

A note on the group lasso and a sparse group lasso

JEROME FRIEDMAN ^{*}
TREVOR HASTIE [†]
and ROBERT TIBSHIRANI [‡]

February 11, 2010

(With corrections; original version Jan 5, 2010)

Abstract

We consider the group lasso penalty for the linear model. We note that the standard algorithm for solving the problem assumes that the model matrices in each group are orthonormal. Here we consider a more general penalty that blends the lasso (L_1) with the group lasso (“two-norm”). This penalty yields solutions that are sparse at both the group and individual feature levels. We derive an efficient algorithm for the resulting convex problem based on coordinate descent. This algorithm can also be used to solve the general form of the group lasso, with non-orthonormal model matrices.

1 Introduction

In this note, we consider the problem of prediction using a linear model. Our data consist of \mathbf{y} , a vector of N observations, and \mathbf{X} , a $N \times p$ matrix of features.

Suppose that the p predictors are divided into L groups, with p_ℓ the number in group ℓ . For ease of notation, we use a matrix \mathbf{X}_ℓ to represent the predictors corresponding to the ℓ th group, with corresponding coefficient vector β_ℓ . Assume that \mathbf{y} and \mathbf{X} has been centered, that is, all variables have mean zero.

^{*}Dept. of Statistics, Stanford Univ., CA 94305, jhf@stanford.edu

[†]Depts. of Statistics, and Health, Research & Policy, Stanford Univ., CA 94305, hastie@stanford.edu

[‡]Depts. of Health, Research & Policy, and Statistics, Stanford Univ, tibs@stanford.edu

In an elegant paper, Yuan & Lin (2007) proposed the group lasso which solves the convex optimization problem

$$\min_{\beta \in \mathbb{R}^p} \left(\|\mathbf{y} - \sum_{\ell=1}^L \mathbf{X}_\ell \beta_\ell\|_2^2 + \lambda \sum_{\ell=1}^L \sqrt{p_\ell} \|\beta_\ell\|_2 \right), \quad (1)$$

where the $\sqrt{p_\ell}$ terms accounts for the varying group sizes, and $\|\cdot\|_2$ is the Euclidean norm (not squared). This procedure acts like the lasso at the group level: depending on λ , an entire group of predictors may drop out of the model. In fact if the group sizes are all one, it reduces to the lasso. Meier et al. (2008) extend the group lasso to logistic regression.

The group lasso does not, however, yield sparsity within a group. That is, if a group of parameters is non-zero, they will all be non-zero. In this note we propose a more general penalty that yields sparsity at both the group and individual feature levels, in order to select groups and predictors within a group. We also point out that the algorithm proposed by Yuan & Lin (2007) for fitting the group lasso assumes that the model matrices in each group are orthonormal. The algorithm that we provide for our more general criterion also works for the standard group lasso with non-orthonormal model matrices.

We consider the *sparse group lasso* criterion

$$\min_{\beta \in \mathbb{R}^p} \left(\|\mathbf{y} - \sum_{\ell=1}^L \mathbf{X}_\ell \beta_\ell\|_2^2 + \lambda_1 \sum_{\ell=1}^L \|\beta_\ell\|_2 + \lambda_2 \|\beta\|_1 \right). \quad (2)$$

where $\beta = (\beta_1, \beta_2, \dots, \beta_\ell)$ is the entire parameter vector. For notational simplicity we omit the weights $\sqrt{p_\ell}$. Expression (2) is the sum of convex functions and is therefore convex. Figure 1 shows the constraint region for the group lasso, lasso and sparse group lasso. A similar penalty involving both group lasso and lasso terms is discussed in Peng et al. (2009). When $\lambda_2 = 0$, criterion (2) reduces to the group lasso, whose computation we discuss next.

2 Computation for the group lasso

Here we briefly review the computation for the group lasso of Yuan & Lin (2007). In the process we clarify a confusing issue regarding orthonormality of predictors within a group.

