive power. However, when the regressor is high dimensional, it is typical that many predictors are highly correlated. We show that in such situations, it is possible to improve stability and computational efficiency of the procedure further by introducing an orthogonalization step. At each selection step, variables potentially available to be selected in the model are screened on the basis of their correlation with variables already in the model, thus preventing unnecessary duplication. The new strategy, called the Selection Technique in Orthogonalized Regression Models (STORM), turns out to be extremely successful in reducing the model dimension further and also leads to improved predicting power. We also consider an aggressive version of the STORM, where a potential predictor will be permanently removed from further consideration if its regression coefficient is estimated as zero at any stage. We shall carry out a detailed simulation study to compare the newly proposed method with existing ones and analyze a real dataset. Statistical Analysis and Data Mining 6: 557–564, 2013

Keywords: forward selection; orthogonalization; high dimensional regression; LASSO

1. INTRODUCTION

In modern applications of statistics and data mining, linear regression models with extremely high dimensional regressors are commonly encountered. Typically the dimension of the regressor variable far exceeds the available sample size, posing serious challenges in the analysis of such data. In particular, the data matrix becomes singular and the least squares estimator is not uniquely defined. Usually, the majority of the regressor variables are not relevant, leading to a sparse structure in the model. However, it is not known beforehand which variables are actually relevant for the response variable. This problem is addressed through a variable selection step, which screens the variables before they can enter in the model. The variable selection step actually allows a fairly accurate estimation of the regression function in such high dimensional low sample size (HDLSS) situations. Variable selection has many other benefits, such as the ability to work with a sparse model, which has much better interpretability compared with a regression model having a lot of predictors.

Variable selection methods mainly fall in two categories. The first one is a recursive selection method such as forward selection, backward selection, and stepwise selection. In a forward selection procedure, variables are added one by one to build up the model, and a stopping rule is used, based on some criterion such as the mean squared error (MSE), adjusted R^2, Mallow's C_p-metric, prediction sum of squares (PRESS), Akaike information criterion (AIC), or the Bayesian information criterion (BIC); see ref [1] for their definitions. This strategy often runs into problems in HDLSS situations. For high dimensional data, Bühlmann [2] and Wang [3] studied forward regression

* *Correspondence to:* Subhashis Ghosal
(sghosal@stat.ncsu.edu)

for variable screening. Backward selection is a variable removal procedure, which starts with the full model and sequentially deletes redundant variables. Since estimation in the full model is not feasible in HDLSS situations if using the least squares method, the backward selection method is typically not favored. A stepwise selection allows both addition and removal of variables in each step depending on predictive performance. The second category of variable selection methods uses the idea of penalization to regularize a basic estimator such as the least squares estimator. Ridge regression [4] is among the first in the linear regression setting, where a quadratic penalty term shrinks the coefficient estimates towards zero, thus reducing variability in the estimated coefficients at the expense of introducing a small bias. The ridge regression is unable to do any variable selection unless aided with a hard thresholding step to filter out coefficients that are too small. More recently, it has been found that some other penalty terms with edged contours can do both shrinkage and selection at the same time. Examples of such procedures include the nonnegative garrote [5], the least absolute selection and shrinkage operator (LASSO) [6], the SCAD penalized estimator [7], the adaptive LASSO [8], and the elastic net [9]. A common feature of the associated penalty functions is that they are not differentiable at zero, thus allowing the minimizer of the penalized sum of squares to have many components exactly zero, inducing sparsity in the estimator. In the literature, there are also numerous Bayesian variable selection methods for high dimensional linear regression, such as refs [10–12].

Recently, Hwang et al. [13] introduced a method that combines the idea of both forward selection and penalization. It is known that, although the typical penalized regression estimators do not have closed forms in general, many of them have such expressions if the regressor were one-dimensional. Hence it is possible to use such an expression in the recursion of a forward selection procedure. Hwang et al. [13] proposed the Forward Iterative Regression and Selection Technique (FIRST). The essential difference with a traditional forward selection procedure is that the additional shrinkage penalty term, at each selection step, filters out many redundant variables before they can enter the model, thus keeping the effective model size low, and the resulting model easily interpretable. This approach can be used with any standard penalization method, such as the LASSO, the adaptive LASSO, the nonnegative garrote, the naive elastic net etc. Moreover, a post selection ordinary least squares method often gives a better prediction result, by reducing the finite sample estimation bias caused by the shrinkage. Through an extensive simulation study, they showed that the FIRST procedure leads to much sparser models compared with existing variable selection methods, without compromising the predictive power of the FIRST.

