The Annals of Statistics
2011, Vol. 39, No. 1, 1–47
DOI: 10.1214/09-AOS776

SUPPORT UNION RECOVERY IN HIGH-DIMENSIONAL MULTIVARIATE REGRESSION[1]

BY GUILLAUME OBOZINSKI, MARTIN J. WAINWRIGHT AND MICHAEL I. JORDAN

INRIA—Willow Project-Team Laboratoire d'Informatique de l'Ecole Normale Supérieure, University of California, Berkeley and University of California, Berkeley

In multivariate regression, a K-dimensional response vector is regressed upon a common set of p covariates, with a matrix $B^* \in \mathbb{R}^{p\times K}$ of regression coefficients. We study the behavior of the *multivariate group Lasso*, in which block regularization based on the ℓ_1/ℓ_2 norm is used for *support union recovery*, or recovery of the set of s rows for which B^* is nonzero. Under high-dimensional scaling, we show that the multivariate group Lasso exhibits a threshold for the recovery of the exact row pattern with high probability over the random design and noise that is specified by the sample complexity parameter $\theta(n, p, s) := n/[2\psi(B^*)\log(p - s)]$. Here n is the sample size, and $\psi(B^*)$ is a *sparsity-overlap function* measuring a combination of the sparsities and overlaps of the K-regression coefficient vectors that constitute the model. We prove that the multivariate group Lasso succeeds for problem sequences (n, p, s) such that $\theta(n, p, s)$ exceeds a critical level θ_u, and fails for sequences such that $\theta(n, p, s)$ lies below a critical level θ_ℓ. For the special case of the standard Gaussian ensemble, we show that $\theta_\ell = \theta_u$ so that the characterization is sharp. The sparsity-overlap function $\psi(B^*)$ reveals that, if the design is uncorrelated on the active rows, ℓ_1/ℓ_2 regularization for multivariate regression never harms performance relative to an ordinary Lasso approach and can yield substantial improvements in sample complexity (up to a factor of K) when the coefficient vectors are suitably orthogonal. For more general designs, it is possible for the ordinary Lasso to outperform the multivariate group Lasso. We complement our analysis with simulations that demonstrate the sharpness of our theoretical results, even for relatively small problems.

1. Introduction. The development of efficient algorithms for estimation of large-scale models has been a major goal of statistical learning research in the last decade. There is now a substantial body of work based on ℓ_1-regularization dating back to the seminal work of Tibshirani (1996) and Donoho and collaborators [Chen, Donoho and Saunders (1998); Donoho and Huo (2001)]. The bulk of

Received June 2009; revised November 2009.

[1] Supported in part by NSF Grants DMS-06-05165 and CCF-05-45862 to MJW, and by NSF Grant 05-09559 and DARPA IPTO Contract FA8750-05-2-0249 to MIJ.

AMS 2000 subject classifications. Primary 62J07; secondary 62F07.

Key words and phrases. LASSO, block-norm, second-order cone program, sparsity, variable selection, multivariate regression, high-dimensional scaling, simultaneous Lasso, group Lasso.

this work has focused on the standard problem of linear regression, in which one makes observations of the form

$$y = X\beta^* + w, \tag{1}$$

where $y \in \mathbb{R}^n$ is a real-valued vector of observations, $w \in \mathbb{R}^n$ is an additive zero-mean noise vector and $X \in \mathbb{R}^{n \times p}$ is the design matrix. A subset of the components of the unknown parameter vector $\beta^* \in \mathbb{R}^p$ are assumed nonzero; the goal is to identify these coefficients and (possibly) estimate their values. This goal can be formulated in terms of the solution of the penalized optimization problem

$$\arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \|y - X\beta\|_2^2 + \lambda_n \|\beta\|_0 \right\}, \tag{2}$$

where $\|\beta\|_0$ counts the number of nonzero components in β and where $\lambda_n > 0$ is a regularization parameter. Unfortunately, this optimization problem is computationally intractable, a fact which has led various authors to consider the convex relaxation [Tibshirani (1996); Chen, Donoho and Saunders (1998)]

$$\arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \|y - X\beta\|_2^2 + \lambda_n \|\beta\|_1 \right\}, \tag{3}$$

in which $\|\beta\|_0$ is replaced with the ℓ_1 norm $\|\beta\|_1$. This relaxation, often referred to as the Lasso [Tibshirani (1996)], is a quadratic program, and can be solved efficiently by various methods [e.g., Boyd and Vandenberghe (2004); Osborne, Presnell and Turlach (2000); Efron et al. (2004)].

A variety of theoretical results are now in place for the Lasso, both in the traditional setting where the sample size n tends to infinity with the problem size p fixed [Knight and Fu (2000)], as well as under high-dimensional scaling, in which p and n tend to infinity simultaneously, thereby allowing p to be comparable to or even larger than n [e.g., Meinshausen and Bühlmann (2006); Wainwright (2009b); Meinshausen and Yu (2009); Bickel, Ritov and Tsybakov (2009)]. In many applications, it is natural to impose *sparsity constraints* on the regression vector β^*, and a variety of such constraints have been considered. For example, one can consider a "hard sparsity" model in which β^* is assumed to contain at most s nonzero entries or a "soft sparsity" model in which β^* is assumed to belong to an ℓ_q ball with $q < 1$. Analyses also differ in terms of the loss functions that are considered. For the model or variable selection problem, it is natural to consider the zero–one loss associated with the problem of recovering the unknown support set of β^*. Alternatively, one can view the Lasso as a shrinkage estimator to be compared to traditional least squares or ridge regression; in this case, it is natural to study the ℓ_2-loss $\|\widehat{\beta} - \beta^*\|_2$ between the estimate $\widehat{\beta}$ and the ground truth. In other settings, the prediction error $\mathbb{E}[(Y - X^T \widehat{\beta})^2]$ may be of primary interest, and one tries to show risk consistency (namely, that the estimated model predicts as well as the best sparse model, whether or not the true model is sparse).

A number of alternatives to the Lasso have been explored in recent years, and in some cases stronger theoretical results have been obtained [Fan and Li (2001); Frank and Friedman (1993); Huang, Horowitz and Ma (2008)]. However, the resulting optimization problems are generally nonconvex and thus difficult to solve in practice. The Lasso remains a focus of attention due to its combination of favorable statistical and computational properties.

1.1. *Block-structured regularization.* While the assumption of sparsity at the level of individual coefficients is one way to give meaning to high-dimensional ($p \gg n$) regression, there are other structural assumptions that are natural in regression, and which may provide additional leverage. For instance, in a hierarchical regression model, groups of regression coefficients may be required to be zero or nonzero in a blockwise manner; for example, one might wish to include a particular covariate and all powers of that covariate as a group [Yuan and Lin (2006); Zhao, Rocha and Yu (2009)]. Another example arises when we consider variable selection in the setting of multivariate regression: multiple regressions can be related by a (partially) shared sparsity pattern, such as when there are an underlying set of covariates that are "relevant" across regressions [Obozinski, Taskar and Jordan (2010); Argyriou, Evgeniou and Pontil (2006); Turlach, Venables and Wright (2005); Zhang et al. (2008)]. Based on such motivations, a recent line of research [Bach, Lanckriet and Jordan (2004); Tropp (2006); Yuan and Lin (2006); Zhao, Rocha and Yu (2009); Obozinski, Taskar and Jordan (2010); Ravikumar et al. (2009)] has studied the use of *block-regularization schemes*, in which the ℓ_1 norm is composed with some other ℓ_q norm ($q > 1$), thereby obtaining the ℓ_1/ℓ_q norm defined as a sum of ℓ_q norms over groups of regression coefficients. The best known examples of such block norms are the ℓ_1/ℓ_∞ norm [Turlach, Venables and Wright (2005); Zhang et al. (2008)] and the ℓ_1/ℓ_2 norm [Obozinski, Taskar and Jordan (2010)].

In this paper, we investigate the use of ℓ_1/ℓ_2 block-regularization in the context of high-dimensional multivariate linear regression, in which a collection of K scalar outputs are regressed on the same design matrix $X \in \mathbb{R}^{n\times p}$. Representing the regression coefficients as an $p \times K$ matrix B^*, the multivariate regression model takes the form

$$Y = XB^* + W, \tag{4}$$

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. In addition, we assume a hard-sparsity model for the regression coefficients in which column k of the coefficient matrix B^* has nonzero entries on a subset

$$S_k := \{i \in \{1, \dots, p\} \mid \beta^*_{ik} \neq 0\} \tag{5}$$

of size $s_k := |S_k|$. In many applications it is natural to expect that the supports S_k should overlap. In that case, instead of estimating the support of each regression separately, it might be beneficial to first estimate the set of variables which are relevant to any of the multivariate responses and to estimate only subsequently the individual supports within that set. Thus we focus on the problem of recovering the union of the supports, namely the set $S := \bigcup_{k=1}^{K} S_k$, corresponding to the subset of indices $i \in \{1, \dots, p\}$ that are involved in at least one regression. We consider a range of problems in which variables can be relevant to all, some, only one or none of the regressions, and we investigate if and how the overlap of the individual supports and the relatedness of individual regressions benefit or hinder estimation of the *support union*.

The *support union problem* can be understood as the generalization of the problem of variable selection to the group setting. Rather than selecting specific components of a coefficient vector, we aim to select specific rows of a coefficient matrix. We thus also refer to the support union problem as the *row selection problem*. We note that recovering S, although not equivalent to recovering each of the distinct individual supports S_k, addresses the essential difficulty in recovering those supports. Indeed, as we show in Section 2.2, given a method that returns the row support S with $|S| \ll p$ (with high probability), it is straightforward to recover the individual supports S_k by ordinary least-squares and thresholding.

