DEPARTMENT OF STATISTICS
North Carolina State University
2501 Founders Drive, Campus Box 8203
Raleigh, NC 27695-8203

Institute of Statistics Mimeo Series No. 2583

Simultaneous regression shrinkage, variable selection and 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.

(email: bondell@stat.ncsu.edu)

Simultaneous regression shrinkage, variable selection and 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.

(email: bondell@stat.ncsu.edu)

Abstract

In this paper, a new method called the OSCAR (Octagonal Shrinkage and Clustering Algorithm for Regression) is proposed to simultaneously select variables and perform supervised clustering in the context of linear regression. The technique is based on penalized least squares with a geometrically intuitive penalty function that, like the LASSO penalty, 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 clusters represented by a single coefficient. These resulting clusters can then be investigated further to discover what contributes to the group having a similar behavior. The OSCAR then enjoys sparseness in terms of the number of unique coefficients in the model. The proposed procedure is shown to compare favorably to the existing shrinkage and variable selection techniques in terms of both prediction error and reduced model complexity.

Keywords: Clustering; Correlation; Grouping effect; Penalization; Regression; Shrinkage; Variable selection

1 Introduction

Consider the usual linear regression model with a data set of 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 jth predictor, $j = 1, ..., p$. The vector of predicted responses $\hat{\mathbf{y}}$ is given by

$$\hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}_1 + ... + \hat{\beta}_p \mathbf{x}_p. \tag{1.1}$$

Interest focuses on finding the vector $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, ..., \hat{\beta}_p)$ that is best under a given criterion, such as prediction accuracy.

Ordinary least squares (OLS) computes $\hat{\boldsymbol{\beta}}$ by minimizing the residual sum of squares. However, it is well known that OLS can often behave poorly in terms of prediction accuracy on future samples. Cases of particular interest are those involving highly correlated predictors and those of high dimension, including the increasingly more common problem where the number of predictors greatly outnumbers the sample size. In addition to prediction accuracy, a parsimonious model is typically preferred over a more complicated model due to its simplicity and interpretability. Hence, another goal is to perform model selection by reducing the dimension of the vector $\hat{\boldsymbol{\beta}}$, which is not directly accomplished by OLS.

Penalization techniques have been introduced to improve upon the prediction accuracy and interpretation of OLS. Ridge regression (Hoerl and Kennard, 1970) minimizes the sum of squared residuals subject to a bound on the L_2 norm of the coefficients. While ridge regression often achieves better prediction accuracy by shrinking the OLS coefficients, particularly in the highly correlated predictor situation, it cannot produce a parsimonious model, as it naturally keeps all predictors. The LASSO (Tibshirani, 1996) instead imposes a bound on the L_1 norm. This technique does both shrinkage and variable selection due to the nature of the constraint region which often results

in several coefficients becoming identically zero. Although it is a highly successful technique, two drawbacks of the LASSO are

1. In the $p > n$ case, the LASSO can select at most n variables. In many situations, including those involving microarray data where $p >> n$, this can be a limiting feature of a variable selection method.

2. If there is a group of highly correlated variables, the LASSO tends to arbitrarily select only one from the group.

Zou and Hastie (2005) introduce the Elastic Net, using a weighted combination of the L_1 and L_2 norms to alleviate these two issues. The Elastic Net has the ability to choose all p variables if necessary, and tends to choose the correlated variables as a group. However, this grouping selection now creates a less parsimonious model, as more coefficients are required to represent the additional variables. The natural question then arises, is it possible to maintain, or even improve on, the parsimony of the LASSO, while still alleviating the above two drawbacks, and even improving prediction accuracy?

Additionally, supervised clustering techniques are also becoming increasingly popular, particularly in large-scale gene expression studies. Supervised clustering refers to the use of a response variable to determine a meaningful clustering of the features, such as a group of genes sharing a common pathway. Jörnsten and Yu (2003) and Dettling and Bühlmann (2004) discuss techniques to perform gene clustering along with subject classification.

To simultaneously shrink coefficients to zero (variable selection) and force equality of coefficients (clustering), Tibshirani et al. (2005) introduce the Fused LASSO, which as in the LASSO places a constraint on the L_1 norm, but additionally constrains the sum of the absolute successive differences of the coefficients. This penalization technique

only applies when the variables have a natural ordering to them (or one enforces an ordering), and does not perform automated variable clustering to unordered features.

