Least-Angle Regression and LASSO for Large Datasets

Chris Fraley and Tim Hesterberg

October 1, 2007

Abstract

Least-Angle Regression and the LASSO (ℓ_1-penalized regression) offer a number of advantages over procedures such as stepwise or ridge regression in variable selection applications, including prediction accuracy, stability and interpretability. We discuss formulations of these algorithms that extend to datasets in which the number of observations could be so large that it would not be possible to access the predictors as a unit in computations. Our methods require a single pass through the data for orthogonal transformation, effectively reducing the dimension of the computations required to obtain the regression coefficients and residual sums-of-squares.

Keywords: regression, regularization, ℓ_1 penalty, lasso, scalable, massive datasets, tall datasets.

1 Introduction

This paper addresses the problem of least-angle regression (Efron et al. 2004) and its extension to the LASSO method (Tibshirani 1996) for linear models with very large numbers of observations. We write the model as:

$$y = \begin{pmatrix} \mathbf{1} & X \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} + \epsilon, \tag{1.1}$$

where y is a numeric response, $\mathbf{1}$ a column of ones, X an $n \times p$ numeric matrix of predictors, β an intercept, b a vector of p additional regression coefficients, and ϵ a vector of residuals. We assume that:

- the number of observations n may be so large that X (and possibly y) could not be stored in memory.
- the number of predictors p is sufficiently small that matrices with dimensions of order p could be held in memory and used in computations.
- X and y are stored in such a way that the rows of X can be accessed in a sequential blockwise fashion, and the corresponding components of the response y can be accessed with each block of rows of X.

Least-angle regression and LASSO were originally described in terms of X and y, and an implementation suitable for datasets of moderate size has been made available (Efron and Hastie 2003).

The main idea in this paper is to do all of the necessary calculations in memory using an orthogonal transformation of the matrix of predictors and the response that reduces the dimension of the computations. The necessary transformations can be accumulated in a single pass through the data, which need not reside in memory, but is assumed blockwise accessible. The process is a memory-efficient version of the QR decomposition that is used for linear regression. Here we show how it can also be applied to least-angle regression and its LASSO extension.

The paper is organized as follows. The basic LARS algorithm is described in Section 2. We show in Section 3 how to express all of the necessary computations without accessing X or y after initial factorization and transformation to matrices and vectors with dimensions of order p. In Section 4 we show how to obtain this factorization and transformation using sequential blockwise access to rows of X along with simultaneous access to corresponding entries of the response y. The results of Section 3 and Section 4 allow extension of LARS to very large blockwise-accessible datasets. Section 5 summarizes the results and provides further discussion. Derivations of some of the relations used in the algorithms are given in the appendix, as well as information on a scalable S-PLUS implementation.

2 Algorithm Description

Efron et al. (2003, 2004) develop least angle regression as a variable-selection technique for linear models. There are three basic variants: *least-angle regression (LAR)*, *ℓ_1 penalty (LASSO)*, and *forward stagewise*, which are collectively referred to as *LARS*. Although the methods described here can be applied to all of these algorithms, we discuss only LAR and LASSO since the forward stagewise variant can require many steps and is not widely used.

In Efron et al. (2003, 2004), the predictors and response are transformed to have mean 0. This effectively assumes an intercept in the model, as shown in the appendix. Efron et al. (2004) also assume that the predictors have unit Euclidean length, but we do not make this assumption here. We write the model as

$$\tilde{y} = \widetilde{X} b + \epsilon, \tag{2.2}$$

where $\tilde{y}$ denotes the response with its mean subtracted, and $\widetilde{X}$ denotes the matrix of predictors X, each with its mean subtracted. Later we show how to implement the methods in reduced dimensions, without first centering or scaling the variables.

The methods proceed iteratively in a series of steps. We make the following definitions:

- b_k: the vector of coefficients at (the end of) step k,
- $\mathcal{A}_k$: the active set of predictors during step k,
- p_k: the number of active predictors,
- $\widetilde{X}_k$: the columns of $\widetilde{X}$ corresponding to $\mathcal{A}_k$,
- P_k: a $p_k \times p$ matrix of zeroes and ones such that $\widetilde{X}_k = \widetilde{X} P_k^T$,
- $r_k \equiv \tilde{y} - \widetilde{X} b_k$: the vector of partial residuals at step k.

Initially all coefficients b_0 are zero, and $\mathcal{A}_0$ is empty. A series of models is fitted in which predictors are successively added to or dropped from the active set, and coefficients are updated. The final model has a maximal set of linearly-independent predictors.

At the kth step, the active set is updated, and a new set of coefficients b_k are obtained by determining a step length λ_k along a direction d_k from b_{k-1}:

$$b_k \leftarrow b_{k-1} + \lambda_k d_k, \tag{2.3}$$

with $0 \leq \lambda_k \leq 1$. The active set and step length calculations differ for LAR and LASSO, and we compare these below.