The subgradient equations (see e.g. Bertsekas (1999)) for the group lasso are

$$-\mathbf{X}_\ell^T (y - \sum_{\ell} \mathbf{X}_\ell \beta_\ell) + \lambda \cdot s_\ell = 0; \quad \ell = 1, 2, \dots, L, \quad (3)$$

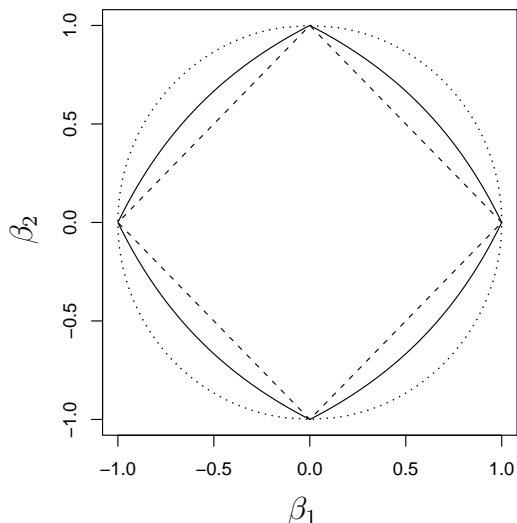


Figure 1: Contour lines for the penalty for the group lasso (dotted), lasso (dashed) and sparse group lasso penalty (solid), for a single group with two predictors.

where $s_\ell = \beta_\ell / \|\beta_\ell\|$ if $\beta_\ell \neq 0$ and s_ℓ is a vector with $\|s_\ell\|_2 < 1$ otherwise. Let the solutions be $\hat{\beta}_1, \hat{\beta}_2 \dots \hat{\beta}_\ell$. If

$$\|\mathbf{X}_\ell^T (y - \sum_{k \neq \ell} \mathbf{X}_k \hat{\beta}_k)\| < \lambda \quad (4)$$

then $\hat{\beta}_\ell$ is zero; otherwise it satisfies

$$\hat{\beta}_\ell = (\mathbf{X}_\ell^T \mathbf{X}_\ell + \lambda / \|\hat{\beta}_\ell\|)^{-1} \mathbf{X}_\ell^T r_\ell \quad (5)$$

where

$$\mathbf{r}_\ell = \mathbf{y} - \sum_{k \neq \ell} \mathbf{X}_k \hat{\beta}_k$$

Now if we assume that $\mathbf{X}_\ell^T \mathbf{X}_\ell = \mathbf{I}$, and let $v_\ell = \mathbf{X}_\ell^T r_\ell$, then (5) simplifies to $\hat{\beta}_\ell = (1 - \lambda / \|v_\ell\|) v_\ell$. This leads to an algorithm that cycles through the groups k , and is a blockwise coordinate descent procedure. It is given in Yuan & Lin (2007).

If however the predictors are not orthonormal, one approach is to orthonormalize them before applying the group lasso. However this will not generally provide a solution to the original problem. In detail, if $\mathbf{X}_\ell = \mathbf{U} \mathbf{D} \mathbf{V}^T$, then the columns of $\mathbf{U} = \mathbf{X}_\ell \mathbf{V} \mathbf{D}^{-1}$ are orthonormal. Then $\mathbf{X}_\ell \beta_\ell = \mathbf{U} \mathbf{V} \mathbf{D}^{-1} \beta_\ell = \mathbf{U} [\mathbf{V} \mathbf{D}^{-1} \beta_\ell] = \mathbf{U} \beta_{\ell*}$.

But $\|\hat{\beta}_\ell^*\| = \|\beta_\ell\|$ only if $\mathbf{D} = \mathbf{I}$. This will not be true in general, e.g. if \mathbf{X} is a set of dummy variables for a factor, this is true only if the number of observations in each category is equal.