In this paper, a new method called the OSCAR (Octagonal Shrinkage and Clustering Algorithm for Regression) is proposed to create the desired grouping effect for correlated predictors without adding complexity to the final model. The 'octagonal' part of the name comes from the geometric interpretation of the procedure, which is discussed in Section 2.1.

Just as the LASSO forces some coefficients identically to zero, whereas ridge regression does not, the OSCAR forces some of the grouped coefficients identically equal (up to a change in sign, if negatively correlated), whereas the ridge-type penalty in the Elastic Net does not. This exact equality of the grouped variables' coefficients allows for a sparse representation in terms of the resulting complexity of the model. So that, in addition to the variable selection via shrinking coefficients to zero, the OSCAR simultaneously accomplishes the supervised clustering task by yielding a single coefficient to determine a cluster of variables that combine to have a single effect on the response. These resulting clusters can then be investigated further to discover what contributes to the group having a similar behavior.

The remainder of the paper is organized as follows. In Section 2, the OSCAR is formulated as a constrained least squares problem and the geometric interpretation of this constraint region is discussed. In addition, a Bayesian perspective of the estimator and a quantification of the grouping property are given. Computational issues, including choosing the tuning parameters, are discussed in Section 3. In Section 4, it is shown 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 a data set investigating the association between plant diversity and soil characteristics.

2 The OSCAR

2.1 Formulation

Assume that the response has been centered and each predictor has been standardized so that

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

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 defined as

$$\begin{gathered} \hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \textstyle\sum_{j=1}^{p} \beta_j \mathbf{x}_j||^2 \\ \text{subject to} \\ \textstyle\sum_{j=1}^{p} |\beta_j| + c \ \sum_{j<k} \max\{|\beta_j|, |\beta_k|\} \leq t, \end{gathered} \tag{2.1}$$

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 clustering. Overall, the OSCAR optimization formulation encourages a sparse solution in terms of the number of unique non-zero coefficients.

The geometric interpretation of the constrained least squares solutions illustrates how this penalty simultaneously encourages sparsity and clustering. Aside from a constant, the contours of the sum-of-squares loss function,

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^0)^T \mathbf{X}^T \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^0), \tag{2.2}$$

are ellipses centered at the OLS solution, $\hat{\boldsymbol{\beta}}^0$. Since the predictors are standardized, when $p = 2$ the principal axis of the contours are at $\pm 45°$ 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; note that the ridge 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.

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 clustering is more likely to occur if the predictors are highly correlated. 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.

Note that choosing $c = 0$ in the OSCAR yields the LASSO, which gives pure shrinkage and no clustering, while letting $c \to \infty$ gives a square penalty region and pure clustering. 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 OSCAR as a Bayes estimate

The OSCAR formulation as a constrained optimization problem (2.1) can be written in the penalized form

$$\begin{aligned} \hat{\boldsymbol{\beta}} = & \ \arg\min_{\boldsymbol{\beta}} \ \left|\left|\mathbf{y} - \textstyle\sum_{j=1}^{p} \beta_j \mathbf{x}_j\right|\right|^2 + \lambda \left[\textstyle\sum_{j=1}^{p} |\beta_j| + c \ \textstyle\sum_{j<k} \max\{|\beta_j|, |\beta_k|\}\right] \\ = & \ \arg\min_{\boldsymbol{\beta}} \ ||\mathbf{y} - \textstyle\sum_{j=1}^{p} \beta_j \mathbf{x}_j||^2 + \lambda \textstyle\sum_{j=1}^{p} \{c(j-1)+1\}|\beta|_{(j)}, \end{aligned} \tag{2.3}$$

with $|\beta|_{(1)} \leq |\beta|_{(2)} \leq ... \leq |\beta|_{(p)}$, and there exists a direct correspondence between λ and the bound t. The relationship between the two constants is further discussed in the next subsection.