Statistical Analysis and Data Mining DOI:10.1002/sam

Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ be a standard linear regression model, where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}} \in \mathbb{R}^n$ is a vector of responses, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^p$ is a vector of parameters, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}}$ is an n-dimensional random error vector with uncorrelated mean zero and equal variance components. The intercept may be excluded from the regression model since the data can be centered. In what follows, we standardize the columns of X to have length 1. Note that the ith observation may be decomposed as $Y_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i, i = 1, \ldots, n$, where $\boldsymbol{X}_1^{\mathrm{T}}, \ldots, \boldsymbol{X}_n^{\mathrm{T}}$ are the rows of $\boldsymbol{X}$. The most common variable selection method LASSO minimizes $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1$, where $\|\cdot\|_1$ is the L_1-norm. The resulting estimator does not have a closed form for $p > 1$, and is usually obtained from the *least angle regression* (LARS) algorithm [14]. However, for $p = 1$, the expression of the LASSO is explicit, namely, with $\hat{\beta}_j = \boldsymbol{X}_j^T\boldsymbol{Y}$ standing for the ordinary least squares estimator for linear regression of Y on X_j,

$$\hat{\beta}_j^L = \begin{cases} \hat{\beta}_j - \frac{\lambda}{2}, & \text{if } \hat{\beta}_j \geq \lambda/2, \\ 0, & \text{if } |\hat{\beta}_j| < \lambda/2, \\ \hat{\beta}_j + \frac{\lambda}{2}, & \text{if } \hat{\beta}_j \leq -\lambda/2. \end{cases} \tag{1}$$

Although $p = 1$ is only a hypothetical situation, the explicit formula of the LASSO in this case allows us to repeat it sequentially as in a forward selection procedure, resulting in the estimator FIRST. Indeed, the same idea can be applied on all modifications of the LASSO penalty mentioned above.

When there is a relatively high degree of correlation, it seems that considering a forward selection procedure based on all variables is somewhat redundant, since a variable which is nearly a linear combination of variables already selected in the model has very little to add in the model. It therefore seems sensible to orthogonalize variables before a forward selection procedure is performed, and eliminate those which are nearly linear combinations of variables already selected. Apart from making the chosen model more sparse, such an orthogonalization step may, in principle, enhance prediction performance by reducing the complexity and variance of the final estimator. The resulting procedure will be called the Selection Technique in Orthogonalized Regression Models (STORM). An aggressive version of the procedure is obtained by permanently removing any potential predictor from further consideration if its regression coefficient is estimated as zero at any stage. Since it is hard to keep track of the linear relations resulting from the orthogonalization step, we only keep track of the selected variables and finally estimate the coefficients by a post selection least squares method. Since the selection step makes the selected model low dimensional even in HDLSS situations, the least squares estimator will be unique. We conduct a comprehensive simulation study to assess the

comparative performances of the STORM, the FIRST and other variable selection techniques in linear regression models. We find that, in all the simulating settings, the new proposal, the STORM, outperforms the FIRST and other variable selection techniques available in the literature both in terms of finding the sparsest model and in terms of estimation error.

The paper is organized as follows. In Section 2, we describe the procedure STORM and give a step-by-step algorithm. In Section 3, we conduct a comprehensive simulation experiment to assess the relative performance of the STORM and other methods. In Section 4, we analyze a high dimensional real dataset.

2. METHODOLOGY

In this section, we define the proposed Selection Technique in Orthogonalized Regression Models (STORM). Consider the linear model $\boldsymbol{Y}_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$, $i = 1, \ldots, n$, with p-dimensional regressor variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Without loss of generality, we center the responses Y_i for $i = 1, \ldots n$, such that $\sum_{i=1}^n Y_i = 0$, and standardize all predictor variables $\{X_1, \ldots, X_p\}$, i.e., $\sum_{i=1}^n X_{ij} = 0$ and $\sum_{i=1}^n X_{ij}^2 = 1$.

2.1. Algorithm

For classical forward selection approaches, predictors are added in the model sequentially, one at a time. At each step, the variable whose inclusion in the model improves the current prediction performance the most then enters the model. The STORM procedure also builds the model by sequentially including one variable at a time but in a different fashion. At each step, the one-dimensional penalization problem is solved for each candidate variable and then the variable which produces the smallest residual sum of squares (RSS) is selected. To handle possible high correlation among the predictors, the STORM orthogonalizes the prediction variables at each step before performing the selection. Let $\mathcal{J}_0 = \{1, \ldots, p\}$ denote the initial set of variables available for selection and $\mathcal{I}_0 = \varnothing$ represent the initial set of variables which have already been selected (i.e., none).

