Estimation Consistency of the Group Lasso and its Applications

Han Liu
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213

Jian Zhang
Department of Statistics
Purdue University
West Lafayette, IN, 47907-2066

Abstract

We extend the ℓ_2-consistency result of (Meinshausen and Yu 2008) from the Lasso to the group Lasso. Our main theorem shows that the group Lasso achieves estimation consistency under a mild condition and an asymptotic upper bound on the number of selected variables can be obtained. As a result, we can apply the nonnegative garrote procedure to the group Lasso result to obtain an estimator which is simultaneously estimation and variable selection consistent. In particular, our setting allows both the number of groups and the number of variables per group increase and thus is applicable to high-dimensional problems. We also provide estimation consistency analysis for a version of the sparse additive models with increasing dimensions. Some finite-sample results are also reported.

1 Introduction

Recently many regularization-based methods have been proposed for the purpose of variable selection in high-dimensional regression. The Lasso (Tibshirani, 1996; Chen et al., 1998) is the most popular one due to its computational feasibility and amenability to theoretical analysis. One well-known result is that the Lasso estimator is not variable selection consistent if the irrepresentable condition fails (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2007), which means the correct sparse subset of the relevant variables can not be identified asymptotically with large probability. However, a recent result from Meinshausen and Yu (2009) shows that even the variable selection fails, the Lasso estimator can still be ℓ_2-consistent in estimation. Which means, even if the exact sparsity pattern might not be recovered, the estimator can still be a good approximation to the truth. This also suggests that, for Lasso, estimation consistency might be easier to achieve than variable selection consistency.

In this paper we are interested in building similar results for the grouped variable selection problems. Grouped variables often appear in real world applications. For example, in many data mining problems we encode categorical variables using a set of dummy variables and as a result they form a group. Another example is additive model, where each component function can be represented using its basis expansions which can be treated as a group. For such problems, it is more natural and suitable to select groups of variables instead of individual ones.

One of our contributions is to extend the fixed design consistency analysis in (Meinshausen and Yu, 2009) from the Lasso to the group Lasso (Yuan and Lin, 2006), which can be viewed as an extension of the Lasso for the grouped variables by replacing the ℓ_1-regularization with the sum of ℓ_2-norm regularization. This extension is non-trivial since the analysis in (Meinshausen and Yu, 2009) utilizes several properties only hold for the ℓ_1-regularization, e.g. piecewise linear solution path and the number of nonzero entries is bounded by the sample size etc. Besides the ℓ_2-consistency result proved in Theorem 1, the optimal rate of convergence and an upper bound of the number of selected variables are also obtained in Corollary 1 and Lemma 6. Furthermore, we use the group Lasso as an initial estimator and apply the nonnegative garrote (Yuan and Lin, 2007) to obtain an estimator in Definition 3 which is simultaneously estimation and variable selection consistent. Since our analysis allows both the number of groups and the number of variables per group increase with the sample size, these results can be extended to the infinite-dimensional cases to

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5.

provide consistency results for a version of the Sparse Additive Models (Ravikumar et al., 2007) from Definition 4.

Some related work on the group Lasso include (Bach, 2008; Meier et al., 2007; Obozinski et al., 2008), which provide either risk analysis or variable selection results using random design. Our fixed design estimation consistency result is complementary to them. A previous work on the nonnegative garrote has been done by Yuan and Lin (2007), but they mainly focus on fixed dimension instead of the increasing dimension as in our case. The sparse additive models proposed in (Ravikumar et al., 2007) focus more on the consistency analysis of risk and variable selection, which are complementary to our estimation consistency analysis.

2 The Group Lasso

We consider the problem of recovering a high-dimensional vector $\beta \in \mathbf{R}^{m_n}$ using a sample of independent pairs $(X_{1\bullet}, Y_1), \ldots, (X_{n\bullet}, Y_n)$ from a multiple linear regression model,

$$Y = X\beta + \epsilon.$$

Here Y is the $n \times 1$ response vector and X represents the observed $n \times m_n$ design matrix whose i-th row vector is denoted by $X_{i\bullet}$. β is the true unknown coefficient vector that we want to recover, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ is an $n \times 1$ vector of i.i.d. noise with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

We are interested in the situation where all the variables are naturally partitioned into p_n groups. Suppose the number of variables in the j-th group is d_j, then by definition we have $m_n = \sum_{j=1}^{p_n} d_j$. We can rewrite this linear model as $Y = X\beta + \epsilon = \sum_{j=1}^{p_n} X_j\beta_j + \epsilon$, where X_j is an $n \times d_j$ matrix corresponding to the j-th group (which could be either categorical or continuous) and β_j is the corresponding $d_j \times 1$ coefficient subvector. Therefore, we have $X = (X_1, \ldots, X_{p_n})$ and $\beta = (\beta_1^T, \ldots, \beta_{p_n}^T)^T$. Both X and Y are assumed to be centered at zero to simplify notation. We also use $X_{\underline{j}}$ to represent the j-th column in the design matrix X and assume that all columns in the design matrix are standardized, i.e. $\frac{1}{n}\|X_{\underline{j}}\|_{\ell_2}^2 = 1, \underline{j} = 1, \ldots, m_n$. Similar to the notation of $X_{\underline{j}}$, we use $\beta_{\underline{j}}$ ($\underline{j} = 1, \ldots, m_n$) to denote the j-th entry of the vector β. Since we are mainly interested in the high-dimensional setting, we assume $p_n \gg n$. Furthermore, we also allow the group size d_j to increase with n at a rate $d_j = o(n)$ and define $\bar{d}_n = \max_j d_j$ to be the upper bound of the group size for each n. In the following we suppress the subscript n if no confusion.

Given the design matrix X and the response vector Y, the group Lasso estimator is defined as the solution of the following convex optimization problem:

$$\widehat{\beta}^{\lambda_n} = \arg\min_{\beta} \frac{1}{2}\|Y - X\beta\|_{\ell_2}^2 + \lambda_n \sum_{j=1}^{p_n} \sqrt{d_j}\|\beta_j\|_{\ell_2} \qquad (1)$$

where λ_n is a positive number which penalizes complex model, and $\sqrt{d_j}$ is multiplied over each group to compensate for different group sizes.

The following proposition is directly obtained from (Yuan and Lin, 2006), which provides the Karush-Kuhn-Tucker (KKT) optimality condition for convex optimization problems.