Whereas ridge regression can be viewed as the posterior mode corresponding to independent Gaussian priors on the coefficients, the LASSO can be viewed as the posterior mode with respect to independent exponential priors on the absolute coefficients (i.e. a double exponential, or Laplacian, prior on the coefficients themselves). The

OSCAR penalized regression problem can also be represented as the posterior mode with respect to a prior on the absolute coefficients. The particular prior is a special case of an absolutely continuous multivariate exponential distribution introduced by Weinman (1966) and discussed by Block (1975). This multivariate exponential distribution corresponds to the absolutely continuous part of a subclass of the proposal of Marshall and Olkin (1967). Using the representation in (2.3) and some algebra to obtain a form matching that of Block (1975), it follows that for each pair (λ, c), the density for the OSCAR prior is given by

$$f(\boldsymbol{\beta}|\lambda, c) = K \ \exp\left\{-\frac{\lambda}{2}\sum_{j=1}^{p}\gamma_j \ (p-j+1)\left(|\beta|_{(j)} - |\beta|_{(j-1)}\right)\right\}, \tag{2.4}$$

where $|\beta|_{(0)} \equiv 0$, $\gamma_j = 2+(p+j-2)c$, and the normalizing constant $K = \left(\frac{\lambda}{4}\right)^p \left(\prod_{j=1}^{p}\gamma_j\right)$. Note that there is an additional factor of $1/2^p$ absorbed in the normalizing constant, just as in the double exponential, to give equal probability to all combinations of sign.

2.3 Exact grouping property

In this section, an explicit relation between the choice of the constraint bound t and the penalization parameter λ is 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 is then given by Theorem 1.

Consider the representation of the OSCAR in terms of the penalized least squares criterion (2.3) 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

$$\mathbf{x}_g^* = \sum_{j \in \mathcal{G}_g} \text{sign}(\hat{\beta}_j) \ \mathbf{x}_j. \tag{2.5}$$

This transformation amounts to combining the variables with identical magnitudes of the coefficients by a simple (signed) summation of their values. Form the corresponding summed weights

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

The criterion in (2.3) can be written explicitly in terms of this "active set" of covariates, as

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \ ||\mathbf{y} - \sum_{g=1}^{G} \theta_g \mathbf{x}_g^*||^2 + \lambda \sum_{g=1}^{G} w_g \theta_g, \tag{2.6}$$

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 $g = 1, ..., G$

$$-2\mathbf{x}_g^{*T}(\mathbf{y} - \mathbf{X}^*\hat{\boldsymbol{\theta}}) + \lambda w_g = 0. \tag{2.7}$$

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 (2.7) 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, (2.7) yields

$$\lambda = 2\mathbf{x}_g^{*T}(\mathbf{y} - \mathbf{X}^*\hat{\boldsymbol{\theta}})/w_g. \tag{2.8}$$

The octagonal shape of the constraint region in Figure 1 graphically depicts the exact grouping property of the OSCAR optimization criterion. The following theorem, similar in spirit to that of the approximate grouping property of the Elastic Net, quantifies this exact grouping property in terms of the correlation between covariates.

THEOREM 1 *Given data* $(\mathbf{y}, \mathbf{X})$ *and parameter* λ*, with centered response* $\mathbf{y}$ *and standardized predictors* $\mathbf{X}$*, let* $\hat{\boldsymbol{\beta}}(\lambda, c)$ *be the OSCAR estimate using the tuning parameters* (λ, c)*. Assume that the predictors are signed so that* $\hat{\beta}_i(\lambda, c) \geq 0$ *for all* i*. Let* $\rho_{ij} = \mathbf{x}_i^T \mathbf{x}_j$ *be the sample correlation between covariates* i *and* j*.*

Suppose that both $\hat{\beta}_i(\lambda, c) > 0$ *and* $\hat{\beta}_j(\lambda, c) > 0$ *are distinct from the other* $\hat{\beta}_k$*. Then there exists* c_0 *such that*

$$0 < c_0 \leq 2\lambda^{-1}|\mathbf{y}|\sqrt{2(1-\rho_{ij})}$$

and

$$\hat{\beta}_i(\lambda, c) = \hat{\beta}_j(\lambda, c), \text{ for all } c > c_0.$$

The proof of Theorem 1 is based on the score equations in (2.7), and is given in the Appendix.

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 (2.5), then ρ_{ij} represents the correlation between the groups.

