J. R. Statist. Soc. B (2006)
68, *Part* 1, *pp.* 49–67

Model selection and estimation in regression with grouped variables

Ming Yuan

Georgia Institute of Technology, Atlanta, USA

and Yi Lin

University of Wisconsin—Madison, USA

[Received November 2004. Revised August 2005]

Summary. 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 multifactor analysis-of-variance problem as the most important and well-known example. Instead of selecting factors by stepwise backward elimination, we focus on the accuracy of estimation and consider extensions of the lasso, the LARS algorithm and the non-negative garrotte for factor selection. The lasso, the LARS algorithm and the non-negative garrotte 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 between these methods. Simulations and real examples are used to illustrate the methods.

Keywords: Analysis of variance; Lasso; Least angle regression; Non-negative garrotte; Piecewise linear solution path

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 multifactor analysis-of-variance (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 cost of measurement. In this paper we propose and study several methods that produce accurate prediction while selecting a subset of important factors.

Address for correspondence: Ming Yuan, School of Industrial and Systems Engineering, Georgia Institute of Technology, 755 First Drive NW, Atlanta, GA 30332, USA.
E-mail: myuan@isye.gatech.edu

 1369–7412/06/68049

Consider the general regression problem with J factors:

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

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

Each of the factors in equation (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 equation (1.1).

Our goal is to select important factors for accurate estimation in equation (1.1). This amounts to deciding whether to set the vector β_j to zero vectors for each j. In the well-studied special case of multifactor ANOVA models with balanced design, we 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 equation (1.1) including the ANOVA problem with unbalanced design are appearing increasingly 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 stepwise procedures, can be used in model (1.1). In best subset selection, an estimation accuracy criterion, such as the Akaike information criterion or C_p, is evaluated on each candidate model and the model that is associated with the smallest score is selected as the best model. This is impractical for even moderate numbers 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 equation (1.1) is when $p_1 = \ldots = p_J = 1$. This is the most studied model selection problem. Several 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; Efron *et al.*, 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 non-negative garrotte which is shown to be more accurate and stable. Tibshirani (1996) proposed the popular lasso, which is defined as

$$\hat{\beta}^{\mathrm{LASSO}}(\lambda) = \arg\min_{\beta} (\|Y - X\beta\|^2 + \lambda \|\beta\|_{l_1}), \tag{1.2}$$

where λ is a tuning parameter and $\|\cdot\|_{l_1}$ stands for the vector l_1-norm. The l_1-norm penalty induces sparsity in the solution. Efron *et al.* (2004) proposed least angle regression selection (LARS) and showed that LARS and the lasso are closely related. These methods proceed in two steps. First a solution path that is indexed by a certain tuning parameter is built. Then the

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

Although the lasso and LARS enjoy great computational advantages and excellent performance, they are designed for selecting individual input variables, not for general factor selection in equation (1.1). When directly applied to model (1.1), they tend to make selection based on the strength of individual derived input variables rather than the strength of groups of input variables, often resulting in selecting more factors than necessary. Another drawback of using the lasso and LARS in equation (1.1) is that the solution depends on how the factors are orthonormalized, i.e. if any factor X_j is reparameterized through a different set of orthonormal contrasts, we may obtain 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 extensions of the lasso and LARS for factor selection in equation (1.1), which we call the group lasso and group LARS. We show that these natural extensions improve over the lasso and LARS in terms of factor selection and enjoy superior performance to that of traditional methods for factor selection in model (1.1). We study the relationship between the group lasso and 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 whereas the solution path of group LARS is. Also considered is a group version of the non-negative garrotte. We compare these factor selection methods via simulations and a real example.

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 an 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, group LARS and the group non-negative garrotte 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 Sections 7 and 8. A summary and discussions are given in Section 9. Technical proofs are relegated to Appendix A.

2. Group lasso

For a vector $\eta \in R^d$, $d \geqslant 1$, and a symmetric $d \times d$ 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

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

where $\lambda \geqslant 0$ is a tuning parameter. Bakin (1999) proposed expression (2.1) as an extension of the lasso for selecting groups of variables and proposed a computational algorithm. A similar formulation was adopted by Lin and Zhang (2003) where X_j and K_j were chosen respectively to

