REGRESSION PERFORMANCE OF GROUP LASSO FOR ARBITRARY DESIGN MATRICES

Marco F. Duarte,[1,*] *Waheed U. Bajwa,*[2,*] *and Robert Calderbank*[1,2]

[1]Department of Computer Science
[2] Department of Electrical and Computer Engineering
Duke University, Durham, NC 27708
E-mails: {marco.duarte,w.bajwa,robert.calderbank}@duke.edu

ABSTRACT

In many linear regression problems, explanatory variables are activated in groups or clusters; group lasso has been proposed for regression in such cases. This paper studies the non-asymptotic regression performance of group lasso using ℓ_1/ℓ_2 regularization for arbitrary (random or deterministic) design matrices. In particular, the paper establishes under a statistical prior on the set of nonzero coefficients that the ℓ_1/ℓ_2 group lasso has a near-optimal regression error for all but a vanishingly small set of models. The analysis in the paper relies on three easily computable metrics of the design matrix – coherence, block coherence, and spectral norm. Remarkably, under certain conditions on these metrics, the ℓ_1/ℓ_2 group lasso can perform near-ideal regression even if the model order scales almost linearly with the number of rows of the design matrix. This is in stark contrast with prior work on the regression performance of the ℓ_1/ℓ_2 group lasso that only provides linear scaling of the model order for the case of random design matrices.

Keywords— Group sparsity, linear regression, group lasso, coherence, block coherence

1. INTRODUCTION

The lasso [18] and group lasso [22] are popular algorithms in the signal processing and statistics communities. In signal processing, these algorithms allow for efficient sparse approximations of arbitrary signals in overcomplete dictionaries. In statistics, they facilitate efficient variable selection and reliable regression under the linear model assumption. In both cases, there is now ample empirical evidence to suggest that an appropriately regularized group lasso can outperform the lasso whenever there is a natural grouping of the dictionary elements/regression variables in terms of their contributions to the observations [1, 22].

Our goal in this paper is to analytically characterize the regression performance of the group lasso algorithm using ℓ_1/ℓ_2 regularization for the case in which one can have far more regression variables than observations. Analytical characterization of group lasso in this "underdetermined" setting has received some attention lately in the statistics literature [1, 13–16]. However, prior analytical work on the performance of group lasso either studies an asymptotic regime [1, 14–16], focuses on random design matrices [1, 15], and/or relies on metrics that are computationally expensive to evaluate [13, 14, 16]. Recently, Candés and Plan [4] successfully circumvented somewhat similar shortcomings of the performance analysis for the lasso by imposing a probabilistic model on the vector of regression coefficients. Specifically, [4] showed that under mild, computable conditions on arbitrary (random or deterministic) design matrices, the lasso can perform near-optimally in terms of the regression error with very high probability for the following model: (i) locations of the nonzero regression coefficients are chosen uniformly at random; (ii) "signs" of nonzero regression coefficients are statistically independent; and (iii) nonzero regression coefficients have zero median.

In this paper, we study the regression performance of the group lasso algorithm using ℓ_1/ℓ_2 regularization in the underdetermined case under a generalization of the probabilistic framework of [4] to the group case. Specifically, our framework assumes that: (i) locations of the groups of nonzero regression coefficients are chosen uniformly at random; (ii) "directions" of the groups of nonzero regression coefficients are statistically independent; and (iii) nonzero regression coefficients have zero median. Our main contribution here is proving under this model that the group lasso[1] can also perform near-optimally in terms of the regression error with very high probability under mild, computable conditions on arbitrary design matrices. To the best of our knowledge, these are the first results for group lasso that are non-asymptotic in nature, applicable to arbitrary design matrices through easily computable metrics, and still allow for near-optimal scaling of the number of observations with the number of groups of nonzero regression coefficients. Our proof techniques are natural extensions of the ones used in [4] for the lasso and rely on our recent result concerning the conditioning of random block-subdictionaries of matrices [2], an extension of a result by Tropp [20] that facilitated the analysis in [4].

This paper is organized as follows. Section 2 provides background and notation. Section 3 provides our result and Section 4

The work is supported in part by NSF under grant DMS-0914892, by ONR under grant N00014-08- 1-1110, and by AFOSR under grants FA9550-09-1-0643 and FA9550-09-1- 0551. MFD is also supported by NSF Supplemental Funding DMS-0439872 to UCLA-IPAM, P.I. R. Caflisch.