3 Computation and cross-validation

3.1 A computational algorithm

A computational algorithm is now discussed to compute the OSCAR estimate for a given set of tuning parameters (t, c). Let $\beta_j = \beta_j^+ - \beta_j^-$ with both β_j^+ and β_j^- being nonnegative, and only one is nonzero. Then $|\beta_j| = \beta_j^+ + \beta_j^-$. There are now $2p$ coefficients,

but at least p are identically zero. Suppose now that the covariates were ordered so that the components of the solution to the OSCAR optimization problem (2.1) was in order of non-decreasing magnitude. Then the optimization problem in (2.1) can be rewritten as

$$\begin{gathered}\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \textstyle\sum_{j=1}^{p} \beta_j \mathbf{x}_j||^2 \\ \text{subject to} \\ |\beta_1| \leq |\beta_2| \leq ... \leq |\beta_p| \\ \textstyle\sum_{j=1}^{p} \{c\ (j-1)+1\}(\beta_j^+ + \beta_j^-) \leq t \\ \boldsymbol{\beta}^+ \geq 0 \\ \boldsymbol{\beta}^- \geq 0.\end{gathered} \tag{3.1}$$

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. A sequential quadratic programming algorithm is proposed to solve this problem 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.

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

The algorithm is based on the idea 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.

One could instead start the algorithm with a set of constraints as opposed to the single constraint. 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 smaller than p. Although this makes computation at least feasible, further work at developing computational algorithms are still needed, particularly for very large p.

3.2 Choosing the tuning parameters

Following Tibshirani (1996), considering $(\mathbf{X}, Y)$ jointly under the linear model

$$Y = \mathbf{X}^T \boldsymbol{\beta} + \epsilon,$$

with $E(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, the prediction error is given by

$$\mathrm{PE} = E(Y - \mathbf{X}^T \hat{\boldsymbol{\beta}})^2 = \mathrm{ME} + \sigma^2. \tag{3.2}$$

The mean-squared error (ME) for this model takes the form

$$\mathrm{ME} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T V (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \tag{3.3}$$

where V is the population covariance matrix for $\mathbf{X}$.

In practice, neither PE or ME can be computed directly because both $\boldsymbol{\beta}$ and V are unknown. However, if a validation set is available, one can estimate the prediction error and σ^2 and hence estimate the mean-squared error via the decomposition in (3.2).

Lacking a validation set one can use five-fold or ten-fold cross-validation to estimate the prediction error.

To minimize an estimate of the prediction error, a two-dimensional grid search must be performed over the two parameters (c, t). For a given constant c, one can use $s \in (0, 1]$ to represent the proportion of the least squares value of the corresponding constraint term as the second parameter, i.e.,

$$s = \frac{t}{\sum_{j=1}^{p}\{c(j-1)+1\}|\beta_j^0|}, \tag{3.4}$$

where β_j^0 represents the jth least squares coefficient (or some other method) ordered by increasing magnitude. In using this parametrization, it has been found in practice that the optimal s, as a function of c, is typically close to the one obtained for the LASSO ($c = 0$), regardless of the optimal c. Therefore, it is easiest to start at the LASSO solution and search locally on a restricted grid for the optimal tuning parameters, (c, s).

An alternative to cross-validation would be to use a generalized cross-validation (GCV) statistic or a criteria such as AIC, BIC, or C_p to estimate the prediction error, based solely on fitting the model a single time. In using this form of model selection criteria one would use the estimated degrees of freedom as in Efron et al. (2004).

For the LASSO, 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 blocks of coefficients and use this as a measure of model complexity. The natural estimate of the degrees of freedom for the OSCAR is

$$\mathrm{df}(\hat{y}) = G, \tag{3.5}$$

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 needed in the final

model.

4 Simulation study

A simulation study was run to show that the OSCAR performs well in terms of both prediction accuracy and parsimony compared with the LASSO, Elastic Net, and Ridge Regression. Four examples are considered in this simulation. In each example, data is simulated from the regression model

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

The first three examples were used in the original LASSO paper (Tibshirani, 1996). The fourth is very similar to the grouped variable situation described in the Elastic Net paper (Zou and Hastie, 2005), except that the correlation within a group is not as high as $\rho \approx 0.99$ used there.

For each example, 100 data sets were simulated. 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 as given by (3.3), using the true values for $\boldsymbol{\beta}$ and V.

The four scenarios are given by:

1. In example one, $n = 20$ for each of the training and validation sets and there are eight predictors. The true parameters are $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and $\sigma = 3$, with the covariance matrix V given by $\mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_j) = 0.5^{|i-j|}$.