The direction d_k comes from the least-squares estimate based on the active set. Let

$$z_k = \operatorname*{argmin}_z \|y - \widetilde{X}_k z\| \tag{2.4}$$

be the least-squares coefficients based on the current active set. Equivalently, z_k satisfies the normal equations

$$\widetilde{X}_k^T \widetilde{X}_k z_k = \widetilde{X}_k^T y. \tag{2.5}$$

The direction d_k is the vector of length p which is equal to $z_k - P_k b_{k-1}$ in the entries corresponding to the active predictors, and 0 elsewhere:

$$d_k = P_k^T (z_k - P_k b_{k-1}).$$

The step length λ_k to be taken from b_{k-1} along d_k to get the new set of coefficients is determined by the algorithm (LAR or LASSO), based on the correlations of the predictors with the partial residuals:

$$cor(\widetilde{X}, r_k) \equiv \frac{\mathcal{D}_X^{-1} \widetilde{X}^T r_k}{\|r_k\|_2},$$

where

$$\mathcal{D}_X \equiv diag\left(\|\widetilde{X}_{*1}\|_2, \|\widetilde{X}_{*2}\|_2, \ldots, \|\widetilde{X}_{*p}\|_2 \right).$$

The correlations are proportional to the inner products between the partial residuals and predictors, standardized by their norms. Dropping the denominator for simplicity, we work with

$$c_k = \mathcal{D}_X^{-1} \widetilde{X}^T r_k \tag{2.6}$$

and with the more general

$$c(b) = \mathcal{D}_X^{-1} \widetilde{X}^T (\tilde{y} - \widetilde{X} b).$$

At the start of a step, the active set is defined to be the predictors corresponding to the correlations that have the largest magnitude. These correlations would vanish if a unit step were taken in the direction d_k. The first predictors to enter the active set are those with the largest absolute correlations with $\tilde{y}$. Subsequently, the differences in step length and active variable selection for the different LAR and LASSO are:

- **Least-Angle Regression**

  The step length λ_k in LAR is the smallest step such that one or more predictors that are not in the current active set $\mathcal{A}_k$ have correlation equal in magnitude to the correlations of members of $\mathcal{A}_k$ at $b_{k-1} + \lambda d_k$. Those predictor(s) are added to the active set for the following step. The required computations are straightforward because

  $$c(b_{k-1} + \lambda d_k) = \mathcal{D}_X^{-1} \widetilde{X}^T \left[\tilde{y} - \widetilde{X}(b_{k-1} + \lambda d_k) \right] \tag{2.7}$$

  is a linear function of λ. In LAR, predictors never leave the active set once they are added.

- **LASSO and ℓ_1 Penalty Regression**

  A *LASSO* (Tibshirani 1996) solution minimizes

  $$\min_b \|\tilde{y} - \widetilde{X}_k b\|_2^2 + \theta_k \|\mathcal{D}_X^{-1} b\|_1$$

for some $\theta_k > 0$. Coefficients are scaled in the ℓ_1 penalty term for consistency with Tibshirani (1996) and Efron et al. (2004), where columns of $\widetilde{X}$ are normalized.

At a LASSO solution, correlations corresponding to non-zero components of b_k are maximal in magnitude and have sign equal to the corresponding element of b_k (e.g. Osborne et al. 2000). In the LASSO modification of least-angle regression, the step length λ_k is the smallest step such that either $b_{k-1} + \lambda d_k$ changes sign for one or more of the predictors in the current active set, or the LAR step criterion holds (Efron et al. 2004). In either case, $b_k = b_{k-1} + \lambda_k d_k$ satisfies the LASSO optimality conditions. When the LASSO step is determined by a sign change, the predictors at which the coefficients change sign are zero and are dropped from the active set at the next iteration (no new predictors are added).

3 Alternative Implementation in Reduced Dimensions

In this section, we show how to implement LAR and LASSO in reduced the dimensions using a Cholesky factor of $\widetilde{X}^T\widetilde{X}$ and the corresponding transformation of $\tilde{y}$. Later, in Section 4, we describe the one-pass blockwise transformation of $\begin{pmatrix} \mathbf{1} & X \end{pmatrix}$ and y to the forms needed for the methods described here.

