High-dimensional union support recovery in multivariate regression

Guillaume Obozinski
Department of Statistics
UC Berkeley
gobo@stat.berkeley.edu

Martin J. Wainwright
Department of Statistics
Dept. of Electrical Engineering and Computer Science
UC Berkeley
wainwright@stat.berkeley.edu

Michael I. Jordan
Department of Statistics
Department of Electrical Engineering and Computer Science
UC Berkeley
jordan@stat.berkeley.edu

Abstract

We study the behavior of block ℓ_1/ℓ_2 regularization for multivariate regression, where a K-dimensional response vector is regressed upon a fixed set of p covariates. The problem of *union support recovery* is to recover the subset of covariates that are active in at least one of the regression problems. Studying this problem under high-dimensional scaling (where the problem parameters as well as sample size n tend to infinity simultaneously), our main result is to show that exact recovery is possible once the order parameter given by $\theta_{\ell_1/\ell_2}(n, p, s) := n/[2\psi(B^*) \log(p - s)]$ exceeds a critical threshold. Here n is the sample size, p is the ambient dimension of the regression model, s is the size of the union of supports, and $\psi(B^*)$ is a *sparsity-overlap function* that measures a combination of the sparsities and overlaps of the K-regression coefficient vectors that constitute the model. This sparsity-overlap function reveals that block ℓ_1/ℓ_2 regularization for multivariate regression never harms performance relative to a naive ℓ_1-approach, and can yield substantial improvements in sample complexity (up to a factor of K) when the regression vectors are suitably orthogonal relative to the design. We complement our theoretical results with simulations that demonstrate the sharpness of the result, even for relatively small problems.

1 Introduction

A recent line of research in machine learning has focused on regularization based on block-structured norms. Such structured norms are well-motivated in various settings, among them kernel learning [3, 6], grouped variable selection [11], hierarchical model selection [12], simultaneous sparse approximation [9], and simultaneous feature selection in multi-task learning [5]. Block-norms that compose an ℓ_1-norm with other norms yield solutions that tend to be sparse like the Lasso [8], but the structured norm also enforces *blockwise sparsity*, in the sense that parameters within blocks are more likely zero (or non-zero) simultaneously.

The focus of this paper is the model selection consistency of block-structured regularization in the setting of multivariate regression. Our goal is to perform model or variable selection, by which we mean extracting the subset of relevant covariates that are active in at least one regression. We refer to this problem as the *support union problem*. In line with a large body of recent work in statistical machine learning (e.g., [2, 7, 13, 10]), our analysis is high-dimensional in nature, meaning that we allow the model dimension p (as well as other structural parameters) to grow along with the sample size n. A great deal of work has focused on the case of ordinary ℓ_1-regularization (Lasso) [2, 10, 13], showing for instance that the Lasso can recover the support of a sparse signal even when $p \gg n$.

Some more recent work has studied consistency issues for block-regularization schemes, including classical analysis (p fixed) of the group Lasso [1], and high-dimensional analysis of the predictive risk of block-regularized logistic regression [4]. Although there have been various empirical demonstrations of the benefits of block regularization, to the best of our knowledge, there has not yet been theoretical analysis of such improvements. In this paper, our goal is to understand the following question: under what conditions does block regularization lead to a quantifiable improvement in statistical efficiency, relative to more naive regularization schemes? Here statistical efficiency is assessed in terms of the *sample complexity*, meaning the minimal sample size n required to recover the support union; we wish to know how this scales as a function of problem parameters. Our main contribution is to provide a function quantifying the benefits of block regularization schemes for the problem of multivariate linear regression, showing in particular that, under suitable structural conditions on the data, the block-norm regularization we consider never harms performance relative to naive ℓ_1-regularization and can lead to substantial gains in sample complexity.

More specifically, we consider the following problem of multivariate linear regression: a group of K scalar outputs are regressed on the same design matrix $X \in \mathbb{R}^{n \times p}$. Representing the regression coefficients as a $p \times K$ matrix B^*, the regression model takes the form

$$Y \quad = \quad XB^* + W, \tag{1}$$

where $Y \in \mathbb{R}^{n \times K}$ and $W \in \mathbb{R}^{n \times K}$ are matrices of observations and zero-mean noise respectively and B^* has columns $\beta^{*(1)}, \ldots, \beta^{*(K)}$ which are the parameter vectors of each univariate regression.

