Statistica Sinica **22** (2012), 983-1001
doi:http://dx.doi.org/10.5705/ss.2011.075

STANDARDIZATION AND THE GROUP LASSO PENALTY

Noah Simon and Robert Tibshirani

Stanford University

Abstract: We re-examine the original Group Lasso paper of Yuan and Lin (2007). The form of penalty in that paper seems to be designed for problems with uncorrelated features, but the statistical community has adopted it for general problems with correlated features. We show that for this general situation, a Group Lasso with a different choice of penalty matrix is generally more effective. We give insight into this formulation and show that it is intimately related to the uniformly most powerful invariant test for inclusion of a group. We demonstrate the efficacy of this method– the "standardized Group Lasso"– over the usual group lasso on real and simulated data sets. We also extend this to the Ridged Group Lasso to provide within group regularization as needed. We discuss a simple algorithm based on group-wise coordinate descent to fit both this standardized Group Lasso and Ridged Group Lasso.

Key words and phrases: Lasso, group Lasso, penalized regression, regularization, standardization, high dimensional data.

1. Introduction

Consider the usual linear regression framework. Our data consists of an n-response vector y, and an n by p matrix of features, X. In many recent applications $p >> n$: a case where standard linear regression fails. In these scenarios we often have the prior belief that few of the measured covariates are "important." A number of different approaches have attempted to incorporate this prior belief. One widely used approach of Tibshirani (1996) regularized the problem by penalizing the ℓ_1 norm of the solution. This approach, known as the lasso, minimizes

$$\frac{1}{2}\left\|y - X\beta\right\|_2^2 + \lambda||\beta||_1. \tag{1.1}$$

The ℓ_1 norm penalty promotes sparsity in the solution vector $\hat{\beta}$. Suppose, further, that our predictor variables were divided into m different groups— for example in categorical data, we may want to group by question. We are given these group memberships and rather than sparsity in individual elements of β, we would like a solution that uses only a few of the groups. Yuan and Lin (2007) proposed the Group Lasso criterion for this problem: find

$$\min_{\beta} \frac{1}{2}\left\|y - \sum_{l=1}^{m} X^{(l)}\beta^{(l)}\right\|_2^2 + \lambda \sum_{l=1}^{m} ||W_l\beta^{(l)}||_2, \tag{1.2}$$

where $X^{(l)}$ is the submatrix of X with columns corresponding to the predictors in group l, $\beta^{(l)}$ is the coefficient vector of that group, and W_l is some penalty matrix. In the original paper, they chose $W_\ell = \sqrt{p_l} I$, where p_l is the number of covariates in group l, yielding the criterion

$$\min_\beta \frac{1}{2}\left\| y - \sum_{l=1}^{m} X^{(l)}\beta^{(l)} \right\|_2^2 + \lambda \sum_{l=1}^{m} \sqrt{p_\ell}||\beta^{(l)}||_2. \tag{1.3}$$

This criterion exploits the non-differentiability of $||\beta^{(l)}||_2$ at $\beta^{(l)} = 0$; setting groups of coefficients to exactly 0. The sparsity of the solution is determined by the magnitude of the tuning parameter λ.

In proposing (1.3), Yuan and Lin (2007) assume that the data is orthonormal, or sphered within each group (i.e. that $X^{(l)\top} X^{(l)} = I$ for each l), and provide an algorithm for that case. For group matrices that are non-orthonormal, this requires sphering before application of the Group Lasso. However they do not point out the fact that this normalization changes the problem. Specifically, suppose that we start with non-orthonormal matrices X_ℓ at (1.3), If we sphere each X_ℓ and re-express (1.3) in terms of the new data, we get a problem that is not equivalent to the original one.

In the subsequent literature on the Group Lasso, there has been much confusion about orthonormalizing within groups. Many works explicitly do not sphere (Puig, Wiesel, and Hero (2009), Foygel and Drton (2010), Jacob, Obozinski, and Vert (2009), Hastie, Tibshirani, and Friedman (2008), among others), and many more make no mention of normalization. For the remainder of this paper we refer to the solution to (1.3) without orthonormalization within group as the *unstandardized* Group Lasso.

In this paper we consider the following simple variant which we henceforth refer to as the *standardized* Group Lasso

$$\min_\beta \frac{1}{2}\left\| y - \sum_{l=1}^{m} X^{(l)}\beta^{(l)} \right\|_2^2 + \lambda_2 \sum_{l=1}^{m} \sqrt{p_l}\left\| X^{(l)}\beta^{(l)} \right\|_2. \tag{1.4}$$

This is the Group Lasso with penalty matrix changed from $W_l = \sqrt{p_l} I$ to $W_l = \sqrt{p_l} X^{(l)}$.

For problems with no overdetermined groups (all $p_l \leq n$), it turns out that the standardized Group Lasso is exactly equivalent to orthonormalizing within groups, running the unstandardized Group Lasso, then transforming the coefficients back to the original basis (it is, in fact, just the vanilla Group Lasso with standardization). This gives us a nice interpretation of sphering within groups: it is equivalent to penalizing the fit of each group $X^{(l)}\beta^{(l)}$ rather than the individual coefficients.

This is not a new idea for model selection: Ravikumar et al. (2009) proposed a similar but more general criterion to fit additive models in a sparse way (spAM)

$$\min_{\beta} \frac{1}{2}\Big\|y - \sum_{l=1}^{m} f_l(X)\Big\|_2^2 + \lambda_2 \sum_{l=1}^{m} \Big\|f_l(X)\Big\|_2, \tag{1.5}$$

where f_i are flexible functions of X used to estimate y. In particular, if we consider each $X^{(l)}$ as a linear basis for some function, spAM reduces to a very similar problem

$$\min_{\beta} \frac{1}{2}\Big\|y - \sum_{l=1}^{m} X^{(l)}\beta^{(l)}\Big\|_2^2 + \lambda_2 \sum_{l=1}^{m} \Big\|X^{(l)}\beta^{(l)}\Big\|_2. \tag{1.6}$$

Though not a new idea, it appears that this connection to orthonormalization has been largely overlooked, and the unstandardized Group Lasso is in common use.

