Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors

Patrick Breheny
Department of Biostatistics
University of Iowa

Jian Huang
Department of Statistics and Actuarial Sciences
Department of Biostatistics
University of Iowa

February 6, 2015

Abstract

Penalized regression is an attractive framework for variable selection problems. Often, variables possess a grouping structure, and the relevant selection problem is that of selecting groups, not individual variables. The group lasso has been proposed as a way of extending the ideas of the lasso to the problem of group selection. Nonconvex penalties such as SCAD and MCP have been proposed and shown to have several advantages over the lasso; these penalties may also be extended to the group selection problem, giving rise to group SCAD and group MCP methods. Here, we describe algorithms for fitting these models stably and efficiently. In addition, we present simulation results and real data examples comparing and contrasting the statistical properties of these methods.

1 Introduction

In regression modeling, explanatory variables can often be thought of as grouped. To represent a categorical variable, we may introduce a group of indicator functions. To allow flexible modeling of the effect of a continuous variable, we may introduce a series of basis functions. Or the variables may simply be grouped because the analyst considers them to be similar in some way, or because a scientific understanding of the problem implies that a group of covariates will have similar effects.

Taking this grouping information into account in the modeling process should improve both the interpretability and the accuracy of the model. These gains are likely to be particularly important in high-dimensional settings where sparsity and variable selection play important roles in estimation accuracy.

Penalized likelihood methods for coefficient estimation and variable selection have become widespread since the original proposal of the lasso (Tibshirani, 1996). Building off of earlier work by Bakin (1999), Yuan and Lin (2006) extended the ideas of penalized regression to the problem of grouped covariates. Rather than penalizing individual covariates, Yuan and Lin proposed penalizing norms of groups of coefficients, and called their method the group lasso.

The group lasso, however, suffers from the same drawbacks as the lasso. Namely, it is not consistent with respect to variable selection and tends to over-shrink large coefficients. These shortcomings arise because the rate of penalization of the group lasso does not change with the magnitude of the group coefficients, which leads to biased estimates of large coefficients. To compensate for overshrinkage, the group lasso tends to reduce the level of penalization, allowing spurious coefficients to enter the model.

The smoothly clipped absolute deviation (SCAD) penalty and the minimax concave penalty (MCP) were developed in an effort to achieve what the lasso could not: simultaneous selection consistency and asymptotic unbiasedness (Fan and Li, 2001; Zhang, 2010). This achievement is known as the *oracle property*, so named because it implies that the model is asymptotically equivalent to the fit of a maximum likelihood model in which the identities of the truly nonzero coefficients are known in advance. These properties extend to group SCAD and group MCP models, as shown in Wang et al. (2008) and Huang et al. (2012).

However, group SCAD and group MCP have not been widely used or studied in comparison with the group lasso, largely due to a lack of efficient and publicly available algorithms for fitting these models. Published articles on the group SCAD (Wang et al., 2007, 2008) have used a local quadratic approximation for fitting these models. The local quadratic approximation was originally proposed by Fan and Li (2001) to fit SCAD models. However, by relying on a quadratic approximation, the approach is incapable of producing naturally sparse estimates, and therefore cannot take advantage of the computational benefits provided by sparsity. This, combined with the fact that solving the local quadratic approximation

problem requires the repeated factorization of large matrices, makes the algorithm very inefficient for fitting large regression problems. Zou and Li (2008) proposed a local linear approximation for fitting SCAD models and demonstrated its superior efficiency to local quadratic approximations. This algorithm was further improved upon by Breheny and Huang (2011), who demonstrated how a coordinate descent approach may used to fit SCAD and MCP models in a very efficient manner capable of scaling up to deal with very large problems.

Here, we show how the approach of Breheny and Huang (2011) may be extended to fit group SCAD and group MCP models. We demonstrate that this algorithm is very fast and stable, and we provide a publicly available implementation in the `grpreg` package, (`http://cran.r-project.org/web/packages/grpreg/index.html`). In addition, we provide examples involving both simulated and real data which demonstrate the potential advantages of group SCAD and group MCP over the group lasso.

2 Group descent algorithms

We consider models in which the relationship between the explanatory variables, which consist of J non-overlapping groups, and the outcome is specified in terms of a linear predictor $\boldsymbol{\eta}$:

$$\boldsymbol{\eta} = \beta_0 + \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j, \tag{2.1}$$

where $\mathbf{X}_j$ is the portion of the design matrix formed by the predictors in the jth group and the vector $\boldsymbol{\beta}_j$ consists of the associated regression coefficients. Letting K_j denote the number of members in group j, $\mathbf{X}_j$ is an $n \times K_j$ matrix with elements (x_{ijk}), the value of kth covariate in the jth group for the ith subject. Covariates that do not belong to any group may be thought of as a group of one.

The problem of interest involves estimating a vector of coefficients $\boldsymbol{\beta}$ using a loss function L which quantifies the discrepancy between y_i and η_i combined with a penalty that encourages sparsity and prevents overfitting; specifically, we estimate $\boldsymbol{\beta}$ by minimizing

$$Q(\boldsymbol{\beta}) = L(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) + \sum_{j=1}^{J} p_\lambda \left(\left\| \boldsymbol{\beta}_j \right\| \right), \tag{2.2}$$

where $p(\cdot)$ is a penalty function applied to the Euclidean norm (L_2 norm) of $\boldsymbol{\beta}_j$. The penalty is indexed by a regularization parameter λ, which controls the tradeoff between loss and penalty. It is not necessary for λ to be the same for each group; *i.e.*, we may consider a collection of regularity parameters $\{\lambda_j\}$. For example, in practice there are often variables known to be related to the outcome and therefore which we do not wish to include in selection or penalization. The above framework and algorithms which follow may be easily extended to include such variables by setting $\lambda_j = 0$.

In this section, we primarily focus on linear regression, where $\mathbb{E}(\mathbf{y}) = \boldsymbol{\eta}$ and L is the least squares loss function, but take up the issue of logistic regression in Section 2.4, where L arises from the binomial likelihood. For linear regression, we assume without loss of generality that all variables have been centered to have zero mean; in this case, $\widehat{\beta}_0 = 0$ and may be ignored.

2.1 Orthonormalization

The algorithm for fitting group lasso models originally proposed by Yuan and Lin (2006) requires the groups $\mathbf{X}_j$ to be orthonormal, as does the extension to logistic regression proposed in Meier et al. (2008). It is somewhat unclear from these papers, however, whether orthonormality is a necessary precondition for applying the group lasso, and if not, how one should go about applying the group lasso to non-orthonormal groups.

This question was explored in a recent work by Simon and Tibshirani (2011). In that paper, the authors provide a number of compelling arguments, both theoretical and empirical, that for the group lasso, proper standardization involves orthonormalizing the groups prior to penalizing their L_2 norms. In particular, they demonstrate that the resulting variable selection procedure is closely connected to uniformly most powerful invariant testing.

As we demonstrate in this article, orthonormalizing the groups also produces tremendous advantages in terms of developing algorithms to fit group penalized models. In addition to greatly reducing the computational burden associated with fitting group lasso models, orthonormalization also leads to straightforward extensions for fitting group SCAD and group MCP models. Importantly, as we will see, this orthonormalization can be accomplished without loss of generality since the resulting solutions can be transformed back to the original scale after fitting the model. Thus, it is not necessary for an analyst fitting these group penalization models to have orthonormal groups or to worry about issues of orthonormality when applying these algorithms or using our `grpreg` software.

Taking the singular value decomposition of the Gram matrix of the jth group, we have

$$\frac{1}{n}\mathbf{X}_j^T\mathbf{X}_j = \mathbf{Q}_j\mathbf{\Lambda}_j\mathbf{Q}_j^T,$$

where $\mathbf{\Lambda}_j$ is a diagonal matrix containing the eigenvalues of $n^{-1}\mathbf{X}_j^T\mathbf{X}_j$ and $\mathbf{Q}_j$ is an orthonormal matrix of its eigenvectors. Now, we may construct a linear transformation $\widetilde{\mathbf{X}}_j = \mathbf{X}_j\mathbf{Q}_j\mathbf{\Lambda}_j^{-1/2}$ with the following properties:

$$\frac{1}{n}\widetilde{\mathbf{X}}_j^T\widetilde{\mathbf{X}}_j = \mathbf{I} \tag{2.3}$$

$$\widetilde{\mathbf{X}}_j\widetilde{\boldsymbol{\beta}}_j = \mathbf{X}_j(\mathbf{Q}_j\mathbf{\Lambda}_j^{-1/2}\widetilde{\boldsymbol{\beta}}_j), \tag{2.4}$$

where $\mathbf{I}$ is the identity matrix and $\widetilde{\boldsymbol{\beta}}_j$ is the solution to (2.2) on the orthonormalized scale. In other words, if we have the solution to the orthonormalized problem, we may easily transform back to the original problem with $\boldsymbol{\beta}_j = \mathbf{Q}_j\mathbf{\Lambda}_j^{-1/2}\widetilde{\boldsymbol{\beta}}_j$. This procedure is not terribly expensive from a computational standpoint, as the decompositions are being applied only to the groups, not the entire design matrix, and the inverses are of course trivial to compute because $\mathbf{\Lambda}_j$ is diagonal. Furthermore, the decompositions need only be computed once initially, not with every iteration.

Note that this procedure may be applied even when $\mathbf{X}_j$ is not full-rank by omitting the zero eigenvalues and their associated eigenvectors. In this case, $\mathbf{Q}_j$ is a $K_j \times r_j$ matrix and $\mathbf{\Lambda}_j$ is an $r_j \times r_j$ matrix, where r_j denotes the rank of $\mathbf{X}_j$. Given these modifications, $\mathbf{\Lambda}_j$ is invertible and $\widetilde{\mathbf{X}}_j = \mathbf{X}_j\mathbf{Q}_j\mathbf{\Lambda}_j^{-1/2}$ still possesses properties (2.3) and (2.4). Note, however, that $\widetilde{\mathbf{X}}_j$ now contains only r_j columns and by extension, $\widetilde{\boldsymbol{\beta}}_j$ contains only r_j elements. Thus, we avoid the problem of incomplete rank by fitting the model in a lower-dimensional parameter space, then transforming back to the original dimensions (note that applying the reverse transformation results in a $\boldsymbol{\beta}_j$ with appropriate dimension K_j). In our experience, it is not uncommon for non-full-rank groups to arise in problems for which one wishes to use group penalized models, and the SVD approach we describe here handles such cases naturally. For example, in the special case where two members of a group are identical, $\mathbf{x}_{jk} = \mathbf{x}_{j\ell}$, the approach ensures that their coefficients, $\hat{\beta}_{jk}$ and $\hat{\beta}_{j\ell}$, will be equal for all values of λ.