Proposition 1 *The necessary and sufficient condition for $\widehat{\beta} = (\widehat{\beta}_1^T, \ldots, \widehat{\beta}_{p_n}^T)^T$ to be a solution to (1) is*

$$\begin{aligned} -X_j^T(Y - X\widehat{\beta}) + \frac{\lambda_n\sqrt{d_j}\widehat{\beta}_j}{\|\widehat{\beta}_j\|_{\ell_2}} = \mathbf{0}, \quad \forall \widehat{\beta}_j \neq \mathbf{0}, \qquad (2) \\ \|X_j^T(Y - X\widehat{\beta})\|_{\ell_2} \leq \lambda_n\sqrt{d_j}, \quad \forall \widehat{\beta}_j = \mathbf{0}. \end{aligned}$$

It is well-known that under mild conditions, the Lasso solution has no more than n nonzero entries even if $p_n > n$ (Osborne et al., 2000). This is no longer true for the group Lasso. However, a slightly different result can be obtained.

Lemma 1 *In equation (1) with $\lambda_n > 0$, a solution $\widehat{\beta}^{\lambda_n}$ exists such that the number of nonzero groups $|S(\widehat{\beta})|$ is upper bounded by n, the number of data points, where $S(\beta) = \{j : \widehat{\beta}_j \neq \mathbf{0}\}$.*

Proof: Suppose there is a solution $\widehat{\beta}$ which has $|S(\widehat{\beta})| > n$ number of nonzero groups, in the following we will show that we can always construct another solution $\widetilde{\beta}$ such that $|S(\widetilde{\beta})| = |S(\widehat{\beta})| - 1$.

Without loss of generality, we assume the first $|S(\widehat{\beta})|$ groups of variables in $\widehat{\beta}$ are nonzero, i.e. $\widehat{\beta}_j \neq \mathbf{0}$ for $j = 1, \ldots, |S(\widehat{\beta})|$. Since $X\widehat{\beta} = \sum_{j=1}^{|S(\widehat{\beta})|} X_j\widehat{\beta}_j \in \mathbf{R}^{n\times 1}$ and $|S(\widehat{\beta})| > n$, the set of vectors $X_1\widehat{\beta}_1, \ldots, X_{|S(\widehat{\beta})|}\widehat{\beta}_{|S(\widehat{\beta})|}$ are linearly dependent. No lose of generality assume

$$X_1\widehat{\beta}_1 = \alpha_2 X_2\widehat{\beta}_2 + \ldots + \alpha_{|S(\widehat{\beta})|}X_{|S(\widehat{\beta})|}\widehat{\beta}_{|S(\widehat{\beta})|}.$$

Now define $\widetilde{\beta}_j = \mathbf{0}$ for $j = 1$ and $j > |S(\widehat{\beta})|$, and $\widetilde{\beta}_j = (1+\alpha_j)\widehat{\beta}_j$ for $j = 2, \ldots, |S(\widehat{\beta})|$, and it is straightforward to check that $\widetilde{\beta}$ satisfies the KKT condition in Proposition 1 and thus is also a group Lasso solution.

Remark 1 *Even though the solution of the group Lasso may not be unique especially when $p > n$, a compact solution $\widehat{\beta}$ with $|S(\widehat{\beta})| \leq n$ can always be constructed as in the proof of Lemma 1.*

3 ℓ_2-Consistency of the Group Lasso

Recall that for linear models, an estimator $\widehat{\beta}$ is called ℓ_2-consistent if $\|\widehat{\beta}-\beta\|_{\ell_2}=o_P(1)$. In this section we obtain the ℓ_2-consistency result for the group Lasso estimator. The main result is Theorem 1, which builds an upper bound for the ℓ_2-distance $\|\widehat{\beta}-\beta\|_{\ell_2}$. Equation (17) from Corollary 1 establishes the concrete rate. Another result on the asymptotic upper bound of the number of selected variables and its implications is provided in Lemma 6 and Remark 2.

We mainly consider the case when $p_n \gg n$ but $s_n = |S(\beta)| = o(n)$, let $C=\frac{1}{n}X^TX$ be the sample covariance matrix, and we start with some definitions which are useful in the proof.

Definition 1 *The m-sparse minimum and maximum eigenvalues of C are $\phi_{\min}(m)=\min_{\beta:\|\beta\|_{\ell_0}\le m}\frac{\beta^TC\beta}{\beta^T\beta}$ and $\phi_{\max}(m)=\max_{\beta:\|\beta\|_{\ell_0}\le m}\frac{\beta^TC\beta}{\beta^T\beta}$. Also, denote $\phi_{\max}$ as $\phi_{\max}=\phi_{\max}\left((s_n+n)\bar{d}_n\right)$.*

Definition 2 *Denote $Y(\xi)=X\beta+\xi\epsilon$ as a de-noised model with level ξ ($0\le\xi\le 1$), we define*

$$\widehat{\beta}^{\lambda,\xi}=\arg\min_{\beta}\|Y(\xi)-X\beta\|_{\ell_2}^2+\lambda_n\sum_{j=1}^{p_n}\sqrt{d_j}\|\beta_j\|_{\ell_2} \quad (3)$$

to be the de-denoised estimator at noise level ξ.

The key assumption is given below, more detailed discussion of such a condition can be found in (Meinshausen and Yu, 2009).

Assumption 1 *There exists a positive sequence e_n, the so-called sparsity multiplier sequence, such that*

$$\liminf_{n\to\infty} e_n\phi_{\min}\left(e_n^2 s_n\right)\ge 18\phi_{\max}.$$

Theorem 1 (Convergence in ℓ_2-norm) *Under assumption 1 with a positive sequence e_n. If $\lambda_n \asymp \sigma e_n\sqrt{n\log m_n}$. For the group Lasso solution constructed in Lemma 1, There exists a constant $M>0$ such that, with probability tending to 1 for $n\to\infty$,*

$$\|\widehat{\beta}^{\lambda_n}-\beta\|_{\ell_2}^2\le M\frac{s_n\bar{d}_n\log m_n}{n}\frac{e_n^2}{\phi_{\min}^2(e_n^2 s_n\bar{d}_n)}.$$

Proof: Our proof extends (Meinshausen and Yu, 2009) to the case of grouped variables. Let $\beta^\lambda=\widehat{\beta}^{\lambda,0}$, the variance related part is $\|\widehat{\beta}^\lambda-\beta^\lambda\|_{\ell_2}^2$ and the bias related part is $\|\beta^\lambda-\beta\|_{\ell_2}^2$. The ℓ_2-consistency can be obtained by bounding the bias and variance terms. i.e.

