Electronic Journal of Statistics
Vol. 11 (2017) 4220–4257
ISSN: 1935-7524
DOI: 10.1214/17-EJS1317

Within group variable selection through the Exclusive Lasso

Frederick Campbell

Department of Statistics, Rice University

and

Genevera I. Allen

Department of Statistics, Rice University,
Department of Electrical and Computer Engineering, Rice University,
Department of Computer, Science Rice University,
Department of Pediatrics-Neurology, Baylor College of Medicine,
Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital
e-mail: gallen@rice.edu

Abstract: Many data sets consist of variables with an inherent group structure. The problem of group selection has been well studied, but in this paper, we seek to do the opposite: our goal is to select at least one variable from each group in the context of predictive regression modeling. This problem is NP-hard, but we propose the tightest convex relaxation: a composite penalty that is a combination of the ℓ_1 and ℓ_2 norms. Our so-called Exclusive Lasso method performs structured variable selection by ensuring that at least one variable is selected from each group. We study our method's statistical properties and develop computationally scalable algorithms for fitting the Exclusive Lasso. We study the effectiveness of our method via simulations as well as using NMR spectroscopy data. Here, we use the Exclusive Lasso to select the appropriate chemical shift from a dictionary of possible chemical shifts for each molecule in the biological sample.

Keywords and phrases: Structured variable selection, composite penalty, NMR spectroscopy, Exclusive Lasso.

Received April 2016.

1. Introduction

In regression problems with a predefined group structure, we seek to accurately predict the response using a small subset of variables consisting of at least one variable from each group. This structured sparsity assumption arises in a number of genomics and proteomics problems. Existing sparse regression methods, however, do not directly enforce the desired structure. To this end, we develop methodology for sparse regression with the Exclusive Lasso, a convex penalty first introduced by (Zhou et al., 2010) in the context of multi-task learning. Similar to the Group Lasso (Yuan and Lin, 2006), the Exclusive Lasso penalty is a composite penalty (Zhao et al., 2009) that uses both an ℓ_1 norm and an ℓ_2

norm. Loosely, the penalty performs selection within group by applying separate lasso penalties to each group. At the group level, the penalty is a ridge penalty preventing entire groups of coefficients from being set to zero. The group structure informs the type of regularization, utilizing the ℓ_1 and ℓ_2 norms within and between groups respectively. Throughout the literature, there are methods for structured sparsity with strong theoretical guarantees and fast algorithms (Obozinski and Bach, 2012; Halabi and Cevher, 2014; Genovese et al., 2012; Wainwright, 2009; Beck and Teboulle, 2009). The Exclusive Lasso however, has received little attention and has not yet been developed for applications outside of multi-task learning or carefully studied as a statistical method for sparse regression.

Consider a motivating example from genomics. Gene set analysis seeks to group genes based on genomic pathways and associate these gene sets with clinical outcomes. Commonly, the group lasso penalty has been used to select entire gene sets (Ma et al., 2007; Simon et al., 2013). Yet, one may also be interested in understanding how all the gene sets are related to a response and select the most representative gene from each gene set. Selecting one variable from each group cannot be easily achieved using existing techniques such as the Lasso or Marginal Regression (Tibshirani, 1996). If the Lasso's incoherence condition and β-min condition are satisfied and Marginal Regression's faithfulness assumption is satisfied, then both methods recover the correct variables without any knowledge of the group structure (Genovese et al., 2012; Wainwright, 2009). However, data rarely satisfies these assumptions. Consider that if two variables are correlated with each other, the Lasso often selects one instead of both variables. When whole groups are correlated, the Lasso may only select variables in one group as opposed to variables across multiple groups. Similarly, if the variables most correlated with the response are in the same group, Marginal Regression will ignore the true variables in other groups. In our example, genes are grouped together because they belong to the same pathway and hence are highly correlated. In these situations, the fact that the Lasso and Marginal regression are agnostic to the group structure hurts their ability to select a reasonable set of variables across all predefined groups. If we know that this group structure is inherent to our problem, then complex real world data motivate the need to develop new structured variable selection methods that directly select variables within each group. Although the Exclusive Lasso penalty can yield this sparsity structure (Obozinski and Bach, 2012), this penalty has not been developed statistically for sparse regression. Specifically, there are no algorithms in the literature to fit the Exclusive Lasso. Its statistical properties such as consistency, sparsistency, and degrees of freedom, are not well understood. We address these concerns and evaluate the effectiveness of the method using an example inspired by yet another real world problem arising in proteomics.

2. Exclusive Lasso

We study the Exclusive Lasso penalty in the context of penalized regression when there are predefined groups. Consider the linear model where the response

is a linear combination of the variables subject to Gaussian noise: $y = X\beta^* + \epsilon$ where ϵ is i.i.d Gaussian. For notational convenience, we assume the response is centered to eliminate an intercept term. We assume β^* is structured such that its indices are divided into non-overlapping predefined groups and that the support of β^* is distributed across all groups. We allow the support set within a group to be as small as one element and as large as the entire group. We can write this as two structural assumptions.

Assumption (1): There exists a collection of non-overlapping predefined groups denoted, $\mathcal{G}$, such that $\underset{g\in\mathcal{G}}{\cup} g = \{1, \ldots, p\}$, $\underset{g,g'}{\cap} = \emptyset$ for all pairs of groups $g, g' \in \mathcal{G}$.

Assumption (2): The support set S of the true parameter β^* is non-empty in each group such that for all $g \in \mathcal{G}$ we have $S \cap g \neq \emptyset$ and $\beta_i^* \neq 0$ for all $i \in S$.

Throughout the paper we study

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2}\|y - X\beta\|_2^2 + \frac{\lambda}{2}\sum_{g\in\mathcal{G}} \left(\sum_{i\in g} |\beta_i|\right)^2 \tag{1}$$

and its equivalent constrained optimization problem. Occasionally, we refer to the penalty as $P(\beta)$ where $P(\beta) = \frac{1}{2}\sum_{g\in\mathcal{G}} \left(\|\beta_g\|_1\right)^2$. Again, the optimization problem highlights that the Exclusive Lasso penalty is the ℓ_1-norm within groups and the ℓ_2-norm between. Figure 1 highlights the connection between the penalties.

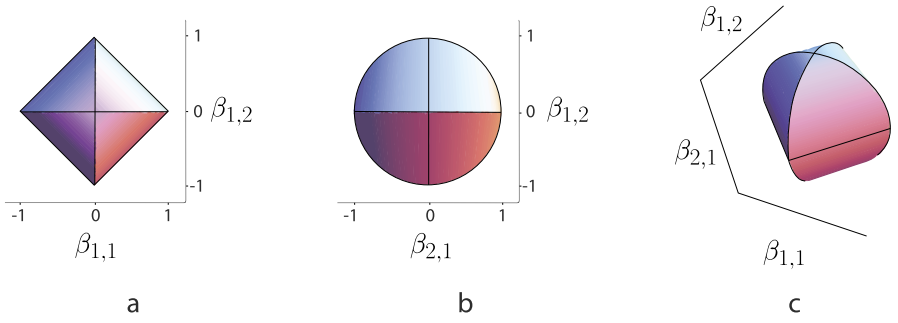

FIG 1. *The unit ball for the Exclusive Lasso penalty in three dimensions for* $\beta = (\beta_{11}, \beta_{12}, \beta_{21})$*, where the first two variables are in group one and the third variable is in group two. Figure (a) shows* β_{11} *vs* β_{12} *highlighting its connection to the* ℓ_1 *norm. Figure (b) shows* β_{21} *vs* β_{12} *highlighting its connection to the* ℓ_2 *norm. Figure (c) shows that the penalty inherits properties from both penalties.*

3. Optimality conditions

In order to understand the behavior of the Exclusive Lasso, we study its optimality conditions. Specifically, we use first order conditions to show that our

estimate behaves like an adaptively shrunken ridge problem that encourages exclusivity within groups. Throughout this section we adopt the notation for the subgradient of a convex function found in Rockafellar and Wets (2009). Recall the subdifferential of a function f at point x, denoted $\partial f(x)$, is the collection of all subgradients of f at x. We use the subdifferential to characterize the active set by deriving two expressions for the Exclusive Lasso estimate $\hat{\beta}$. Because problem (1) is convex, an optimal point satisfies $-X^T(y - X\,\hat{\beta}) + \lambda z = 0$ where z is an element of the subdifferential such that

$$z_i \in \partial P(\hat{\beta}) = \begin{cases} sign(\hat{\beta}_i) \| \, \hat{\beta}_g \, \|_1 & \text{if } \hat{\beta}_i \neq 0, i \in g \\ \left[- \| \, \hat{\beta}_g \, \|_1, \ \| \, \hat{\beta}_g \, \|_1 \right] & \text{if } \hat{\beta}_i = 0, i \in g. \end{cases} \tag{2}$$

Alternatively, we can express the sub gradient as the product of a matrix and a vector. If we let $M_g = sign(\hat{\beta}_{s \cap g}) sign(\hat{\beta}_{s \cap g})^T$ and let M_S be a block diagonal matrix with matrices M_g on the diagonal, then the sub gradient restricted to the support set $\hat{S}$ of $\hat{\beta}$ will be $z_{\hat{S}} = M_{\hat{S}}\, \hat{\beta}_{\hat{S}}$.

Note that the matrix $M_{\hat{S}}$ depends on the support set as the block diagonal matrices are defined by the nonzero elements of $\hat{\beta}$ in each group.

Proposition (1): Let $\hat{S}$ be the support set of $\hat{\beta}$, then

$$\hat{\beta}_{\hat{S}} = (X_{\hat{S}}^T X_{\hat{S}} + \lambda M_{\hat{S}})^{\dagger} X_{\hat{S}}^T y \text{ and } \hat{\beta}_{\hat{S}^c} = 0.$$

Here † denotes the Moore-Penrose pseudoinverse. The matrix $M_{\hat{S}}$ distinguishes the Exclusive Lasso from similar estimates like Ridge Regression. It is a block diagonal matrix that is only equivalent to the identity matrix when there is exactly one nonzero variable in each group. At this point, the Exclusive Lasso behaves like a Ridge Regression estimate on the nonzero indices that it has selected.

This characterization describes the behavior of the nonzero variables but it does not describe the behavior of the entire active set as we vary λ. To derive a second characterization of $\hat{\beta}$, we note that the optimality conditions imply that every nonzero variable in the same group has an equal correlation with the residual $X_i^T(y - X\,\hat{\beta})$. This allows us to determine when variables enter and exit the active set. There is always at least one nonzero variable in each group because the ℓ_2-norm at the group level ensures that the norm of each group $\| \, \hat{\beta}_g \, \|$ is always greater then 0. Another variable only enters the active set once its correlation with the residual is equal to the correlation shared by the other nonzero variables in the same group. We call the set $\mathcal{E} = \{i : |X_i^T(y - X\,\hat{\beta})| / \| \, \hat{\beta}_g \, \|_1 = \lambda\}$ the "group weighted equicorrelation set" because of its resemblance to the equicorrelation set described in Efron et al. (2004). We can use this set to derive an explicit formula for $\hat{\beta}$.

Proposition (2): If $\mathcal{E}$ is the group weighted equicorrelation set, i is in group g, γ' is a vector such that $\gamma_i' = \| \, \hat{\beta}_g \, \|_1 - | \, \hat{\beta}_i \, |$, $s \in \{-1, 1\}^{| \, \mathcal{E} \, |}$ is a vector of signs that satisfies the optimality conditions and $\mathcal{E}^c$ is the compliment of the set $\mathcal{E}$ then,

$$\hat{\beta}_{\mathcal{E}} = (X_{\mathcal{E}}^T X_{\mathcal{E}} + \lambda I)^{-1} [X_{\mathcal{E}}^T y - \lambda \gamma' s] \text{ and } \hat{\beta}_{\mathcal{E}^c} = 0. \tag{3}$$

The expression points to the general behavior of the penalty. For the non-zero indices, the first term is a ridge regression estimate $(X_{\mathcal{E}}^T X_{\mathcal{E}} + \lambda I)^{-1} X_{\mathcal{E}}^T y$. The second term $(X_{\mathcal{E}}^T X_{\mathcal{E}} + \lambda I)^{-1} \lambda \gamma' s$ adaptively shrinks the variables to zero competitively within each group. In the case where all groups have exactly one non-zero element, the Exclusive Lasso is a ridge regression estimate, ensuring that the Exclusive Lasso always selects at least one non-zero element in each group. This characterization also helps us see that for λ large enough, our method usually selects exactly one non-zero element in each group, but this is not guaranteed. See Appendix E for more details.

Before proceeding, we use a small simulated example to compare the behavior of the Lasso to the behavior of the Exclusive Lasso. We let $y = X\beta^* + \epsilon$ where $\epsilon \sim N(0, 1)$. The design matrix $X \in \mathbf{R}^{20\times 30}$ is multivariate normal with covariance that encourages correlation between groups and within groups. The incoherence condition is not satisfied with $\||X_{S^c}^T X_S (X_S^T X_S)^{-1}|\|_\infty = 2.603$. There are five groups and β^* is nonzero for one variable in each group. In Figure 2, we show the Exclusive Lasso and Lasso regularization paths for this example. In the figure, the solid lines are the truly nonzero variables and each color represents a different group. The Exclusive Lasso sends variables to zero until there is exactly one nonzero variable in each group whereas the Lasso eventually sends all variables to zero. Further, the Lasso does not enforce the group structure; the first five variables to enter the regularization path only represent three of the five groups. Because of this, the Lasso misses several true variables. The regularization path also highlights the Exclusive Lasso's connection to Ridge Regression; five variables will never go to zero. Overall, our work highlights that the Exclusive Lasso behaves like an adaptively shrunken ridge problem that forces competition or exclusivity within groups.

4. Statistical theory

We study prediction consistency and selection consistency for the Exclusive Lasso.