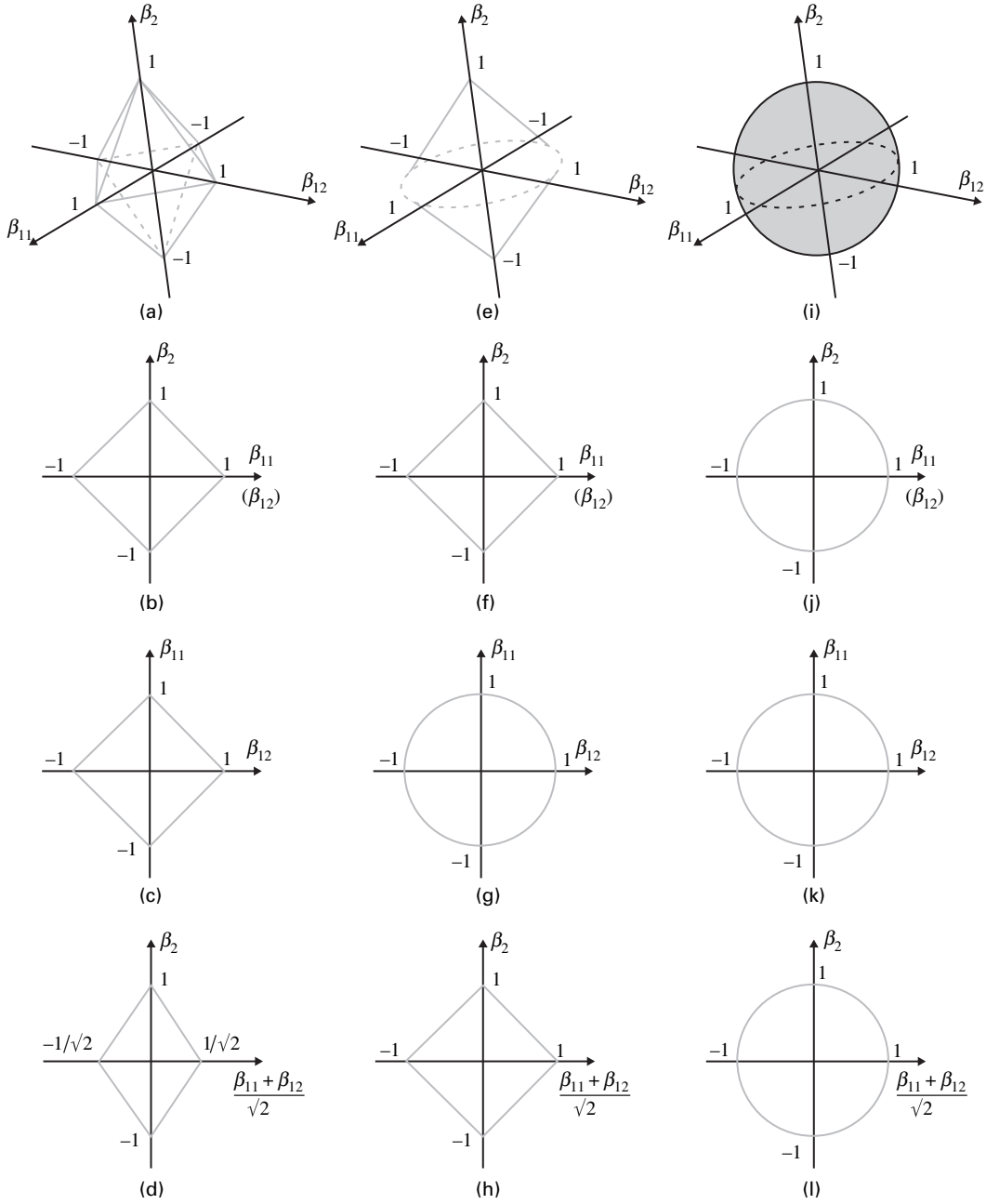

Fig. 1. (a)–(d) l_1-penalty, (e)–(h) group lasso penalty and (i)–(l) l_2-penalty

be basis functions and the reproducing kernel of the functional space induced by the jth factor. It is clear that expression (2.1) reduces to the lasso when $p_1 = \ldots = p_J = 1$. The penalty function that is used in expression (2.1) is intermediate between the l_1-penalty that is used in the lasso and the l_2-penalty that is used in ridge regression. This is illustrated in Fig. 1 in the case that all K_js 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 β_2. Figs 1(a), 1(e) and 1(i) depict the contour of the penalty functions. Fig. 1(a) corresponds to the l_1-penalty $|\beta_{11}| + |\beta_{12}| + |\beta_2| = 1$, Fig. 1(e) corresponds to $\|\beta_1\| + |\beta_2| = 1$ and Fig. 1(i) 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 Figs 1(b)–1(d), 1(f)–1(h) and 1(j)–1(l). As shown in Fig. 1, the l_1-penalty treats the three co-ordinate directions differently from other directions, and this encourages sparsity in individual coefficients. The l_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_js. An obvious choice would be $K_j = I_{p_j}$, $j = 1, \ldots, J$. In the implementation of the group lasso in this paper, we choose to set $K_j = p_j I_{p_j}$. Note that under both choices the solution that is given by the group lasso does not depend on the particular sets of orthonormal contrasts that are 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 that is given by ANOVA tests. This will become clear in later discussions.

Bakin (1999) proposed a sequential optimization algorithm for expression (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 1. Let $K_j = p_j I_{p_j}$, $j = 1, \ldots, J$. A necessary and sufficient condition for $\beta = (\beta_1', \ldots, \beta_J')'$ to be a solution to expression (2.1) is

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

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

Recall that $X_j' X_j = I_{p_j}$. It can be easily verified that the solution to expressions (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 $\beta_{-j} = (\beta_1', \ldots, \beta_{j-1}', \mathbf{0}', \beta_{j+1}', \ldots, \beta_J')$. The solution to expression (2.1) can therefore be obtained by iteratively applying equation (2.4) to $j = 1, \ldots, J$.

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

3. Group least angle regression selection

LARS (Efron *et al.*, 2004) was proposed for variable selection in equation (1.1) with $p_1 = \ldots = p_J = 1$ and the algorithm can be described roughly as follows. Starting with all coefficients equal to 0, the LARS algorithm 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 algorithm takes only the largest step that is 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 that is spanned by the two variables has equal angle with the two variables, and the LARS algorithm proceeds in this direction until a third variable 'earns its way into the most correlated set'. The LARS algorithm then proceeds in the direction of the projection of the current residual on the

space that is 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 algorithm comes from the fact that the LARS path is piecewise linear.

When all the factors in equation (1.1) have the same number of derived input variables ($p_1 = \ldots = p_J$, though they may not be equal to 1), a natural extension of 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 an n-vector r and a factor that is represented by X_j as the angle between the vector r and the space that is 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 that is spanned by the columns of X_j. Therefore $\cos^2\{\theta(r, X_j)\}$ is the proportion of the total variation sum of squares 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 the zero vector, 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 that is spanned by the factor until some other factor (say X_{j_2}) has as small an angle with the current residual, i.e.