For the remainder of this article, we will assume that this process has been applied and that the model fitting algorithms we describe are being applied to orthonormal groups (by "orthonormal groups", we mean groups for which $n^{-1}\mathbf{X}_j^T\mathbf{X}_j = \mathbf{I}$, not that groups $\mathbf{X}_j$ and $\mathbf{X}_k$ are orthogonal to each other). Consequently, we drop the tildes on $\widetilde{\mathbf{X}}$ and $\widetilde{\boldsymbol{\beta}}$ in subsequent sections.

It is worth noting that penalizing the norm of $\widetilde{\boldsymbol{\beta}}_j$ is not equivalent to penalizing the norm of the original coefficients. As pointed out in Simon and Tibshirani (2011) and Huang et al. (2012),

$$\begin{aligned}\left\|\widetilde{\boldsymbol{\beta}}_j\right\| &= \sqrt{\frac{1}{n}\boldsymbol{\beta}_j^T\mathbf{X}_j^T\mathbf{X}_j\boldsymbol{\beta}_j} \\ &\propto \left\|\boldsymbol{\eta}_j\right\|,\end{aligned}$$

where $\boldsymbol{\eta}_j = \mathbf{X}_j\boldsymbol{\beta}_j$. In other words, orthonormalizing the groups is equivalent to applying an L_2 penalty on the scale of the linear predictor. The idea of penalizing on the scale of the linear predictors is also explored in Ravikumar et al. (2009). The two penalties are equivalent in the non-grouped case, provided that the standard normalization $n^{-1}\sum_i x_{ijk}^2 = 1$ has been applied. However, for grouped regression models, this normalization at the coefficient level is inadequate; orthonormalization at the group level is appropriate.

2.2 Group lasso

In this section, we describe the group lasso and algorithms for fitting group lasso models. The group lasso estimator, originally proposed by Yuan and Lin (2006), is defined as the value $\widehat{\boldsymbol{\beta}}$ that minimizes

$$Q(\boldsymbol{\beta}) = \frac{1}{2n}\left\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right\|^2 + \lambda\sum_j \sqrt{K_j}\left\|\boldsymbol{\beta}_j\right\|. \tag{2.5}$$

The idea behind the penalty is to apply a lasso penalty to the Euclidean (L_2) norm of each group, thereby encouraging sparsity and variable selection at the group level. The solution $\widehat{\boldsymbol{\beta}}$ has the property that if group j is selected, then $\hat{\beta}_{jk} \neq 0$ for all k, otherwise $\hat{\beta}_{jk} = 0$ for all k. The correlation between $\mathbf{X}_j$ and the residuals will tend to be larger if group j has more elements; the presence of the $\sqrt{K_j}$ term in the penalty compensates for this, thereby normalizing across groups of different sizes. As discussed in Simon and Tibshirani (2011), this results in variable selection which is roughly equivalent to

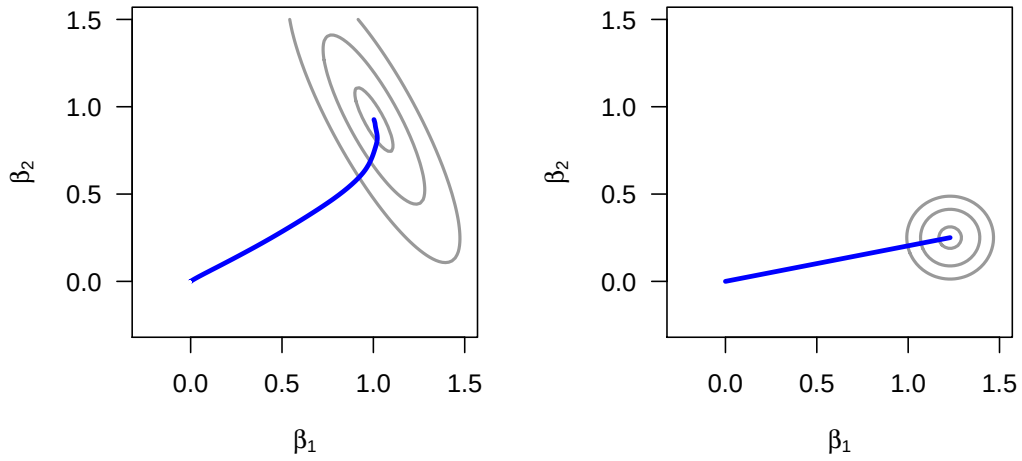


Figure 1: The impact of orthonormalization on the solution to the group lasso. Contour lines for the likelihood (least squares) surface are drawn, centered around the OLS solution, as well as the solution path for the group lasso as λ goes from 0 to ∞. *Left*: Non-orthonormal $\mathbf{X}$. *Right*: Orthonormal $\mathbf{X}$.

the uniformly most powerful invariant test for inclusion of the jth group. In what follows, we will absorb the $\sqrt{K_j}$ term into λ and use $\lambda_j = \lambda\sqrt{K_j}$.

Yuan and Lin (2006) also propose an algorithm which they base on the "shooting algorithm" of Fu (1998). Here, we refer to this type of algorithm as a "group descent" algorithm. The idea behind the algorithm is the same as that of coordinate descent algorithms (Friedman et al., 2007; Wu and Lange, 2008), with the modification that the optimization of (2.5) takes place repeatedly with respect to a group $\boldsymbol{\beta}_j$ rather than an individual coordinate β_j.

Below, we present the group descent algorithm for solving (2.5) to obtain the group lasso estimator. The algorithm is essentially the same as Yuan and Lin's, although (a) we make explicit its generalization to the case of non-orthonormal groups using the approach described in Section 2.1, (b) we restate the algorithm to more clearly illustrate the connections with coordinate descent algorithms, and (c) we employ techniques developed in the coordinate descent literature to speed up the implementation of the algorithm considerably. As we will see, this presentation of the algorithm also makes clear how to easily extend it to fit group SCAD and group MCP models in the following sections.

We begin by noting that the subdifferential (Bertsekas, 1999) of Q with respect to $\boldsymbol{\beta}_j$ is given by

$$\partial Q(\boldsymbol{\beta}_j) = \begin{cases} -\mathbf{z}_j + \boldsymbol{\beta}_j + \lambda_j \boldsymbol{\beta}_j / \left\|\boldsymbol{\beta}_j\right\| & \text{if } \boldsymbol{\beta}_j \neq \mathbf{0} \\ -\mathbf{z}_j + \lambda_j \mathbf{v} & \text{if } \boldsymbol{\beta}_j = \mathbf{0} \end{cases} \tag{2.6}$$

where $\mathbf{z}_j = \mathbf{X}_j^T(\mathbf{y} - \mathbf{X}_{-j}\boldsymbol{\beta}_{-j})$ is the least squares solution, $\mathbf{X}_{-j}$ is the portion of the design matrix that remains after $\mathbf{X}_j$ has been excluded, $\boldsymbol{\beta}_{-j}$ are its associated regression coefficients, and $\mathbf{v}$ is any vector satisfying $\|\mathbf{v}\| \leq 1$. The main implication of (2.6) is that, by orthonormalizing the groups, $\mathbf{X}_j$ drops out of the equation and the multivariate problem of optimizing with respect to $\boldsymbol{\beta}_j$ is reduced to a univariate problem, as the solution must lie on the line segment joining $\mathbf{0}$ and $\mathbf{z}_j$. The geometry of this problem is illustrated in Figure 1. As the figure illustrates, orthonormalization renders the direction of $\widehat{\boldsymbol{\beta}}_j$ invariant with respect to λ, thereby enabling us to break the solution down into two components: determining the direction of $\boldsymbol{\beta}_j$ and determining its length.

Furthermore, determining the length of $\boldsymbol{\beta}_j$ is equivalent to solving a one-dimensional lasso problem, which has a simple, closed-form solution given by the soft-thresholding operator (Donoho and Johnstone, 1994):

$$S(z, \lambda) = \begin{cases} z - \lambda & \text{if } z > \lambda \\ 0 & \text{if } |z| \leq \lambda \\ z + \lambda & \text{if } z < -\lambda. \end{cases} \tag{2.7}$$

With a slight abuse of notation, we extend this definition to a vector-valued argument $\mathbf{z}$ as follows:

$$S(\mathbf{z}, \lambda) = S(\|z\|, \lambda)\frac{\mathbf{z}}{\|\mathbf{z}\|}, \tag{2.8}$$

where $\mathbf{z}/\|\mathbf{z}\|$ is the unit vector in the direction of $\mathbf{z}$. In other words, $S(\mathbf{z}, \lambda)$ acts on a vector $\mathbf{z}$ by shortening it towards $\mathbf{0}$ by an amount λ, and if the length of $\mathbf{z}$ is less than λ, the vector is shortened all the way to $\mathbf{0}$.

The multivariate soft-thresholding operator is the solution to the single-group group lasso, just as the univariate soft-thresholding operator is the solution to the single-variable lasso problem. This leads to Algorithm 1, which is exactly the same as the coordinate descent algorithm described in Friedman et al. (2010b) for fitting lasso-penalized regression models, only with multivariate soft-thresholding replacing univariate soft-thresholding. Both algorithms are essentially modified backfitting algorithms, with soft-thresholding replacing the usual least squares updating.