* Equal contribution; author ordering was determined by a fair coin toss.

[1]We refer to the group lasso algorithm using ℓ_1/ℓ_2 regularization as "group lasso" throughout the rest of the paper for brevity.

contrasts our result with related prior work.

2. BACKGROUND AND NOTATION

We consider a vector of observations $y \in \mathbb{R}^n$ corresponding to the classical linear model $y = X\beta + z$, where X denotes the design matrix containing one regression variable per column, β denotes the vector of regression coefficients for these variables, and z denotes the modeling error. Here, we assume (without loss of generality) that X has unit-norm columns and we treat z as an independent and identically distributed (i.i.d.) Gaussian vector with variance σ^2.

The key distinguishing feature of our model is that we assume there is a natural grouping of the regression variables. For the sake of exposition, we consider p equal-sized groups of the regressors, leading to the block representation $\beta = [\beta_1^T \ \beta_2^T \ \dots \ \beta_p^T]^T$, where $\beta_i \in \mathbb{R}^m$, $1 \leq i \leq p$, denote different groups of size m in β. We define the $\ell_{q,r}$ norm of a vector $\beta \in \mathbb{R}^{pm}$ containing p blocks of size m entries each as

$$\|\beta\|_{q,r} = \left(\sum_{i=1}^{p} \|\beta_i\|_q^r \right)^{1/r},$$

with the standard modification for $q, r = \infty$. The group lasso solution for estimating β from y under this setup can then be written as [22]

$$\widehat{\beta} = \arg \min_{\beta \in \mathbb{R}^{pm}} \frac{1}{2}\|y - X\beta\|_2^2 + 2\lambda\sigma\sqrt{m}\|\beta\|_{2,1}.$$

3. REGRESSION PERFORMANCE OF GROUP LASSO

In this section, we provide performance guarantees for group lasso for the underdetermined case, $n < pm$, using the metric of regression error: $\|X\beta - X\widehat{\beta}\|_2$. In order to make this problem well-posed and tractable, we assume that the vector of regression coefficients $\beta \in \mathbb{R}^{pm}$ is k-block sparse with $\#\{i : \beta_i \neq \mathbf{0}\} = k \ll p$ and we impose a statistical prior on β. Specifically, we assume that: (i) block support of β, $I = \{i : \beta_i \neq \mathbf{0}\}$, has a uniform distribution over all k-subsets of $\{1, \dots, p\}$; (ii) "directions" of the nonzero blocks of β are statistically independent: $\mathbb{P}\left(\bigcap_{i \in I} \overline{\text{sign}}(\beta_i) \in \mathcal{A}_i\right) = \prod_{i \in I} \mathbb{P}\left(\overline{\text{sign}}(\beta_i) \in \mathcal{A}_i\right)$, where $\overline{\text{sign}}(\beta_i) = \beta_i/\|\beta_i\|_2$ denotes the unit-norm vector pointing in the direction of β_i in $\mathbb{R}^m$; and (iii) nonzero regression coefficients have zero median: $\mathbb{E}(\text{sign}(\beta)) = \mathbf{0}$, where $\text{sign}(\cdot)$ denotes the entry-wise sign operator.

The main result of this paper relies on three easily computable metrics of the design matrix, namely, coherence, block coherence, and spectral norm of X. The coherence of a matrix $X \in \mathbb{R}^{n \times pm}$ with unit norm columns is defined as

$$\mu = \max_{1 \leq i, i' \leq p, 1 \leq j, j' \leq m, (i,j) \neq (i',j')} |\langle X_{ij}, X_{i'j'} \rangle|,$$

where X_{ij} denotes the j^{th} column of the i^{th} block of $X = [X_1 \ \dots \ X_p]$. Similarly, the block coherence of X is defined as

$$\mu_B = \max \left\{ \max_{1 \leq i, i' \leq p, i \neq i'} \|X_i^* X_{i'}\|_2, \max_{1 \leq i \leq p} \|X_i^* X_i - I\|_2 \right\},$$

where X_i denotes the i^{th} block of X. Note here that X_i^* denotes the adjoint of X_i rather than a submatrix of X^*. We now state our main theorem, which is motivated by the analysis in [4].