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

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

When the p_js are not all equal, some adjustment to the above group LARS algorithm is needed to take into account the different number of derived input variables in the groups. Instead of choosing the factors on the basis of 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 a balanced design.

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

Step 1: start from $\beta^{[0]} = 0$, $k = 1$ and $r^{[0]} = Y$.

Step 2: compute the current 'most correlated set'

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

Step 3: compute the current direction γ which is a $p = \Sigma p_j$ dimensional vector with $\gamma_{\mathcal{A}_k^c} = 0$ and

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

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

Step 4: for every $j \notin \mathcal{A}_k$, compute how far the group LARS algorithm will progress in direction γ before X_j enters the most correlated set. This can be measured by an $\alpha_j \in [0, 1]$ 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 $\mathcal{A}_k$.

Step 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 $\alpha = 1$.

Step 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 $\alpha = 1$.

Equation (3.2) is a quadratic equation of α_j and can be solved easily. Since j' is from the current most correlated set, the left-hand side of equation (3.2) is less than the right-hand side when $\alpha_j = 0$. However, by the definition of γ, the right-hand side is 0 when $\alpha_j = 1$. Therefore, at least one of the solutions to equation (3.2) must lie between 0 and 1. In other words, α_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, i.e. the solution after the final step is the ordinary least square estimate. With probability 1, this is reached in J steps.

4. Group non-negative garrotte

Another method for variable selection in equation (1.1) with $p_1 = \ldots = p_J = 1$ is the non-negative garrotte that was proposed by Breiman (1995). The non-negative garrotte estimate of β_j is the least square estimate $\hat{\beta}_j^{\mathrm{LS}}$ scaled by a constant $d_j(\lambda)$ given by

$$d(\lambda) = \arg\min_d \left(\tfrac{1}{2}\|Y - Zd\|^2 + \lambda \sum_{j=1}^{J} d_j \right) \qquad \text{subject to } d_j \geqslant 0, \forall j, \tag{4.1}$$

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

The non-negative garrotte can be naturally extended to select factors in equation (1.1). In this case $\hat{\beta}_j^{\mathrm{LS}}$ is a vector, and we scale every component of vector $\hat{\beta}_j^{\mathrm{LS}}$ 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

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

The (group) non-negative garrotte solution path can be constructed by solving the quadratic programming problem (4.2) for all λs, as was done in Breiman (1995). It can be shown (see Yuan and Lin (2005)) that the solution path of the non-negative garrotte is piecewise linear, and this can be used to construct a more efficient algorithm for building the (group) non-negative garrotte solution path. The algorithm is quite similar to the modified LARS algorithm for the lasso, with a complicating factor being the non-negativity constraints in equation (4.2).

Step 1: start from $d^{[0]} = 0$, $k = 1$ and $r^{[0]} = Y$.

Step 2: compute the current active set

$$\mathcal{C}_1 = \arg\max_j (Z'_j r^{[k-1]} / p_j).$$

Step 3: compute the current direction γ, which is a p-dimensional vector defined by $\gamma_{\mathcal{C}_k^c} = 0$ and

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

Step 4: for every $j \notin \mathcal{C}_k$, compute how far the group non-negative garrotte will progress in direction γ before X_j enters the active set. This can be measured by an α_j such that

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

where j' is arbitrarily chosen from $\mathcal{C}_k$.
Step 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 non-negative, measures how far the group non-negative garrotte will progress before d_j becomes 0.
Step 6: if $\alpha_j \leqslant 0$, $\forall j$, 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 $\mathcal{C}_{k+1} = \mathcal{C}_k \cup \{j^*\}$; otherwise update $\mathcal{C}_{k+1} = \mathcal{C}_k - \{j^*\}$.
Step 7: set $r^{[k]} = Y - Zd^{[k]}$ and $k = k+1$. Go back to step 3 until $\alpha = 1$.

5. Similarities and differences

Efron *et al.* (2004) showed that there is a close connection between the lasso and 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, group LARS and the group non-negative garrotte, and we pin-point the similarities and differences between these procedures.

We start with the simple special case where the design matrix $X = (X_1, \ldots, X_J)$ is orthonormal. The ANOVA with balanced design problem is of this situation. For example, a two-way ANOVA with number of levels I and J can be formulated as equation (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 equation (2.4), it is easy to see that, when X is orthonormal, the group lasso estimator with tuning parameter λ can be given as

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

As λ descends from ∞ to 0, the group lasso follows a piecewise linear solution path with changepoints 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 group LARS when X is orthonormal. In contrast, when X is orthonormal, the non-negative garrotte solution is

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

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

Now we turn to the general case. Whereas group LARS and the group non-negative garrotte have piecewise linear solution paths, it turns out that in general the solution path of the group lasso is not piecewise linear.

Theorem 1. The solution path of the group lasso is piecewise linear if and only if any group lasso solution $\hat{\beta}$ can be written as $\hat{\beta}_j = c_j\beta_j^{\mathrm{LS}}$, $j = 1, \ldots, J$, 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 paths based on piecewise linearity for the group lasso. Thus, the group lasso is computationally more expensive in large scale problems than group LARS and

the group non-negative garrotte, whose solution paths can be built very efficiently by taking advantage of their piecewise linear property.