Algorithm 1 Group descent algorithm for the group lasso

repeat
  for $j = 1, 2, \ldots, J$
    $\mathbf{z}_j = \mathbf{X}_j^T \mathbf{r} + \boldsymbol{\beta}_j$
    $\boldsymbol{\beta}_j' \leftarrow S\left(\mathbf{z}_j, \lambda_j\right)$
    $\mathbf{r}' \leftarrow \mathbf{r} - \mathbf{X}_j(\boldsymbol{\beta}_j' - \boldsymbol{\beta}_j)$
until convergence

In Algorithm 1, $\boldsymbol{\beta}_j$ refers to the current (*i.e.*, most recently updated) value of coefficients in the jth group prior to the execution of the for loop; during the loop, $\boldsymbol{\beta}_j$ is updated to $\boldsymbol{\beta}_j'$. The same notation is applied to $\mathbf{r}$, where $\mathbf{r}$ refers to the residuals: $\mathbf{r} = \mathbf{y} - \sum_j \mathbf{X}_j \boldsymbol{\beta}_j$. The "$\leftarrow$" refers to the fact that $\boldsymbol{\beta}_j$ and $\mathbf{r}$ are being continually updated; at convergence, $\widehat{\boldsymbol{\beta}}$ consists of the final updates $\{\boldsymbol{\beta}_j\}$. The expression $\mathbf{z}_j = \mathbf{X}_j^T \mathbf{r} + \boldsymbol{\beta}_j$ is derived from

$$\mathbf{z}_j = \mathbf{X}_j^T(\mathbf{y} - \mathbf{X}_{-j}\boldsymbol{\beta}_{-j}) = \mathbf{X}_j^T \mathbf{r} + \boldsymbol{\beta}_j; \tag{2.9}$$

the algorithm is implemented in this manner because it is more efficient computationally to update $\mathbf{r}$ than to repeatedly calculate the partial residuals $\mathbf{y} - \mathbf{X}_{-j}\boldsymbol{\beta}_{-j}$, especially in high dimensions.

The computational efficiency of Algorithm 1 is clear: no complicated numerical optimization steps or matrix factorizations or inversions are required, only a small number of simple arithmetic operations. This efficiency is possible only because the groups $\{\mathbf{X}_j\}$ are made to be orthonormal prior to model fitting. Without this initial orthonormalization, we cannot obtain the simple closed-form solution (2.8), and the updating steps required to fit the group lasso become considerably more complicated, as in Friedman et al. (2010a), Foygel and Drton (2010), and Puig et al. (2011).

2.3 Group MCP and group SCAD

We have just seen how the group lasso may be viewed as applying the lasso/soft-thresholding operator to the length of each group. Not only does this formulation lead to a very efficient algorithm, it also makes it clear how to extend other univariate penalties to the group setting. Here, we focus on two popular alternative univariate penalties to the lasso: SCAD, the smoothly clipped absolute deviation penalty (Fan and Li, 2001) and MCP, the minimax concave penalty (Zhang, 2010).

The two penalties are similar in motivation, definition, and performance. The penalties are defined on $[0, \infty)$ for $\lambda > 0$ as follows, and plotted on the left side of Figure 2:

$$\text{SCAD:} \quad p_{\lambda,\gamma}(\theta) = \begin{cases} \lambda\theta & \text{if } \theta \le \lambda \\ \frac{\gamma\lambda\theta - 0.5(\theta^2+\lambda^2)}{\gamma-1} & \text{if } \lambda < \theta \le \gamma\lambda \\ \frac{\lambda^2(\gamma^2-1)}{2(\gamma-1)} & \text{if } \theta > \gamma\lambda \end{cases} \tag{2.10}$$

$$\text{MCP:} \quad p_{\lambda,\gamma}(\theta) = \begin{cases} \lambda\theta - \frac{\theta^2}{2\gamma} & \text{if } \theta \le \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2 & \text{if } \theta > \gamma\lambda \end{cases}. \tag{2.11}$$

To have a well-defined minimum, we must have $\gamma > 1$ for MCP and $\gamma > 2$ for SCAD. Although originally proposed for univariate penalized regression, these penalties may be extended to the grouped-variable selection problem by substituting (2.10) and (2.11) into (2.2), as has been proposed in Wang et al. (2007) and discussed in Huang et al. (2012). We refer to these penalized regression models as the Group SCAD and Group MCP methods, respectively.

The rationale behind the penalties can be understood by considering their derivatives, which appear in the middle panel of Figure 2. MCP and SCAD begin by applying the same rate of penalization as the lasso, but continuously relax that penalization until the point at which $\theta = \gamma\lambda$, where the rate of penalization has fallen all the way to 0. The aim of both penalties is to achieve the variable selection properties of the lasso, but to introduce less bias towards zero among the true nonzero coefficients. The only difference between the two is that MCP reduces the rate of penalization immediately, while SCAD remains flat for a while before moving towards zero.

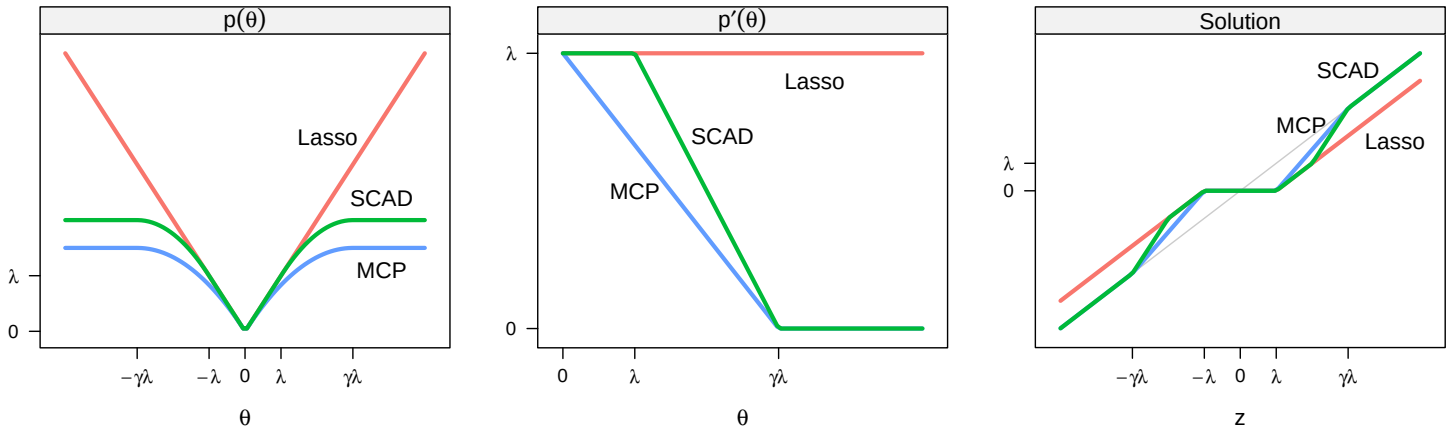


Figure 2: Lasso, SCAD, and MCP penalty functions, derivatives, and univariate solutions. The panel on the left plots the penalties themselves, the middle panel plots the first derivative of the penalty, and the right panel plots the univariate solutions as a function of the ordinary least squares estimate. The light gray line in the rightmost plot is the identity line. Note that none of the penalties are differentiable at $\beta_j = 0$.

The rationale behind the penalties can also be understood by considering their univariate solutions. Consider the simple linear regression of $\mathbf{y}$ upon $\mathbf{x}$, with unpenalized least squares solution $z = n^{-1}\mathbf{x}^T\mathbf{y}$. For this simple linear regression problem, the MCP and SCAD estimators have the following closed forms:

$$\widehat{\beta} = F(z, \lambda, \gamma) = \begin{cases} \frac{S(z,\lambda)}{1-1/\gamma} & \text{if } |z| \leq \gamma\lambda \\ z & \text{if } |z| > \gamma\lambda, \end{cases} \tag{2.12}$$

$$\widehat{\beta} = F_S(z, \lambda, \gamma) = \begin{cases} S(z,\lambda) & \text{if } |z| \leq 2\lambda \\ \frac{S(z,\gamma\lambda/(\gamma-1))}{1-1/(\gamma-1)} & \text{if } 2\lambda < |z| \leq \gamma\lambda \\ z & \text{if } |z| > \gamma\lambda. \end{cases} \tag{2.13}$$

Noting that $S(z, \lambda)$ is the univariate solution to the lasso, we can observe by comparison that MCP and SCAD scale the lasso solution upwards toward the unpenalized solution by an amount that depends on γ. For both MCP and SCAD, when $|z| > \gamma\lambda$, the solution is scaled up fully to the unpenalized least squares solution. These solutions are plotted in the right panel of Figure 2; the figure illustrates how the solutions, as a function of z, make a smooth transition between the lasso and least squares solutions. This transition is essential to the oracle property described in the introduction.

As $\gamma \to \infty$, the MCP and lasso solutions are identical. As $\gamma \to 1$, the MCP solution becomes the hard thresholding estimate $zI_{|z|>\lambda}$. Thus, in the univariate sense, the MCP produces the "firm shrinkage" estimator of Gao and Bruce (1997); hence the $F(\cdot)$ notation. The SCAD solutions are similar, of course, but not identical, and thus involve a "SCAD-modified firm thresholding" operator which we denote $F_S(\cdot)$. In particular, the SCAD solutions also have soft-thresholding as the limiting case when $\gamma \to \infty$, but do not have hard thresholding as the limiting case when $\gamma \to 2$.

We extend these two firm-thresholding operators to multivariate arguments as in (2.8), with $F(\cdot)$ or $F_S(\cdot)$ taking the place of $S(\cdot)$, and note that $F(\mathbf{z}_j, \lambda, \gamma)$ and $F_S(\mathbf{z}_j, \lambda, \gamma)$ optimize the objective functions for Group MCP and Group SCAD, respectively, with respect to $\boldsymbol{\beta}_j$. An illustration of the nature of these estimators is given in Figure 3. We note the following: (1) All estimators carry out group selection, in the sense that, for any value of λ, the coefficients belonging to a group are either wholly included or wholly excluded from the model. (2) The group MCP and group SCAD methods eliminate some of the bias towards zero introduced by the group lasso. In particular, at $\lambda \approx 0.2$, they produce the same estimates as a least squares regression model including only the nonzero covariates (the "oracle" model). (3) Group MCP makes a smoother transition from $\mathbf{0}$ to the unpenalized solutions than group SCAD. This is the "minimax" aspect of the penalty. Any other penalty that makes the transition between these two extremes must have some region (*e.g.* $\lambda \in [0.7, 0.5]$ for group SCAD) in which its solutions are changing more abruptly than those of group MCP.