In contrast to the unstandardized Group Lasso solution, the standardized solution behaves like a continuous version of stepwise regression with grouped variables. In particular in the case of orthogonality between groups, we show in Section 3 that the standardized Group Lasso chooses groups roughly according to the uniformly most powerful invariant test and chooses the same groups as grouped subset selection, while the unstandardized Group Lasso does not (these connections will be discussed in more depth in Section 3). If the size of each group is one, the standardized Group Lasso gives the usual lasso solution for X with column norms standardized to one.

In this paper we explore properties of criterion (1.4). We show that in general this is a more natural extension of the lasso to the group setting than the unstandardized Group Lasso. We describe a simple modification of the algorithm in Yuan and Lin (2007) to fit this for general X. We also show the efficacy of the criterion on real and simulated data, and show that it decreases subset selection and estimation error as compared to the unstandardized group lasso. Furthermore, for data that requires within group regularization as well, we propose the ridged Group Lasso, a variant which fits a thresholded ridge operator within each group.

2. Solution Properties

To better understand the advantages of our approach, we begin by characterizing the solution to

$$\min_{\beta} \frac{1}{2}\Big\|y - \sum_{l=1}^{m} X^{(l)}\beta^{(l)}\Big\|_2^2 + \lambda_2 \sum_{l=1}^{m} \sqrt{p_l}\Big\|X^{(l)}\beta^{(l)}\Big\|_2. \tag{2.1}$$

We first note that in order to ensure a unique solution we need all $X^{(l)}$ to have full column rank — in particular this means $p_l \leq n$ for $l = 1, \ldots, m$. For the most part, we assume this to be the case. In Section 4.3, we propose a revision in the case that some $X^{(l)}$ is rank deficient.

If each $X^{(l)}$ has full column rank, then we can decompose it into $X^{(l)} = U^{(l)}R^{(l)}$, where $U^{(l)}$ is an $n \times p_l$ matrix with orthonormal columns and $R^{(l)}$ a $p_l \times p_l$ invertible matrix. We can rewrite our criterion as

$$\min_{\theta} \frac{1}{2}\left\| y - \sum_{l=1}^{m} U^{(l)}\theta^{(l)} \right\|_2^2 + \lambda_2 \sum_{l=1}^{m} \sqrt{p_l} \left\| U^{(l)}\theta^{(l)} \right\|_2, \tag{2.2}$$

where

$$\hat{\beta}^{(l)} = \left(R^{(l)}\right)^{-1} \hat{\theta}^{(l)}. \tag{2.3}$$

Now, noting that $\|U^{(l)}\theta^{(l)}\|_2 = \|\theta^{(l)}\|_2$, we can simplify our criterion to

$$\min_{\theta} \frac{1}{2}\left\| y - \sum_{l=1}^{m} U^{(l)}\theta^{(l)} \right\|_2^2 + \lambda_2 \sum_{l=1}^{m} \sqrt{p_l} \left\| \theta^{(l)} \right\|_2, \tag{2.4}$$

where (2.3) still holds. This reduces things to the orthogonal case of Yuan and Lin (2007); as in their work, if we consider the Karush-Kuhn optimality conditions, we see that

$$\hat{\theta}^{(k)} = \left(1 - \frac{\sqrt{p_k}\lambda}{||U^{(k)\top} r^{(-k)}||_2}\right)_+ U^{(k)\top} r^{(-k)}$$

with $r^{(-k)} = y - \sum_{l \neq k} X^{(l)}\hat{\beta}^{(l)}$ the kth partial residual. Transforming back to the original coordinates, this becomes

$$\hat{\beta}^{(k)} = \left(1 - \frac{\sqrt{p_k}\lambda}{||P_{\mathrm{col}(X^{(k)})} r^{(-k)}||_2}\right)_+ \left(X^{(k)\top} X^{(k)}\right)^{-1} X^{(k)\top} r^{(-k)}, \tag{2.5}$$

where $P_{\mathrm{col}(X^{(k)})}$ is the orthogonal projection operator onto the column space of $X^{(k)}$. We use these optimality conditions in Sections 3 and 4 to explore properties of the standardized Group Lasso solution, and in Section 5 to give an algorithm for finding the solution.

3. Connections to Other Problems

From the optimality conditions in (2.5), we can draw connections between the standardized Group Lasso and other statistical testing and estimation procedures: usual linear regression, uniformly most powerful invariant testing, sparse additive models, and the grouped ℓ_0 subset selection problem.

3.1. Connection to linear regression

Condition (2.5) is very similar to the optimality conditions for usual linear regression,

$$\hat{\beta}^{(k)} = \left(X^{(k)\top} X^{(k)}\right)^{-1} X^{(k)\top} r^{(-k)}.$$

Contrasting this with (2.5), we see that for the standardized Group Lasso, instead of fully fitting to the residual (as in regression), we soft threshhold the norm of the fit, but keep the direction unchanged. This parallels the optimality conditions for the usual lasso, which is just a soft-threshholding of the univariate fit Friedman, Hastie, and Tibshirani (2009). This relationship is not paralleled in the unstandardized case.

3.2. Connection to UMPI testing

One of the strongest arguments in favor of the standardized group lasso is its connection to uniformly most powerful invariant testing. Assume we have fit a standard linear regression model on $X = \left(X^{(1)}\, X^{(2)}\, \ldots\, X^{(m-1)}\right)$ and are deciding whether or not to add a new group, $X^{(m)}$. If the variance of the noise σ^2 were known, then the uniformly most powerful invariant test of $\mathrm{H} : \hat{\beta}^{(m)} = 0$ at level α is to reject H if

$$\left\|\hat{y}_m - \hat{y}_{m-1}\right\|_2^2 \geq \sigma^2 \chi^2_{\mathrm{ncol}X^{(m)},1-\alpha}, \tag{3.1}$$