To illustrate the similarities and differences between the three algorithms, we consider a simple example with two covariates X_1 and X_2 that are generated from a bivariate normal distribution with $\mathrm{var}(X_1) = \mathrm{var}(X_2) = 1$ and $\mathrm{cov}(X_1, X_2) = 0.5$. The response is then generated as

$$Y = X_1^3 + X_1^2 + X_1 + \tfrac{1}{3}X_2^3 - X_2^2 + \tfrac{2}{3}X_2 + \varepsilon,$$

where $\varepsilon \sim N(0, 3^2)$. We apply the group lasso, group LARS and the group non-negative garrotte to the data. This is done by first centring the input variables and the response variable and orthonormalizing the design matrix corresponding to the same factor, then applying the algorithms that are given in Sections 2–4, and finally transforming the estimated coefficients back to the original scale. Fig. 2 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 for linear, quadratic and cubic terms are represented by full, broken and dotted lines respectively.

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

$$\text{fraction}(\beta) = \sum_j \|\beta_j\| \sqrt{p_j} \Big/ \sum_j \|\beta_j^{\mathrm{LS}}\| \sqrt{p_j}.$$

For the group non-negative garrotte,

$$\text{fraction}(d) = \sum_j p_j d_j \Big/ \sum_j p_j.$$

For group LARS,

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

where β 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 that was introduced above as the x-scale, we can preserve the piecewise linearity of the group LARS and non-negative garrotte solution paths.

Obvious non-linearity is noted in the group lasso solution path. It is also interesting 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, group LARS or the group non-negative garrotte has been constructed, we choose our final estimate in the solution path according to the accuracy of prediction, 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.

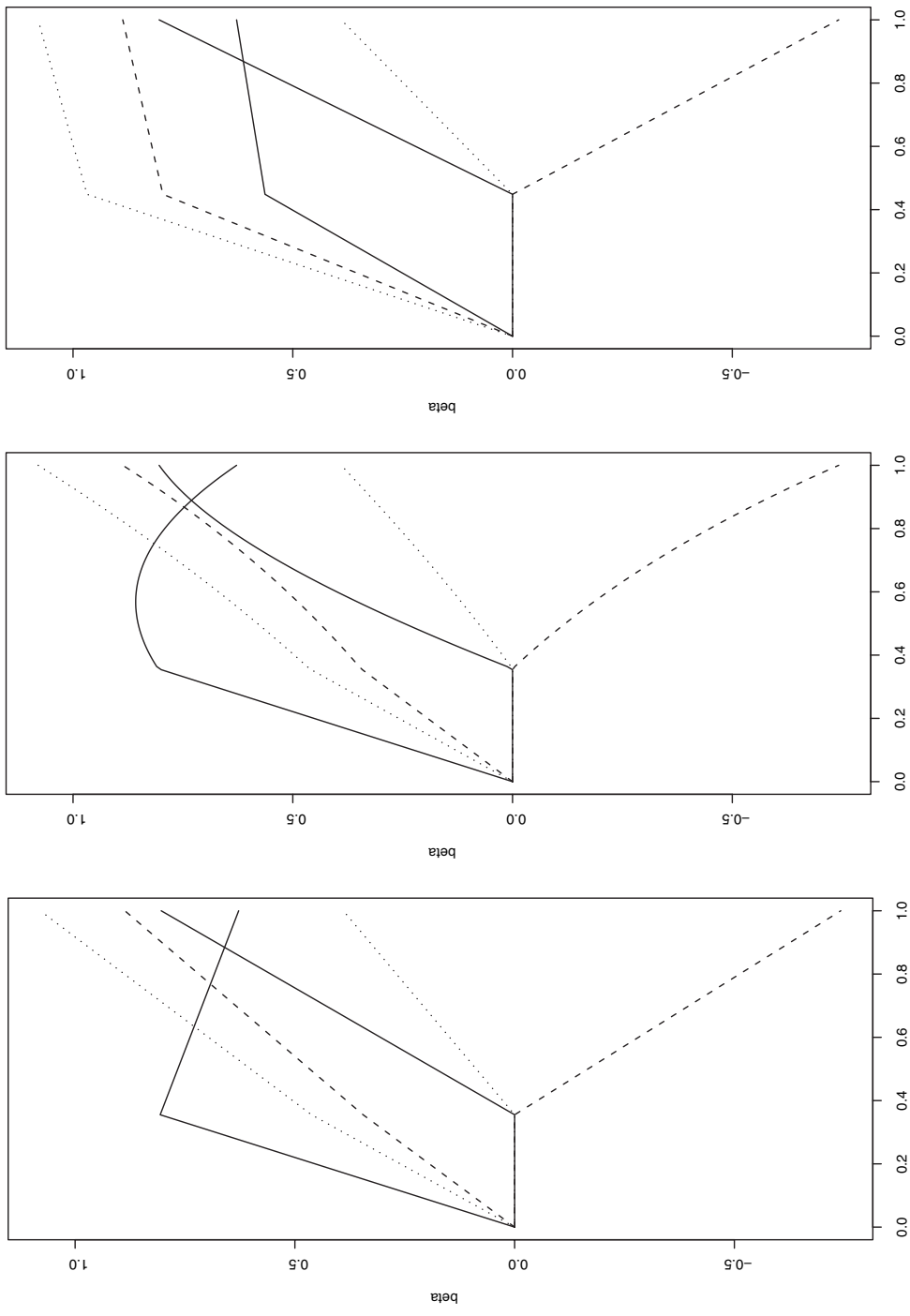

Fig. 2. (a) Group LARS, (b) group lasso and (c) group non-negative garrotte solutions

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

$$C_p(\hat{\mu}) = \frac{\|Y - \hat{\mu}\|^2}{\sigma^2} - n + 2\,\mathrm{df}_{\mu,\sigma^2}, \tag{6.1}$$

where

$$\mathrm{df}_{\mu,\sigma^2} = \sum_{i=1}^{n} \mathrm{cov}(\hat{\mu}_i, Y_i)/\sigma^2. \tag{6.2}$$

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

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

for group LARS,

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

and, for the non-negative garrotte,

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

Similarly to Efron *et al.* (2004), for 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 that were given in Efron *et al.* (2004).