We are interested in recovering the *union of the supports* of individual regressions, more specifically if $S_k = \Big\{ i \in \{1, \ldots, p\}, \beta_i^{*(k)} \neq 0 \Big\}$ we would like to recover $S = \cup_k S_k$. The Lasso is often presented as a relaxation of the so-called ℓ_0 regularization, i.e. the count of the number of non-zero parameter coefficients, a quite intractable non-convex function. More generally, block-norm regularizations can be thought of as the relaxation of some non-convex regularization which counts the number of covariates i for which at least one of the univariate regression parameter $\beta_i^{*(k)}$ is non-zero. More specifically, let β_i^* denote the i^{th} row of B^*, and define, for $q \geq 1$,

$$\|B^*\|_{\ell_0/\ell_q} = |\{i \in \{1, \ldots, p\}, \; \|\beta_i^*\|_q > 0\}| \qquad \text{and} \qquad \|B^*\|_{\ell_1/\ell_q} = \sum_{i=1}^{p} \|\beta_i^*\|_q$$

All ℓ_0/ℓ_q norms define the same function, but differ conceptually in that they lead to different ℓ_1/ℓ_q relaxations. In particular the ℓ_1/ℓ_1 regularization is the same as the usual Lasso. The other conceptually most natural block-norms are ℓ_1/ℓ_2 and ℓ_1/ℓ_∞. While ℓ_1/ℓ_∞ is of interest, it seems intuitively to be relevant essentially to situations where the support is exactly the same for all regressions, an assumption that we are not willing to make.

In the current paper, we focus on the ℓ_1/ℓ_2 case and consider the estimator $\widehat{B}$ obtained by solving the following disguised second-order cone program:

$$\min_{B \in \mathbb{R}^{p \times K}} \left\{ \frac{1}{2n} \, |\!|\!| Y - XB |\!|\!|_F^2 + \lambda_n \, \|B\|_{\ell_1/\ell_2} \right\}, \tag{2}$$

where $|\!|\!| M |\!|\!|_F := (\sum_{i,j} m_{ij}^2)^{1/2}$ denotes the Frobenius norm. We study the support union problem under high-dimensional scaling, meaning that the number of observations n, the ambient dimension p and the size of the *union of supports* s can all tend to infinity. The main contribution of this paper is to show that under certain technical conditions on the design and noise matrices, the model selection performance of block-regularized ℓ_1/ℓ_2 regression (2) is governed by the *control parameter* $\theta_{\ell_1/\ell_2}(n, p\,; B^*) := \frac{n}{2\,\psi(B^*)\log(p-s)}$, where n is the sample size, p is the ambient dimension, $s = |S|$ is the size of the union of the supports, and $\psi(\cdot)$ is a *sparsity-overlap function* defined below. More precisely, the probability of correct union support recovery converges to one for all sequences (n, p, s, B^*) such that the control parameter $\theta_{\ell_1/\ell_2}(n, p\,; B^*)$ exceeds a fixed critical threshold $\theta_{\text{crit}} < +\infty$. Note that θ_{ℓ_1/ℓ_2} is a measure of the sample complexity of the problem—that is, the sample size required for exact recovery as a function of the problem parameters. Whereas the ratio $(n/\log p)$ is standard for high-dimensional theory on ℓ_1-regularization (essentially due to covering numberings of ℓ_1 balls), the function $\psi(B^*)$ is a novel and interesting quantity, which measures both the sparsity of the matrix B^*, as well as the overlap between the different regression tasks (columns of B^*).

In Section 2, we introduce the models and assumptions, define key characteristics of the problem and state our main result and its consequences. Section 3 is devoted to the proof of this main result, with most technical results deferred to the appendix. Section 4 illustrates with simulations the sharpness of our analysis and how quickly the asymptotic regime arises.

1.1 Notations

For a (possibly random) matrix $M \in \mathbb{R}^{p \times K}$, and for parameters $1 \leq a \leq b \leq \infty$, we distinguish the ℓ_a/ℓ_b block norms from the (a,b)-operator norms, defined respectively as

$$\|M\|_{\ell_a/\ell_b} := \left\{ \sum_{i=1}^{p} \left(\sum_{k=1}^{K} |m_{ik}|^b \right)^{\frac{a}{b}} \right\}^{\frac{1}{a}} \qquad \text{and} \qquad |\!|\!| M |\!|\!|_{a,\,b} := \sup_{\|x\|_b = 1} \|Mx\|_a, \tag{3}$$

although ℓ_∞/ℓ_p norms belong to both families (see Lemma B.0.1). For brevity, we denote the spectral norm $|\!|\!| M |\!|\!|_{2,\,2}$ as $|\!|\!| M |\!|\!|_2$, and the ℓ_∞-operator norm $|\!|\!| M |\!|\!|_{\infty,\,\infty} = \max_i \sum_j |M_{ij}|$ as $|\!|\!| M |\!|\!|_\infty$.

2 Main result and some consequences

The analysis of this paper applies to multivariate linear regression problems of the form (1), in which the noise matrix $W \in \mathbb{R}^{n \times K}$ is assumed to consist of i.i.d. elements $W_{ij} \sim N(0, \sigma^2)$. In addition, we assume that the measurement or design matrices X have rows drawn in an i.i.d. manner from a zero-mean Gaussian $N(0, \Sigma)$, where $\Sigma \succ 0$ is a $p \times p$ covariance matrix.