where $\hat{y}_i$ is the prediction from the linear fit on $(X^{(1)}X^{(2)}\ldots X^{(i)})$. The squared difference in (3.1) is exactly $||P_{\mathrm{col}(X^{(m)})} r^{(-m)}||_2^2$. In the standardized group lasso, we also decide group inclusion based on the magnitude of $||P_{\mathrm{col}(X^{(m)})} r^{(-m)}||_2^2$, however we only infinitesimally fit each group at each step. Thus, in some sense, the standardized Group Lasso is a continuous analogue to group stepwise regression. Notice that $||P_{\mathrm{col}(X^{(m)})} r^{(-m)}||_2^2$ is highly dependent on the group size p_m. Under $\beta^{(m)} = 0$ (ie $X^{(m)}$ not in the model), we have that

$$E\left[\left\|P_{\mathrm{col}(X^{(m)})} r^{(-m)}\right\|_2^2\right] = p_m \sigma^2,$$

so for comparability of groups we include a factor $\sqrt{p_m}$ in our penalty. Because σ^2 is included in all terms, estimating it is unnecessary. For example if we use cross-validation to choose λ, we implicity estimate σ^2.

3.3. Connection to spAM

Ravikumar et al. (2009) suggest a similar "fit penalized" framework for additive models,

$$\min_{f_l} \frac{1}{2}\left\|y - \sum_{l=1}^{m} f_l(X)\right\|_2^2 + \lambda_2 \sum_{l=1}^{m} \left\|f_l(X)\right\|_2.$$

In fact, if one considers each f_l to be a function with basis $X^{(l)}$, their approach reduces to

$$\min_{\beta} \frac{1}{2}\left\|y - \sum_{l=1}^{m} X^{(l)}\beta^{(l)}\right\|_2^2 + \lambda_2 \sum_{l=1}^{m} \left\|X^{(l)}\beta^{(l)}\right\|_2.$$

a very similar criteria to the standardized Group Lasso. However, because they do not make the connection to testing, they miss the $\sqrt{p_l}$ factor in the penalty term. The $\sqrt{p_l}$ factor is important to make sure that groups with larger column spaces play on an even footing with smaller groups — though their is work that does something similar using kernel norms Yu, Wainwright, and Raskutti (2010) and smoothness penalties Meier, Van De Geer, and Buhlmann (2009). Koltchinskii and Yuan (2008) and Koltchinskii and Yuan (2010) also give similar kernel formalizations as well as impressive reconstruction results.

3.4. Connection to group ℓ_0 subset selection

The Group Lasso is often used as a surrogate for the group ℓ_0 subset selection problem — to find

$$\hat{\beta} = \operatorname*{argmin}_{\beta} L(\beta),$$
$$\text{s.t. Number of nonzero groups} \leq c.$$

For the standardized Group Lasso this is very appropriate. To illustrate this consider the case of equal group sizes ($p_i = p_j$ for all i, j) and orthogonality between groups ($X^{(i)\top}X^{(j)} = 0$ for any $i \neq j$). We have

$$L(\beta) = \frac{1}{2}\left\|y - \sum_{l=1}^{m} X^{(l)}\beta^{(l)}\right\|_2^2 \tag{3.2}$$

$$= \frac{1}{2}\left\|y_\perp\right\|^2 + \sum_{l=1}^{m} \left\|P_{\text{col}(X^{(l)})}(y) - X^{(l)}\beta^{(l)}\right\|_2^2, \tag{3.3}$$

where $y_\perp$ is the projection of y onto the orthogonal complement of all the $X^{(l)}$. If we solve the group subset selection problem (in general a combinatorially hard problem) we get

$$\hat{\beta}^{(k)} = \begin{cases} \left(X^{(k)\top}X^{(k)}\right)^{-1} X^{(k)\top} y : ||P_{\text{col}(X^{(k)})}(y)||_2 > \lambda_c \\ 0 \qquad\qquad\qquad\qquad\qquad\quad : \text{otherwise} \end{cases}.$$

This is a hard thresholding of the groups based on how well they fit the response, where λ_c is defined such that exactly c groups surpass the threshold. This is easy to see from (3.2) — the groups we include contribute 0 to the sum, and the

groups we do not contribute $||P_{\mathrm{col}(X^{(k)})}(y)||_2^2$. Referring to (2.5), we see that the standardized Group Lasso solution is

$$\hat{\beta}^{(k)} = \begin{cases} \left(1 - \frac{\sqrt{p_k}\lambda_c}{\left\|P_{\mathrm{col}\left(X^{(k)}\right)}(y)\right\|_2}\right)\left(X^{(k)\top}X^{(k)}\right)^{-1}X^{(k)\top}y & : ||P_{\mathrm{col}\left(X^{(k)}\right)}(y)||_2 > \sqrt{p_k}\lambda_c \\ 0 & : \text{otherwise} \end{cases}$$

and, in particular, has the same order of active groups entering the model as the group subset problem (though the penalty parameters are off by a constant factor). Furthermore, the group solutions are given by soft-thresholding the ℓ_2-norm of the unrestricted fits. This soft versus hard thresholding parallels the relationship between the regular lasso and the best subset selection problem in the ungrouped orthogonal case.

4. Comparison with the Unstandardized Group Lasso

The unstandardized Group Lasso lacks many of the attractive qualities of its standardized counterpart. As shown in Meier, van de Geer, and Bühlmann (2008), the unstandardized Group Lasso chooses groups for which

$$\frac{1}{\sqrt{p_k}}||X^{(k)\top}r^{(-k)}||_2 > \lambda_c,$$

the "average" covariance is large. This ignores the multivariate nature of the problem, and more obviously causes problems if the covariates are on different scales. Furthermore, it has no relation to the UMPI test or to group ℓ_0 subset selection.

One disadvantage of the standardized Group Lasso is that it can only handle problems with $X^{(l)}$ full column rank for all l (this is not a limitation of the unstandardized Group Lasso). However in Section 6 we suggest a modification to remedy this problem.