$$\|\widehat{\beta}^\lambda-\beta\|_{\ell_2}^2\le 2\|\widehat{\beta}^\lambda-\beta^\lambda\|_{\ell_2}^2+2\|\beta^\lambda-\beta\|_{\ell_2}^2.$$

Let $K=\{k:\beta_k\neq\mathbf{0},k=1,...,p_n\}$ represent the set of index for all the groups with nonzero coefficient vectors. The cardinality of nonzero groups is again denoted by $s_n=|K|$. Then, the solution β^λ can, for each value of λ, be written as $\beta^\lambda=\beta+\gamma^\lambda$, where $\gamma^\lambda=\arg\min_\eta f(\eta)$ with $f(\eta)$ defined as

$$\begin{aligned} f(\eta) &= n\eta^TC\eta+\lambda\sum_{k\in K^c}\sqrt{d_k}\|\eta_k\|_{\ell_2} \\ &+ \lambda\sum_{k\in K}\sqrt{d_k}\left(\|\eta_k+\beta_k\|_{\ell_2}-\|\beta_k\|_{\ell_2}\right) \end{aligned} \quad (4)$$

The next lemma bound the ℓ_2-norm of γ^λ.

Lemma 2 *Under assumption 1 with a positive sequence e_n. The ℓ_2-norm of γ^{λ_n}, as defined in (4), is bounded for sufficiently large values of n by $\|\gamma^\lambda\|_{\ell_2}\le 17.5\lambda\sqrt{s_n\bar{d}_n}/(n\phi_{\min}(e_ns_n\bar{d}_n))$.*

Proof : For the notational simplicity, we use γ instead of γ^λ. Let $\gamma(K)$ be the vector with sub-vectors $\gamma_k(K)=\gamma_k\mathbf{1}\{k\in K\}$. That is, $\gamma(K)$ is the bias of the coefficients from the truly nonzero groups. And similarly for $\gamma(K^c)=\gamma_k\mathbf{1}\{k\notin K\}$. Therefore, we have that $\gamma=\gamma(K)+\gamma(K^c)$. Since $f(\eta)=0$ in (4) when $\eta=0$, we have $f(\gamma^\lambda)\le 0$. Together with the fact that $\eta^TC\eta\ge 0$ for any η, we have $\sum_{k\in K^c}\sqrt{d_k}\|\gamma_k\|_{\ell_2}\le\sum_{k\in K}\sqrt{d_k}\|\gamma_k\|_{\ell_2}$. Also, we have

$$\sum_{k\in K}\sqrt{d_k}\|\gamma_k\|_{\ell_2}\le\sqrt{\sum_{k\in K}d_k}\|\gamma(K)\|_{\ell_2}\le\sqrt{s_n\bar{d}_n}\|\gamma\|_{\ell_2} \quad (5)$$

From (5) and its previous inequality, we have

$$\sum_{k=1}^{p_n}\sqrt{d_k}\|\gamma_k\|_{\ell_2}\le 2\sqrt{s_n\bar{d}_n}\|\gamma\|_{\ell_2}. \quad (6)$$

Since $f(\gamma)<0$, and ignoring the non-negative term $\lambda\sum_{k\in K^c}\sqrt{d_k}\|\eta_k\|_{\ell_2}$, it follows that

$$n\gamma^TC\gamma\le\lambda\sqrt{s_n\bar{d}_n}\|\gamma\|_{\ell_2}. \quad (7)$$

Now, we bound the term $\gamma^TC\gamma$ from below and plugging the result into (7) will yield the desired upper bound on the ℓ_2-norm of γ. Let $\|\gamma_{(1)}\|_{\ell_2}\ge\|\gamma_{(2)}\|_{\ell_2}\ge\cdots\ge\|\gamma_{(p)}\|_{\ell_2}$ be the ordered block entries of γ. Let $\{u_n\}_{n\in\mathbf{N}}$ be a sequence of positive integers, such that $1\le u_n\le p_n$ and define the set of "u_n-largest groups" as $U=\{k:\|\gamma_k\|_{\ell_2}\ge\|\gamma_{(u_n)}\|_{\ell_2}\}$. Define analogously as before $\gamma(U)$ and $\gamma(U^c)$. The quantity $\gamma^TC\gamma$ can be written as $(\gamma(U)+\gamma(U^c))^TC(\gamma(U)+\gamma(U^c))=\|a+b\|_{\ell_2}^2$ with $a=X\gamma(U)/\sqrt{n}$ and $b=X\gamma(U^c)/\sqrt{n}$. Then

$$\gamma^TC\gamma=a^Ta+2b^Ta+b^Tb\ge\left(\|a\|_{\ell_2}-\|b\|_{\ell_2}\right)^2. \quad (8)$$

Further, we derive the bound for $\|\gamma(U^c)\|_{\ell_2}$ as a function of u_n. Aassuming that $\ell=\sum_{k=1}^p\|\gamma_k\|_{\ell_2}$, it holds

for every $k = 1, \cdots, p_n$ that $\gamma_{(k)} \leq \ell/k$. Therefore

$$\|\gamma(U^c)\|_{\ell_2}^2 \leq \left(\sum_{k=1}^{p_n} \|\gamma_k\|_{\ell_2}\right)^2 \sum_{k=u_n+1}^{p_n} \frac{1}{k^2} \leq \left(\sum_{k=1}^{p_n} \sqrt{d_k}\|\gamma_k\|_{\ell_2}\right)^2 \frac{1}{u_n}.$$

Further from (6), we get $\|\gamma(U^c)\|_{\ell_2}^2 \leq (4s_n\bar{d}_n\|\gamma\|_{\ell_2}^2)\frac{1}{u_n}$. Since $\gamma(U)$ has at most $\sum_{k\in U} d_k$ non-zero coefficients

$$\begin{aligned}\|a\|_{\ell_2}^2 &\geq \phi_{\min}\left(\sum_{k\in U} d_k\right)\left(\|\gamma\|_{\ell_2}^2 - \|\gamma(U^c)\|_{\ell_2}^2\right)^2 \\ &\geq \phi_{\min}\left(\sum_{k\in U} d_k\right)\left(1 - \frac{4s_n\bar{d}_n}{u_n}\right)\|\gamma\|_{\ell_2}^2. \quad (9)\end{aligned}$$