If computational complexity were not a concern, the natural way to perform row selection for B^* would be by solving the optimization problem

$$\arg \min_{B \in \mathbb{R}^{p \times K}} \left\{ \frac{1}{2n} |||Y - XB|||_F^2 + \lambda_n \|B\|_{\ell_0/\ell_q} \right\}, \tag{6}$$

where $B = (\beta_{ik})_{1 \le i \le p, 1 \le k \le K}$ is a $p \times K$ matrix, the quantity $||| \cdot |||_F$ denotes the Frobenius norm,[1] and the "norm" $\|B\|_{\ell_0/\ell_q}$ counts the number of rows in B that have nonzero ℓ_q norm. As before, the ℓ_0 component of this regularizer yields a nonconvex and computationally intractable problem, so that it is natural to consider the relaxation

$$\arg \min_{B \in \mathbb{R}^{p \times K}} \left\{ \frac{1}{2n} |||Y - XB|||_F^2 + \lambda_n \|B\|_{\ell_1/\ell_q} \right\}, \tag{7}$$

where $\|B\|_{\ell_1/\ell_q}$ is the block ℓ_1/ℓ_q norm

$$\|B\|_{\ell_1/\ell_q} := \sum_{i=1}^{p} \left(\sum_{j=1}^{K} {\beta_{ij}}^q \right)^{1/q} = \sum_{i=1}^{p} \|\beta_i\|_q. \tag{8}$$

The relaxation (7) is a natural generalization of the Lasso; indeed, it specializes to the Lasso in the case $K = 1$. For later reference, we also note that setting $q = 1$ leads to the use of the ℓ_1/ℓ_1 block norm in the relaxation (7). Since this norm decouples across both the rows and columns, this particular choice is equivalent to solving K separate Lasso problems, one for each column of the $p \times K$

[1]The Frobenius norm of a matrix A is given by $|||A|||_F := \sqrt{\sum_{i,j} A_{ij}^2}$.

regression matrix B^*. A more interesting choice is $q = 2$, which yields a block ℓ_1/ℓ_2 norm that couples together the columns of B. Regularization with the ℓ_1/ℓ_2 norm is commonly referred to as the *group Lasso* in the setting of univariate regression [Yuan and Lin (2006)]. We thus refer to ℓ_1/ℓ_2 regularization in the multivariate setting as the *multivariate group Lasso*. Note that the multivariate group Lasso can be viewed as a special case of the group Lasso, in that it involves a specific grouping of regression coefficients, but the multivariate setting brings new statistical issues to the fore. As we discuss in Appendix B, the multivariate group Lasso can be cast as a *second-order cone program* (SOCP). This is a family of convex optimization problems that can be solved efficiently with interior point methods [Boyd and Vandenberghe (2004)] and includes quadratic programs as a particular case.

Some recent work has addressed certain statistical aspects of block-regularization schemes. Meier, van de Geer and Bühlmann (2008) perform an analysis of risk consistency with block-norm regularization. Bach (2008) provides an analysis of block-wise support recovery for the kernelized group Lasso in the classical, fixed p setting. In the high-dimensional setting, Ravikumar et al. (2009) studies the consistency of block-wise support recovery for the group Lasso for fixed design matrices, and their result is generalized by Liu and Zhang (2008) to block-wise support recovery in the setting of general ℓ_1/ℓ_q regularization, again for fixed design matrices. However, these analyses do not discriminate between various values of q, yielding the same qualitative results and the same convergence rates for $q = 1$ as for $q > 1$. Our focus, which is motivated by the empirical observation that the group Lasso and the multivariate group Lasso can outperform the ordinary Lasso [Bach (2008); Yuan and Lin (2006); Zhao, Rocha and Yu (2009); Obozinski, Taskar and Jordan (2010)], is precisely the distinction between $q = 1$ and $q > 1$ (specifically $q = 2$). We note that in concurrent work Negahban and Wainwright (2008) have studied a related problem of support recovery for the ℓ_1/ℓ_∞ relaxation.

The distinction between $q = 1$ and $q = 2$ is also significant from an optimization-theoretic point of view. In particular, the SOCP relaxations underlying the multivariate group Lasso ($q = 2$) are generally tighter than the quadratic programming relaxation underlying the Lasso ($q = 1$); however, the improved accuracy is generally obtained at a higher computational cost [Boyd and Vandenberghe (2004)]. Thus we can view our problem as an instance of the general question of the relationship of statistical efficiency to computational efficiency: does the qualitatively greater amount of computational effort involved in solving the multivariate group Lasso always yield greater statistical efficiency? More specifically, can we give theoretical conditions under which solving the generalized Lasso problem (7) has greater statistical efficiency than naive strategies based on the ordinary Lasso? Conversely, can the multivariate group Lasso ever be worse than the ordinary Lasso?

With this motivation, this paper provides a detailed analysis of model selection consistency of the multivariate group Lasso (7) with ℓ_1/ℓ_2-regularization. Statistical efficiency is defined in terms of the scaling of the sample size n, as a function of the problem size p and of the sparsity structure of the regression matrix B^*, required for consistent row selection. Our analysis is high-dimensional in nature, allowing both n and p to diverge, and yielding explicit error bounds as a function of p. As detailed below, our analysis provides affirmative answers to both of the questions above. First, we demonstrate that under certain structural assumptions on the design and regression matrix B^*, the multivariate group Lasso is always guaranteed to outperform the ordinary Lasso, in that it correctly performs row selection for sample sizes for which the Lasso fails with high probability. Second, we also exhibit some problems (though arguably not generic) for which the multivariate group Lasso will be outperformed by the naive strategy of applying the Lasso separately to each of the K columns, and taking the union of supports.

1.2. *Our results*. 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 (7) is governed by the *sample complexity function*

$$\theta_{\ell_1/\ell_2}(n, p; B^*) := \frac{n}{2\psi(B^*)\log(p-s)}, \tag{9}$$

where n is the sample size, p is the ambient dimension, $s = |S|$ is the number of rows that are nonzero and $\psi(\cdot)$ is a *sparsity-overlap function*. Our use of the term "sample complexity" for θ_{ℓ_1/ℓ_2} reflects the role it plays in our analysis as the rate at which the sample size must grow in order to obtain consistent row selection as a function of the problem parameters. More precisely, for scalings (n, p, s, B^*) such that $\theta_{\ell_1/\ell_2}(n, p; B^*)$ exceeds a fixed critical threshold $\theta_u \in (0, +\infty)$, we show that the probability of correct row selection by the ℓ_1/ℓ_2 multivariate group Lasso converges to one, and conversely, for scalings such that $\theta_{\ell_1/\ell_2}(n, p; B^*)$ is below another threshold θ_ℓ, we show that the multivariate group Lasso fails with high probability.

Whereas the ratio $(\log p)/n$ is standard for the high-dimensional theory of ℓ_1-regularization, the function $\psi(B^*)$ is a novel and interesting quantity, one which measures both the sparsity of the matrix B^* as well as the overlap between the different regressions, represented by the columns of B^* [see equation (16) for the precise definition of $\psi(B^*)$]. As a particular illustration, consider the special case of a univariate regression with $K = 1$, in which the convex program (7) reduces to the ordinary Lasso (3). In this case, if the design matrix is drawn from the standard Gaussian ensemble [i.e., $X_{ij} \sim N(0, 1)$, i.i.d.], we show that the sparsity-overlap function reduces to $\psi(B^*) = s$, corresponding to the support size of the single coefficient vector. We thus recover as a corollary a previously known result [Wainwright (2009b)]: namely, the Lasso succeeds in performing exact support recovery once the ratio $n/[s\log(p-s)]$ exceeds a certain critical threshold. At the other extreme, for a genuinely multivariate problem with $K > 1$ and s nonzero rows, again for a standard Gaussian design, when the regression matrix is "suitably orthonormal" relative to the design (see Section 2 for a precise definition), the sparsity-overlap function is given by $\psi(B^*) = s/K$. In this case, ℓ_1/ℓ_2 block-regularization has sample complexity lower by a factor of K relative to the naive approach of solving K separate Lasso problems. Of course, there is also a range of behavior between these two extremes, in which the gain in sample complexity varies smoothly as a function of the sparsity-overlap $\psi(B^*)$ in the interval $[\frac{s}{K}, s]$. On the other hand, we also show that for suitably correlated designs, it is possible that the sample complexity $\psi(B^*)$ associated with ℓ_1/ℓ_2 block-regularization is larger than that of the ordinary Lasso (ℓ_1/ℓ_1) approach.

The remainder of the paper is organized as follows. In Section 2, we provide a precise statement of our main results (Theorems 1 and 2), discuss some of their consequences and illustrate the close agreement between our theoretical results and simulations. Sections 3 and 4 are devoted to the proofs of Theorems 1 and 2, respectively, with the arguments broken down into a series of steps. More technical results are deferred to the appendices. We conclude with a brief discussion in Section 5.

1.3. *Notation.* We collect here some notation used throughout the paper. For a (possibly random) matrix $M \in \mathbb{R}^{p\times K}$, we define the Frobenius norm $|||M|||_F := (\sum_{i,j} m_{ij}^2)^{1/2}$, and for parameters $1 \le a \le b \le \infty$, the ℓ_a/ℓ_b block norm is defined as follows:

$$\|M\|_{\ell_a/\ell_b} := \left\{\sum_{i=1}^{p}\left(\sum_{k=1}^{K}|m_{ik}|^b\right)^{a/b}\right\}^{1/a}. \tag{10}$$

These vector norms on matrices should be distinguished from the (a,b)-operator norms

$$|||M|||_{a,b} := \sup_{\|x\|_b=1} \|Mx\|_a \tag{11}$$

(although some norms belong to both families; see Lemma 7 in Appendix E). Important special cases of the latter include the spectral norm $|||M|||_{2,2}$ (also denoted $|||M|||_2$), and the ℓ_∞-operator norm $|||M|||_{\infty,\infty} = \max_{i=1,\dots,p} \sum_{j=1}^{K} |M_{ij}|$, denoted $|||M|||_\infty$ for short.