Theorem 1. *Suppose that the vector of regression coefficients β is drawn according to the statistical model described earlier. If $k \leq C_0 p/\|X\|_2^2 \log(pm)$, and the matrix X satisfies $\mu \leq 1/m$ and $\mu_B \leq C_1/\log(pm)$ for some positive numerical constants C_0 and C_1, then the group lasso estimate $\widehat{\beta}$ computed with $\lambda = \sqrt{2\log(pm)}$ obeys*

$$\|X\beta - X\widehat{\beta}\|_2^2 \leq Cmk\sigma^2 \log(pm)$$

with probability at least $1 - (pm)^{-1}(2\pi \log(pm))^{-1/2} - 8(pm)^{-2\log 2}$. Here, $C > 0$ is a constant independent of the problem parameters.

Proof. We mirror the procedure of the proof of Theorem 1.2 in [4]. The proof uses the following lemma, proven in [7].

Lemma 1. *The group lasso estimate obeys*

$$\|X^*(y - X\widehat{\beta})\|_{2,\infty} \leq 2\lambda\sigma\sqrt{m}.$$

We also borrow the following theorem from [2].

Theorem 2. *Define random variables $\delta_1, \dots, \delta_p$ that are independent and identically distributed (i.i.d.) Bernoulli with parameter $\delta := k/p$, and form a block subdictionary $X_{I'} = [X_i : \delta_i = 1]$. Then, for $q = 2\log(pm)$, we have the bound*

$$[\mathbb{E}\|X_{I'}^* X_{I'} - \text{Id}\|_2^q]^{1/q} \leq 20\mu_B \log(pm) + \delta\|X\|_2^2 + 9\sqrt{\delta \log(pm)(1 + (m-1)\mu)}\|X\|_2.$$

We assume that $\sigma = 1$ without loss of generality and establish three conditions that together imply the theorem:

- *Invertibility.* The submatrix $X_I^* X_I$ is invertible and obeys $\|(X_I^* X_I)^{-1}\|_2 \leq 2$.
- *Orthogonality.* The vector z obeys $\|X^* z\|_{2,\infty} \leq \sqrt{2m} \cdot \lambda$.
- *Complementary size.* The following inequality holds:

$$2\lambda\sqrt{m}\|X_{I^C}^* X_I (X_I^* X_I)^{-1} \overline{\text{sign}}(\beta_I)\|_{2,\infty} + \|X_{I^C}^* X_I (X_I^* X_I)^{-1} X_I^* z\|_{2,\infty} \leq (2 - \sqrt{2})\lambda\sqrt{m}.$$

It is possible to show that the three conditions hold with the specified probability [7]. Since $\widehat{\beta}$ minimizes the group lasso objective function, we must have

$$\frac{1}{2}\|y - X\widehat{\beta}\|_2^2 + 2\lambda\sqrt{m}\|\widehat{\beta}\|_{2,1} \leq \frac{1}{2}\|y - X\beta\|_2^2 + 2\lambda\sqrt{m}\|\beta\|_{2,1}.$$

Set $h = \widehat{\beta} - \beta$, and note that

$$\|y - X\widehat{\beta}\|_2^2 = \|(y - X\beta) - Xh\|_2^2 = \|Xh\|_2^2 + \|y - X\beta\|_2^2 - 2\langle Xh, y - X\beta \rangle.$$

Plugging this identity with $z = y - X\beta$ into the above inequality and rearranging the terms gives

$$\frac{1}{2}\|Xh\|_2^2 \leq \langle Xh, z\rangle + 2\lambda\sqrt{m}(\|\beta\|_{2,1} - \|\widehat{\beta}\|_{2,1}). \quad (1)$$

Next, break up h into h_I and $h_{I^C} = \widehat{\beta}_{I^C}$ and rewrite (1) as

$$\frac{1}{2}\|Xh\|_2^2 \leq 2\lambda\sqrt{m}(\|\beta_I\|_{2,1} - \|\beta_I + h_I\|_{2,1} - \|h_{I^C}\|_{2,1}) + \langle h, X^* z\rangle. \quad (2)$$

For each $i \in I$, we have

$$\|\widehat{\beta}_i\|_2 = \|\beta_i + h_i\|_2 \geq \|\beta_i\|_2 - \frac{\langle h_i, \beta_i\rangle}{\|\beta_i\|_2} \geq \|\beta_i\|_2 - \left\langle h_i, \frac{\beta_i}{\|\beta_i\|_2}\right\rangle \geq \|\beta_i\|_2 - \langle h_i, \overline{\mathrm{sign}}(\beta_i)\rangle.$$

