DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 1095

November 9, 2004

# Model Selection and Estimation in Regression with Grouped Variables[1]

Ming Yuan[2] and Yi Lin[3]

Key words: ANOVA, LASSO, LARS, Nonnegative Garrote, Piecewise linear solution path.

[1]This work was supported in part by National Science Foundation grant DMS-0134987.
[2]Email: yuanm@stat.wisc.edu
[3]Email: yilin@stat.wisc.edu

# Model Selection and Estimation in Regression with Grouped Variables[4]

Ming Yuan and Yi Lin

(November 9, 2004)

**Abstract**

We consider the problem of selecting grouped variables (factors) for accurate prediction in regression. Such a problem arises naturally in many practical situations with the multi-factor ANOVA problem as the most important and well known example. Instead of selecting factors by stepwise backward elimination, we focus on estimation accuracy and consider extensions of the LASSO, the LARS, and the nonnegative garrote for factor selection. The LASSO, the LARS, and the nonnegative garrote are recently proposed regression methods that can be used to select individual variables. We study and propose efficient algorithms for the extensions of these methods for factor selection, and show that these extensions give superior performance to the traditional stepwise backward elimination method in factor selection problems. We study the similarities and the differences among these methods. Simulations and real examples are used to illustrate the methods.

**Key words:** ANOVA, LASSO, LARS, Nonnegative Garrote, Piecewise linear solution path.

[4]Ming Yuan is Assistant Professor, School of Industrial and System Engineering, Georgia Institute of Technology, Atlant, GA 30332 (E-mail: yuanm@stat.wisc.edu); and Yi Lin is Associate Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706 (E-mail: yilin@stat.wisc.edu). This work was supported in part by National Science Foundation grant DMS-0134987.

# 1 Introduction

In many regression problems we are interested in finding important explanatory factors in predicting the response variable, where each explanatory factor may be represented by a group of derived input variables. The most common example is the multi-factor ANOVA problem, in which each factor may have several levels and can be expressed through a group of dummy variables. The goal of ANOVA is often to select important main effects and interactions for accurate prediction, which amounts to the selection of groups of derived input variables. Another example is the additive model with polynomial or nonparametric components. In both situations, each component in the additive model may be expressed as a linear combination of a number of basis functions of the original measured variable. In such cases the selection of important measured variables corresponds to the selection of groups of basis functions. In both of these two examples, variable selection typically amounts to the selection of important factors (groups of variables) rather than individual derived variables, as each factor corresponds to one measured variable and is directly related to the measurement cost. In this paper we propose and study several methods that produce accurate prediction while selecting a subset of important factors.

Consider the general regression problem with $J$ factors:

$$Y = \sum_{j=1}^{J} X_j \beta_j + \epsilon, \tag{1.1}$$

where $Y$ is a $n \times 1$ vector, $\epsilon \sim N_n(0, \sigma^2 I)$, $X_j$ is a $n \times p_j$ matrix corresponding to the $j$th factor, and $\beta_j$ is a coefficient vector of size $p_j$, $j = 1, ..., J$. To eliminate the intercept from (1.1), throughout this paper, we center the response variable and each input variable so that the observed mean is zero. To simplify description, we further assume that each $X_j$ is orthonormalized. That is, $X_j' X_j = I_{p_j}$, $j = 1, ..., J$. This can be done through Gram-Schmidt orthonormalization, and different orthonormalizations corresponds to reparametrizing the factor through different orthonormal contrasts. Denoting $X = (X_1, X_2, ..., X_J)$ and $\beta = (\beta_1', ..., \beta_J')'$, equation (1.1) can be written as $Y = X\beta + \epsilon$.

Each of the factors in (1.1) can be categorical or continuous. The traditional ANOVA model is the special case in which all the factors are categorical and the additive model is a special case in which all the factors are continuous. It is clearly possible to include both categorical and continuous factors in (1.1).

Our goal is to select important factors for accurate estimation in (1.1). This amounts to deciding whether to set the vector $\beta_j$ to zero vector for each $j$. In the well studied special case of multi-factor ANOVA model with balanced design, one can construct an ANOVA table for hypothesis testing by partitioning the sums of squares. The columns in the full design matrix $X$ are orthogonal, thus the test results are independent of the order in which the hypotheses are tested. More general cases of (1.1) including the ANOVA problem with unbalanced design are appearing more and more frequently in practice. In such cases the columns of $X$ are no longer orthogonal, and there is no unique partition of the sums of squares. The test result on one factor depends on the presence (or absence) of other factors. Traditional approaches to model selection, such as the best subset selection and the stepwise procedures can be used in model (1.1). In the best subset selection, an estimation accuracy criterion, such as AIC or $C_p$, is evaluated on each candidate model and the model associated with the smallest score is selected as the best model. This is impractical for even moderate number of factors since the number of candidate models grows exponentially as the number of factors increases. The stepwise methods are computationally more attractive, and can be conducted with an estimation accuracy criterion or through hypothesis testing. However, these methods often lead to locally optimal solutions rather than globally optimal solutions.

A commonly considered special case of (1.1) is when $p_1 = ... = p_J = 1$. This is the most studied model selection problem. A number of new model selection methods have been introduced for this problem in recent years (George and McCulloch, 1993; Foster and George, 1994; Breiman, 1995; Tibshirani, 1996; George and Foster, 2000; Fan and Li, 2001; Shen and Ye, 2002; and Efron, Johnston, Hastie and Tibshirani, 2004). In particular, Breiman (1995) showed that the traditional subset selection methods are not satisfactory in terms of prediction accuracy and stability, and proposed the nonnegative garrote which is shown to be more accurate and stable. Tibshirani (1996) proposed the popular LASSO, which is defined as:

$$\hat{\beta}^{LASSO}(\lambda) = \arg\min_{\beta}(\|Y - X\beta\|^2 + \lambda\|\beta\|_{\ell_1}), \tag{1.2}$$

where $\lambda$ is a tuning parameter, and $\|\cdot\|_{\ell_1}$ stands for the vector $\ell_1$ norm. The $\ell_1$ norm penalty induces sparsity in the solution. Efron et. al. (2004) proposed the least angle regression (LARS) and showed that the LARS and the LASSO are closely related. These methods proceed in two steps. First a solution path indexed by certain tuning parameter is built.

Then the final model is selected on the solution path by cross validation or using a criterion such as the $C_p$. As shown in Efron et. al. (2004), the solution paths of the LARS and the LASSO are piecewise linear, and thus can be computed very efficiently. This gives the LARS and the LASSO tremendous computational advantages when compared with other methods. Rosset and Zhu (2004) studied several related piecewise-linear-solution-path algorithms.

While the LASSO and the LARS enjoy great computational advantages and excellent performance, they are designed for selecting individual input variables, not for general factor selection in (1.1). When directly applied to model (1.1), they tend to make selection based on the strength of individual derived input variable rather than the strength of groups of input variables, often resulting in selecting more factors than necessary. Another drawback of using the LASSO and the LARS in (1.1) is that the solution depends on how the factors are orthonormalized. That is, if any factor $X_j$ is reparametrized through a different set of orthonormal contrasts, we may get a different set of factors in the solution. This is undesirable since our solution to a factor selection and estimation problem should not depend on how the factors are represented. In this paper we consider the extensions of the LASSO and the LARS for factor selection in (1.1), which we call group LASSO and group LARS. We show these natural extensions improve over the LASSO and LARS in terms of factor selection, and enjoys superior performance to that of traditional methods for factor selection in model (1.1). We study the relationship between the group LASSO and the group LARS, and show that they are equivalent when the full design matrix $X$ is orthogonal, but can be different in more general situations. In fact, a somewhat surprising result is that the solution path of the group LASSO is generally not piecewise linear while the solution path of the group LARS is. We also consider a group version of the nonnegative garrote. We show that the nonnegative garrote has a piecewise linear solution path, and we propose a new efficient algorithm for computing the nonnegative garrote solution path. We compare these factor selection methods via simulations and a real example.

In order to select the final models on the solution paths of the group selection methods, we introduce an easily computable $C_p$ criterion. The form of the criterion is derived in the special case of orthogonal design matrix, but has a reasonable interpretation in general. Simulations and real examples show that the $C_p$ criterion works very well.

The later sections are organized as follows. We introduce the group LASSO, the group

LARS, and the group nonnegative garrote in Sections 2-4. In Section 5 we consider the connection between the three algorithms. Section 6 is on the selection of tuning parameters. Simulation and a real example are given in Section 7 and 8. A summary and discussions are given in Section 9. Technical proofs are relegated to the appendix.