3.1 Cholesky and QR Factorizations

A $p \times p$ upper triangular matrix $\widetilde{R}$ satisfying

$$\widetilde{R}^T\widetilde{R} = \widetilde{X}^T\widetilde{X},$$

is called a Cholesky factor [1]of $\widetilde{X}$. The Cholesky factor $\widetilde{R}$ can be obtained by applying a series of orthogonal Householder transformations to $\widetilde{X}$. There is an associated QR factorization of $\widetilde{X}$ in which

$$\widetilde{X} = \widetilde{Q}\widetilde{R},$$

where $\widetilde{Q}$ is an $n \times p$ matrix with orthogonal columns. In ordinary least squares, the coefficients b for the regression

$$\tilde{y} \sim \widetilde{X}$$

satisfy the normal equations

$$\widetilde{X}^T\widetilde{X}b = \widetilde{X}^T\tilde{y},$$

or equivalently

$$\widetilde{R}b = \widetilde{Q}^T\tilde{y}.$$

The latter has advantages for numerical accuracy, and also for efficiency in applications like LAR and LASSO. It is not necessary to form the matrix $\widetilde{Q}$, since the product $\widetilde{Q}^T y$ can be accumulated by sequential application of the transformations used to form $\widetilde{R}$. For extensive discussion of the QR factorization and its use in regression computations, see Chapter 5 of Golub and Van Loan (1996). In Section 4, we show how to obtain the Cholesky factor via blockwise application of Householder transformations.

[1] The Cholesky factor is unique up to the signs of the rows.

3.2 Directions and Correlations

Let $\widetilde{R}$ be an upper triangular Cholesky factor of $\widetilde{X}$, and let

$$\breve{y} \equiv \widetilde{Q}^T \tilde{y}$$

be the transformation of $\tilde{y}$ obtained by forming $\widetilde{R}$ from $\widetilde{X}$ via sequential orthogonal transformations. To obtain LAR/LASSO directions, we need the solution z_k to the normal equations (2.5) for the current active set:

$$\widetilde{X}_k^T \widetilde{X}_k z_k = \widetilde{X}_k^T y. \tag{3.8}$$

If $\widetilde{R}_k$ is the Cholesky factorization of $\widetilde{X}_k$, and $\breve{y}_k$ is the corresponding transformation of y, then (3.8) is equivalent to

$$\widetilde{R}_k z_k = \breve{y}_k. \tag{3.9}$$

Use of (3.9) has numerical advantages over (3.8), since growth in roundoff error is bounded by the square root of the condition number of $\widetilde{X}_k^T \widetilde{X}_k$ rather than the condition number.[2] The scaled correlations (2.6) are given by

$$c_k = \mathcal{D}_X^{-1} \widetilde{X}^T r_k = \mathcal{D}_X^{-1} \widetilde{X}^T (\tilde{y} - \widetilde{X} b_k) = \mathcal{D}_X^{-1} \widetilde{R}^T \left(\breve{y} - \widetilde{R}[b_{k-1} + \lambda_k d_k] \right). \tag{3.10}$$

Step lengths for LAR and LASSO are computed in reduced dimensions from (3.10) instead of (2.7).

Because $\widetilde{X}$ and $\widetilde{R}$ are related through orthogonal transformation of columns,

$$\|\widetilde{R}_{*j}\|_2 = \|\widetilde{X}_{*j}\|_2, \quad j = 1, \ldots, p,$$

where $\widetilde{R}_{*j}$ and $\widetilde{X}_{*j}$, represent the jth column of $\widetilde{R}$ and $\widetilde{X}$, respectively, so that

$$\mathcal{D}_X = diag \left(\|\widetilde{R}_{*1}\|_2, \|\widetilde{R}_{*2}\|_2, \ldots, \|\widetilde{R}_{*p}\|_2 \right).$$

Our methods compute $\widetilde{R}$ in one pass through the data, which also gives us an efficient way to compute $\mathcal{D}_X$ (which is needed for scaling the correlations) at the outset.[3]

3.3 Cholesky Factor and Transformed Response for Reduced Sets of Predictors

The Cholesky $\widetilde{R}_k$ of the reduced set of predictors $\widetilde{X}_k$ can be obtained from the Cholesky factor $\widetilde{R}$ of the full set of predictors $\widetilde{X}$ via orthogonal Householder transformations as illustrated in Figure 1. These Householder transformations must also be applied to the vector $\breve{y}$ to obtain the vector $\breve{y}_k$ needed in to compute the direction. Although it is not necessary, we assume that the columns of $\widetilde{X}_k$ appear in the same order as they would in $\widetilde{X}$, and that the reduced Cholesky factor is formed from the original accumulated Cholesky factor, rather than the Cholesky factor used in the $(k-1)$st step. However, if at stage k, the active set consists only of leading columns of $\widetilde{X}$ or of $\widetilde{X}_{k-1}$, then no update is necessary since $\widetilde{R}_k$ and $\breve{y}_k$ would consist of the corresponding leading columns of $\widetilde{R}$ or $\widetilde{R}_{k-1}$ and the corresponding leading elements of $\breve{y}$ or $\breve{y}_{k-1}$, respectively.