Prediction consistency

We focus on establishing prediction consistency under weak assumptions that are likely to be satisfied in practice. To this end, we begin by studying prediction consistency using the framework presented in Chatterjee (2013) for the Lasso. We make three assumptions and focus on bounding the mean squared prediction error.

Assumption (3): The training data $X \in \mathbb{R}^{n\times p}$ is generated by a probability distribution with covariance Σ, and the entries of X are bounded so that $|X_{ij}| \leq M$. The testing data $X_0 \in \mathbb{R}^p$ is an additional independent observation from the same distribution.

Assumption (4): The value of the penalty evaluated at the true parameter is bounded so that $P(\beta^*) \leq K$.

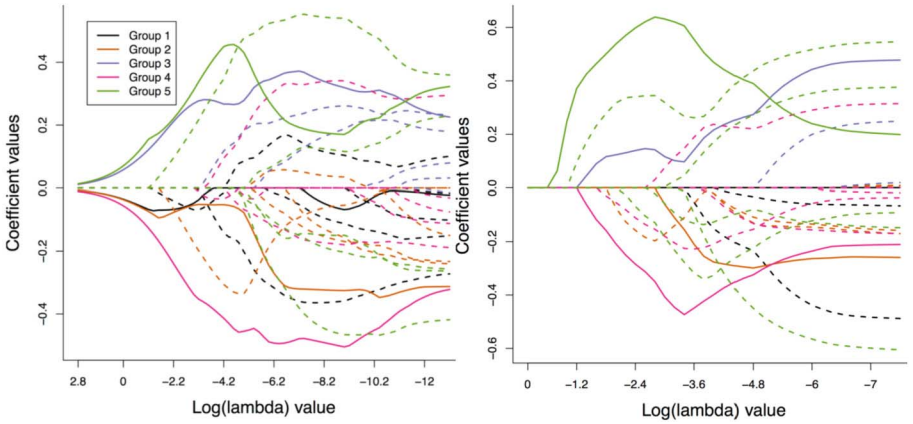

FIG 2. *The regularization paths for the Exclusive Lasso (left) and the Lasso (right) in a small simulation consisting of five groups with one true variable in each group. The Exclusive Lasso behaves like an adaptively regularized Ridge Regression estimate sending variables to zero until only one variable from each group is nonzero. The Lasso sends variables to zero without considering the group structure. Note that the first five variables to enter the model for the Lasso represent only groups 3, 4 and 5, where as the Exclusive Lasso has five variables, at least one from each group, that are in the model for all* λ

.

Assumption (5): The response is generated by the linear model $Y = X\beta^* + \epsilon$ where $\epsilon \overset{iid}{\sim} N(0, \sigma^2)$ and $Y_0 = X_0^T\beta^* + \epsilon_0$ such that $\epsilon_0 \overset{iid}{\sim} N(0, \sigma^2)$.

Definition (1): The mean squared prediction error is $\text{MSPE}(\hat{\beta}) = \mathbf{E}(Y_0^* - \hat{Y}_0)^2$ if $Y_0^* = X_0\beta^*$ and $\hat{Y}_0 = X_0\,\hat{\beta}$ where $\hat{\beta}$ is estimated using the training data X.

Definition (2): The estimated mean squared prediction error is $\text{eMSPE}(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i^* - \hat{Y}_i)^2$ if $Y_i^* = X_i\beta^*$ and $\hat{Y}_i = X_i\,\hat{\beta}$.

Let $|\mathcal{G}|$ be the number of groups; then using assumptions (3)–(5) we show that the Exclusive Lasso is prediction consistent.

Theorem (1): Under assumptions (3)–(5), the mean squared prediction error of $\hat{\beta}$ is bounded such that

$$\text{MSPE}(\hat{\beta}) \leq 2(K + |\mathcal{G}|)M\sigma\sqrt{\frac{2\log(2p)}{n}} + 8(K + |\mathcal{G}|)^2M^2\sqrt{\frac{2\log(2p^2)}{n}}. \quad (4)$$

We can also bound the estimated mean squared prediction error.

Theorem (2): Under assumptions (3)–(5) the estimated mean squared prediction error of $\hat{\beta}$ is bounded such that

$$\mathbf{E}[\text{eMSPE}(\hat{\beta})] \leq 2(K + |\mathcal{G}|)M\sigma\sqrt{\frac{2\log(2p)}{n}}. \quad (5)$$

Assumptions (3)–(5) are not difficult to satisfy in practice with real data. Many data sets will satisfy assumption (3). If we believe the data truly arises from a linear model then assumptions (4) and (5) will be satisfied as well. These assumptions are similar to those used to bound prediction consistency of the Lasso and are much easier to satisfy in practice than assumptions for other consistency results like sparsistency (Greenshtein et al., 2004; Chatterjee, 2013; Wainwright, 2009).

Theorem 1 shows that the Exclusive Lasso is consistent in terms of the norm $\|x\|_\Sigma$. The group structure in the penalty appears in the bound as the cardinality of the collection of groups. When there is only one group, this result reduces to a bound of essentially the same order $O(K^2\sqrt{\log(p)}/\sqrt{n})$ as the bound for the Lasso under similarly minimal assumptions (Bühlmann and Van De Geer, 2011). This suggests that we can allow n, p and the number of groups to scale together and still ensure that the estimate is prediction consistent. Additionally, if we do not consider M constant we can let the norm of the columns of X vary with n, p and $|\mathcal{G}|$ as well. We use this result to justify using the Exclusive Lasso, even in the high-dimensional setting, for prediction when a small number of variables are desired in each group.

Alternatively if the data has mean zero we can interpret the the MSPE using the covariance matrix Σ. In this setting, $\text{MSPE}(\hat{\beta}) = \mathbf{E}\| \hat{\beta} - \beta^*\|_\Sigma^2$ where $\|x\|_A = x^T A x$ and $\text{eMSPE}(\hat{\beta}) = \| \hat{\beta} - \beta^*\|_{\hat{\Sigma}}^2$ where $\hat{\Sigma} = X^T X/n$.

If we add another assumption on the eigenvalues of Σ we can show that the Exclusive Lasso estimate is consistent using the ℓ_2 norm.

Corollary (1): For centered data X, if the smallest eigenvalue of the covariance matrix Σ is bounded below by $c > 0$ then the Exclusive Lasso estimate is consistent in the ℓ_2-norm:

$$\mathbf{E}\| \hat{\beta} - \beta^*\|_2^2 \leq \frac{2}{c}(K + |\mathcal{G}|)M\sigma\sqrt{\frac{2\log(2p)}{n}} + \frac{8}{c}(K + |\mathcal{G}|)^2 M^2 \sqrt{\frac{2\log(2p^2)}{n}}. \quad (6)$$

Our additional assumption requires the covariance matrix to be strictly positive definite which is much more restrictive then our previous assumptions on Σ. Tighter bounds are likely attainable using more restrictive assumptions on X. For example, an assumption related to the subspace compatibility constant or the restricted eigenvalue condition will likely yield improved bounds, but would also severely limit the correlation structure permitted in the data.

Selection consistency

Next, we investigate under which conditions the Exclusive Lasso can estimate the true support with high probability. Our analysis shares important similarities to selection consistency, or "sparsistency", results for the Lasso, but there are key differences that arise as a result of our penalty and our group-wise sparsity assumptions.

Assumption (6): Assume that the columns of X are standardized so that $\|X_j\|_1 = 1$.

Assumption (7): Let $d_{min} \leq \cdots \leq d_{max}$ be the eigenvalues of $X_S^T X_S$ and assume that $0 < C_{min}^2 \leq d_{min} \leq d_{max} \leq C_{max}^2 < \infty$ where C_{min} and C_{max} are constants.

Assumption (8): Assume $\|X_{S^c}^T X_S (X_S^T X_S + \lambda)^{-1}\|_{\infty,\infty} < \alpha$ for $\alpha = \frac{k C_{max}^2 \lambda}{C_{max}^2 + \lambda}$ where λ is the regularization parameter and k is the number of groups.

Assumption (9): Assume $\|\beta^*\|_1 > \lambda L \sqrt{2log(n)}$ where L is a constant that depends on C_{max}, C_{min} and λ.

Assumption (10): Assume there is exactly one non-zero variable in each group.

Theorem (3): If there exists a regularization parameter $\lambda > 0$ that yields an Exclusive Lasso estimate with exactly one non-zero variable per group and X, y and λ satisfy assumptions (1), (2) and (6)–(10), then the support of the Exclusive Lasso estimate S is equal to the true support S^* so that $S = S^*$ with probability greater then or equal to $1 - 1/n$.

See Appendix B for the proof of Theorem 3.

Our result shows the conditions under which the Exclusive Lasso is model selection consistent when selecting one true variable per group. We leave signed selection consistency to later work. Notice that several of our assumptions and conditions are similar to those found in other sparsistency results but differ slightly due to the structure of our penalty. First, Assumption (6) and Assumption (9) resemble common standardization and beta-min assumptions respectively. Notice that where the standardization condition appears to be a cosmetic difference, the beta-min condition is weaker than the condition presented in Wainwright (2009). Our condition requires the aggregate signal strength to be above a certain threshold instead of each variable individually. Next, Assumption (7) is a standard bound on the size of the eigenvalues ensuring the solution is unique. Finally, Assumption (8) is similar to the irrepresentable condition of Zhao and Yu (2006) which states that $\|X_{S^c}^T X_S (X_S^T X_S)^{-1}\|_{\infty,\infty} < 1$, but differs in two ways. First for the Lasso, the irrepresentable condition must be strictly less than one; for the Exclusive Lasso, this must be less than α. Notice that α can be much greater than one, especially for large values of λ that are typical for ensuring only one variable is selected per group. Second, notice that the irrepresentable condition differs from Assumption (8) in that the empirical covariance between the true support is shrunken towards λI, thus further decreasing the value of our irrepresentable condition-like term. Putting these two together, we see that the Exclusive Lasso will be selection consistent in situations where there is much stronger correlation both within the true support and between the non-support and support than those for which the Lasso is selection consistent. This theoretical finding is consistent with our simulation studies which indicate that the Exclusive Lasso outperforms the Lasso in terms of model selection when there is strong correlation within or between groups. Overall, Theorem (3) gives

fairly weak conditions under which the Exclusive Lasso is sparsistent. Future research will extended these results to consider cases with more than one variable per group and characterize the conditions under which it is possible to find a λ that yields exactly one variable per group. (See Appendix E for additional discussion of the later).

5. Estimation

There are currently no fast algorithms developed specifically to fit the Exclusive Lasso regression problem. Because of the composite nature of our penalty, standard sparse algorithms such as coordinate descent, proximal gradient descent, and alternating direction method of multipliers (ADMM) cannot be applied to fit our method in a straightforward manner without further investigation. However, careful investigation shows that we can develop a coordinate descent algorithm even though our penalty violates the separability assumption typically necessary for proving convergence. This leads to a coordinate descent algorithm for fitting the Exclusive Lasso problem as well as an algorithm to compute the proximal operator that allows us to develop additional first-order algorithms like ADMM and proximal gradient descent. In this section, we present our coordinate descent algorithm for fitting the Exclusive Lasso problem. A coordinate descent scheme to compute the proximal operator for the Exclusive Lasso is given in Appendix H where we also develop a proximal gradient algorithm using this proximal operator as an example.

The Exclusive Lasso penalty is not a separable function meaning it cannot be decomposed into functions of each individual variable, $P(\beta) \neq \sum_{j=1}^{p} P_j(\beta_j)$. This makes computing closed-form updates of all variables at once impossible (see the subgradient equations in 2). Coordinate descent however, is a natural algorithm to fit the Exclusive Lasso because it iteratively updates one variable at a time while holding all others fixed. We present our coordinate descent scheme to fit the Exclusive Lasso in Algorithm 1. Here, $S(x, \lambda) = sign(x)(|x| - \lambda)_+$ is the usual soft-thresholding function. Note that in the algorithm, we let g be the group for index j and we use $g \setminus j$ to denote the set of indices in group g without index j.

Input: $\beta^0 \in \mathbf{R}^p, \epsilon > 0$
Output: $\hat{\beta} \in \mathbf{R}^p$
Pre-compute $X^T y$ and $X^T X$.
while $\|\beta^{k+1} - \beta^k\| > \epsilon$ **do**
  for $j \in 1$ ***to*** p **do**
    $\tilde{z} = X_j^T (y - \sum_{l \neq j} X_l \beta_l^k)$
    $\tilde{\lambda} = \lambda \sum_{l \in g \setminus j} |\beta_l^k|$
    $\beta_j^{k+1} = S\left(\frac{\tilde{z}}{X_j^T X_j + \lambda}, \frac{\tilde{\lambda}}{(X_j^T X_j + \lambda)} \right)$
return β

Algorithm 1: Exclusive Lasso Coordinate Descent Algorithm

The Exclusive Lasso coordinate descent updates, resemble the coordinate descent updates for the Group Lasso presented by Yuan and Lin (2006) and that of the Lasso (Wu and Lange, 2008). The correlation between the j^{th} variable, X_j, and the current residual is shrunken using the soft thresholding operator, which is similar to the Lasso updates. And, like the Group Lasso, other variables in the same group affect the amount of shrinkage. Overall the algorithm is not significantly more complicated or computationally intensive than the analogous coordinate descent algorithm for the Lasso, despite the added complexity of the Exclusive Lasso penalty. Additionally, our experience suggests that the coordinate descent algorithm is computationally fast when there is exactly one nonzero variable in each group.

As we have mentioned, the Exclusive Lasso penalty is non-separable which means that we cannot invoke standard convergence guarantees for coordinate descent schemes (Tseng, 2001) without additional investigation. Nevertheless, we prove that the Exclusive Lasso problem enjoys certain regularity conditions that we can use to ensure that our coordinate descent algorithm converges to the global minimum:

Theorem (4): The Exclusive Lasso coordinate descent algorithm converges to the global minimum of (1).