The proposed STORM algorithm is described as follows:

1. Initialization Stage

(1.1) Initial setup: Set $m = 1$, and let $Y_{i,1} = Y_i$ and $X_{ij,1} = X_{ij}$ for $i = 1, \ldots, n$.

(1.2) Initial regression: Calculate the ordinary least square estimates, $\hat{\beta}_{1,1}, \ldots, \hat{\beta}_{p,1}$, by $\hat{\beta}_{j,1} = \sum_{i=1}^n Y_{i,1}X_{ij,1}$, for $j = 1, \ldots, p$.

(1.3) Initial shrinkage: Calculate the LASSO estimates,

$$\hat{\beta}_{j,1}^L = \begin{cases} \hat{\beta}_{j,1} - \lambda/2, & \text{if } \hat{\beta}_{j,1} > \lambda/2, \\ 0, & \text{if } \left|\hat{\beta}_{j,1}\right| \le \lambda/2, \\ \hat{\beta}_{j,1} + \lambda/2, & \text{if } \hat{\beta}_{j,1} < -\lambda/2. \end{cases} \tag{2}$$

(1.4) Initializing selection set: Let $\mathcal{K}_1 = \mathcal{J}_0$, and calculate $c_{jk,1} = \sum_{i=1}^n X_{ij,1}X_{ik,1}$, for all $j, k \in \mathcal{K}_1$.

(1.5) Initial selection: Select $X_{j_1^*}$ as the most effective variable in the first step, where

$$\begin{aligned} j_1^* &= \arg\min_j \sum_{i=1}^n (Y_{i,1} - \hat{\beta}_{j,1}^L X_{ij,1})^2 \\ &= \arg\max_j \left\{ 2\hat{\beta}_{j,1}\hat{\beta}_{j,1}^L - \left(\hat{\beta}_{j,1}^L\right)^2 \right\}. \end{aligned}$$

(1.6) First updating: Let

$$\begin{aligned} R_1 &= \sum_{i=1}^n Y_{i,1}^2 - \sum_{i=1}^n (Y_{i,1} - \hat{\beta}_{j,1}^L X_{ij,1})^2 \\ &= 2\hat{\beta}_{j_1^*,1}\hat{\beta}_{j_1^*,1}^L - \left(\hat{\beta}_{j_1^*,1}^L\right)^2, \\ \mathcal{J}_1 &= \mathcal{K}_1 \setminus \{j_1^*\}, \\ \mathcal{I}_1 &= \{j_1^*\}. \end{aligned}$$

2. Recursion Stage

(2.1) Recursion step: Set m to $m + 1$, and let

$$Y_{i,m+1} = Y_{i,m} - \hat{\beta}_{j_m^*,m}^L X_{ij_m^*,m}, \quad i = 1, \ldots, n.$$

(2.2) Recursion correlation: Let $Z_{ij,m+1} = X_{ij,m} - c_{jj_m^*,m}\, X_{ij_m^*,m}$, then define

$$\mathcal{K}_m = \left\{ j \in \mathcal{J}_m : \sum_{i=1}^n Z_{ij,m+1}^2 \ge \eta \right\}.$$

(2.3) Recursion data update: For all $j \in \mathcal{K}_m$, compute

$$X_{ij,m+1} = \frac{Z_{ij,m+1}}{\left[\sum_{i=1}^n Z_{ij,m+1}^2\right]^{1/2}}, \tag{3}$$

and for all $j, k \in \mathcal{K}_m$, compute

$$c_{jk,m+1} = \frac{c_{jk,m} - c_{jj^*_m,m}\, c_{kj^*_m,m}}{\sqrt{1 - c^2_{jj^*_m,m}}\sqrt{1 - c^2_{kj^*_m,m}}}. \quad (4)$$

(2.4) Recursion regression: For $j \in \mathcal{K}_m$, calculate regression coefficients

$$\hat{\beta}_{j,m+1} = \frac{\hat{\beta}_{j,m} - \hat{\beta}_{j^*_m,m}\, c_{jj^*_m,m}}{1 - c^2_{jj^*_m,m}}. \quad (5)$$

(2.5) Recursion shrinkage: Calculate the LASSO estimates,

$$\hat{\beta}^L_{j,m+1} = \begin{cases} \hat{\beta}_{j,m+1} - \lambda/2, & \text{if } \hat{\beta}_{j,m+1} > \lambda/2, \\ 0, & \text{if } \left|\hat{\beta}_{j,m+1}\right| \le \lambda/2, \\ \hat{\beta}_{j,m+1} + \lambda/2, & \text{if } \hat{\beta}_{j,m+1} < -\lambda/2. \end{cases} \quad (6)$$