In addition to the usual Landau notation $\mathcal{O}$ and o, we write $a_n = \Omega(b_n)$ for sequences such that $\frac{b_n}{a_n} = o(1)$. We also use the notation $a_n = \Theta(b_n)$ if both $a_n = \mathcal{O}(b_n)$ and $b_n = \mathcal{O}(a_n)$ hold.

2. Main results and some consequences. The analysis of this paper considers the multivariate group Lasso estimator, obtained as a solution to the SOCP

$$\arg\min_{B\in\mathbb{R}^{p\times K}} \left\{\frac{1}{2n}|||Y - XB|||_F^2 + \lambda_n \|B\|_{\ell_1/\ell_2}\right\} \tag{12}$$

for random ensembles of multivariate linear regression problems, each of the form (4), where the noise matrix $W \in \mathbb{R}^{n \times K}$ is assumed to consist of i.i.d. elements $W_{ij} \sim N(0, \sigma^2)$. We consider random design matrices X with each row 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. Although the block-regularized problem (12) need not have a unique solution in general, a consequence of our analysis is that in the regime of interest, the solution is unique, so that we may talk unambiguously about the estimated support $\widehat{S}$. The main object of study in this paper is the probability $\mathbb{P}[\widehat{S} = S]$, where the probability is taken both over the random choice of noise matrix W and random design matrix X. We study the behavior of this probability as elements of the triplet (n, p, s) tend to infinity.

2.1. *Notation and assumptions.* More precisely, our main result applies to sequences of models indexed by $(n, p(n), s(n))$, an associated sequence of $p \times p$ covariance matrices and a sequence $\{B^*\}$ of coefficient matrices with row support

$$S := \{i \mid \beta_i^* \neq 0\} \tag{13}$$

of size $|S| = s = s(n)$. We use S^c to denote its complement (i.e., $S^c := \{1, \dots, p\} \backslash S$). We let

$$b^*_{\min} := \min_{i \in S} \|\beta_i^*\|_2 \tag{14}$$

correspond to the minimal ℓ_2 row-norm of the coefficient matrix B^* over its nonzero rows. Given an observed pair (Y, X) from the model (4), the goal is to estimate the row support S of the matrix B^*.

We impose the following conditions on the covariance Σ of the design matrix:

(A1) *Bounded eigenspectrum*: There exist fixed constants $C_{\min} > 0$ and $C_{\max} < +\infty$ such that all eigenvalues of the $s \times s$ matrix Σ_{SS} are contained in the interval $[C_{\min}, C_{\max}]$.

(A2) *Irrepresentable condition*: There exists a fixed parameter $\gamma \in (0, 1]$ such that

$$|||\Sigma_{S^cS}(\Sigma_{SS})^{-1}|||_\infty \leq 1 - \gamma.$$

(A3) *Self-incoherence*: $|||(\Sigma_{SS})^{-1}|||_\infty \leq D_{\max}$ for some $D_{\max} < +\infty$.

The lower bound involving $C_{\min}$ in assumption (A1) prevents excess dependence among elements of the design matrix associated with the support S; conditions of this form are required for model selection consistency or ℓ_2 consistency of the Lasso. The upper bound involving $C_{\max}$ in assumption (A1) is not needed for proving success but only failure of the multivariate group Lasso. The irrepresentable condition and self-incoherence assumptions are also well known from previous work on variable selection consistency of the Lasso [Meinshausen and Bühlmann (2006); Tropp (2006); Zhao and Yu (2006)]. Although such assumptions are not needed in analyzing ℓ_2 or risk consistency, they are known to be necessary for variable selection consistency of the Lasso. Indeed, in the absence of such conditions, it is always possible to make the Lasso fail, even with an arbitrarily large sample size. [See, however, Meinshausen and Yu (2009) for methods that weaken the irrepresentable condition.] Note that these assumptions are trivially satisfied by the standard Gaussian ensemble $\Sigma = I_{p\times p}$, with $C_{\min} = C_{\max} = 1$, $D_{\max} = 1$ and $\gamma = 1$. More generally, it can be shown that various matrix classes (e.g., Toeplitz matrices, tree-structured covariance matrices, bounded off-diagonal matrices) satisfy these conditions [Meinshausen and Bühlmann (2006); Zhao and Yu (2006); Wainwright (2009b)].

We require a few pieces of notation before stating the main results. For an arbitrary matrix $B_S \in \mathbb{R}^{s\times K}$ with ith row $\beta_i \in \mathbb{R}^{1\times K}$, we define the matrix $\zeta(B_S) \in \mathbb{R}^{s\times K}$ with ith row

$$\zeta(\beta_i) := \frac{\beta_i}{\|\beta_i\|_2}, \tag{15}$$

when $\beta_i \neq 0$, and we set $\zeta(\beta_i) = 0$ otherwise. With this notation, the *sparsity-overlap function* is given by

$$\psi(B) := |||\zeta(B_S)^T(\Sigma_{SS})^{-1}\zeta(B_S)|||_2, \tag{16}$$

where $|||\cdot|||_2$ denotes the spectral norm. We use this sparsity-overlap function to define the *sample complexity parameter*, which captures the effective sample size

$$\theta_{\ell_1/\ell_2}(n, p; B^*) := \frac{n}{2\psi(B^*)\log(p-s)}. \tag{17}$$

In the following two theorems, we consider a random design matrix X drawn with i.i.d. $N(0, \Sigma)$ row vectors, where Σ satisfies assumptions (A1)–(A3), and an observation matrix Y specified by model (4). In order to capture dependence induced by the design covariance matrix, for any positive semidefinite matrix $Q \succeq 0$, we define the quantities

$$\rho_\ell(Q) := \frac{1}{2}\min_{i\neq j}[Q_{ii} + Q_{jj} - 2Q_{ij}] \tag{18a}$$

and

$$\rho_u(Q) := \max_i Q_{ii}. \tag{18b}$$

We note that by definition, we have $\rho_\ell(Q) \le \rho_u(Q)$ whenever $Q \succeq 0$. Our bounds are stated in terms of these quantities as applied to the conditional covariance matrix

$$\Sigma_{S^cS^c|S} := \Sigma_{S^cS^c} - \Sigma_{S^cS}(\Sigma_{SS})^{-1}\Sigma_{SS^c}.$$

Our first result is an achievability result, showing that the multivariate group Lasso succeeds in recovering the row support and yields consistency in ℓ_∞/ℓ_2 norm. We state this result for sequences of regularization parameters $\lambda_n =$

$\sqrt{\frac{f(p)\log p}{n}}$, where $f(p) \underset{p\to+\infty}{\longrightarrow} +\infty$ is any function such that $\lambda_n \to 0$. We also assume that n is sufficiently large such that $s/n < 1/2$.

THEOREM 1. *Suppose that we solve the multivariate group Lasso with specified regularization parameter sequence λ_n for a sequence of problems indexed by (n, p, B^*, Σ) that satisfy assumptions* (A1)–(A3), *and such that, for some $\nu > 0$,*

$$\theta_{\ell_1/\ell_2}(n, p; B^*) = \frac{n}{2\psi(B^*)\log(p-s)} > (1+\nu)\frac{\rho_u(\Sigma_{S^cS^c|S})}{\gamma^2}. \tag{19}$$

Then for universal constants $c_i > 0$ (i.e., independent of n, p, s, B^*, Σ), with probability greater than $1 - c_2\exp(-c_3 K\log s) - c_0\exp(-c_1\log(p-s))$, the following statements hold:

(a) *The multivariate group Lasso has a unique solution $\widehat{B}$ with row support $S(B)$ that is contained within the true row support $S(B^*)$, and moreover satisfies the bound*

$$\|\widehat{B} - B^*\|_{\ell_\infty/\ell_2} \le \underbrace{\sqrt{\frac{8K\log s}{C_{\min} n}} + \lambda_n D_{\max} + \frac{6\lambda_n}{C_{\min}}\sqrt{\frac{s^2}{n}}}_{\rho(n,s,\lambda_n)}. \tag{20}$$

(b) *If $\frac{\rho}{b^*_{\min}} = o(1)$, the estimate of the row support, $S(\widehat{B}) := \{i \in \{1, \dots, p\} \mid \widehat{\beta}_i \neq 0\}$, specified by this unique solution is equal to the row support set $S(B^*)$ of the true model.*

Note that the theorem is naturally separated into two distinct but related claims. Part (a) guarantees that the method produces *no false inclusions* and, moreover, bounds the maximum ℓ_2-error across the rows. Part (b) requires some additional assumptions—namely, the restriction $\frac{\rho}{b^*_{\min}} = o(1)$ ensuring that the error ρ is of lower order than the minimum ℓ_2-norm $b^*_{\min}$ across rows—but also guarantees the stronger result of no false exclusions as well, so that the method recovers the row support exactly. Note that the probability of these events converges to one only if both $(p - s)$ and s tend to infinity, which might seem counter-intuitive initially (since problems with larger support sets s might seem harder). However, as we discuss at the end of Section 3.3, this dependence can be removed at the expense of a slightly slower convergence rate for $\|\widehat{B} - B^*\|_{\ell_\infty/\ell_2}$.

Our second main theorem is a negative result, showing that the multivariate group Lasso fails with high probability if the rescaled sample size θ_{ℓ_1/ℓ_2} is below a critical threshold. In order to clarify the phrasing of this result, note that Theorem 1 can be summarized succinctly as guaranteeing that there is a unique solution $\widehat{B}$ with the correct row support that satisfies $\|\widehat{B} - B^*\|_{\ell_\infty/\ell_2} = o(b^*_{\min})$. The following result shows that such a guarantee cannot hold if the sample size n scales too slowly relative to p, s and the other problem parameters.