4.1. One active group

To look more closely at the differences between the two criteria, we delve slightly further into the orthogonal case and consider a generative model with one active group:

$$y = X^{(1)}\beta + \epsilon,$$

where $\epsilon \sim N(0, I)$, the $X_i^{(k)}$ are all unit vectors, and the empircal correlations between group are all 0, ie. $X^{(k)^\top}X^{(l)} = \mathbf{0}$ for $k \neq l$. We are interested in differences between the standardized and unstandardized Group Lasso under

this simplified scheme — we compare the ways by which the two criteria choose the first group to enter the model.

For the standardized Group Lasso, we choose the group with the largest $(1/\sqrt{p_k})\left\|P_{\mathrm{col}\left(X^{(k)}\right)}(y)\right\|_2$. For $k \neq 1$ (the null groups) we have

$$\frac{1}{p_k}||P_{\mathrm{col}\left(X^{(k)}\right)}(y)||_2^2 = \frac{1}{p_k}||P_{\mathrm{col}\left(X^{(k)}\right)}\left(X^{(k)}\beta\right) + P_{\mathrm{col}\left(X^{(k)}\right)}(\epsilon)||_2^2 \tag{4.1}$$

$$= \frac{1}{p_k}||P_{\mathrm{col}\left(X^{(k)}\right)}(\epsilon)||_2^2 \tag{4.2}$$

$$\sim \frac{1}{p_k}\chi^2_{p_k}, \tag{4.3}$$

where $\chi^2_{p_k}$ is a chi-square distribution with p_k degrees of freedom.

For the active group we have

$$\frac{1}{p_1}||P_{\mathrm{col}\left(X^{(1)}\right)}(y)||_2^2 = \frac{1}{p_1}||P_{\mathrm{col}\left(X^{(1)}\right)}\left(X^{(1)}\beta\right) + P_{\mathrm{col}\left(X^{(1)}\right)}(\epsilon)||_2^2 \tag{4.4}$$

$$= \frac{1}{p_1}||X^{(1)}\beta + P_{\mathrm{col}\left(X^{(1)}\right)}(\epsilon)||_2^2 \tag{4.5}$$

$$\sim \frac{1}{p_1}\chi^2_{p_1}\left(||X^{(1)}\beta||_2^2\right) \tag{4.6}$$

with $\chi^2_{p_1}\left(||X^{(1)}\beta||_2^2\right)$ a chi-square distribution with p_1 degrees of freedom and noncentrality parameter $||X^{(1)}\beta||_2^2$.

To find the correct group, the standardized Group Lasso needs that (4.6) be greater than (4.3) for all non-active groups ($k \geq 2$). This criterion is invariant under reparametrization within group. In particular these equations are reminiscent of the size and power of the uniformly most powerful invariant test for inclusion of a group, as discussed in Section 3.2.

For the unstandardized Group Lasso, the first group is selected to have the largest $(\sqrt{p_k})^{-1}\left\|{X^{(k)}}^\top y\right\|_2$. For $k \neq 1$ (the null groups) we have

$$\frac{1}{p_k}||{X^{(k)}}^\top y||_2^2 = \frac{1}{p_k}||{X^{(k)}}^\top \epsilon||_2^2 \tag{4.7}$$

$$= \frac{1}{p_k}\epsilon^\top X^{(k)}{X^{(k)}}^\top \epsilon \tag{4.8}$$

$$= \frac{1}{p_k}\sum_{i=1}^{p_k} d_{k,i}\langle u_{k,i}, \epsilon\rangle^2 \tag{4.9}$$

$$\sim \frac{1}{p_k}\sum_{i=1}^{p_k} d_{k,i}\chi^2_1, \tag{4.10}$$

where $\{u_{k,j}\}_{j=1}^{p_k}$ are the eigenvectors of $X^{(k)}X^{(k)^\top}$ with corresponding eigenvalues $\{d_{k,j}\}_{j=1}^{p_k}$ (these are the same eigenvalues as for $X^{(k)^\top}X^{(k)}$, the sample correlation matrix). Note that while the expectation of this quantity is $\text{Trace}(X^{(k)^\top}X^{(k)}) = p_k$ regardless of the correlation structure, its variance $2\sum_j d_{k,j}^2$ is greatly increased by correlation.

For the first group we have

$$\frac{1}{p_1}||X^{(1)^\top}y||_2^2 = \frac{1}{p_1}||X^{(1)^\top}(X^{(1)}\beta + \epsilon)||_2^2 \tag{4.11}$$

$$= \frac{1}{p_1}\sum_{i=1}^{p_1} d_{1,i}\langle u_{1,i}, X^{(1)}\beta + \epsilon\rangle^2. \tag{4.12}$$

To find the correct group the unstandardized Group Lasso needs that (4.12) be greater than (4.10) for all non-active groups ($k \geq 2$). We see that this is more likely to happen if $\langle u_{1,i}, X^{(1)}\beta\rangle$ is large for large $d_{1,i}$. This means that to have power we need the mean, $X^{(1)}\beta$, to be in a direction similar to the majority of columns of $X^{(1)}$. This is an unsatisfactory criterion — in particular it is highly dependent on the parametrization of the group. The closer the columns of $X^{(k)}$ are to orthogonal (within each group k), the closer the standardized and unstandardized solutions (for a near orthogonal matrix the $d_{k,j}$ are all near 1).

5. Fitting the Model

Fitting the standardized Group Lasso is straightforward and fast and is, in fact, significantly easier to fit than the unstandardized Group Lasso. The algorithm we discuss is a block coordinate descent algorithm, optimizing over each group given fixed coefficient values for all others. Since our problem is convex and the non-differentiable part of our objective is group separable, our algorithm converges to the global minimum (Tseng, 2001). We have seen in (2.5) that if we choose $k \leq m$ and fix $\beta^{(l)}$ for all $l \neq k$, then

$$\hat{\beta}^{(k)} = \left(1 - \frac{\sqrt{p_k}\lambda}{||P_{\text{col}(X^{(k)})}r^{(-k)}||_2}\right)_+ (X^{(k)\top}X^{(k)})^{-1}X^{(k)\top}r^{(-k)} \tag{5.1}$$

where, again, $r^{(-k)}$ is the partial residual $r^{(-k)} = y - \sum_{l\neq k} X^{(l)}\beta^{(l)}$. This gives us the following algorithm.