See Appendix C for the proof of Theorem (4). Also, see Appendix H where we develop a coordinate descent scheme to solve the Exclusive Lasso proximal operator and use this in a proximal gradient descent method.

6. Model selection

In practice, we need a data-driven method to select the regularization parameter and regulate the amount of sparsity within each group. To this end, we provide an estimate of the degrees of freedom that will allow us to use the Bayesian information criteria and the extended Bayesian information criteria for model selection (Schwarz et al., 1978; Chen and Chen, 2008). While other general model selection procedures like cross validation and stability selection can be employed, these tend to over-select variables for the Exclusive Lasso; see details in Appendix F.

We leverage techniques used by Stein (1981) and Tibshirani et al. (2012) to calculate the degrees of freedom and provide an unbiased estimate of the degrees of freedom. Previously, we defined the matrix $M_{\hat{S}}$ as a block diagonal matrix where each nonzero block M_g is the outer product of the sign vector of the estimate, $M_g = sign(\hat{\beta}_{\hat{S}\cap g})sign(\hat{\beta}_{\hat{S}\cap g})^T$. This leads to our statement of the degrees of freedom for $\hat{y}$:

Theorem (5): For any design matrix X and regularization parameter $\lambda \geq 0$, if y is normally distributed, then the degrees of freedom for $\hat{y} = X\hat{\beta}$ is

$$\nu(\hat{y}) = \mathbf{E}\left[trace(X_{\hat{S}}(X_{\hat{S}}^T X_{\hat{S}} + \lambda M_{\hat{S}})^{\dagger} X_{\hat{S}}^T)\right].$$

An unbiased estimate of the degrees of freedom is then

$$\hat{\nu}(\hat{y}) = trace[X_{\hat{S}}(X_{\hat{S}}^T X_{\hat{S}} + \lambda M_{\hat{S}})^{\dagger} X_{\hat{S}}^T]. \tag{7}$$

To verify this result, we compare our unbiased estimate of the degrees of freedom to simulated degrees of freedom following the set up outlined in Efron et al. (2004) and Zou et al. (2007). Empirically, our unbiased estimate of the degrees of freedom closely matches the simulated degrees of freedom Figure 3.

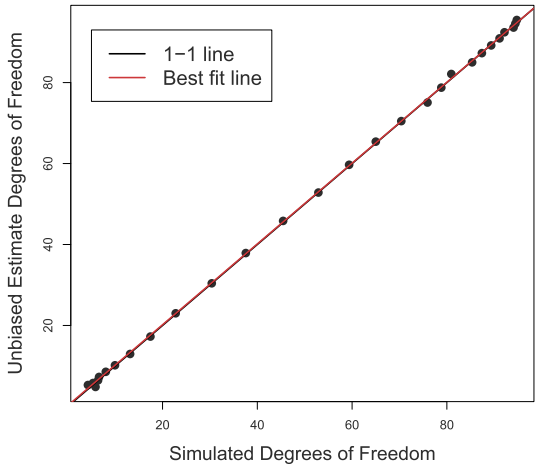

FIG 3. *Comparison of our estimate for the degrees of freedom to the simulated degrees of freedom. The simulated degrees of freedom matches the estimated degrees of freedom very closely.*

Our degrees of freedom estimate allows one to select λ for the Exclusive Lasso using the Bayesian information criteria or the extended Bayesian information criteria in the high dimensional setting. If we want exactly one variable per group, then usually we can select λ sufficiently large to achieve this. In other cases, we suggest using the Bayesian information criteria or the extended Bayesian information criteria to select λ and then threshold the estimate within each group.

7. Simulation study

We study the empirical performance of our Exclusive Lasso through two sets of simulation studies: first, for selecting one variable per group and second, for selecting a small number of variables per group. We examine three situations with moderate to large amounts of correlation between groups and within groups. We omit the low correlation setting from the simulations because they correspond to design matrices that are nearly orthogonal, satisfying both the Incoherence condition and the Faithfulness condition in which all methods perform well.

This is not representative of the types of real data for which we would need to use the Exclusive Lasso. Situations requiring the Exclusive Lasso will have large correlations based on group structure.

In the first simulations, we simulate data using the model $y = X\beta^* + \epsilon$ where $\epsilon \overset{iid}{\sim} N(0,1)$ and β^* is the true parameter. The variables are divided into five equal sized groups and the true parameter is nonzero at one index in each group and zero otherwise. We use three design matrices each with $n = 100$ observations and $p = 100$ variables, to test the robustness of the Exclusive Lasso to within group correlation and between group correlation. All three matrices are drawn from a multivariate normal distribution with a Toeplitz covariance matrix with entries $\Sigma_{ij} = w^{|i-j|}$ for variables in the same group, and $\Sigma_{ij} = b^{|i-j|}$ for variables in different groups. The first covariance matrix uses constant $b = .9$ and $w = .9$ to simulate high correlation within groups and high correlation between groups. The second covariance matrix uses $b = .6$ and $w = .9$ so that the correlation between groups is lower then the correlation within groups, resulting in high correlation within group and medium correlation between groups. The third covariance matrix uses constants $w = .6$ and $b = .6$ so that there is medium correlation both between group and within group.

We compare two versions of our Exclusive Lasso as described in the previous section. First, we use a regularization parameter λ, large enough to ensure that the method selects exactly one element per group. In these simulations, $\lambda = \max_i |X_i^T y|$ was large enough to ensure the correct structure was estimated; we refer to this as the Exclusive Lasso. The second estimate, the Thresholded Exclusive Lasso, chooses the regularization parameter λ that minimizes the BIC and then thresholds in each group keeping the index with the largest magnitude. We also compare our method to competitors and logical extensions of competitors in the literature. We base three comparison methods on the Lasso: First, we take the largest regularization parameter that yields exactly five nonzero coefficients (Lasso); second, we take the largest λ that has nonzero indices in each group and then threshold group-wise to keep the coefficient in each group with the largest magnitude (Thresholded Lasso); third, we take the first coefficient along the Lasso regularization path to enter the active set from each group (Thresholded Regularization Path). We use Marginal Regression where we take the five indices that maximize $|X_i^T y|$ (Marginal Regression) and we take the one coefficient in each group that maximizes $|X_i^T y|$ for $i \in g$ (Group-wise Marginal Regression). Finally, we use Elastic Net (Elastic Net) and Ridge Regression with regularization parameters that minimize their respective BICs. We threshold Ridge Regression so that the correct number of variables are estimated (Thresholded Ridge Regression). For all methods we select a set of variables $\hat{S}$, and then use the data matrix restricted to this set $X_{\hat{S}}$ to calculate an Ordinary Least Square estimate $\hat{\beta}_{\hat{S}}$. The prediction error is calculated using $\hat{\beta}_{\hat{S}}$. Results in terms of prediction error and variable selection recovery are given in Table 1.

The Exclusive Lasso outperforms all other methods at all levels of correlation, likely because it selects more variables that are truly nonzero. We observe that the thresholded estimators generally perform better then the non thresh-

TABLE 1

Comparison of variable selection methods with exactly one true variable in each group.

Within group correlation=.9, Between group correlation =.9

	Exclusive Lasso	Lasso	Marginal Regression	Group-wise Marginal Regression	Thresholded Exclusive Lasso
True Vars (SE)	**4.40 (0.67)**	2.54 (1.01)	1.06 (0.31)	2.98 (0.89)	4.28 (0.86)
False Vars (SE)	**0.60 (0.67)**	2.46 (1.01)	3.94 (0.31)	2.02 (0.89)	0.72 (0.86)
Pred Err (SE)	**1.08 (0.08)**	1.36 (0.11)	1.68 (0.10)	1.19 (0.12)	1.08 (0.11)
	Thresholded Lasso	Thresholded Regularization Path	Elastic Net	Thresholded Ridge	
True Vars (SE)	3.82 (1.00)	3.40 (0.95)	1.16 (0.42)	3.20 (0.88)	
False Vars (SE)	1.18 (1.00)	1.60 (0.95)	3.84 (0.42)	1.80 (0.88)	
Pred Err (SE)	1.12 (0.11)	1.14 (0.11)	1.67 (0.11)	1.34 (0.17)	

Within group correlation=.9, Between group correlation =.6

	Exclusive Lasso	Lasso	Marginal Regression	Group-wise Marginal Regression	Thresholded Exclusive Lasso
True Vars (SE)	**4.86 (0.45)**	3.24 (0.92)	1.80 (0.67)	3.32 (0.62)	4.30 (0.81)
False Vars (SE)	**0.14 (0.45)**	1.76 (0.92)	3.20 (0.67)	1.68 (0.62)	0.70 (0.81)
Pred Err (SE)	**1.07 (0.07)**	1.28 (0.17)	1.62 (0.16)	1.16 (0.09)	1.09 (0.08)
	Thresholded Lasso	Thresholded Regularization Path	Elastic Net	Thresholded Ridge	
True Vars (SE)	4.30 (0.74)	3.66 (0.77)	1.92 (0.70)	3.30 (0.91)	
False Vars (SE)	0.70 (0.74)	1.34 (0.77)	3.08 (0.70)	1.70 (0.91)	
Pred Err (SE)	1.07 (0.07)	1.12 (0.09)	1.60 (0.16)	1.33 (0.18)	

Within group correlation=.6, Between group correlation =.6

	Exclusive Lasso	Lasso	Marginal Regression	Group-wise Marginal Regression	Thresholded Exclusive Lasso
True Vars (SE)	**5.00 (0.00)**	4.40 (0.61)	3.52 (0.50)	4.08 (0.60)	4.80 (0.83)
False Vars (SE)	**0.00 (0.00)**	0.60 (0.61)	1.48 (0.50)	0.92 (0.60)	0.20 (0.83)
Pred Err (SE)	**1.04 (0.08)**	1.16 (0.16)	1.61 (0.15)	1.17 (0.13)	1.06 (0.17)
	Thresholded Lasso	Thresholded Regularization Path	Elastic Net	Thresholded Ridge	
True Vars (SE)	4.60 (0.78)	4.68 (0.78)	3.62 (0.53)	4.36 (0.78)	
False Vars (SE)	0.40 (0.78)	0.32 (0.78)	1.38 (0.53)	0.64 (0.78)	
Pred Err (SE)	1.10 (0.16)	1.08 (0.15)	1.32 (0.17)	1.15 (0.17)	

olded estimators. These simulations highlight the Exclusive Lasso's robustness to moderate and large amounts of correlation, which is important considering we expect variables in the same group to be similar and possibly highly correlated with each other.

In the second set of simulations, we also simulate data using the model $y = X\beta^* + \epsilon$ where $\epsilon \sim N(0, 1)$ and β^* is the true parameter for $n = p = 100$. In these simulations, the variables are divided into the same five equal-sized groups but the true parameter can be nonzero at more than one index in each group. Specifically, there are seven nonzero coefficients distributed so that three groups have exactly one nonzero index and two groups have two nonzero indices each. We simulate the design matrices in the same way we simulate design matrices in the first set of simulations to have varying levels of between and within group correlation.

We compare five methods: the Exclusive Lasso, the Lasso, the Lasso applied independently to each group, the Elastic Net, and Thresholded Ridge Regression. For the Lasso and Exclusive Lasso we use EBIC to select the regularization parameters. For the Elastic Net, Group-wise Lasso, and Thresholded Ridge Regression, the EBIC resulted in solutions that were often entirely zero so we used the BIC to select the regularization parameters. We used the oracle number of variables to select additional parameters such as the threshold value for Thresholded Ridge Regression. When we apply the Lasso separately to each group, we use separate regularization parameters chosen independently with BIC.

Results in terms of prediction error and variable selection are presented in Table 2. The Exclusive Lasso performs the best at support recovery and performs comparably in terms of prediction error. We attribute the Exclusive Lasso's success to its ability to find estimates with the correct sparsity structure and its ability to tolerate relatively high levels of correlation between variables. In each simulation study presented, we are violating the irrepresentable condition needed to guarantee that the Lasso selects the true support. We also note that although Thresholded Ridge Regression performs well, especially when we tune it to select the oracle number of variables, its performance falls short of the Exclusive Lasso because it does not select variables that respect the group structure. Overall, our results indicate that the Exclusive Lasso outperforms competing methods for selecting one or a small number of variables per group when there is strong correlation between and within groups. Additional simulation studies can be found in Appendix G.

8. NMR spectroscopy study

Finally, we illustrate an application of the Exclusive Lasso for selecting the chemical shift of molecules in Nuclear Magnetic Resonance (NMR) spectroscopy. NMR spectroscopy is a high-throughput technology used to study the complete metabolic profile of a biological sample by measuring a molecule's interaction with an external magnetic field (De Graaf, 2013; Cavanagh et al., 1995). This technology produces a spectrum where the chemical components of each

TABLE 2
Comparison of variable selection methods with multiple true variables in a group.