Hence an alternative approach is needed. In the non-orthonormal case, we can think of equation (5) as a ridge regression, with the ridge parameter depending on $\|\hat{\beta}_\ell\|$. A complicated scalar equation can be derived for $\|\hat{\beta}_\ell\|$ from (5); then substituting into the right-hand side of (5) gives the solution. However this is not a good approach numerically, as it can involve dividing by the norm of a vector that is very close to zero. It is also not guaranteed to converge. In the next section we provide a better solution to this problem, and to the sparse group lasso.

3 Computation for the sparse group lasso

The criterion (1) is separable so that block coordinate descent can be used for its optimization. Therefore we focus on just one group ℓ , and denote the predictors by $\mathbf{X}_\ell = Z = (Z_1, Z_2, \dots, Z_k)$, the coefficients by $\beta_\ell = \theta = (\theta_1, \theta_2, \dots, \theta_k)$ and the residual by $\mathbf{r} = \mathbf{y} - \sum_{k \neq \ell} X_k \beta_k$. The subgradient equations are

$$-Z_j^T (\mathbf{r} - \sum_j Z_j \theta_j) + \lambda_1 s_j + \lambda_2 t_j = 0 \quad (6)$$

for $j = 1, 2, \dots, k$ where $s_j = \theta_j / \|\theta\|$ if $\theta_j \neq 0$ and s is a vector satisfying $\|s\|_2 \leq 1$ otherwise, and $t_j \in \text{sign}(\theta_j)$, that is $t_j = \text{sign}(\theta_j)$ if $\theta_j \neq 0$ and $t_j \in [-1, 1]$ if $\theta_j = 0$. Letting $\mathbf{a} = \mathbf{X}_\ell^T \mathbf{r}$, then a necessary and sufficient condition for θ to be zero is that the system of equations

$$a_j = \lambda_1 s_j + \lambda_2 t_j \quad (7)$$

have a solution with $\|s\|_2 \leq 1$ and $t_j \in [-1, 1]$. We can determine this by minimizing

$$J(t) = (1/\lambda_1^2) \sum_{j=1}^k (a_j - \lambda_2 t_j)^2 = \sum_{j=1}^k s_j^2 \quad (8)$$

with respect to the $t_j \in [-1, 1]$ and then checking if $J(\hat{t}) \leq 1$. The minimizer is easily seen to be

$$\hat{t}_j = \begin{cases} \frac{a_j}{\lambda_2} & \text{if } \left| \frac{a_j}{\lambda_2} \right| \leq 1, \\ \text{sign}\left(\frac{a_j}{\lambda_2}\right) & \text{if } \left| \frac{a_j}{\lambda_2} \right| > 1. \end{cases}$$

Now if $J(\hat{t}) > 1$, then we must minimize the criterion

$$\frac{1}{2} \sum_{i=1}^N \left(r_i - \sum_{j=1}^k Z_{ij} \theta_j \right)^2 + \lambda_1 \|\theta\|_2 + \lambda_2 \sum_{j=1}^k |\theta_j| \quad (9)$$

This is the sum of a convex differentiable function (first two terms) and a separable penalty, and hence we can use coordinate descent to obtain the global minimum.

Here are the details of the coordinate descent procedure. For each j let $\mathbf{w}_j = \mathbf{r} - \sum_{k \neq j} Z_k \hat{\theta}_k$. Then $\hat{\theta}_j = 0$ if $|Z_j^T \mathbf{w}_j| < \lambda_2$. This follows easily by examining the subgradient equation corresponding to (9). Otherwise if $|Z_j^T \mathbf{w}_j| \geq \lambda_2$ we minimize (9) by a one-dimensional search over θ_j . We use the `optimize` function in the R package, which is a combination of golden section search and successive parabolic interpolation.