THEOREM 2. *Consider problem sequences indexed by* (n, p, B^*, Σ) *that satisfy assumptions* (A1)–(A2), *and with minimum value* $b^*_{\min}$ *such that* ${b^*_{\min}}^2 = \Omega(\frac{\log p}{n})$, *and suppose that we solve the multivariate group Lasso with any positive regularization sequence* $\{\lambda_n\}$. *Then there exist* $\nu > 0$ *and universal constants* $c_i > 0$ *such that if the sample size is lower bounded as*

$$\theta_{\ell_1/\ell_2}(n, p; B^*) = \frac{n}{2\psi(B^*)\log(p-s)} < (1-\nu)\frac{\rho_\ell(\Sigma_{S^cS^c|S})}{(2-\gamma)^2},$$

then with probability greater than $1 - c_0 \exp\{-c_1 \min(\frac{Kn}{s}, \frac{\theta_\ell}{2}\log(p-s))\}$, *there is no solution* $\widehat{B}$ *of the multivariate group Lasso that has the correct row support and satisfies the bound* $\|\widehat{B} - B^*\|_{\ell_\infty/\ell_2} = o(b^*_{\min})$.

The proof of this claim is provided in Section 4. We note that information-theoretic methods [Wainwright (2009a)] imply that no method (including the multivariate group Lasso) can perform exact support recovery unless $n/s \to +\infty$, so that the probability given in Theorem 2 converges to one under the given conditions. Note that Theorems 1 and 2 in conjunction imply that the rescaled sample size $\theta_{\ell_1/\ell_2}(n, p; B^*) = \frac{n}{2\psi(B^*)\log(p-s)}$ captures the behavior of the multivariate group Lasso for support recovery and estimation in block ℓ_∞/ℓ_2 norm. For the special case of random design matrices drawn from the standard Gaussian ensemble (i.e., $\Sigma = I_{p\times p}$), the given scalings are sharp:

COROLLARY 1. *For the standard Gaussian ensemble*, *the multivariate group Lasso undergoes a sharp threshold at the level* $\theta_{\ell_1/\ell_2}(n, p, B^*) = 1$. *More specifically*, *for any* $\delta > 0$:

(a) *For problem sequences* (n, p, B^*) *such that* $\theta_{\ell_1/\ell_2}(n, p, B^*) > 1 + \delta$, *the multivariate group Lasso succeeds with high probability.*
(b) *Conversely*, *for sequences such that* $\theta_{\ell_1/\ell_2}(n, p, B^*) < 1 - \delta$, *the multivariate group Lasso fails with high probability.*

PROOF. In the special case $\Sigma = I_{p\times p}$, it is straightforward to verify that all the assumptions are satisfied: in particular, we have $C_{\min} = C_{\max} = 1$, $D_{\max} = 1$ and $\gamma = 1$. Moreover, a short calculation shows that $\rho_u(I) = \rho_\ell(I) = 1$. Consequently, the thresholds given in the sufficient condition (19) and the necessary condition (21) are both equal to one. □

2.2. *Efficient estimation of individual supports*. The preceding results address exact recovery of the *support union* of the regression matrix B^*. As demonstrated by the following procedure and the associated corollary of Theorem 1, once the row support has been recovered, it is straightforward to recover the individual supports of each column of the regression matrix via the additional steps of performing ordinary least squares and thresholding.

Efficient multi-stage estimation of individual supports:

(1) estimate the *support union* with $\hat{S}$, the support union of the solution $\widehat{B}$ of the multivariate group Lasso;
(2) compute the restricted ordinary least squares (ROLS) estimate,

$$\widetilde{B}_{\hat{S}} := \arg\min_{B_{\hat{S}}} |||Y - X_{\hat{S}} B_{\hat{S}}|||_F \tag{21}$$

for the restricted multivariate problem;
(3) compute the matrix $T(\widetilde{B}_{\hat{S}})$ obtained by thresholding $\widetilde{B}_{\hat{S}}$ at the level $2\sqrt{\frac{2\log(K|\hat{S}|)}{C_{\min} n}}$, and estimate the individual supports by the nonzero entries of $T(\widetilde{B}_{\hat{S}})$.

The following result, which is proved in Appendix A, shows that under the assumptions of Theorem 1, the additional post-processing applied to the support union estimate will recover the individual supports with high probability:

COROLLARY 2. *Under assumptions* (A1)–(A3) *and the additional assumptions of Theorem* 1, *if for all individual nonzero coefficients* $\beta^*_{ik}, i \in S, 1 \leq k \leq K$, *we have* $|\beta^*_{ik}| \geq 2\sqrt{\frac{4\log(Ks)}{C_{\min} n}}$, *then with probability greater than* $1 - \Theta(\exp(-c_0 K \log s))$ *the above two-step estimation procedure recovers the individual supports of* B^*.

2.3. *Some consequences of Theorems* 1 *and* 2. We begin by making some simple observations about the sparsity-overlap function.

LEMMA 1. (a) *For any design satisfying assumption* (A1), *the sparsity-overlap* $\psi(B^*)$ *obeys the bounds*

$$\frac{s}{C_{\max} K} \leq \psi(B^*) \leq \frac{s}{C_{\min}}; \tag{22}$$

(b) *if* $\Sigma_{SS} = I_{s \times s}$, *and if the columns* $(Z^{(k)*})$ *of the matrix* $Z^* = \zeta(B^*)$ *are orthogonal*, *then the sparsity-overlap function is* $\psi(B^*) = \max_{k=1,\dots,K} \|Z^{(k)*}\|_2^2$.

The proof of this claim is provided in Appendix C. Based on this lemma, we now study some special cases of Theorems 1 and 2. The simplest special case is the univariate regression problem ($K = 1$), in which case the quantity $\zeta(\beta^*)$ [as defined in equation (15)] simply yields an s-dimensional sign vector with elements $z^*_i = \operatorname{sign}(\beta^*_i)$. [Recall that the sign function is defined as $\operatorname{sign}(0) = 0$, $\operatorname{sign}(x) = 1$

if $x > 0$ and $\operatorname{sign}(x) = -1$ if $x < 0$.] In this case, the sparsity-overlap function is given by $\psi(\beta^*) = {z^*}^T(\Sigma_{SS})^{-1}z^*$, and as a consequence of Lemma 1(a), we have $\psi(\beta^*) = \Theta(s)$. Consequently, a simple corollary of Theorems 1 and 2 is that the Lasso succeeds once the ratio $n/(2s\log(p-s))$ exceeds a certain critical threshold, determined by the eigenspectrum and incoherence properties of Σ, and it fails below a certain threshold. This result matches the necessary and sufficient conditions established in previous work on the Lasso [Wainwright (2009b)].

We can also use Lemma 1 and Theorems 1 and 2 to compare the performance of the multivariate group Lasso to the following (arguably naive) strategy for row selection using the ordinary Lasso.

Row selection using ordinary Lasso:

(1) Apply the ordinary Lasso separately to each of the K univariate regression problems specified by the columns of B^*, thereby obtaining estimates $\widehat{\beta}^{(k)}$ for $k = 1, \dots, K$.
(2) For $k = 1, \dots, K$, estimate the support of individual columns via $\widehat{S}_k := \{i \mid \widehat{\beta}_i^{(k)} \neq 0\}$.
(3) Estimate the row support by taking the union: $\widehat{S} = \bigcup_{k=1}^K \widehat{S}_k$.

To understand the conditions governing the success/failure of this procedure, note that it succeeds if and only if for each nonzero row $i \in S = \bigcup_{k=1}^K S_k$, the variable $\widehat{\beta}_i^{(k)}$ is nonzero for at least one k, and for all $j \in S^c = \{1, \dots, p\} \backslash S$, the variable $\widehat{\beta}_j^{(k)} = 0$ for all $k = 1, \dots, K$. From our understanding of the univariate case, we know that for $\theta_u = \frac{\rho_u(\Sigma_{S^cS^c|S})}{\gamma^2}$, the condition

$$n \geq 2\theta_u \max_{k=1,\dots,K} \psi(\beta_S^{*(k)}) \log(p - s_k) \geq 2\theta_u \max_{k=1,\dots,K} \psi(\beta_S^{*(k)}) \log(p - s) \tag{23}$$

is sufficient to ensure that the ordinary Lasso succeeds in row selection. Conversely, for $\theta_\ell = \frac{\rho_\ell(\Sigma_{S^cS^c|S})}{(2-\gamma)^2}$, if the sample size is upper bounded as

$$n < 2\theta_\ell \max_{k=1,\dots,K} \psi(\beta_S^{*(k)}) \log(p - s),$$

then there will exist some $j \in S^c$ such for at least one $k \in \{1, \dots, K\}$, there holds $\widehat{\beta}_j^{(k)} \neq 0$ with high probability, implying failure of the ordinary Lasso.

A natural question is whether the multivariate group Lasso, by taking into account the couplings across columns, always outperforms (or at least matches) the naive strategy. The following result, proven in Appendix D, shows that if the design is uncorrelated on the support, then indeed this is the case.

COROLLARY 3 (Multivariate group Lasso versus ordinary Lasso). *Assume $\Sigma_{SS} = I_{s \times s}$. Then for any multivariate regression problem, row selection using* *the ordinary Lasso strategy requires, with high probability, at least as many samples as the ℓ_1/ℓ_2 multivariate group Lasso. In particular, the relative efficiency of multivariate group Lasso versus ordinary Lasso is given by the ratio*

$$\frac{\max_{k=1,\dots,K} \psi(\beta_S^{*(k)}) \log(p - s_k)}{\psi(B_S^*) \log(p - s)} \geq 1. \tag{24}$$

We consider the special case of identical regressions, for which a result can be stated for any covariance design and the case of "orthonormal" regressions which illustrates Corollary 3.