(2.6) Recursion selection: Select $X_{j^*_{m+1}}$ as the most effective variable in the $(m+1)^{\text{th}}$ step, where

$$\begin{aligned} j^*_{m+1} &= \arg\min_{j \in \mathcal{K}_m} \sum_{i=1}^{n} (Y_{i,m} - \hat{\beta}^L_{j,m+1} X_{ij,m+1})^2 \\ &= \arg\max_{j \in \mathcal{K}_m} \left\{ 2\hat{\beta}_{j,m+1}\hat{\beta}^L_{j,m+1} - \left(\hat{\beta}^L_{j,m+1}\right)^2 \right\}. \end{aligned}$$

(2.7) Recursion updating: Let

$$\begin{aligned} R_{m+1} &= \sum_{i=1}^{n} Y^2_{i,m+1} - \sum_{i=1}^{n} (Y_{i,m+1} \\ &\quad - \hat{\beta}^L_{j,m+1} X_{ij,m+1})^2 \\ &= 2\,\hat{\beta}_{j^*_{m+1},m+1}\,\hat{\beta}^L_{j^*_{m+1},m+1} \\ &\quad - \left(\hat{\beta}^L_{j^*_{m+1},m+1}\right)^2, \\ \mathcal{J}_{m+1} &= \mathcal{K}_m \setminus \{j^*_{m+1}\}, \\ \mathcal{I}_{m+1} &= \mathcal{I}_m \cup \{j^*_{m+1}\}. \end{aligned}$$

3. Stopping Rule: Repeat 'Recursion Stage' until $R_{m+1} < \delta$.

4. Final Step: Once stopped, save the current set of selected variables, $\mathcal{I}_m$, and compute the ordinary least squares (OLS) regression of Y on $\{X_j : j \in \mathcal{I}_m\}$. This gives the final model and prediction formula.

Statistical Analysis and Data Mining DOI:10.1002/sam

2.2. Properties

The STORM algorithm outlined above has the following properties:

(i) The residual sum of squares decreases in every iteration.

(ii) The maximum number of steps allowed is the number of linearly independent variables, which is bounded by $\min(p, n) \le n$. In particular, the proposed forward selection algorithm automatically terminates.

(iii) Because of the orthogonalization step, no variable can be repeated more than once, whether or not the original regressors are orthogonal.

(iv) In the orthogonal design case, the procedure will reduce to the OLS procedure based on variables selected by the LASSO.

With the exception of property (iii) above, similar properties are also shared by the FIRST procedure. Below, we briefly give arguments why these properties hold.

Proof of (i): At iteration step m, the most effective variable $X_{j^*_m}$ is selected by minimizing the objective criterion $\sum_{i=1}^{n}(Y_{i,m} - X_{ij,m}\hat{\beta}^L_{j,m})^2$ among all $j \in \mathcal{K}_m$, with notations as introduced in Subsection 2.1. Thus, by the definition of $\hat{\beta}^L_{j,m}$,

$$\begin{aligned} \sum_{i=1}^{n} Y^2_{i,m+1} &= \sum_{i=1}^{n} (Y_{i,m} - \hat{\beta}^L_{j^*_m,m} X_{ij^*_m,m})^2 \\ &\le \sum_{i=1}^{n} (Y_{i,m} - \hat{\beta}^L_{j^*_m,m} X_{ij^*_m,m})^2 + \lambda|\hat{\beta}^L_{j^*_m,m}| \\ &\le \sum_{i=1}^{n} (Y_{i,m} - 0 \times X_{ij^*_m,m})^2 + \lambda|0| \\ &= \sum_{i=1}^{n} Y^2_{i,m}. \end{aligned}$$

Therefore the residual sum of squares decreases in each step.

Proof of (ii) and (iii): At each step, due to the orthogonalization step, the residual variable is orthogonal to the variables which have been already selected in the previous steps. In particular, if the jth variable at stage $(m+1)$ is exactly a linear combination of already selected variables, then the residual variable obtained by step (2.2) is identically zero, and hence $j \notin \mathcal{K}_m$, ruling out the selection of variable j in the predictor. Thus in particular, no variable

can be repeated, proving (iii) as well as justifying the second operation in step (2.7). Further, as there can be at most n linearly independent predictors while observing a sample of size n, property (ii) follows.