This leads to the following algorithm:

Algorithm for the sparse group lasso

1. Start with $\hat{\beta} = \beta_0$
2. In group ℓ , define $\mathbf{r} = \mathbf{y} - \sum_{k \neq \ell} \mathbf{X}_k \beta_k$, $\mathbf{X}_\ell = (Z_1, Z_2, \dots, Z_k)$, $\beta_\ell = (\theta_1, \theta_2, \dots, \theta_k)$ and $\mathbf{w}_j = (w_1, w_2, \dots, w_N) = \mathbf{r} - \sum_{k \neq j} Z_k \theta_k$. Check if $J(\hat{t}) \leq 1$ according to (8) and if so set $\hat{\beta}_\ell = 0$. Otherwise for $j = 1, 2, 3 \dots k, 1, 2, 3 \dots$, if $|Z_j^T \mathbf{w}_j| < \lambda_2$ then $\hat{\theta}_j = 0$; if instead $|Z_j^T \mathbf{w}_j| \geq \lambda_2$ then minimize

$$\frac{1}{2} \sum_{i=1}^N (w_i - \sum_{j=1}^k Z_{ij} \theta_j)^2 + \lambda_1 \|\theta\|_2 + \lambda_2 \sum_{j=1}^k |\theta_j| \quad (10)$$

over θ_j by a one-dimensional optimization. This cyclic optimization for $j = 1, 2, 3 \dots k, 1, 2, 3 \dots$ is iterated until convergence.

3. Iterate the entire step (2) over groups $\ell = 1, 2, \dots, L$ until convergence.

If λ_2 is zero, we instead use condition (4) for the group-level test and we don't need to check the condition $|Z_j^T \mathbf{w}_j| < \lambda_2$. With these modifications, this algorithm also gives a effective method for solving the group lasso with non-orthogonal model matrices.

Note that in the special case where $\mathbf{X}_\ell^T \mathbf{X}_\ell = I$, with $\mathbf{X}_\ell = (Z_1, Z_2, \dots, Z_k)$ then its is easy to show that

$$\hat{\theta}_j = \left(\|S(Z_j^T \mathbf{y}, \lambda_2)\|_2 - \lambda_1 \right)_+ \frac{S(Z_j^T \mathbf{y}, \lambda_2)}{\|S(Z_j^T \mathbf{y}, \lambda_2)\|_2} \quad (11)$$

and this reduces to the algorithm of Yuan & Lin (2007).

4 An example

We generated $n = 200$ observations with $p = 100$ predictors, in ten blocks of ten. The second fifty predictors all have coefficients of zero. The number of non-zero coefficients in the first five blocks of 10 are (10, 8, 6, 4, 2, 1) respectively, with coefficients equal to ± 1 , the sign chosen at random. The predictors are standard Gaussian with correlation 0.2 within a group and zero otherwise. Finally, Gaussian noise with standard deviation 4.0 was added to each observation.

Figure 2 shows the signs of the estimated coefficients from the lasso, group lasso and sparse group lasso, using a well chosen tuning parameter for each method (we set $\lambda_1 = \lambda_2$ for the sparse group lasso). The corresponding misclassification rates for the groups and individual features are shown in Figure 3. We see that the sparse group lasso strikes an effective compromise between the lasso and group lasso, yielding sparseness at the group and individual predictor levels.

Authors note: After completion of this note, we became aware the related recent work of Puig et al. (2009) on multidimensional shrinkage thresholding operators.

References

- Bertsekas, D. (1999), *Nonlinear programming*, Athena Scientific.
- Meier, L., van de Geer, S. & Bühlmann, P. (2008), ‘The group lasso for logistic regression’, *Journal of the Royal Statistical Society B* **70**, 53–71.
- Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R. & Wang, P. (2009), ‘Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer’, *Annals of Applied Statistics (to appear)*.
- Puig, A., Wiesel, A. & Hero, A. (2009), ‘A multidimensional shrinkage-thresholding operator,’ statistical signal processing, in ‘SSP ’09. IEEE/SP 15th Workshop on Statistical Signal Processing’, pp. 113–116.
- Yuan, M. & Lin, Y. (2007), ‘Model selection and estimation in regression with grouped variables’, *Journal of the Royal Statistical Society, Series B* **68**(1), 49–67.