Suppose that we partition the full set of covariates into the support set S and its complement S^c, with $|S| = s$, $|S^c| = p - s$. Consider the following block decompositions of the regression coefficient matrix, the design matrix and its covariance matrix:

$$B^* = \begin{bmatrix} B^*_S \\ B^*_{S^c} \end{bmatrix}, \quad X = [X_S \; X_{S^c}], \quad \text{and} \quad \Sigma = \begin{bmatrix} \Sigma_{SS} & \Sigma_{SS^c} \\ \Sigma_{S^cS} & \Sigma_{S^cS^c} \end{bmatrix}.$$

We use β^*_i to denote the i^{th} row of B^*, and assume that the sparsity of B^* is assessed as follows:

$(A0)$ **Sparsity:** The matrix B^* has row support $S := \{i \in \{1, \ldots, p\} \mid \beta^*_i \neq 0\}$, with $s = |S|$.

In addition, we make the following assumptions about the covariance Σ of the design matrix:

$(A1)$ **Bounded eigenspectrum:** There exist a constant $C_{\min} > 0$ (resp. $C_{\max} < +\infty$) such that all eigenvalues of Σ_{SS} (resp. Σ) are greater than $C_{\min}$ (resp. smaller than $C_{\max}$).

$(A2)$ **Mutual incoherence:** There exists $\gamma \in (0, 1]$ such that $|\!|\!| \Sigma_{S^cS}(\Sigma_{SS})^{-1} |\!|\!|_\infty \leq 1 - \gamma$.

$(A3)$ **Self incoherence:** There exists a constant $D_{\max}$ such that $|\!|\!| (\Sigma_{SS})^{-1} |\!|\!|_\infty \leq D_{\max}$.

Assumption $A1$ is a standard condition required to prevent excess dependence among elements of the design matrix associated with the support S. The mutual incoherence assumption $A2$ is also well-known from previous work on model selection with the Lasso [9, 13]. These assumptions are trivially satisfied by the standard Gaussian ensemble ($\Sigma = I_p$) with $C_{\min} = C_{\max} = D_{\max} = \gamma = 1$. More generally, it can be shown that various matrix classes satisfy these conditions [13, 10].

2.1 Statement of main result

With the goal of estimating the union of supports S, our main result is a set of sufficient conditions using the following procedure. Solve the block-regularized problem (2) with regularization parameter $\lambda_n > 0$, thereby obtaining a solution $\widehat{B} = \widehat{B}(\lambda_n)$. Use this solution to compute an estimate of the support union as $\widehat{S}(\widehat{B}) := \left\{ i \in \{1, \ldots, p\} \mid \widehat{\beta}_i \neq 0 \right\}$. This estimator is unambiguously defined if the solution $\widehat{B}$ is unique, and as part of our analysis, we show that the solution $\widehat{B}$ is indeed unique with high probability in the regime of interest. We study the behavior of this estimator for a sequence of linear regressions indexed by the triplet (n, p, s), for which the data follows the general model presented in the previous section with defining parameters $B^*(n)$ and $\Sigma(n)$ satisfying $A0$-$A3$. As (n, p, s) tends to infinity, we give conditions on the triplet and properties of B^* for which $\widehat{B}$ is unique, and such that $\mathbb{P}[\widehat{S} = S] \rightarrow 1$.

The central objects in our main result are the sparsity-overlap function, and the sample complexity parameter, which we define here. For any vector $\beta_i \neq 0$, define $\zeta(\beta_i) := \frac{\beta_i}{\|\beta_i\|_2}$. We extend the function ζ to any matrix $B_S \in \mathbb{R}^{s \times K}$ with non-zero rows by defining the matrix $\zeta(B_S) \in \mathbb{R}^{s \times K}$ with i^{th} row $[\zeta(B_S)]_i = \zeta(\beta_i)$. With this notation, we define the *sparsity-overlap function* $\psi(B)$ and the *sample complexity parameter* $\theta_{\ell_1/\ell_2}(n, p\,; B^*)$ as

$$\psi(B) := \left\| \zeta(B_S)^T (\Sigma_{SS})^{-1} \zeta(B_S) \right\|_2 \qquad \text{and} \qquad \theta_{\ell_1/\ell_2}(n, p\,; B^*) := \frac{n}{2\,\psi(B^*) \log(p-s)}. \tag{4}$$

Finally, we use $b^*_{\min} := \min_{i \in S} \|\beta^*_i\|_2$ to denote the minimal ℓ_2 row-norm of the matrix B^*_S. With this notation, we have the following result:

Theorem 1. *Consider a random design matrix X drawn with i.i.d. $N(0, \Sigma)$ row vectors, an observation matrix Y specified by model (1), and a regression matrix B^* such that $(b^*_{\min})^2$ decays strictly more slowly than $\frac{f(p)}{n} \max\{s, \log(p-s)\}$, for any function $f(p) \rightarrow +\infty$. Suppose that we solve the block-regularized program (2) with regularization parameter $\lambda_n = \Theta\left(\sqrt{f(p) \log(p)/n}\right)$.*

