Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR

Howard D. Bondell and Brian J. Reich

Department of Statistics, North Carolina State University,

Raleigh, NC 27695-8203, U.S.A.

April 6, 2007

Summary. Variable selection can be challenging, particularly in situations with a large number of predictors with possibly high correlations, such as gene expression data. In this paper, a new method called the OSCAR (Octagonal Shrinkage and Clustering Algorithm for Regression) is proposed to simultaneously select variables while grouping them into predictive clusters. In addition to improving prediction accuracy and interpretation, these resulting groups can then be investigated further to discover what contributes to the group having a similar behavior. The technique is based on penalized least squares with a geometrically intuitive penalty function that shrinks some coefficients to exactly zero. Additionally, this penalty yields exact equality of some coefficients, encouraging correlated predictors that have a similar effect on the response to form predictive clusters represented by a single coefficient. The proposed procedure is shown to compare favorably to the existing shrinkage and variable selection techniques in terms of both prediction error and model complexity, while yielding the additional grouping information.

Key words: Correlation; Penalization; Predictive group; Regression; Shrinkage; Supervised clustering; Variable selection.

email: bondell@stat.ncsu.edu

1. Introduction

Variable selection for regression models with many covariates is a challenging problem that permeates many disciplines. Selecting a subset of covariates for a model is particularly difficult if there are groups of highly-correlated covariates. As a motivating example, consider a recent study of the association between soil composition and forest diversity in the Appalachian Mountains of North Carolina. For this study, there are 15 soil characteristics potentially to be used as predictors, of which there are seven that are highly correlated. Based on a sample of 20 forest plots, the goal is to identify the important soil characteristics.

Penalized regression has emerged as a highly-successful technique for variable selection. For example, the LASSO (Tibshirani, 1996) imposes a bound on the L_1 norm of the coefficients. This results in both shrinkage and variable selection due to the nature of the constraint region which often results in several coefficients becoming identically zero. However, a major stumbling block for the LASSO is that if there are groups of highly-correlated variables, it tends to arbitrarily select only one from each group. These models are difficult to interpret because covariates that are strongly-associated with the outcome are not included in the predictive model.

Supervised clustering, or determining meaningful groups of predictors that form predictive clusters, can be beneficial in both prediction and interpretation. In the soil data, several of the highly correlated predictors are related to the same underlying factor, the abundance of positively charged ions, and hence can be combined into a group. However, just combining them beforehand can dilute the group's overall signal, as not all of them may be related to the response in the same manner. As another example, consider a gene expression study in which several genes sharing a common pathway may be combined to form a grouped predictor. For the classification problem in which the goal is to discriminate between categories, Jörnsten and Yu (2003) and Dettling and Bühlmann (2004) perform supervised gene clustering along with subject classification. These techniques are based on creating a new predictor which is just the average of the grouped predictors, called a 'super gene' for gene expression data by Park et al. (2006). This form of clustering aids in prediction as the process of averaging reduces the variance. It also suggests a possible structure among the predictor variables that can be further investigated.

For a continuous response, Hastie et al. (2001) and Park et al. (2006) first perform hierarchical clustering on the predictors, and, for each level of the hierarchy, take the cluster averages as the new set of potential predictors for the regression. After clustering, the response is then used to select a subset of these candidate grouped predictors via either stepwise selection or using the LASSO.

An alternative and equivalent view of creating new predictors from the group averages is to consider each predictor in a group as being assigned identical regression coefficients. This paper takes this alternative point of view which allows supervised clustering to be directly incorporated into the estimation procedure via a novel penalization method. The new method called the OSCAR (Octagonal Shrinkage and Clustering Algorithm for Regression) performs variable selection for regressions with many highly-correlated predictors. The OSCAR simultaneously eliminates extraneous variables and performs supervised clustering on the important variables.

Other penalized regression methods have been proposed for grouped predictors (Tibshirani et al., 2005; Yuan and Lin, 2006; and Zou and Yuan; 2006), however all of these methods presuppose the grouping structure, e.g., the number of groups or the corresponding sizes. The OSCAR uses a new type of penalty region which is octagonal in shape and requires no initial information regarding the grouping structure. The nature of the penalty region encourages both sparsity and equality of coefficients for correlated predictors having similar relationships with the response. The exact equality of coefficients obtained via this penalty creates grouped predictors as in the supervised clustering techniques. These predictive clusters can then be investigated further to discover what contributes to the group having a similar behavior. Hence, the procedure can also be used as an exploratory tool in a data analysis. Often this structure can be explained by an underlying characteristic, as in the soil example where a group of variables are all related to the abundance of positively charged ions.

The remainder of the paper is organized as follows. Section 2, formulates the OSCAR as a constrained least squares problem and the geometric interpretation of this constraint region is discussed. Computational issues, including choosing the tuning parameters, are discussed in Section 3. Section 4, shows that the OSCAR compares favorably to the existing shrinkage and variable selection techniques in terms of both prediction error and reduced model complexity. Finally, the OSCAR is applied to the soil data in Section 5.

2. The OSCAR

2.1 *Formulation*

Consider the usual linear regression model with observed data on n observations and p predictor variables. Let $\mathbf{y} = (y_1, ..., y_n)^T$ be the vector of responses and $\mathbf{x}_j = (x_{1j}, ..., x_{nj})^T$ denote the j^{th} predictor, $j = 1, ..., p$. Assume that the response has been centered and each predictor has been standardized so that

$$\textstyle\sum_{i=1}^n y_i = 0, \quad \sum_{i=1}^n x_{ij} = 0 \quad \text{and} \quad \sum_{i=1}^n x_{ij}^2 = 1 \quad \text{for all } j = 1, ..., p.$$

Since the response is centered, the intercept is omitted from the model.

As with previous approaches, the OSCAR is constructed via a constrained least squares problem. The choice of constraint used here is on a weighted combination of the L_1 norm and a pairwise L_∞ norm for the coefficients. Specifically, the constrained least squares optimization problem for the OSCAR is given by

$$\begin{gathered}\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \textstyle\sum_{j=1}^{p} \beta_j \mathbf{x}_j||^2 \\ \text{subject to} \\ \textstyle\sum_{j=1}^{p} |\beta_j| + c \ \sum_{j<k} \max\{|\beta_j|, |\beta_k|\} \leq t,\end{gathered} \tag{1}$$

where $c \geq 0$ and $t > 0$ are tuning constants with c controlling the relative weighting of the norms and t controlling the magnitude. The L_1 norm encourages sparseness, while the pairwise L_∞ norm encourages equality of coefficients. Overall, the OSCAR optimization formulation encourages a parsimonious solution in terms of the number of unique non-zero coefficients. Although the correlations between predictors does not directly appear in the penalty term, it is shown both graphically and later in Theorem 1 that the OSCAR implicitly encourages grouping of highly correlated predictors.