2. Example two is the same as example one, except that $\beta_j = 0.85$ for all j.

3. In example three, $n = 100$ for each of the training and validation sets and there are 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 the covariance matrix V 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. Tibshirani (1996) uses $2V$ instead of V.

4. In example four, $n = 50$ for each of the training and validation sets 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:

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

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.

***** TABLES 1 AND 2 GO HERE ******

Table 1 summarizes the prediction error for the four examples, while Table 2 summarizes the sparseness and complexity of the model in terms of the unique non-zero coefficients required in the chosen model. This corresponds to the estimate of degrees of freedom in the OSCAR case, as well as the LASSO case. In all examples, the OSCAR produces the least complex model. Meanwhile, the simulations show that the OSCAR is highly competitive in prediction. Its mean squared error is either best or second best in all four examples. In all cases, the OSCAR outperforms the LASSO and is beaten by the Elastic Net only in the fourth example. However as Table 2 shows, the resulting complexity of the Elastic Net model is much greater in this example, owing to the exact grouping effect of the OSCAR. Although ridge regression performs very well in examples two and three, its performance in the first and last examples is poor, and it does not perform variable selection. Overall, the OSCAR appears to compare favorably with the existing approaches at both prediction and model complexity.

5 Analysis of the soil 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.

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.

***** FIGURE 3 GOES HERE ******

Using five-fold cross-validation, the best LASSO model includes seven predictors, including two moderately correlated cation covariates: CEC and potassium (Table 3). The LASSO solution paths as a function of s, the proportion of the OLS norm, for the seven cation-related covariates are plotted in Figure 4a. 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.

***** TABLE 3 GOES HERE ******

***** FIGURE 4 GOES HERE ******

The five-fold cross-validation OSCAR solution (Table 3) 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.

***** FIGURE 5 GOES HERE ******

The pairs of tuning parameters selected by both five-fold cross validation and GCV each have $c = 4$, therefore Figure 5 plots the OSCAR solution paths for fixed $c = 4$ as a function of the proportion of the OLS norm, s, as given in (3.4). Ten-fold and leave-one-out cross-validation along with the AIC and BIC criteria were also used for comparison and the resulting model choices are shown in Table 3. Note that the BIC criterion chose the same model as GCV. 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 because the magnitude of its parameter estimate 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.

6 Discussion

This paper has introduced a new procedure for creating sparsity in regression while simultaneously performing supervised clustering. 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. However, this will result in further computational burden. 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 it may be desirable to only include in the constraint the pairwise L_∞ norms for predictors among the same group. An example of this would be in the ANOVA setting. Here it may make sense to reduce the number of levels of each factor separately. To accomplish this, one may modify the constraint to only include pairwise L_∞ norms for those levels corresponding to a single factor. The effectiveness of this idea on both main effects and interactions is under further investigation.

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. However, in addition to adding or removing variables at each step, more possibilities must be considered as variables may group together or split apart as well. Further research into a more efficient computational algorithm is warranted, particularly upon extension to more complicated models.

Acknowledgement

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

Appendix

Proof of Theorem 1.

Suppose that $\hat{\beta}_i(\lambda, c) \neq \hat{\beta}_j(\lambda, c)$, then from (2.7) one obtains

$$-2\mathbf{x}_i^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \lambda w_i = 0, \tag{A.1}$$

and

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

Subtracting (A.1) from (A.2) yields

$$-2(\mathbf{x}_j^T - \mathbf{x}_i^T)(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \lambda(w_j - w_i) = 0. \tag{A.3}$$

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

$$|w_j - w_i| \leq 2\lambda^{-1}|\mathbf{y}|\sqrt{2(1 - \rho_{ij})}. \tag{A.4}$$

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.

References

Block, H. W. (1975), Continuous multivariate exponential extensions, in *Reliability and Failure Tree Analysis*, Ed. R. E. Barlow, J. B. Fussel, and N. Singpurwalla, pp. 285-306, Philadelphia: SIAM.

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

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

Hoerl, A. E. and Kennard, R. (1970), Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55-67.

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, *J. Amer. Statist. Assoc.*, **62**, 30-44.