Proof of (iv): If the variables are already orthogonal, then steps (2.2) and (2.3) are redundant, and the procedure will select variables according to the decreasing order of nonzero $|\hat{\beta}_j^L|$, or equivalently, according to the decreasing order of nonzero $|\hat{\beta}_j|$. It is well known that, the LASSO in this special case also coincides with this procedure. Thus, in view of step 4 in the algorithm, the final procedure will be the same as ordinary least squares estimate applied on the variables retained by the LASSO.

An analog of Theorem 2.1 in ref [13] is also valid for the STORM; namely in the orthogonal design case, the STORM is consistent for model selection. This follows from two observations—the LASSO has the model selection consistency property, and the STORM coincides with the LASSO followed by OLS.

2.3. Tuning

There are three parameters in the STORM determining its performance, namely, δ, η, and λ. The role of δ is to determine when the recursion stops. We find that, as long as δ is small enough, the choice of δ affects the predictive performance very little. The main motivation of using η is that it effectively discourages considering for selection the predictors which are highly correlated with those already in the model, since there is little extra gain in reducing the residual sum of squares. Also, the use of η assures the numerical stability in the orthogonalization step.

The most important parameter in the procedure is λ, which needs to be tuned carefully to assure a proper selection and estimation result. This is in line with the FIRST procedure introduced by Hwang et al. [13]. Compared with the FIRST, one additional parameter we need to tune is η, which controls the thresholding of orthogonalized variables and therefore its value may significantly affect the chosen model and the estimated coefficients. Since the ideal value of η is not known, some tuning criterion such as AIC, BIC, or cross validation may be used. In our numerical experiments, we use a joint five-fold cross validation, i.e., a set of possible values of λ and η are chosen over a grid and the prediction error is estimated by cross-validation. The value minimizing the estimated error is then chosen in the final estimation stage using all data. In practice, we recommend tuning both λ and η in the STORM, but computing time may be significantly saved by simply selecting a sensible value for η without cross-validation.

2.4. Extensions

Although the STORM procedure described above is based on the one-dimensional LASSO, the idea can be extended to use other shrinkage estimators such as the nonnegative garrote or the naive elastic net as the basis. This only requires a simple change in the update equation (6). For the nonnegative garrote procedure, the updating formula, starting from the least squares estimator, is

$$\hat{\beta}_{j,m+1}^{\mathrm{NG}} = \begin{cases} \hat{\beta}_{j,m+1} - \left|\frac{\lambda}{2\hat{\beta}_{j,m+1}}\right|, & \text{if } \hat{\beta}_{j,m+1} > \sqrt{\lambda/2}, \\ 0, & \text{if } \left|\hat{\beta}_{j,m+1}\right| \le \sqrt{\lambda/2}, \\ \hat{\beta}_{j,m+1} + \left|\frac{\lambda}{2\hat{\beta}_{j,m+1}}\right|, & \text{if } \hat{\beta}_{j,m+1} < -\sqrt{\lambda/2}. \end{cases} \tag{7}$$

The naive elastic net procedure [9] employs the combination of the L_1- and the L_2-penalty, which obtains

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\};$$

here $\lambda_1 > 0$ and $\lambda_2 > 0$ are two tuning parameters. To implement the STORM-based naive elastic net procedure, we need to replace Eq. (6) by the following

$$\hat{\beta}_{j,m+1}^{\mathrm{EN}} = \begin{cases} \frac{\hat{\beta}_{j,m+1} - \lambda_1/2}{1+\lambda_2}, & \text{if } \hat{\beta}_{j,m+1} > \lambda_1/2, \\ 0, & \text{if } \left|\hat{\beta}_{j,m+1}\right| \le \lambda_1/2, \\ \frac{\hat{\beta}_{j,m+1} + \lambda_1/2}{1+\lambda_2}, & \text{if } \hat{\beta}_{j,m+1} < -\lambda_1/2. \end{cases} \tag{8}$$

Note that the corrected elastic net estimate [9] for a one-dimensional predictor coincides with the lasso and hence does not lead to a different estimator.

3. SIMULATION

We now demonstrate the performance of the STORM method under various settings. We concentrate on large p and small n situations and generate the data from a high-dimensional sparse linear model,

$$Y_i = \boldsymbol{X}_i^{\mathrm{T}}\beta + \sigma\epsilon_i, \ i = 1, \ldots, n,$$

where $\boldsymbol{X}_i \in \mathbb{R}^p$, $p > n$, $\sigma > 0$ is a scale parameter, and ϵ_i's are independent and identically distributed (i.i.d.) errors from $N(0, 1)$. For each method, we generate a training data

set to fit the model, a validation data set to select the tuning parameter values, and finally an independent test data set to assess the prediction accuracy of the resulting estimators. The training and validating data sets are of each size n, while the test data set is always of size 1000. For each experiment, we run 100 Monte-Carlo simulations and report the average results along with the Monte-Carlo errors.