While given mathematically by (1), the form of the constrained optimization problem is directly motivated more from the geometric interpretation of the constraint region, rather than the penalty itself. This geometric interpretation of the constrained least squares solutions illustrates how this penalty simultaneously encourages sparsity and grouping. Aside from a constant, the contours of the sum-of-squares loss function,

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^0)^T \mathbf{X}^T \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^0), \tag{2}$$

are ellipses centered at the Ordinary Least Squares (OLS) solution, $\hat{\boldsymbol{\beta}}^0$. Since the predictors are standardized, when $p = 2$ the principal axis of the contours are at ±45° to the horizontal. As the contours are in terms of $\mathbf{X}^T\mathbf{X}$, as opposed to $(\mathbf{X}^T\mathbf{X})^{-1}$, positive correlation would yield contours that are at $-45°$ whereas negative correlation gives the reverse.

*** FIGURE 1 GOES HERE ***

In the (β_1, β_2) plane, intuitively, the solution is the first time that the contours of the sum-of-squares loss function hit the constraint region. The left panel of Figure 1 depicts the shape of the constraint region for the LASSO and the Elastic Net (Zou and Hastie, 2005), which uses a mixture of L_1 and L_2 penalties. Note that the ridge regression contours (not shown) are circles centered at the origin. As the contours are more likely to hit at a vertex, the non-differentiability of the LASSO and Elastic Net at the axes encourage sparsity, with the LASSO doing so to a larger degree due to the linear boundary. Meanwhile, if two variables were highly correlated, the Elastic Net would more often include both into the model, as opposed to only including one of the two.

The right panel of Figure 1 illustrates the constraint region for the OSCAR for various values of the parameter c. From this figure, the reason for the octagonal term in the name is now clear. The shape of the constraint region in two dimensions is exactly an octagon. With vertices on the diagonals along with the axes, the OSCAR encourages both sparsity and equality of coefficients to varying degrees, depending on the strength of correlation, the value of c, and the location of the OLS solution.

*** FIGURE 2 GOES HERE ***

Figure 2 shows that with the same OLS solution, grouping is more likely to occur if the predictors are highly correlated. This implicit relationship to correlation is also quantified later in Theorem 1. Figure 2a shows that if the correlation between predictors is small ($\rho = 0.15$), the sum-of-squares contours first intersect the constraint region on the vertical axis, giving a sparse solution with $\hat{\beta}_1 = 0$. In comparison, the right panel shows that with the same OLS solution, if the predictors are highly correlated ($\rho = 0.85$), the two coefficients reach equality, and thus the predictors form a group.

Remark: By construction, considering the mirror image, i.e. negative correlation, the coefficients would be set equal in magnitude, differing in sign. This would correspond to using the difference between the two predictors as opposed to the sum, possibly denoting a pair of competing predictors, or that a sign change is appropriate, if applicable.

Note that choosing $c = 0$ in the OSCAR yields the LASSO, which gives only sparsity and no clustering, while letting $c \to \infty$ gives a square penalty region and only clustering with no variable selection. Varying c changes the angle formed in the octagon from the extremes of a diamond ($c = 0$), through various degrees of an octagon to its limit as a square, as in two dimensions, $-1/(c+1)$ represents the slope of the line in the first quadrant that intersects the $y-$axis. In all cases, it remains a convex region.

Remark: Note that the pairwise L_∞ is used instead of the overall L_∞. Although in two-dimensions they accomplish the identical task, their behaviors in $p > 2$ dimensions are quite different. Using an overall L_∞ only allows for the possibility of a single clustered group which must contain the largest coefficient, as it shrinks from top down. Defining the OSCAR through the pairwise L_∞ allows for multiple groups of varying sizes, as its higher dimensional constraint region has vertices and edges corresponding to each of these more complex possible groupings.

2.2 *Exact grouping property*

The OSCAR formulation as a constrained optimization problem (1) can be written in the penalized form

$$\begin{aligned}\hat{\boldsymbol{\beta}} = &\ \arg\min_{\boldsymbol{\beta}} \left|\left|\mathbf{y} - \textstyle\sum_{j=1}^{p} \beta_j \mathbf{x}_j\right|\right|^2 + \lambda \left[\textstyle\sum_{j=1}^{p} |\beta_j| + c\ \textstyle\sum_{j<k} \max\{|\beta_j|, |\beta_k|\}\right] \\ = &\ \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \textstyle\sum_{j=1}^{p} \beta_j \mathbf{x}_j||^2 + \lambda \textstyle\sum_{j=1}^{p} \{c(j-1)+1\}|\beta|_{(j)},\end{aligned} \qquad (3)$$

with $|\beta|_{(1)} \leq |\beta|_{(2)} \leq ... \leq |\beta|_{(p)}$, and there exists a direct correspondence between λ and the bound t.

An explicit relation between the choice of the constraint bound t and the penalization parameter λ is now given. This allows for computation using an algorithm as discussed in Section 3 derived via the constraint representation, while also considering properties that can be derived via the equivalent penalized representation. Furthermore, a quantification of the exact grouping property of the OSCAR solution in terms of correlation is then given by

Theorem 1.

Consider the representation of the OSCAR in terms of the penalized least squares criterion (3) with penalty parameter λ. Suppose that the set of covariates $(\mathbf{x}_1, ..., \mathbf{x}_p)$ are ordered such that their corresponding coefficient estimates satisfy $0 < |\hat{\beta}_1| \leq ... \leq |\hat{\beta}_Q|$ and $\hat{\beta}_{Q+1} = ... = \hat{\beta}_p = 0$. Let $0 < \hat{\theta}_1 < ... < \hat{\theta}_G$ denote the G unique nonzero values of the set of $|\hat{\beta}_j|$, so that $G \leq Q$.