[2]The condition number of a positive semi-definite matrix is its largest eigenvalue divided by its smallest eigenvalue. This ratio goes to infinity as the matrix nears singularity.

[3]Alternatively, R can be scaled initially by its column norms and the coefficients correspondingly transformed after the procedure is completed.

$$\begin{array}{ccccc} \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & \times \end{array} \quad \overset{\substack{remove \\ predictors \\ 2\ and\ 4}}{\longrightarrow} \quad \begin{array}{c|c|c} \times & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & \times \end{array} \quad \overset{Householder}{\longrightarrow} \quad \begin{array}{ccc} \times & \times & \times \\ 0 & \otimes & \otimes \\ 0 & 0 & \otimes \\ 0 & 0 & \times \\ 0 & 0 & \times \end{array} \quad \overset{Householder}{\longrightarrow} \quad \begin{array}{ccc} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \otimes \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}$$

Figure 1: Restoration of triangular form via Householder transformations after removal of some predictors. predictors are kept in their original order.

3.4 Detecting Linear Dependence

The approach to linear models and LAR/LASSO via orthogonal factorization has an advantage over the normal equations in terms of detection of ill-conditioning and linear dependence in predictors, which is typically revealed by small diagonal elements in the resulting upper-triangular factor. When redundant predictors are detected, they can be eliminated using the updating process described in Section 3.3. In the case of LASSO, there should be the option to attempt to reintroduce predictors that are dropped due to ill-conditioning if one or more of the leading predictors present when the ill-conditioning was detected has been dropped from the active set.

3.5 Residual Sum of Squares

Values of the residual sum of squares $\left\| y - \begin{pmatrix} \mathbf{1} & X \end{pmatrix} \begin{pmatrix} \beta_k \\ b_k \end{pmatrix} \right\|_2^2$ are needed for regression diagnostics, such as the C_p statistic proposed in Efron et al. (2004) for variable section. If we have the upper-triangular matrix R from a QR factorization of $\begin{pmatrix} \mathbf{1} & X \end{pmatrix}$:

$$\begin{pmatrix} \mathbf{1} & X \end{pmatrix} = QR,$$

and the accumulated transformed response $Q^T y$, the residual sum of squares can be expressed in terms of R and $Q^T y$ as follows:

$$\left\| y - \begin{pmatrix} \mathbf{1} & X \end{pmatrix} \begin{pmatrix} \beta_k \\ b_k \end{pmatrix} \right\|_2^2 = \left\| Q^T y - R \begin{pmatrix} \beta_k \\ b_k \end{pmatrix} \right\|_2^2,$$

with no need to access X or y, or to explicitly form Q.

3.6 Obtaining R and $\widetilde{R}$

In the appendix, we show that if R is a Cholesky factor of $\begin{pmatrix} \mathbf{1} & X \end{pmatrix}$, then a Cholesky factor of $\widetilde{X}$ (the matrix $\widetilde{R}$ of Section 3) is available as a submatrix of R. We also show that the transformed response $\check{y} = \tilde{Q}^T \tilde{y}$ of Section 3 is a subvector if $Q^T y$. Hence orthogonal factorization of $\begin{pmatrix} \mathbf{1} & X \end{pmatrix}$ and simultaneous transformation of y suffices both for computing LAR/LASSO coefficients and for computation of the associated residual sums of squares.

4 Blockwise Cholesky Factorization

The first step in LAR/LASSO for large data sets is to form the upper triangular Cholesky factor R of $\begin{pmatrix} \mathbf{1} & X \end{pmatrix}^T \begin{pmatrix} \mathbf{1} & X \end{pmatrix}$ in one row-wise pass through the data X. As mentioned in Section 3.6, we

$$\begin{array}{ccc} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ \hline \times & \times & \times \\ \vdots & \vdots & \vdots \\ \times & \times & \times \end{array} \overset{Householder}{\longrightarrow} \begin{array}{ccc} \otimes & \otimes & \otimes \\ 0 & \times & \times \\ 0 & 0 & \times \\ \hline 0 & \otimes & \otimes \\ \vdots & \vdots & \vdots \\ 0 & \otimes & \otimes \end{array} \overset{Householder}{\longrightarrow} \begin{array}{ccc} \times & \times & \times \\ 0 & \otimes & \otimes \\ 0 & 0 & \times \\ \hline 0 & 0 & \otimes \\ \vdots & \vdots & \vdots \\ 0 & 0 & \otimes \end{array} \overset{Householder}{\longrightarrow} \begin{array}{ccc} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \otimes \\ \hline 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{array}$$