Theorem 2. Consider model (1.1) with the design matrix X being orthonormal. For any estimate on the solution path of the group lasso, group LARS or the group non-negative garrotte, we have $\mathrm{df} = E(\widetilde{\mathrm{df}})$.

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 with that of fivefold cross-validation and is sometimes better. Fivefold cross-validation is computationally much more expensive.

7. Simulation

In this section, we compare the prediction performance of group LARS, the group lasso and the group non-negative garrotte, as well as that of LARS and the lasso, the ordinary least squares estimate and the traditional backward stepwise method based on the Akaike information

criterion. The backward stepwise method has commonly been used in the selection of grouped variables, with multifactor 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.

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

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

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

(b) In model II, 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 model I. The true regression equation is

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

with signal-to-noise ratio 3. 100 observations were collected for each simulated data set.

(c) Model III is a more sophisticated version of the example from Section 5. 17 random variables $Z_1, \ldots, Z_{16}$ and W were independently generated from a standard normal distribution. The covariates are then defined as $X_i = (Z_i + W)/\sqrt{2}$. The response follows

$$Y = X_3^3 + X_3^2 + X_3 + \tfrac{1}{3}X_6^3 - X_6^2 + \tfrac{2}{3}X_6 + \varepsilon,$$

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

(d) In model IV covariates $X_1, \ldots, X_{20}$ were generated in the same fashion as in model III. Then the last 10 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

$$Y = X_3^3 + X_3^2 + X_3 + \tfrac{1}{3}X_6^3 - X_6^2 + \tfrac{2}{3}X_6 + 2\, I(X_{11} = 0) + I(X_{11} = 1) + \varepsilon,$$

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

For each data set, the group LARS, the group lasso, the group non-negative garrotte and the LARS solution paths were computed. The group lasso solution path is computed by evaluating on 100 equally spaced λs 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

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

and the estimate with tuning parameter that is chosen by the approximate C_p was recorded. It is worth pointing out that the oracle estimate, $\arg\min\{\mathrm{ME}(\hat{\beta})\}$, is only computable in simulations, not real examples. Also reported is the performance of the full least squares estimate and the stepwise method. Only main effects were considered except for the second model where

Table 1. Results for the four models that were considered in the simulation†

	Results for the following methods:									
	Group LARS		*Group garrotte*		*Group lasso*		*LARS*		*Least squares*	
	Oracle	C_p	*Oracle*	C_p	*Oracle*	C_p	*Oracle*	C_p	*Full*	*Stepwise*
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 (ms)	168.2		97		2007.3		380.8		1.35	167.05
	(19.82)		(13.6)		(265.24)		(40.91)		(3.43)	(29.9)
Model II										
Model error	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)
Number of factors	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 (ms)	126.85		83.85		2692.25		452		2.1	99.85
	(15.35)		(12.63)		(429.56)		(32.95)		(4.08)	(21.32)
Model III										
Model error	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)
Number of factors	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 (ms)	124.4		71.9		3364.2		493.2		2.15	195
	(9.06)		(7.39)		(562.5)		(15.78)		(4.12)	(18.51)
Model IV										
Model error	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)
Number of factors	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 (ms)	159.5		88.4		5265.55		530.6		2.2	305.4
	(8.67)		(8.47)		(715.28)		(30.68)		(4.15)	(23.87)

†Reported are the average model error, average number of factors in the selected model and average central processor unit (CPU) computation time, over 200 runs, for group LARS, the group non-negative garrotte, the group lasso, LARS, the full least squares estimator and the stepwise method.

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 central processor unit time consumed for constructing the solution path. The results that are 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 that were selected by LARS are larger than those selected by other methods (other than the full least squares). This is to be expected since LARS selects individual derived variables and, once a derived variable has been included in the model, the corresponding factor is present in the model. Therefore LARS often produces unnecessarily large models in factor selection problems. The models that are selected by the stepwise method are smaller than those selected by

Table 2. p-values of the paired t-tests comparing the estimation error of the various methods

Method	*p-values for the following models:*							
	Model I		*Model II*		*Model III*		*Model IV*	
	LARS (C_p)	*Stepwise*	*LARS* (C_p)	*Stepwise*	*LARS* (C_p)	*Stepwise*	*LARS* (C_p)	*Stepwise*
Group LARS (C_p)	0.0000	0.0000	0.0000	0.0000	0.5829	0.0007	0.0162	0.0017
Group garrotte (C_p)	0.3386	0.0000	0.5887	0.0003	0.4717	0.0000	0.0122	0.0000
Group lasso (C_p)	0.0000	0.0000	0.0000	0.0001	0.3554	0.0000	0.0001	0.0001

other methods. The models that are selected by the group methods are similar in size, though the group non-negative garrotte seems to produce slightly smaller models. The group non-negative garrotte is fastest to compute, followed by group LARS, the stepwise method and 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 on the model errors that were obtained for the 200 runs. The p-values of the paired t-tests (two sided) are given in Table 2. In all four examples, group LARS (with C_p) and the group lasso (with C_p) perform significantly better than the traditional stepwise method. The group non-negative garrotte performs significantly better than the stepwise method in three of the four examples, but the stepwise method is significantly better than the group non-negative garrotte in example 2. In example 3, the difference between the three group methods and LARS is not significant. In examples 1, 2 and 4, group LARS and the group lasso perform significantly better than LARS. The performance of the group non-negative garrotte and that of LARS are not significantly different in examples 1, 2 and 3, but the non-negative garrotte significantly outperforms LARS in example 4. We also report in Table 1 the minimal estimation error over the solution paths for each of the group methods. It represents the estimation error of the oracle estimate and is a lower bound to the estimation error of any model that is picked by data-adaptive criteria on the solution path.