For each $g = 1, ..., G$, let

$$\mathcal{G}_g = \{j \ : \ |\hat{\beta}_j| = \hat{\theta}_g\}$$

denote the set of indices of the covariates that correspond to that value for the absolute coefficient. Now construct the grouped $n \times G$ covariate matrix $\mathbf{X}^* \equiv [\mathbf{x}_1^* \ ... \ \mathbf{x}_G^*]$ with

$$\mathbf{x}_g^* = \sum_{j \in \mathcal{G}_g} \text{sign}(\hat{\beta}_j) \ \mathbf{x}_j. \tag{4}$$

This transformation amounts to combining the variables with identical magnitudes of the coefficients by a simple (signed) summation of their values, as in forming a new predictor from the group mean. Form the corresponding summed weights

$$w_g = \textstyle\sum_{j \in \mathcal{G}_g} \{c \ (j-1) + 1\}.$$

The criterion in (3) can be written explicitly in terms of this "active set" of covariates, as

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \ ||\mathbf{y} - \sum_{g=1}^{G} \theta_g \mathbf{x}_g^*||^2 + \lambda \sum_{g=1}^{G} w_g \theta_g, \tag{5}$$

with $0 < \theta_1 < ... < \theta_G$. In a neighborhood of the solution, the ordering, and thus the weights, remain constant and as the criteria is differentiable on the active set, one obtains for each $g = 1, ..., G$

$$-2\mathbf{x}_g^{*T}(\mathbf{y} - \mathbf{X}^* \hat{\boldsymbol{\theta}}) + \lambda w_g = 0. \tag{6}$$

This vector of score equations corresponds to those in Zou et al. (2004) and Zou and Hastie (2005) after grouping and absorbing the sign of the coefficient into the covariate.

Equation (6) allows one to obtain the corresponding value of λ for a solution obtained from a given choice of t, i.e. for all values of g, (6) yields

$$\lambda = 2\mathbf{x}_g^{*T}(\mathbf{y} - \mathbf{X}^*\hat{\boldsymbol{\theta}})/w_g. \tag{7}$$

The octagonal shape of the constraint region in Figure 1 graphically depicts the exact grouping property of the OSCAR optimization criterion. The following theorem quantifies this exact grouping property in terms of the correlation between covariates, showing that the equality of two coefficients is easier to obtain as the correlation between the two predictors increases, in that less penalty is needed on the L_∞ norm to do so.

Theorem. Set $\lambda_1 \equiv \lambda$ and $\lambda_2 \equiv c\lambda$ in the Lagrangian formulation given by (3). Given data $(\mathbf{y}, \mathbf{X})$ with centered response $\mathbf{y}$ and standardized predictors $\mathbf{X}$, let $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)$ be the OSCAR estimate using the tuning parameters (λ_1, λ_2). Assume that the predictors are signed so that $\hat{\beta}_i(\lambda_1, \lambda_2) \geq 0$ for all i. Let $\rho_{ij} = \mathbf{x}_i^T\mathbf{x}_j$ be the sample correlation between covariates i and j.

For a given pair of predictors $\mathbf{x}_i$ and $\mathbf{x}_j$, suppose that both $\hat{\beta}_i(\lambda_1, \lambda_2) > 0$ and $\hat{\beta}_j(\lambda_1, \lambda_2) > 0$ are distinct from the other $\hat{\beta}_k$. Then there exists $\lambda_0 \geq 0$ such that if $\lambda_2 > \lambda_0$ then

$$\hat{\beta}_i(\lambda_1, \lambda_2) = \hat{\beta}_j(\lambda_1, \lambda_2), \text{ for all } \lambda_1 > 0.$$

Furthermore, it must be that

$$\lambda_0 \leq 2||\mathbf{y}||\sqrt{2(1-\rho_{ij})}.$$

The proof of Theorem 1 is based on the score equations in (6), and is given in Appendix A.

In the above notation, λ_2 controls the degree of grouping. As λ_2 increases, any given pair of predictors will eventually group. However, the $\sqrt{2(1-\rho_{ij})}$ term shows that highly correlated predictors are more likely to be grouped. In particular, if two predictors were identical ($\rho = 1$), they will be grouped for any $\lambda_2 > 0$, i.e., any form of the OSCAR penalty other than the special case of the LASSO.

Remark: In Theorem 1, the requirement of the distinctness of $\hat{\beta}_i$ and $\hat{\beta}_j$ is not as restrictive as may first appear. The $\mathbf{x}_i$ and $\mathbf{x}_j$ may themselves already represent grouped covariates as in (4), then ρ_{ij} represents the correlation between the groups.

3. Computation and cross-validation

3.1 *Computation*

A computational algorithm is now discussed to compute the OSCAR estimate for a given set of tuning parameters (t, c). Write $\beta_j = \beta_j^+ - \beta_j^-$ with both β_j^+ and β_j^- being non-negative, and only one is nonzero. Then $|\beta_j| = \beta_j^+ + \beta_j^-$. Introduce the additional $p(p-1)/2$ variables η_{jk} for $1 \leq j < k \leq p$, for the pairwise maxima. Then the optimization problem in (1) is equivalent to

$$\begin{gathered}\text{Minimize: } \tfrac{1}{2}\|\mathbf{y} - \textstyle\sum_{j=1}^{p}(\beta_j^+ - \beta_j^-)\mathbf{x}_j\|^2 \\ \text{subject to} \\ \textstyle\sum_{j=1}^{p}(\beta_j^+ + \beta_j^-) + c\ \textstyle\sum_{j<k}\eta_{jk} \leq t, \\ \eta_{jk} \geq \beta_j^+ + \beta_j^-, \eta_{jk} \geq \beta_k^+ + \beta_k^- \text{ for each } 1 \leq j < k \leq p, \\ \beta_j^+ \geq 0, \beta_j^- \geq 0 \text{ for all } j = 1, ..., p,\end{gathered} \tag{8}$$

where the minimization is with respect to the expanded parameter vector $(\boldsymbol{\beta}^+, \boldsymbol{\beta}^-, \boldsymbol{\eta})$.

This is now a quadratic programming problem with $(p^2 + 3p)/2$ total parameters and $p^2 + p + 1$ total linear constraints. The constraint matrix is very large, but it is extremely sparse. The optimization has been performed using the quadratic programming algorithm SQOPT (Gill et al., 2005), designed specifically for large-scale problems with sparse matrices. Problems with a few hundred predictors are directly computable using this algorithm.