It is straightforward to extend Algorithm 1 to fit Group SCAD and Group MCP models; all that is needed is to replace the multivariate soft-thresholding update with a multivariate firm-thresholding update. The group updates for all three

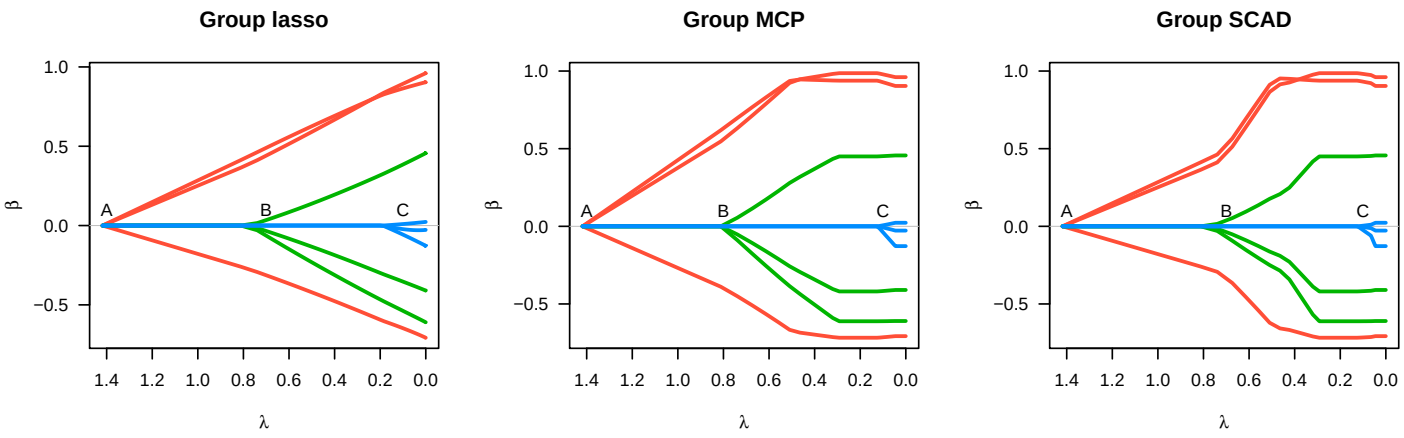


Figure 3: Representative solution paths for the group lasso, group MCP, and group SCAD methods. In the generating model, groups A and B have nonzero coefficients and while those belonging to group C are zero.

methods are listed below:

$$\text{Group lasso} : \boldsymbol{\beta}_j \leftarrow S(\mathbf{z}_j, \lambda_j) = S(\|\mathbf{z}_j\|, \lambda_j)\frac{\mathbf{z}_j}{\|\mathbf{z}_j\|}$$
$$\text{Group MCP} : \boldsymbol{\beta}_j \leftarrow F(\mathbf{z}_j, \lambda_j, \gamma) = F(\|\mathbf{z}_j\|, \lambda_j, \gamma)\frac{\mathbf{z}_j}{\|\mathbf{z}_j\|}$$
$$\text{Group SCAD} : \boldsymbol{\beta}_j \leftarrow F_S(\mathbf{z}_j, \lambda_j, \gamma) = F_S(\|\mathbf{z}_j\|, \lambda_j, \gamma)\frac{\mathbf{z}_j}{\|\mathbf{z}_j\|};$$

recall that $\lambda_j = \lambda\sqrt{K_j}$. Note that all three updates are simple, closed-form expressions. Furthermore, as each update minimizes the objective function with respect to $\boldsymbol{\beta}_j$, the algorithms possess the descent property, meaning that they decrease the objective function with every iteration. This, along with the fact that Q is a strictly convex function with respect to $\boldsymbol{\beta}_j$ (see Lemma 1 in the Appendix for details) leads to attractive convergence properties, which we now state formally.

Proposition 1. *Let $\boldsymbol{\beta}^{(m)}$ denote the value of the fitted regression coefficient at the end of iteration m. At every iteration of the proposed group descent algorithms for linear regression models involving group lasso, group MCP, or group SCAD penalties,*

$$Q(\boldsymbol{\beta}^{(m+1)}) \leq Q(\boldsymbol{\beta}^{(m)}).$$

Furthermore, every limit point of the sequence $\{\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \ldots\}$ is a stationary point of Q.

For the group lasso penalty, the objective function being minimized is convex, and thus the above proposition establishes convergence to the global minimum. For group MCP and group SCAD, whose objective function is a sum of convex and nonconvex components, convergence to a local minimum is possible.

A similar algorithm was explored in She (2012), who proposed the same updating steps as above, although without an initial orthonormalization. Interestingly, She showed that even without orthonormal groups, the updating steps described above still produce a sequence $\boldsymbol{\beta}^{(m)}$ converging to a stationary point of the objective function. Without orthonormal groups, however, the updates are not exact group-wise solutions. In other words, in the single-group case, our approach (with an initial orthonormalization) produces exact solutions in one step, whereas the approach in She (2012) requires multiple iterations to converge to the same solution. This leads to a considerable loss of efficiency, as we will see in Section 3.

In conclusion, the algorithms we present here for fitting group lasso, group MCP, and group SCAD models are both fast and stable. We examine the empirical running time of these algorithms in Section 3.

2.4 Logistic regression

It is possible to extend the algorithms described above to fit group-penalized logistic regression models as well, where the loss function is the negative log-likelihood of a binomial distribution:

$$L(\boldsymbol{\eta}) = \frac{1}{n}\sum_i L_i(\eta_i) = -\frac{1}{n}\sum_i \log \mathbb{P}(y_i|\eta_i).$$

Recall, however, that the simple, closed form solutions of the previous sections were possible only with orthonormalization. The iteratively reweighted least squares (IRLS) algorithm typically used to fit generalized linear models (GLMs) introduces a $n^{-1}\mathbf{X}^T\mathbf{W}\mathbf{X}$ term into the score equation (2.6), where $\mathbf{W}$ is an $n \times n$ diagonal matrix of weights. Because $n^{-1}\mathbf{X}^T\mathbf{W}\mathbf{X} \neq \mathbf{I}$, the group lasso, group MCP, and group SCAD solutions will lack the simple closed forms of the previous section.

However, we may preserve the sphericity of the likelihood surface (Figure 1) through the application of a majorization-minimization (MM) approach (Lange et al., 2000; Hunter and Lange, 2004). In the context of penalized logistic regression, this approach was proposed by Krishnapuram et al. (2005), who referred to it as a *bound optimization algorithm*. The application of the method depends on the ability to bound the Hessian of the loss function with respect to the linear predictor $\boldsymbol{\eta}$. Let $v = \max_i \sup_\eta \{\nabla^2 L_i(\eta)\}$, so that $v\mathbf{I} - \nabla^2 L(\boldsymbol{\eta})$ is a positive semidefinite matrix at all points $\boldsymbol{\eta}$. For logistic regression, where

$$\pi_i = \mathbb{P}(Y_i = 1|\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}},$$

we may easily obtain $v = 1/4$, since $\nabla^2 L_i(\eta) = \pi(1 - \pi)$.

Bounding the loss function in this manner allows us to define

$$\tilde{L}(\boldsymbol{\eta}|\boldsymbol{\eta}^*) = L(\boldsymbol{\eta}^*) + (\boldsymbol{\eta} - \boldsymbol{\eta}^*)^T \nabla L(\boldsymbol{\eta}^*) + \frac{v}{2}(\boldsymbol{\eta} - \boldsymbol{\eta}^*)^T(\boldsymbol{\eta} - \boldsymbol{\eta}^*)$$

such that the function $\tilde{L}(\boldsymbol{\eta}|\boldsymbol{\eta}^*)$ has the following two properties:

$$\begin{aligned} \tilde{L}(\boldsymbol{\eta}^*|\boldsymbol{\eta}^*) &= L(\boldsymbol{\eta}^*) \\ \tilde{L}(\boldsymbol{\eta}|\boldsymbol{\eta}^*) &\geq L(\boldsymbol{\eta}). \end{aligned}$$

Thus, $\tilde{L}(\boldsymbol{\eta}|\boldsymbol{\eta}^*)$ is a majorizing function of $L(\boldsymbol{\eta})$. The theory underlying MM algorithms then ensures that Algorithm 2, which consists of alternately executing the majorizing step and the minimization steps, will retain the descent property of the previous sections, which we formally state below.

Proposition 2. *Let $\boldsymbol{\beta}^{(m)}$ denote the value of the fitted regression coefficient at the end of iteration m. At every iteration of the proposed group descent algorithms for logistic regression involving group lasso, group MCP, or group SCAD penalties,*

$$Q(\boldsymbol{\beta}^{(m+1)}) \leq Q(\boldsymbol{\beta}^{(m)}).$$

Furthermore, provided that no elements of $\boldsymbol{\beta}$ tend to $\pm\infty$, every limit point of the sequence $\{\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \ldots\}$ is a stationary point of Q.

As with linear regression, this result implies convergence to a global minimum for the group lasso, but allows convergence to local minima for group MCP and group SCAD. Note, however, that unlike linear regression, in logistic regression maximum likelihood estimates can occur at $\pm\infty$ (this is often referred to as *complete separation*). In practice, this is not a concern for large values of λ, but saturation of the model will certainly occur when $p > n$ and λ is small. Our implementation in `grpreg` terminates the path-fitting algorithm if saturation is detected, based on a check of whether $> 99\%$ of the null deviance has been explained by the model.