# 2 Group LASSO

For a vector $\eta \in R^d$, $d \geq 1$, and a symmetric $d$ by $d$ positive definite matrix $K$, we denote

$$||\eta||_K = (\eta' K \eta)^{1/2}.$$

We write $||\eta|| = ||\eta||_{I_d}$ for brevity. Given positive definite matrices $K_1, \ldots, K_J$, the group LASSO estimate is defined as the solution to

$$\frac{1}{2}\left|\left|Y - \sum_{j=1}^{J} X_j \beta_j\right|\right|^2 + \lambda \sum_{j=1}^{J} ||\beta_j||_{K_j}, \tag{2.1}$$

where $\lambda \geq 0$ is a tuning parameter. Bakin (1999) proposed (2.1) as an extension of the LASSO for selecting groups of variables and proposed a computational algorithm. It is clear that (2.1) reduces to the LASSO when $p_1 = ... = p_J = 1$. The penalty function used in (2.1) is intermediate between the $\ell_1$ penalty used in the LASSO and $\ell_2$ penalty used in ridge regression. This is illustrated in Figure 1 in the case that all $K_j$'s are identity matrices. Consider a case in which there are two factors, and the corresponding coefficients are a 2-vector $\beta_1 = (\beta_{11}, \beta_{12})'$ and a scalar $\beta_2$. The top panels of Figure 1 depict the contour of the penalty functions. The leftmost panel corresponds to the $\ell_1$ penalty $|\beta_{11}| + |\beta_{12}| + |\beta_2| = 1$, the central panel corresponds to $||\beta_1|| + |\beta_2| = 1$, and the rightmost panel corresponds to $||(\beta_1', \beta_2)'|| = 1$. The intersections of the contours with planes $\beta_{12} = 0$ (or $\beta_{11} = 0$), $\beta_2 = 0$, and $\beta_{11} = \beta_{12}$, are shown in the next three rows of Figure 1. As shown in Figure 1, the $\ell_1$ penalty treats the three coordinate directions differently from other directions, and this encourages sparsity in individual coefficients. The $\ell_2$ penalty treats all directions equally, and does not encourage sparsity. The group LASSO encourages sparsity at the factor level.

There are many reasonable choices for the kernel matrices $K_j$'s. An obvious choice would be $K_j = I_{p_j}$, $j = 1, ..., J$. In the implementation of the group LASSO in this paper, we choose to set $K_j = p_j I_{p_j}$. Notice that under both choices the solution given by the group LASSO

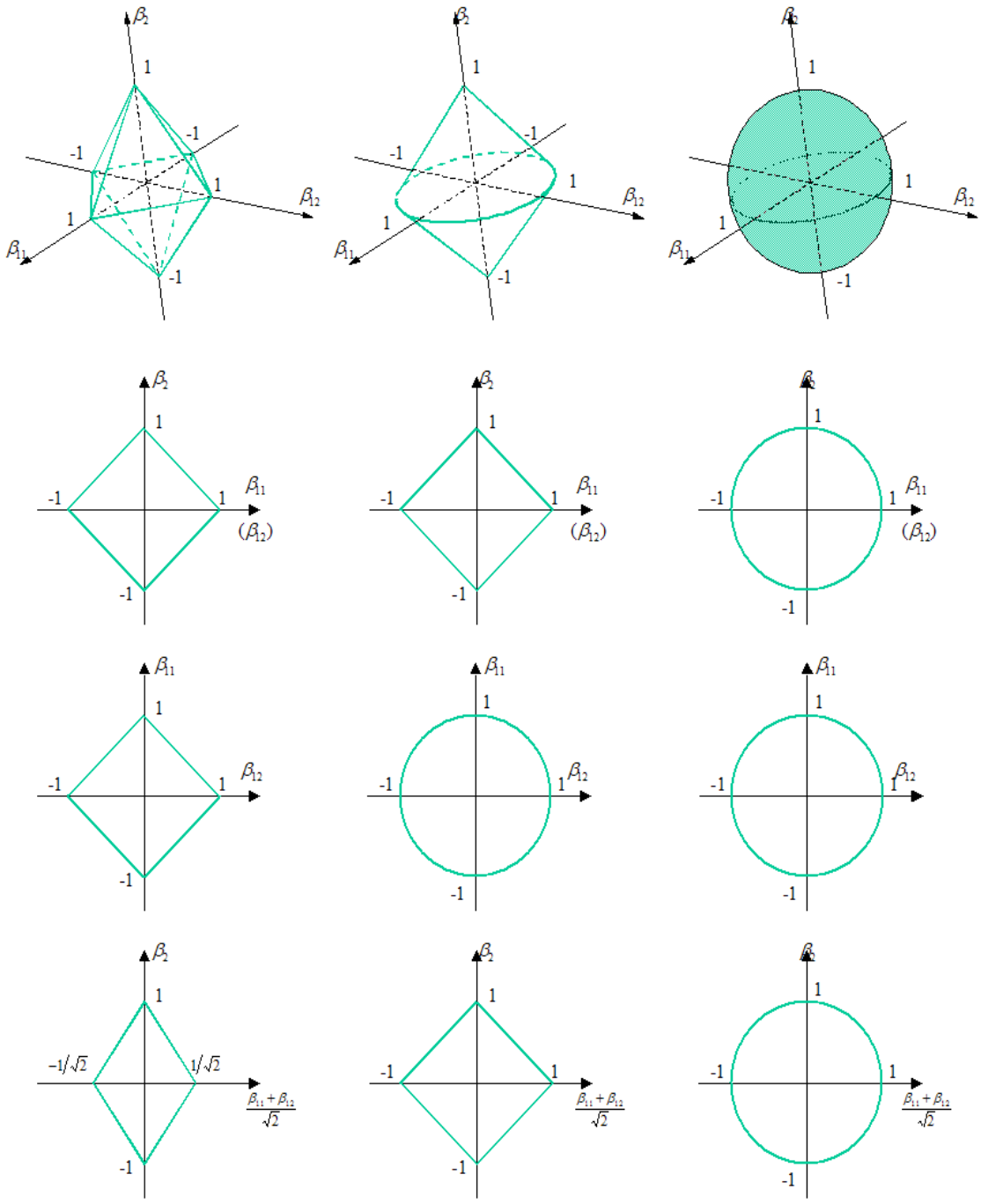


Figure 1: The $\ell_1$ penalty (left panels), Group LASSO penalty (central panels) and $\ell_2$ penalty (right panels)

does not depend on the particular sets of orthonormal contrasts that is used to represent the factors. We prefer the latter since in the ANOVA with balanced design case the resulting solution is similar to the solution given by ANOVA tests. This will become clear in later discussions.

Bakin (1999) proposed a sequential optimization algorithm for (2.1). In this paper, we introduce a more intuitive approach. Our implementation of the group LASSO is an extension of the shooting algorithm (Fu, 1999) for the LASSO. It is motivated by the following proposition, which is a direct consequence of the Karush-Kuhn-Tucker conditions.

**Proposition 2.1** *Let $K_j = p_j I_{p_j}$, $j = 1, ..., J$. A necessary and sufficient condition for $\beta = (\beta_1', ..., \beta_J')'$ to be a solution to (2.1) is*

$$-X_j'(Y - X\beta) + \frac{\lambda\sqrt{p_j}\beta_j}{||\beta_j||} = \mathbf{0} \qquad \forall \beta_j \neq \mathbf{0} \tag{2.2}$$

$$\left|\left|-X_j'(Y - X\beta)\right|\right| \leq \lambda\sqrt{p_j} \qquad \forall \beta_j = \mathbf{0} \tag{2.3}$$

Recall that $X_j'X_j = I_{p_j}$. It can be easily verified that the solution to (2.2) and (2.3) is

$$\beta_j = \left(1 - \frac{\lambda\sqrt{p_j}}{||S_j||}\right)_+ S_j, \tag{2.4}$$

where $S_j = X_j'(Y - X\beta_{-j})$, with $\beta_{-j} = (\beta_1', ..., \beta_{j-1}', \mathbf{0}', \beta_{j+1}', ..., \beta_J')$ . The solution to (2.1) can therefore be obtained by iteratively applying (2.4) to $j = 1, \ldots, J$.

The algorithm is found to be very stable and usually reaches reasonable convergence tolerance within a few iterations. However, the computational burden increases dramatically as the number of predictors increases.

# 3 Group LARS

The LARS (Efron et. al., 2004) was proposed for variable selection in (1.1) with $p_1 = ... = p_J = 1$ and the algorithm can be described roughly as follows. Starting with all coefficients equal to zero, the LARS finds the input variable that is most correlated with the response variable and proceeds on this direction. Instead of taking a full step towards the projection of $Y$ on the variable, as would be done in a greedy algorithm, the LARS only takes the largest step possible in this direction until some other input variable has as much correlation

with the current residual. At this point the projection of the current residual on the space spanned by the two variables has equal angle with the two variables, and the LARS proceeds in this direction until a third variable "earns its way into the most correlated set". The LARS then proceeds in the direction of the projection of the current residual on the space spanned by the three variables, a direction that has equal angle with the three input variables, until a fourth variable enters, etc. The great computational advantage of the LARS comes from the fact that the LARS path is piecewise linear.

When all the factors in (1.1) have the same number of derived input variables ($p_1 = ... = p_J$, though they may not be equal to one), a natural extension of the LARS for factor selection that retains the piecewise linear property of the solution path is the following. Define the angle $\theta(r, X_j)$ between a $n$-vector $r$ and a factor represented by $X_j$ as the angle between the vector $r$ and the space spanned by the column vectors of $X_j$. It is clear that this angle does not depend on the set of orthonormal contrasts representing the factor, and that it is the same as the angle between $r$ and the projection of $r$ in the space spanned by the columns of $X_j$. Therefore $\cos^2(\theta(r, X_j))$ is the proportion of the total variation sum of square in $r$ that is explained by the regression on $X_j$, i.e. the $R^2$ when $r$ is regressed on $X_j$. Since $X_j$ is orthonormal, we have $\cos^2(\theta(r, X_j)) = \|X_j'r\|^2/\|r\|^2$. Starting with all coefficient vectors equal to zero vector, the Group LARS finds the factor (say $X_{j_1}$) that has the smallest angle with $Y$ (i.e. $\|X_{j_1}'Y\|^2$ is the largest), and proceeds in the direction of the projection of $Y$ on the space spanned by the factor until some other factor (say $X_{j_2}$) has as small an angle with the current residual. That is,

$$\|X_{j_1}'r\|^2 = \|X_{j_2}'r\|^2, \tag{3.1}$$

where $r$ is the current residual. At this point the projection of the current residual on the space spanned by the columns of $X_{j_1}$ and $X_{j_2}$ has equal angle with the two factors, and the Group LARS proceeds in this direction. Notice that as the Group LARS marches on, the direction of projection of the residual on the space spanned by the two factors does not change. The Group LARS continues on this direction until a third factor $X_{j_3}$ has the same angle with the current residual as the two factors with the current residual. The Group LARS then proceeds in the direction of the projection of the current residual on the space spanned by the three factors, a direction that has equal angle with the three factors, until a fourth factor enters, etc.