From Lemma 1, $\gamma(U^c)$ has at most n non-zero groups,

$$\|b\|_{\ell_2}^2 \leq \phi_{\max}(n\bar{d}_n)\|\gamma(U^c)\|_{\ell_2}^2 \leq \frac{4s_n\bar{d}_n}{u_n}\|\gamma\|_{\ell_2}^2. \quad (10)$$

Plugging (9) and (10) into (8), combine with the facts $\sum_{k\in U} d_k \leq \bar{d}_n u_n$ and $\phi_{\max} \geq \phi_{\min}(u_n)$, we have

$$\gamma^T C\gamma \geq \phi_{\min}(u_n\bar{d}_n)\|\gamma\|_{\ell_2}^2\left(1 - 4\sqrt{\frac{\bar{d}_n s_n \phi_{\max}}{u_n\phi_{\min}(u_n\bar{d}_n)}}\right). \quad (11)$$

Choose $u_n = e_n s_n\bar{d}_n$, from the assumption, we have that $(s_n\bar{d}_n\phi_{\max})/(e_n s_n\bar{d}_n\phi_{\min}(e_n^2 s_n\bar{d}_n)) < \frac{1}{18}$. From (7) and $\phi_{\min}(e_n^2 s_n\bar{d}_n) \leq \phi_{\min}(e_n s_n\bar{d}_n)$, we get

$$\frac{\lambda\sqrt{s_n\bar{d}_n}\|\gamma\|_{\ell_2}}{n} \geq \gamma^T C\gamma \geq \phi_{\min}(u_n\bar{d}_n)\|\gamma\|_{\ell_2}^2\left(1 - \sqrt{\frac{4}{18}}\right)$$

Therefore, $\|\gamma\|_{\ell_2} \leq 17.5\lambda\sqrt{s_n\bar{d}_n}/(n\phi_{\min}(e_n s_n\bar{d}_n))$, for large n. The proves the desired lemma. Q.E.D.

Under the assumption of the theorem, when $\lambda_n \asymp \sigma e_n\sqrt{n\log m_n}$ and the fact that $\phi_{\min}(e_n^2 s_n\bar{d}_n) \leq \phi_{\min}(e_n s_n\bar{d}_n)$, we immediately get that

$$\|\gamma^\lambda\|_{\ell_2}^2 \leq (17.5)^2\sigma^2\frac{s_n\bar{d}_n\log m_n}{n}\frac{e_n^2}{\phi_{\min}^2(e_n^2 s_n\bar{d}_n)}.$$

Next we bound the variance term. For every subset $M \subset \{1, \cdots, m_n\}$ with $|M| \leq n$, denote $\widehat{\theta}^M \in \mathbf{R}^{|M|}$ the restricted least square estimator of the noise ϵ,

$$\widehat{\theta}^M = (X_M^T X_M)^{-1} X_M^T \epsilon \quad (12)$$

The next lemma bounds the ℓ_2-norm of this estimator, which is useful to bound the variance of the group Lasso estimator.

Lemma 3 *Let $\bar{m}_n$ be a sequence with $\bar{m}_n = o(n)$ and $\bar{m}_n \to \infty$ for $n \to \infty$. Then it holds with probability converging to 1 for $n \to \infty$*

$$\max_{M:|M|\leq\bar{m}_n} \|\widehat{\theta}^M\|_{\ell_2}^2 \leq \frac{2\log m_n}{n}\frac{\bar{m}_n}{\phi_{\min}^2(\bar{m}_n)}\sigma^2.$$

The ℓ_2-norm of the restricted estimator $\widehat{\theta}^M$ is uniformly over all sets M with $|M| \leq \bar{m}_n$.

Proof : From (12), for every M with $|M| \leq \bar{m}_n$,

$$\|\widehat{\theta}^M\|_{\ell_2}^2 \leq \frac{1}{n^2\phi_{\min}^2(\bar{m}_n)}\|X_M^T\epsilon\|_{\ell_2}^2. \quad (13)$$

We want to show that, with probability tending to 1,

$$\max_{M:|M|\leq\bar{m}_n} \|X_M^T\epsilon\|_{\ell_2}^2 \leq 2\sigma^2\bar{m}_n n\log m_n.$$

Since $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, we have, with probability converging to 1, for $n \to \infty$, that $\max_{\underline{j}\in\{1,\ldots,m_n\}} |X_{\underline{j}}^T\epsilon|^2$ is bounded from above by $2\sigma^2 n\log m_n$. Therefore, with probability tending to 1 for $n \to \infty$,

$$\begin{aligned}\max_{M:|M|\leq\bar{m}_n} \|X_M^T\epsilon\|_{\ell_2}^2 &\leq \bar{m}_n \max_{\underline{j}\in\{1,\ldots,m_n\}} |X_{\underline{j}}^T\epsilon|^2 \\ &\leq 2\sigma^2\bar{m}_n n\log m_n.\end{aligned}$$

This proves the given lemma. Q.E.D.

For the following analysis, we define $\mathcal{A}_{\lambda,\xi}$ to be

$$\mathcal{A}_{\lambda,\xi} = \left\{j : \lambda\frac{\sqrt{d_j}\widehat{\beta}_j}{\|\widehat{\beta}_j\|_{\ell_2}} = X_j^T(Y(\xi) - X\widehat{\beta})\right\} \quad (14)$$

which represents the set of active groups for the denoised version problem in (3).

The following lemma shows that the variance of the group Lasso estimator can be bounded by the variances of the restricted OLS estimators $\widehat{\theta}^M$.

Lemma 4 *If, for a fixed value of λ, the number of active variables of the de-noised estimators $\widehat{\beta}^{\lambda,\xi}$ is for every $0 \leq \xi \leq 1$ bounded by m', then*

$$\|\widehat{\beta}^{\lambda,0} - \widehat{\beta}^\lambda\|_{\ell_2}^2 \leq C \cdot \max_{M:|M|\leq m'} \|\widehat{\theta}^M\|_{\ell_2}^2$$

with C as a generic constant.

Proof : A solution path approach as in (Meinshausen and Yu, 2009) is adopted. Let $M(\xi) \equiv \mathcal{A}_{\lambda,\xi}$ as in (14). Similar as the linear Lasso case, the estimator $\widehat{\beta}^{\lambda,\xi}$ and also the gradient $X_j^T(Y(\xi) - X\widehat{\beta})$ are continuous functions in both λ and ξ. Let $0 = \xi_1 < \cdots < \xi_{J+1} = 1$ be the points of discontinuity of $M(\xi)$. At these locations, variables either join the active set or are dropped from the active set.