Figure 2: An illustration of the changes in sparsity structure in applying successive Householder transformations in updating the current upper triangular factor (above line) from a new block of rows of predictors (below line). Each arrow corresponds to application of a Householder transformation, $\times$ represents a potentially nonzero entry, and $\otimes$ represents an entry that would typically change value after the Householder transformation is applied. In general, k Householder transformations are needed to process a block with k columns.

show in the appendix that it suffices to form the Cholesky factorization of $\begin{pmatrix} \mathbf{1} & X \end{pmatrix}$ via orthogonal transformation and the corresponding transformation of y to obtain the triangular factor $\tilde{R}$ and the transformed response $\check{y}$ of Section 3. Only a limited number of rows of X will be available at a time, so we accumulate R by applying a succession of orthogonal Householder transformations using the factorization accumulated so far and the currently available rows of X. The corresponding entries of the response y must be available as the rows of X are processed in order to accumulate $\check{y}$, the transformation of y needed for the LAR and LASSO procedures.

Formation of a triangular Cholesky factor for a crossproduct matrix via blockwise orthogonal transformation is a straightforward extension of in-memory techniques used for the QR factorization (e.g. Chapter 5 of Golub and Van Loan 1996). Details useful from the point of view of implementation are not widely available, so we include a description here. We assume the data $X \mid y$ is read in blocks. The procedure as applied to a block of data is illustrated in Figure 2.

The accumulated upper-triangular matrix R is held separately in memory from the blocks of X as they are accessed. The computations involved in formation and application of each Householder transformation can be arranged to exploit this. The effect of an individual Householder transformation can be viewed as follows:

$$\left(\begin{array}{c} r_j^T \\ \hline \mathcal{B}_j \end{array}\right) = \left(\begin{array}{cc} \rho_j & s_j^T \\ \hline b_j & B_j \end{array}\right) \overset{Householder}{\longrightarrow} \left(\begin{array}{cc} \tilde{\rho}_j & \tilde{s}_j^T \\ \hline 0 & \mathcal{B}_{j+1} \end{array}\right) \quad j = 1, \ldots, p+1,$$

where r_j is the nonzero portion of a row of the accumulated R, and $\mathcal{B}_j$ consists of the corresponding columns of the current block of data. Both r_j and $\mathcal{B}_j$ are modified with each application of a Householder transformation.

The details of applying a Householder transformation in this context are as follows:

$$\begin{aligned}\left(I - \frac{2}{h_j^T h_j} h_j h_j^T\right)\begin{pmatrix} r_j^T \\ \mathcal{B}_j \end{pmatrix} &= \left(I - 2u_j u_j^T\right)\begin{pmatrix} r_j^T \\ \mathcal{B}_j \end{pmatrix} \quad (u_j \text{ is the unit vector in the direction of } h_j) \\ &= \begin{pmatrix} r_j^T \\ \mathcal{B}_j \end{pmatrix} - 2\begin{pmatrix} \omega_j \\ w_j \end{pmatrix}\begin{pmatrix} \omega_j & w_j^T \end{pmatrix}\begin{pmatrix} r_j^T \\ \mathcal{B}_j \end{pmatrix} \\ &= \begin{pmatrix} r_j^T \\ \mathcal{B}_j \end{pmatrix} - 2\begin{pmatrix} \omega_j\left[\omega_j r_j^T + w_j^T \mathcal{B}_j\right] \\ w_j\left[\omega r_j^T + w_j^T \mathcal{B}_j\right] \end{pmatrix},\end{aligned}$$

which gives the following algorithm for computation:

1. Let h_j be a Householder transformation that maps the first column of $\begin{pmatrix} r_j^T \\ B_j \end{pmatrix}$ to a vector whose first element is its only nonzero element:

$$h_j \leftarrow \begin{pmatrix} \rho_j + sign(\rho_j)\left\|\begin{pmatrix} \rho_j \\ b_j \end{pmatrix}\right\|_2 \\ b_j \end{pmatrix}.$$

2. $\begin{pmatrix} \omega_j \\ w_j \end{pmatrix} \leftarrow \frac{h_j}{\|h_j\|_2}$
3. $v_j \leftarrow 2\left(\omega_j r_j + \mathcal{B}_j^T w_j\right)$
4. $r_j^T \leftarrow r_j^T - \omega_j v_j^T$
5. $\begin{pmatrix} 0 & \mathcal{B}_{j+1} \end{pmatrix} \leftarrow \mathcal{B}_j - w_j v_j^T$ (first column vanishes)

Since an intercept is included in the regression, the first Householder transformation in each block involves the intercept, which is not explicitly stored. In this case,

$$h_1 \leftarrow \begin{pmatrix} \rho_1 + sign(\rho_1)\,\|\mathbf{1}_{n_B}\|_2 \\ \mathbf{1}_{n_B} \end{pmatrix} = \begin{pmatrix} \rho_1 + sign(\rho_1)\sqrt{n_B} \\ \mathbf{1}_{n_B} \end{pmatrix},$$

where n_B is the number of rows in the block.