Within group correlation=.9, Between group correlation =.9

	Exclusive Lasso	Lasso	Group-wise Lasso	Elastic Net	Thresholded Ridge
True Vars (SE)	**6.70 (0.46)**	6.10 (0.93)	3.04 (0.70)	2.96 (0.40)	4.48 (0.58)
False Vars (SE)	**0.30 (0.46)**	6.66 (2.33)	8.04 (2.65)	12.08 (0.90)	2.52 (0.58)
Pred Err (SE)	1.19 (0.10)	**1.18 (0.46)**	1.74 (0.13)	1.88 (0.09)	1.46 (0.19)

Within group correlation=.9, Between group correlation =.6

	Exclusive Lasso	Lasso	Group-wise Lasso	Elastic Net	Thresholded Ridge
True Vars (SE)	**6.80 (0.40)**	6.44 (0.81)	4.54 (0.68)	3.86 (0.40)	4.54 (0.81)
False Vars (SE)	**0.20 (0.40)**	5.28 (2.08)	5.72 (1.44)	9.76 (1.22)	2.46 (0.81)
Pred Err (SE)	1.20 (0.11)	**1.11 (0.11)**	1.42 (0.17)	1.78 (0.11)	1.50 (0.21)

Within group correlation=.6, Between group correlation =.6

	Exclusive Lasso	Lasso	Group-wise Lasso	Elastic Net	Thresholded Ridge
True Vars (SE)	**7.00 (0.00)**	7.00 (0.00)	3.98 (0.25)	3.00 (0.00)	5.50 (0.76)
False Vars (SE)	**0.00 (0.00)**	6.72 (17.65)	5.12 (1.89)	3.08 (0.67)	1.50 (0.76)
Pred Err (SE)	**1.31 (0.13)**	1.34 (1.31)	1.68 (0.10)	1.87 (0.09)	1.40 (0.22)

molecule resonate at a particular ppm. Ideally, NMR spectroscopy would allow us to identify and quantify the concentrations of all molecules in a given biological sample, however this is challenging for numerous reasons discussed in (Ebbels et al., 2011; Weljie et al., 2006; Zhang et al., 2009). Our work accurately quantifies the relative concentrations of known molecules in a sample by accounting for positional uncertainty inherent to NMR spectroscopy data. This positional uncertainty is known as a "chemical shift" and is the phenomena where every molecules' chemical signature is subject to a random translation in ppm due to the external physical environment of the sample (De Graaf, 2013). We model the chemical shifts by creating an expanded dictionary of shifted molecules where we consider each molecule and its associated shifts as a group, allowing us to use the Exclusive Lasso to identify the best shift of each molecule.

We create an expanded dictionary using reference measurements for thirty-three unique molecules. The dictionary, $X \in \mathbf{R}_{+}^{4000\times(33*11)}$, consists of spectra for thirty-three molecules and ten artificial positional shifts for each molecule, five left and five right. These shifts are no more then .05ppm greater than or less than the reference measurement yielding eleven possible positions for each molecule. We use one randomly selected shift for each molecule, hence simulating the positional uncertainty found in real data. The columns of this expanded dictionary are strongly correlated with each other. Molecules are correlated with their ten shifts as well as other molecules with similar chemical structures. If we consider each molecule and its shifts a group, this results in a data set that has high correlation between groups as well as high correlation within each group.

We simulate an NMR spectroscopy signal y by simulating a random shift and concentrations that are chosen to recreate the crowding that is common in NMR data. Often, molecules will resonate at similar frequencies, causing peaks to overlap (De Graaf, 2013). Informally, this yields signals that appear smoother with less pronounced peaks because of the crowding. We use weights that recreate this effect in the region between .5 and 0 ppm. We then simulate our signal using positive noise so that $y = X_{S^*}\beta^* + \epsilon$ where ϵ is the absolute value of Gaussian noise thereby ensuring y is non-negative. S^* is the true set of shifts.

If all molecules and their shifts are known, the observed NMR signal will be a linear combination of the molecules. In general, we must estimate the shifts and we evaluate the effectiveness of our estimates through the mean squared error $\|\beta^* - \hat{\beta}_{\hat{S}}\|_2^2$ where $\hat{\beta}_{\hat{S}}$ is a least squares estimator based on the dictionary restricted to $\hat{S}$ an estimated set of shifts. This measures how accurately we can estimate the concentrations and is more useful than prediction error for a researcher analyzing NMR spectroscopy data. We compare four methods. First the Exclusive Lasso with $\lambda = \max_i X_i^T y$. Second, the Lasso with λ large enough to estimate one variable in each group. For both methods, if there are more than one variable in each group we threshold so that there is exactly one variable in each group. Next, we use Marginal Regression which selects the variable in each group with the highest correlation with the response. Lastly, we use an estimator that does not estimate chemical shifts. We refer to this method as the OLS estimator.

Among all methods, the Exclusive Lasso performs best at quantifying molecule concentrations under positional uncertainty. This case study highlights a real example where there is high correlation both within and between pre-defined groups. Consistent with our simulation studies, the Exclusive Lasso performs best in these situations.

TABLE 3
Comparison of variable selection methods for NMR spectroscopy.

	Mean Squared Error (SE)	Prediction Error(SE)
Exclusive Lasso	**1.07(.03)**	**1.34e-04(9.80e-07)**
OLS regression	2.87(.06)	2.61e-04(1.16e-06)
Marginal Regression	1.16(.23)	1.45e-04(1.84e-05)
Lasso	2.09(.14)	8.03e-05(1.09e-05)

9. Discussion

In this work, we focus on statistical questions important to the practitioner, but there are several directions for future work. Investigating overlapping or hierarchical group structures, and inference are important open questions. Additionally, extending the selection consistency results presented in this work to signed selection consistency would highlight connections or differences with the Lasso. One could also use the Exclusive Lasso penalty with other loss functions

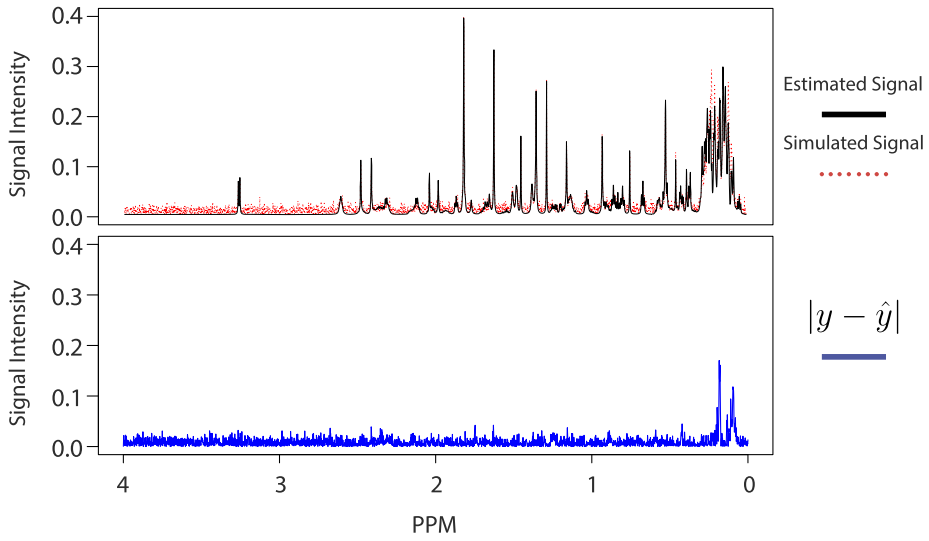

FIG 4. *A comparison of the simulated NMR signal and the signal estimated using the Exclusive Lasso. The estimate recovers most of the peaks suggesting it is selecting a useful set of shifts. The estimate also zeros out most of the noise in the simulated signal.*

such as that of generalized linear models. Additionally, there are many possible applications of our method besides NMR spectroscopy such as creating index funds in finance, and selecting genes from functional groups or pathways, among others. Overall, the Exclusive Lasso is an effective method for within group variable selection in sparse regression; an R-package will be made available for others to utilize our method.

Appendix A: Proof of Theorems 1 and 2

The proof of Theorems 1 and 2 follows the proof technique presented in Chatterjee (2013). There are several differences due to the structure of our penalty however the assumptions are the same. We assume that X_i is a p dimensional random variable with covariance Σ. We assume the entries of X_i are bounded so that $\|X_i\|_\infty \leq M$ and that the data we observe $(Y_1, X_1) \dots (Y_n, X_n)$ is independent and identically distributed. Let (Y, X) be the vector and matrix of all n observations. We let X_0 be an independent observation from the same distribution of X. We also assume the value of the penalty evaluated at the true parameter is bounded so that $P(\beta^*) \leq K$ and that the response is generated by the linear model $Y = X\beta^* + \epsilon$ where $\epsilon \sim N(0, \sigma^2 I)$. Let $Y_0 = X_0\beta^* + \epsilon_0$ where $\epsilon_0 \sim N(0, \sigma^2)$. Let $\mathcal{G}$ be a collection of predefined non overlapping groups such that $\bigcup_{g \in \mathcal{G}} g = \{1 \dots p\}$.

Instead of the Exclusive Lasso penalty, we work with the equivalent constrained optimization problem

$$\hat{\beta} = \underset{\beta:P(\beta)\leq K}{\operatorname{argmin}} \sum_{i=1}^{n}(Y_i - X_i\beta)^2.$$

Let $C = \{X\beta : P(\beta) \leq K\}$. By definition, $\hat{Y}$ is the projection of Y onto the set C. For constrained optimization problems first order necessary conditions for an optimal solution state that for all d in the linear tangent cone a solution to the problem x^* necessarily satisfies $f'(x^*; d) \geq 0$. In our case the linear tangent cone is the set $T_\ell(\hat{Y}) = \{(x - \hat{Y}) : x \in C\}$ so an optimal solution satisfies $\langle -(Y - \hat{Y}), (x - \hat{Y})\rangle \geq 0$ for all $x \in C$. This follows because $\hat{Y}$ is the solution to an optimization problem and the inequality is a necessary property of any solution to the problem. See Theorem 6.12 in Rockafellar and Wets (2009). Letting $x = Y^*$ we can rewrite $\langle (Y - \hat{Y}), (Y^* - \hat{Y})\rangle \leq 0$ as the inequality

$$\begin{aligned}\|Y^* - \hat{Y}\|_2^2 &\leq \langle (Y - Y^*), (\hat{Y} - Y^*)\rangle \\ &= \sum_{i=1}^{n}\epsilon_i\left(\sum_{j=1}^{p}(\hat{\beta}_j - \beta_j^*)X_{i,j}\right) \\ &= \sum_{j=1}^{p}(\hat{\beta}_j - \beta_j^*)\left(\sum_{i=1}^{n}\epsilon_i X_{i,j}\right).\end{aligned}$$

At this point our bound holds for any convex constraint region because it relies on the optimality conditions for convex nonlinear optimization problems. The bound also closely resembles the basic inequality for the Lasso because at this point, we have not used any information about the penalty. The next bound is where our bound begins to differ from the bound for the Lasso. We use our assumption $P(\beta^*) \leq K$ the definition of $\hat{\beta}$ and the structure of the penalty to bound $\sum_{j=1}^{p}(\hat{\beta}_j - \beta_j^*)$ so that

$$\begin{aligned}\sum_{j=1}^{p}(\hat{\beta}_j - \beta_j^*) &\leq \sum_{j=1}^{p}|\hat{\beta}_j - \beta_j^*| \\ &\leq \sum_{j=1}^{p}|\hat{\beta}_j| + |\beta_j^*| \\ &\leq 2(K + |\mathcal{G}|).\end{aligned}$$

In the last line, if the norm of the group $\|\hat{\beta}_g\|_1$ is greater then 1 then it is bounded by $\|\hat{\beta}_g\|_1^2$ otherwise it is bounded by 1.

This implies that if we let $U_j = \sum_{i=1}^{n}\epsilon_i X_{i,j}$ then

$$\|Y^* - \hat{Y}\|^2 \leq 2(K + |\mathcal{G}|)\max_{1\leq j\leq p}|U_j|.$$

Because $U_j \sim N\left(0, \sigma^2 \sum_{i=1}^{n} X_{i,j}^2\right)$ we have the bound

$$\mathbf{E}(\max_{1\leq j\leq p} |U_j|) \leq M\sigma\sqrt{2n\log(2p)}.$$

See Lemma 3 in Chatterjee (2013) for proof of the bound. Therefore

$$\mathbf{E}\|Y^* - \hat{Y}\|^2 \leq 2(K + |\mathcal{G}|)M\sigma\sqrt{2n\log(2p)},$$

and we have Theorem 2:

$$\mathbf{E}[eMSPE(\hat{\beta})] \leq 2(K + |\mathcal{G}|)M\sigma\sqrt{\frac{2\log(2p)}{n}}.$$

We use this result to prove Theorem 1. Because X_0 is independent of X, the pair (X_0, Y_0) is independent of $\hat{\beta}$ which was fit with the training data X. This yields

$$\mathbf{E}[(Y_0^* - \hat{Y}_0)^2|X] = \sum_{j,k=1}^{p} (\beta_j^* - \hat{\beta}_j)(\beta_k^* - \hat{\beta}_k)\mathbf{E}(X_{0j}X_{0k}).$$

Note that

$$\frac{1}{n}\|Y^* - \hat{Y}\|^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j,k=1}^{p} (\beta_j^* - \hat{\beta}_j)(\beta_k^* - \hat{\beta}_k)X_{ij}X_{ik}.$$

Combining these two expressions yields

$$\begin{aligned}\mathbf{E}[(Y_0^* - \hat{Y}_0)^2|X] - \frac{1}{n}\|Y^* - \hat{Y}\|^2 \\ = \sum_{j,k=1}^{p} (\beta_j^* - \hat{\beta}_j)(\beta_k^* - \hat{\beta}_k)[\mathbf{E}(X_{0j}X_{0k}) - \frac{1}{n}\sum_{i=1}^{n} X_{ij}X_{ik}].\end{aligned}$$

We then define $V_{j,k} = [\mathbf{E}(X_{0j}X_{0k}) - \frac{1}{n}\sum_{i=1}^{n} X_{ij}X_{ik}]$. Note that $V_{j,k}$ is the mean of $[\mathbf{E}(X_{0j}X_{0k}) - X_{ij}X_{ik}]$. Each of the n terms is bounded so that $[\mathbf{E}(X_{0j}X_{0k}) - X_{ij}X_{ik}] \leq 2M^2$. By Hoeffding's inequality

$$\mathbf{E}(\max_{1\leq j,k\leq p} |V_{j,k}|) \leq 2M^2\sqrt{\frac{2\log(2p^2)}{n}}.$$

We use a version of Hoeffding's inequality that is rather uncommon so we refer the interested reader to the appendix of Chatterjee (2013) for a derivation of the result.