This is due to the projection of h_i on $\mathrm{span}\{\beta_i\}$ having magnitude $\frac{\langle h_i, \beta_i\rangle}{\|\beta_i\|_2}$. Thus, we can write $\|\widehat{\beta}_I\|_{2,1} \geq \|\beta_I\|_{2,1} - \langle h_I, \overline{\mathrm{sign}}(\beta_I)\rangle$. Merging this inequality with (2) gives us

$$\begin{aligned}\frac{1}{2}\|Xh\|_2^2 &\leq \langle Xh, z\rangle + 2\lambda\sqrt{m}(\langle h_I, \overline{\mathrm{sign}}(\beta_I)\rangle - \|h_{I^C}\|_{2,1}), \\ &= \langle h, X^* z\rangle + 2\lambda\sqrt{m}(\langle h_I, \overline{\mathrm{sign}}(\beta_I)\rangle - \|h_{I^C}\|_{2,1}), \\ &= \langle h_I, X_I^* z\rangle + \langle h_{I^C}, X_{I^C}^* z\rangle \\ &\quad + 2\lambda\sqrt{m}(\langle h_I, \overline{\mathrm{sign}}(\beta_I)\rangle - \|h_{I^C}\|_{2,1}). \end{aligned} \quad (3)$$

We will need a brief lemma extending Hölder's inequality to the block norms defined earlier. Its proof is a simple exercise.

Lemma 2. *For all $a, b \in \mathbb{R}^{pm}$, $\langle a, b\rangle \leq \|a\|_{2,1}\|b\|_{2,\infty}$.*

The orthogonality condition and the lemma above implies

$$\langle h_{I^C}, X_{I^C}^* z\rangle \leq \|h_{I^C}\|_{2,1}\|X_{I^c} z\|_{2,\infty} \leq \sqrt{2m}\cdot\lambda\|h_{I^C}\|_{2,1}.$$

Merging this result with (3) results in

$$\frac{1}{2}\|Xh\|_2^2 \leq \langle h_I, v\rangle - (2-\sqrt{2})\lambda\sqrt{m}\|h_{I^C}\|_{2,1}, \quad (4)$$

where $v = X_I^* z - 2\lambda\sqrt{m}\cdot\overline{\mathrm{sign}}(\beta_I)$. We aim to bound each of the terms on the right hand side independently. For the first term, we have

$$\begin{aligned}\langle h_I, v\rangle &= \langle (X_I^* X_I)^{-1} X_I^* X_I h_I, v\rangle = \langle X_I^* X_I h_I, (X_I^* X_I)^{-1} v\rangle \\ &= \langle X_I^* X_I h, (X_I^* X_I)^{-1} v\rangle - \langle X_I^* X_I h_{I^C}, (X_I^* X_I)^{-1} v\rangle.\end{aligned}$$

Denote the two terms on the right hand side as A_1 and A_2, respectively. For A_1 we use Lemma 2 to obtain

$$A_1 \leq \|(X_I^* X_I)^{-1} v\|_{2,1}\|X_I^* X_I h\|_{2,\infty}.$$

Now we bound these two terms. For the first term,we get

$$\begin{aligned}\|(X_I^* X_I)^{-1} v\|_{2,1} &\leq \sqrt{k}\|(X_I^* X_I)^{-1} v\|_2 \\ &\leq \sqrt{k}\|(X_I^* X_I)^{-1}\|_2\|v\|_2 \leq 2k\|v\|_{2,\infty},\end{aligned}$$

due to the invertibility condition. Using the orthogonality condition, we get

$$\begin{aligned}\|v\|_{2,\infty} &= \|X_I^* z - 2\lambda\sqrt{m}\cdot\overline{\mathrm{sign}}(\beta_I)\|_{2,\infty} \\ &\leq \|X_I^* z\|_{2,\infty} + 2\lambda\sqrt{m} \leq (2+\sqrt{2})\lambda\sqrt{m}.\end{aligned}$$

For the second term, we use Lemma 1 and the orthogonality condition to get

$$\begin{aligned}\|X_I^* X_I h\|_{2,\infty} &\leq \|X_I^*(X\beta - y)\|_{2,\infty} + \|X_I^*(y - X\widehat{\beta})\|_{2,\infty} \\ &\leq \|X_I^* z\|_{2,\infty} + \|X_I^*(y - X\widehat{\beta})\|_{2,\infty} \\ &\leq (2+\sqrt{2})\lambda\sqrt{m}.\end{aligned}$$