We compare the STORM method and its aggressive version (aggrSTORM) with the FIRST followed by OLS (FIRST+OLS), the adaptive FIRST followed by OLS (aFIRST+OLS), and the LASSO. Recall that the aggrSTORM excludes the participation of a predictor j for which $\hat{\beta}_j^L = 0$ at any step, in the subsequent steps and therefore can save computing time significantly. The LARS algorithm is used to implement the LASSO, which is available in *R*. The optimal tuning parameters for each method are chosen with a grid search method using the validation data set. Both the STORM and the aggrSTORM require tuning of both λ and η, which is done using a two-dimensional two-stage grid search.

3.1. Simulation Settings

We examine four experiment settings by varying sample sizes, error variance (or signal strength), and correlation scenarios among the covariates. The following is a detailed description of the examples.

(a) In Example 1, we set $p = 1000$ and $n = 100$ or 500. The covariates $X_1, \ldots, X_p$ are taken to be i.i.d. $N(0, 1)$ variables. The error variance is $\sigma^2 = 1$ for both cases. The true coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ has only 10 nonzero coefficients (3, 3, 3, 3, 1.5, 1.5, 1.5, 2, 2, 2), and the remaining coefficients are zero. We equally space the locations of the nonzero values in the coefficient vector. More specifically, β_1, β_{101}, β_{201}, and β_{301} are 3, β_{401}, β_{501}, and β_{601} are 1.5, and β_{701}, β_{801}, and β_{901} are 2.

(b) Example 2 has the same settings as Example 1, but we introduce moderate correlation among the covariates. The correlation between two covariates, X_i and X_j, is $\mathrm{corr}(X_i, X_j) = \rho^{|i-j|}$. We set $\rho = 0.5$. For this example, we consider $n = 100$ and two error variances: $\sigma^2 = 1$ and $\sigma^2 = 4$.

(c) Example 3 is the same as Example 2, except we increase the correlation between covariates to $\rho = 0.75$.

(d) In Example 4, we increase the number of covariates to $p = 2500$. The coefficient vector still has only 10 nonzero coefficients, which are identical to the nonzero coefficient values in Example 1. The locations of the nonzero values are again equally spaced throughout the coefficient vector. For this example, we consider $n = 100$ and $\sigma^2 = 1$.

(e) Example 5 introduces moderate correlation between the covariates in Example 4. We set $\rho = 0.5$, $n = 100$, and again consider the two error variances: $\sigma^2 = 1$ and $\sigma^2 = 4$.

(f) Example 6 is identical to Example 5, but we increase the correlation to $\rho = 0.75$.

Table 1. Simulation results for Example 1 ($p = 1000$, $\sigma^2 = 1$, independent covariates)

	Test error		False negative		False positive	
Method	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
LASSO (LARS)	2.574 (0.084)	1.152 (0.006)	0	0	53.38	66.49
FIRST+OLS	1.316 (0.041)	1.022 (0.005)	0	0	4.74	0
aFIRST+OLS	1.629 (0.067)	1.019 (0.005)	0	0	18.35	0
STORM	1.119 (0.009)	1.016 (0.005)	0	0	0.03	0
aggrSTORM	1.400 (0.076)	1.015 (0.005)	0.03	0	5.20	0

Table 2. Simulation results for Example 2 ($p = 1000$, $n = 100$, correlated covariates with $\rho = 0.5$)

	Test error		False negative		False positive	
Method	$\sigma^2 = 1$	$\sigma^2 = 4$	$\sigma^2 = 1$	$\sigma^2 = 4$	$\sigma^2 = 1$	$\sigma^2 = 4$
LASSO (LARS)	2.530 (0.065)	10.682 (0.339)	0	0.09	50.01	52.79
FIRST+OLS	1.606 (0.158)	7.719 (0.254)	0	0.14	8.38	21.73
aFIRST+OLS	1.958 (0.166)	8.317 (0.328)	0.09	0.30	15.08	24.7
STORM	1.117 (0.008)	5.401 (0.389)	0	0.15	0.14	0.81
aggrSTORM	1.395 (0.062)	6.635 (0.295)	0	0.34	6.42	6.31