3.2 *Choosing the tuning parameters*

Choosing the tuning parameters (c, t) can be done via minimizing an estimate of the out-of-sample prediction error. If a validation set is available, this can be estimated directly.

Lacking a validation set one can use five-fold or ten-fold cross-validation, or a technique such as generalized cross-validation (GCV), AIC, BIC, or C_p to estimate the prediction error. In using this form of model selection criteria one would need to use the estimated degrees of freedom as in Efron et al. (2004).

For the LASSO, it is known that the number of non-zero coefficients is an unbiased estimate of the degrees of freedom (Efron et al., 2004; Zou et al., 2004). For the fused LASSO, Tibshirani et al. (2005) estimate the degrees of freedom by the number of non-zero distinct blocks of coefficients. Thus, the natural estimate of the degrees of freedom for the OSCAR is the number of distinct non-zero values of $\{|\hat{\beta}_1|, ..., |\hat{\beta}_p|\}$. This gives a measure of model complexity for the OSCAR in terms of the number of coefficients in the model.

4. Simulation study

A simulation study was run to examine the performance of the OSCAR under various conditions. Five setups are considered in this simulation. The setups are similar to those used in both Tibshirani (1996) and Zou and Hastie (2005). In each example, data is simulated from the regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

For each example, 100 data sets were generated. Each data set consisted of a training set of size n, along with an independent validation set of size n used solely to select the tuning parameters. For each of the 100 data sets, the models were fit on the training data only. For each procedure, the model fit with tuning parameter(s) yielding the lowest prediction error on the validation set was selected as the final model. For these tuning parameters, the estimated coefficients based on the training set are then compared in terms of the mean-squared error and the resulting model complexity. For the simulations, the mean-squared error (MSE) is calculated as in Tibshirani (1996) via

$$\text{MSE} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T V (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \tag{9}$$

where V is the population covariance matrix for $\mathbf{X}$, with prediction error given by MSE + σ^2.

The five scenarios are given by:

1. In example one, $n = 20$ and there are $p = 8$ predictors. The true parameters are $\boldsymbol{\beta} = (3, 2, 1.5, 0, 0, 0, 0, 0)^T$ and $\sigma = 3$, with covariance given by $\text{Cov}(\mathbf{x}_i, \mathbf{x}_j) = 0.7^{|i-j|}$. The first three variables are moderately correlated and similar in effect sizes, while the remaining five are unimportant and also somewhat correlated.

2. Example two is the same as example one, except that $\beta_j = (3, 0, 0, 1.5, 0, 0, 0, 2)^T$. Now the important variables have little correlation with one another, but they are more correlated with the unimportant predictors.

3. Example three is the same as example one, except that $\beta_j = 0.85$ for all j, creating a non-sparse underlying model.

4. In example four, $n = 100$ and there are $p = 40$ predictors. The true parameters are

$$\boldsymbol{\beta} = (\underbrace{0, ..., 0}_{10}, \underbrace{2, ..., 2}_{10}, \underbrace{0, ..., 0}_{10}, \underbrace{2, ..., 2}_{10})^T$$

and $\sigma = 15$, with covariance given by $\text{Cov}(\mathbf{x}_i, \mathbf{x}_j) = 0.5$ for $i \neq j$ and $\text{Var}(\mathbf{x}_i) = 1$ for all i.

5. In example five, $n = 50$ and there are again 40 predictors. The true parameters are

$$\boldsymbol{\beta} = (\underbrace{3, ..., 3}_{15}, \underbrace{0, ..., 0}_{25})^T$$

and $\sigma = 15$. The predictors were generated as:

$$
\begin{gathered}
\mathbf{x}_i = Z_1 + \epsilon_i^x, \;\; Z_1 \sim N(0,1), \;\; i = 1, ..., 5 \\
\mathbf{x}_i = Z_2 + \epsilon_i^x, \;\; Z_2 \sim N(0,1), \;\; i = 6, ..., 10 \\
\mathbf{x}_i = Z_3 + \epsilon_i^x, \;\; Z_3 \sim N(0,1), \;\; i = 11, ..., 15 \\
\mathbf{x}_i \sim N(0,1), \;\; i = 16, ..., 40
\end{gathered}
$$

where ϵ_i^x are independent identically distributed $N(0, 0.16)$, $i = 1, ..., 15$. In this model the three equally important groups have pairwise correlations $\rho \approx 0.85$, and there are 25 pure noise features.

*** TABLE 1 GOES HERE ***

Table 1 summarizes both the mean squared error and complexity of the model in terms of the number of unique non-zero coefficients required in the chosen model. In all examples, the OSCAR produces the least complex model by collapsing some of the predictors into groups. Meanwhile, the simulations show that the OSCAR is highly competitive in prediction. Its mean squared error is either best or second best in all five examples.

Although the values of the coefficients are the same for examples 1 and 2, the OSCAR generally chooses a smaller model for example 1, as can be seen from the number of degrees of freedom in Table 1. This is due to the interplay between the correlation and the values of the coefficients. This is to be expected, as in example 1, variables with similar coefficients are also highly correlated so the grouping mechanism of the OSCAR is more likely to group both the first 3 coefficients together, as well as group the remaining 5 unimportant variables together at zero.

The Elastic Net also performs well in terms of prediction error, particularly in cases such as examples 1, 2, and 5 in which there is higher correlation and the true vector is sparse.

Particularly in example 5, the Elastic Net's median MSE is lower than the rest, although upon looking at the quantiles, the distribution of MSE in the 100 samples is somewhat similar to the OSCAR. However, the exact grouping effect of the OSCAR allows for the identification of a group structure among the predictors that is not accomplished by the Elastic Net, as seen in the resulting number of coefficients in the model. The loss in prediction error using the OSCAR for this model could come from the large number of unimportant variables combined with the smaller sample size resulting in some of the unimportant variables being smoothed towards the important ones a bit more. In example 3, when all of the predictors are important and equal in effect, the OSCAR and ridge regression perform extremely well in mean squared error, while the OSCAR also performs grouping. The coefficients for this example were also varied to allow for unequal, but similar effects and the results were similar, thus omitted. Overall, the OSCAR appears to compete well with the existing approaches in terms of mean squared error in all cases studied, while yielding the additional grouping information to accomplish the supervised clustering task that is not built into the other procedures.