For any sequence (n, p, B^*) such that the ℓ_1/ℓ_2 control parameter $\theta_{\ell_1/\ell_2}(n, p\,; B^*)$ exceeds the critical threshold $\theta_{\text{crit}}(\Sigma) := \frac{C_{\max}}{\gamma^2}$, then with probability greater than $1 - \exp(-\Theta(\log p))$,

- *(a) the block-regularized program (2) has a unique solution $\widehat{B}$, and*
- *(b) its support set $\widehat{S}(\widehat{B})$ is equal to the true support union S.*

Remarks: For the standard Gaussian ensemble ($\Sigma = I_p$), the critical threshold is simply $\theta_{\text{crit}}(\Sigma) = 1$. More generally, the critical threshold $\theta_{\text{crit}}(\Sigma)$ depends on the maximum and minimum eigenvalues $C_{\max}$ and $C_{\min}$, as well as the mutual incoherence parameter γ (see Assumptions $A1$ and $A2$). A technical condition that we require on the regularization parameter is

$$\frac{\lambda_n^2 n}{\log(p-s)} \rightarrow \infty \tag{5}$$

which is satisfied by the choice given in the statement.

2.2 Some consequences of Theorem 1

It is interesting to consider some special cases of our main result. The simplest special case is the univariate regression problem ($K = 1$), in which case the the function $\zeta(\beta^*)$ outputs an s-dimensional sign vector with elements $z^*_i = \text{sign}(\beta^*_i)$, so that $\psi(\beta^*) = {z^*}^T (\Sigma_{SS})^{-1} z^* = \Theta(s)$. Consequently, the order parameter of block ℓ_1/ℓ_2-regression for univariate regresion is given by $\Theta(n/(2s\log(p-s))$, which matches the scaling established in previous work on the Lasso [10].

More generally, given our assumption $(A1)$ on Σ_{SS}, the sparsity overlap $\psi(B^*)$ always lies in the interval $[C_{\min}\frac{s}{K}, C_{\max}s]$. At the most pessimistic extreme, suppose that $B^* := \beta^* \vec{1}_K^T$—that is, B^* consists of K copies of the same coefficient vector $\beta^* \in \mathbb{R}^p$, with support of cardinality $|S| = s$. We then have $[\zeta(B^*)]_{ij} = \text{sign}(\beta^*_i)/\sqrt{K}$, from which we see that $\psi(B^*) = {z^*}^T (\Sigma_{SS})^{-1} z^*$, with z^* again the s-dimensional sign vector with elements $z^*_i = \text{sign}(\beta^*_i)$, so that there is no benefit in sample complexity relative to the naive strategy of solving separate Lasso problems and constructing the union of individually estimated supports. This might seem a pessimistic result, since under model (1), we essentially have Kn observations of the coefficient vector β^* with the same design matrix but K independent noise realizations. However, the thresholds as well as the rates of convergence in high-dimensional results such as Theorem 1 are not determined by the noise variance, but rather by the number of interfering variables $(p - s)$.

At the most optimistic extreme, consider the case where $\Sigma_{SS} = I_s$ and (for $s > K$) suppose that B^* is constructed such that the columns of the $s \times K$ matrix $\zeta(B^*)$ are all orthogonal. Under this condition, we have

Corollary 1 (Orthonormal tasks)**.** *If the columns of the matrix $\zeta(B^*)$ are all orthogonal with equal length and $\Sigma_{SS} = I_{s\times s}$ then the block-regularized problem* (2) *succeeds in union support recovery once the sample complexity parameter $n/(2\frac{s}{K}\log(p-s))$ is larger than 1.*

For the standard Gaussian ensemble, it is known [10] that the Lasso fails with probability one for all sequences such that $n < (2-\nu)s\log(p-s)$ for any arbitrarily small $\nu > 0$. Consequently, Corollary 1 shows that under suitable conditions on the regression coefficient matrix B^*, ℓ_1/ℓ_2 can provides a K-fold reduction in the number of samples required for exact support recovery.

As a third illustration, consider, for $\Sigma_{SS} = I_{s\times s}$, the case where the supports S_k of individual regression problems are all disjoint. The sample complexity parameter for each of the individual Lassos is $n/(2s_k\log(p-s_k))$ where $|S_k| = s_k$, so that the sample size required to recover the union support from individual Lassos scales as $n = \Theta(2(\max_k s_k)\log(p-s))$. However, if the supports are all disjoint, then the columns of the matrix $Z_S^* = \zeta(B_S^*)$ are orthogonal, and ${Z_S^*}^T Z_S^* = \text{diag}(s_1,\ldots,s_K)$ so that $\psi(B^*) = \max_k s_k$ and the sample complexity is the same. In other words, even though there is no sharing of variables at all there is surprisingly no penalty from regularizing jointly with the ℓ_1/ℓ_2-norm. However, this is not always true if $\Sigma_{SS} \neq I_{s\times s}$ and in some cases ℓ_1/ℓ_2-regularization can have higher sample complexity than separate Lassos.