So we get $A_1 \leq 2(2+\sqrt{2})^2\lambda^2 mk$. For A_2, we have from Lemma 2 that

$$\begin{aligned}|A_2| &\leq \|h_{I^C}\|_{2,1}\|X_{I^C}^* X_I (X_I^* X_I)^{-1} v\|_{2,\infty} \\ &\leq (2-\sqrt{2})\lambda\sqrt{m}\|h_{I^C}\|_{2,1},\end{aligned}$$

because of the complementary size condition. Using now these bounds on A_1, A_2, we have

$$\langle h_I, v\rangle \leq 2(2+\sqrt{2})^2\lambda^2 mk + (2-\sqrt{2})\lambda\sqrt{m}\|h_{I^C}\|_{2,1}.$$

Plugging this into (4) gives

$$\frac{1}{2}\|X(\beta - \widehat{\beta})\|_2^2 \leq 2(2+\sqrt{2})^2\lambda^2 mk,$$

proving the theorem. □

4. DISCUSSION AND RELATED WORK

Note that since β has mk nonzero regression coefficients, Theorem 1 states that the group lasso results in near-optimal regression error (modulo the logarithmic factor) of $O(mk\sigma^2\log(pm))$ provided the coherence and block coherence of the design matrix are not too high. Equally importantly, the theorem states that if the design matrix is an approximately tight frame, $\|X\|_2^2 \approx \frac{pm}{n}$, then this regression error can be achieved as long as the number of nonzero regression coefficients satisfies $mk = O(n/\log(pm))$. Summarizing, our result establishes that the group lasso performs near-optimal regression even when the number of nonzero regression coefficients scales almost linearly with the number of observations, provided X is an approximately tight frame and its coherence and block coherence are not too high. Example design matrices satisfying these requirements include random Gaussian matrices and deterministic matrices designed from Grassmanian packings [3].

In terms of relation with previous work, there have been other efforts in the recent past to establish near-optimal performance of the group lasso in the underdetermined setting [1, 13–16]. However, there are three key aspects of our work that set it apart from these and similarly related works. First, our results are completely non-asymptotic in nature. Second, our results are applicable to arbitrary design matrices through the metrics of coherence, block coherence, and spectral norm, all of which are easily computable in polynomial time. Third, our results allow for near-optimal scaling of the number of observations with the number of groups of nonzero regression coefficients for matrices that are approximately tight frames. Note also that the key enabling factor that makes our results possible is a weak statistical prior on the vector of regression coefficients β, in contrast with prior work on the group lasso focused on deterministic β.

There is also a line of work in compressive sensing and sparse approximation literature that can be thought of as a special case of the problem studied here. In that work, termed the multiple measurement vector (MMV) [5] or multivariate linear regression [17] problem, it is assumed that a total of m correlated vectors $B = [\beta_1\ \beta_2\ \ldots\ \beta_m]$ are observed using a single design matrix $X \in \mathbb{R}^{d\times p}$ to obtain a set of observation vectors $Y = XB + Z$, where $Z \in \mathbb{R}^{d\times m}$ denotes the observation noise. The key distinguishing feature of the MMV setup is the assumption that the vectors $\{\beta_i\}_{i=1}^m$ share the same support, so that B has only a small number k of nonzero rows.

Interestingly, it is possible to express the MMV problem in terms of a group linear regression problem studied in this paper. Denote by $\mathrm{vect}(A)$ a column vector obtained by stacking the columns of the matrix A. Next, define $y' = \mathrm{vect}(Y^T)$, $\beta' = \mathrm{vect}(B^T)$, and $z' = \mathrm{vect}(Z^T)$ as the vectorized versions of the observations, the sparse vectors, and the noise, respectively. Additionally, define the expanded design matrix $X' = X^T \otimes \mathrm{Id}$, where $\otimes$ denotes the Kronecker product. It then follows that the MMV problem can be equivalently expressed as $y' = X'\beta' + z'$, where $\beta' \in \mathbb{R}^{pm}$ has a total of k nonzero blocks. Thus, the MMV problem can be viewed as a special case of the group linear regression problem where the design matrix X' has a particular Kronecker structure. In this special case, the proof of Theorem 1 relies on our recent result concerning the conditioning of random block-subdictionaries of matrices [2], but conditioning of random block-subdictionaries of X' is trivially guaranteed by the random subdictionaries result of [20] because of the special structure of X', which leads to a stronger variant of Theorem 1.