When $p_j$'s are not all equal, some adjustment to the above Group LARS algorithm is needed to take into account of the different number of derived input variables in the groups. Instead of choosing the factors based on the angle of the residual $r$ with the factors $X_j$, or equivalently, on $\|X_j'r\|^2$, we can base the choice on $\|X_j'r\|^2/p_j$. There are other reasonable choices of the scaling, we have taken this particular choice in the implementation in this paper since it gives similar results to the ANOVA test in the special case of ANOVA with balanced design.

To sum up, our group version of the LARS algorithm proceeds in the following way:

**Algorithm – Group LARS**

(1) Start from $\beta^{[0]} = 0$, $k = 1$ and $r^{[0]} = Y$

(2) Compute the current "most correlated set"

$$\mathcal{A}_1 = \arg\max_j ||X_j'r^{[k-1]}||^2/p_j$$

(3) Compute the current direction $\gamma$ which is a $p = \sum p_j$ dimensional vector with $\gamma_{\mathcal{A}_k^c} = 0$ and

$$\gamma_{\mathcal{A}_k} = \left(X_{\mathcal{A}_k}'X_{\mathcal{A}_k}\right)^- X_{\mathcal{A}_k}'r^{[k-1]},$$

where $X_{\mathcal{A}_k}$ denotes the matrix comprised of the columns of $X$ corresponding to $\mathcal{A}_k$.

(4) For every $j \notin \mathcal{A}_k$, compute how far the group LARS will progress in direction $\gamma$ before $X_j$ enters the most correlated set. This can be measured by a $\alpha_j \in [0, 1]$ such that

$$||X_j'(r^{[k-1]} - \alpha_j X\gamma)||^2/p_j = ||X_{j'}'(r^{[k-1]} - \alpha_j X\gamma)||^2/p_{j'}, \tag{3.2}$$

where $j'$ is arbitrarily chosen from $\mathcal{A}_k$.

(5) If $\mathcal{A}_k \neq \{1, \ldots, J\}$, let $\alpha = \min_{j \notin \mathcal{A}_k} \alpha_j \equiv \alpha_{j^*}$ and update $\mathcal{A}_{k+1} = \mathcal{A} \cup \{j^*\}$. Otherwise, set $\alpha = 1$.

(6) Update $\beta^{[k]} = \beta^{[k-1]} + \alpha\gamma$, $r^{[k]} = Y - X\beta^{[k]}$ and $k = k + 1$. Go back to step (3) until $\alpha = 1$.

Note that (3.2) is a quadratic equation of $\alpha_j$ and can be solved easily. Since $j'$ is from the current most correlated set, the left side of (3.2) is less than the right hand side when

$\alpha_j = 0$. On the other hand, by the definition of $\gamma$, the right hand side is 0 when $\alpha_j = 1$. Therefore, at least one of the solutions to (3.2) must lie between 0 and 1. In other words, $\alpha_j$ in Step (4) is always well defined. The algorithm stops after $\alpha = 1$, at which time the residual is orthogonal to the columns of $X$. That is, the solution after the final step is the ordinary least square estimate. With probability one, this is reached in $J$ steps.

# 4 Group Nonnegative Garrote

Another method for variable selection in (1.1) with $p_1 = \ldots = p_J = 1$ is the nonnegative garrote proposed by Breiman (1995). The nonnegative garrote estimate of $\beta_j$ is the least square estimate $\widehat{\beta}_j^{LS}$ scaled by a constant $d_j(\lambda)$ given by

$$d(\lambda) = \arg\min_d \frac{1}{2} ||Y - Zd||^2 + \lambda \sum_{j=1}^{J} d_j \quad \text{subject to} \quad d_j \geq 0, \forall j, \tag{4.1}$$

where $Z = (Z_1, \ldots, Z_J)$ and $Z_j = X_j \widehat{\beta}_j^{LS}$.

The nonnegative garrote can be naturally extended to select factors in (1.1). In this case $\widehat{\beta}_j^{LS}$ is a vector, and we scale every component of vector $\widehat{\beta}_j^{LS}$ by the same constant $d_j(\lambda)$. To take into account the different number of derived variables in the factor, we define $d(\lambda)$ as

$$d(\lambda) = \arg\min_d \frac{1}{2} ||Y - Zd||^2 + \lambda \sum_{j=1}^{J} p_j d_j \quad \text{subject to} \quad d_j \geq 0, \forall j. \tag{4.2}$$

The (group) nonnegative garrote solution path can be constructed by solving the quadratic programming problem (4.2) for all $\lambda'$s, as done in Breiman (1995). We show that the solution path of the nonnegative garrote is piecewise linear, and use this to construct a more efficient algorithm of building the (group) nonnegative garrote solution path. The following algorithm is quite similar to the modified LARS algorithm for the LASSO, with a complicating factor being the nonnegative constraints in (4.2).

**Algorithm – Group Nonnegative Garrote**

(1) Start from $d^{[0]} = 0$, $k = 1$ and $r^{[0]} = Y$

(2) Compute the current active set

$$\mathcal{C}_1 = \arg\max_j Z_j' r^{[k-1]}/p_j$$

(3) Compute the current direction $\gamma$, which is a $p$ dimensional vector defined by $\gamma_{\mathcal{C}_k^c} = 0$ and

$$\gamma_{\mathcal{C}_k} = \left(Z'_{\mathcal{C}_k} Z_{\mathcal{C}_k}\right)^- Z'_{\mathcal{C}_k} r^{[k-1]}$$

(4) For every $j \notin \mathcal{C}_k$, compute how far the group nonnegative garrote will progress in direction $\gamma$ before $X_j$ enters the active set. This can be measured by a $\alpha_j$ such that

$$Z'_j \left(r^{[k-1]} - \alpha_j Z\gamma\right)/p_j = Z'_{j'} \left(r^{[k-1]} - \alpha_j Z\gamma\right)/p_{j'} \tag{4.3}$$

where $j'$ is arbitrarily chosen from $\mathcal{C}_k$.

(5) For every $j \in \mathcal{C}_k$, compute $\alpha_j = \min(\beta_j, 1)$ where $\beta_j = -d_j^{[k-1]}/\gamma_j$, if nonnegative, measures how far the group nonnegative garrote will progress before $d_j$ becomes zero.

(6) If $\alpha_j \leq 0$, $\forall j$ or $\min_{j:\alpha_j>0}\{\alpha_j\} > 1$, set $\alpha = 1$. Otherwise, denote $\alpha = \min_{j:\alpha_j>0}\{\alpha_j\} \equiv \alpha_{j^*}$. Set $d^{[k]} = d^{[k-1]} + \alpha\gamma$. If $j^* \notin \mathcal{C}_k$, update $\mathcal{C}_{k+1} = \mathcal{C}_k \cup \{j^*\}$; else update $\mathcal{C}_{k+1} = \mathcal{C}_k - \{j^*\}$.

(7) Set $r^{[k]} = Y - Zd^{[k]}$ and $k = k + 1$. Go back to step (3) until $\alpha = 1$.

**Theorem 4.1** *Under the "one at a time" condition discussed below, the trajectory of this algorithm coincides with (group) nonnegative garrote solution path.*

The same condition as we assumed in Theorem 4.1, referred to as "one at a time", was used in deriving the connection between the LASSO and the LARS by Efron et. al (2004). With the current notation, the condition states that $j^*$ in Step (6) is uniquely defined. This assumption basically means that (i) the addition occurs only for one factor a time at any stage of the above algorithm; (ii) no factor vanishes at the time of addition; and (iii) no two factors vanishes simultaneously. This is generally true in practice and can always be enforced by slightly perturbing the response. For more detailed discussions, the readers are referred to Efron et. al. (2004).

Since

$$\sum_j Z'_j Y = (\beta^{LS})' X'Y = Y'X(X'X)^{-1}X'Y > 0,$$

we have $\max_j Z'_j r^{[k-1]}/p_j > 0$ in Step (2). A careful examination of the proof of Theorem 4.1 reveals that $\max_j Z'_j r^{[k-1]}/p_j > 0$ is monotonically decreasing as the algorithm progresses

and $\mathcal{C}_k$ maintains the collection of factors which maximize $Z_j' r^{[k-1]}/p_j$. The stopping rule in Step (7) makes sure that the algorithm ends when $\max_j Z_j' r^{[k-1]}/p_j = 0$.

Breiman (1995) conjectured that the models produced by the nonnegative garrote are nested in that the model corresponding to a smaller $\lambda$ always contains the model corresponding to a larger $\lambda$. This amounts to stating that $j^* \in \mathcal{C}_k$ never takes place in Step (6). However, we found this conjecture not true although $j^* \in \mathcal{C}_k$ happens only very rarely in our simulation. A counterexample can be obtained from the authors.