5. Real data

The data for this example come from a study of the associations between soil characteristics and rich-cove forest diversity in the Appalachian Mountains of North Carolina. Twenty 500 m^2 plots were surveyed. The outcome is the number of different plant species found within the plot and the fifteen soil characteristics used as predictors of forest diversity are listed in Figure 3. The soil measurements for each plot are the average of five equally-spaced measurements taken within the plot. The predictors were first standardized before performing the analysis. Since this data set has only $p = 15$ predictors, it allows for an in-depth illustration of the behavior of the OSCAR solution.

*** FIGURE 3 GOES HERE ***

Figure 3 shows that there are several highly correlated predictors. The first seven covariates are all related to the abundance of positively charged ions, i.e., cations. Percent base saturation, cation exchange capacity (CEC), and the sum of cations are all summaries of the abundance of cations; calcium, magnesium, potassium, and sodium are all examples of cations. Some of the pairwise absolute correlations between these covariates are as high as 0.95. The correlations involving potassium and sodium are not quite as high as the others. There is also strong correlation between sodium and phosphorus, and between soil pH and exchangeable acidity, two measures of acidity. Additionally, the design matrix for these predictors is not full rank, as the sum of cations is derived as the sum of the four listed elements.

*** TABLE 2 GOES HERE ***

*** FIGURE 4 GOES HERE ***

Using five-fold cross-validation, the best LASSO model includes seven predictors, including two moderately correlated cation covariates: CEC and potassium (Table 2). The LASSO solution paths as a function of s, the proportion of the OLS L_1 norm, for the seven cation-related covariates are plotted in Figure 4a, while the remaining eight are plotted in Figure 4b. As the penalty decreases, the first two cation-related variables to enter the model are CEC and potassium. As the penalty reaches 15% of the OLS norm, CEC abruptly drops out of the model and is replaced by calcium, which is highly correlated with CEC ($\rho = 0.94$). Potassium remains in the model after the addition of calcium, as the correlation between the two is not as extreme ($\rho = 0.62$). Due to the high collinearity, the method for choosing the tuning parameter in the LASSO greatly affects the choice of the model; five-fold cross validation includes CEC, whereas generalized cross-validation (GCV) instead includes calcium. Clearly, at least one of the highly correlated cation covariates should be included in the model, but the LASSO is unsure about which one.

The five-fold cross-validation OSCAR solution (Table 2) includes all seven predictors selected by the LASSO along with two additional cation covariates: the sum of cations and calcium. The OSCAR solution groups the four selected cation covariates together, giving a model with six distinct non-zero parameters. The cation covariates are highly correlated and are all associated with the same underlying factor. Therefore, taking their sum as a derived predictor, rather than treating them as separate covariates and arbitrarily choosing a representative, may provide a better measure of the underlying factor and thus a more informative and better predictive model. Note that since the LASSO is a special case of the OSCAR with $c = 0$, the grouped OSCAR solution has smaller cross-validation error than the LASSO solution.

The pairs of tuning parameters selected by both five-fold cross validation and GCV each have $c = 4$, therefore Figures 4c and 4d plot the OSCAR solution paths for fixed $c = 4$ as a function of the proportion of the penalty's value at the OLS solution, denoted by s. Ten-fold and leave-one-out cross-validation along with the AIC and BIC criteria were also used and the results were similar. As with the LASSO, CEC is the first cation-related covariate to enter the model as the penalty decreases. However, rather than replacing CEC with calcium as the penalty reaches 15% of the OLS norm, these parameters are fused, along with the sum of cations and potassium.

Soil pH is also included in the group for the GCV solution. Although pH is not as strongly associated with the cation covariates (Figure 3), it is included in the group chosen by GCV (but not from five-fold cross validation) because the magnitude of its parameter estimate at that stage is similar to the magnitude of the cation groups estimate. The OSCAR penalty occasionally results in grouping of weakly correlated covariates that have similar magnitudes, producing a smaller dimensional model. However, by further examining the solution paths in Figures 4c and 4d, it is clear that the more correlated variables tend to remain grouped, whereas others only briefly join the group and are then pulled elsewhere. For example, the GCV solution groups Copper and Manganese, but the solution paths of these two variables' coefficients are only temporarily set equal as they cross. This example shows that more insight regarding the predictor relationships can be uncovered from the solution paths. It is also worth noting that for the eight covariates that are not highly correlated, the OSCAR and LASSO solution paths are similar, as may be expected.

6. Discussion

This paper has introduced a new procedure for variable selection in regression while simultaneously performing supervised clustering. The resulting clusters can then be further investigated to determine what relationship among the predictors and response may be responsible for the grouping structure.

The OSCAR penalty can be applied to other optimization criteria in addition to least squares regression. Generalized linear models with this penalty term on the likelihood are possible via quadratic approximation of the likelihood. Extensions to lifetime data, in which difficulties due to censoring often arise, is another natural next step. In some situations there may be some natural potential groups among the predictors, so one would only include the penalty terms corresponding to predictors among the same group. Examples would include ANOVA or nonparametric regression via a set of basis functions.

In the spirit of other penalized regression techniques, the OSCAR solution also has an interpretation as the posterior mode for a particular choice of prior distribution. The OSCAR prior corresponds to a member of the class of multivariate exponential distributions proposed by Marshall and Olkin (1967).

The quadratic programming problem can be large and many standard solvers may have difficulty solving it directly. In the absence of a more efficient solver such as the SQOPT algorithm used by the authors, Web Appendix B discusses a sequential method that will often alleviate this problem.

Based on recent results of Rosset and Zhu (2006), for each given c, the solution path for the OSCAR as a function of the bound t, should be piecewise linear. A modification of the Least Angle Regression (LARS) algorithm that gives the entire solution path for a fixed c, as it does for $c = 0$ would be desirable to dramatically improve computation. However, in addition to adding or removing variables at each step, more possibilities must be considered as variables can group together or split apart as well. Further research into a more efficient computational algorithm is warranted, particularly upon extension to more complicated models.

Acknowledgements

The authors are grateful to Clay Jackson of the Department of Forestry at North Carolina State University for providing the soil data.

Appendix A

Proof of Theorem 1.