Finally

$$\mathbf{E}[(Y_0^* - \hat{Y}_0)^2|X] - \frac{1}{n}\|Y^* - \hat{Y}\|^2 \leq 4(K + |\mathcal{G}|)^2 \max_{1\leq j,k\leq p} |V_{j,k}|.$$

Combining our results and noting that the right-hand side is deterministic yields Theorem 1

$$\mathbf{E}(Y_0^* - \hat{Y}_0)^2 \leq 2(K + |\mathcal{G}|)M\sigma\sqrt{\frac{2\log(2p)}{n}} + 8(K + |\mathcal{G}|)^2M^2\sqrt{\frac{2\log(2p^2)}{n}}.$$

Proof of Corollary 1

If the smallest eigenvalue of Σ is bounded below by $c > 0$ we can bound $\|\hat{\beta} - \beta^*\|_2$ by the MSPE such that $\|\hat{\beta} - \beta^*\|_2^2 \leq \frac{1}{c}\|\hat{\beta} - \beta^*\|_\Sigma^2$ which follows from the definition of the Rayleigh quotient showing that $\|\hat{\beta} - \beta^*\|_2^2$ goes to 0 as $MSPE(\hat{\beta})$ goes to 0.

Appendix B: Proof of Theorem 3

Roughly, our proof follows two steps.

1. We construct a pair $\hat{\beta}, \hat{z}$ that satisfy the stationarity conditions of
$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2}\|y - X\beta\|_2^2 + \frac{\lambda}{2}\sum_{g\in\mathcal{G}}(\|\beta_g\|_1)^2. \tag{8}$$
while requiring $\hat{\beta}_{S^c} = 0$ and $\hat{z} \in \partial P(\hat{\beta})$. This implies that the pair is a solution to the original problem and that $\hat{\beta}$ has the true support.
2. We show that the constructed solution is the unique solution.

If we were analyzing the Lasso, steps one and two would only be enough to ensure that we have recovered a subset of the true support. For the Exclusive Lasso, verifying that the complement of the support set of the Exclusive Lasso estimate is equal to the complement of the true support set is enough to ensure that we have recovered the true support because the Exclusive Lasso cannot send all variables in a group to zero. Under our assumption of exactly one truly nonzero variable in each group this is enough to establish selection consistency. We leave signed support recovery to later work.

Constructing $\{\hat{\beta}, \hat{z}\}$

We solve the following optimization problem given that we know the true support S. We call it the Exclusive Lasso oracle problem:

$$\hat{\beta}_S = \underset{\beta:\beta_{S^c}=0}{\operatorname{argmin}} \frac{1}{2}\|y - X\beta\|_2^2 + \frac{\lambda}{2}\sum_{g\in\mathcal{G}}(\|\beta_g\|_1)^2 \tag{9}$$

We take the subgradient and set it equal to zero.

$$\begin{pmatrix} X_S^T X_S & X_S^T X_{S^c} \\ X_{S^c}^T & X_{S^c}^T X_{S^c} \end{pmatrix} \begin{pmatrix} \hat{\beta}_S - \beta^* \\ 0 \end{pmatrix} - \begin{pmatrix} X_S^T \epsilon \\ X_{S^c}^T \epsilon \end{pmatrix} + \lambda \begin{pmatrix} \hat{z}_S \\ \hat{z}_{S^c} \end{pmatrix} = 0 \tag{10}$$

We know that when $\hat{\beta}_i$ is nonzero the subgradient of $\hat{z}_S$ is $\text{sign}(\hat{\beta}_i)\|\hat{\beta}_g\|$. Under the assumption that there is exactly one true variable in each group we can simplify so that $\hat{z}_i = \hat{\beta}_i$ for each index i in the support set S.

The term $\hat{z}_{S^c}$ is more complicated. We solve the above equations to get an expression for $\hat{z}_{S^c}$. Rewriting the above as two equations gives:

$$\begin{aligned} X_S^T X_S(\hat{\beta}_S - \beta^*) - X_S^T \epsilon + \lambda \hat{z}_S = 0 \\ X_{S^c}^T X_S(\hat{\beta}_S - \beta^*) - X_{S^c}^T \epsilon + \lambda \hat{z}_{S^c} = 0 \end{aligned} \tag{11}$$

In the first equation $\hat{z}_S = \hat{\beta}_S$ so we use the first equation to solve for $\hat{\beta}_S$. This yields the expression

$$\hat{\beta}_S = (X_S^T X_S + \lambda I)^{-1} X_S^T (X_S \beta^* + \epsilon)$$

We use this expression of the estimate to solve for $\hat{z}_{S^c}$ in the second equation. Plugging in the estimate and rearranging terms yields

$$\hat{z}_{S^c} = \frac{1}{\lambda} X_{S^c}^T [I - X_S (X_S^T X_S + \lambda I)^{-1} X_S^T](X_S \beta^* + \epsilon)$$

We need the $\hat{z}_{S^c}$ that we have constructed to be an element of the subdifferential $\partial P(\hat{\beta})$. For every index $j \in S^c$ and in group g we must satisfy the inequality

$$|\hat{z}_j| < |\hat{\beta}_g|. \tag{12}$$

We propose a series of inequalities that will imply inequality 12. We consider

$$\begin{aligned} |\hat{z}_j| &\leq \|\hat{z}_{S^c}\|_\infty \\ &= \|\frac{1}{\lambda} X_{S^c}^T [I - X_S (X_S^T X_S + \lambda I)^{-1} X_S^T](X_S \beta^* + \epsilon)\|_\infty \\ &\leq \frac{1}{\lambda} \left(\|X_{S^c}^T (X_S \beta^* + \epsilon)\|_\infty + \|X_{S^c}^T X_S (X_S^T X_S + \lambda I)^{-1} X_S^T (X_S \beta^* + \epsilon)\|_\infty \right) \\ &\leq \frac{1}{\lambda} \big(|||X_{S^c}^T|||_{\infty,\infty} \|(X_S \beta^* + \epsilon)\|_\infty \\ &\quad + |||X_{S^c}^T X_S (X_S^T X_S + \lambda I)^{-1}|||_{\infty,\infty} \|X_S^T (X_S \beta^* + \epsilon)\|_\infty \big) \\ &\leq \frac{1}{\lambda} \big(|||X_{S^c}^T|||_{\infty,\infty} \|X_S \beta^* + \epsilon\|_\infty \\ &\quad + |||X_{S^c}^T X_S (X_S^T X_S + \lambda I)^{-1}|||_{\infty,\infty} |||X_S^T|||_{\infty,\infty} \|X_S \beta^* + \epsilon\|_\infty \big) \\ &\leq \frac{1}{\lambda} \left(\|X_S \beta^* + \epsilon\|_\infty + |||X_{S^c}^T X_S (X_S^T X_S + \lambda I)^{-1}|||_{\infty,\infty} \|X_S \beta^* + \epsilon\|_\infty \right) \\ &< \frac{1}{\lambda} (1 + \alpha) \|X_S \beta^* + \epsilon\|_\infty \end{aligned}$$

$$\leq \frac{1}{\lambda}(1+\alpha)\|X_S\beta^*\|_\infty + \frac{1}{\lambda}(1+\alpha)\|\epsilon\|_\infty$$
$$\leq \frac{1}{\lambda}(1+\alpha)|||X_S|||_{1,\infty}\|\beta^*\|_1 + \frac{1}{\lambda}(1+\alpha)\|\epsilon\|_\infty$$
$$\leq \frac{1}{\lambda}(1+\alpha)\|\beta^*\|_1 + \frac{1}{\lambda}(1+\alpha)\|\epsilon\|_\infty$$

The inequalities follow from the definition of induced matrix norms and in the fifth inequality by assumption. We assume that our design matrix satisfies an inequality similar to the irrepresentable condition

$$|||X_{S^c}^T X_S(X_S^T X_S + \lambda I)^{-1}|||_{\infty,\infty} < \alpha.$$

Looking at the right hand side of inequality 12, we note that $\hat{\beta}_{g_{min}} \leq |\hat{\beta}_{g_{min}}| \leq |\hat{\beta}_g|$ where g_{min} is the index of $\hat{\beta}$ with minimum magnitude. Combining this with our definition of $\hat{\beta}_S$ and rearranging terms so the random variable ϵ only appears on the left hand side gives us an inequality

$$\frac{1+\alpha}{\lambda}\|\epsilon\|_\infty - [(X_S^T X_S + \lambda I)^{-1} X_S^T]_{g_{min}}\epsilon$$
$$\leq [(X_S^T X_S + \lambda I)^{-1} X_S^T]_{g_{min}} X_S\beta^* - \frac{1+\alpha}{\lambda}\|\beta^*\|_1$$

that when satisfied implies that inequality 12 holds. Because ϵ is $N(0, I_n)$ we show that the inequality holds with high probability. For Lipschitz continuous function f with constant L and multivariate Gaussian z with mean 0 and identity covariance Massart (2007) gives the following inequality

$$\mathbb{P}(f(z) - \mathbb{E}f(z) \geq t) \leq exp\left(-\frac{t^2}{2L^2}\right). \tag{13}$$

For convenience let $(X_S^T X_S + \lambda I)^{-1} X_S^T = A$. Using this we rewrite the left hand side

$$\frac{1+\alpha}{\lambda}\|\epsilon\|_\infty - [(X_S^T X_S + \lambda I)^{-1} X_S^T]_{g_{min}}\epsilon = \underset{i=1:n}{max}\left\{\frac{1+\alpha}{\lambda}|e_i^T\epsilon| - e_{g_{min}}^T A\epsilon\right\}.$$

For any i we can compute the Lipschitz constant as follows

$$\left|\frac{1+\alpha}{\lambda}|e_i^T x| - \frac{1+\alpha}{\lambda}|e_i^T y| - e_{g_{min}}^T A(x-y)\right|$$
$$\leq \left|\frac{1+\alpha}{\lambda}|e_i^T x| - \frac{1+\alpha}{\lambda}|e_i^T y|\right| + \left|e_{g_{min}}^T A(x-y)\right|$$
$$\leq \left|\frac{1+\alpha}{\lambda}|e_i^T (x-y)|\right| + \left|e_{g_{min}}^T A(x-y)\right|$$
$$\leq \frac{1+\alpha}{\lambda}\|e_i\|_2\|(x-y)\|_2 + \left|e_{g_{min}}^T A(x-y)\right|$$
$$\leq \frac{1+\alpha}{\lambda}\|(x-y)\|_2 + \|e_{g_{min}}^T A\|_2\|(x-y)\|_2$$

$$\leq \frac{1+\alpha}{\lambda}\|(x-y)\|_2 + \|e_{g_{min}}^T\|_2\|A\|_{2,2}\|(x-y)\|_2$$
$$\leq \frac{1+\alpha}{\lambda}\|(x-y)\|_2 + \frac{C_1}{C_1^2+\lambda}\|(x-y)\|_2$$
$$\leq \left(\frac{1+\alpha}{\lambda} + \frac{C_1}{C_1^2+\lambda}\right)\|x-y\|_2$$

where $C_1 = \underset{C\in\{C_{min},C_{max}\}}{argmax} \frac{|C|}{C^2+\lambda}$ is a constant that lets us bound the maximum singular value of A and e_i and $e_{g_{min}}$ are canonical vectors with a 1 in the i or g_{min} index respectively and 0 otherwise.

Plugging in the value for α gives us a Lipschitz constant of

$$L = \frac{1}{\lambda} + \frac{kC_{max}^2}{C_{max}^2+\lambda} + \frac{C_1}{C_1^2+\lambda}$$

Combining this with inequalities 12 and 13 we get

$$\begin{aligned}
&P(\|\hat{z}_{S^c}\|_\infty < |\hat{\beta}_{g_{min}}|) \\
&\geq P\left(\underset{i=1:n}{max}\left\{\frac{1+\alpha}{\lambda}|e_i^T\epsilon| - e_{g_{min}}^T A\epsilon\right\} \leq a_{g_{min}}^T X_S\beta^* - \frac{1+\alpha}{\lambda}\|\beta^*\|_1\right) \\
&= 1 - P\left(\underset{i=1:n}{\cup}\left\{\frac{1+\alpha}{\lambda}e_i^T\epsilon - e_{g_{min}}^T A\epsilon \geq a_{g_{min}}^T X_S\beta^* - \frac{1+\alpha}{\lambda}\|\beta^*\|_1\right\}\right) \\
&- P\left(\underset{i=1:n}{\cup}\left\{-\frac{1+\alpha}{\lambda}e_i^T\epsilon - e_{g_{min}}^T A\epsilon \geq -\left(a_{g_{min}}^T X_S\beta^* + \frac{1+\alpha}{\lambda}\|\beta^*\|_1\right)\right\}\right) \\
&\geq 1 - 2exp\left(-\frac{\left(a_{g_{min}}^T X_S\beta^* - \frac{1+\alpha}{\lambda}\|\beta^*\|_1\right)^2}{L^2} + log(n)\right) \\
&\geq 1 - 2exp\left(-\frac{\left(\frac{1}{\lambda}\|\beta^*\|_1\right)^2}{L^2} + log(n)\right).
\end{aligned}$$

The inequality is a result of the union bound and applying the concentration inequality. In the last inequality, we note that the exponential term is maximized when the numerator is minimized and show that

$$\frac{1}{\lambda}\|\beta^*\|_1 \leq \frac{1+\alpha}{\lambda}\|\beta^*\|_1 - a_{g_{min}}^T X_S\beta^*.$$

We first note that

$$\begin{aligned}
a_{g_{min}}^T X_S\beta^* &= \sum_{j=1}^{p}\sum_{l=1}^{k} V_{g_{min},l}V_{j,l}\frac{d_l^2}{d_l^2+\lambda}\beta_j^* \\
&\leq \sum_{j=1}^{p}\sum_{l=1}^{k}\frac{d_l^2}{d_l^2+\lambda}\beta_j^*
\end{aligned}$$

$$
\begin{aligned}
&\leq k\frac{C_{max}^2}{C_{max}^2+\lambda}\sum_{j=1}^{p}\beta_j^* \\
&\leq \frac{\alpha}{\lambda}\sum_{j=1}^{p}\beta_j^* \\
&\leq \frac{\alpha}{\lambda}|\sum_{j=1}^{p}\beta_j^*| \\
&\leq \frac{\alpha}{\lambda}\sum_{j=1}^{p}|\beta_j^*| \\
&= \frac{\alpha}{\lambda}\|\beta^*\|_1.
\end{aligned}
$$

This implies that the term $\frac{\alpha}{\lambda}\|\beta^*\|_1 - a_{g_{min}}^T X_S\beta^*$ is positive and $\frac{1}{\lambda}\|\beta^*\|_1 \leq \frac{1+\alpha}{\lambda}\|\beta^*\|_1 - a_{g_{min}}^T X_S\beta^*$.