# 5 Similarities and Differences

Efron et. al. (2004) showed that there is a close connection between the LASSO and the LARS, and the LASSO solution can be obtained with a slightly modified LARS algorithm. It is of interest to study whether a similar connection exists between the group versions of these methods. In this section, we compare the Group LASSO, the Group LARS and the group nonnegative garrote, and pinpoint the similarities and differences among these procedures.

We start with the simple special case where the design matrix $X = (X_1, ..., X_J)$ is orthonormal. The ANOVA with balanced design is of this situation. For example, a two-way ANOVA with number of levels $I$ and $J$ can be formulated as (1.1) with $p_1 = I-1$, $p_2 = J-1$, and $p_3 = (I-1)(J-1)$ corresponding to the two main effects and one interaction. The design matrix $X$ would be orthonormal in the balanced design case.

From (2.4), it is easy to see that when $X$ is orthonormal, the group LASSO estimator with tuning parameter $\lambda$ can be given as

$$\widehat{\beta}_j = \left(1 - \frac{\lambda\sqrt{p_j}}{\left|\left|X_j'Y\right|\right|}\right)_+ X_j'Y, \qquad j = 1, ..., J. \tag{5.1}$$

As $\lambda$ descends from $+\infty$ to 0, the group LASSO follows a piecewise linear solution path with change points at $\lambda = ||X_j'Y||/\sqrt{p_j}$, $j = 1, \ldots, J$. It is easy to see that this is identical to the solution path of the group LARS when $X$ is orthonormal. On the other hand, when $X$ is orthonormal, the nonnegative garrote solution is

$$\widehat{\beta}_j = \left(1 - \frac{\lambda p_j}{||X_j'Y||^2}\right)_+ X_j'Y, \tag{5.2}$$

which is different from the solution path of the LASSO or the LARS.

Now we turn to the general case. While the group LARS and the group nonnegative garrote have piecewise linear solution paths, it turns out that in general the solution path of the group LASSO is not piecewise linear.

**Theorem 5.1** *The solution path of the group LASSO is piecewise linear if and only if any group LASSO solution* $\widehat{\beta}$ *can be written as* $\widehat{\beta}_j = c_j \beta_j^{LS}$, $j = 1, \ldots, J$ *for some scalars* $c_1, \ldots, c_J$.

The condition for the group LASSO solution path to be piecewise linear as stated above is clearly satisfied if each group has only one predictor or if $X$ is orthonormal. But in general, this condition is rather restrictive and is seldom met in practice. This precludes the possibility of the fast construction of solution path based on piecewise linearity for the group LASSO. Thus, the group LASSO is computationally more expensive in large scale problems than the group LARS and the group nonnegative garrote, whose solution paths can be built very efficiently by taking advantage of their piecewise linear property.

To illustrate the similarities and differences among the three algorithms, we consider a simple example with 2 covariates $X_1, X_2$ generated from a bivariate normal distribution with $var(X_1) = var(X_2) = 1$ and $cov(X_1, X_2) = 0.5$. The response is then generated as

$$Y = \left(X_1^3 + X_1^2 + X_1\right) + \left(\frac{1}{3}X_2^3 - X_2^2 + \frac{2}{3}X_2\right) + \epsilon,$$

where $\epsilon \sim N(0, 3^2)$. We apply the group LASSO, the group LARS and the group nonnegative garrote to the data. This is done by first centering the input variables and the response variable and orthonormalizing the design matrix corresponding to the same factor, then applying the algorithms given in Sections 2-4, and finally transforming the estimated coefficients back to the original scale. The following plot gives the resulting solution paths. Each line in the plot corresponds to the trajectory of an individual regression coefficient. The path of the estimated coefficients from the same group are represented in the same color.

The x-axis in Figure 2 is the fraction of progress measuring how far the estimate has marched on the solution path. More specifically, for the group LASSO,

$$fraction(\beta) = \frac{\sum_j \sqrt{p_j}||\beta_j||}{\sum_j \sqrt{p_j}||\beta_j^{LS}||}.$$

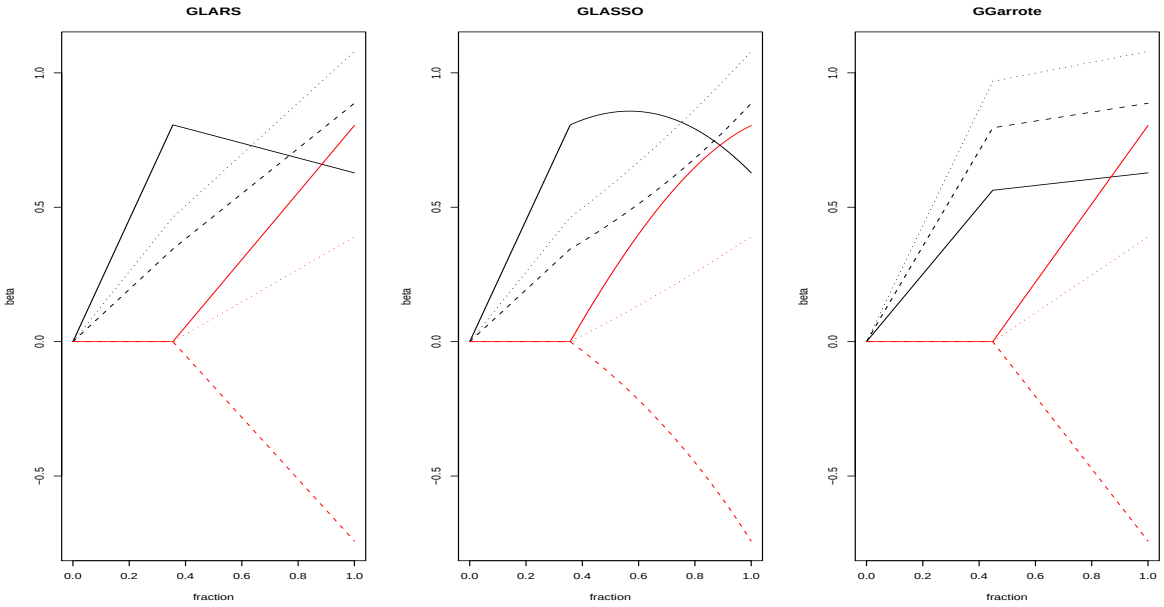


Figure 2: Group LARS (left panel), Group LASSO (central panel) and group nonnegative garrote solution paths (right panel)

For the group nonnegative garrote,

$$fraction(d) = \sum_j p_j d_j / \sum_j p_j.$$

For the group LARS,

$$fraction(\beta) = \frac{\sum_{k=1}^{K} \left( \sum_{j=1}^{J} \sqrt{p_j} ||\beta_j^{[k]} - \beta_j^{[k-1]}|| \right) + \sum_{j=1}^{J} \sqrt{p_j} ||\beta_j - \beta_j^{[K]}||}{\sum_{k=1}^{J} \left( \sum_{j=1}^{J} \sqrt{p_j} ||\beta_j^{[k]} - \beta_j^{[k-1]}|| \right)},$$

where $\beta$ is an estimate between $\beta^{[K]}$ and $\beta^{[K+1]}$. The fraction of progress amounts to a one-to-one map from the solution path to the unit interval $[0, 1]$. Using the fraction introduced above as x-scale, we are able to preserve the piecewise linearity of the group LARS and nonnegative garrote solution paths.

Obvious nonlinearity is noted in the group LASSO solution path. It is also interesting to notice that even though the group LASSO and group LARS are different, their solution paths look quite similar in this example. According to our experience, this is usually true as long as $\max_j p_j$ is not very big.

# 6 Tuning

Once the solution path of the Group LASSO, the Group LARS, or the Group nonnegative garrote is constructed, we choose our final estimate in the solution path according to prediction accuracy, which depends on the unknown parameters and needs to be estimated. In this section we introduce a simple approximate $C_p$ type criterion to select the final estimate on the solution path.

It is well known that in Gaussian regression problems, for an estimate $\widehat{\mu}$ of $\mu = E(Y|X)$, an unbiased estimate of the true risk $E(||\widehat{\mu} - \mu||^2/\sigma^2)$ is

$$C_p(\widehat{\mu}) = \frac{||Y - \widehat{\mu}||^2}{\sigma^2} - n + 2df_{\mu,\sigma^2}, \tag{6.1}$$

where

$$df_{\mu,\sigma^2} = \sum_{i=1}^{n} cov(\widehat{\mu}_i, Y_i)/\sigma^2. \tag{6.2}$$

Since the definition of the degrees of freedom involves the unknowns, in practice, it is often estimated through bootstrap (Efron et. al., 2004) or some data perturbation methods (Shen and Ye, 2002). To reduce the computation cost, Efron et. al. (2004) introduced a simple explicit formula for the degrees of freedom of the LARS which they show is exact in the case of orthonormal design and more generally, when a positive cone condition is satisfied. Here we take the strategy of deriving simple formulas in the special case of orthonormal design, and then test the formulas as approximations in more general case through simulations. The same strategy has also been used in the original LASSO paper (Tibshirani, 1996). We propose the following approximations to $df$. For the group LASSO,

$$\tilde{df}(\widehat{\mu}(\lambda) \equiv X\beta) = \sum_j I(||\beta_j|| > 0) + \sum_j \frac{||\beta_j||}{||\beta_j^{LS}||}(p_j - 1); \tag{6.3}$$

for the group LARS,

$$\tilde{df}(\widehat{\mu}_k \equiv X\beta^{[k]}) = \sum_j I(||\beta_j^{[k]}|| > 0) + \sum_j \left( \frac{\sum_{l<k} ||\beta_j^{[l+1]} - \beta_j^{[l]}||}{\sum_{l<J} ||\beta_j^{[l+1]} - \beta_j^{[l]}||} \right) (p_j - 1); \tag{6.4}$$

and for the nonnegative garrote,

$$\tilde{df}(\widehat{\mu}(\lambda) \equiv Zd) = 2\sum_j I(d_j > 0) + \sum_j d_j(p_j - 2). \tag{6.5}$$

Similar to Efron et. al. (2004), for the group LARS we confine ourselves to the models corresponding to the turning points on the solution path. It is worth noting that if each factor contains only one variable, formula (6.3) reduces to the approximate degrees of freedom given in Efron et. al. (2004).