Fix some j with $1 \le j \le J$. Denote by M_j be the set of active groups $M(\xi)$ for any $\xi \in (\xi_j, \xi_{j+1})$. Assuming

$$\forall \xi \in (\xi_j, \xi_{j+1}) : \|\widehat{\beta}^{\lambda,\xi} - \widehat{\beta}^{\lambda,\xi_j}\|_{\ell_2} \le C(\xi - \xi_j)\|\widehat{\theta}^{M_j}\|_{\ell_2} \quad (15)$$

is true, where $\widehat{\theta}^{M_j}$ is the restricted OLS estimator of noise, as in (12). The claim then follows from a piecewise bound of the difference of the de-noised solutions at different noise levels. That is

$$\begin{aligned} \|\widehat{\beta}^{\lambda,0} - \widehat{\beta}^{\lambda,1}\|_{\ell_2} &\le \sum_{j=1}^{J} \|\widehat{\beta}^{\lambda,\xi_j} - \widehat{\beta}^{\lambda,\xi_{j+1}}\|_{\ell_2} \\ &\le C \cdot \max_{M:|M|\le m} \|\widehat{\theta}^{M}\|_{\ell_2} \sum_{j=1}^{J} (\xi_{j+1} - \xi_j) \\ &= C \cdot \max_{M:|M|\le m} \|\widehat{\theta}^{M}\|_{\ell_2}. \end{aligned}$$

It thus remains to show the correctness of (15), which follows from an application of lemma 5, by replacing $\widehat{x}_1$, $\widehat{x}_2$, $\widehat{y}_1$ and $\widehat{y}_2$ with $\xi\widehat{\theta}^{M_j}$, $\xi_j\widehat{\theta}^{M_j}$, $\widehat{\beta}^{\lambda,\xi}$ and $\widehat{\beta}^{\lambda,\xi_j}$, respectively. The key observation is that the submatrix X_{M_j} should be full rank to make $\widehat{\theta}^{M_j}$ well-defineded. Q.E.D.

Lemma 5 *For $x \in \mathbf{R}^q$, Suppose $\widehat{x}_1 = \arg\min_x f_1(x)$ and $\widehat{x}_2 = \arg\min_x f_2(x)$ where $f_1(x) = \frac{1}{2}x^T Ax + b^T x$ with $A \in \mathbf{R}^{q\times q}$ positive definite and $b \in \mathbf{R}^q$. Also, $f_2(x) = f_1(x) + c^T x = \frac{1}{2}x^T Ax + b^T x + c^T x$ with $c \in \mathbf{R}^q$. Let $g_1(x) = f_1(x) + h(x)$ and $g_2(x) = f_2(x) + h(x)$ where $h(x)$ is a convex function with respect to x and everywhere subdifferentiable, and define $\widehat{y}_1 = \arg\min_y g_1(y)$ and $\widehat{y}_2 = \arg\min_y g_2(y)$. Then we have*

$$\|\widehat{y}_2 - \widehat{y}_1\|_{\ell_2} \le \gamma\|\widehat{x}_2 - \widehat{x}_1\|_{\ell_2}.$$

Proof : First we have $f_1'(\widehat{x}_1) = A\widehat{x}_1 + b = \mathbf{0}$ and $f_2'(\widehat{x}_2) = A\widehat{x}_2 + b + c = \mathbf{0}$, which leads to $\widehat{x}_2 - \widehat{x}_1 = -A^{-1}c$ and thus $\|\widehat{x}_2 - \widehat{x}_1\|_{\ell_2} = \|A^{-1}c\|_{\ell_2} \ge \lambda_{\min}(A^{-1})\|c\|_{\ell_2}$.

Since $\widehat{y}_1$ and $\widehat{y}_2$ are minimizers of $g_1(.)$ and $g_2(.)$ respectively, by the optimality condition we have $A\widehat{y}_1 + b + y_1^* = \mathbf{0}$ and $A\widehat{y}_2 + b + c + y_2^* = \mathbf{0}$, where y_1^* is a subgradient of h at $\widehat{y}_1$ and y_2^* is a subgradient of h at $\widehat{y}_2$. Since $h(.)$ is a convex function, we have

$$h(z_1) \ge h(\widehat{y}_1) + \langle y_1^*, z_1 - \widehat{y}_1\rangle$$

and

$$h(z_2) \ge h(\widehat{y}_2) + \langle y_2^*, z_2 - \widehat{y}_2\rangle$$

for arbitrary $z_1, z_2 \in \mathbf{R}^q$. By setting $z_1 = \widehat{y}_2$ and $z_2 = \widehat{y}_1$ and combining with the optimality condition we have $(\widehat{y}_2 - \widehat{y}_1)^T A(\widehat{y}_2 - \widehat{y}_1) + c^T(\widehat{y}_2 - \widehat{y}_1) \le 0$. It follows that

$$\begin{aligned} \lambda_{\min}(A)\|\widehat{y}_2 - \widehat{y}_1\|_{\ell_2}^2 &\le (\widehat{y}_2 - \widehat{y}_1)^T A(\widehat{y}_2 - \widehat{y}_1) \\ &\le -c^T(\widehat{y}_2 - \widehat{y}_1) \le \|c\|\|\widehat{y}_2 - \widehat{y}_1\|_{\ell_2} \end{aligned}$$

which implies that $\|\widehat{y}_2 - \widehat{y}_1\|_{\ell_2} \le \|c\|/\lambda_{min}(A)$. Finally by combining the upper and lower bounds of $\|c\|$ we have $\|\widehat{y}_2 - \widehat{y}_1\|_{\ell_2} \le \gamma\|\widehat{x}_2 - \widehat{x}_1\|_{\ell_2}$ with $\gamma = (\lambda_{\min}(A)\lambda_{\min}(A^{-1}))^{-1}$. Q.E.D.

The next lemma provides an asymptotic upper bound on the number of selected variables, the proof of which is similar to Lemma 4 in (Meinshausen and Yu, 2009).