While we can provide a regression error performance guarantee for the MMV problem via group lasso, the significant body of literature on MMV [6, 8–12, 17, 19, 21] focuses either on model selection (support detection) or on estimation (reconstruction) error. Additionally, most of these works either study an asymptotic regime, focus on random design matrices, or rely on metrics that are either computationally expensive to evaluate or which do not allow for near-optimal scaling of the number of observations with the number of nonzero rows of B. The notable exception to this is a recent work by Eldar and Rauhut [10], which provides guarantees for exact recovery of B from Y in a noiseless setting. However, studying the regression error – the focus of this paper – is vacuous in a noiseless setting.

We conclude by noting that for $m = 1$ in the group linear regression problem, in which case the group lasso reduces to the standard lasso, Theorem 1 reduces to that obtained in [4] for the lasso. Further, while it is possible to make use of [2] together with the analysis in [4] to analyze the performance of the standard lasso for group regression error, this would require imposing statistical independence on the signs of nonzero regression coefficients even within the groups in β. We plan to highlight these and other subtle, but important, differences between the lasso and the group lasso in the future.

5. REFERENCES

[1] F. Bach. Consistency of the group lasso and multiple kernel learning. *J. Machine Learning Research*, 9(6):1179–1225, June 2008.

[2] W. Bajwa, R. Calderbank, and M. F. Duarte. On the conditioning of random block subdictionaries. Technical Report TR-2010-06, Duke University, Department of Computer Science, Durham, NC, Sept. 2010.

[3] R. Calderbank, R. H. Hardin, E. M. Rains, P. W. Shor, and N. J. A. Sloane. A group-theoretic framework for the construction of packings in Grassmanian spaces. *J. Algebraic Combinatorics*, 9(2):129–140, Mar. 1999.

[4] E. J. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *Annals of Statistics*, 37(5A):2145–2177, Oct. 2009.

[5] S.F. Cotter, B.D. Rao, E. Kjersti, and K. Kreutz-Delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Trans. Signal Processing*, 53(7):2477–2488, July 2005.

[6] M. E. Davies and Y. C. Eldar. Rank awareness in joint sparse recovery. Apr. 2010. submitted to *IEEE Trans. Info. Theory*.

[7] M. F. Duarte, W. U. Bajwa, and R. Calderbank. The performance of group lasso for linear regression of grouped variables. Technical Report TR-2010-10, Duke University, Dept. Computer Science, Durham, NC, Feb. 2011.

[8] Y. C. Eldar, P. Kuppinger, and H. Bölcksei. Block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Trans. Signal Processing*, 58(6):3042–3054, June 2010.

[9] Y. C. Eldar and M. Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Trans. Info. Theory*, 55(11):5302–5316, 2009.

[10] Y. C. Eldar and H. Rauhut. Average case analysis of multichannel sparse recovery using convex relaxation. *IEEE Trans. Info. Theory*, 6(1):505–519, Jan. 2010.

[11] M. Fornasier and H. Rauhut. Recovery algorithms for vector valued data with joint sparsity constraints. *SIAM J. Numer. Anal.*, 46(2):577–613, 2008.

[12] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst. Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms. *J. Fourier Anal. Appl.*, 14(5):655–687, 2008.

[13] J. Huang and T. Zhang. The benefit of group sparsity. *Annals of Statistics*, 38(4):1978–2004, Aug. 2010.

[14] H. Liu and J. Zhang. Estimation consistency of the group lasso and its applications. In *Int. Conf. Artificiall Intelligence and Statistics (AISTATS)*, pages 376–383, Clearwater Beach, FL, Apr. 2009.

[15] L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *J. Royal Statist. Soc. B*, 70(1):53–71, Jan. 2008.

[16] Y. Nardi and A. Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Statistics*, 2:605–633, 2008.

[17] G. Obozinski, M. J. Wainwright, and M. I. Jordan. Support union recovery in high-dimensional multivariate regression. *Annals of Statistics*, 39(1):1–47, Jan. 2011.

[18] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc B*, 58(1):267–288, 1996.

[19] J. A. Tropp. Algorithms for simultaneous sparse approximation. Part II: Convex relaxation. *Signal Processing*, 86, Apr. 2006.

[20] J. A. Tropp. Norms of random submatrices and sparse approximation. *C. R. Acad. Sci. Paris, Ser. I*, 346(23–24):1271–1274, 2008.

[21] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Processing*, 86:572–588, Apr. 2006.

[22] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. Royal Statist. Soc. B*, 68(1):49–67, Feb. 2006.