**Theorem 6.1** *Consider model (1.1) with the design matrix $X$ being orthonormal. For any estimate on the solution path of the group LASSO, the group LARS or the group nonnegative garrote, we have $df = E(\tilde{df})$.*

Empirical evidence suggests that these approximations work fairly well for correlated predictors. In our experience, the performance of this approximate $C_p$ criterion is generally comparable to that of five fold cross validation, and is sometimes better. Notice five fold cross validation is computationally much more expensive.

# 7 Simulation

In this section, we compare the prediction performance of the group LARS, the group LASSO, and the group nonnegative garrote, as well as that of the LARS/LASSO, the ordinary least squares estimate, and the traditional backward stepwise method based on AIC. The backward stepwise method has commonly been used in the selection of grouped variables, with the multi-factor ANOVA as a well known example.

Four models were considered in the simulations. In the first we consider fitting an additive model involving categorical factors. In the second we consider fitting an ANOVA model with all the two way interactions. In the third we fit an additive model of continuous factors. Each continuous factor is represented through a third order polynomial. The last model is an additive model involving both continuous and categorical predictors. Each continuous factor is represented by a third order polynomial.

(I) Fifteen latent variables $Z_1, \ldots, Z_{15}$ were first simulated according to a centered multivariate normal distribution with covariance between $Z_i$ and $Z_j$ being $0.5^{|i-j|}$. Then $Z_i$ is trichotomized as $0, 1, 2$ if it is smaller than $\Phi^{-1}(1/3)$, larger than $\Phi^{-1}(2/3)$ or in between. The response $Y$ was then simulated from

$$Y = 1.8I\,(Z_1 = 1) - 1.2I\,(Z_1 = 0) + I\,(Z_3 = 1) + 0.5I\,(Z_3 = 0) + I\,(Z_5 = 1) + I\,(Z_5 = 0) + \epsilon,$$

where $I(\cdot)$ is the indicator function and the regression noise $\epsilon$ is normally distributed with variance $\sigma^2$ chosen so that the signal to noise ratio is 1.8. 50 observations were collected for each run.

(II) In this example, both main effects and second order interactions were considered. Four categorical factors $Z_1, Z_2, Z_3$ and $Z_4$ were first generated as in (I). The true regression equation is

$$\begin{aligned} Y &= 3I(Z_1 = 1) + 2I(Z_1 = 0) + 3I(Z_2 = 1) + 2I(Z_2 = 0) + I(Z_1 = 1, Z_2 = 1) \\ &\quad +1.5I(Z_1 = 1, Z_2 = 0) + 2I(Z_1 = 0, Z_2 = 1) + 2.5I(Z_1 = 0, Z_2 = 0) + \epsilon, \end{aligned}$$

with signal to noise ratio 3. 100 observations were collected for each simulated dataset.

(III) This example is a more sophisticated version of the example from Section 5. Sixteen random variables $Z_1, \ldots, Z_{16}$ and $W$ were independently generated from a standard normal distribution. The covariates is then defined as $X_i = (Z_i + W)/\sqrt{2}$. The response follows

$$Y = \left(X_3^3 + X_3^2 + X_3\right) + \left(\frac{1}{3}X_6^3 - X_6^2 + \frac{2}{3}X_6\right) + \epsilon,$$

where $\epsilon \sim N(0, 2^2)$. 100 observations were collected for each run.

(IV) Twenty covariates $X_1, \ldots, X_{20}$ were generated in the same fashion as in (III). Then the last ten covariates $X_{11}, \ldots, X_{20}$ were trichotomized as in the first two models. This gives us a total of 10 continuous covariates and 10 categorical covariates. The true regression equation is given by

$$Y = \left(X_3^3 + X_3^2 + X_3\right) + \left(\frac{1}{3}X_6^3 - X_6^2 + \frac{2}{3}X_6\right) + 2I(X_{11} = 0) + I(X_{11} = 1) + \epsilon,$$

where $\epsilon \sim N(0, 2^2)$. For each run, we collected 100 observations.

For each dataset, the group LARS (GLARS), the group LASSO (GLASSO), the group nonnegative garrote (GGarrote), and the LARS (LARS) solution paths were computed. The group LASSO solution path is computed by evaluating on 100 equally spaced $\lambda'$s between 0 and $\max_j ||X_j' Y||/\sqrt{p_j}$. On each solution path, the performance of both the "oracle" estimate which minimizes the true model error defined as

$$ME(\widehat{\beta}) = (\widehat{\beta} - \beta)' E(X'X)(\widehat{\beta} - \beta),$$

and the estimate with tuning parameter chosen by the approximate $C_p$ was recorded. Also reported is the performance of the the full least square estimate and the stepwise method. Only main effects were considered except for the second model where second order interactions are also included. Table 1 summarizes the model error, model sizes in terms of the number of factors (or interaction) selected, and the CPU time consumed for constructing for the solution path. The results reported in Table 1 are averages based on 200 runs. The numbers in parentheses are standard deviations based on the 200 runs.

Several observations can be made from Table 1. In all four examples, the models selected by the LARS are larger than those selected by other methods (other than the full least squares). This is to be expected since the LARS selects individual derived variables, and once a derived variable is included in the model, the corresponding factor is present in the model. Therefore the LARS often produces unnecessarily large models in factor selection problems. The models selected by the stepwise method are smaller than those selected by other methods. The models selected by the group methods are similar in size, though the group nonnegative garrote seems to produce slightly smaller models. The group nonnegative garrote is fastest to compute, followed by the group LARS, the stepwise method, and the LARS. The group LASSO is the slowest to compute.

To compare the performance of the group methods with that of the other methods, we conducted head to head comparisons by performing paired t-tests at 0.05 level. The p-values of the paired t-tests (two sided) are given in Table 2. In all four examples, the group LARS (with $C_p$) and the group LASSO (with $C_p$) perform significantly better than the traditional stepwise method. The group nonnegative garrote performs significantly better than the stepwise method in three of the four examples, but the stepwise method is significantly better than the group nonnegative garrote in Example 2. In Example 3, the difference among the three group methods and the LARS is not significant. In examples 1, 2 and 4, the group LARS and the group LASSO perform significantly better than the LARS. The performance of the group nonnegative garrote and that of the LARS are not significantly different in examples 1, 2 and 3, but the nonnegative garrote significantly outperform the LARS in example 4. We also report in Table 1 the minimal estimation error over the solution paths for each of the group methods. This is only computable in simulations, not real example. It represents the estimation error of the ideal (oracle) estimator which minimizes the true

model error on the solution path, and is a lower bound to the estimation error of any model picked by data adaptive criteria on the solution path.

# 8 Real Example

We re-examine the birthweight dataset from Hosmer and Lemeshow (1989) with the group methods. The birthwt dataset records the birthweights of 189 babies and 8 predictors concerning the mom. Among the eight predictors, 2 are continuous: mother's age in years, mother's weight in pounds at last menstrual period; and 6 are categorical: mother's race (white, black or other), smoking status during pregnancy (yes or no), number of previous premature labors (0, 1 or $\geq$2), history of hypertension (yes or no), presence of uterine irritability (yes or no), number of physician visits during the first trimester (0, 1, 2 or $\geq$3). The data were collected at Baystate Medical Center, Springfield, Massachusetts during 1986. Preliminary analysis suggests that nonlinear effects of both mother's age and weight may exist. To incorporate this into analysis, we model both effects using third order polynomials.

For validation purpose, we randomly selected three quarters of the observations (151 cases) for model fitting, and reserve the rest of the data as the test set. Figure 3 gives the solution paths of the group LARS, the group LASSO, and the group nonnegative garrote. The x-axis is defined as before and the y-axis represents the group score defined as the $\ell_2$ norm of the fitted value for a factor. As Figure 3 shows, the solution paths are quite similar. All these methods suggest that number of physician visits should be excluded from the final model. In addition to this variable, the backward stepwise method excludes two more factors: mother's weight and history of hypertension. The prediction errors of the selected models on the test set are reported Table 3. The group LARS, the group LASSO, and the group nonnegative garrote all perform better than the stepwise method. The performance of the LARS depends on how the categorical factors are represented, therefore the LARS was not included in this study.