Table 3. Simulation results for Example 3 ($p = 1000$, $n = 100$, correlated covariates with $\rho = 0.75$)

	Test error		False negative		False positive	
Method	$\sigma^2 = 1$	$\sigma^2 = 4$	$\sigma^2 = 1$	$\sigma^2 = 4$	$\sigma^2 = 1$	$\sigma^2 = 4$
LASSO (LARS)	2.562 (0.070)	9.911 (0.253)	0	0.11	49.84	49.18
FIRST+OLS	2.371 (0.251)	9.111 (0.579)	0.31	0.87	10.31	15.71
aFIRST+OLS	2.153 (0.155)	8.701 (0.370)	0.24	0.80	15.76	19.38
STORM	1.503 (0.137)	7.198 (0.440)	0.07	0.66	5.16	4.49
aggrSTORM	1.775 (0.103)	7.093 (0.276)	0.15	0.79	9.21	5.14

Table 4. Simulation results for Example 4 ($p = 2500$, $\sigma^2 = 1$, $n = 100$, independent covariates)

Method	Test error	False negative	False positive
LASSO (LARS)	4.447 (0.282)	0.02	63.75
FIRST+OLS	1.681 (0.085)	0	11.16
aFIRST+OLS	2.526 (0.191)	0.10	21.31
STORM	1.126 (0.008)	0	0.02
aggrSTORM	1.552 (0.102)	0	7.74

3.2. Experiments and Results

We compare four different methods with regard to their prediction error and variable selection performance. For prediction performance, we report the mean squared error evaluated on the test set. For variable selection, we report two types of selection errors: 'False Negative' defined as the (average) number of nonzero coefficients which are estimated as zero, and 'False Positive' defined as the number of zero coefficients which are not estimated as zero.

From Tables 1–6, we observe that the STORM consistently has the best prediction performance, in terms of the 'Test Error' for prediction. In terms of the selection, the STORM also shows advantages with small False Negative and False Positive numbers. The aggrSTORM generally performs the second best among all the methods under comparison.

Figure 1 presents a plot where the relative performance of the STORM against the FIRST is compared, showing the worth of the orthogonalization step, as correlation varies.

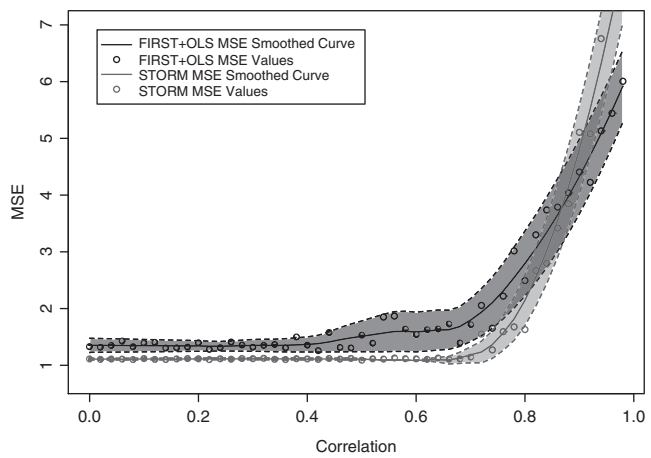

Fig. 1 Comparison of the STORM and the FIRST procedures for varying degrees of covariate correlations ranging from $\rho = 0$ to $\rho = 0.98$. Each data point corresponds to the average MSE for $N = 75$ Monte-Carlo trials with $n = 100$, $p = 1000$, $\sigma^2 = 1$, and true covariate vector, $\boldsymbol{\beta}$, identical to that from the previous examples. Lines are created using LOESS on the average MSE values for each ρ value. Dotted lines correspond to the LOESS 95% CI for the average MSE at each ρ value. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

4. REAL DATA ANALYSIS

We explore how the STORM performs when applied to real data by considering the gene expression data used in ref [15]. The dataset features 31 099 probe sets and 120 observations. The computation is made more manageable by performing two stages of prescreening. First, we remove the 3815 probe sets whose maximum values are not greater than the 25th percentile of the entire probe set. Then in the second stage, we select the 3000 probe sets with the largest variance of the 27 283 remaining probe sets. In our analysis, these 3000 probe sets are used as predictors, with the response being the RMA expression value for the probe 1389163_at, which has recently been found to cause Bardet-Biedl syndrome [15].