Simple Algorithm for the standardized Group Lasso

1. Set $r = y$, $\hat{\beta}^{(k)} = 0$ for all k.
2. Cyclically iterate through the groups until convergence; for each group (k) execute the following.

(a) Update $r^{(-k)}$ by
$$r^{(-k)} = r + X^{(k)}\hat{\beta}^{(k)}.$$
(b) Solve for $\hat{\beta}^{(k)}$ by
$$\hat{\beta}^{(k)} = \left(1 - \frac{\sqrt{p_k}\lambda}{||P_{\text{col}(X^{(k)})}r^{(-k)}||_2}\right)_+ (X^{(k)\top}X^{(k)})^{-1}X^{(k)\top}r^{(-k)}.$$
(c) Update r by
$$r = r^{(-k)} - X^{(k)}\hat{\beta}^{(k)}.$$

In contrast, the unstandardized Group Lasso implicitly applies a ridge penalty within nonzero groups. Puig, Wiesel, and Hero (2009) and Foygel and Drton (2010) use clever dual arguments to show that the ridge penalty can be found with a 1-dimensional line search, however this still increases complexity, and slows down the algorithm.

This algorithm is very similar to the original algorithm of Yuan and Lin (2007) for the orthogonalized case, however, we work in the original coordinate system. Because the only variables that change at each step in our algorithm are $\hat{\beta}^{(k)}$ and $r^{(-k)}$, we can speed this algorithm up by pre-multiplying and storing $(X^{(k)\top}X^{(k)})^{-1}X^{(k)\top}$ and $P_{\text{col}(X^{(k)})}$. Taking this further, notice that solving for $\hat{\beta}^{(k)}$ at each step is unnecessary, we only need $X^{(k)}\hat{\beta}^{(k)}$, the fit of group k. If we work in an orthogonal basis for each group then we can solve for the fit at each iteration in time np_i rather than n^2 (recall that $p_i < n$). This leads to a new algorithm, similar to the old, with slightly more bookkeeping and speed.

More Efficient Algorithm for the standardized Group Lasso

1. QR factorize $X^{(l)}$ giving Q_l and R_l for $l = 1, \ldots, m$ (with Q_l of dimension $n \times p_i$).
2. Initialize the fit vectors $\tilde{F}_i = \underline{0}$ for $i = 1, \ldots, m$, and the residual vector $r = y$.
3. Cyclically iterate through the groups until convergence; for each group, k, execute a-e.
   (a) Set $r^{(-k)} = r - \tilde{F}_k$.
   (b) Solve for the coefficient vector in the orthogonalized basis by
   $$\theta^{(k)} = \left(1 - \frac{\sqrt{p_k}\lambda}{||Q_k Q_k^\top r^{(-k)}||_2}\right)_+ Q_k^\top r^{(-k)}. \tag{5.2}$$
   (c) Solve for $F_k = \left(X^{(k)}\hat{\beta}^{(k)}\right)$ by
   $$F_k = Q_k\theta^{(k)}. \tag{5.3}$$

(d) Update the residuals $r = r^{(-k)} + F_k$.

(e) Update the "old fit" $\tilde{F}_k = F_k$.

4. After convergence, transform to the original coordinates by $\hat{\beta}^{(k)} = R_k^{-1}\theta^{(k)}$.

The details of this algorithm are the same as in Yuan and Lin (2007). It can also be seen as the group analog to the coordinate descent algorithm described in Friedman, Hastie, and Tibshirani (2010); as in the univariate case we can use warm starts, active sets, and other bells and whistles to speed up this algorithm. However, we do have the disadvantage that we must calculate a QR decomposition to all groups at the start (not just the active set). Ravikumar et al. (2009) discuss a similar algorithm for fitting sparse additive models — this type of algorithm is a very natural approach for solving problems of this nature (where the penalty term matches the fit term).

5.1. Computational complexity

Computational complexity is somewhat tricky to calculate for iterative algorithms as it depends greatly on the number of iterations. The noniterative components, the QR decomposition and backsolve, are $O(np_i^2)$ and $O(p_i^2)$ per group, respectively, and so are $O(n\sum p_i^2)$ overall.

Within each iteration, we must update (5.2) for each group (which is $O(np_i)$ due to the QR decomposition) and calculate new residuals by (5.3) (also $O(np_i)$). Thus, within each iteration, $O(n\sum p_i) = O(np)$ calculations are required. In practice we often find a set of active variables after only a few passes, iterate over those few groups until convergence, then check the KKT conditions to make sure no new groups should enter; most passes take only $O(n\sum p_i)$, where this sum is over groups in the eventual fit. While the overall calculation appears to be dominated by the QR decomposition, often the groups are reasonably small and most of the time is spent iterating.

5.2. Generalization to other likelihoods

As with the unstandardized Group Lasso, one can generalize the standardized Group Lasso to any log-concave likelihood, L, by

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ -\log L(\beta; X) + \lambda_2 \sum_{l=1}^{m} \sqrt{p_l} \left\| X^{(l)}\beta^{(l)} \right\|_2 \right\}.$$

For example a penalized logistic model would look like

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \log\left(1 + e^{x_i^\top \beta}\right) - y^\top X\beta + \lambda_2 \sum_{l=1}^{m} \sqrt{p_l} \left\| X^{(l)}\beta^{(l)} \right\|_2 \right\}.$$

Because the penalty function is convex and the negative log-likelihood is convex, problems of this type are convex. Unfortunately this criterion no longer corresponds to the UMPI test; however, because one can approximate the negative log-likelihood by quadratic loss, the penalty is still roughly correct from a statistical testing viewpoint (it corresponds to more of a score test). Because the Hessian of our likelihood is no longer constant as it was in the Gaussian case (it has changed from $X^\top X$ to $X^\top W(\beta)X$ with $W(\beta)$ diagonal in the case of exponential families), we are not using the exact score statistic. Instead we disregard W and just implicitly approximate it with a multiple of the identity — in exponential families this is reasonable as W is just the estimated inverse variances of the observations, and for these very overdetermined problems one might want to shrink those variances towards each other. To more exactly use the score statistic, our penalty would be $\sum ||W(\beta^{(k)})^{1/2}X^{(k)}\beta^{(k)}||_2$ but this is no longer convex, and while one might consider using a penalty of this type, it is beyond the scope of this paper.