8. Real example

We re-examine the birth weight data set from Hosmer and Lemeshow (1989) with the group methods. The birth weight data set records the birth weights of 189 babies and eight predictors concerning the mother. Among the eight predictors, two are continuous (mother's age in years and mother's weight in pounds at the last menstrual period) and six are categorical (mother's race (white, black or other), smoking status during pregnancy (yes or no), number of previous premature labours (0, 1 or 2 or more), 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 3 or more)). The data were collected at Baystate Medical Center, Springfield, Massachusetts, during 1986.

A preliminary analysis suggests that non-linear effects of both mother's age and weight may exist. To incorporate these into the analysis, we model both effects by using third-order polynomials.

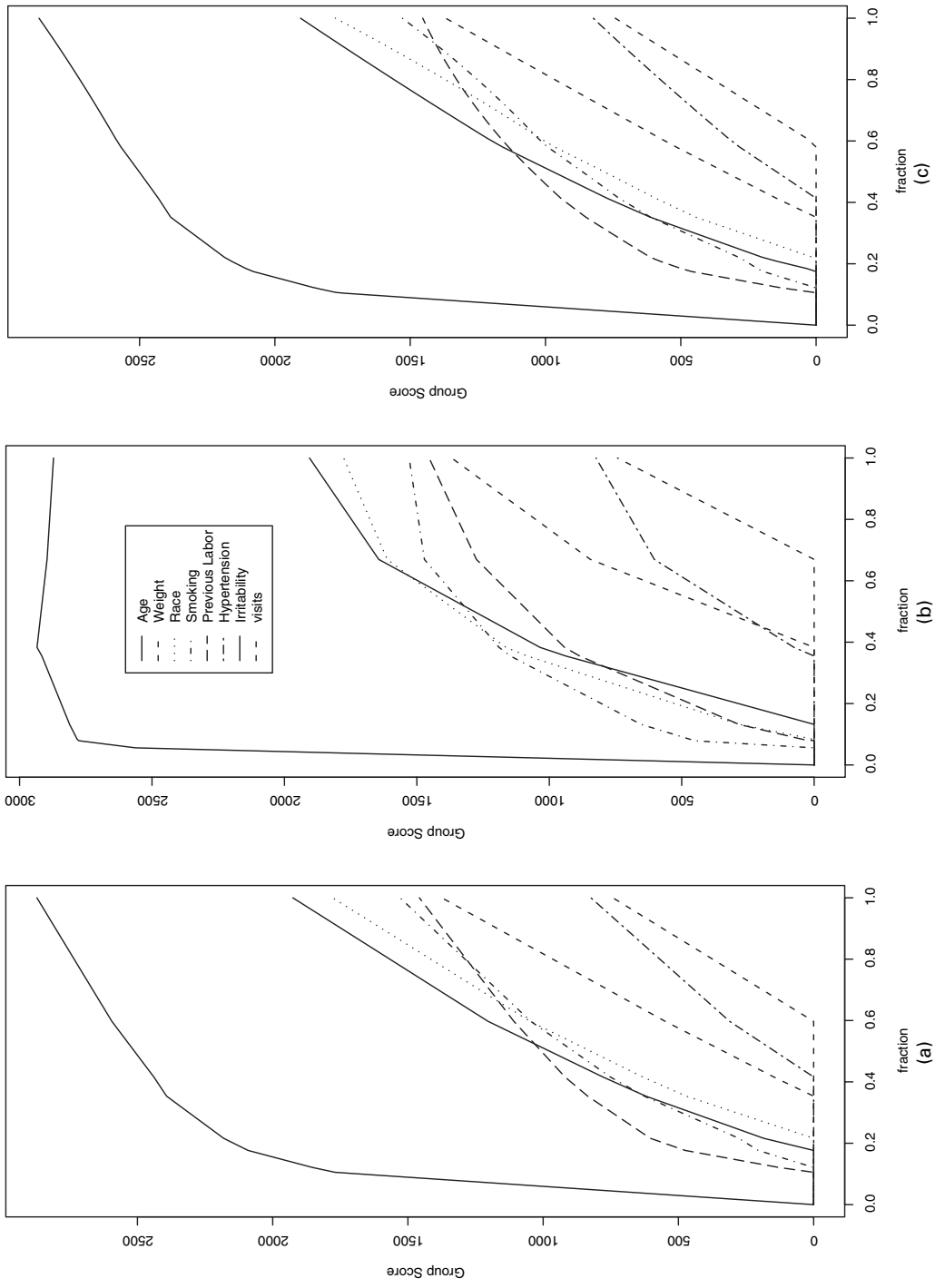

Fig. 3. Solution paths for the birth weight data: (a) group LARS; (b) group garrotte; (c) group lasso

Table 3. Test set prediction error of the models selected by group LARS, the group non-negative garrotte, the group lasso and the stepwise method

Method	*Prediction error*
Group LARS (C_p)	609092.8
Group garrotte (C_p)	579413.6
Group lasso (C_p)	610008.7
Stepwise	646664.1

For validation, we randomly selected three-quarters of the observations (151 cases) for model fitting and reserve the rest of the data as the test set. Fig. 3 gives the solution paths of group LARS, the group lasso and the group non-negative garrotte. The x-axis is defined as before and the y-axis represents the group score defined as the l_2-norm of the fitted value for a factor. As Fig. 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 in Table 3. Group LARS, the group lasso and the group non-negative garrotte all perform better than the stepwise method. The performance LARS depends on how the categorical factors are represented, and therefore LARS was not included in this study.