3 Proof of Theorem 1

In addition to previous notations, the proofs use the shorthands: $\widehat{\Sigma}_{SS}{=}\frac{1}{n}X_S^T X_S$, $\widehat{\Sigma}_{S^cS}{=}\frac{1}{n}X_{S^c}^T X_S$ and $\Pi_S = X_S(\widehat{\Sigma}_{SS})^{-1}X_S^T$ denotes the orthogonal projection onto the range of X_S.

High-level proof outline: At a high level, our proof is based on the notion of what we refer to as a *primal-dual witness*: we first formulate the problem (2) as a second-order cone program (SOCP), with the same primal variable B as in (2) and a dual variable Z whose rows coincide at optimality with the subgradient of the ℓ_1/ℓ_2 norm. We then construct a primal matrix $\widehat{B}$ along with a dual matrix $\widehat{Z}$ such that, under the conditions of Theorem 1, with probability converging to 1:

(a) The pair $(\widehat{B}, \widehat{Z})$ satisfies the Karush-Kuhn-Tucker (KKT) conditions of the SOCP.

(b) In spite of the fact that for general high-dimensional problems (with $p \gg n$), the SOCP need not have a unique solution a priori, a strict feasibility condition satisfied by the dual variables $\widehat{Z}$ guarantees that $\widehat{B}$ is the unique optimal solution of (2).

(c) The support union $\hat{S}$ of $\widehat{B}$ is identical to the support union S of B^*.

At the core of our constructive procedure is the following convex-analytic result, which characterizes an optimal primal-dual pair for which the primal solution $\widehat{B}$ correctly recovers the support set S:

Lemma 1. *Suppose that there exists a primal-dual pair $(\widehat{B}, \widehat{Z})$ that satisfy the conditions:*

$$\begin{aligned}
\widehat{Z}_S &= \zeta(\widehat{B}_S) && \text{(6a)}\\
\widehat{\Sigma}_{SS}(\widehat{B}_S - B_S^*) - \frac{1}{n}X_S^T W &= -\lambda_n \widehat{Z}_S && \text{(6b)}\\
\lambda_n \left\|\widehat{Z}_{S^c}\right\|_{\ell_\infty/\ell_2} &:= \left\|\widehat{\Sigma}_{S^cS}(\widehat{B}_S - B_S^*) - \frac{1}{n}X_{S^c}^T W\right\|_{\ell_\infty/\ell_2} < \lambda_n && \text{(6c)}\\
\widehat{B}_{S^c} &= 0. && \text{(6d)}
\end{aligned}$$

Then $(\widehat{B}, \widehat{Z})$ is the unique optimal solution to the block-regularized problem, with $\widehat{S}(\widehat{B}) = S$ by construction.

Appendix A proves Lemma 1, with the strict feasibility of $\widehat{Z}_{S^c}$ given by (6c) to certify uniqueness.

3.1 Construction of primal-dual witness

Based on Lemma 1, we construct the primal dual pair $(\widehat{B}, \widehat{Z})$ as follows. First, we set $\widehat{B}_{S^c} = 0$, to satisfy condition (6d). Next, we obtain the pair $(\widehat{B}_S, \widehat{Z}_S)$ by solving a restricted version of (2):

$$\widehat{B}_S = \arg\min_{B_S \in \mathbb{R}^{s\times K}} \left\{ \frac{1}{2n}\left|\left|\left| Y - X \begin{bmatrix} B_S \\ 0_{S^c} \end{bmatrix} \right|\right|\right|_F^2 + \lambda_n \|B_S\|_{\ell_1/\ell_2} \right\}. \tag{7}$$

Since $s < n$, the empirical covariance (sub)matrix $\widehat{\Sigma}_{SS} = \frac{1}{n} X_S^T X_S$ is strictly positive definite with probability one, which implies that the restricted problem (7) is strictly convex and therefore has a unique optimum $\widehat{B}_S$. We then choose $\widehat{Z}_S$ to be the solution of equation (6b). Since any such matrix $\widehat{Z}_S$ is also a dual solution to the SOCP (7), it must be an element of the subdifferential $\partial \|\widehat{B}_S\|_{\ell_1/\ell_2}$. It remains to show that this construction satisfies conditions (6a) and (6c). In order to satisfy condition (6a), it suffices to show that $\widehat{\beta}_i \neq 0,\ i \in S$. From equation (6b) and since $\widehat{\Sigma}_{SS}$ is invertible, we may solve as follows

$$(\widehat{B}_S - B_S^*) \quad = \quad \left(\widehat{\Sigma}_{SS}\right)^{-1} \left[\frac{X_S^T W}{n} - \lambda_n \widehat{Z}_S\right] \; =: \; U_S. \tag{8}$$

For any row $i \in S$, we have $\|\widehat{\beta}_i\|_2 \geq \|\beta_i^*\|_2 - \|U_S\|_{\ell_\infty/\ell_2}$. Thus, it suffices to show that the following event occurs with high probability

$$\mathcal{E}(U_S) \quad := \quad \left\{ \|U_S\|_{\ell_\infty/\ell_2} \leq \frac{1}{2}\, b^*_{\min} \right\} \tag{9}$$

to show that no row of $\widehat{B}_S$ is identically zero. We establish this result later in this section.