5.3. Fitting the model for other losses

We can fit the standardized Group Lasso for a general log-concave likelihood using nearly the same algorithm as for Gaussian loss (combined with a penalized quasi-Newton step). For a given estimate, $\tilde{\beta}$, of $\hat{\beta}$, we can write a quadratic approximation, $Q(\beta, \tilde{\beta})$, of the log likelihood, $\ell(\beta)$, centered about $\tilde{\beta}$ as

$$Q\left(\beta, \tilde{\beta}\right) = \ell\left(\tilde{\beta}\right) + \left(X\beta - X\tilde{\beta}\right)^\top \ell'\left(\tilde{\beta}\right) + \frac{1}{2}\left(X\beta - X\tilde{\beta}\right)^\top \ell''\left(\tilde{\beta}\right)\left(X\beta - X\tilde{\beta}\right),$$

where ℓ', and ℓ'' are the gradient and Hessian of ℓ with respect to $X\beta$. We can majorize this by dominating the negative Hessian, $-\ell''$, by tI, the Lipschitz constant of the negative log-likelihood times the identity, and add in our penalty term. This majorization is important because the penalized weighted least squares problem is markedly less straightforward to solve than the unweighted. Thus, within each "Newton" step we solve

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} - \left(X\beta - X\tilde{\beta}\right)^\top \ell'\left(\tilde{\beta}\right) + \frac{t}{2}\left(X\beta - X\tilde{\beta}\right)^\top \left(X\beta - X\tilde{\beta}\right) + \lambda_2 \sum_{l=1}^{m} \sqrt{p_l}\left\|X^{(l)}\beta^{(l)}\right\|_2.$$

This reduces to

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{t}{2}\left\|X\tilde{\beta} - \frac{1}{t}\ell'\left(\tilde{\beta}\right) - X\beta\right\|_2^2 + \lambda_2 \sum_{l=1}^{m} \sqrt{p_l}\left\|X^{(l)}\beta^{(l)}\right\|_2 \tag{5.4}$$

which we can solve efficiently with groupwise descent. After solving (5.4), we use our new $\hat{\beta}$ to center the next Newton step. This "majorize-minimize" style of optimization has been discussed for similar problems in Tseng and Yun (2009) and Meier, van de Geer, and Bühlmann (2008), among others. As in the Gaussian case, this algorithm converges to the global minimum (Nesterov, 2007).

6. Numerical Comparisons

In this section we compare the standardized and unstandardized Group Lasso on simulated data. We simulated responses via the model

$$y = \sum_{l=1}^{g} X^{(l)}\beta^{(l)} + \sigma\epsilon$$

with g groups with nonzero coefficients in the model ($= 1, 2, 3$ in our simulations) and G total groups (all equally sized; in the categorical case, this is the number of categorical predictors).

We chose X, β, and σ according to two different schemes. The first scheme was meant to approximate problems with grouped continuous variables (eg. gene pathways). Here, X was simulated as multivariate normal with between-group correlation ψ and within-group correlation ρ. For each active group, l, We set $\beta^{(l)} = (-2, -1, 0, 1, 2, 0, \ldots, 0)$, and σ so that the signal to noise ratio was 1.

In the second set of simulations, we used categorical predictors. For each simulation, each predictor had the same number of levels ($p_i = 3, 5$, or 7 for different simulations). The probability of an observation being from a particular level of a factor was given according to one of two schemes, either equal probability of being in each level, or an overwhelming probability of being in the first level ($\text{prob}_1 = (1 + p_i)/2p_i$), with the rest of the probability split evenly between the other levels. In both schemes there was independence between factors. For a given simulation, the scheme for all groups was the same (either "equal probability", or "overwhelm"). Regardless of scheme, β for the nonzero groups was given as follows

$$\begin{aligned}
&\beta = (0, -1, 1) \qquad \text{for } p_i = 3,\\
&\beta = (0, -1, 1, 0, 0) \qquad \text{for } p_i = 5,\\
&\beta = (0, -2, -1, 1, 2, 0, 0) \qquad \text{for } p_i = 7.
\end{aligned}$$

The leading 0 means that in the "overwhelm" scheme, the majority classification has no effect. In all of the categorical simulations σ was set such that the signal to noise ratio was 4.

We ran the both standardized and unstandardized Group Lasso on these data and checked if the first g groups to enter each fit matched the true g groups

in the model. We repeated this 100 times for each of several values of the different parameters and calculated the proportion of times the true groups were correctly identified. Table 1 has the results for the continuous predictors and Table 2 has the results for the categorical predictors.

Referring to Table 1 one can see that standardized Group Lasso performs uniformly better than the unstandardized for our continuous predictors. As expected, this difference is particularly pronounced when there is high within group correlation. However there is not nearly as substantial a difference in Table 2. This is unsurprising — categorical variables are already orthogonalized within group, so the standardized Group Lasso only scales columns to have the same norm. In the case of equal probability per level the columns have nearly the same norm without normalization, so the standardized and unstandardized Group Lasso should have very similar solutions. In the case of "overwhelm" (most observations belonging to a single level), the unstandardized Group Lasso gives this level most of the importance, while the standardized gives all levels equal importance. This is potentially useful, as sometimes the effects of being in the "minority" groups are what is of interest, and the unstandardized Group Lasso will tend to miss these.

6.1. Data example

We compared the prediction accuracy of the standardized and unstandardized Group Lasso on the freely available horse colic dataset of Frank and Asuncion (2010). The data consist of horse surgery measurements — 22 covariates for 300 horses, trying to predict a binary endpoint (whether or not a lesion is a surgical issue). We used the penalized logistic model discussed in Section 5.2.