Writing $\tilde{L}(\boldsymbol{\eta}|\boldsymbol{\eta}^*)$ in terms of $\boldsymbol{\beta}$, we have

$$\tilde{L}(\boldsymbol{\beta}) \propto \frac{v}{2n}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^T(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}),$$

where $\tilde{\mathbf{y}} = \boldsymbol{\eta}^* + (\mathbf{y} - \boldsymbol{\pi})/v$ is the pseudo-response vector. Thus, the gradient of $\tilde{L}(\boldsymbol{\eta}|\boldsymbol{\eta}^*)$ with respect to $\boldsymbol{\beta}_j$ is given by

$$\nabla \tilde{L}(\boldsymbol{\beta}_j) = -v\mathbf{z}_j + v\boldsymbol{\beta}_j, \tag{2.14}$$

where, as before, $\mathbf{z}_j = \mathbf{X}_j^T(\tilde{\mathbf{y}} - \mathbf{X}_{-j}\boldsymbol{\beta}_{-j})$ is the unpenalized (maximum likelihood) solution for $\boldsymbol{\beta}_j$.

The presence of the scalar v in the score equations affects the updating equations; however, as the majorized loss function remains spherical with respect to $\boldsymbol{\beta}$, the updating equations still have simple, closed form solutions:

$$\begin{aligned} \text{Group lasso} : \boldsymbol{\beta}_j &\leftarrow \frac{1}{v} S(v\mathbf{z}_j, \lambda_j) = \frac{1}{v} S(v \left\|\mathbf{z}_j\right\|, \lambda_j) \frac{\mathbf{z}_j}{\left\|\mathbf{z}_j\right\|} \\ \text{Group MCP} : \boldsymbol{\beta}_j &\leftarrow \frac{1}{v} F(v\mathbf{z}_j, \lambda_j, \gamma) = \frac{1}{v} F(v \left\|\mathbf{z}_j\right\|, \lambda_j, \gamma) \frac{\mathbf{z}_j}{\left\|\mathbf{z}_j\right\|} \\ \text{Group SCAD} : \boldsymbol{\beta}_j &\leftarrow \frac{1}{v} F_S(v\mathbf{z}_j, \lambda_j, \gamma) = \frac{1}{v} F_S(v \left\|\mathbf{z}_j\right\|, \lambda_j, \gamma) \frac{\mathbf{z}_j}{\left\|\mathbf{z}_j\right\|} \end{aligned}$$

Algorithm 2 Group descent algorithm for logistic regression with a group lasso penalty

repeat
  $\boldsymbol{\eta} \leftarrow \mathbf{X}\boldsymbol{\beta}$
  $\boldsymbol{\pi} \leftarrow \{e^{\eta_i}/(1+e^{\eta_i})\}_{i=1}^n$
  $\tilde{\mathbf{r}} \leftarrow (\mathbf{y}-\boldsymbol{\pi})/v$
  for $j = 1, 2, \ldots, J$
    $\mathbf{z}_j = \mathbf{X}_j^T\tilde{\mathbf{r}} + \boldsymbol{\beta}_j$
    $\boldsymbol{\beta}_j' \leftarrow S\left(v\mathbf{z}_j, \lambda_j\right)/v$
    $\tilde{\mathbf{r}}' \leftarrow \tilde{\mathbf{r}} - \mathbf{X}_j(\boldsymbol{\beta}_j' - \boldsymbol{\beta}_j)$
until convergence

Algorithm 2 is presented for the group lasso, but is easily modified to fit group MCP and group SCAD models by substituting the appropriate expression into the updating step for $\boldsymbol{\beta}_j$.

Note that Proposition 2 does not necessarily follow for other generalized linear models, as the Hessian matrices for other exponential families are typically unbounded. One possibility is to set $v \leftarrow \max_i\{\nabla^2 L_i(\eta_i^*)\}$ at the beginning of each iteration as a pseudo-upper bound. As this is not an actual upper bound, an algorithm based on it is not guaranteed to possess the descent property. Nevertheless, the approach seems to work well in practice. The authors have tested the approach on the group-penalized Poisson regression models and did not observe any problems with convergence, although we have not studied these models as extensively as the logistic regression model.

2.5 Path-fitting algorithm

The above algorithms are presented from the perspective of fitting a penalized regression model for a single value of λ. Usually, we are interested in obtaining $\widehat{\boldsymbol{\beta}}$ for a range of λ values, and then choosing among those models using either cross-validation or some form of information criterion. The regularization parameter λ may vary from a maximum value $\lambda_{\max}$ at which all penalized coefficients are 0 down to $\lambda = 0$ or to a minimum value $\lambda_{\min}$ beyond which the model becomes excessively large. When the objective function is strictly convex, the estimated coefficients vary continuously with $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ and produce a path of solutions regularized by λ. Examples of such paths were seen in Figure 3.

Algorithms 1 and 2 are iterative and require initial values; the fact that $\widehat{\boldsymbol{\beta}} = \mathbf{0}$ at $\lambda_{\max}$ provides an efficient approach to choosing those initial values. Group lasso, group MCP, and group SCAD all have the same value of $\lambda_{\max}$; namely, $\lambda_{\max} = \max_j\{\|\mathbf{z}_j\|\}$ for linear regression or $\lambda_{\max} = \max_j\{v\,\|\mathbf{z}_j\|\}$ for logistic regression, where the $\{\mathbf{z}_j\}$ are computed with respect to the intercept-only model (or, if unpenalized covariates are present, with respect to the residuals of the fitted model including all of the unpenalized covariates). Thus, by starting at $\lambda_{\max}$ where $\widehat{\boldsymbol{\beta}} = 0$ is available in closed form and proceeding towards $\lambda_{\min}$, using $\widehat{\boldsymbol{\beta}}$ from the previous value of λ as the initial value for the next value of λ, we can ensure that the initial values will never be far from the solution, a helpful property often referred to as "warm starts" in the path-fitting literature.

3 Algorithm efficiency

Here, we briefly comment on the efficiency of the proposed algorithms. Regardless of the penalty chosen, the most computationally intensive steps in Algorithm 1 are the calculation of the inner products $\mathbf{X}_j^T\mathbf{r}$ and $\mathbf{X}_j^T(\boldsymbol{\beta}_j' - \boldsymbol{\beta}_j)$, each of which requires $O(nK_j)$ operations. Thus, one full pass over all the groups requires $O(2np)$ operations. The fact that this approach scales linearly in p allows it to be efficiently applied to high-dimensional problems.

Of course, the entire time required to fit the model depends on the number of iterations, which in turn depends on the data and on λ. Broadly speaking, the dominant factor influencing the number of iterations is the number of nonzero groups at that value of λ, since no iteration is required to solve for groups that remain fixed at zero. Consequently, when fitting a regularized path, a disproportionate amount of time is spent at the least sparse end of the path, where λ is small.

Table 1 compares our implementation (`grpreg`) with two other publicly available `R` packages for fitting group lasso models over increasingly large data sets: the `grplasso` package (Meier et al., 2008) and the `standGL` package (Simon and Tibshirani, 2011). We note that (a) the `grpreg` implementation appears uniformly more efficient than the others, and (b) that group MCP and group SCAD tend to be slightly faster than group lasso. Presumably, this is because their solution paths tend to be more sparse.

It is worth noting that all three of these packages can handle $p > n$ problems; however, for the purposes of timing, we chose to restrict our attention to problems in which the entire path can be computed. Otherwise, different implementations may terminate the fitting process at different points along the path, which would prevent a fair comparison of computing times.

Table 1: Comparison of `grpreg` with other publicly available group lasso packages. The median time, over 100 independent data sets, required to fit the entire solution path over a grid of 100 λ values is reported in seconds. Each group consisted of 10 variables; thus p ranges over $\{10, 100, 1000\}$ across the columns.

			n=50 J=1[a]	n=500 J=10[b]	n=5000 J=100[c]
Linear regression	`grpreg`	Group lasso	0.01	0.1	18
	`grpreg`	Group MCP	< 0.01	0.1	11
	`grpreg`	Group SCAD	< 0.01	0.1	11
	`grplasso`	Group lasso	0.17	1.9	61
	`standGL`	Group lasso	0.04	1.3	153
Logistic regression	`grpreg`	Group lasso	0.01	0.2	105
	`grpreg`	Group MCP	0.01	0.1	45
	`grpreg`	Group SCAD	0.01	0.1	78
	`grplasso`	Group lasso	0.52	5.1	202
	`standGL`	Group lasso	0.40	6.4	400

[a] SE ≤ 0.02
[b] SE ≤ 0.1
[c] SE ≤ 3

Finally, let us compare these results to those presented in She (2012). For the algorithm presented in that paper, the author reports an average time of 32 minutes to estimate the group SCAD regression coefficients when $n = 100$ and $p = 500$. For the same size problem, our approach required a mere 0.35 seconds.

4 Simulation studies

In this section, we compare the performance of group lasso, group MCP, and group SCAD using simulated data. First, a relatively basic setting is used to illustrate the primary advantages of group MCP and group SCAD over group lasso. We then attempt to mimic two settings in which the methodology might be used: to allow flexible semiparametric modeling of continuous variables and in genetic association studies, which involve large numbers of categorical variables. We use the term "null predictor" to refer to a covariate whose associated regression coefficient is zero in the true model.

In all of the studies, five-fold cross-validation was used to choose the regularization parameter λ. Group SCAD and group MCP have an additional tuning parameter, γ. In principle, one may attempt to select optimal values of γ using, for example, cross-validation over a two-dimensional grid or using an information criterion approach. Here, we fix $\gamma = 3$ for group MCP and $\gamma = 4$ for group SCAD, roughly in line with the default recommendations suggested in Fan and Li (2001) and Zhang (2010) in the non-grouped case.