Suppose that $\hat{\beta}_i(\lambda_1, \lambda_2) \neq \hat{\beta}_j(\lambda_1, \lambda_2)$, and both are non-zero, then by differentiation one obtains

$$-2\mathbf{x}_i^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \lambda w_i = 0, \tag{10}$$

and

$$-2\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \lambda w_j = 0. \tag{11}$$

Subtracting (10) from (11) yields

$$-2(\mathbf{x}_j^T - \mathbf{x}_i^T)(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \lambda(w_j - w_i) = 0. \tag{12}$$

Since $\mathbf{X}$ is standardized, $||\mathbf{x}_j^T - \mathbf{x}_i^T||^2 = 2(1-\rho_{ij})$. This together with the fact that $||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2 \leq ||\mathbf{y}||^2$ gives

$$|w_j - w_i| \leq 2\lambda^{-1}||\mathbf{y}||\sqrt{2(1-\rho_{ij})}. \tag{13}$$

However, by construction of the weights, $|w_j - w_i| \geq c$, with equality holding if the two are adjacent in the coefficient ordering. Hence if $c > 2\lambda^{-1}||\mathbf{y}||\sqrt{2(1-\rho_{ij})}$, one obtains a contradiction. This completes the proof.

Appendix B

An alternative computational algorithm is now discussed to compute the OSCAR estimate for a given set of tuning parameters (t, c). This is a sequential algorithm that requires solving a series of quadratic programming problems that are increasing in size instead of one large problem. As before, write $\beta_j = \beta_j^+ - \beta_j^-$ with both β_j^+ and β_j^- being non-negative, and only one is nonzero. Then $|\beta_j| = \beta_j^+ + \beta_j^-$.

Suppose now that the covariates were ordered so that the components of the solution to the OSCAR optimization problem was in order of non-decreasing magnitude. Then the optimization problem can be rewritten as

$$\begin{array}{c} \hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \sum_{j=1}^p \beta_j \mathbf{x}_j||^2 \\ \text{subject to} \\ |\beta_1| \leq |\beta_2| \leq ... \leq |\beta_p| \\ \sum_{j=1}^p \{c\ (j-1)+1\}(\beta_j^+ + \beta_j^-) \leq t \\ \boldsymbol{\beta}^+ \geq 0 \\ \boldsymbol{\beta}^- \geq 0. \end{array} \tag{14}$$

Note that the weighted linear combination being bounded is using weights that increase with increasing magnitude of the component. Due to the nature of the weights, the ordering constraint can instead be incorporated by placing the same bound, t, on each of the $p!$ possible weighted linear combinations. This follows immediately from the fact that given two ordered vectors, $\mathbf{w}$ and $\mathbf{v}$, so that $w_1 < w_2 < ... < w_p$ and $v_1 < v_2 < ... < v_p$, clearly

$\mathbf{w}^T\mathbf{v} \geq \mathbf{w}_*{}^T\mathbf{v}$, where $\mathbf{w}_*$ is any other permutation of $\mathbf{w}$. However, this gives a quadratic programming problem with an almost surely overwhelming $p! + 2p$ linear constraints. Instead of directly solving the large quadratic programming problem, a sequential quadratic programming algorithm proceeds as follows:

1. Solve the quadratic programming problem with $2p + 1$ constraints using the ordering of coefficients obtained from least squares (or some other method).

2. If the solution does not maintain the same ordering, add the linear constraint corresponding to the new ordering and solve the more restrictive quadratic programming problem (with one more constraint).

3. Repeat until the ordering remains constant. Any additional constraint will no longer affect the solution.

The algorithm is based on the fact that, for a given set of constraints based on orderings, if the minimizer of the quadratic programming problem has components that are ordered in the same way as one in the current set, this solution automatically satisfies the remaining constraints. This again follows immediately from the nature of the weights.

Since the feasible region at each step is contained in the feasible region from the previous step, the algorithm is guaranteed to converge to the final fully constrained solution. Although arrival at the final solution could, in theory, require inclusion of all $p!$ constraints, in testing the algorithm through a number of examples, the number of constraints needed to obtain the final solution is typically on the order of p or less.

References

Dettling, M. and Bühlmann, P. (2004). Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis* **90**, 106-131.

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

Gill, P. E., Murray, W., and Saunders, M. A. (2005). Users guide for SQOPT 7: a Fortran package for large-scale linear and quadratic programming. *Technical Report NA 05-1, Department of Mathematics, University of California, San Diego.*

Hastie, T., Tibshirani, R., Botstein, D., and Brown, P. (2001). Supervised harvesting of expression trees. *Genome Biology* **2** (1), 3.1-3.12.

Jörnsten, R. and Yu, B. (2003). Simultaneous gene clustering and subset selection for sample classification via MDL. *Bioinformatics* **19**, 1100-1109.

Marshall, A. W. and Olkin, I. (1967). A multivariate exponential distribution. *Journal of the American Statistical Association* **62**, 30-44.

Park, M. Y., Hastie, T., and Tibshirani, R. (2006). Averaged gene expressions for regression. *Biostatistics*, Advanced access preprint available online.

Rosset, S. and Zhu, J. (2006), Piecewise linear regularized solution paths, *Annals of Statistics* **35**, to appear.

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society* B **67**, 91-108.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society* B **68**, 49-67.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society* B **67**, 301-320.

Zou, H., Hastie, T., and Tibshirani, R. (2004). On the degrees of freedom of the lasso. *Technical report, Department of Statistics, Stanford University.*

Zou, H. and Yuan, M. (2006). The F_∞-norm support vector machine. *Technical report 646, School of Statistics, University of Minnesota.*

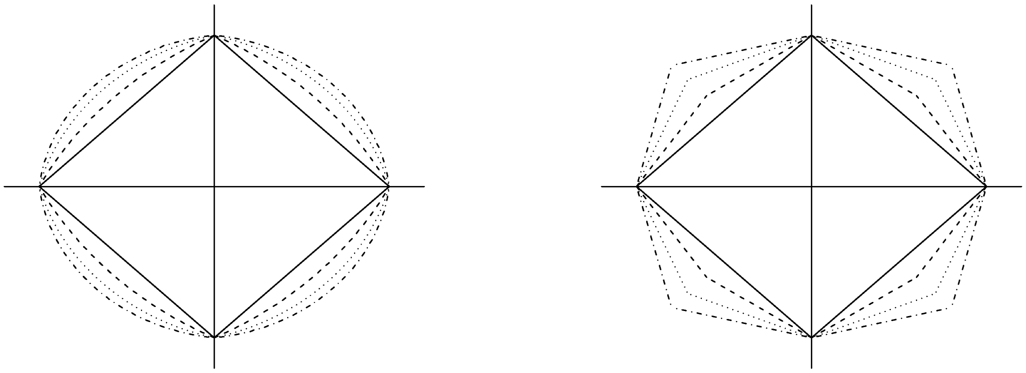
Table 1: *Simulation Study. Median mean-squared errors for the simulated examples based on 100 replications with standard errors estimated via the bootstrap in parentheses. The 10th and 90th percentiles of the 100 MSE values are also reported. The median number of unique non-zero coefficients in the model is denoted by Median Df, while the 10th and 90th percentiles of this distribution are also reported.*