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

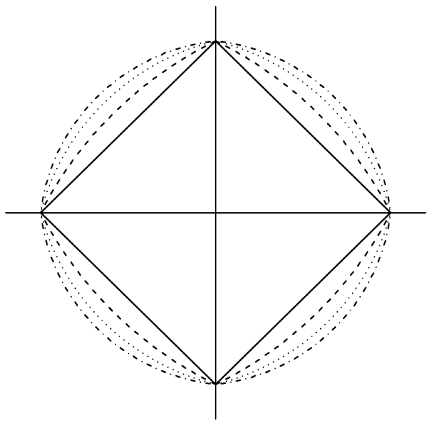
Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005), Sparsity and smoothness via the fused lasso, *J. R. Statist. Soc.* B, **67**, 91-108.

Weinman, D. G. (1966), A multivariate extension of the exponential distribution, Ph.D. thesis, Arizona State University.

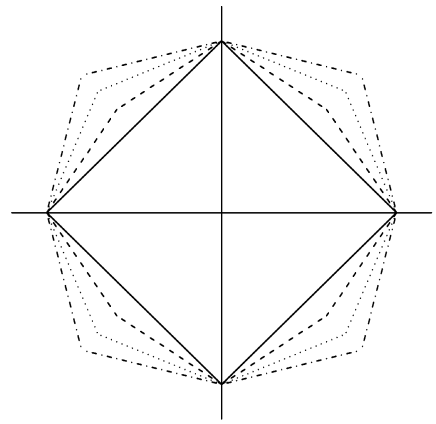
Zou, H. and Hastie, T. (2005), Regularization and variable selection via the elastic net, *J. R. Statist. Soc.* B, **67**, 301-320.

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

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.

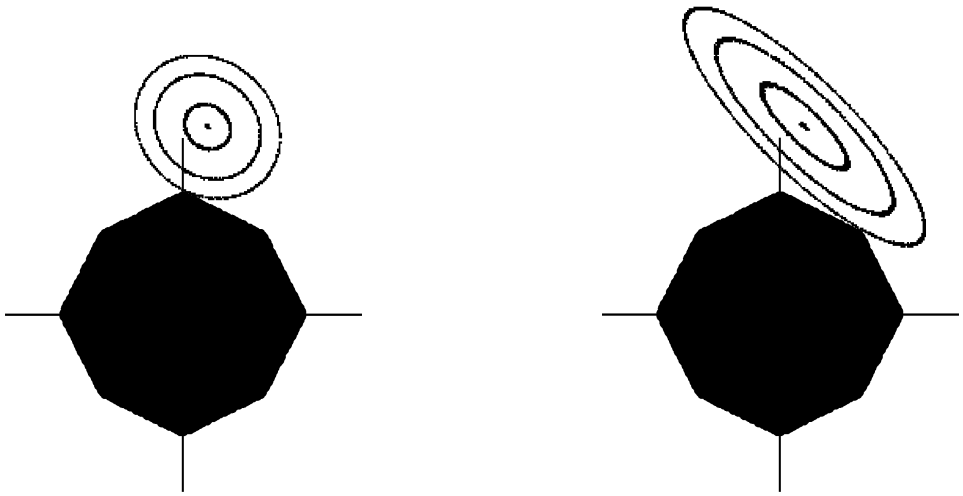


(a) Constraint region for the Lasso (solid line), along with three choices of tuning parameter for the Elastic Net.



(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.