EXAMPLE 1 (Identical regressions). 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^* being an s-dimensional sign vector with elements $z_i^* = \text{sign}(\beta_i^*)$. Consequently, we have the equality $\psi(B^*) = \psi(\beta^*)$, so that under our analysis there is no benefit in using the multivariate group Lasso relative to the strategy of solving separate Lasso problems and constructing the union of individually estimated supports. This behavior may seem counter-intuitive, since under the model (4) we essentially have Kn observations of the coefficient vector β^* with the same design matrix but K independent noise realizations, which could help to reduce the effective noise variance from σ^2 to σ^2/K if the fact that the regressions are identical is known. It must be borne in mind, however, that in our high-dimensional analysis the noise variance does not grow as the dimensionality grows, and thus asymptotically the noise is dominated by the interference between the covariates, which grows as $(p-s)$. It is thus this high-dimensional interference that dominates the rates given in Theorems 1 and 2.

In contrast to this pessimistic example, we now turn to the most optimistic extreme:

EXAMPLE 2 ("Orthonormal" regressions). Suppose that $\Sigma_{SS} = I_{s\times s}$ and (for $s > K$) suppose that B^* is constructed such that the columns of the $s \times K$ matrix $\zeta(B^*)$ are orthogonal and with equal norm (which implies their norm equals $\sqrt{\frac{s}{K}}$). Under these conditions, we claim that the sample complexity of multivariate group Lasso is smaller than that of the ordinary Lasso by a factor of $1/K$. Indeed, using Lemma 1(b), we observe that

$$K\psi(B^*) = K\|Z^{(1)*}\|^2 = \sum_{k=1}^{K} \|Z^{(k)*}\|^2 = \text{tr}(Z^{*T} Z^*) = \text{tr}(Z^* Z^{*T}) = s,$$

because $Z^* Z^{*T} \in \mathbb{R}^{s\times s}$ is the Gram matrix of s unit vectors in $\mathbb{R}^k$ and its diagonal elements are therefore all equal to 1. Consequently, the multivariate group Lasso recovers the row support with high probability for sequences such that

$$\frac{n}{2(s/K)\log(p-s)} > 1,$$

which allows for sample sizes K times smaller than the ordinary Lasso approach.

Corollary 3 and the subsequent examples address the case of uncorrelated design ($\Sigma_{SS} = I_{s\times s}$) on the row support S, for which the multivariate group Lasso is never worse than the ordinary Lasso in performing row selection. The following example shows that if the supports are disjoint, the ordinary Lasso has the same sample complexity as the multivariate group Lasso for uncorrelated design $\Sigma_{SS} = I_{s\times s}$, but can be better than the multivariate group Lasso for designs Σ_{SS} with suitable correlations:

COROLLARY 4 (Disjoint supports). *Suppose that the support sets S_k of individual regression problems are disjoint. Then for any design covariance Σ_{SS}, we have*

$$\max_{1\le k\le K} \psi(\beta_S^{(k)*}) \overset{\text{(a)}}{\le} \psi(B_S^*) \overset{\text{(b)}}{\le} \sum_{k=1}^{K} \psi(\beta_S^{(k)*}). \tag{25}$$

PROOF. First note that, since all supports are disjoint, $Z_i^{(k)*} = \operatorname{sign}(\beta_{ik}^*)$, so that $Z_S^{(k)*} = \zeta(\beta_S^{(k)*})$. Inequality (b) is then immediate because we have $|||Z_S^{*T}\Sigma_{SS}^{-1}Z_S^*|||_2 \le \operatorname{tr}(Z_S^{*T}\Sigma_{SS}^{-1}Z_S^*)$. To establish inequality (a), we note that

$$\begin{aligned}\psi(B^*) &= \max_{x\in\mathbb{R}^K:\|x\|\le 1} x^T Z_S^{*T}\Sigma_{SS}^{-1}Z_S^* x \ge \max_{1\le k\le K} e_k^T Z_S^{*T}\Sigma_{SS}^{-1}Z_S^* e_k \\ &\ge \max_{1\le k\le K} Z_S^{(k)*T}\Sigma_{SS}^{-1}Z_S^{(k)*}.\end{aligned}$$

□

A caveat in interpreting Corollary 4, and more generally in comparing the performance of the ordinary Lasso and the multivariate group Lasso, is that for a general covariance matrix Σ_{SS}, assumptions (A2) and (A3) required by Theorem 1 do not induce the same constraints on the covariance matrix Σ when applied to the multivariate problem as when applied to the individual regressions. Indeed, in the latter case, (A2) would require $\max_k |||\Sigma_{S_k^c S_k}\Sigma_{S_k S_k}^{-1}|||_\infty \le 1-\gamma$ and (A3) would require $\max_k |||\Sigma_{S_k S_k}^{-1}|||_\infty \le D_{\max}$. Thus (A3) is a stronger assumption in the multivariate case but (A2) is not.

We illustrate Corollary 4 with an example.

EXAMPLE 3 (Disjoint support with uncorrelated design). Suppose that $\Sigma_{SS} = I_{s\times s}$, and the supports are disjoint. In this case, we claim that the sample complexity of the ℓ_1/ℓ_2 multivariate group Lasso is the same as the ordinary Lasso.

If the individual regressions have disjoint support, then $Z^*_S = \zeta(B^*_S)$ has only a single nonzero entry per row and therefore the columns of Z^* are orthogonal. Moreover, $Z^*_{ik} = \operatorname{sign}(\beta_i^{(k)*})$. By Lemma 1(b), the sparsity-overlap function $\psi(B^*)$ is equal to the largest squared column norm. But $\|Z^{(k)*}\|^2 = \sum_{i=1}^s \operatorname{sign}(\beta_i^{(k)*})^2 = s_k$. Thus, the sample complexity of the multivariate group Lasso is the same as the ordinary Lasso in this case.[2]

Finally, we consider an example that illustrates the effect of correlated designs:

EXAMPLE 4 (Effects of correlated designs). To illustrate the behavior of the sparsity-overlap function in the presence of correlations in the design, we consider the simple case of two regressions with support of size 2. For parameters ϑ_1 and $\vartheta_2 \in [0, \pi]$ and $\mu \in (-1, +1)$, consider regression matrices B^* such that $B^* = \zeta(B^*_S)$ and

$$\zeta(B^*_S) = \begin{bmatrix} \cos(\vartheta_1) & \sin(\vartheta_1) \\ \cos(\vartheta_2) & \sin(\vartheta_2) \end{bmatrix} \quad \text{and} \quad \Sigma_{SS}^{-1} = \begin{bmatrix} 1 & \mu \\ \mu & 1 \end{bmatrix}. \tag{26}$$

Setting $M^* = \zeta(B^*_S)^T \Sigma_{SS}^{-1} \zeta(B^*_S)$, a simple calculation shows that

$$\operatorname{tr}(M^*) = 2\bigl(1 + \mu \cos(\vartheta_1 - \vartheta_2)\bigr) \quad \text{and} \quad \det(M^*) = (1 - \mu^2) \sin(\vartheta_1 - \vartheta_2)^2,$$

so that the eigenvalues of M^* are

$$\mu^+ = (1 + \mu)\bigl(1 + \cos(\vartheta_1 - \vartheta_2)\bigr) \quad \text{and} \quad \mu^- = (1 - \mu)\bigl(1 - \cos(\vartheta_1 - \vartheta_2)\bigr),$$

and $\psi(B^*) = \max(\mu^+, \mu^-)$. On the other hand, with

$$\tilde{z}_1 = \zeta(\beta^{(1)*}) = \begin{pmatrix} \operatorname{sign}(\cos(\vartheta_1)) \\ \operatorname{sign}(\cos(\vartheta_2)) \end{pmatrix} \quad \text{and} \quad \tilde{z}_2 = \zeta(\beta^{(2)*}) = \begin{pmatrix} \operatorname{sign}(\sin(\vartheta_1)) \\ \operatorname{sign}(\sin(\vartheta_2)) \end{pmatrix}$$

we have

$$\psi(\beta^{(1)*}) = \tilde{z}_1^T \Sigma_{SS}^{-1} \tilde{z}_1 = \mathbf{1}_{\{\cos(\vartheta_1) \neq 0\}} + \mathbf{1}_{\{\cos(\vartheta_2) \neq 0\}} + 2\mu \operatorname{sign}(\cos(\vartheta_1)\cos(\vartheta_2)),$$
$$\psi(\beta^{(2)*}) = \tilde{z}_2^T \Sigma_{SS}^{-1} \tilde{z}_2 = \mathbf{1}_{\{\sin(\vartheta_1) \neq 0\}} + \mathbf{1}_{\{\sin(\vartheta_2) \neq 0\}} + 2\mu \operatorname{sign}(\sin(\vartheta_1)\sin(\vartheta_2)).$$

Figure 1 provides a graphical comparison of these sample complexity functions. The function $\tilde{\psi}(B^*) = \max(\psi(\beta^{(1)*}), \psi(\beta^{(2)*}))$ is discontinuous on $\mathcal{S} = \frac{\pi}{2}\mathbb{Z} \times \mathbb{R} \cup \mathbb{R} \times \frac{\pi}{2}\mathbb{Z}$, and, as a consequence, so is its difference with $\psi(B^*)$. Note that, for fixed ϑ_1 or fixed ϑ_2, some of these discontinuities are *removable discontinuities* of the induced function on the other variable, and these discontinuities therefore

[2] In making this assertion, we are ignoring any difference between $\log(p - s_k)$ and $\log(p - s)$, which is valid, for instance, in the regime of sublinear sparsity, when $s_k/p \to 0$ and $s/p \to 0$.