If we want the probability of recovering the true support to be at least $1-1/n$ then we need $\|\beta^*\|_1$ to be on the order of $k\sqrt{log(n)}$. Note that

$$
\begin{aligned}
-log(n) &> -\left(\|\beta^*\|_1/\lambda L\right)^2 + log(n) \\
0 &> -\left(\|\beta^*\|_1/\lambda L\right)^2 + 2log(n) \\
\left(\|\beta^*\|_1/\lambda L\right)^2 &> 2log(n) \\
\|\beta^*\|_1/\lambda L &> \sqrt{2log(n)} \\
\|\beta^*\|_1 &> \lambda L\sqrt{2log(n)}
\end{aligned}
$$

implies that we recover the correct support with probability that goes to 1 as n goes to infinity and it implies a minimum signal strength similar to the beta-min condition in typical sparsistency proofs.

Therefore with probability at least $1 - 1/n$ the inequality $\|\hat{z}_{S^c}\|_\infty < \hat{\beta}_{g_{min}}$ holds.

Uniqueness

In the previous section, we establish strict dual feasibility and therefore that the Exclusive Lasso finds no false positives. Assumption (8) ensures that the restricted program has a unique solution. Together these results imply the solution to the original problem is unique.

Appendix C: Proof of Theorem 4

Our coordinate descent algorithm minimizes the following optimization problem

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\frac{1}{2}\|y - X\beta\|_2^2 + \lambda P(\beta).$$

We show that the assumptions for Theorem 4.1 from Tseng (2001) hold for the problem above. For a function of the form

$$f(x) = g(x) + h(x),$$

where g is convex and differentiable and h is convex but not necessarily differentiable, verifying the assumptions involves showing that:

1. the differential part of our function g satisfies assumption (A1) from Tseng (2001), where assumption A1 states that the domain of g is open and g is Gateux differentiable,
2. the function f is a regular function,
3. the level set $X_0 = \{x : f(x) \leq f(x^0)\}$ is compact and that f is continuous on X_0,
4. for every pair $i, k \in \{1 \dots p\}$ it follows that f is jointly pseudo convex in x_i and x_k.

First we state several definitions. We say direction d is a vector in $\mathbf{R}^n$. We allow d_k to be the scalar in the k^{th} position in the vector $(0 \dots 0, d_k, 0 \dots 0)$. We abuse notation if the meaning is unambiguous, and also let d_k denote the entire vector with 0s in all positions except for the k^{th} position. It is typical to define first order optimality conditions in terms of the Gateaux derivative. We however use the more general forward variation defined as follows:

Definition *For a function f the forward variation in direction d at x is*

$$f'_+(x; d) = \lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t}$$

The Gateaux derivative exists if both the forward and backward variation exist and are equal. Tseng uses the Gateaux derivative to define his optimality conditions but for our unconstrained convex non-differentiable problem it is necessary and sufficient for a minimizer of f to satisfy $f'_+(x; d) \geq 0$ for all $d \in \mathbf{R}^n$. We also use a notion called regularity. Note that this is the same definition of regularity given in Tseng (2001) communicated here for convenience. Throughout the rest of the paper we use the forward variation and the directional derivative interchangeably.

Definition A function f is regular at x if $f'(x; d) \geq 0$ for all d such that $f'(x; d_k) \geq 0$.

Regularity ensures that if we have a point that minimizes f coordinate-wise, then the point minimizes the function f.

Definition A function f is pseudoconvex if $f(x + d) \geq f(x)$ whenever $x \in dom(f)$ and $f'(x; d) \geq 0$.

Assertion 1: The differential part of our function g satisfies assumption (A1) from Tseng (2001).

Proof. If we let

$$g(\beta) = \frac{1}{2}\|y - X\beta\|_2^2,$$

its domain is $\mathbf{R}^p$ which is an open set. We must also show that $g(\beta) = \frac{1}{2}\|y - X\beta\|_2^2$ is Gateux-differentiable on $\mathbf{R}^n$. Then

$$\begin{aligned} g'(\beta; d) &= \lim_{t \downarrow 0} \frac{g(\beta + td) - g(\beta)}{t} \\ &= -d^T X^T (y - X\beta) \\ &= \nabla g(x)^T d. \end{aligned}$$

A similar argument holds as $t \uparrow 0$. □

Assertion 2: The function f is a regular function.

Proof. Our goal is to show that if we have a point β that minimizes f point wise i.e. that $f'(\beta; d_k) \geq 0$ for all d_k then we have a point that minimizes f and satisfies the standard first order necessary and sufficient condition for optimality $f'(\beta; d) \geq 0$ for all $d \in \mathbf{R}^p$. We know that $g(\beta) = \frac{1}{2}\|y - X\beta\|_2^2$ is Gateux-differentiable on $\mathbf{R}^p$.

Next we show that the entire function $f(\beta) = g(\beta) + h(\beta)$ is regular. Assume that the point β minimizes f point wise therefore satisfying:

$$f'(\beta; (0...0, d_k, 0...0)) \geq 0$$

for all d_k. Then it follows that

$$\begin{aligned} f'(\beta; d) &= \nabla g(\beta)^T d + \lim_{t \downarrow 0} \frac{(\sum_{i=1}^n |\beta_i + td_i|)^2 - (\sum_{i=1}^n |\beta_i|)^2}{t} \\ &= \nabla g(\beta)^T d + \lim_{t \downarrow 0} \frac{\left(\sum_{i=1}^n |\beta_i + td_i| - \sum_{i=1}^n |\beta_i|\right)\left(\sum_{i=1}^n |\beta_i + td_i| + \sum_{i=1}^n |\beta_i|\right)}{t} \\ &= \nabla g(\beta)^T d + \lim_{t \downarrow 0} \frac{\left(\sum_{i=1}^n |\beta_i + td_i| - \sum_{i=1}^n |\beta_i|\right)}{t} \lim_{t \downarrow 0} \left(\sum_{i=1}^n |\beta_i + td_i| + \sum_{i=1}^n |\beta_i|\right) \\ &= \nabla g(\beta)^T d + \lim_{t \downarrow 0} \frac{\sum_{i=1}^n |\beta_i + td_i| - \sum_{i=1}^n |\beta_i|}{t} 2\|\beta\| \\ &\geq \nabla g(\beta)^T d + \sum_{i=1}^n \lim_{t \downarrow 0} \frac{|\beta_i + td_i| - |\beta_i|}{t} 2\|\beta\| \\ &= \sum_{i=1}^n f'(\beta; (0, \ldots, 0, d_k, 0, \ldots, 0)) \\ &\geq 0. \end{aligned}$$

□

Assertion 3: The level set $X_0 = \{x : f(x) \leq f(x^0)\}$ is compact and that f is continuous on X_0.

Proof. We show that the function is continuous by showing that the penalty is continuous and that the differentiable part of the objective function is continuous. Let $x, y \in X_0$ then there exists a δ such that for

$$|x - y| \leq \delta$$

it follows that

$$|P(x) - P(y)| \leq \epsilon.$$

To find δ consider

$$\begin{aligned} |P(x) - P(y)| &\leq P(x - y) \\ &= \frac{1}{2}\sum_{g \in \mathcal{G}} (\sum_{i \in g} |x_i - y_i|)^2 \\ &\leq \frac{1}{2}\sum_{g \in \mathcal{G}} (\sum_{i \in g} \delta_i)^2. \end{aligned}$$

Note that the first line follows from the reverse triangle inequality. If $i \in g$ then for any $\epsilon > 0$ we can define δ such that $\delta_i = \frac{\sqrt{2\epsilon}}{n_g \sqrt{|\mathcal{G}|}}$ which shows that the penalty is continuous on the set.

The term $\|y - X\beta\|_2^2$ is Lipschitz continuous with parameter $L = \lambda_{max}(X^T X)$. Therefore f is continuous because the sum of continuous functions is a continuous function. Using Theorem 1.6 of Rockafellar and Wets (2009), continuity implies that the level sets are closed.

The level sets also must be bounded. For any fixed vector β_0, we define the level set as

$$X_0 = \{\beta : \|y - X\beta\|_2^2 + \lambda P(\beta) \leq \|y - X\beta_0\|_2^2 + \lambda P(\beta_0)\}.$$

If we let $\|y - X\beta_0\|_2^2 + \lambda P(\beta_0) = \alpha$ we can consider a vector of the form $\beta_\alpha = (0, \ldots, 0, \sqrt{\frac{2(|\alpha|+1)}{\lambda}}, 0, \ldots, 0)$. Our penalty evaluated at this vector gives $\lambda P(\beta_\alpha) = |\alpha| + 1 > \alpha$. Since $\|y - X\beta\| \geq 0$ for all $x \in \mathbf{R}^n$ the objective function $f(\beta_\alpha) > \alpha$. This implies that for all $\beta \in X_0$, the absolute value of each index of β must be less then $\sqrt{\frac{2(|\alpha|+1)}{\lambda}}$. Therefore the level sets are bounded.

By the Heine-Borel theorem since X_0 a closed bounded subset of $\mathbf{R}^n$ it is compact. □

Assertion 4: For every pair $i, k \in \{1 \ldots p\}$ it follows that f is jointly pseudo-convex in β_i and β_k.

Proof. For any pair of indices $i, k \in \{1 \ldots p\}$ the function

$$\|y - X\beta\|_2^2 + \frac{\lambda}{2} P(\beta)$$

is jointly convex in β_i and β_k. Suppose indices i and k are in the same group. We can rewrite the objective function as

$$f_1(\beta_i, \beta_k) = \|y - X_{i,k}\beta_{i,k} + c\|_2^2 - \frac{\lambda}{2}\sum_{g\in\mathcal{G}}(\sum_{j\in g}|\beta_j|)^2$$
$$= \|y - X_{i,k}\beta_{i,k} - c_0\|_2^2 + \frac{\lambda}{2}(|\beta_i| + |\beta_k| + c_1)^2 + c_2$$

where c_0, c_1, c_2 are terms constant in β_i and β_k, $X_{i,k} = (X_i, X_k)$ are i^{th} and k^{th} columns and $\beta_{i,k} = (\beta_i, \beta_k)$. The function $\|y - X_{i,k}\beta_{i,k} - c_0\|_2^2$ has a positive semidefinite Hessian so it is convex. We appeal to the definition of convexity for the penalty. When the elements are in the same group

$$\begin{aligned}\frac{\lambda}{2}(\|tx + (1-t)y\|_1 + c_1)^2 &\leq \frac{\lambda}{2}(\|tx\|_1 + \|(1-t)y\|_1 + c_1)^2 \\ &\leq \frac{\lambda}{2}(\|tx\|_1 + tc_1)^2 + \frac{\lambda}{2}(\|(1-t)y\|_1 + (1+t)c_1)^2 \\ &\leq \frac{\lambda}{2}t^2(\|x\|_1 + c_1)^2 + \frac{\lambda}{2}(1+t)^2(\|(y\|_1 + c_1)^2 \\ &\leq \frac{\lambda}{2}t(\|x\|_1 + c_1)^2 + \frac{\lambda}{2}(1+t)(\|y\|_1 + c_1)^2.\end{aligned} \tag{14}$$

If i, k are in different groups then

$$\lambda P(tx + (1-t)y) = \frac{\lambda}{2}(|tx_1 + (1-t)y_1 + c_1|)^2 + (|tx_2 + (1-t)y_2 + c_2|)^2 + c_3$$

which is the sum of two convex functions and therefore convex.

Therefore the function f is convex in every pair of indices which implies that it is pseudoconvex in every pair of indices. □

Given that the objective function satisfies all of the assumptions for Tseng (2001) Theorem 4.1 we can say that our coordinate descent algorithm converges to a stationary point. Because our function is convex the stationary point is a global minimum.

Appendix D: Proof of Theorem 5

For a continuous and almost differentiable function g, Steins formula

$$df(g) = \mathbf{E}[(\nabla * g)(y)]$$

defines the degrees of freedom for normal random variables in terms of the function $(\nabla * g)$. The function $(\nabla * g)$ known as the divergence is defined for $g : \mathbf{R}^n \to \mathbf{R}^n$ as

$$(\nabla * g)(y) = \sum_{i=1}^{n}\frac{\partial g_i}{\partial y_i}.$$

To derive the degrees of freedom for the Exclusive Lasso problem we need to prove that the estimate is a continuous and almost differentiable function of y. Tibshirani provides a lemma stating that for a convex set $C \subset \mathbf{R}^n$ the projection map P_C and the map $I - P_C$ are continuous and almost differentiable.

For proof see Tibshirani et al. (2012).