Turning to condition (6c), by substituting expression (8) for the difference $(\widehat{B}_S - B_S^*)$ into equation (6c), we obtain a $(p-s) \times K$ random matrix V_{S^c}, whose row $j \in S^c$ is given by

$$V_j \quad := \quad X_j^T \left([\Pi_S - I_n] \frac{W}{n} - \lambda_n \frac{X_S}{n} \, (\widehat{\Sigma}_{SS})^{-1} \widehat{Z}_S \right). \tag{10}$$

In order for condition (6c) to hold, it is necessary and sufficient that the probability of the event

$$\mathcal{E}(V_{S^c}) \quad := \quad \left\{ \|V_{S^c}\|_{\ell_\infty/\ell_2} < \lambda_n \right\} \tag{11}$$

converges to one as n tends to infinity.

Correct inclusion of supporting covariates: We begin by analyzing the probability of $\mathcal{E}(U_S)$.

Lemma 2. *Under assumption $A3$ and conditions (5) of Theorem 1, with probability $1 - \exp(-\Theta(\log s))$, we have*

$$\|U_S\|_{\ell_\infty/\ell_2} \leq \mathcal{O}\left(\sqrt{(\log s)/n}\right) + \lambda_n \, \left(D_{\max} + \mathcal{O}\left(\sqrt{s^2/n}\right)\right).$$

This lemma is proved in in the Appendix. With the assumed scaling $n = \Omega\left(s \log(p-s)\right)$, and the assumed slow decrease of $b^*_{\min}$, which we write explicitly as $(b^*_{\min})^2 \geq \frac{1}{\varepsilon_n^2} \frac{f(p)\max\{s, \log(p-s)\}}{n}$ for some $\varepsilon_n \to 0$, we have

$$\frac{\|U_S\|_{\ell_\infty/\ell_2}}{b^*_{\min}} \quad \leq \quad \mathcal{O}(\varepsilon_n) \tag{12}$$

so that the conditions of Theorem 1 ensure that $\mathcal{E}(U_S)$ occurs with probability converging to one.

Correct exclusion of non-support: Next we analyze the event $\mathcal{E}(V_{S^c})$. For simplicity, in the following arguments, we drop the index S^c and write V for V_{S^c}. In order to show that $\|V\|_{\ell_\infty/\ell_2} < \lambda_n$ with probability converging to one, we make use of the decomposition

$$\frac{1}{\lambda_n} \|V\|_{\ell_\infty/\ell_2} \leq \sum_{i=1}^{3} T_i' \qquad \text{where} \qquad T_1' := \frac{1}{\lambda_n} \|\mathbb{E}\left[V \mid X_S\right]\|_{\ell_\infty/\ell_2},$$

$$T_2' := \frac{1}{\lambda_n} \|\mathbb{E}\left[V | X_S, W\right] - \mathbb{E}\left[V | X_S\right]\|_{\ell_\infty/\ell_2} \qquad \text{and} \qquad T_3' := \frac{1}{\lambda_n} \|V - \mathbb{E}\left[V | X_S, W\right]\|_{\ell_\infty/\ell_2}.$$

Lemma 3. *Under assumption $A2$, $T_1' \leq 1 - \gamma$. Under conditions (5) of Theorem 1, $T_2' = o_p(1)$.*

Therefore, to show that $\frac{1}{\lambda_n} \|V\|_{\ell_\infty/\ell_2} < 1$ with high probability , it suffices to show that $T_3' < \gamma$ with high probability. Until now, we haven't appealed to the sample complexity parameter $\theta_{\ell_1/\ell_2}(n, p\,; B^*)$. In the next section, we prove that $\theta_{\ell_1/\ell_2}(n, p\,; B^*) > \theta_{\text{crit}}(\Sigma)$ implies that $T_3' < \gamma$ with high probability.

Lemma 4. *Conditionally on W and X_S, we have*

$$\left(\|V_j - \mathbb{E}\left[V_j \mid X_S,\, W\right]\|_2^2 \,|\, W, X_S \right) \;\; \overset{d}{=} \;\; \left(\Sigma_{S^c\,|\,S}\right)_{jj} \xi_j^T M_n \xi_j.$$

where $\xi_j \sim N(\vec{0}_K, I_K)$ and where the $K \times K$ matrix $M_n = M_n(X_S, W)$ is given by

$$M_n \quad := \quad \frac{\lambda_n^2}{n} \widehat{Z}_S^T (\widehat{\Sigma}_{SS})^{-1} \widehat{Z}_S + \frac{1}{n^2} W^T (\Pi_S - I_n) W. \qquad (13)$$

But the covariance matrix M_n is itself concentrated. Indeed,

Lemma 5. *Under the conditions (5) of Theorem 1, for any $\delta > 0$, the following event $\mathcal{T}(\delta)$ has probability converging to 1:*

$$\mathcal{T}(\delta) \;:=\; \left\{ \|\!|M_n|\!\|_2 \leq \lambda_n^2 \, \frac{\psi(B^*)}{n} \, (1+\delta) \right\} \qquad (14)$$

For any fixed $\delta > 0$, we have $\mathbb{P}[T_3' \geq \gamma] \leq \mathbb{P}[T_3' \geq \gamma \mid \mathcal{T}(\delta)] + \mathbb{P}[\mathcal{T}(\delta)^c]$. but, from lemma 5, $\mathbb{P}[\mathcal{T}(\delta)^c] \to 0$, so that it suffices to deal with the first term.