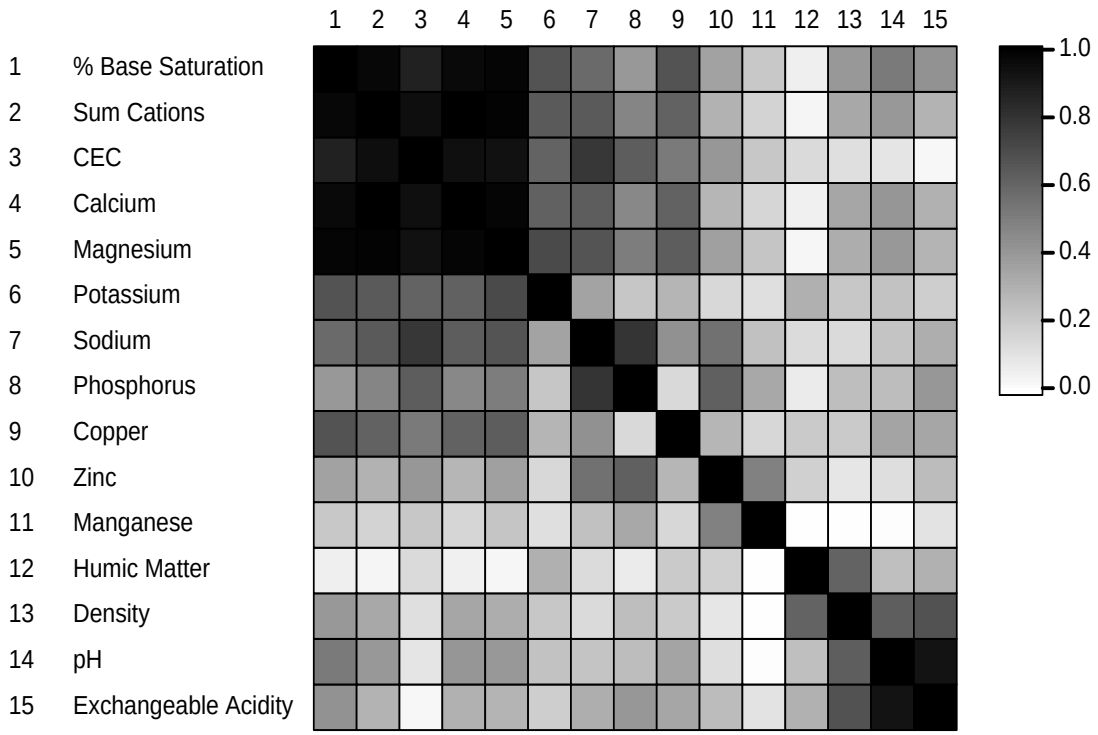
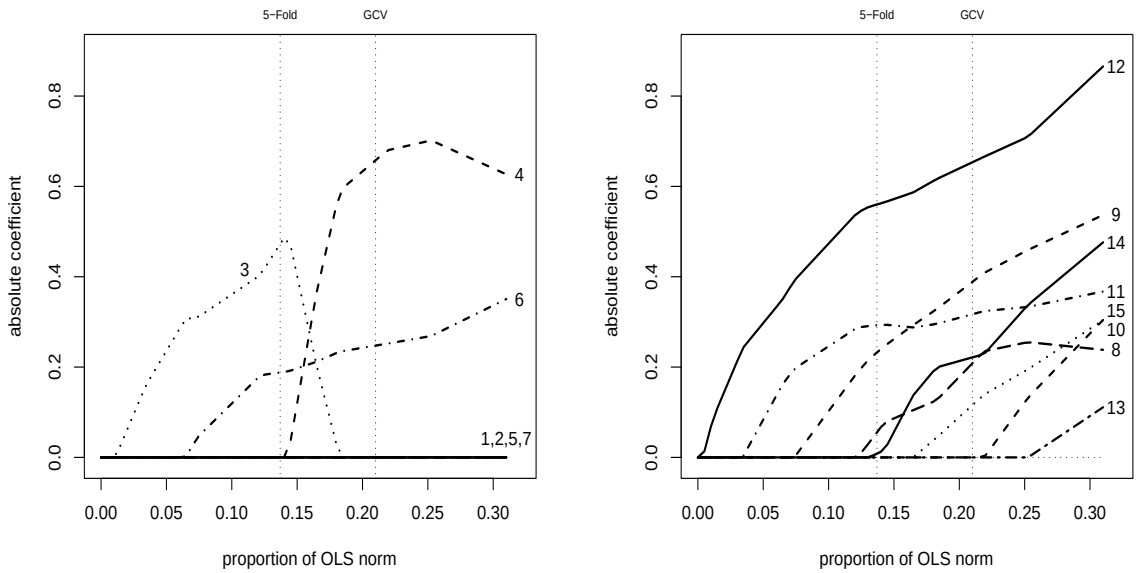