5 Summary and Discussion

A Google Scholar search in September 2007 shows over 300 citations of Efron et al. (2004), and nearly 700 citations of Tibshirani (1996). These deal with issues such as categorical and grouped variables, extension to nonlinear models, and cases in which there are many more predictors than observations, but not with datasets with very large numbers of observations.

Our approach is a scalable extension to least-angle regression and the LASSO method. Formation of a triangular Cholesky factor for a crossproduct matrix via blockwise orthogonal transformation is a straightforward extension of in-memory techniques used for the QR factorization (e.g. Chapter 5 of Golub and Van Loan 1996); our contribution is its adaption and application to least-angle regression and LASSO. Although our methods do not apply to adaptive linear and nonlinear models because iterative evaluation of functions of the predictors would be required, our approach allows considerable flexibility, since it applies to any model of the form

$$y \sim (\mathbf{1} \quad \phi_1(X) \quad \ldots \quad \phi_m(X)),$$

where $\phi_1, \ldots \phi_m$ are (possibly nonlinear) functions of the predictors, m is of order p, and the functions are fixed in advance. Such functions could include higher-order terms and interactions between predictors, for example.

Fan et al. (2007) and Fan and Cheng (2007) also propose a blockwise approach to regression and variable selection for massive datasets, but they apply their methods to each block separately. Their approach requires synthesis of the blockwise results, and more than one pass is needed to determine an appropriate blocksize. In our proposed methodology, the results are independent (except for roundoff error) of the order of the observations as well as of the block size, which is limited only by the size of memory.

For information on an available S-PLUS implementation of our methods that uses the `bigdata` library to handle massive datasets by blockwise processing and on ongoing work on software for generalized least-angle regression, see

`http:www.insightful.com/Hesterberg/glars`

References

Efron, B. and Hastie, T. (2003) *LARS software for R and Splus.* `http://www-stat.stanford.edu/~hastie/Papers/LARS`.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *Annals of Statistics*, **32**, 407–451.

Fan, T.-H. and Cheng, K.-F. (2007) Tests and variables selection for regression analysis for massive datasets. *Data Knowledge and Engineering*, **63**, 809–817.

Fan, T.-H., Lin, D. K. J. and Cheng, K.-F. (2007) Regression analysis for massive datasets. *Data Knowledge and Engineering*, **61**, 554–562.

Golub, G. H. and Van Loan, C. F. (1996) *Matrix Computations.* Baltimore: Johns Hopkins University Press, 3rd edn.

Osborne, M. R., Presnell, B. and Turlach, B. A. (2000) On the LASSO and its dual. *Journal of Computational and Graphical Statistics*, **9**, 319–337.

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

Acknowledgments

This work was supported by NIH under NIH SBIR Phase I 1R43GM074313-01 and Phase II 2R44GM074313-02 awards (Tim Hesterberg, P.I.).

A Derivations

In this section, we derive some of the relations on which we have based our scalable version of least-angle regression and LASSO.

Let X be an $n \times p$ matrix, and y and n-vector. Consider the following regression problems:

$$y \sim \begin{pmatrix} \mathbf{1} & X \end{pmatrix} \tag{A.11}$$

and

$$\tilde{y} \sim \widetilde{X}, \tag{A.12}$$

where $\mathbf{1}$ is the vector of length n in which all components are equal to 1, and $\tilde{y}$ represents y with its mean subtracted out

$$\tilde{y} \equiv y - \bar{y}\mathbf{1},$$

and $\widetilde{X}$ represents X with column means subtracted out, that is,

$$\widetilde{X} \equiv X - \frac{\mathbf{1}\mathbf{1}^T}{n} X = \left(I - \frac{\mathbf{1}\mathbf{1}^T}{n} \right) X.$$

It is easy to see that if $\mathbf{b} = \begin{pmatrix} \beta \\ b \end{pmatrix}$ are ordinary least-squares regression coefficients for (A.11), then b are ordinary least-squares regression coefficients for (A.12), and

$$\beta = \frac{1^T(y - Xb)}{n} = \bar{y} - \left(\bar{X}_{*1} \quad \bar{X}_{*2} \quad \ldots \quad \bar{X}_{*p} \right)^T b, \tag{A.13}$$

where $\bar{X}_{*j}$ is the mean of the jth column of X.