Given that $(\Sigma_{S^c\,|\,S})_{jj} \leq (\Sigma_{S^cS^c})_{jj} \;\leq\; C_{\max}$ for all j, on the event $\mathcal{T}(\delta)$, we have

$$\max_{j \in S^c} \, (\Sigma_{S^c\,|\,S})_{jj} \, \xi_j^T M_n \xi_j \leq C_{\max} \; \|\!|M_n|\!\|_2 \max_{j \in S^c} \|\xi_j\|_2^2 \leq C_{\max} \; \lambda_n^2 \; \frac{\psi(B^*)}{n} \max_{j \in S^c} \|\xi_j\|_2^2 \quad \text{and}$$

$$\mathbb{P}[T_3' \geq \gamma \,|\, \mathcal{T}(\delta)] \;\leq\; \mathbb{P}\left[\max_{j \in S^c} \|\xi_j\|_2^2 \geq 2t^*(n, B^*)\right] \quad \text{with} \quad t^*(n, B^*) \;:=\; \frac{1}{2} \frac{\gamma^2}{C_{\max}} \; \frac{n}{\psi(B^*) \; (1+\delta)}.$$

Finally using the union bound and a large deviation bound for χ^2 variates we get the following condition which is equivalent to the condition of Theorem 1: $\theta_{\ell_1/\ell_2}(n, p\,; B^*) > \theta_{\text{crit}}(\Sigma)$:

Lemma 6. $\mathbb{P}\left[\max_{j \in S^c} \|\xi_j\|_2^2 \geq 2t^*(n, B^*)\right] \to 0$ *if* $t^*(n, B^*) > (1+\nu)\log(p-s)$ *for some* $\nu > 0$.

4 Simulations

In this section, we illustrate the sharpness of Theorem 1 and furthermore ascertain how quickly the predicted behavior is observed as n, p, s grow in different regimes, for two regression tasks (i.e. $K = 2$). In the following simulations, the matrix B^* of regression coefficients is designed with entries β_{ij}^* in $\{-1/\sqrt{2}, 1/\sqrt{2}\}$ to yield a desired value of $\psi(B^*)$. The design matrix X is sampled from the standard Gaussian ensemble. Since $|\beta_{ij}^*| = 1/\sqrt{2}$ in this construction, we have $B_S^* = \zeta(B_S^*)$, and $b_{\min}^* = 1$. Moreover, since $\Sigma = I_p$, the sparsity-overlap $\psi(B^*)$ is simply $\|\!|\, \zeta(B^*)^T \zeta(B^*) |\!\|_2$. From our analysis, the sample complexity parameter θ_{ℓ_1/ℓ_2} is controlled by the "interference" of irrelevant covariates, and not by the variance of a noise component.

We consider linear sparsity with $s = \alpha p$, for $\alpha = 1/8$, for various ambient model dimensions $p \in \{32, 256, 1024\}$. For each value of p, we perform simulations varying the sample size n to match corresponding values of the basic Lasso sample complexity parameter, given by $\theta_{\text{Las}} := n/(2s\log(p-s))$, in the interval $[0.25, 1.5]$. In each case, we solve the block-regularized problem (2) with sample size $n = 2\theta_{\text{Las}} s \log(p-s)$ using the regularization parameter $\lambda_n = \sqrt{\log(p-s)\;(\log s)/n}$. In all cases, the noise level is set at $\sigma = 0.1$.

For our construction of matrices B^*, we choose both p and the scalings for the sparsity so that the obtained values for s that are multiples of four, and construct the columns $Z^{(1)*}$ and $Z^{(2)*}$ of the matrix $B^* = \zeta(B^*)$ from copies of vectors of length 4. Denoting by $\otimes$ the usual matrix tensor product, we considered :

Identical regressions: We set $Z^{(1)*} = Z^{(2)*} = \frac{1}{\sqrt{2}} \vec{1}_s$, so that the sparsity-overlap is $\psi(B^*) = s$.

Orthogonal regression: Here B^* is constructed with $Z^{(1)*} \perp Z^{(2)*}$, so that $\psi(B^*) = \frac{s}{2}$, the most favorable situation. To achieve this, we set $Z^{(1)*} = \frac{1}{\sqrt{2}} \vec{1}_s$ and $Z^{(2)*} = \frac{1}{\sqrt{2}} \vec{1}_{s/2} \otimes (1, -1)^T$.