In Section 4.1, we evaluate model accuracy by root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{p}\sum_{j,k}(\beta_{jk} - \widehat{\beta}_{jk})^2}.$$

In Sections 4.2 and 4.3, because the model fit to the data is not always the same as the generating model, we focus on root model error (RME) instead:

$$\text{RME} = \sqrt{\frac{1}{n}\sum_{i}(\mu_i - \hat{\mu}_i)^2},$$

where μ_i and $\hat{\mu}_i$ denote the true and estimated mean of observation i given $\mathbf{x}_i$. Note that the model error, which is also discussed in Fan and Li (2001), is equal in expected value to the prediction error minus the irreducible error σ^2. In all simulations, errors follow a standard Gaussian distribution and results are averaged over 1,000 independently generated data sets.

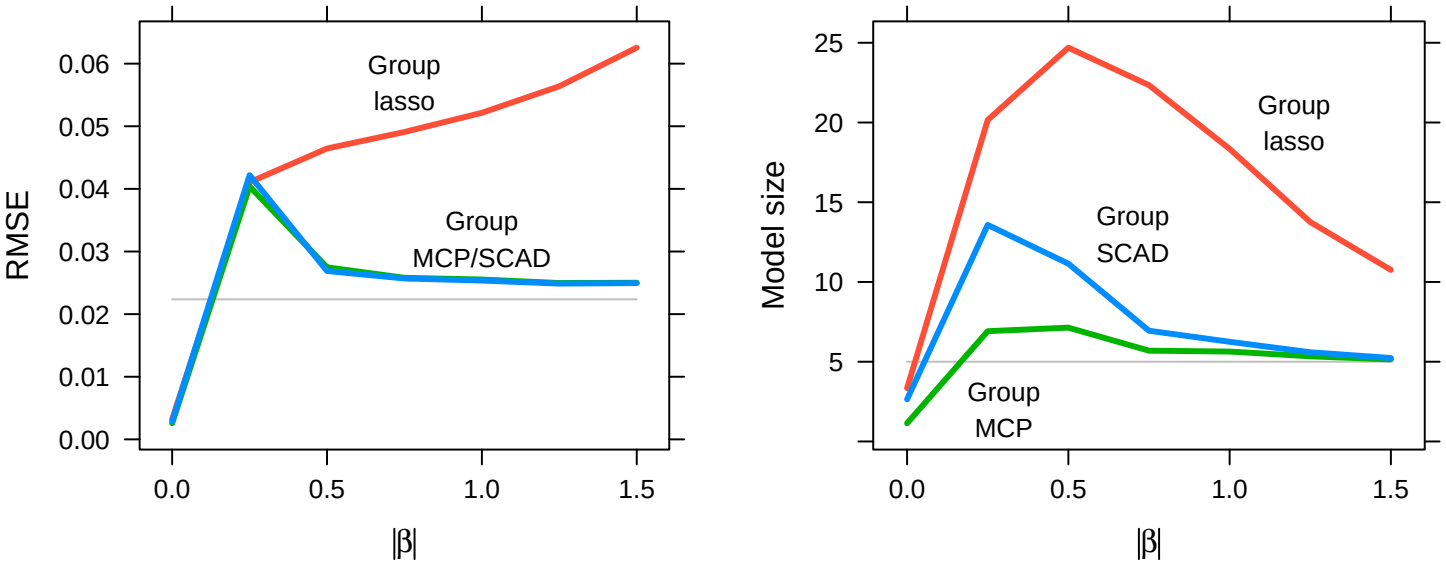


Figure 4: The impact of increasing coefficient magnitude on group regularization methods. Model size is given in terms of number of groups (*i.e.*, the number of variables in the model is four times the amount shown). The faint gray line on the left is the theoretically optimal RMSE than can be achieved in this setting. The faint gray line on the right is the true model size.

4.1 Basic

We begin with a very straightforward study designed to illustrate the basic shortcomings of the group lasso in comparison with group MCP and group SCAD. The design matrix consists of 100 groups, each with 4 elements. In five of these groups, the coefficients are $\pm\beta$; in the others, the true regression coefficients are zero. Covariate values were generated from the standard normal distribution. We fixed the sample size at 100 (*i.e.*, n=100, p=400) and varied $|\beta|$ between 0 and 1.5. In principle, group lasso should struggle when $|\beta|$ is large, as it cannot alleviate the problem of bias towards zero for large coefficients without lowering λ and thereby allowing null predictors to enter the model. Indeed, as Figure 4 illustrates, this is exactly what occurs.

For small values of the regression coefficients, all three group regularization methods perform similarly. As we increase the magnitude of these coefficients, however, group MCP and group SCAD begin to estimate $\boldsymbol{\beta}$ with an error approaching the theoretically optimal value, while group lasso performs increasingly poorly. Furthermore, group MCP and group SCAD select much smaller models and approach the true model size much faster than group lasso, which selects far too many variables.

Comparing group MCP and group SCAD, the two are nearly identical in terms of estimation accuracy. However, group MCP selects a considerably more sparse model, and has better variable selection properties. Thus, although the two methods behave similarly in an asymptotic setting, group MCP seems to have somewhat better finite-sample properties.

In the left panel of Figure 4, we include an "oracle" RMSE for reference; we now clarify what exactly we mean by this. An oracle model (one which knows in advance which coefficients are zero and which are nonzero) would be able to achieve a mean squared error of zero for the zero-coefficient variables and a total MSE of $\text{tr}\{(\mathbf{X}^T\mathbf{X})^{-1}\}$ for the nonzero variables. In our simulation $\mathbf{X}$ was random, with $\mathbb{E}(\mathbf{X}\mathbf{X}^T) = \mathbf{I}$. Thus, the total MSE of the oracle model is approximately $\text{tr}(\mathbf{I}_0/n)$ where $\mathbf{I}_0$ is the identity matrix with dimension equal to that of the nonzero coefficients, with RMSE $\sqrt{s/n}$, where s is the sparsity fraction (*i.e.*, the fraction of coefficients that are nonzero). Note that in any finite sample, the columns of $\mathbf{X}$ will be correlated, so even the oracle model cannot achieve this RMSE for finite sample sizes; the gray line in Figure 4 is thus the optimal RMSE that could be theoretically be achieved in this setting.

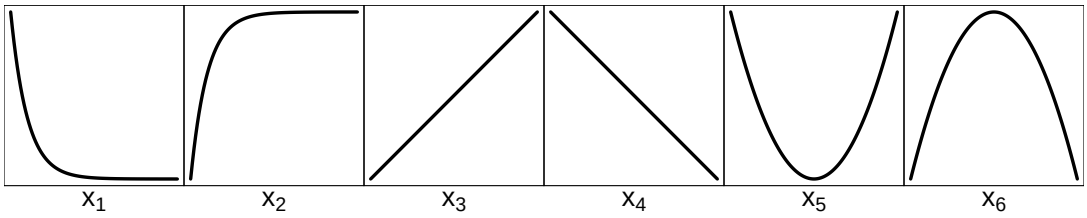
4.2 Semiparametric regression

Our next simulation involves groups of covariates constructed by taking basis expansions of continuous variables to allow for flexible covariate effects in semiparametric modeling. The sample size was 200 and the data consisted of 100 variables, each of which were generated as independent uniform (0,1) variates. The first six variables had potentially nonlinear effects given

by the following equations:

$$
\begin{aligned}
f_1(x) &= 2(e^{-10x} - e^{-10})/(1 - e^{-10}) - 1 \\
f_2(x) &= -2(e^{-10x} - e^{-10})/(1 - e^{-10}) + 1 \\
f_3(x) &= 2x - 1 \\
f_4(x) &= -2x + 1 \\
f_5(x) &= 8(x - 0.5)^2 - 1 \\
f_6(x) &= -8(x - 0.5)^2 + 1;
\end{aligned}
$$

the other 94 had no effect on the outcome. The scaling of the functions is to ensure that each variable attains a minimum and maximum of $(-1, 1)$ over the domain of x and thus that the effects of all six variables are roughly comparable. Visually the effect of the six nonzero variables is illustrated below:



For model fitting, each variable was represented using a 6-term B-spline basis expansion (*i.e.*, $\mathbf{X}$ had dimensions $n = 200$, $p = 600$). In addition to the three group selection methods, models were also fit using the lasso (*i.e.*, ignoring grouping). Results are given in Table 2.

Table 2: Prediction and variable selection accuracy for the semiparametric regression simulation.

	Lasso	**Group Lasso**	**Group MCP**	**Group SCAD**
RME[a]	0.73	0.59	0.50	0.52
Variables selected[b]	31.5	29.3	10.4	23.1

[a] SE ≤ 0.002
[b] SE ≤ 0.4

Certainly, all three group selection approaches greatly outperform the lasso here. However, as in the previous section, group MCP and group SCAD are able to achieve superior prediction accuracy while selecting more parsimonious models. Also as in the previous simulation, group MCP and group SCAD perform similarly as far as prediction accuracy, but group MCP is seen to have better finite-sample variable selection properties — recall that the true number of variables in the model is only 6.

4.3 Genetic association study

Finally, we carry out a simulation designed to mimic a small genetic association study involving single nucleotide polymorphisms (SNPs). Briefly, a SNP is a location on the genome in which multiple versions (alleles) may be present. A SNP may take on three values $\{0, 1, 2\}$, depending on the number of minor alleles present. The effect is not necessarily linear — for example, if the allele has a recessive effect, the phenotype associated with $x = 0$ and $x = 1$ are identical, while the phenotype associated with $x = 2$ is different. In such studies, it is desirable to have a method which is robust to different mechanisms of action, yet powerful enough to actually detect important SNPs, as the number of SNPs is typically rather large.

We simulated data involving 250 subjects and 500 SNPs, each of which was represented with 2 indicator functions (*i.e.*, n=250, p=1,000). Three of the variables had an effect on the phenotype (one dominant effect, one recessive effect, one additive effect); the other 497 did not. In addition to the three group selection methods, we included for comparison two versions of the lasso: one applied to all $p = 1,000$ variables and ignoring grouping, the other assuming an additive effect for each genotype. Note that for the second approach, $p = 500$ as we estimate only a single coefficient for each SNP. The results of this simulation are given in Table 3.