Lemma 6 (Bound on # of selected variables) *For $\lambda \ge \sigma e_n\sqrt{n\log m_n}$, the maximal number of selected variables, $\sup_{0\le\xi\le 1}\sum_{k\in\mathcal{A}_{\lambda,\xi}} d_k$, is bounded, with probability converging to 1 for $n \to \infty$, by*

$$\sup_{0\le\xi\le 1}\sum_{k\in\mathcal{A}_{\lambda,\xi}} d_k \le e_n^2 s_n \bar{d}_n. \quad (16)$$

Follow from Lemmas 3, 4, and 6, the next lemma bounds the variance part of the group Lasso estimator:

Lemma 7 *Under the conditions of Theorem 1, with probability tending to 1 for $n \to \infty$*

$$\|\beta^\lambda - \widehat{\beta}^\lambda\|_{\ell_2}^2 \le 2C\sigma^2 \frac{s_n\bar{d}_n \log m_n}{n} \frac{e_n^2}{\phi_{\min}^2(e_n^2 s_n \bar{d}_n)}.$$

From all above, the proof of theorem 1 finishes by combining Lemma 2 and Lemma 7. Q.E.D.

Corollary 1 (rate of convergence) *Let the assumptions of Theorem 1 be satisfied. Assume that there exist constants $0 < \kappa_{\min} \le \kappa_{\max} < \infty$ such that $\liminf_{n\to\infty} \phi_{\min}(s_n\bar{d}_n \log n) \ge \kappa_{\min}$ and $\limsup_{n\to\infty} \phi_{\max} \le \kappa_{\max}$. Then, for $\lambda \asymp \sigma\sqrt{n\log m_n}$, there exists a constant $M > 0$ such that, with probability tending to 1 for $n \to \infty$,*

$$\|\beta - \widehat{\beta}^{\lambda_n}\|_{\ell_2}^2 \le M\sigma^2 \frac{s_n\bar{d}_n \log m_n}{n}. \quad (17)$$

Proof: directly follows from theorem 1 by choosing a constant positive sequence e_n. Q.E.D.

Remark 2 *Lemma 6 implies that, with high probability, at most $e_n^2 s_n \bar{d}_n$ variables will be chosen by the group Lasso estimator. When $e_n = \log n$, we see that, up to a logarithm factor, the selected variables is of the same order of magnitude as the number of true nonzero coefficients, which is $s_n\bar{d}_n$. We also see that under the conditions of Corollary 1, the group Lasso estimator is ℓ_2-consistent if $s_n\bar{d}_n \log m_n/n \to 0$. It implies that, if $s_n = O(1)$, up to a logarithmic factor, the number of variables within each group can increase almost as fast as sample size n. If we have $s_n\bar{d}_n = O(1)$, the number of groups p_n can increase almost as fast as $o(\exp(n))$.*

4 Applications of the Main Results

4.1 Group Nonnegative Garrote

One application of the previous result is that when using the group Lasso solution as the initial estimator, we can build a two-step estimator that is both estimation and variable selection consistent by applying the nonnegative garotte procedure. The main result here is adapted from (Yuan and Lin, 2007), in which they assume the initial estimator is ℓ_∞-norm consistent and derive the result for fixed dimensionality. In our case, Theorem 1 guarantees ℓ_2-norm consistency, which is stronger than the ℓ_∞-norm consistency as in (Yuan and Lin, 2007). As a consequence, we achieve a variable selection consistency result with increasing dimensions using the following defined group nonnegative garrote procedure.

Definition 3 (Group nonnegative garrotte) *Assuming $\widehat{\beta}^{\text{init}} = \widehat{\beta}^{\lambda}$ is the group Lasso solution obtained from (1). Define $Z_j = X_j\widehat{\beta}_j^{\lambda}$. For some $\gamma_n > 0$, the group nonnegative garotte estimator is defined as*

$$\widehat{\alpha}^{\mathbf{NG}}(\gamma) = \arg\min_{\alpha} \|Y - Z\alpha\|_{\ell_2}^2 + \gamma_n \sum_{j=1}^{p_n} \alpha_j \tag{18}$$

where $\alpha_j \geq 0, j = 1, \ldots, p_n$.

Equation (18) is a quadratic programming problem with the polyhedral type constraint regions, therefore the whole solution path can be solved efficiently (Yuan and Lin, 2007). Theorem 2 establishes the variable selection consistency result for the group nonnegative garrote estimators with increasing dimensions. The proof can be found in (Zhang et al., 2008).

Theorem 2 *Under the conditions of Theorem 1, when applying the group nonnegative garotte estimator as in (18) to the group Lasso solution from (1), there exists a sequence of γ_n, such that*

$$\mathbf{P}\left(S(\widehat{\alpha}^{\mathbf{NG}}(\gamma)) = S(\beta)\right) \to 1. \tag{19}$$

and the final estimator $\widetilde{\beta} \equiv (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_{p_n})^T$ with $\widetilde{\beta}_j = \widehat{\beta}_j\widehat{\alpha}_j^{NG}(\gamma)$ is also estimation consistent.

4.2 Sparse Additive Models

Sparse additive models (or SpAM) are first introduced by Ravikumar et al. (2007). These models combine the smoothness assumptions in nonparametric regression with sparsity assumptions in high dimensional linear models. Consider an additive model $Y_i = \sum_{j=1}^{p_n} f_j(X_{ij}) + \epsilon_i$ where each function f_j can be represented by an orthonormal basis $\mathcal{B}_j = \{\psi_{j1}, \psi_{j2}, \ldots\}$ for the second-order Sobolev space $\mathcal{H}_j$. If we assume the true model is sparse, i.e. only a small number of component functions are nonzero, then using a truncated basis of size d_n, a version of the SpAM estimate f_j as $\widehat{f}_j(x_j) = \sum_{k=1}^{d_n} \widehat{\beta}_{jk}\psi_k(x_j)$, where $\widehat{\beta}_{jk}$ is the solution to the minimization problem

$$\min_{\beta} \sum_{i=1}^{n} \left\{ \left(Y_i - \sum_{j=1}^{p_n} \sum_{k=1}^{d_n} \beta_{jk}\psi_{jk}(X_{ij}) \right)^2 + \lambda_n \sum_{j=1}^{p_n} \sqrt{\sum_{k=1}^{d_n} \beta_{jk}^2} \right\} \tag{20}$$

where λ_n is the regularization parameter. Let β_j be the d_n dimensional vector $\{\beta_{jk}, k = 1, \ldots, d_n\}$, and Ψ_j the $n \times d_n$ matrix, $\Psi_j[i,k] = \psi_{jk}(X_{ij})$, equation (20) can be written as a group Lasso problem

$$\widehat{\beta} = \arg\min_{\beta} \|Y - \sum_{j=1}^{p_n} \Psi_j\beta_j\|_{\ell_2}^2 + \lambda_n^* \sum_{j=1}^{p_n} \sqrt{d_n}\|\beta_j\|_{\ell_2} \tag{21}$$

with $\lambda_n^* = \lambda_n/\sqrt{d_n}$. This estimator is essentially nonparametric since d_n should increase with the sample size. From Theorem 1 and Corollary 1 , we can obtain the following consistency result (the proof is omited).