Intermediate angles: In this intermediate case, the columns $Z^{(1)*}$ and $Z^{(2)*}$ are at a 60° angle, which leads to $\psi(B^*) = \frac{3}{4}s$. We set $Z^{(1)*} = \frac{1}{\sqrt{2}} \vec{1}_s$ and $Z^{(2)*} = \frac{1}{\sqrt{2}} \vec{1}_{s/4} \otimes (1, 1, 1, -1)^T$.

Figure 1 shows plots of all three cases and the reference Lasso case for the three different values of the ambient dimension and the two types of sparsity described above. Note how the curves all undergo a threshold phenomenon, with the location consistent with the predictions of Theorem 1.

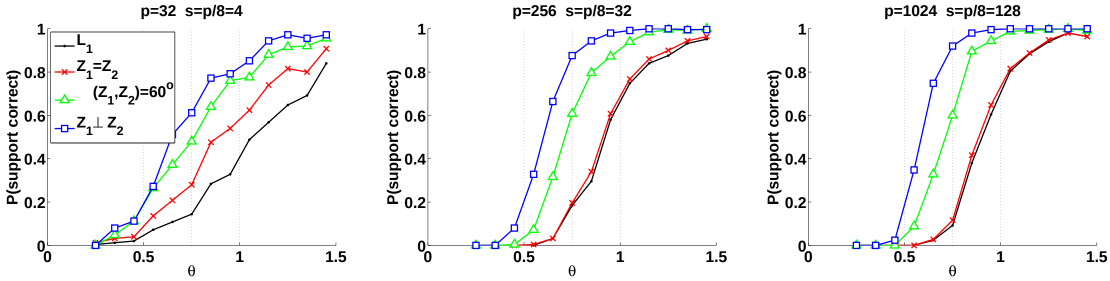

Figure 1. Plots of support recovery probability $\mathbb{P}[\widehat{S} = S]$ versus the basic ℓ_1 control parameter $\theta_{\text{Las}}{=}n/[2s\log(p-s)]$ for linear sparsity $s{=}p/8$, and for increasing values of $p \in \{32, 256, 1024\}$ from left to right. Each graph shows four curves corresponding to the case of independent ℓ_1 regularization (black), and for ℓ_1/ℓ_2 regularization, the cases of identical regression (red), intermediate angles (green), and orthogonal regressions (blue). As plotted in dotted vertical lines, Theorem 1 predicts that identical case should succeed for $\theta_{\text{Las}}{>}1$ (identical to the ordinary Lasso), intermediate case for $\theta_{\text{Las}} {>} 0.75$, and orthogonal case for $\theta_{\text{Las}} {>} 0.50$. The leftward shift of these curves confirms this theoretical prediction.

5 Discussion

We have studied union support recovery under high-dimensional scaling with the ℓ_1/ℓ_2 regularization, and shown that its sample complexity is determined by the function $\psi(B^*)$. The latter integrates the sparsity of each univariate regression with the overlap of all the supports and the discrepancies between each of the vectors of parameter estimated. In favorable cases, for K regressions, the sample complexity for ℓ_1/ℓ_2 is K times smaller than the Lasso sample complexity. Moreover, this gain is not obtained at the expense of an assumption of shared support over the data, that could be violated. Rather, the regularization seems "adaptive" in sense that, at least for standard Gaussian designs, it doesn't perform worse than the Lasso if the supports are disjoint. This seems to hold with some generality (although not universally) over choices of the design covariance matrix; this adaptivity should be characterized in future work.

References

[1] F. Bach. Consistency of the group Lasso and multiple kernel learning. Technical report, INRIA - Département d'Informatique, Ecole Normale Supérieure, 2008.

[2] D. Donoho, M. Elad, and V. M. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Info Theory*, 52(1):6–18, January 2006.

[3] G. Lanckriet, F. Bach, and M. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proc. Int. Conf. Machine Learning (ICML)*. Morgan Kaufmann, 2004.

[4] L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. Technical report, Mathematics Department, Swiss Federal Institute of Technology Zürich, 2007.

[5] G. Obozinski, B. Taskar, and M. Jordan. Joint covariate selection for grouped classification. Technical report, Statistics Department, UC Berkeley, 2007.

[6] M. Pontil and C.A. Michelli. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6:1099–1125, 2005.

[7] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. SpAM: sparse additive models. In *Neural Info. Proc. Systems (NIPS) 21*, Vancouver, Canada, December 2007.

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

[9] J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Info Theory*, 52(3):1030–1051, March 2006.

[10] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity using using ℓ_1-constrained quadratic programs. Technical Report 709, Department of Statistics, UC Berkeley, 2006.

[11] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 1(68):4967, 2006.

[12] P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. Technical report, Statistics Department, UC Berkeley, 2007.

[13] P. Zhao and B. Yu. Model selection with the lasso. *J. of Machine Learning Research*, pages 2541–2567, 2007.