We limited our analysis to covariates with less than 50% missing (excluding 2). Of the remaining 20 variables, 14 were categorical and 6 were continuous. We grouped together each set of indicator variables corresponding to a given categorical covariate, and assigned each continuous variable its own group. There were a number of missing values which we imputed by mode imputation for categorical variables and mean imputation for continuous variables. We were interested in the methods' performaces in the case that $n \sim p$ (and $n < p$) so we considered only a subset of the horses. We first ran our analysis on the 67 horses with fewer than 3 missing covariates (46 of whom had surgical lesions), and then on the 32 with fewer than 2 missing covariates (24 of whom had surgical lesions). We ran 5-fold cross validation with each method to estimate the prediction accuracy of each approach.

Referring to Figure 1 we see that on the smaller subset of 32 horses, the unstandardized group lasso performed poorly, choosing the null model (only the intercept), while the standardized Group Lasso improved from 24 (75%) correct

Table 1. Proportions of correct nonzero group identifications for standardized and unstandardized Group Lasso out of 100 simulated data sets for continuous predictors.

		Correlation (ψ, ρ)			
		(0, 0.2)	(0, 0.8)	(0.167, 0.33)	(0.33, 0.67)
$N = 50$, $p = 200$, $G = 10$					
1 group	SGL	0.97	0.93	0.96	0.91
	UGL	0.63	0.07	0.48	0.14
2 groups	SGL	0.36	0.41	0.3	0.33
	UGL	0.12	0.05	0.19	0.05
3 groups	SGL	0.16	0.14	0.11	0.10
	UGL	0.11	0.01	0.04	0.03
$N = 50$, $p = 100$, $G = 20$					
1 group	SGL	1.00	1.00	1.00	1.00
	UGL	0.97	0.05	0.91	0.41
2 groups	SGL	0.75	0.75	0.76	0.79
	UGL	0.41	0.01	0.34	0.09
3 groups	SGL	0.27	0.28	0.29	0.34
	UGL	0.13	0.00	0.08	0.02
$N = 100$, $p = 400$, $G = 40$					
1 group	SGL	1.00	1.00	1.00	1.00
	UGL	0.99	0.02	0.92	0.26
2 groups	SGL	0.97	0.94	0.93	0.94
	UGL	0.61	0.00	0.38	0.01
3 groups	SGL	0.49	0.47	0.48	0.49
	UGL	0.18	0.00	0.16	0.00

guesses with just the intercept, up to 26 (81%) with its "optimal" model. On the larger problem, both methods chose non-null models, however the standardized Group Lasso was able to correctly categorize 60 (89%) horses at its peak as opposed to the 57 (85%) of the unstandardized. While these differences are not large, the groups were categorical and the levels were reasonably well balanced, so one would not expect large differences. However we do see improvement with the standardized criterion.

7. Ridged Group Lasso

As noted in Section 4, sometimes we may be using very large groups (eg. gene pathways in gene expression data), and regularization within group is necessary. In particular the solution to (2.2) is undefined if any $X^{(l)}$ is column rank deficient (which necessarily happens if any $p_i > n$). One should also note that in this case the $X^{(l)}$ cannot all be orthogonalized, so it would be impossible to orthogonalize then run the original proposal of Yuan and Lin (2007). However, rather than just using the unstandardized Group Lasso, we would like to combine penalizing

Table 2. Proportions of correct nonzero group identifications for standardized and unstandardized Group Lasso out of 100 simulated data sets for categorical predictors.

		Scheme	
		Equal Probs	Overwhelm
	$N = 200$, Levels $= 7$, $G = 10$		
1 group	SGL	0.81	0.94
	UGL	0.83	0.57
2 groups	SGL	0.33	0.45
	UGL	0.33	0.21
3 groups	SGL	0.12	0.14
	UGL	0.09	0.06
	$N = 150$, Levels $= 5$, $G = 20$		
1 group	SGL	0.86	0.78
	UGL	0.87	0.53
2 groups	SGL	0.32	0.33
	UGL	0.32	0.13
3 groups	SGL	0.10	0.04
	UGL	0.09	0.00
	$N = 100$, Levels $= 3$, $G = 30$		
1 group	SGL	0.92	0.87
	UGL	0.92	0.72
2 groups	SGL	0.28	0.32
	UGL	0.32	0.16
3 groups	SGL	0.05	0.08
	UGL	0.06	0.02

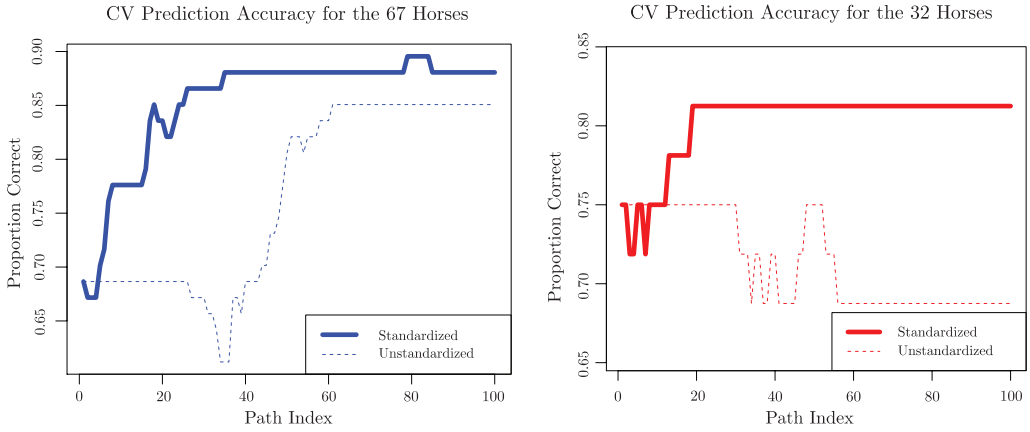

Figure 1. Plots of cross-validated prediction accuracy for regularization path in 32 and 67 horse subsets for standardized and unstandardized Group Lasso.