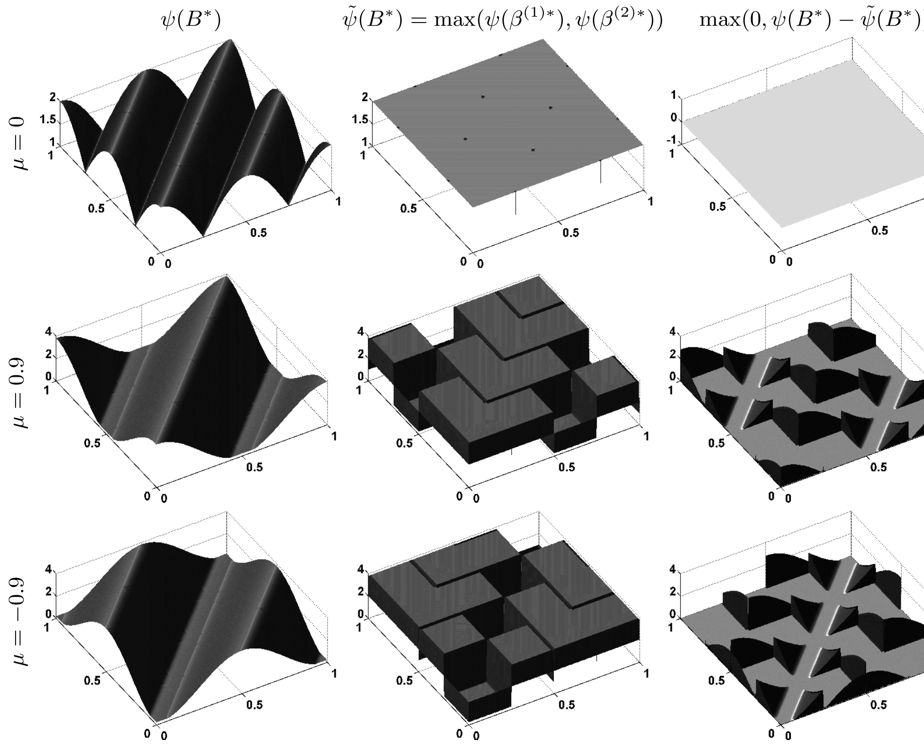

FIG. 1. *Comparison of sparsity-overlap functions for* ℓ_1/ℓ_2 *and the Lasso. For the pair* $\frac{1}{2\pi}(\vartheta_1, \vartheta_2)$, *we represent in each row of plots*, *corresponding respectively to* $\mu = 0$ (*top*), 0.9 (*middle*) *and* -0.9 (*bottom*), *from left to right*, *the quantities*: $\psi(B^*)$ (*left*), $\max(\psi(\beta^{(1)*}), \psi(\beta^{(2)*}))$ (*center*) *and* $\max(0, \psi(B^*) - \max(\psi(\beta^{(1)*}), \psi(\beta^{(2)*})))$ (*right*). *The latter indicates when the inequality* $\psi(B^*) \leq \max(\psi(\beta^{(1)*}), \psi(\beta^{(2)*}))$ *does not hold and by how much it is violated.*

create needles, slits or flaps in the graph of the function $\tilde{\psi}$. Denote by $\mathcal{R}^+$ and $\mathcal{R}^-$ the sets

$$\mathcal{R}^+ = \{(\vartheta_1, \vartheta_2) | \min[\cos(\vartheta_1)\cos(\vartheta_2), \sin(\vartheta_1)\sin(\vartheta_2)] > 0\},$$
$$\mathcal{R}^- = \{(\vartheta_1, \vartheta_2) | \max[\cos(\vartheta_1)\cos(\vartheta_2), \sin(\vartheta_1)\sin(\vartheta_2)] < 0\}$$

on which $\tilde{\psi}(B^*)$ reaches its minimum value when $\mu \geq 0.5$ and $\mu \leq 0.5$, respectively (see middle and bottom center plots in Figure 4). For $\mu = 0$, the top center graph illustrates that $\tilde{\psi}(B^*)$ is equal to 2 except for the cases of matrices B^*_S with disjoint support, corresponding to the discrete set $\mathcal{D} = \{(k\frac{\pi}{2}, (k \pm 1)\frac{\pi}{2}), k \in \mathbb{Z}\}$ for which it equals 1. The top rightmost graph illustrates that, as shown in Corollary 3, the inequality always holds for an uncorrelated design. For $\mu > 0$, the inequality $\psi(B^*) \leq \max(\psi(\beta^{(1)*}), \psi(\beta^{(2)*}))$ is violated only on a subset of $\mathcal{S} \cup \mathcal{R}^-$; and for $\mu < 0$, the inequality is symmetrically violated on a subset of $\mathcal{S} \cup \mathcal{R}^+$ (see Figure 4).

2.4. *Illustrative simulations.* In this section, we present the results of simulations that illustrate the sharpness of Theorems 1 and 2, and furthermore demonstrate how quickly the predicted behavior is observed as elements of the triple (n, p, s) grow in different regimes. We explore the case of two regressions (i.e., $K = 2$) which share an identical support set S with cardinality $|S| = s$ in Section 2.4.1 and consider a slightly more general case in Section 2.4.3.

2.4.1. *Threshold effect in the standard Gaussian case.* The first set of experiments was designed to reveal the threshold effect predicted by Theorems 1 and 2. The design matrix X is sampled from the standard Gaussian ensemble, with i.i.d. entries $X_{ij} \sim N(0, 1)$. We consider two types of sparsity:

- logarithmic sparsity, where $s = \alpha \log(p)$, for $\alpha = 2/\log(2)$, and
- linear sparsity, where $s = \alpha p$, for $\alpha = 1/8$

for various ambient model dimensions $p \in \{16, 32, 64, 256, 512, 1024\}$. For a given triplet (n, p, s), we solve the block-regularized problem (12) with the regularization parameter $\lambda_n = \sqrt{\log(p - s)(\log s)/n}$. For each fixed (p, s) pair, we measure the sample complexity in terms of a parameter θ, in particular letting $n = \theta s \log(p - s)$ for $\theta \in [0.25, 1.5]$. We let the matrix $B^* \in \mathbb{R}^{p\times 2}$ of regression coefficients have entries β^*_{ij} in $\{-1/\sqrt{2}, 1/\sqrt{2}\}$, choosing the parameters to vary the angle between the two columns, thereby obtaining various desired values of $\psi(B^*)$. Since $\Sigma = I_{p\times p}$ for the standard Gaussian ensemble, the sparsity-overlap function $\psi(B^*)$ is simply the maximal eigenvalue of the Gram matrix $\zeta(B^*_S)^T \zeta(B^*_S)$. Since $|\beta^*_{ij}| = 1/\sqrt{2}$ by construction, we are guaranteed that $B^*_S = \zeta(B^*_S)$, that the minimum value $b^*_{\min} = 1$, and, moreover, that the columns of $\zeta(B^*_S)$ have the same Euclidean norm.

To construct parameter matrices B^* that satisfy $|\beta^*_{ij}| = 1/\sqrt{2}$, we choose both p and the sparsity scalings so that the obtained values for s are multiples of four. We then construct the columns $Z^{(1)*}$ and $Z^{(2)*}$ of the matrix $B^* = \zeta(B^*)$ from copies of vectors of length four. Denoting by $\otimes$ the usual matrix tensor product, we consider the following 4-vectors:

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

Orthonormal regressions: Here B^* is constructed with $Z^{(1)*} \perp Z^{(2)*}$, so that $\psi(B^*) = \frac{s}{2}$, the most favorable situation. In order to achieve this orthonormality, 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 $\psi(B^*) = \frac{3}{4}s$. Specifically, 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 2 shows plots for linear sparsity (left column) and logarithmic sparsity (right column) in all three cases and where the multivariate group Lasso was used