# 9 Discussion

The group LARS, the group LASSO, and the group nonnegative garrote are natural extensions of the LARS, the LASSO and the nonnegative garrote. While the LARS, the LASSO

| | GLARS | | GGarrote | | GLASSO | | LARS | | LS | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Oracle | $C_p$ | Oracle | $C_p$ | Oracle | $C_p$ | Oracle | $C_p$ | FULL | STEP |
| | | | | | Model I | | | | | |
| Model Error | 0.83 | 1.31 | 0.99 | 1.79 | 0.82 | 1.31 | 1.17 | 1.72 | 4.72 | 2.39 |
| | (0.4) | (1.06) | (0.62) | (1.34) | (0.38) | (0.95) | (0.47) | (1.17) | (2.28) | (2) |
| Number of factors | 7.79 | 8.32 | 5.41 | 7.63 | 8.48 | 8.78 | 10.14 | 10.44 | 15 | 5.94 |
| | (1.84) | (2.94) | (1.82) | (3.05) | (2.05) | (3.4) | (2.5) | (3.07) | (0) | (2.29) |
| CPU Time | 168.2 | | 97 | | 2007.3 | | 380.8 | | 1.35 | 167.05 |
| (msec) | (19.82) | | (13.6) | | (265.24) | | (40.91) | | (3.43) | (29.9) |
| | | | | | Model II | | | | | |
| Mean ME | 0.09 | 0.11 | 0.13 | 0.17 | 0.09 | 0.12 | 0.13 | 0.17 | 0.36 | 0.15 |
| | (0.04) | (0.05) | (0.08) | (0.13) | (0.04) | (0.07) | (0.05) | (0.11) | (0.14) | (0.13) |
| Mean Size | 5.67 | 5.36 | 5.68 | 5.83 | 6.72 | 6.29 | 8.46 | 8.03 | 10 | 4.15 |
| | (1.16) | (1.62) | (1.81) | (2.12) | (1.42) | (2.03) | (1.09) | (1.39) | (0) | (1.37) |
| CPU Time | 126.85 | | 83.85 | | 2692.25 | | 452 | | 2.1 | 99.85 |
| (msec) | (15.35) | | (12.63) | | (429.56) | | (32.95) | | (4.08) | (21.32) |
| | | | | | Model III | | | | | |
| Mean ME | 1.71 | 2.13 | 1.47 | 2.02 | 1.6 | 2.04 | 1.68 | 2.09 | 7.86 | 2.52 |
| | (0.82) | (1.14) | (0.93) | (2.1) | (0.78) | (1.15) | (0.88) | (1.4) | (3.21) | (2.22) |
| Mean Size | 7.45 | 7.46 | 4.87 | 4.44 | 8.88 | 7.94 | 11.05 | 9.34 | 16 | 4.3 |
| | (1.99) | (2.99) | (1.47) | (3.15) | (2.42) | (3.73) | (2.58) | (3.37) | (0) | (2.11) |
| CPU Time | 124.4 | | 71.9 | | 3364.2 | | 493.2 | | 2.15 | 195 |
| (msec) | (9.06) | | (7.39) | | (562.5) | | (15.78) | | (4.12) | (18.51) |
| | | | | | Model IV | | | | | |
| Mean ME | 1.89 | 2.14 | 1.68 | 2.06 | 1.78 | 2.08 | 1.92 | 2.25 | 6.01 | 2.44 |
| | (0.73) | (0.87) | (0.84) | (1.21) | (0.7) | (0.92) | (0.79) | (0.99) | (2.06) | (1.64) |
| Mean Size | 10.84 | 9.75 | 6.43 | 6.08 | 12.05 | 10.26 | 14.34 | 12.08 | 20 | 5.73 |
| | (2.3) | (3.24) | (1.97) | (3.54) | (2.86) | (3.81) | (2.95) | (3.83) | (0) | (2.26) |
| CPU Time | 159.5 | | 88.4 | | 5265.55 | | 530.6 | | 2.2 | 305.4 |
| (msec) | (8.67) | | (8.47) | | (715.28) | | (30.68) | | (4.15) | (23.87) |

Table 1: Results for the four models considered in the simulation. Reported are the average model error, average number of factors in the selected model, and average computation time, over 200 runs, for the group LARS, the group nonnegative garrote, the group LASSO, the LARS, the full least square estimator, and the stepwise method.

| | Model I | | Model II | | Model III | | Model IV | |
|---|---|---|---|---|---|---|---|---|
| | LARS (Cp) | STEP | LARS (Cp) | STEP | LARS (Cp) | STEP | LARS (Cp) | STEP |
| GLARS (Cp) | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.5829 | 0.0007 | 0.0162 | 0.0017 |
| GGarrote (Cp) | 0.3386 | 0.0000 | 0.5887 | 0.0003 | 0.4717 | 0.0000 | 0.0122 | 0.0000 |
| GLASSO (Cp) | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.3554 | 0.0000 | 0.0001 | 0.0001 |

Table 2: The p-values of the paired t-tests comparing the estimation error of different methods.

| | GLARS (Cp) | GGarrote (Cp) | GLASSO (Cp) | STEP |
|---|---|---|---|---|
| Prediction Error | 609092.8 | 579413.6 | 610008.7 | 646664.1 |

Table 3: The test set prediction error of the models selected by the group LARS, the group nonnegative garrote, the group LASSO, and the stepwise method.

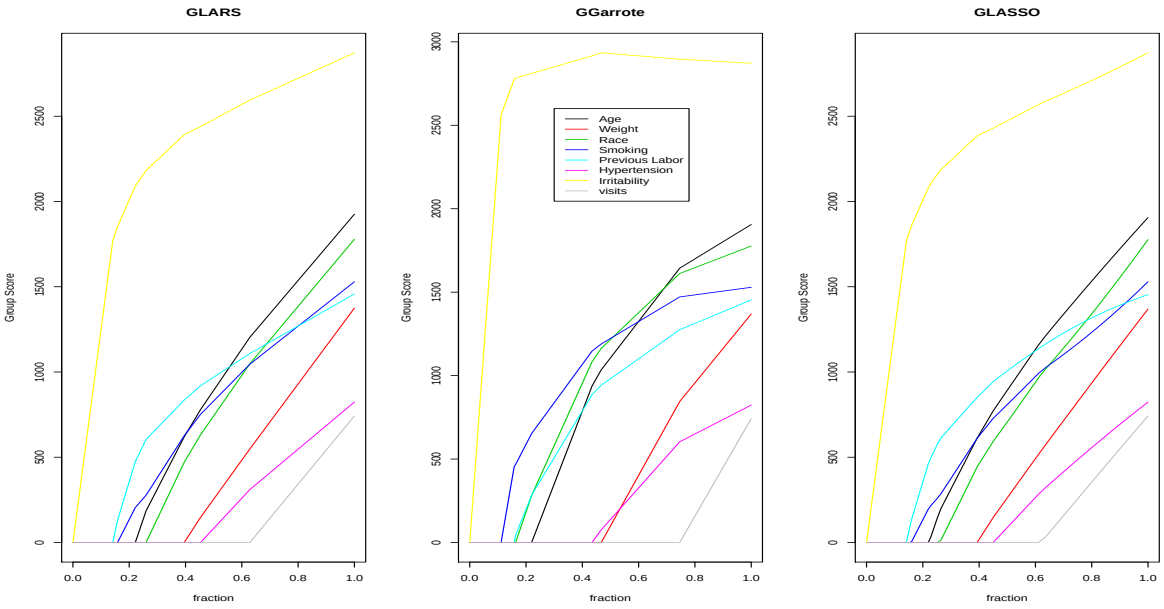


Figure 3: Solution path for the birthwight data

and the nonnegative garrote are very successful in selecting individual variables, their group counterparts are more suitable for factor selection. These new group methods can be used in ANOVA problems with general design, and tend to outperform the traditional stepwise backward elimination method. The group LASSO enjoys excellent performance, but as shown in Section 5, its solution path in general is not piecewise linear and therefore requires intensive computation in large scale problems. The group LARS proposed in Section 3 has comparable performance to that of the group LASSO, and can be computed quickly due to its piecewise linear solution path. The group nonnegative garrote can be computed the fastest among the methods considered in this paper, through a new algorithm taking advantage of the piecewise linearity of its solution. However, due to its explicit dependence on the full least squares estimates, in problems where the sample size is small relative to the total number of variables, the nonnegative garrote may perform suboptimal. In particular, the nonnegative garrote can not be directly applied to problems where the total number of variables exceeds the sample size, whereas the other two group methods can.

# References


Breiman, L. (1995), Better subset regression using the nonnegative garrote, *Technometrics*, **37**, 373-384.

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

Fan, J. and Li, R. (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist. Assoc.*, **96** 1348-1360.

Foster, D. P. and George, E. I. (1994), The risk inflation criterion for multiple regression, *Ann. Statist.*, **22**, 1947-1975.

Fu, W. J. (1999), Penalized regressions: the bridge versus the lasso. *J. Comput. Graph. Statist.*, **7(3)** 397–416.

Bakin, S. (1999) Adaptive Regression and Model Selection in Data Mining Problems.

*unpublished PhD thesis*, Australian National University.

George, E. I. (2000), The variable selection problem, *J. Amer. Statist. Assoc.*, **95**, 1304-1308.

George, E. I. and Foster, D. P. (2000), Calibration and empirical Bayes variable selection, *Biometrika*, **87**, 731-747.

George, E. I. and McCulloch, R. E. (1993), Variable selection via Gibbs sampling, *J. Amer. Statist. Assoc.*, **88**, 881-889.

Hosmer, D.W. and Lemeshow, S. (1989), *Applied Logistic Regression.* New York: Wiley.

Rosset, S. and Zhu, J. (2004), Piecewise Linear Regularized Solution Paths. Technical Report. (available at `http://www-stat.stanford.edu/~saharon/`)

Shen, X. and Ye, J. (2002), Adaptive model selection, *J. Amer. Statist. Assoc.*, **97**, 210-221.

Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *J. Royal. Statist. Soc. B.*, **58**, 267-288.

# APPENDIX

*Proof of Theorem 4.1* Karush-Kuhn-Tucker Theorem suggests that a necessary and sufficient condition for a point $d$ to be on the solution path of (4.2) is that there exists a $\lambda \geq 0$ such that for any $j = 1, \ldots, J$,

$$\{-Z_j'(Y - Zd) + \lambda p_j\}d_j = 0 \tag{A.1}$$

$$-Z_j'(Y - Zd) + \lambda p_j \geq 0 \tag{A.2}$$

$$d_j \geq 0 \tag{A.3}$$

In the following we show that (A.1)-(A.3) are satisfied by any point on the solution path constructed by the algorithm; and any solution to (A.1)-(A.3) for certain $\lambda \geq 0$ is also on the constructed solution path.

We verify (A.1)-(A.3) for the solution path by induction. Obviously, they are satisfied by $d^{[0]}$. Now suppose that they hold for any point prior to $d^{[k]}$. It suffices to show that they are also true for any point between $d^{[k]}$ and $d^{[k+1]}$. There are three possible actions at step $k$: (i) a variable is added to active set: $j^* \notin \mathcal{C}_k$; (ii) a variable is deleted from the active set: $j^* \in \mathcal{C}_k$; and (iii) $\alpha = 1$. It is easy to see that (A.1)-(A.3) will continue to hold for any point between $d^{[k]}$ and $d^{[k+1]}$ if $\alpha = 1$. Now we consider the other two possibilities.

First consider additions. Without loss of generality, assume that $\mathcal{C}_k - \mathcal{C}_{k-1} = \{1\}$. Note that a point between $d^{[k]}$ and $d^{[k+1]}$ can be expressed as $d^\alpha \equiv d^{[k]} + \alpha\gamma$, where $\alpha \in (0, \alpha_1]$ and $\gamma$ is a vector defined by $\gamma_{\mathcal{C}_k^c} = \mathbf{0}$ and

$$\gamma_{\mathcal{C}_k} = (Z'_{\mathcal{C}_k} Z_{\mathcal{C}_k})^{-1} Z'_{\mathcal{C}_k} r^{[k]}. \tag{A.4}$$

It is not hard to show that (A.1) and (A.2) are true for $d^\alpha$. It now suffices to check (A.3). By the construction of the algorithm, it boils down to verify that $\gamma_1 > 0$.

By the definition of $\mathcal{C}_k$ and $\mathcal{C}_{k-1}$, we know that for any $j \in \mathcal{C}_{k-1}$,

$$Z'_j r^{[k-1]}/p_j > Z'_1 r^{[k-1]}/p_1 \tag{A.5}$$

$$Z'_j r^{[k]}/p_j = Z'_1 r^{[k]}/p_1 \tag{A.6}$$

Therefore,

$$Z'_1(r^{[k-1]} - r^{[k]})/p_1 < Z'_j(r^{[k-1]} - r^{[k]})/p_j.$$

Because there exists a positive constant $b$ such that $r^{[k-1]} - r^{[k]} = bZ_{\mathcal{C}_{k-1}}(Z'_{\mathcal{C}_{k-1}} Z_{\mathcal{C}_{k-1}})^{-1} Z'_{\mathcal{C}_{k-1}} r^{[k-1]}$, one concludes that

$$Z'_1 Z_{\mathcal{C}_{k-1}}(Z'_{\mathcal{C}_{k-1}} Z_{\mathcal{C}_{k-1}})^{-1} Z'_{\mathcal{C}_{k-1}} r^{[k-1]}/p_1 < Z'_j Z_{\mathcal{C}_{k-1}}(Z'_{\mathcal{C}_{k-1}} Z_{\mathcal{C}_{k-1}})^{-1} Z'_{\mathcal{C}_{k-1}} r^{[k-1]}/p_j.$$

Write $s = (p_1, \ldots, p_1, p_2, \ldots, p_2, \ldots, p_J, \ldots, p_J)'$. Since $Z'_{\mathcal{C}_{k-1}} r^{[k-1]} = (Z'_j r^{[k-1]}/p_j) s_{\mathcal{C}_{k-1}}$, we have

$$Z'_1 Z_{\mathcal{C}_{k-1}}(Z'_{\mathcal{C}_{k-1}} Z_{\mathcal{C}_{k-1}})^{-1} s_{\mathcal{C}_{k-1}} < p_1. \tag{A.7}$$

Together with (A.4),

$$\gamma_1 = \frac{\{p_1 - Z'_1 Z_{\mathcal{C}_{k-1}} \left(Z'_{\mathcal{C}_{k-1}} Z_{\mathcal{C}_{k-1}}\right)^{-1} s_{\mathcal{C}_{k-1}}\} Z'_j r^{[k]}}{\{Z'_1 Z_1 - Z'_1 Z_{\mathcal{C}_{k-1}} \left(Z'_{\mathcal{C}_{k-1}} Z_{\mathcal{C}_{k-1}}\right)^{-1} Z'_{\mathcal{C}_{k-1}} Z_1\} p_j} > 0, \tag{A.8}$$

Now let us consider the case of deletion. Without loss of generality, assume that $\mathcal{C}_{k-1} - \mathcal{C}_k = \{1\}$. In this case, a point between $d^{[k]}$ and $d^{[k+1]}$ can still be expressed as $d^\alpha \equiv d^{[k]} + \alpha\gamma$,

where $\alpha \in (0, \alpha_1]$ and $\gamma$ is still defined by (A.4). It is readily to show that (A.1) and (A.3) are true with $\lambda = Z_j'(Y - Zd^\alpha)/p_j$ where $j$ is arbitrarily chosen from $\mathcal{C}_k$. It suffices to verify (A.2). By the construction of the solution path, it suffices to show that (A.2) holds for $j = 1$.

Note that any point between $d^{[k-1]}$ and $d^{[k]}$ can be written as $d^{[k-1]} + c\tilde{\gamma}$, where $c > 0$ and $\tilde{\gamma}$ is given by $\tilde{\gamma}_{\mathcal{C}_{k-1}^c} = \mathbf{0}$ and

$$\tilde{\gamma}_{\mathcal{C}_{k-1}} = (Z_{\mathcal{C}_{k-1}}' Z_{\mathcal{C}_{k-1}})^{-1} Z_{\mathcal{C}_{k-1}}' r^{[k-1]}. \tag{A.9}$$

Clearly, $\tilde{\gamma}_1 < 0$. Similar to (A.8), we have

$$\tilde{\gamma_1} = \frac{\{p_1 - Z_1' Z_{\mathcal{C}_k} \left(Z_{\mathcal{C}_k}' Z_{\mathcal{C}_k}\right)^{-1} s_{\mathcal{C}_k}\} Z_j' r^{[k]}}{\{Z_1' Z_1 - Z_1' Z_{\mathcal{C}_k} \left(Z_{\mathcal{C}_k}' Z_{\mathcal{C}_k}\right)^{-1} Z_{\mathcal{C}_k}' Z_1\} p_j}. \tag{A.10}$$

where $j$ is arbitrarily chosen from $\mathcal{C}_k$. Therefore,

$$Z_1' Z_{\mathcal{C}_k} \left(Z_{\mathcal{C}_k}' Z_{\mathcal{C}_k}\right)^{-1} s_{\mathcal{C}_k} = (p_j / Z_j' r^{[k]}) Z_1' Z\gamma < p_1.$$

In other words, $Z_1' Z\gamma / p_1 < Z_j' r^{[k]} / p_j = Z_j' Z\gamma / p_j$. Since $Z_1' r^{[k]} / p_1 = Z_j' r^{[k]} / p_j$, we conclude that $Z_1'(Y - Zd^\alpha)/p_1 < Z_j'(Y - Zd^\alpha)/p_j = \lambda$.

Next, we need to show that for any $\lambda \geq 0$, the solution to (A.1)-(A.3) is on the solution path. By the continuity of the solution path and the uniqueness of the solution to (4.2), it is evident that for any $\lambda \in [0, \max Z_j' Y / p_j]$, the solution to (A.1)-(A.3) is on the path. The proof is now completed by the fact that for any $\lambda > \max Z_j' Y / p_j$, the solution to (A.1)-(A.3) is $\mathbf{0}$ which is also on the solution path. ∎

*Proof of Theorem 5.1* The "if" part is true because in this case, (2.1) is equivalent to the LASSO formulation for $c_j'$s; and the solution path of the LASSO is piecewise linear. The proof of the "only if" part relies on the following lemma.