Lemma The estimate $X\,\hat{\beta} = (I - P_C)y$ for the set

$$C = \{u \in \mathbf{R}^n : P^*(X^T u) \leq \alpha\}$$

where

$$P^*(\beta) = \sqrt{\sum_{g \in \mathcal{G}} \|\beta_g\|_\infty^2}$$

is the dual norm of the square root of our penalty and α is a constant.

Proof. The dual norm of a norm $\|z\|$ is defined as the norm $\|x\|^*$ such that $\|z\| = \sup\{\langle x, z\rangle : \|x\|^* \leq 1\}$. Note that for the square root of our penalty

$$\sqrt{P(\hat{\beta})} = \left\langle \frac{sign(\hat{\beta})\| \hat{\beta}_{g_i} \|_1}{\sqrt{\sum_g \| \hat{\beta} \|_1^2}}, \hat{\beta} \right\rangle$$

This means that our dual norm is the norm such that $P^*(\frac{sign(\hat{\beta})\| \hat{\beta}_{g_i} \|_1}{\sqrt{\sum_g \| \hat{\beta} \|_1^2}}) \leq 1$ which holds for the norm

$$P^*(\beta) = \sqrt{\sum_{g \in \mathcal{G}} \|\beta_g\|_\infty^2}$$

We show that $\theta = y - X\,\hat{\beta}$ is equal to the projection of y onto the set C. The projection $\theta = P_C(y)$ can be characterized as a point θ satisfying the first order optimality conditions for the constrained optimization problem $\min_{\theta \in C} \|y - \theta\|_2^2$. The first order optimality conditions are

$$f'(\theta; d) \geq 0$$
$$\langle y - \theta, \theta - u\rangle \geq 0$$

for all $u \in C$.

We must verify that $f'(\theta; d) \geq 0$. If we let $\theta = y - X\,\hat{\beta}(y)$ then

$$\langle y - \theta, \theta - u\rangle = \langle X\,\hat{\beta}, y - X\,\hat{\beta} - u\rangle \tag{15}$$
$$= \langle X\,\hat{\beta}, y - X\,\hat{\beta}\rangle - \langle X^T u, \hat{\beta}\rangle \tag{16}$$
$$= \frac{\alpha}{2}\sqrt{P(\hat{\beta})} - \langle X^T u, \hat{\beta}\rangle \tag{17}$$
$$= \max_{P^*(w) \leq \frac{\alpha}{2}} \langle w, \hat{\beta}\rangle - \langle X^T u, \hat{\beta}\rangle \tag{18}$$
$$\geq 0. \tag{19}$$

Line 3 follows from the fact that there exists a regularization parameter such that the necessary conditions for the Exclusive Lasso problem are exactly the same as the necessary conditions for the optimization problem that uses the square root of the Exclusive Lasso penalty. Notice that if we let $\alpha = 2\lambda P(\hat{\beta})^{\frac{1}{2}}$ then $\lambda \partial P(\hat{\beta}) = \alpha \partial \sqrt{P(\hat{\beta})}$. This implies that $\hat{\beta}$ necessarily satisfies

$$-X^T(y - X\,\hat{\beta}) + \alpha \partial \sqrt{P(\hat{\beta})} = 0.$$

Taking the inner product with $\hat{\beta}$ yields

$$(X\,\hat{\beta})^T(y - X\,\hat{\beta}) = \frac{\alpha}{2}\sqrt{P(\hat{\beta})}.$$

Line 5 follows for the set $C = \{u \in \mathbf{R}^n : P^*(X^T u) \leq \frac{\alpha}{2}\}$ proving that $y - X\,\hat{\beta}$ is equal to the projection of y onto the set C. This implies that $X\,\hat{\beta} = (I - P_C)y$. □

Combining Lemmas 1 and 2 yields that the exclusive lasso estimate is continuous and almost differentiable. Next we define $\hat{\beta}$ in terms of the support set S. First recall the KKT conditions

$$-X^T(y - X\,\hat{\beta}) + \lambda z = 0$$

where

$$z_i = \begin{cases} sign(\hat{\beta}_i)\|\,\hat{\beta}_g\,\|_1 & : \hat{\beta}_i \neq 0, i \in g \\ \left[-\|\,\hat{\beta}_g\,\|_1, \|\,\hat{\beta}_g\,\|_1\right] & : \hat{\beta}_i = 0. \end{cases}$$

Note that we can rewrite the sub gradient for the indices $i \in g \cap S$. If we let $s_{g\cap S} = sign(\hat{\beta}_{g\cap S})$

$$z_{g\cap S} = s_{g\cap S} s_{g\cap S}^T\,\hat{\beta}_{g\cap S}\,.$$

We can write the sub gradient over the indices of the support as

$$z_S = M_S\,\hat{\beta}_S$$

where M_S is a block diagonal matrix with the matrices $\{s_{g\cap S}s_{g\cap S}^T : g \in \mathcal{G}\}$ on the diagonal.

We can rewrite the KKT conditions with respect to the support set

$$-\begin{bmatrix} X_S^T \\ X_{S^c}^T \end{bmatrix}\left(y - \begin{bmatrix} X_S & X_{S^c} \end{bmatrix}\hat{\beta}\right) + \lambda\begin{pmatrix} z_S \\ z_{S^c} \end{pmatrix} = 0.$$

This is equal to

$$\begin{aligned} -X_S^T y + X_S^T X_S\,\hat{\beta}_S + \lambda z_S &= 0 \\ -X_{S^c}^T y + X_{S^c}^T X_S\,\hat{\beta}_S + \lambda z_{S^c} &= 0. \end{aligned}$$

We then solve for $\hat{\beta}_S$ using $z_S = M_S\,\hat{\beta}_S$ yielding

$$\hat{\beta}_S = (X_S^T X_S + \lambda M_S)^\dagger X_S^T y.$$

Note that we are relying on the fact that we have already proved the existence of a solution to the optimization problem in the proof for Theorem 4. This gives us an estimate $\hat{y} = X_S(X_S^T X_S + \lambda M_S)^\dagger X_S^T y$. The divergence is therefore

$$(\nabla * X\,\hat{\beta})(y) = trace[X_S(X_S^T X_S + \lambda M_S)^\dagger X_S^T]$$

which is equal to the sum of the eigenvalues.

Appendix E: Selecting one variable per group

For specific values of X and y the Exclusive Lasso will select more than one variable per group for all values of the regularization parameter λ. This means that although the Exclusive Lasso is designed to select exactly one element per group we cannot guarantee the Exclusive Lasso will enforce the correct structure. Consider an example. Suppose we characterize the Exclusive Lasso estimate using the equicorrelation set. Recall the equicorrelation set

$$\mathcal{E} = \left\{ i : \frac{|X_i^T(y - X\,\hat{\beta})|}{\|\,\hat{\beta}_g\,\|_1} = \lambda \right\}$$

If we let s be a vector such that $s_i = sign(\hat{\beta}_i)$ for $i \in \mathcal{E}$ and γ be a vector such that $\gamma_i = \|\,\hat{\beta}_{g_i}\,\|_1$ where g_i is the group for an index $i \in \mathcal{E}$. Let $\bar{\gamma}$ be a vector such that $\bar{\gamma}_i = \|\,\hat{\beta}_{g_i}\,\|_1 - |\,\hat{\beta}_i\,|$ then we can solve for $\hat{\beta}$.

$$\begin{aligned} X_{\mathcal{E}}^T(y - X_{\mathcal{E}}\,\hat{\beta}_{\mathcal{E}}) &= \lambda\gamma s \\ &= \lambda\bar{\gamma}s + \lambda\,\hat{\beta}_{\mathcal{E}} \\ \hat{\beta}_{\mathcal{E}} &= (X_{\mathcal{E}}^T X_{\mathcal{E}} + \lambda I)^{-1}[X_{\mathcal{E}}^T y - \lambda\bar{\gamma}s]. \end{aligned}$$

Let $X = I_2$ and we let $y^T = (1, 1)$ then because X is orthonormal the estimate simplifies to

$$\hat{\beta}_{\mathcal{E}} = \frac{1}{1+\lambda}y - \frac{\lambda}{1+\lambda}\gamma' s.$$

In this case $\hat{\beta}_1 = \hat{\beta}_2$ so the term $\frac{\lambda}{1+\lambda}\gamma' s$ is going to shrink both indices equally for all λ. This prevents the estimate from selecting exactly one element in each group.

We conjecture that conditions on X and y for this to occur can be formalized, but this is beyond the scope of this work. Intuitively, this behavior occurs when two or more variables get shrunken equally. As such, this behavior is relatively rare in practice. If it does occur and one variable per group is desired, we propose to use BIC to select λ and apply group-wise thresholding.

Appendix F: Selecting the regularization parameter

We compare our unbiased estimate of the degrees of freedom to simulated degrees of freedom following the set up outlined in Efron et al. (2004) and Zou et al. (2007). These works use Stein's unbiased risk estimation to estimate the degrees of freedom of an estimate of the form $\hat{y} = Hy$ where y is Gaussian.

In our simulation, we let β^* be the true parameter and we simulate y^*, B times such that $y = X\beta^* + \epsilon$. For simplicity we let $\epsilon \sim N(0, 1)$. We then calculate an estimate for the covariance using constant $c = 0$, again for simplicity. Because y is standard Gaussian with $\sigma^2 = 1$, the simulated degrees of freedom is

$$df(\hat{y}) = \sum_{i=1}^{n} \widehat{cov}(\hat{y}_i, y_i)/\sigma^2$$

where we simulate the covariances according to $\widehat{cov}(\hat{y}_i, y_i) = \frac{1}{B}\sum_{b=1}^{B}(\hat{y}_i - a_i)(y_i^* - X\beta_i^*)$ where a_i is a fixed known constant. In our work we let $a_i = [X\beta^*]_i$. In our simulations we set $B = 500$ and found that empirically our estimate of the degrees of freedom matched the simulated degrees of freedom quite well.

In this section we include the following table that lists the point estimates and their standard errors.

Regularization Parameter	Simulated DF (Standard Error)	Estimated DF (Standard Error)
0.01	94.76 (1.35)	95.46 (1.45)
0.01	94.45 (1.36)	94.56 (1.55)
0.01	94.07 (1.34)	93.59 (1.69)
0.02	92.16 (1.32)	92.43 (1.84)
0.02	91.12 (1.30)	90.92 (2.02)
0.03	89.36 (1.29)	89.18 (2.27)
0.04	87.39 (1.25)	87.27 (2.37)
0.06	85.33 (1.24)	85.00 (2.54)
0.07	80.94 (1.17)	82.12 (2.62)
0.10	78.83 (1.15)	78.75 (2.94)
0.14	75.92 (1.13)	75.07 (3.07)
0.18	70.38 (1.04)	70.48 (3.33)
0.25	65.00 (0.98)	65.41 (3.49)
0.33	59.37 (0.91)	59.69 (3.84)
0.45	52.89 (0.84)	52.83 (3.82)
0.61	45.45 (0.74)	45.83 (3.75)
0.82	37.57 (0.65)	37.89 (3.68)
1.11	30.39 (0.59)	30.41 (3.55)
1.49	22.77 (0.52)	23.01 (3.17)
2.01	17.44 (0.52)	17.26 (2.59)
2.72	13.16 (0.55)	12.96 (2.07)
3.67	9.96 (0.64)	10.16 (1.73)
4.95	8.02 (0.78)	8.55 (1.51)
6.69	6.60 (0.99)	7.28 (1.21)
9.03	6.39 (1.29)	6.43 (0.96)
12.18	5.32 (1.66)	5.78 (0.84)
16.44	4.25 (2.15)	5.29 (0.64)
22.20	5.85 (2.74)	4.81 (0.54)

Appendix G: Additional simulations

Here, we provide additional simulations that explore the Exclusive Lasso under several conditions: unequal group sizes, unequal signal strengths, and a combination of the unequal group sizes and signal strengths. Our simulation set-up largely follows from those used for Table 1 with the following modifications. When the group sizes are unequal, the groups consist of 10, 30, 20, 30, and 10 variables. When the group sizes are equal, each group consists of 20 variables. When the signal strengths are unequal, the values for the true parameters are 1, -4, 2, 1, and -1. When the signal strengths are of equal, the magnitude of the true parameters is 0.7 for all true variables. For these simulations, we let w = .8 and b = .8, again denoting the within and between group correlations

of the autoregressive covariance matrix. Results are presented in Table 4. Overall, the Exclusive Lasso seems to perform equally well for unequal group sizes, but is effected by differences in signal strength. In this situation, however, the Thresholded Exclusive Lasso outperforms all other methods.

Appendix H: Proximal gradient descent

We develop a method for efficiently computing the proximal operator of the Exclusive Lasso penalty. The proximal operator facilitates the development of common first order algorithms like proximal gradient descent and alternating direction method of multipliers (ADMM) but it is impossible to compute in closed form. We develop an algorithm for approximating the proximal operator based on coordinate descent. In this section, we use our proximal operator algorithm to develop a proximal gradient descent algorithm and prove that it converges.

Recall the proximal operator problem

$$prox_P(z) = \underset{\beta}{\operatorname{argmin}} \frac{1}{2}\|\beta - z\|_2^2 + \frac{\lambda}{2}\sum_{g\in\mathcal{G}} \|\beta_g\|_1^2. \tag{20}$$

For the Exclusive Lasso penalty, there is no closed form solution, so we use a coordinate descent algorithm to compute the proximal operator. Coordinate updates for the proximal operator algorithm are extremely similar to the updates for the Exclusive Lasso regression problem.

Lemma (1): For proximal operator $prox_P(z)$ where P is our Exclusive Lasso penalty, if $S(z,\lambda) = sign(z)(|z| - \lambda)_+$ and j is in group g then the coordinate wise updates are:

$$\beta_j^{k+1} = S\left(\frac{1}{1+\lambda} z_j, \frac{\lambda}{1+\lambda}\sum_{l\in g\setminus j} |\beta_l^k|\right). \tag{21}$$

This algorithm converges:

Corollary (2): The proximal operator coordinate descent algorithm converges to the global minimum of the optimization problem given in (20).