Example		Med. MSE	MSE	MSE	Med. Df	Df	Df
		(Std. Err.)	10th Perc.	90th Perc.		10th Perc.	90th Perc.
1	Ridge	2.31 (0.18)	0.98	4.25	8	8	8
	Lasso	1.92 (0.16)	0.68	4.02	5	3	8
	Elastic Net	1.64 (0.13)	0.49	3.26	5	3	7.5
	Oscar	1.68 (0.13)	0.52	3.34	4	2	7
2	Ridge	2.94 (0.18)	1.36	4.63	8	8	8
	Lasso	2.72 (0.24)	0.98	5.50	5	3.5	8
	Elastic Net	2.59 (0.21)	0.95	5.45	6	4	8
	Oscar	2.51 (0.22)	0.96	5.06	5	3	8
3	Ridge	1.48 (0.17)	0.56	3.39	8	8	8
	Lasso	2.94 (0.21)	1.39	5.34	6	4	8
	Elastic Net	2.24 (0.17)	1.02	4.05	7	5	8
	Oscar	1.44 (0.19)	0.51	3.61	5	2	7
4	Ridge	27.4 (1.17)	21.2	36.3	40	40	40
	Lasso	45.4 (1.52)	32.0	56.4	21	16	25
	Elastic Net	34.4 (1.72)	24.0	45.3	25	21	28
	Oscar	25.9 (1.26)	19.1	38.1	15	5	19
5	Ridge	70.2 (3.05)	41.8	103.6	40	40	40
	Lasso	64.7 (3.03)	27.6	116.5	12	9	18
	Elastic Net	40.7 (3.40)	17.3	94.2	17	13	25
	Oscar	51.8 (2.92)	14.8	96.3	12	9	18

Table 2: *Estimated coefficients for the soil data example.*

	OSCAR	OSCAR	LASSO	LASSO
Variable	(5-fold CV)	(GCV)	(5-fold CV)	(GCV)
% Base Saturation	0	-0.073	0	0
Sum Cations	-0.178	-0.174	0	0
CEC	-0.178	-0.174	-0.486	0
Calcium	-0.178	-0.174	0	-0.670
Magnesium	0	0	0	0
Potassium	-0.178	-0.174	-0.189	-0.250
Sodium	0	0	0	0
Phosphorus	0.091	0.119	0.067	0.223
Copper	0.237	0.274	0.240	0.400
Zinc	0	0	0	-0.129
Manganese	0.267	0.274	0.293	0.321
Humic Matter	-0.541	-0.558	-0.563	-0.660
Density	0	0	0	0
pH	0.145	0.174	0.013	0.225
Exchangeable Acidity	0	0	0	0

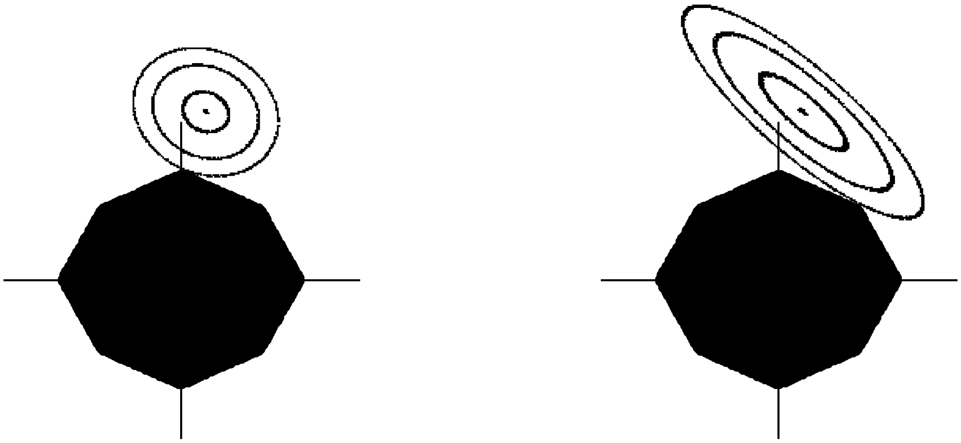
Figure 1: Graphical representation of the constraint region in the (β_1, β_2) plane for the LASSO, Elastic Net, and OSCAR. Note that all are non-differentiable at the axes.



(a) Constraint region for the Lasso (solid line), along with three choices of tuning parameter for the Elastic Net.

(b) Constraint region for the OSCAR for four values of c. The solid line represents $c = 0$, the LASSO.

Figure 2: Graphical representation in the (β_1, β_2) plane. The OSCAR solution is the first time the contours of the sum-of-squares function hits the octagonal constraint region.



(a) Contours centered at OLS estimate, low correlation ($\rho = .15$). Solution occurs at $\hat{\beta}_1 = 0$.

(b) Contours centered at OLS estimate, high correlation ($\rho = .85$). Solution occurs at $\hat{\beta}_1 = \hat{\beta}_2$.

Figure 3: Graphical representation of the correlation matrix of the 15 predictors for the soil data. The magnitude of each pairwise correlation is represented by a block in the grayscale image.