Theorem 3 (Consistency of SpAM) *Assuming the the number of nonzero component functions $s_n = O(1)$, let $d_n = O(n^{1/5})$, $p_n = O(n^{4/5})$ and the true function $f = \sum_j f_j$ with each f_j in the 2-nd order Sobolev space. Choosing $\lambda_n \asymp \sigma\sqrt{\frac{\log p + \log n}{n}}$, we have*

$$\|\widehat{f} - f\|_{L_2}^2 = O_P\left(\frac{\log p_n}{n^{4/5}}\right) \tag{22}$$

where $\widehat{f} = \sum_{j=1}^{p_n} \widehat{f}_j = \sum_{j=1}^{p_n} \Psi_j\widehat{\beta}_j$ with $\widehat{\beta}_j$ from (21).

Definition 4 (Nonnegative garrote SpAM) *For sparse additive models, when using the solution to Equation (21) as an initial estimator and apply the group nonnegative garrote procedure, the final estimator $\widetilde{f} = \sum_{j=1}^{p_n} \Psi_j\widetilde{\beta}_j$ with $\widetilde{\beta}_j$ as in Theorem 2 is called the nonnegative garrote SpAM (or Ng-SpAM).*

From Theorems 2 and 3, it's obvious that Ng-SpAM is both estimation and variable selection consistent.

5 Experimental Results

In this section, we report experiments on both synthetic and real datasets. They provide empirical evidence to our theory as well as to the superior finite-sample performance of the group nonnegative garrote and the Ng-SpAM estimators.

Experiment 1 (Group nonnegative garrote) In this simulation, we compare the group Lasso and the group nonnegative garrote in high-dimensional problems. We use a similar setting as in Zhao and Yu (2007) by taking different (n, p, d, s) combinations with

each of them representing sample size, number of groups, number of variables per group, and the number of nonzero groups. For each (n, p, d, s) combination, we sample 100 times the covariance matrix Σ from a Wishart distribution $\mathsf{Wishart}(pd, I_{pd})$ and the true parameter vector β_j for the j-th nonzero group is $(8 \cdot (0.5)^{j-1}, \ldots, 8 \cdot (0.5)^{j-1})^T$. For each Σ we sample a design matrix X from the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \Sigma)$. The response vector $Y = X\beta + \epsilon$ is then calculated using $\epsilon \sim \mathcal{N}(0, (\sqrt{0.6})^2)$. The noise level σ^2 is set to 0.6 to manifest the asymptotic characterizations.

For each design, 1000 simulations are conducted by repeatedly generating the noise vectors. For the group nonnegative garrote we use the group Lasso as the initial estimator, for which the tuning parameter λ_n is automatically chosen such that there are exactly $\min\{n, p\} - 1$ nonzero groups kept. The tuning parameter γ_n for the second step is chosen optimally over the solution path to find the correct model if possible. For the group Lasso we also select its optimal tuning parameter λ_n^* by searching over the whole path. The advantage of using such "oracle values" is that the simulation results will only depend on different methods.

Table 1: (Experiment 1) variable selection accuracy

(n, p, d, s)	gLasso(sd)	gNg(sd)
$(100, 10, 4, 2)$	0.9917 (0.0270)	1.000 (0.00)
$(100, 16, 5, 3)$	0.9808 (0.0343)	1.000 (0.00)
$(100, 32, 4, 3)$	0.9619 (0.0810)	1.000 (0.00)
$(100, 40, 4, 5)$	0.7068 (0.1121)	1.000 (0.00)
$(100, 50, 4, 6)$	0.4111 (0.1295)	1.000 (0.00)

The results are reported in Table 1, where gLasso represents the group Lasso and gNg represents the group nonnegative garrote. For each (n, p, d, s) combination the mean percentage of the variable selection accuracy over all designs and the corresponding standard deviations are listed. It's obvious that variable selection accuracy of the group Lasso decreases with the increase of p, d and s. This is consistent with the Lasso result as in Zhao and Yu (2007). As a contrast, the group nonnegative garrote achieves a perfect variable selection performance. This suggests that the initial group Lasso estimator are reasonably good in the estimation sense. Since the result from Corollary 1 is a rate argument, it's impossible to verify it quantitatively. We will provide some qualitative justifications based on the nonnegative garrote regularization paths. This is shown in the next experiment. The statistical significance of these results is justified by a paired two-sample t-test.

Experiment 2 (Sparse additive models) We generate $n = 100$ observations from a 10-dimensional sparse additive model with four relevant variables: $Y_i = \sum_{j=1}^{4} f_j(X_{ij}) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 1)$ and $f_1(x) = \exp(-2x)$, $f_2(x) = (2x - 1)^2$, $f_3(x) = \frac{\sin(2\pi x)}{2 - \sin(2\pi x)}$, and $f_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.2\sin^2(2\pi x) + 0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x)$. The covariates are generated as $X_j = (W_j + tU)/(1 + t)$, $j = 1, \ldots, 10$ where $W_1, \ldots, W_{10}$ and U are i.i.d. sampled from Uniform$(-2.5, 2.5)$. Thus, the correlation between X_j and $X_{j'}$ is $t^2/(1 + t^2)$ for $j \neq j'$. Altogether 100 designs with 1000 simulations per design are generated.

t	gLasso − SpAM(sd)	Ng − SpAM(sd)
$t = 0.0$	0.9991 (0.0021)	1.0000 (0.0000)
$t = 0.5$	0.9942 (0.0137)	1.0000 (0.0000)
$t = 1.0$	0.9835 (0.0208)	1.0000 (0.0000)
$t = 1.5$	0.9597 (0.0493)	1.0000 (0.0000)
$t = 2.0$	0.9390 (0.0530)	0.9999 (0.0003)
$t = 2.5$	0.8481 (0.0722)	0.9982 (0.0007)
$t = 3.0$	0.7488 (0.0993)	0.9856 (0.0020)