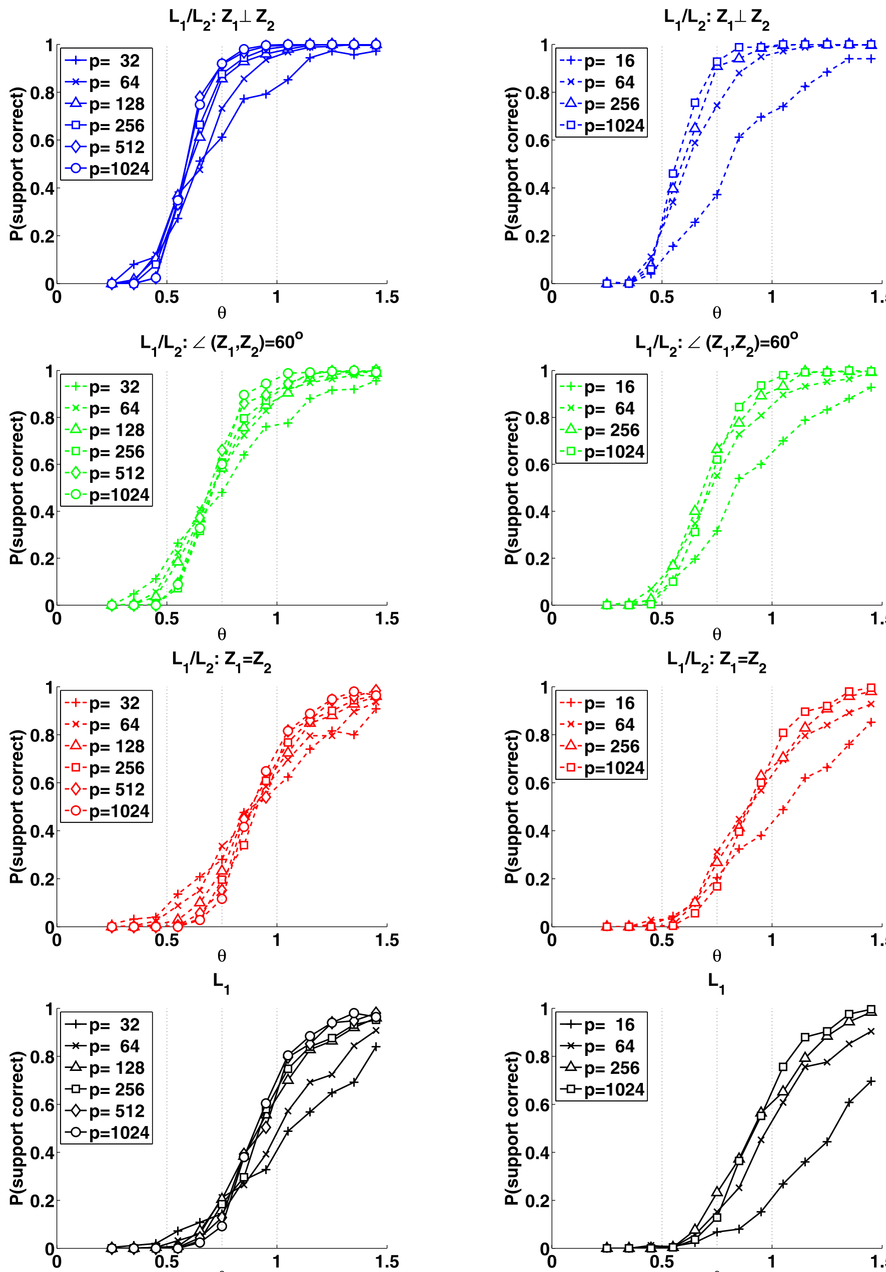

FIG. 2. *Plots of support union recovery probability,* $\mathbb{P}[\widehat{S} = S]$, *versus the control parameter* $\theta = n/[2s\log(p-s)]$ *for two different types of sparsity: linear sparsity in the left column* ($s = p/8$) *and logarithmic sparsity in the right column* ($s = 2\log_2(p)$). *The first three rows are based on using the multivariate group Lasso to estimate the support for the three cases of identical regression, intermediate angles and orthonormal regressions. The fourth row presents results for the Lasso in the case of identical regressions.*

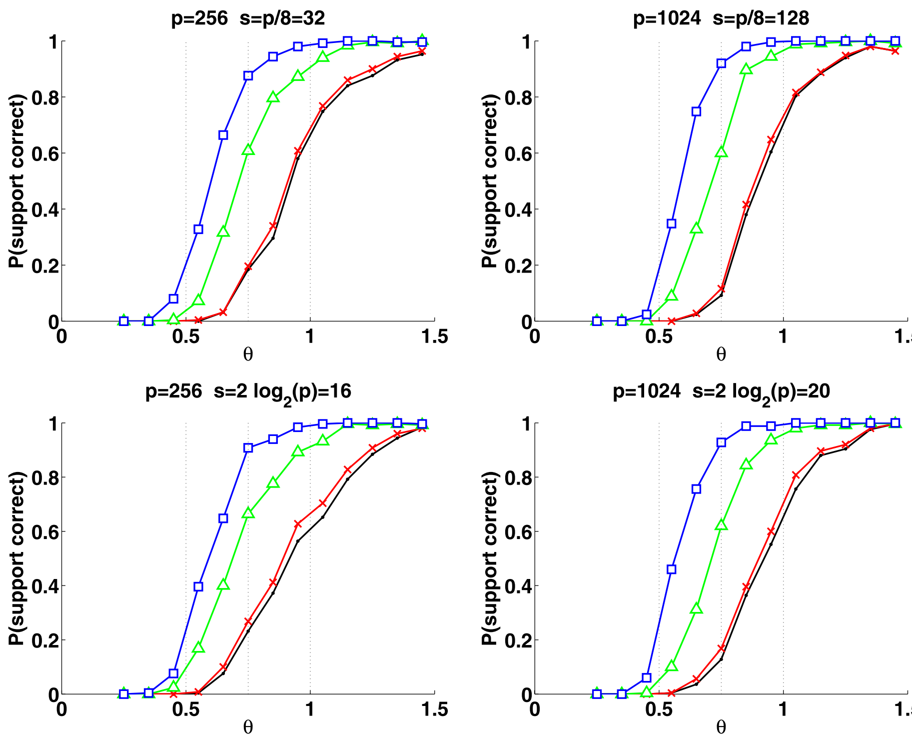

FIG. 3. *Plots of support recovery probability,* $\mathbb{P}[\widehat{S}=S]$, *versus the control parameter* $\theta = n/[2s\log(p-s)]$ *for two different type of sparsity: logarithmic sparsity on top* ($s = \mathcal{O}(\log(p))$) *and linear sparsity on bottom* ($s = \alpha p$), *and for increasing values of* p *from left to right. The noise level is set at* $\sigma = 0.1$. *Each graph shows four curves* (*black, red, green, blue*) *corresponding to the case of independent* ℓ_1 *regularization, and, for* ℓ_1/ℓ_2 *regularization, the cases of identical regression, intermediate angles and orthonormal regressions. Note how curves corresponding to the same case across different problem sizes* p *all coincide, as predicted by Theorems 1 and 2. Moreover, consistent with the theory, the curves for the identical regression group reach* $\mathbb{P}[\widehat{S}=S] \approx 0.50$ *at* $\theta \approx 1$, *whereas the orthonormal regression group reaches* 50% *success substantially earlier.*

(top three rows), as well as the reference Lasso case for the case of identical regressions (bottom row). Each panel plots the success probability, $\mathbb{P}[\widehat{S}=S]$, versus the rescaled sample size $\theta = n/[2s\log(p-s)]$. Under this rescaling, Theorems 1 and 2 predict that the curves should align, and that the success probability should transition to 1 once θ exceeds a critical threshold (dependent on the type of ensemble). Note that for suitably large problem sizes ($p \geq 128$), the curves do align in the predicted way, showing step-function behavior. Figure 3 plots data from the same simulations in a different format. Here the top row corresponds to logarithmic sparsity, and the bottom row to linear sparsity; each panel shows the four different choices for B^*, with the problem size p increasing from left to right. Note how in each panel the location of the transition of $\mathbb{P}[\widehat{S}=S]$ to one shifts from right to left, as we move from the case of identical regressions to intermediate angles to orthogonal regressions.

2.4.2. *Threshold effect with Toeplitz covariance matrices.* The simulations in the previous section involved the standard Gaussian design matrix; in this section, we explore the behavior for design matrices with some dependence structure. In particular, we report results for random designs with rows drawn from a Gaussian with Toeplitz covariance matrix of the form $\Sigma = (\rho^{|i-j|})_{1\le i,j\le p}$, for some parameter $\rho \in [0, 1)$. Zhao and Yu (2006) have shown that such Toeplitz matrices satisfy the irrepresentable conditions required for support consistency. As with our experiments in the standard Gaussian case, we consider the same two regimes (linear and logarithmic), using the same families of regression matrices B^* and the same noise level. We select the support of the regression matrices as a random subset of the p covariates of size s, and draw the design matrices from the Toeplitz ensemble $\rho = 0.5$. For each pair (s, p), we consider a number of observations of the form $n = 2\theta s \log(p)$ for $\theta \in [0.25, 4]$.

Figure 4 is the analog of the previously shown Figure 3: for problems with random designs from the Toeplitz ensemble, it plots the support recovery probability $\mathbb{P}[\widehat{S} = S]$ versus the control parameter $\theta = n/[2s\log(p-s)]$ for two different types of sparsity—logarithmic sparsity on top ($s = \mathcal{O}(\log(p))$) and linear sparsity on bottom ($s = \alpha p$). The four curves (black, red, green, blue) corresponding to the case of independent ℓ_1 regularization, and, for ℓ_1/ℓ_2 regularization, the cases of identical regression, intermediate angles and orthonormal regressions. Qualitatively, note that we observe the same type of transitions as in the standard Gaussian case; moreover, the curves shift from right to left as the angles between the regression columns vary from orthogonal to identical.

2.4.3. *Empirical threshold values.* In this experiment, we aim at verifying more precisely the location of the ℓ_1/ℓ_2 threshold as the regression vectors vary continuously from identical to orthogonal with equal length. We consider the case of matrices B^* of size $s \times 2$ for s even. In Example 4 of Section 2.3, we characterized the value of $\psi(B^*)$ when B^* is a 2×2 matrix.

In order to generate a family of regression matrices with smoothly varying sparsity-overlap function consider the following 2×2 matrix:

$$B_1(\alpha) = \begin{bmatrix} \dfrac{1}{\sqrt{2}} & \dfrac{1}{\sqrt{2}} \\ \cos\left(\dfrac{\pi}{4} + \alpha\right) & \sin\left(\dfrac{\pi}{4} + \alpha\right) \end{bmatrix}. \tag{27}$$

Note that α is the angle between the two *rows* of $B_1(\alpha)$ in this setup. Note, moreover, that the columns of $B_1(\alpha)$ have varying norm.

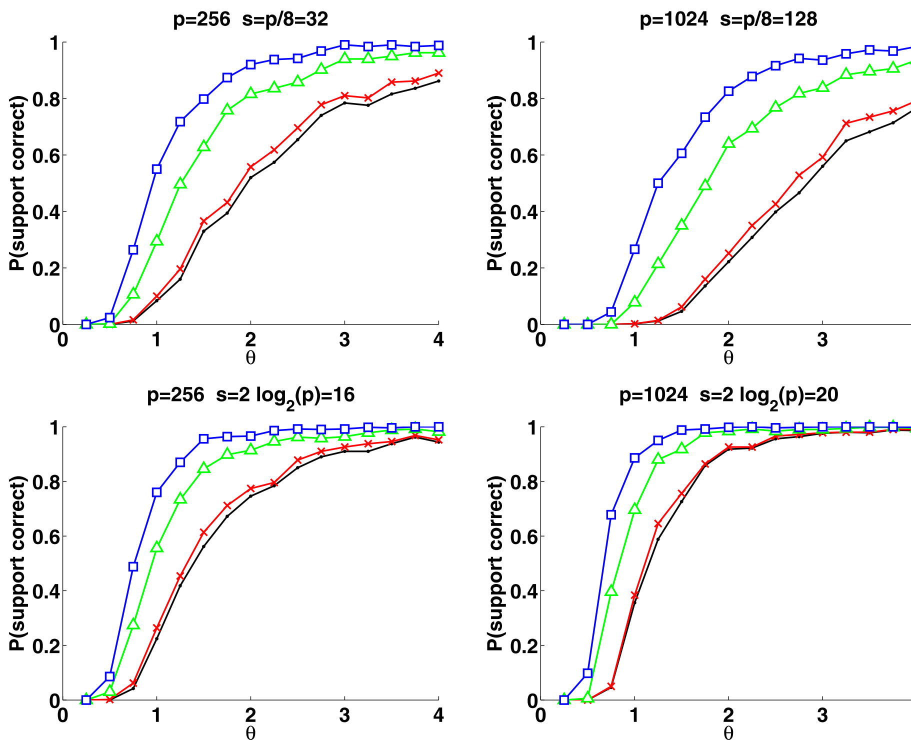

FIG. 4. *Plots of support recovery probability,* $\mathbb{P}[\widehat{S} = S]$, *versus the control parameter* $\theta = n/[2s\log(p-s)]$ *when the covariance matrix is a Toeplitz matrix with parameter* $\rho = 0.5$, *for the same protocol as described in Figure 3.*