The same broad conclusions may be reached here as in the previous simulations. In particular, we note that (1) The group selection methods outperform the variable selection methods that either do not account for grouping or that attempt

Table 3: Prediction and variable selection accuracy for the genetic association study simulation. True discoveries are selected variables that have a truly nonzero effect; false discoveries are selected SNPs that have no effect on the phenotype.

	Lasso	**Additive Lasso**	**Group Lasso**	**Group MCP**	**Group SCAD**
RME[a]	0.38	0.37	0.34	0.25	0.27
True discoveries[b]	2.6	2.2	2.6	2.4	2.5
False discoveries[c]	21.1	16.3	15.5	4.0	12.5

[a] SE ≤ 0.002
[b] SE ≤ 0.02
[c] SE ≤ 0.6

to incorporate grouping in an ad-hoc fashion. (2) Group MCP and group SCAD outperform the group lasso both in terms of prediction accuracy as well as the number of false discoveries. (3) Although group MCP and group SCAD are similar in terms of prediction accuracy, group MCP has significantly better variable selection properties, producing only four false discoveries compared to group SCAD's 12.5.

5 Real data

We give two examples of applying grouped variable selection methods to real data. The first is a gene expression study in rats to determine genes associated with Bardet-Biedl syndrome. The second is a genetic association study to determine SNPs associated with age-related macular degeneration. As in the previous section, we fix $\gamma = 3$ for group MCP, $\gamma = 4$ for group SCAD, and select λ via cross-validation. For the continuous outcome in Section 5.1, we use root cross-validation error, defined analogously to root model error in Section 4, to evaluate predictive accuracy. For the binary outcome in Section 5.2, we use cross-validated misclassification error.

5.1 Bardet-Biedl syndrome gene expression study

The data we analyze here is discussed more fully in Scheetz et al. (2006). Briefly, the data set consists of normalized microarray gene expression data harvested from the eye tissue of 120 twelve-week-old male rats. The outcome of interest is the expression of TRIM32, a gene which has been shown to cause Bardet-Biedl syndrome (Chiang et al., 2006). Bardet-Biedl syndrome is a genetic disease of multiple organ systems including the retina.

Following the approach in Scheetz et al. (2006), 18,976 of the 31,042 probe sets on the array "exhibited sufficient signal for reliable analysis and at least 2-fold variation in expression." These probe sets include TRIM32 and 18,975 other genes that potentially influence its expression. We further restricted our attention to the 5,000 genes with the largest variances in expression (on the log scale) and considered a three-term natural cubic spline basis expansion of those genes, resulting in a grouped regression problem with $n = 120$ and $p = 15,000$. The models selected by group lasso, group MCP, and group SCAD are described in Table 4.

This is an interesting case study in that group MCP selects a very different model from the other two approaches. In particular, group lasso and group SCAD each select a fairly large number of genes, while shrinking each gene's group of coefficients nearly to zero. Group MCP, on the other hand, selects a single gene and returns a fit nearly the same as the least-squares fit for that gene alone. The relationship between probe set 1372928_at and TRIM32 estimated by each model is plotted in Figure 5.

The most important aspect of Figure 5 to note is that the outcome, TRIM32, has a large outlying value almost 10 standard deviations below the mean of the rest of the points. This observation has a large impact on the fit: for all of the genes in Table 4, this subject is also responsible for the lowest/highest or second-lowest/highest value of that gene, and the median absolute correlation for the genes in the table is 0.62. Many of the scatterplots of those genes versus TRIM32 look qualitatively similar to the one in Figure 5.

Faced with this set of genes, group MCP selects a single gene and fits a model that explains 65% of the variance in TRIM32 expression. Group SCAD and group lasso select an ensemble of correlated genes, downweighting the contribution of each gene considerably. Each approach has advantages, depending on the goal of the analysis. The group SCAD/lasso approaches avoid a possibly arbitrary selection of one gene from a highly correlated set, and produce a model with somewhat better predictive ability (root cross-validation error of 0.092 ± 0.04 versus 0.099 ± 0.04), although all three approaches are

Table 4: Genes selected by group lasso/SCAD/MCP, along with the Euclidean norm of the coefficients for each gene's basis expansion.

		Group norm		
Probe Set	**Gene Symbol**	**Group Lasso**	**Group MCP**	**Group SCAD**
1374131_at		0.11		0.13
1383110_at	Klhl24	0.10		0.08
1383749_at	Phospho1	0.02		0.04
1376267_at		0.22		0.23
1377791_at		0.13		0.12
1376747_at		0.28		0.24
1390539_at		0.11		0.12
1384470_at		0.05		0.07
1386032_at	Prkd3	0.03		0.06
1393231_at	Ppp4r2	0.01		0.03
1385798_at		0.03		
1383730_at	Ttc9c	0.02		0.06
1368476_at	Nr3c2	0.05		0.01
1384860_at	Zfp84	0.03		0.07
1372928_at		0.22	1.83	0.20
1381902_at	Zfp292	0.16		0.18
1390574_at		0.02		0.01
1384940_at	Zfp518a	0.10		0.10

within random variability of each other. Group MCP, on the other hand, produces a highly parsimonious model capable of predicting just as well as the group lasso/SCAD models despite using only a single gene. This is potentially a valuable property if the goal is, say, to develop a diagnostic assay and each gene that need to be measured adds cost to the assay.

Finally, it should be noted that although group MCP is behaving in a rather greedy fashion in this example, this is not an inherent aspect of the method. By adjusting the γ parameter, group MCP can be made to resemble the group lasso and group SCAD solutions as well – recall that group lasso can be considered a special case of group MCP with $\gamma = \infty$. Group SCAD, on the other hand, cannot be made to resemble group MCP and is incapable in this case of selecting a highly parsimonious model. Group MCP is considerably more flexible, although of course to pursue this flexibility, proper selection of tuning parameters becomes an important issue. The selection of additional tuning parameters is an important area for further study in both the grouped and non-grouped application of the MC penalty.

5.2 Age-related macular degeneration genetic association study

We analyze data from a case-control study of age-related macular degeneration consisting of 400 cases and 400 controls. We confine our analysis to 532 SNPs that previous biological studies have suggested may be related to the disease. As in Section 4.3, we represent each SNP as a three-level factor depending on the number of minor alleles the individual carries at that locus. This requires the expansion of each SNP into a group of two indicator functions; our design matrix for this example thus has $n = 800$ and $p = 1,064$. Group regularized logistic regression models were then fit to the data; as before, λ was selected via cross-validation. The results are presented in Table 5.

We make the following remarks on Table 5: (1) All three approaches represent significant improvements over the baseline (intercept-only) misclassification error. (2) The group MCP model is slightly superior to the group lasso and group SCAD models, although again, the difference between the three models in terms of predictive accuracy is comparable to the SE. (3) The group MCP and group SCAD approaches produce considerably more parsimonious models without a loss in prediction accuracy. Compared with the gene expression example of the previous section, parsimony is both more desirable and more believable in this example. The SNPs selected here are not highly correlated and thus more likely to represent independent causes than dependent manifestations of a separate underlying cause such as up-regulation of a pathway.

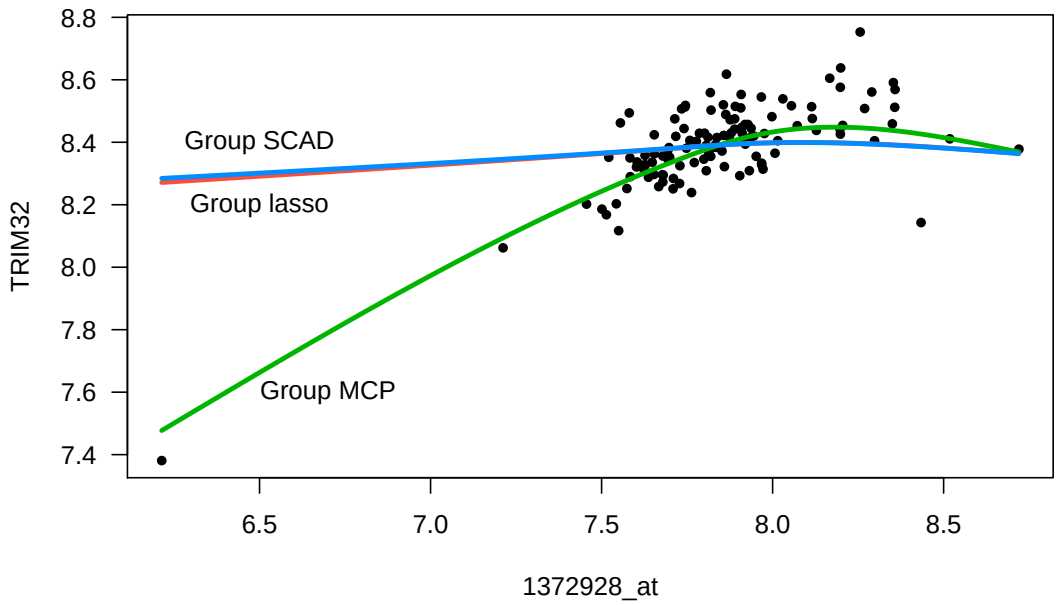


Figure 5: Estimated relationship between probe set 1372928_at and TRIM32 estimated by group lasso, group MCP, and group SCAD. Estimates are superimposed on top of a scatterplot and restricted to pass through the mean expression for each probe set.

Table 5: Application of group regularized methods to age-related macular degeneration study. The number of SNPs selected by the method as well as the cross-validated misclassification error are reported. For comparison, the intercept-only model is also listed ("Baseline").

	SNPs	**Misclassification Error**[a]
Baseline		0.50
grLasso	51	0.41
grMCP	12	0.39
grSCAD	15	0.41

[a] SE = 0.02 for all models

6 Conclusion

Group MCP and group SCAD models are powerful alternatives to the group lasso in problems involving grouped variable selection. However, application and study of these approaches has been limited, especially in high-dimensional problems, due to a lack of efficient algorithms and a lack of publicly available software for fitting these models. In this article, we attempt to remedy this, describing the development of efficient algorithms and proving an implementation via the `R` package `grpreg`.