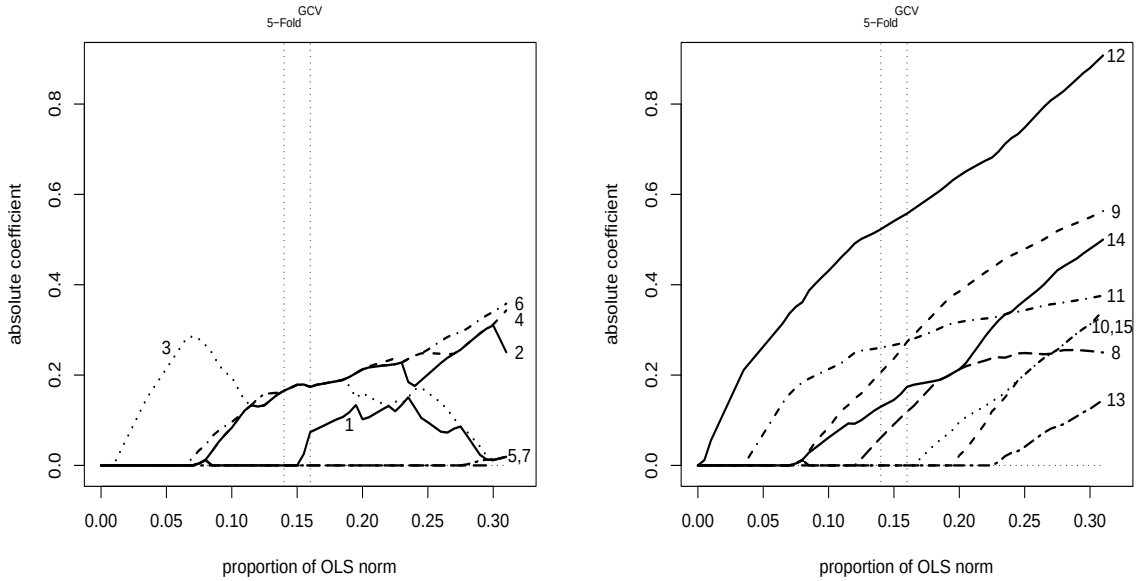
Figure 4: LASSO solution paths for the soil data. Absolute value of the 15 coefficients as a function of s, the proportion of the OLS norm, for the fixed value of $c = 0$, the LASSO. The vertical lines represent the best LASSO models in terms of the GCV and the 5-fold cross-validation criteria.



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

(b) Solution paths for the remaining 8 coefficients.

Figure 5: OSCAR solution paths for the soil data. Absolute value of the 15 coefficients as a function of s, the proportion of the OLS norm, for the value of $c = 4$ as chosen by both GCV and 5-fold cross-validation. The vertical lines represent the selected models based on the two criteria.



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

(b) Solution paths for the remaining 8 coefficients.

Table 1: *Median mean-squared errors for the simulated examples based on 100 replications. Standard errors estimated via the bootstrap are in parentheses.*

	Example 1	Example 2	Example 3	Example 4
Ridge Regression	3.33 (0.39)	2.21 (0.18)	27.4 (1.17)	70.2 (3.05)
Lasso	2.86 (0.25)	3.19 (0.17)	45.4 (1.52)	64.7 (3.03)
Elastic Net	3.07 (0.27)	2.77 (0.17)	34.4 (1.72)	40.7 (3.40)
Oscar	2.75 (0.24)	2.25 (0.19)	25.9 (1.26)	51.8 (2.92)

Table 2: *Median number of unique non-zero coefficients (number of parameters in the final model).*

	Example 1	Example 2	Example 3	Example 4
Ridge Regression	8	8	40	40
Lasso	5	6	21	12
Elastic Net	6	7	25	17
Oscar	5	5	15	12

Table 3: *Models chosen by OSCAR using various selection criteria for soil data example. Each set of variables in parenthesis denotes a group selection. The variables are numbered as in Figure 3. The best LASSO model is also shown for two criteria.*

	Variables Chosen
GCV	(2, 3, 4, 6, 14) (9, 11) (1) (8) (12)
5-fold CV	(2, 3, 4, 6) (8) (9) (11) (12) (14)
10-fold CV	(2, 3, 4, 6, 14) (1) (8) (9) (11) (12)
1-out CV	(2, 3, 4, 6, 14) (1) (8) (9) (11) (12)
AIC	(2, 3, 4, 6, 8, 14) (1) (9) (10) (11) (12)
BIC	(2, 3, 4, 6, 14) (9, 11) (1) (8) (12)
LASSO – GCV	(4) (6) (8) (9) (10) (11) (12) (14)
LASSO – 5-fold CV	(3) (6) (8) (9) (11) (12) (14)