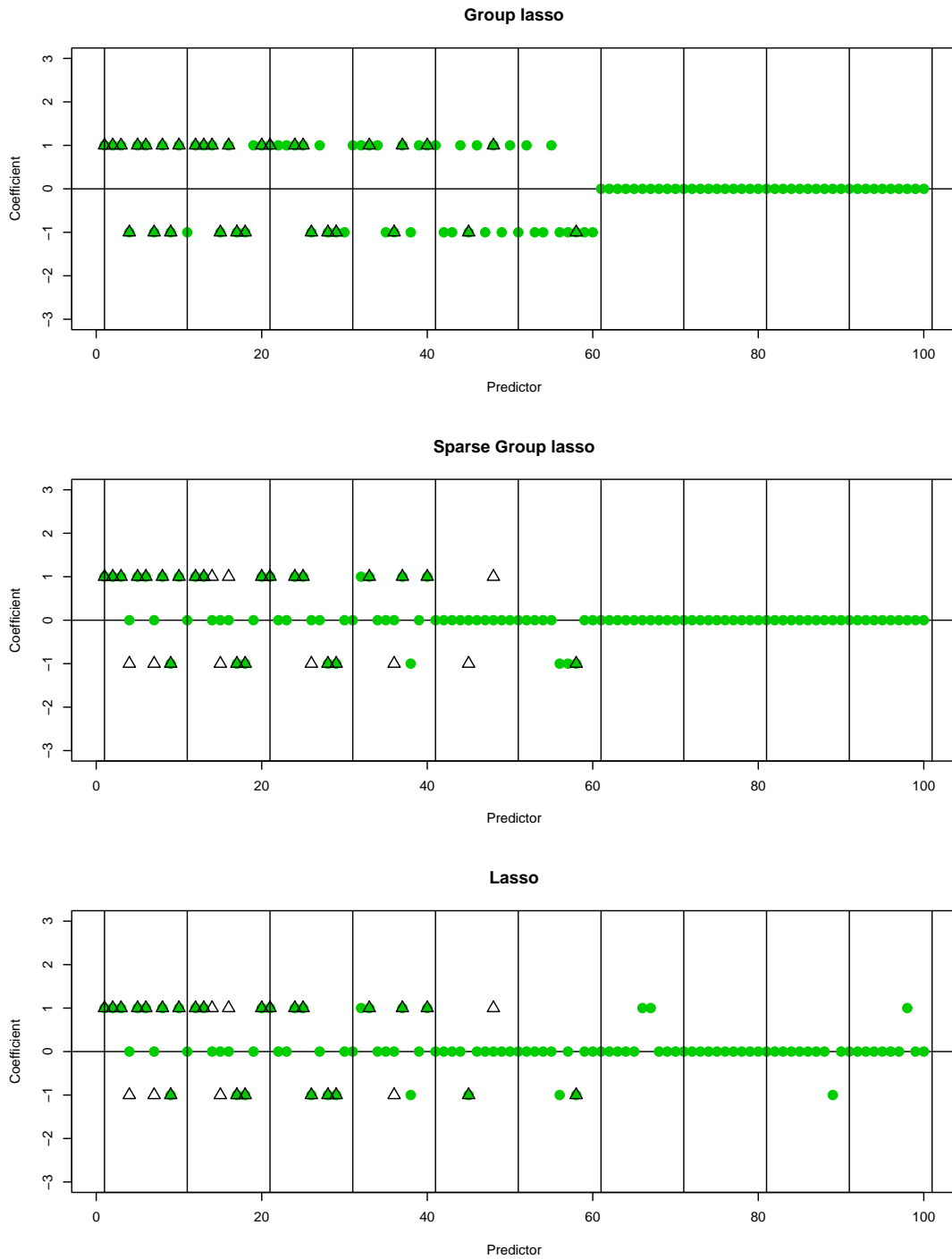


Figure 2: Results for the simulated example. True coefficients are indicated by the open triangles while the filled green circles indicate the sign of the estimated coefficients from each method.