the fit with some regularization within group. In this case we propose using the objective

$$\min_{\beta} \frac{1}{2}\left\|y - \sum_{l=1}^{m} X^{(l)}\beta^{(l)}\right\|_2^2 + \lambda \sum_{l=1}^{m} \sqrt{df_l \left(\left\|X^{(l)}\beta^{(l)}\right\|_2^2 + \delta_l \left\|\beta^{(l)}\right\|_2^2\right)} + \sum_{l=1}^{m} \frac{\delta_l^2}{2} ||\beta^{(l)}||_2^2$$

where

$$df_l = \sum_{i=1}^{p_l} \frac{d_{l,i}^2}{d_{l,i}^2 + \delta_l},$$

for $d_{l,i}$ the ith singular value of $X^{(l)}$. This objective may seem unintuitive, but we can rewrite it in augmented form as

$$\min_{\beta} \frac{1}{2}\left\| \begin{pmatrix} y \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \left[\begin{pmatrix} X^{(1)} \\ \delta_1 I \\ 0 \\ \vdots \\ 0 \end{pmatrix} \beta^{(1)} + \begin{pmatrix} X^{(2)} \\ 0 \\ \delta_2 I \\ 0 \\ \vdots \end{pmatrix} \beta^{(2)} + \ldots + \begin{pmatrix} X^{(m)} \\ 0 \\ \vdots \\ 0 \\ \delta_m I \end{pmatrix} \beta^{(m)} \right] \right\|_2^2 \tag{7.1}$$

$$+ \lambda \left[\sqrt{df_1} \left\| \begin{matrix} X^{(1)} \\ \delta_1 I \\ 0 \\ \vdots \\ 0 \end{matrix} \right\|_2 + \sqrt{df_2} \left\| \begin{matrix} X^{(2)} \\ 0 \\ \delta_2 I \\ 0 \\ \vdots \end{matrix} \right\|_2 + \ldots + \sqrt{df_m} \left\| \begin{matrix} X^{(m)} \\ 0 \\ \vdots \\ 0 \\ \delta_m I \end{matrix} \right\|_2 \right]. \tag{7.2}$$

If we consider the optimality conditions for this model we get

$$\hat{\beta}^{(k)} = \left(1 - \frac{\sqrt{df_k}\lambda}{||\tilde{P}_{\mathrm{col}(X^{(k)})} r^{(k)}||_2}\right)_+ \left(X^{(k)\top} X^{(k)} + \delta_k I\right)^{-1} X^{(k)\top} r^{(-k)},$$

where $\tilde{P}_{\mathrm{col}(X^{(k)})}$ is no longer a projection, but instead

$$\tilde{P}_{\mathrm{col}(X^{(k)})} = X^{(k)} \left(X^{(k)\top} X^{(k)} + \delta_k I\right)^{-1} X^{(k)\top}.$$

This is just a shrunken ridge regression estimate. If $\delta_l > 0$, then we add a ridge penalty to group l, shrinking the covariance estimate of $X^{(l)}$ toward the identity for each l. df_l is defined to be the degrees of freedom of the ridge fit Hastie, Tibshirani, and Friedman (2008). With no prior knowledge of which groups should be in the model, one might consider choosing δ_l so that all groups have equal degrees of freedom. One should also note that if all $\delta_l > 0$, then this model is strictly convex and thus has a unique minimizer. We find this approach

attractive because it maintains the same "soft-thresholding the norm" framework as before, but extends to under-determined models.

8. Conclusion

We have shown that the standardized Group Lasso is the natural extension of the lasso to grouped data. We have proven its efficacy on real and simulated data. We have shown that it compares favorably to the unstandardized Group Lasso. In the case of high dimensionality within group we have extended the standardized Group Lasso to the ridged Group Lasso, and discuss a fast, straightforward algorithm to fit both standardized and ridged Group Lasso models. We will soon make available a public domain `R` package that implements these ideas.

References

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.

Frank, A. and Asuncion, A (2010). UCI machine learning repository. URL `http://archive.ics.uci.edu/ml`.

Friedman, J. H., Hastie, T. and Tibshirani, R. (2009). Applications of the lasso and grouped lasso to the estimation of sparse graphical models. Submitted.

Friedman, J. H., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Software* **33**, 1-22.

Hastie, T., Tibshirani, R. and Friedman, J. (2008). *The Elements of Statistical Learning; Data Mining, Inference and Prediction.* 2nd edition. Springer Verlag, New York.

Jacob, L., Obozinski, G. and Vert, J. (2009). Group Lasso with overlap and graph Lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 433-440. ACM.

Koltchinskii, V. and Yuan M. (2008). Sparse recovery in large ensembles of kernel machines. In *Conference on Learning Theory, COLT*, 229-238. Citeseer, 2008.

Koltchinskii, V. and Yuan M. (2010). Sparsity in multiple kernel learning. *Ann. Statist.*, **38**, 3660-3695.

Meier, L., van de Geer, S. and Bühlmann, P. (2008). The group lasso for logistic regression. *J. Roy. Statist. Soc. Ser. B* **70**, 53-71.

Meier, L., van de Geer, S. and Bühlmann, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37**, 3779-3821.

Nesterov, Y. (2007). *Gradient methods for minimizing composite objective function.* CORE.

Puig, A. T., Wiesel, A. and Hero, A. O. (2009). A multidimensional shrinkage-thresholding operator. In *SSP '09. IEEE/SP 15th Workshop on Statistical Signal Processing*, 113-116.

Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009). Sparse additive models. *J. Roy. Statist. Soc. Ser. B* **71**, 1009-1030.

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

Tseng, P. (2001). Convergence of block coordinate descent method for nondifferentiable maximization. *J. Optim. Theory Appl.*, **109**, 474-494.

Tseng, P. and Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Math. Program.* **117**, 387-423.

Yu, B., Wainwright, M. and Raskutti, G. (2010). Minimax-Optimal Rates for Sparse Additive Models over Kernel Classes via Convex Programming.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68**, 49-67.

Department of Statistics, Stanford University, Stanford CA 94305, USA.

E-mail: nsimon@stanford.edu

Department of Statistics and Department of Health Research and Policy, Stanford University, Stanford CA 94305, USA.

E-mail: rtibs@stanford.edu

(Received March 2011; accepted July 2011)