Suppose that we have a QR factorization of $\begin{pmatrix} \mathbf{1} & X \end{pmatrix}$:

$$\begin{pmatrix} \mathbf{1} & X \end{pmatrix} = \mathcal{QR} = \mathcal{Q} \begin{pmatrix} R \\ O \end{pmatrix}, \tag{A.14}$$

where $\mathcal{Q}$ is an $n \times n$ matrix with orthogonal columns, and $\mathcal{R}$ is an $n \times p$ matrix, and R is a $p \times p$ upper triangular matrix. Then, if

$$\mathcal{Q} = \left(Q \quad Z \right),$$

where Q is $n \times p$, an alternative QR factorization is

$$\begin{pmatrix} \mathbf{1} & X \end{pmatrix} = QR. \tag{A.15}$$

Let R have the following partition

$$R = \begin{pmatrix} \rho & s^T \\ 0 & U \end{pmatrix}. \tag{A.16}$$

In **Proposition 1**, we derive the relationship between Cholesky factors of $\begin{pmatrix} \mathbf{1} & X \end{pmatrix}$ and $\widetilde{X}$. In **Proposition 2**, we derive a relationship between the transformations of y corresponding to a QR factorization of $\begin{pmatrix} \mathbf{1} & X \end{pmatrix}$ and the transformation of $\tilde{y}$ corresponding to a QR factorization of $\widetilde{X}$. In **Proposition 3**, we derive an expression for the regression coefficient of the intercept in (A.11) in terms of the QR factorization of $\begin{pmatrix} \mathbf{1} & X \end{pmatrix}$ and the corresponding transformation of y.

Proposition 1: Suppose that R is an upper triangular Cholesky factor $\begin{pmatrix} \mathbf{1} & X \end{pmatrix}^T \begin{pmatrix} \mathbf{1} & X \end{pmatrix}$:

$$R^T R = \begin{pmatrix} \mathbf{1} & X \end{pmatrix}^T \begin{pmatrix} \mathbf{1} & X \end{pmatrix},$$

R $p \times p$ upper triangular. Let R be partitioned as in (A.16). Then the upper triangular matrix U in (A.16) is a Cholesky factor of $\widetilde{X}$.

Proof.

$$R^T R = \begin{pmatrix} \rho & s^T \\ 0 & U \end{pmatrix}^T \begin{pmatrix} \rho & s^T \\ 0 & U \end{pmatrix} = \begin{pmatrix} \rho^2 & \rho s^T \\ \rho s & U^T U + s s^T \end{pmatrix}, \tag{A.17}$$

and

$$\begin{pmatrix} \mathbf{1} & X \end{pmatrix}^T \begin{pmatrix} \mathbf{1} & X \end{pmatrix} = \begin{pmatrix} \mathbf{1}^T \\ X^T \end{pmatrix} (\mathbf{1} \quad X) = \begin{pmatrix} \mathbf{1}^T\mathbf{1} & \mathbf{1}^T X \\ X^T \mathbf{1} & X^T X \end{pmatrix} = \begin{pmatrix} n & \mathbf{1}^T X \\ X^T \mathbf{1} & X^T X \end{pmatrix}.$$

Since $R^T R = \begin{pmatrix} \mathbf{1} & X \end{pmatrix}^T \begin{pmatrix} \mathbf{1} & X \end{pmatrix}$,

$$\begin{aligned} \rho^2 &= n \\ X^T 1 &= \rho s \\ X^T X &= U^T U + s s^T, \end{aligned} \tag{A.18}$$

from which it follows that

$$\begin{aligned} U^T U &= X^T X - s s^T = X^T X - \frac{X^T 1 1^T X}{\rho^2} = X^T X - \frac{X^T 1 1^T X}{n} \\ &= X^T \left(I - \frac{\mathbf{1}\mathbf{1}^T}{n} \right) X = X^T \left(I - \frac{\mathbf{1}\mathbf{1}^T}{n} \right) \left(I - \frac{\mathbf{1}\mathbf{1}^T}{n} \right) X = \widetilde{X}^T \widetilde{X}, \end{aligned}$$

so that U is a Cholesky factor of $\widetilde{X}$.

Proposition 2: Let y be any n vector, and consider the QR factorization (A.15). Let R be partitioned as in (A.16). and partition $Q^T y$ as

$$Q^T y = \begin{pmatrix} \zeta \\ z \end{pmatrix},$$

where ζ is a scalar. Then z differs from the transformation $\check{y} = \tilde{Q}^T \tilde{y}$ of $\tilde{y}$ would result in forming U from a QR factorization $\tilde{Q} U$ of $\widetilde{X}$ only in the null space of U^T. Moreover, if U is nonsingular (or, equivalently, the columns of $\widetilde{X}$ are linearly independent), $z = \check{y}$.