**Lemma A.1** *Suppose that $\widehat{\beta}$ and $\tilde{\beta}$ are two distinct points on the group LASSO solution path. If any point on the straight line connecting $\widehat{\beta}$ and $\tilde{\beta}$ is also on the group LASSO solution path, then $\widehat{\beta}_j = c_j \tilde{\beta}_j$, $j = 1, \ldots, J$ for some scalars $c_1, \ldots, c_J$.*

Now suppose that the group LASSO solution path is piecewise linear with change points at $\beta^{[0]} = \mathbf{0}, \beta^{[1]}, \ldots, \beta^{[M]} = \beta^{LS}$. Certainly the conclusion of Theorem 5.1 holds for $\beta^{[M]}$. Using Lemma A.1, the proof can then be completed by induction. ∎

*Proof of Lemma A.1* For any estimate $\beta$, define its active set by $\{j : \beta_j \neq \mathbf{0}\}$. Without loss of generality, assume that the active set stays the same for $\alpha\widehat{\beta}+(1-\alpha)\tilde{\beta}$ as $\alpha$ increases from 0 to 1. Denote the set by $\mathcal{E}$. More specifically, for any $\alpha \in [0,1]$,

$$\mathcal{E} = \{j : \alpha\widehat{\beta}_j + (1-\alpha)\tilde{\beta}_j \neq \mathbf{0}\}.$$

Suppose that $\alpha\widehat{\beta} + (1-\alpha)\tilde{\beta}$ is a group LASSO solution with tuning parameter $\lambda_\alpha$. For an arbitrarily $j \in \mathcal{E}$, write

$$C_\alpha = \frac{\lambda_\alpha\sqrt{p_j}}{||\alpha\widehat{\beta}_j + (1-\alpha)\tilde{\beta}_j||}.$$

From (2.2),

$$X_j'\left[Y - X\{\alpha\widehat{\beta} + (1-\alpha)\tilde{\beta}\}\right] = C_\alpha\{\alpha\widehat{\beta}_j + (1-\alpha)\tilde{\beta}_j\}. \tag{A.11}$$

Note that

$$\begin{aligned} X_j'\left[Y - X\{\alpha\widehat{\beta} + (1-\alpha)\tilde{\beta}\}\right] &= \alpha X_j'\left(Y - X\widehat{\beta}\right) + (1-\alpha)X_j'\left(Y - X\tilde{\beta}\right) \\ &= \alpha C_1\widehat{\beta}_j + (1-\alpha)C_0\tilde{\beta}_j. \end{aligned}$$

Therefore, we can re-write (A.11) as

$$\alpha(C_1 - C_\alpha)\widehat{\beta}_j = (1-\alpha)(C_\alpha - C_0)\tilde{\beta}_j \tag{A.12}$$

Assume that the conclusion of Lemma 10.3 is not true. We intend to derive a contradiction by applying (A.12) to two indexes $j_1, j_2 \in \mathcal{E}$ which are defined in the following.

Choose $j_1$ such that $\widehat{\beta}_{j_1} \neq c\tilde{\beta}_{j_1}$ for any scalar $c$. According to (A.12), $C_\alpha$ must be a constant as $\alpha$ varies in $[0,1]$. By the definition of $C_\alpha$, we conclude that $\lambda_\alpha \propto ||\alpha\widehat{\beta}_{j_1} + (1-\alpha)\tilde{\beta}_{j_1}||$. In other words,

$$\lambda_\alpha^2 = \eta||\widehat{\beta}_{j_1} - \tilde{\beta}_{j-1}||^2\alpha^2 + 2\eta(\widehat{\beta}_{j_1} - \tilde{\beta}_{j_1})'\tilde{\beta}_{j_1}\alpha + \eta||\tilde{\beta}_{j_1}||^2 \tag{A.13}$$

for some positive constant $\eta$.

In order to define $j_2$, assume that $\lambda_1 > \lambda_0$ without loss of generality. Then $\sum_j \sqrt{p_j}||\tilde{\beta}_j|| > \sum_j \sqrt{p_j}||\widehat{\beta}_j||$. There exists a $j_2$ such that $\sqrt{p_j}||\tilde{\beta}_{j_2}|| > \sqrt{p_j}||\widehat{\beta}_{j_2}||$. Then for $j_2$, $C_1 > C_0$. Assume that $C_1 - C_\alpha \neq 0$ without loss of generality. By (A.12),

$$\widehat{\beta}_{j_2} = \frac{(1-\alpha)(C_\alpha - C_0)}{\alpha(C_1 - C_\alpha)}\tilde{\beta}_{j_2} \equiv c_{j_2}\tilde{\beta}_{j_2}.$$

Therefore,

$$C_\alpha = \frac{(1-\alpha)C_0 + c_{j_2}\alpha C_1}{1-\alpha+c_{j_2}\alpha} \tag{A.14}$$

Now by definition of $C_\alpha$,

$$\lambda_\alpha = (\alpha C_1 c_{j_2} + (1-\alpha)C_0)||\tilde{\beta}_{j_2}|| \tag{A.15}$$

Combining (A.13) and (A.15), we conclude that

$$\{(\widehat{\beta}_{j_1} - \tilde{\beta}_{j_1})'\tilde{\beta}_{j_1}\}^2 = ||\widehat{\beta}_{j_1} - \tilde{\beta}_{j-1}||^2||\tilde{\beta}_{j_1}||^2,$$

which implies that $\widehat{\beta}_{j_1}/||\widehat{\beta}_{j_1}|| = \tilde{\beta}_{j_1}/||\tilde{\beta}_{j_1}||$. This contradicts our definition of $j_1$. The proof is now completed. ∎

*Proof of Theorem 6.1* Write $\widehat{\beta}_j = (\widehat{\beta}_{j1}, \ldots, \widehat{\beta}_{jp_j})$ and $\beta_j^{LS} = (\beta_{j1}^{LS}, \ldots, \beta_{jp_j}^{LS})'$. For any $\widehat{\beta}$ that depends on $Y$ only through $\beta^{LS}$, Since $X'X = I$, by the chain rule we have

$$\begin{aligned}
tr\left(\frac{\partial \widehat{Y}}{\partial Y}\right) &= tr\{\frac{\partial(X\widehat{\beta})}{\partial Y}\} \\
&= tr\{\frac{\partial(X\widehat{\beta})}{\partial \beta^{LS}}\frac{\partial \beta^{LS}}{\partial Y}\} \\
&= tr\left(X\frac{\partial \widehat{\beta}}{\partial \beta^{LS}}X'\right) \\
&= tr\left(X'X\frac{\partial \widehat{\beta}}{\partial \beta^{LS}}\right) \\
&= tr\left(\frac{\partial \widehat{\beta}}{\partial \beta^{LS}}\right) \\
&= \sum_{j=1}^{J}\sum_{i=1}^{p_j}\left(\frac{\partial \widehat{\beta}_{ji}}{\partial \beta_{ji}^{LS}}\right)
\end{aligned} \tag{A.16}$$

Recall that the group LASSO or the group LARS solution is given by

$$\widehat{\beta}_{ji} = \left(1 - \frac{\lambda\sqrt{p_j}}{||\beta_j^{LS}||}\right)_+ \beta_{ji}^{LS}. \tag{A.17}$$

It implies

$$\frac{\partial \widehat{\beta}_{ji}}{\partial \beta_{ji}^{LS}} = I\left(||\beta_j^{LS}|| > \lambda\sqrt{p_j}\right)\left[1 - \frac{\lambda\sqrt{p_j}\{||\beta_j^{LS}||^2 - \left(\beta_{ji}^{LS}\right\}^2)}{||\beta_j^{LS}||^3}\right]. \tag{A.18}$$

Combining (A.16) and (A.18), we have

$$\begin{aligned}\sum_{l=1}^{n}\frac{\partial\widehat{Y}_l}{\partial Y_l} &= \sum_{j=1}^{J} I\left(||\beta_j^{LS}|| > \lambda\sqrt{p_j}\right)\{p_j - \frac{\lambda\sqrt{p_j}(p_j-1)}{||\beta_j^{LS}||}\} \\ &= \sum_{j=1}^{J} I\left(||\beta_j^{LS}|| > \lambda\sqrt{p_j}\right) + \sum_{j=1}^{J}\left(1-\frac{\lambda\sqrt{p_j}}{||\beta_j^{LS}||}\right)_+(p_j-1) \\ &= \tilde{df}.\end{aligned}$$

Similarly, the nonnegative garrote solution is given as

$$\widehat{\beta}_{ji} = \left(1-\frac{\lambda p_j}{||\beta_j^{LS}||^2}\right)_+ \beta_{ji}^{LS}. \tag{A.19}$$

Therefore,

$$\frac{\partial\widehat{\beta}_{ji}}{\partial\beta_{ji}^{LS}} = I\left(||\beta_j^{LS}|| > \sqrt{\lambda p_j}\right)\left[1-\frac{\lambda p_j\{||\beta_j^{LS}||^2 - 2\left(\beta_{ji}^{LS}\right)^2\}}{||\beta_j^{LS}||^4}\right]. \tag{A.20}$$

As a result of (A.16) and (A.20),

$$\begin{aligned}\sum_{l=1}^{n}\frac{\partial\widehat{Y}_l}{\partial Y_l} &= \sum_{j=1}^{J} I\left(||\beta_j^{LS}|| > \sqrt{\lambda p_j}\right)\{p_j - \frac{\lambda p_j(p_j-2)}{||\beta_j^{LS}||^2}\} \\ &= 2\sum_{j=1}^{J} I\left(||\beta_j^{LS}|| > \sqrt{\lambda p_j}\right) + \sum_{j=1}^{J}\left(1-\frac{\lambda p_j}{||\beta_j^{LS}||^2}\right)_+(p_j-2) \\ &= \tilde{df},\end{aligned}$$

where the last equality holds because $d_j = \left(1-\lambda p_j/||\beta_j^{LS}||^2\right)_+$.

Now, an application of Stein's identity yields

$$df = \sum_{l=1}^{n} cov(\widehat{Y}_l, Y_l)/\sigma^2 = E\left[\sum_{l=1}^{n}\frac{\partial\widehat{Y}_l}{\partial Y_l}\right] = E\left[\tilde{df}\right]. \blacksquare$$