9. Discussion

Group LARS, the group lasso and the group non-negative garrotte are natural extensions of LARS, the lasso and the non-negative garrotte. Whereas LARS, the lasso and the non-negative garrotte 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 method that was proposed in Section 3 has comparable performance with that of the group lasso and can be computed quickly owing to its piecewise linear solution path. The group non-negative garrotte can be computed the fastest among the methods that are considered in this paper, through a new algorithm taking advantage of the piecewise linearity of its solution. However, owing 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 non-negative garrotte may perform suboptimally. In particular, the non-negative garrotte cannot be directly applied to problems where the total number of variables exceeds the sample size, whereas the other two group methods can.

Acknowledgement

Lin's research was supported in part by National Science Foundation grant DMS-0134987.

Appendix A

A.1. Proof of theorem 1

The 'if' part of theorem 1 is true because, in this case, expression (2.1) is equivalent to the lasso formulation for c_js, and the solution path of the lasso is piecewise linear. The proof of the 'only if' part relies on the following lemma.

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

Now suppose that the group lasso solution path is piecewise linear with changepoints at $\beta^{[0]} = \mathbf{0}$, $\beta^{[1]}, \ldots, \beta^{[M]} = \beta^{\mathrm{LS}}$. Certainly the conclusion of theorem 1 holds for $\beta^{[M]}$. Using lemma 1, the proof can then be completed by induction.

A.2. Proof of lemma 1

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

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

Suppose that $\alpha\hat{\beta} + (1-\alpha)\tilde{\beta}$ is a group lasso solution with tuning parameter λ_α. For an arbitrary $j \in \mathcal{E}$, write

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

From equation (2.2),

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

Note that

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

Therefore, we can rewrite equation (A.1) as

$$\alpha(C_1 - C_\alpha)\hat{\beta}_j = (1-\alpha)(C_\alpha - C_0)\tilde{\beta}_j. \tag{A.2}$$

Assume that the conclusion of lemma 1 is not true. We intend to derive a contradiction by applying equation (A.2) to two indices $j_1, j_2 \in \mathcal{E}$ which are defined in the following.

Choose j_1 such that $\hat{\beta}_{j_1} \neq c\tilde{\beta}_{j_1}$ for any scalar c. According to equation (A.2), C_α must be a constant as α varies in [0,1]. By the definition of C_α, we conclude that

$$\lambda_\alpha \propto \|\alpha\hat{\beta}_{j_1} + (1-\alpha)\tilde{\beta}_{j_1}\|.$$

In other words,

$$\lambda_\alpha^2 = \eta\|\hat{\beta}_{j_1} - \tilde{\beta}_{j-1}\|^2\alpha^2 + 2\eta(\hat{\beta}_{j_1} - \tilde{\beta}_{j_1})'\tilde{\beta}_{j_1}\alpha + \eta\|\tilde{\beta}_{j_1}\|^2 \tag{A.3}$$

for some positive constant η.

To define j_2, assume that $\lambda_1 > \lambda_0$ without loss of generality. Then $\Sigma_j\|\tilde{\beta}_j\|\sqrt{p_j} > \Sigma_j\|\hat{\beta}_j\|\sqrt{p_j}$. There is a j_2 such that $\|\tilde{\beta}_{j_2}\|\sqrt{p_j} > \|\hat{\beta}_{j_2}\|\sqrt{p_j}$. Then, for j_2, $C_1 > C_0$. Assume that $C_1 - C_\alpha \neq 0$ without loss of

generality. By equation (A.2),

$$\begin{aligned}\hat{\beta}_{j_2} &= \frac{(1-\alpha)(C_\alpha - C_0)}{\alpha(C_1 - C_\alpha)}\tilde{\beta}_{j_2} \\ &\equiv c_{j_2}\tilde{\beta}_{j_2}.\end{aligned}$$

Therefore,

$$C_\alpha = \frac{(1-\alpha)C_0 + c_{j_2}\alpha C_1}{1-\alpha+c_{j_2}\alpha}. \tag{A.4}$$

Now, by definition of C_α,

$$\lambda_\alpha = \{\alpha C_1 c_{j_2} + (1-\alpha)C_0\}\|\tilde{\beta}_{j_2}\|. \tag{A.5}$$