Proof. Let $\mathbf{b} = \begin{pmatrix} \beta \\ b \end{pmatrix}$ be ordinary least-squares regression coefficients for (A.11). Then,

$$\begin{aligned}
\begin{pmatrix} \mathbf{1} & X \end{pmatrix}^T \begin{pmatrix} \mathbf{1} & X \end{pmatrix} \mathbf{b} &= \begin{pmatrix} n & \mathbf{1}^T X \\ X^T\mathbf{1} & X^T X \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} n\beta + \mathbf{1}^T X b \\ \beta X^T \mathbf{1} + X^T X b \end{pmatrix} = \begin{pmatrix} \mathbf{1} & X \end{pmatrix}^T y \\
&= R^T Q^T y = \begin{pmatrix} \rho & s^T \\ 0 & U \end{pmatrix}^T \begin{pmatrix} \zeta \\ z \end{pmatrix} = \begin{pmatrix} \rho\zeta \\ \zeta w + U^T z \end{pmatrix}.
\end{aligned}$$

Using relations from (A.18), this can be rewritten as

$$\begin{pmatrix} n\beta + \mathbf{1}^T X b \\ \beta X^T \mathbf{1} + X^T X b \end{pmatrix} = \begin{pmatrix} n\beta + \mathbf{1}^T X b \\ \beta X^T \mathbf{1} + \left[U^T U + \dfrac{X^T \mathbf{1}\mathbf{1}^T X}{n} \right] b \end{pmatrix} = \begin{pmatrix} \sqrt{n}\zeta \\ \zeta \dfrac{X^T\mathbf{1}}{\sqrt{n}} + U^T z \end{pmatrix}. \tag{A.19}$$

So that

$$\zeta = \frac{n\beta + \mathbf{1}^T X b}{\sqrt{n}},$$

and

$$\zeta \frac{X^T\mathbf{1}}{\sqrt{n}} = \left[\frac{n\beta + \mathbf{1}^T X b}{\sqrt{n}} \right] \frac{X^T\mathbf{1}}{\sqrt{n}} = \beta X^T \mathbf{1} + \frac{(\mathbf{1}^T X b)(X^T \mathbf{1})}{n} = \beta X^T \mathbf{1} + \frac{(X^T\mathbf{1})(\mathbf{1}^T X b)}{n} = \beta X^T \mathbf{1} + \frac{X^T \mathbf{1}\mathbf{1}^T X}{n} b.$$

Combining with (A.19), this gives

$$U^T U b = U^T z.$$

Now U is a Cholesky factor of $\widetilde{X}$ from **Proposition 2** and b is an ordinary least-squares solution of (A.12), so

$$U^T U b = \widetilde{X}^T \tilde{y} = U^T \tilde{Q}^T \tilde{y} = U^T \breve{y},$$

by the discussion in Section 3.1. Hence

$$U^T z = U^T \breve{y} \quad \text{or} \quad U^T (z - \breve{y}) = 0,$$

so that z and $\breve{y}$ differ only in the null space of U^T. Moreover, if U is nonsingular,

$$z = U^{-T} U^T z = U^{-T} U^T \breve{y} = \breve{y}.$$

Proposition 3: Let $\mathbf{b} = \begin{pmatrix} \beta \\ b \end{pmatrix}$ be ordinary least-squares regression coefficients for (A.11). Let R in the QR factorization (A.14) be partitioned as in (A.16), and let $\mathcal{Q}^T y$ be partitioned as

$$\mathcal{Q}^T y = \begin{pmatrix} Q^T y \\ Z^T y \end{pmatrix} = \begin{pmatrix} \zeta \\ z \\ Z^T y \end{pmatrix},$$

where ζ is a scalar. Then

$$\beta = \frac{\rho(\zeta - s^T b)}{n}.$$

Proof. From (A.14), we have

$$\mathcal{Q}^T \begin{pmatrix} \mathbf{1} & X \end{pmatrix} = \begin{pmatrix} \mathcal{Q}^T\mathbf{1} & \mathcal{Q}^T X \end{pmatrix} \begin{pmatrix} R \\ O \end{pmatrix}.$$

From the partition (A.16), we have

$$\mathcal{Q}^T\mathbf{1} = \begin{pmatrix} \rho \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \mathcal{Q}^T X = \begin{pmatrix} s^T \\ U \\ O \end{pmatrix}.$$

Combining these results with (A.13), we have

$$\beta \;=\; \frac{1^T(y - Xb)}{n} = \frac{1^T\mathcal{Q}\mathcal{Q}^T(y - Xb)}{n} = \frac{1}{n}\begin{pmatrix} \rho \\ 0 \\ 0 \end{pmatrix}^T \left[\begin{pmatrix} \zeta \\ z \\ Z^T y \end{pmatrix} - \begin{pmatrix} s^T \\ U \\ O \end{pmatrix} b \right] = \frac{\rho(\zeta - s^T b)}{n}.$$