Acknowledgments

Jian Huang's work is supported in part by NIH Grants R01CA120988, R01CA142774 and NSF Grants DMS-0805670 and DMS-1208225. The authors would like to thank Rob Mullins for the genetic association data analyzed in Section 5.2, as well as the associate editor and two anonymous reviewers, who provided many helpful remarks that led to considerable refinement of this article.

Appendix

Before proving Proposition 1, we establish the groupwise convexity of all the objective functions under consideration. Note that for group SCAD and group MCP, although they contain nonconvex components and are not necessarily convex overall, the objective functions are still convex with respect to the variables in a single group.

Lemma 1. *The objective function $Q(\boldsymbol{\beta}_j)$ for regularized linear regression is a strictly convex function with respect to $\boldsymbol{\beta}_j$ for the group lasso, for group SCAD with $\gamma > 2$, and for group MCP with $\gamma > 1$.*

Proof. Although $Q(\boldsymbol{\beta}_j)$ is not differentiable, it is directionally twice differentiable everywhere. Let $\nabla^2_{\mathbf{d}} Q(\boldsymbol{\beta}_j)$ denote the second derivative of $Q(\boldsymbol{\beta}_j)$ in the direction d. Then the strict convexity of $Q(\boldsymbol{\beta}_j)$ follows if $\nabla^2_{\mathbf{d}} Q(\boldsymbol{\beta}_j)$ is positive definite at all $\boldsymbol{\beta}_j$ and for all $\mathbf{d}$. Let ξ_* denote the infimum over $\boldsymbol{\beta}_j$ and $\mathbf{d}$ of the minimum eigenvalue of $\nabla^2_{\mathbf{d}} Q(\boldsymbol{\beta}_j)$. Then, after some algebra, we obtain

$$\begin{aligned} \xi_* &= 1 && \text{Group lasso} \\ \xi_* &= 1 - \frac{1}{\gamma - 1} && \text{Group SCAD} \\ \xi_* &= 1 - \frac{1}{\gamma} && \text{Group MCP,} \end{aligned}$$

These quantities are positive under the conditions specified in the lemma. □

We now proceed to the proof of Proposition 1.

Proof of Proposition 1. The descent property is a direct consequence of the fact that each updating step consists of minimizing $Q(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_j$. Lemma 1, along with the fact that the least squares loss function is continuously differentiable and coercive, provide sufficient conditions to apply Theorem 4.1 of Tseng (2001), thereby establishing that every limit point of $\{\boldsymbol{\beta}^{(m)}\}$ is a stationary point of $Q(\boldsymbol{\beta})$.

We further note that $\{\boldsymbol{\beta}^{(m)}\}$ is guaranteed to converge to a unique limit point. Suppose that the sequence possessed two limit points, $\boldsymbol{\beta}'$ and $\boldsymbol{\beta}''$, such that for at least one group, $\boldsymbol{\beta}'_j \neq \boldsymbol{\beta}''_j$. For the transition $\boldsymbol{\beta}' \to \boldsymbol{\beta}''$ to occur, the algorithm must pass through the point $(\boldsymbol{\beta}''_j, \boldsymbol{\beta}'_{-j})$. However, by Lemma 1, $\boldsymbol{\beta}'_j$ is the unique value minimizing $Q(\boldsymbol{\beta}_j | \boldsymbol{\beta}_{-j})$. Thus, $\boldsymbol{\beta}' \to \boldsymbol{\beta}''$ is not allowed by the group descent algorithm and $\{\boldsymbol{\beta}^{(m)}\}$ possesses a single limit point. □

For Proposition 2, involving logistic regression, we proceed similarly, letting $R(\boldsymbol{\beta}|\widetilde{\boldsymbol{\beta}})$ denote the majorizing approximation to $Q(\boldsymbol{\beta})$ at $\widetilde{\boldsymbol{\beta}}$.

Lemma 2. *The majorizing approximation $R(\boldsymbol{\beta}_j|\widetilde{\boldsymbol{\beta}})$ for regularized logistic regression is a strictly convex function with respect to $\boldsymbol{\beta}_j$ at all $\widetilde{\boldsymbol{\beta}}$ for the group lasso, for group SCAD with $\gamma > 5$, and for group MCP with $\gamma > 4$.*

Proof. Proceeding as in the previous lemma, and letting ξ_* denote the infimum over $\widetilde{\boldsymbol{\beta}}$, $\boldsymbol{\beta}_j$ and $\mathbf{d}$ of the minimum eigenvalue of $\nabla^2_{\mathbf{d}} Q(\boldsymbol{\beta}_j)$, we obtain

$$\begin{aligned} \xi_* &= \frac{1}{4} && \text{Group lasso} \\ \xi_* &= \frac{1}{4} - \frac{1}{\gamma - 1} && \text{Group SCAD} \\ \xi_* &= \frac{1}{4} - \frac{1}{\gamma} && \text{Group MCP,} \end{aligned}$$

These quantities are positive under the conditions specified in the lemma. □

Proof of Proposition 2. The proposition makes two claims: descent with every iteration and convergence to a stationary point. To establish descent for logistic regression, we note that because L is twice differentiable, for any point $\boldsymbol{\eta}$ there exists a vector $\boldsymbol{\eta}^{**}$ on the line segment joining $\boldsymbol{\eta}$ and $\boldsymbol{\eta}^*$ such that

$$\begin{aligned} L(\boldsymbol{\eta}) &= L(\boldsymbol{\eta}^*) + (\boldsymbol{\eta} - \boldsymbol{\eta}^*)^T \nabla L(\boldsymbol{\eta}^*) + \frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{\eta}^*)^T \nabla^2 L(\boldsymbol{\eta}^{**})(\boldsymbol{\eta} - \boldsymbol{\eta}^*) \\ &\leq \tilde{L}(\boldsymbol{\eta}|\boldsymbol{\eta}^*) \end{aligned}$$

where the inequality follows from the fact that $v\mathbf{I} - \nabla^2 L(\boldsymbol{\eta}^{**})$ is a positive semidefinite matrix. Descent now follows from the descent property of MM algorithms (Lange et al., 2000) coupled with the fact that each updating step consists of minimizing $R(\boldsymbol{\beta}_j|\widetilde{\boldsymbol{\beta}})$.

To establish convergence to a stationary point, we note that if no elements of $\boldsymbol{\beta}$ tend to $\pm\infty$, then the descent property of the algorithm ensures that the sequence $\boldsymbol{\beta}^{(k)}$ stays within a compact set and therefore possesses a limit point $\widetilde{\boldsymbol{\beta}}$. Then, as in the proof of Proposition 1, Lemma 2 allows us to apply the results of Tseng (2001) and conclude that $\widetilde{\boldsymbol{\beta}}$ must be a stationary point of $R(\boldsymbol{\beta}|\widetilde{\boldsymbol{\beta}})$. Furthermore, because $R(\boldsymbol{\beta}|\widetilde{\boldsymbol{\beta}})$ is tangent to $Q(\boldsymbol{\beta})$ at $\widetilde{\boldsymbol{\beta}}$, $\widetilde{\boldsymbol{\beta}}$ must also be a stationary point of $Q(\boldsymbol{\beta})$. □

References

Bakin, S. (1999). *Adaptive regression and model selection in data mining problems.* Ph.D. thesis, Australian National University.

Bertsekas, D. (1999). *Nonlinear Programming.* 2nd ed. Athena Scientific.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, **5** 232–253.

Chiang, A., Beck, J., Yen, H., Tayeh, M., Scheetz, T., Swiderski, R., Nishimura, D., Braun, T., Kim, K., Huang, J. et al. (2006). Homozygosity mapping with snp arrays identifies trim32, an e3 ubiquitin ligase, as a bardet–biedl syndrome gene (bbs11). *Proceedings of the National Academy of Sciences*, **103** 6287–6292.

Donoho, D. and Johnstone, J. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81** 425–455.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96** 1348–1360.

Foygel, R. and Drton, M. (2010). Exact block-wise optimization in group lasso and sparse group lasso for linear regression. *Arxiv preprint arXiv:1010.3320*.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, **1** 302–332.

Friedman, J., Hastie, T. and Tibshirani, R. (2010a). A note on the group lasso and a sparse group lasso. *Arxiv preprint arXiv:1001.0736*.

Friedman, J. H., Hastie, T. and Tibshirani, R. (2010b). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33** 1–22.

Fu, W. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, **7** 397–416.

Gao, H. and Bruce, A. (1997). Waveshrink with firm shrinkage. *Statistica Sinica*, **7** 855–874.

Huang, J., Breheny, P. and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science*, **27** 481–499.

Hunter, D. and Lange, K. (2004). A tutorial on MM algorithms. *American Statistician*, **58** 30–38.

Krishnapuram, B., Carin, L., Figueiredo, M. and Hartemink, A. (2005). Sparse multinomial logistic regression: fast algorithms and generalization bounds. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **27** 957 –968.

Lange, K., Hunter, D. and Yang, I. (2000). Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, **9** 1–20.

Meier, L., van de Geer, S. and Buhlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society Series B*, **70** 53.

Puig, A., Wiesel, A., Fleury, G. and Hero, A. (2011). Multidimensional shrinkage-thresholding operator and group lasso penalties. *Signal Processing Letters, IEEE*, **18** 363–366.

Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society Series B*, **71** 1009–1030.

Scheetz, T., Kim, K., Swiderski, R., Philp, A., Braun, T., Knudtson, K., Dorrance, A., DiBona, G., Huang, J., Casavant, T. et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, **103** 14429–14434.

She, Y. (2012). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics & Data Analysis*, **56** 2976 – 2990.

Simon, N. and Tibshirani, R. (2011). Standardization and the group lasso penalty. *Statistica Sinica*, **22** 983–1001.

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

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, **109** 475–494.

Wang, L., Chen, G. and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, **23** 1486.

Wang, L., Li, H. and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, **103** 1556–1569.

Wu, T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics*, **2** 224–244.

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

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, **38** 894–942.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, **36** 1509–1533.