Combining equations (A.3) and (A.5), we conclude that

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

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

A.3. Proof of theorem 2

Write $\hat{\beta}_j = (\hat{\beta}_{j1}, \ldots, \hat{\beta}_{jp_j})$ and $\beta_j^{\mathrm{LS}} = (\beta_{j1}^{\mathrm{LS}}, \ldots, \beta_{jp_j}^{\mathrm{LS}})'$. For any $\hat{\beta}$ that depends on Y only through β^{LS}, since $X'X = I$, by the chain rule we have

$$\begin{aligned}\mathrm{tr}\left(\frac{\partial\hat{Y}}{\partial Y}\right) &= \mathrm{tr}\left\{\frac{\partial(X\hat{\beta})}{\partial Y}\right\} \\ &= \mathrm{tr}\left\{\frac{\partial(X\hat{\beta})}{\partial\beta^{\mathrm{LS}}}\frac{\partial\beta^{\mathrm{LS}}}{\partial Y}\right\} \\ &= \mathrm{tr}\left(X\frac{\partial\hat{\beta}}{\partial\beta^{\mathrm{LS}}}X'\right) \\ &= \mathrm{tr}\left(X'X\frac{\partial\hat{\beta}}{\partial\beta^{\mathrm{LS}}}\right) \\ &= \mathrm{tr}\left(\frac{\partial\hat{\beta}}{\partial\beta^{\mathrm{LS}}}\right) \\ &= \sum_{j=1}^{J}\sum_{i=1}^{p_j}\left(\frac{\partial\hat{\beta}_{ji}}{\partial\beta_{ji}^{\mathrm{LS}}}\right).\end{aligned} \tag{A.6}$$

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

$$\hat{\beta}_{ji} = \left(1 - \frac{\lambda\sqrt{p_j}}{\|\beta_j^{\mathrm{LS}}\|}\right)_+\beta_{ji}^{\mathrm{LS}}. \tag{A.7}$$

It implies that

$$\frac{\partial\hat{\beta}_{ji}}{\partial\beta_{ji}^{\mathrm{LS}}} = I(\|\beta_j^{\mathrm{LS}}\| > \lambda\sqrt{p_j})\left[1 - \frac{\lambda\{\|\beta_j^{\mathrm{LS}}\|^2 - (\beta_{ji}^{\mathrm{LS}})^2\}\sqrt{p_j}}{\|\beta_j^{\mathrm{LS}}\|^3}\right]. \tag{A.8}$$

Combining equations (A.6) and (A.8), we have

$$\sum_{l=1}^{n}\frac{\partial\hat{Y}_l}{\partial Y_l}=\sum_{j=1}^{J}I(\|\beta_j^{\mathrm{LS}}\|>\lambda\sqrt{p_j})\left\{p_j-\frac{\lambda(p_j-1)\sqrt{p_j}}{\|\beta_j^{\mathrm{LS}}\|}\right\}$$
$$=\sum_{j=1}^{J}I(\|\beta_j^{\mathrm{LS}}\|>\lambda\sqrt{p_j})+\sum_{j=1}^{J}\left(1-\frac{\lambda\sqrt{p_j}}{\|\beta_j^{\mathrm{LS}}\|}\right)_+(p_j-1).$$
$$=\widetilde{\mathrm{df}}.$$

Similarly, the non-negative garrotte solution is given as

$$\hat{\beta}_{ji}=\left(1-\frac{\lambda p_j}{\|\beta_j^{\mathrm{LS}}\|^2}\right)_+\beta_{ji}^{\mathrm{LS}}. \tag{A.9}$$

Therefore,

$$\frac{\partial\hat{\beta}_{ji}}{\partial\beta_{ji}^{\mathrm{LS}}}=I\{\|\beta_j^{\mathrm{LS}}\|>\sqrt{(\lambda p_j)}\}\left[1-\frac{\lambda p_j\{\|\beta_j^{\mathrm{LS}}\|^2-2(\beta_{ji}^{\mathrm{LS}})^2\}}{\|\beta_j^{\mathrm{LS}}\|^4}\right]. \tag{A.10}$$

As a result of equations (A.6) and (A.10),

$$\sum_{l=1}^{n}\frac{\partial\hat{Y}_l}{\partial Y_l}=\sum_{j=1}^{J}I\{\|\beta_j^{\mathrm{LS}}\|>\sqrt{(\lambda p_j)}\}\left\{p_j-\frac{\lambda p_j(p_j-2)}{\|\beta_j^{\mathrm{LS}}\|^2}\right\}$$
$$=2\sum_{j=1}^{J}I\{\|\beta_j^{\mathrm{LS}}\|>\sqrt{(\lambda p_j)}\}+\sum_{j=1}^{J}\left(1-\frac{\lambda p_j}{\|\beta_j^{\mathrm{LS}}\|^2}\right)_+(p_j-2)$$
$$=\widetilde{\mathrm{df}},$$

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

Now, an application of Stein's identity yields

$$\mathrm{df}=\sum_{l=1}^{n}\mathrm{cov}(\hat{Y}_l,Y_l)/\sigma^2$$
$$=E\left(\sum_{l=1}^{n}\frac{\partial\hat{Y}_l}{\partial Y_l}\right)=E(\widetilde{\mathrm{df}}).$$

References

Bakin, S. (1999) Adaptive regression and model selection in data mining problems. *PhD Thesis*. Australian National University, Canberra.

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. Am. Statist. Ass.*, **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**, 397–416.

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. Am. Statist. Ass.*, **88**, 881–889.

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

Lin, Y. and Zhang, H. H. (2003) Component selection and smoothing in smoothing spline analysis of variance models. *Technical Report 1072*. Department of Statistics, University of Wisconsin, Madison. (Available from `http://www.stat.wisc.edu~yilin/`.)

Rosset, S. and Zhu, J. (2004) Piecewise linear regularized solution paths. *Technical Report*. Stanford University, Stanford. (Available from `http://www-stat.stanford.edu~saharon/`.)

Shen, X. and Ye, J. (2002) Adaptive model selection. *J. Am. Statist. Ass.*, **97**, 210–221.

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

Yuan, M. and Lin, Y. (2005) On the nonnegative garrote estimate. *Statistics Discussion Paper 2005-25*. School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta. (Available from `http://www.isye.gatech.edu~myuan/`.)