Proof This is a special case of the Exclusive Lasso regression problem with $X = I$.

Proximal Gradient Descent

Given a computable proximal operator many first order algorithms can be derived for the Exclusive Lasso. As an example we derive a proximal gradient descent algorithm for the Exclusive Lasso. Loosely, proximal gradient descent

TABLE 4

Comparison of variable selection methods with exactly one true variable in each group. The groups sizes and the magnitude of the true parameters are unequal.

Unequal Signal Strength

	Exclusive Lasso	Lasso	Marginal Regression	Group-wise Marginal Regression	Thresholded Exclusive Lasso
True Vars (SE)	2.00 (0.00)	3.50 (0.71)	1.00 (0.00)	2.00 (0.0 0)	**5.00 (0.00)**
False Vars (SE)	3.00 (0.00)	1.50 (0.71)	4.00 (0.00)	3.00 (0.00)	**0.00 (0.00)**
Pred Err (SE)	1.84 (0.11)	1.48 (0.21)	2.60 (0.17)	1.84 (0.10)	**1.03 (0.07)**
	Thresholded Lasso	Thresholded Regularization Path	Elastic Net	Thresholded Ridge	
True Vars (SE)	3.80 (0.92)	3.50 (0.71)	1.60 (0.52)	3.50 (0.53)	
False Vars (SE)	1.20 (0.92)	1.50 (0.71)	3.40 (0.52)	1.50 (0.53)	
Pred Err (SE)	1.33 (0.20)	1.37 (0.14)	2.18 (0.44)	1.59 (0.13)	

Unequal Group Size

	Exclusive Lasso	Lasso	Marginal Regression	Group-wise Marginal Regression	Thresholded Exclusive Lasso
True Vars (SE)	**4.90 (0.32)**	3.10 (0.57)	2.20 (0.79)	2.70 (0.67)	4.80 (0.32)
False Vars (SE)	**0.10 (0.32)**	1.90 (0.57)	2.80 (0.79)	2.30 (0.67)	0.20 (0.32)
Pred Err (SE)	**1.10 (0.10)**	1.32 (0.08)	2.58 (0.12)	1.25 (0.13)	1.10 (0.10)
	Thresholded Lasso	Thresholded Regularization Path	Elastic Net	Thresholded Ridge	
True Vars (SE)	3.20 (1.23)	3.20 (0.79)	2.30 (0.82)	3.30 (0.67)	
False Vars (SE)	1.80 (1.23)	1.80 (0.79)	2.70 (0.82)	1.70 (0.67)	
Pred Err (SE)	1.27 (0.15)	1.22 (0.15)	1.44 (0.13)	1.33 (0.16)	

Unequal Group Size and Unequal Signal Strength

	Exclusive Lasso	Lasso	Marginal Regression	Group-wise Marginal Regression	Thresholded Exclusive Lasso
True Vars (SE)	2.10 (0.32)	3.50 (0.71)	1.00 (0.00)	2.00 (0.00)	**5.00 (0.00)**
False Vars (SE)	2.90 (0.32)	1.50 (0.71)	4.00 (0.00	3.00 (0.00)	**0.00 (0.00)**
Pred Err (SE)	1.81 (0.18)	1.48 (0.21)	2.60 (0.17)	1.84(0.12)	**1.03 (0.07)**
	Thresholded Lasso	Thresholded Regularization Path	Elastic Net	Thresholded Ridge	
True Vars (SE)	3.80 (0.92)	3.50 (0.71)	1.60 (0.52)	3.50 (0.53)	
False Vars (SE)	1.20 (0.92)	1.50 (0.71)	3.40 (0.52)	1.50 (0.53)	
Pred Err (SE)	1.33 (0.44)	1.37 (0.14)	2.19 (0.44)	1.59 (0.13)	

algorithms proceed by moving in the negative gradient direction of the smooth loss projected onto the set defined by the non-smooth penalty. Proximal gradient descent can be accelerated yielding a convergence rate that is optimal for first order methods. However for simplicity, we present the unaccelerated version. For our proximal gradient descent algorithm, each coordinate update depends on the other coordinates in the same group. Because of this, we can implement this in parallel over the groups, dramatically reducing computation time. We use our coordinate descent algorithm to compute the proximal operator. Using the negative gradient of our ℓ_2 regression loss, our proximal gradient descent update is $\beta^{k+1} = prox_P(\beta^k - (X^T X\beta^k - X^T y)/L)$, where $L = \lambda_{\max}(X^T X)$ is the Lipschitz constant for our squared error loss. Putting everything together, we give an algorithm outline for our Exclusive Lasso proximal gradient descent algorithm in Algorithm 2.

Input: $\beta^0 \in \mathbf{R}^p, \epsilon, \delta > 0$
Output: $\hat{\beta} \in \mathbf{R}^p$
while $\|\beta^{k+1} - \beta^k\| > \epsilon$ **do**
  $z_g = \beta_g^k - \frac{1}{L}(X_g^T X\beta^k - X_g^T y)$
  In parallel for each g:
  Initialize $\tilde{\beta}_g \in \mathbf{R}^{p_g}$
  while $\|\tilde{\beta}_g^{t+1} - \tilde{\beta}_g^t\| > \delta$ **do**
    for $j \in g$ **do**
      $\tilde{\beta}_j^{t+1} = S\left(\frac{1}{\lambda+1}[z_g]_j, \frac{\lambda}{\lambda+1}\sum_{l \in g \setminus j} |\tilde{\beta}_l^t|\right)$
  $\beta_g^{k+1} = \tilde{\beta}_g$
return β

Algorithm 2: A Proximal Gradient Descent Algorithm

We also need to prove convergence of Algorithm 2 which is non standard as we never calculate the proximal operator exactly, resulting in a sequence of errors $\{\epsilon_k\}$. We show that as long as the sequence of errors converges to zero, the proximal gradient descent algorithm will converge.

Theorem (6): The objective values of the proximal gradient descent algorithm converge to the Exclusive Lasso solution at a rate of at least $O(1/k)$ when the sequences $\{\|\epsilon_k\|\}$ and $\{\sqrt{\epsilon_k}\}$ are summable.

Proof Our result depends on work by Schmidt et al. (2011). We seek the convergence rate for the our Exclusive Lasso proximal gradient descent algorithm. In our algorithm at each step k the proximal operator is computed to within a small error ϵ_k such that the iterate $x_k = \epsilon_k + \underset{x}{argmin}\|y - x\|_2^2 + \lambda P(x)$. As long as the sequence of errors is summable the algorithm will converge at a rate of at least $O(1/k)$ when the following assumptions hold. For function $f(x) = g(x) + h(x)$ we assume:

1. the function g is convex with a Lipschitz-continuous gradient,
2. the function h is a lower semi-continuous proper convex function,

3. there exists a point $x^* \in \mathbf{R}$ that minimizes f,
4. the points x_k are ϵ_k-optimal solutions to the proximal operator optimization problem at iteration k.

We must verify that these assumptions hold for the Exclusive Lasso

Assumption 1: In our case $g(\beta) = \frac{1}{2}\|y - X\beta\|_2^2$ so

$$\begin{aligned} |\|y - X\beta_1\|_2 - \|y - X\beta_2\|_2| &\leq \|(y - X\beta_1) - (y - X\beta_2)\|_2 \\ &= \|X(\beta_1 - \beta_2)\|_2 \\ &\leq \|X\|_2\|(\beta_1 - \beta_2)\|_2 \\ &= \lambda_{max}(X^T X)\|(\beta_1 - \beta_2)\|_2 \end{aligned}$$

which implies that g is Lipschitz-continuous with Lipschitz constant $L = \lambda_{max}(X^T X)$ the largest eigenvalue of $X^T X$.

Assumption 2: Because $\|x\|_1$ is continuous for all $x \in \mathbf{R}^n$ and $b(z) = z^2$ is continuous for all $z \in \mathbf{R}$ their composition $\|x\|_1^2$ is continuous at all points in $\mathbf{R}^n$. To show that the penalty is convex we will consider the convexity of $f(x) = \|x\|_1^2$. For $t \in [0, 1]$ and $x, z \in \mathbf{R}^n$

$$\begin{aligned} \|tx + (1-t)z\|_1^2 &\leq (t\|x\|_1 + (1-t)\|z\|)^2 \\ &\leq t\|x\|_1^2 + (1-t)\|z\|_1^2. \end{aligned}$$

Therefore $f(x) = \|x\|_1^2$ is convex. The convexity of $P(\beta)$ follows from the fact that the sum of convex functions is also convex.

The penalty is proper by definition since for all $x \in \mathbf{R}^n$ we have $P(x) \neq \infty$.

Assumption 3: Using Theorem 1.9 from Rockafellar and Wets (2009) we show existence of a solution. We know the level sets $X_\alpha = \{x : f(x) \leq \alpha\}$ are bounded for all $\alpha \in \mathbf{R}$ by Assertion (3) in Appendix C and we have already shown that both g and h are continuous so their sum must also be continuous. Therefore, because the level sets of our function f are bounded, and f is continuous and proper by Theorem 1.9 there exists a minimum to our objective function f.

Assumption 4: This assumption holds by Theorem 4.

Therefore by proposition 1 from Schmidt et al. (2011) the Exclusive Lasso algorithm converges at a rate of $O(1/k)$.

Overall, this particular algorithm compares well to the iterative soft thresholding algorithm, a proximal gradient descent algorithm for the Lasso (Beck and Teboulle, 2009). Although computing the proximal operator is more complicated due to the structure of the penalty, the convergence rate is the same order as the convergence rate for the iterative soft thresholding algorithm. The fact that the iterates are easy to compute and the convergence results are competitive reinforce our empirical observations; despite the additional structure, the Exclusive Lasso proximal gradient descent algorithm compares well to first order methods for the Lasso and other penalized regression problems.

Acknowledgments

The authors would like to thank Dr. Zhandong Liu for helpful conversations while preparing the manuscript and for his help acquiring the NMR spectroscopy data. We also thank the anonymous reviewers and associate editor for comments and suggestions that led to many improvements in this paper. FC acknowledges support from NSF Graduate Research Fellowship Program under grant No.0940902. GIA acknowledges support from NSF DMS-1264058 and NSF DMS- 1554821.

References

Beck, A. and M. Teboulle (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences 2*(1), 183–202. MR2486527

Bühlmann, P. and S. Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media. MR2807761

Cavanagh, J., W. J. Fairbrother, A. G. Palmer III, and N. J. Skelton (1995). *Protein NMR spectroscopy: principles and practice*. Academic Press.

Chatterjee, S. (2013). Assumptionless consistency of the lasso. *arXiv preprint arXiv:1303.5817*.

Chen, J. and Z. Chen (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika 95*(3), 759–771. MR2443189

De Graaf, R. A. (2013). *In vivo NMR spectroscopy: principles and techniques*. John Wiley & Sons.

Ebbels, T. M., J. C. Lindon, and M. Coen (2011). Processing and modeling of nuclear magnetic resonance (nmr) metabolic profiles. In *Metabolic Profiling*, pp. 365–388. Springer.

Efron, B., T. Hastie, I. Johnstone, R. Tibshirani, et al. (2004). Least angle regression. *The Annals of Statistics 32*(2), 407–499. MR2060166

Genovese, C. R., J. Jin, L. Wasserman, and Z. Yao (2012). A comparison of the lasso and marginal regression. *The Journal of Machine Learning Research 13*(1), 2107–2143. MR2956354

Greenshtein, E., Y. Ritov, et al. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli 10*(6), 971–988. MR2108039

Halabi, M. E. and V. Cevher (2014). A totally unimodular view of structured sparsity. *arXiv preprint arXiv:1411.1990*. MR3382112

Ma, S., X. Song, and J. Huang (2007). Supervised group lasso with applications to microarray data analysis. *BMC bioinformatics 8*(1), 1. MR2707737

Massart, P. (2007). *Concentration inequalities and model selection*, Volume 6. Springer. MR2319879

Obozinski, G. and F. Bach (2012). Convex relaxation for combinatorial penalties. *arXiv preprint arXiv:1205.1240*.

Rockafellar, R. T. and R. J.-B. Wets (2009). *Variational analysis*, Volume 317. Springer Science & Business Media.

Schmidt, M., N. L. Roux, and F. R. Bach (2011). Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in Neural Information Processing Systems*, pp. 1458–1466.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464. MR0468014

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics 22*(2), 231–245. MR3173712

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics 9*, 1135–1151. MR0630098

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288. MR1379242

Tibshirani, R. J., J. Taylor, et al. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics 40*(2), 1198–1232.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications 109*(3), 475–494.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on 55*(5), 2183–2202.

Weljie, A. M., J. Newton, P. Mercier, E. Carlson, and C. M. Slupsky (2006). Targeted profiling: quantitative analysis of 1h nmr metabolomics data. *Analytical Chemistry 78*(13), 4430–4442.

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

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(1), 49–67.

Zhang, S., C. Zheng, I. R. Lanza, K. S. Nair, D. Raftery, and O. Vitek (2009). Interdependence of signal processing and analysis of urine 1h nmr spectra for metabolic profiling. *Analytical Chemistry 81*(15), 6080–6088.

Zhao, P., G. Rocha, and B. Yu (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics 37*(6A), 3468–3497.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *Journal of Machine learning research 7*(Nov), 2541–2563.

Zhou, Y., R. Jin, and S. Hoi (2010). Exclusive lasso for multi-task feature selection. In *International Conference on Artificial Intelligence and Statistics*, pp. 988–995.

Zou, H., T. Hastie, R. Tibshirani, et al. (2007). On the degrees of freedom of the lasso. *The Annals of Statistics 35*(5), 2173–2192.