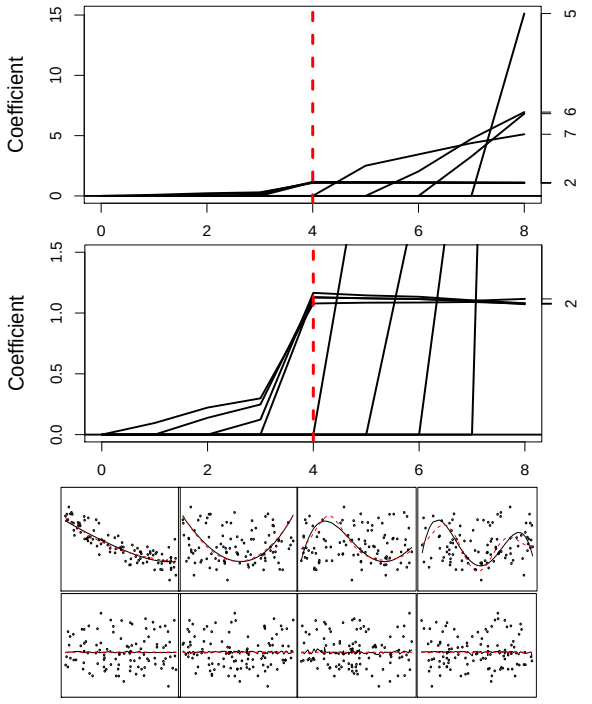

Figure 1: Experiment 2 results: (Upper) Variable selection accuracy table; (Middle) A typical nonnegative garrote regularization path ($t = 2.0$) and its zoomed-in plot on the range [0,1.5], the vertical dashed line separates the relevant and irrelevant variabls; (Lower) The group Lasso fitted component functions (solid curve) and the truth (dashed curve) for the fist 8 dimensions ($t = 2.0$).

For each covariate, we expand it using the Legendre basis and directly applies the group Lasso as in Equation (21) with the group size $d = 4$. This method is

denoted as glasso-SpAM. Choosing the tuning parameter such that there are exactly 8 groups are nonzero, we can apply the group nonnegative garrote on this initial estimator and denote the obtained estimator as Ng-SpAM. The results are reported in Figure 1. The upper variable selection accuracy table is consistent with the previous experiment: the glasso-SpAM performs worse when the correlations become larger. The two middle path plots are from one typical run when $t = 2$, in which glasso-SpAM fails to correctly identify the true model but Ng-SpAM succeeds. From Equation (18), if $\gamma_n = 0$, the final estimator obtained from the group nonnegative garrote corresponds to directly calculating the ordinary least square (OLS) solution using the basis expanded design. From the full and the zoomed-in paths in Figure 1, we see that the nonnegative garrote coefficients for the 4 relevant dimensions finally goes to 1, while those for the irrelevant dimensions shoot to much larger values. This suggests that the initial glasso-SpAM solution, though overselects, is very close to the OLS solution by directly regressing the response on the 4 relevant groups. The lower fitted component function plots also confirm this, with the scales of the 4 irrelevant dimensions being very small.

Experiment 3: (Boston Housing Data) We apply the Ng-SpAM to the corrected Boston Housing data in Ravikumar et al. (2007). The dataset contains 506 records about housing prices in suburbs of Boston. Each record has 10 continuous features which might be useful in describing housing price, and the response variable is the median house price. We consider a sparse additive model and using exactly the same experimental setup as in Ravikumar et al. (2007) except that their methods are replaced by the Ng-SpAM. The results are reported in Figure 2.

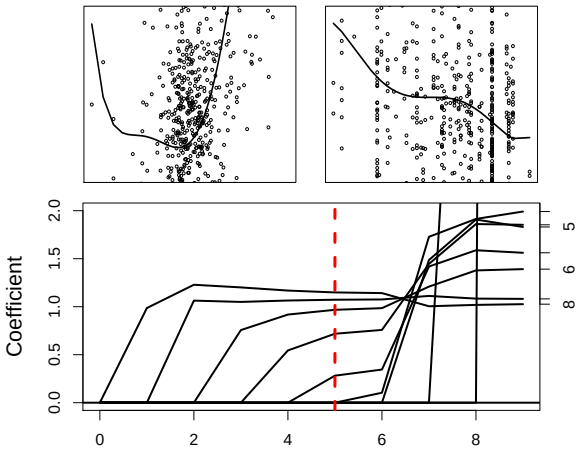

Figure 2: (Experiment 3)The Boston Housing data (Upper) the fitted component functions for the variables rm and lsat; (Lower) The regularization path with the dashed vertical line represents the 5-fold CV cutpoint.

The upper two plots show the fitted component functions for the variables rm and lsat, their shapes are very close to those obtained in Ravikumar et al. (2007). The lower plot illustrates the group nonnegative garrote path with the dashed vertical line represents the model selected by the 5-fold cross-validation. Altogether 5 variables are selected by our method: rm, lsat, ptratio, crim and nox, which is consistent with Ravikumar et al. (2007) except that they treat nox as borderline important. More detailed results and comparisons will be reported elsewhere.

Acknowledgements

This research was supported in part by NSF grant CCF-0625879.

References

F. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9: 1179–1225, 2008.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific and Statistical Computing*, 20:33–61, 1998.

L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B, Methodological*, 70:53–71, 2007.

N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34:1436–1462, 2006.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data". *The Annals of Statistics*, 37(1):246–270, 2009.

G. Obozinski, M. Wainwright, and M. Jordan. High-dimensional union support recovery in multivariate regression. *Advances in Neural Information Processing Systems (NIPS)*, 2008.

M. R. Osborne, B. Presnell, and B. A. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9:319–337, 2000.

P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. Spam: Sparse additive models. In *Advances in Neural Information Processing Systems 20*. MIT Press, 2007.

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

M. Yuan and Y. Lin. On the non-negative garrote estimator. *Journal of the Royal Statistical Society, Series B, Methodological*, 69:143–161, 2007.

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

J. Zhang, X. Jeng, and H. Liu. Some two-step procedures for variable selection in high- dimensional linear regression. *Technical report, Purdue University*, 2008.

P. Zhao and B. Yu. On model selection consistency of lasso. *J. of Mach. Learn. Res.*, 7:2541–2567, 2007.