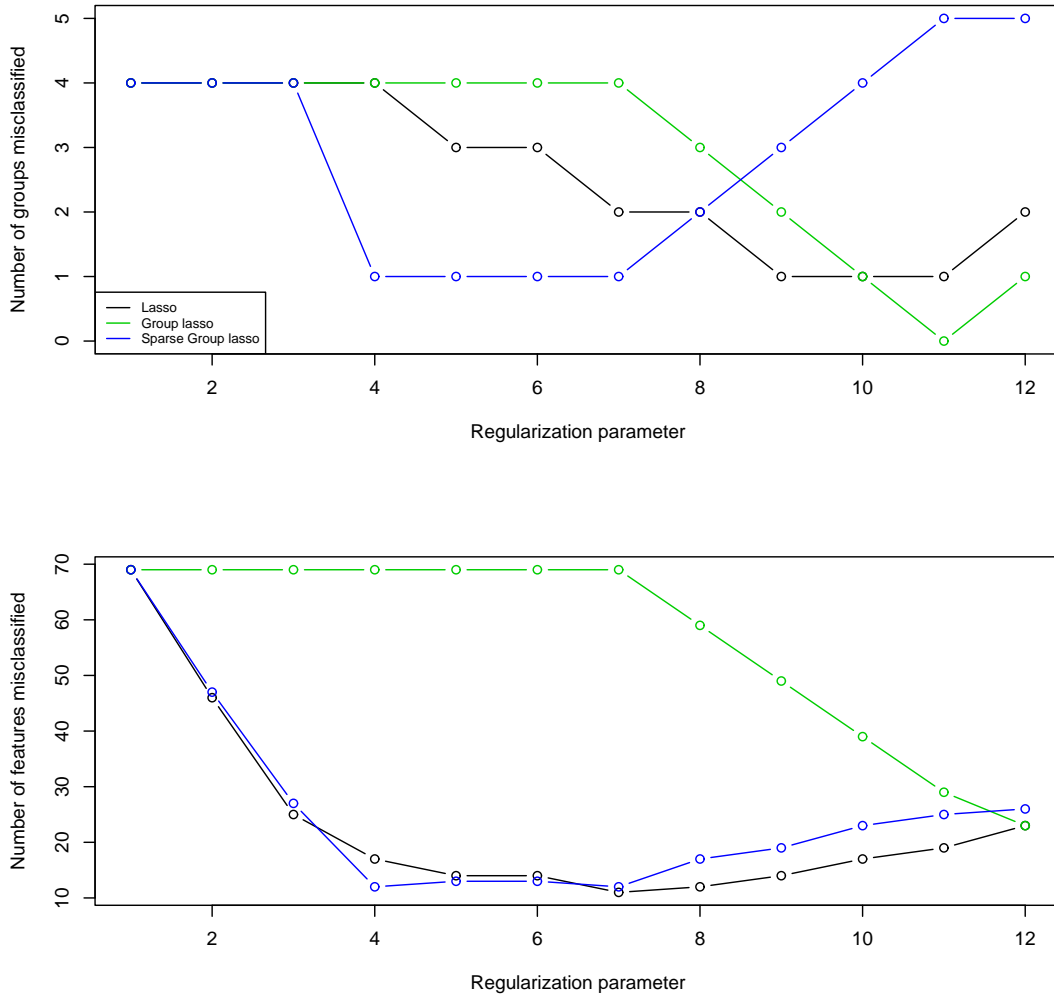


Figure 3: Results for the simulated example. The top panel shows the number of groups that are misclassified as the regularization parameter is varied. A misclassified group is one with at least one nonzero coefficient whose estimated coefficients are all set to zero, or vice versa. The bottom panel shows the number of individual coefficients that are misclassified, that is, estimated to be zero when the true coefficient is nonzero or vice-versa.