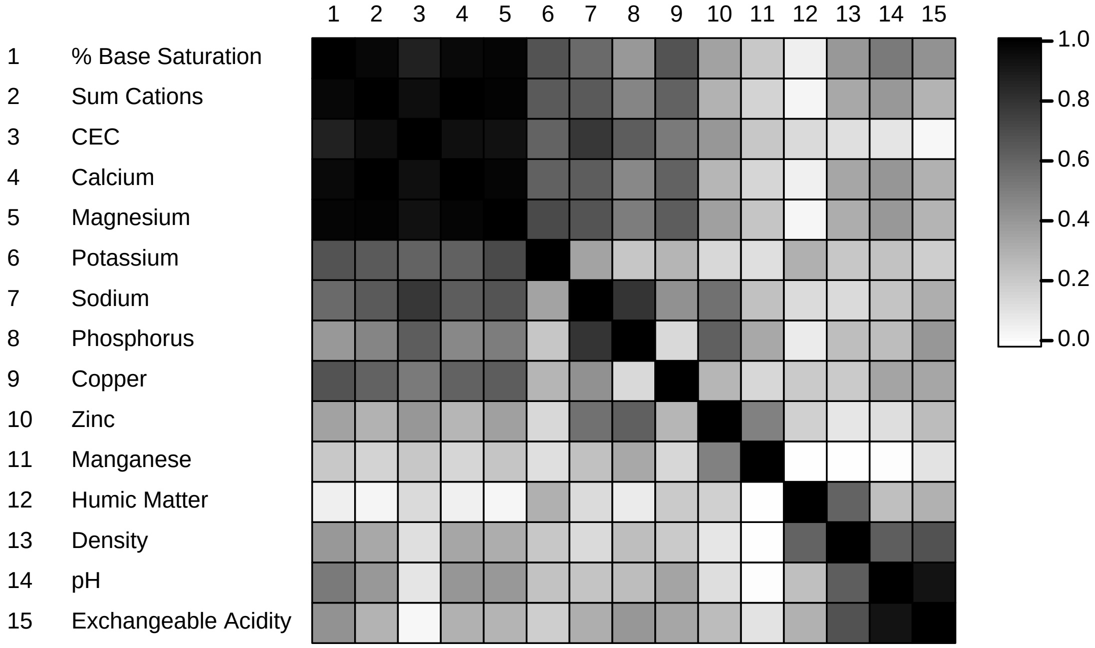
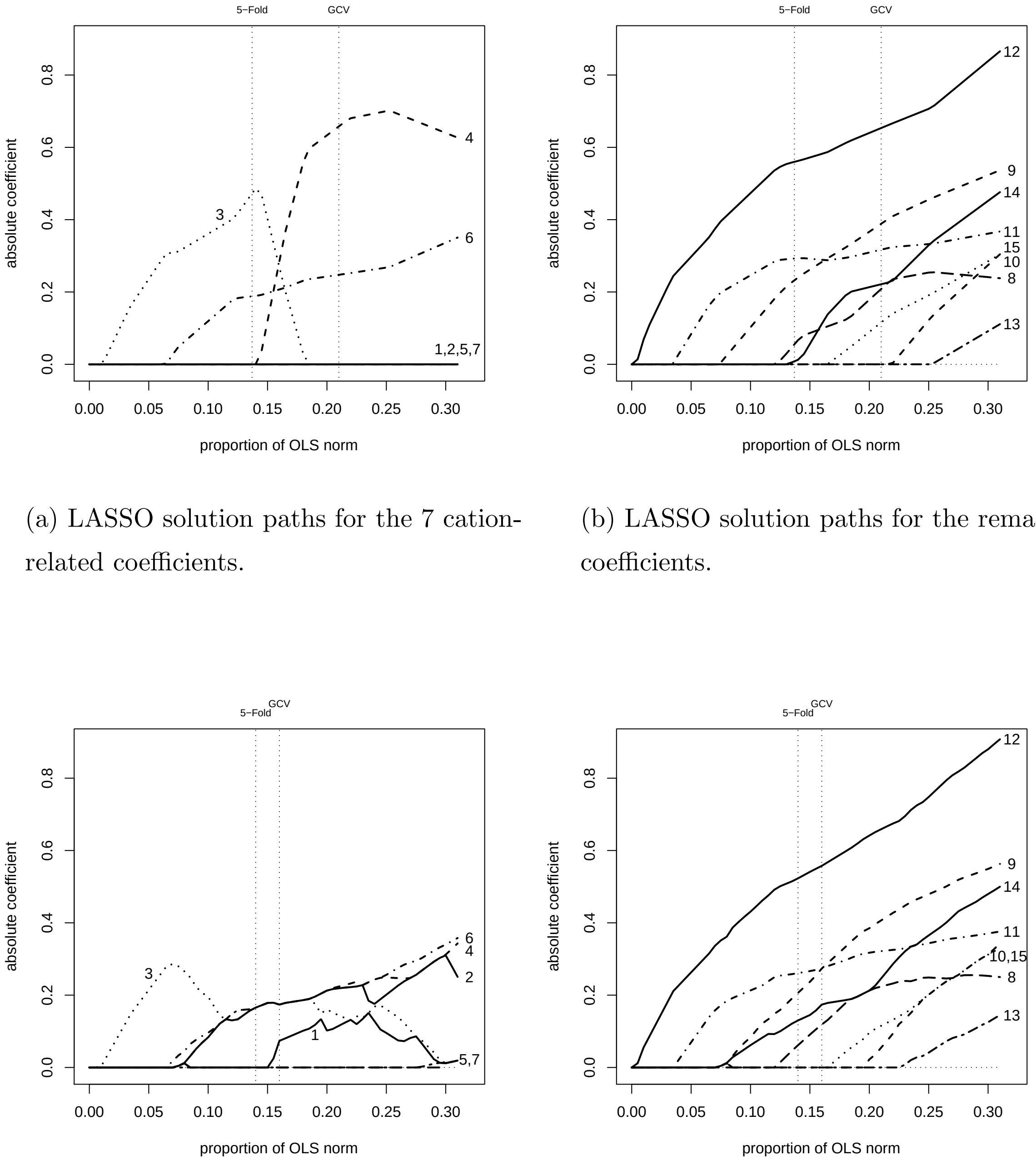


Figure 4: Solution paths for the soil data. Plot of the 15 coefficients as a function of s, the proportion of the penalty evaluated at the OLS solution. The first row uses the fixed value of $c = 0$, the LASSO. The second row uses the value $c = 4$ as chosen by both GCV and 5-fold cross-validation. The vertical lines represent the best models in terms of the GCV and the 5-fold cross-validation criteria for each.



(a) LASSO solution paths for the 7 cation-related coefficients.

(b) LASSO solution paths for the remaining 8 coefficients.

(a) OSCAR solution paths for the 7 cation-related coefficients.

(b) OSCAR solution paths for the remaining 8 coefficients.