We randomly divide the data into a training set of size 100 and test set of size 20. We implement the STORM with $\delta = 0.001$. Five-fold cross validation is conducted with the training data set in order to tune the λ and η parameters. The procedure is repeated 10 times. We consider the FIRST+OLS, the LASSO+OLS, the STORM, and the aggrSTORM for comparison. Table 7 shows that all the methods give competitive prediction performance in terms of test errors, while the STORM and the aggrSTORM select fewer number of genes than others. This suggests that the STORM and the aggrSTORM are able to find a fairly simple regression model which can still predict the response quite accurately.

Table 5. Simulation results for Example 5 ($p = 2500$, $n = 100$, correlated covariates with $\rho = 0.50$)

	Test error		False negative		False positive	
Method	$\sigma^2 = 1$	$\sigma^2 = 4$	$\sigma^2 = 1$	$\sigma^2 = 4$	$\sigma^2 = 1$	$\sigma^2 = 4$
LASSO (LARS)	4.688 (0.383)	14.868 (0.547)	0.05	0.30	65.45	63.19
FIRST+OLS	1.782 (0.130)	9.857 (0.362)	0.03	0.32	12.50	29.4
aFIRST+OLS	2.457 (0.198)	10.902 (0.529)	0.12	0.72	21.35	25.41
STORM	1.11 (0.008)	5.445 (0.243)	0	0.23	0.07	0.9
aggrSTORM	1.61 (0.087)	7.171 (0.254)	0.01	0.26	9.04	11.44

Table 6. Simulation results for Example 6 ($p = 2500$, $n = 100$, correlated covariates with $\rho = 0.75$)

	Test error		False negative		False positive	
Method	$\sigma^2 = 1$	$\sigma^2 = 4$	$\sigma^2 = 1$	$\sigma^2 = 4$	$\sigma^2 = 1$	$\sigma^2 = 4$
FIRST+OLS	3.142 (0.271)	11.045 (0.457)	0.38	1.23	16.01	20.08
aFIRST+OLS	4.032 (0.397)	12.064 (0.537)	0.6	1.58	18.66	18.64
LASSO (LARS)	4.253 (0.331)	13.788 (0.464)	0.05	0.28	62.5	59.18
STORM	1.572 (0.198)	8.185 (0.704)	0.16	1.1	0.84	1.85
aggrSTORM	2.224 (0.144)	8.535 (0.571)	0.32	0.94	9.32	7.52

Table 7. Real data analysis results

Method	Test error	Num. genes selected
FIRST+OLS	0.0261 (0.0058)	50.6 (4.3)
LASSO+OLS	0.0211 (0.0021)	36.4 (4.1)
STORM	0.0284 (0.0017)	30.5 (3.4)
aggrSTORM	0.0207 (0.0025)	25.1 (3.4)

ACKNOWLEDGMENTS

This work is supported by National Security Agency Grant number H98230-12-1-0229 and NSF DMS-1309507.

REFERENCES

[1] J. Rawlings, S. Pantula, and D. Dickey, Applied Regression Analysis, Springer, New York, 2001.
[2] P. Bühlmann, Boosting for high-dimensional linear models, Ann Stat 34 (2006), 559–583.
[3] H. Wang, Forward regression for ultra-high dimensional variable screening, J Am Stat Assoc 104 (2012), 1512–1524.
[4] E. Hoerl and W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technomerics 12 (1970), 55–67.
[5] L. Breiman, Better subset selection using the non-negative garotte, Technometrics 37 (1995), 373–384.
[6] R. Tibshirani, Regression shrinkage and selection via the Lasso, J R Stat Soc 58 (1996), 147–169.
[7] J. Fan and R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J Am Stat Assoc 96 (2001), 1348–1360.
[8] H. Zou, The adaptive Lasso and its oracle properties, J Am Stat Assoc 476 (2006), 1418–1429.
[9] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net", J R Stat Soc Vol. 67 (2005), 301–320.
[10] E. George and R. McCulloch, Variable selection via Gibbs sampling, J Am Stat Assoc 88 (1993), 881–889.
[11] H. Ishwaran and J. Rao, Spike and slab variable selection: frequentist and Bayesian strategies, Ann Stat 33 (2005), 730–773.
[12] C. Hans, A. Dobra, and M. West, Shotgun stochastic search for large p regression, J Am Stat Assoc 102 (2007), 507–516.
[13] W. Hwang, H. H. Zhang, and S. Ghosal, FIRST: Combining forward selection and shrinkage in high dimensional linear regression, Stat Interf 2 (2009), 341–348.
[14] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, Least angle regression, Ann Stat 24 (2004), 407–499.
[15] J. Huang, S. Ma, and C.-H. Zhang, Adaptive Lasso for sparse high-dimensional regression models, Stat Sin 18 (2009), 1603–1618.