We use this base matrix to define the following family of regression matrices $B_S^* \in \mathbb{R}^{s\times 2}$:

$$\mathcal{B}_1 := \left\{ B_{1s}(\alpha) = \vec{1}_{s/2} \otimes B_1(\alpha), \alpha \in \left[0, \frac{\pi}{2}\right]\right\}. \tag{28}$$

For a design matrix drawn from the standard Gaussian ensemble, the analysis of Example 4 in Section 2.3 extends naturally to show that the sparsity-overlap function is $\psi(B_{s1}(\alpha)) = \frac{s}{2}(1 + |\cos(\alpha)|)$. Moreover, as we vary α from 0 to $\frac{\pi}{2}$, the two regressions vary from identical to orthonormal and the sparsity-overlap function decreases from s to $\frac{s}{2}$.

We fix the problem size $p = 2048$ and sparsity $s = \log_2(p) = 22$. For each value of $\alpha \in [0, \frac{\pi}{2}]$, we generate a matrix from the specified family and angle. We then solve the multivariate group Lasso optimization problem (12) with sample size $n = 2\theta s\log(p-s)$ for a range of values of θ in $[0.25, 1.5]$; for each value of θ, we repeat the experiment (generating random design matrix X and observation matrix Y each time) over $T = 500$ trials. Based on these trials, we then estimate the value of $\theta_{50\%}$ for which the exact support is retrieved at least 50% of the time.

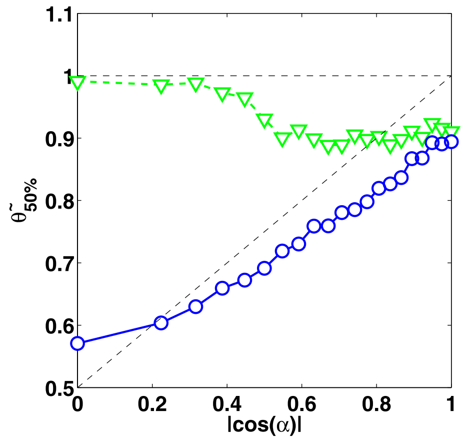

FIG. 5. *Plots of the Lasso sample complexity $\theta = n/[2s\log(p-s)]$ for which the probability of support union recovery exceeds 50% empirically as a function of $|\cos(\alpha)|$ for ℓ_1-based recovery and ℓ_1/ℓ_2-based recovery, where α is the angle between $Z^{(1)*}$ and $Z^{(2)*}$ for the family $\mathcal{B}_1$. We consider the two following methods for performing row selection: Ordinary Lasso (ℓ_1, green triangles) and multivariate group Lasso (blue circles).*

Since $\psi(B^*) = \frac{1+|\cos(\alpha)|}{2}s$, our theory predicts that if we plot $\theta_{50\%}$ versus $|\cos(\alpha)|$, the plot should lie on or below the straight line $\frac{1+|\cos(\alpha)|}{2}$. We also perform the same experiments for row selection using the ordinary Lasso and plot the resulting estimated thresholds on the same axes.

The results are shown in Figure 5. Note first that the curve obtained for $\widehat{S}_{\ell_1/\ell_2}$ (blue circles) coincides roughly with the theoretical prediction, $\frac{1+|\cos(\alpha)|}{2}$ (black dashed diagonal) as regressions vary from orthogonal to identical. Moreover, the estimated $\theta_{50\%}$ of the ordinary Lasso remains above 0.9 for all values of α, close to the theoretical value of 1. However, the curve obtained is not constant, but is roughly sigmoidal with a first plateau close to 1 for $\cos(\alpha) < 0.4$ and a second plateau close to 0.9 for $\cos(\alpha) > 0.5$. The latter coincides with the empirical value of $\theta_{50\%}$ for the univariate Lasso for the first column $\beta^{(1)*}$ (not shown). There are two reasons why the value of $\theta_{50\%}$ for the ordinary Lasso does not match the prediction of the first-order asymptotics: first, for $\alpha = \frac{\pi}{4}$ [corresponding to $\cos(\alpha) = 0.7$], the support of $\beta^{(2)*}$ is reduced by one half and therefore its sample complexity is decreased in that region. Second, the supports recovered by individual Lassos for $\beta^{(1)*}$ and $\beta^{(2)*}$ vary from uncorrelated when $\alpha = \frac{\pi}{2}$ to identical when $\alpha = 0$. It is therefore not surprising that the sample complexity is the same as a single univariate Lasso for $\cos(\alpha)$ large and higher for $\cos(\alpha)$ small, where independent estimates of the support are more likely to include, by union, spurious covariates in the row support.

3. Proof of Theorem 1. In this section, we provide the proof of Theorem 1, which gives sufficient conditions for success of the multivariate group Lasso. Subsequently, in Section 4, we provide the proof for the necessary conditions as given in Theorem 2. For the convenience of the reader, we begin by recapitulating the notation to be used throughout both of these arguments:

- The sets S and S^c are a partition of the set of columns of X, such that $|S| = s$, $|S^c| = p - s$.
- The design matrix is partitioned as $X = [X_S X_{S^c}]$, where $X_S \in \mathbb{R}^{n\times s}$ and $X_{S^c} \in \mathbb{R}^{n\times(p-s)}$.
- The regression coefficient matrix is also partitioned as $B^* = \left[\begin{smallmatrix} B^*_S \\ B^*_{S^c}\end{smallmatrix}\right]$, with $B^*_S \in \mathbb{R}^{s\times K}$ and $B^*_{S^c} = 0 \in \mathbb{R}^{(p-s)\times K}$. We use β^*_i to denote the ith row of B^*.
- The regression model is given by $Y = XB^* + W$, where the noise matrix $W \in \mathbb{R}^{n\times K}$ has i.i.d. $N(0, \sigma^2)$ entries.
- The matrix $Z^*_S = \zeta(B^*_S) \in \mathbb{R}^{s\times K}$ has rows $Z^*_i = \zeta(\beta^*_i) = \frac{\beta^*_i}{\|\beta^*_i\|_2} \in \mathbb{R}^K$.

3.1. *High-level proof outline.* At a high level, the proof is based on the notion of a *primal–dual witness*: we construct a primal matrix $\widehat{B}$ along with a dual matrix $\widehat{Z}$ such that:

(a) the pair $(\widehat{B}, \widehat{Z})$ together satisfy the Karush–Kuhn–Tucker (KKT) conditions associated with the second-order cone program (12), and
(b) this solution certifies that the multivariate group Lasso recovers the union of supports S.

For general high-dimensional problems (with $p \gg n$), the multivariate group Lasso of (12) need not have a unique solution; however, a consequence of our theory is that the constructed solution $\widehat{B}$ is the unique optimal solution under the conditions of Theorem 1.

We begin by noting that the block-regularized problem (12) is convex, and not differentiable for all B. In particular, denoting by β_i the ith row of B, the subdifferential of the ℓ_1/ℓ_2-block norm over row i takes the form

$$[\partial\|B\|_{\ell_1/\ell_2}]_i = \begin{cases} \dfrac{\beta_i}{\|\beta_i\|_2}, & \text{if } \beta_i \neq \vec{0}, \\ Z_i \text{ such that } \|Z_i\|_2 \le 1, & \text{otherwise.} \end{cases} \tag{29}$$

We define the *empirical covariance matrix*

$$\widehat{\Sigma} := \frac{1}{n} X^T X = \frac{1}{n}\sum_{i=1}^n X_i X_i^T, \tag{30}$$

where X_i is the ith row of X. (This definition is natural under our standing assumption of zero mean for the variables X_i; note, however, that our proofs extend readily to the case of nonzero mean, in which case we would center the variables and use the usual definition of the empirical covariance matrix.) We also make use of the shorthand $\widehat{\Sigma}_{SS} = \frac{1}{n}X_S^T X_S$ and $\widehat{\Sigma}_{S^c S} = \frac{1}{n}X_{S^c}^T X_S$ as well as $\Pi_S = X_S(\widehat{\Sigma}_{SS})^{-1}X_S^T$ to denote the projector on the range of X_S.

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 2. *Suppose that there exists a primal–dual pair* $(\widehat{B}, \widehat{Z})$ *that satisfies the conditions*

$$\widehat{Z}_S = \zeta(\widehat{B}_S), \tag{31a}$$

$$-\lambda_n \widehat{Z}_S = \widehat{\Sigma}_{SS}(\widehat{B}_S - B_S^*) - \frac{1}{n}X_S^T W, \tag{31b}$$

$$\lambda_n \|\widehat{Z}_{S^c}\|_{\ell_\infty/\ell_2} := \left\| \widehat{\Sigma}_{S^c S}(\widehat{B}_S - B_S^*) - \frac{1}{n}X_{S^c}^T W \right\|_{\ell_\infty/\ell_2} < \lambda_n, \tag{31c}$$

$$\widehat{B}_{S^c} = 0. \tag{31d}$$

Then $(\widehat{B}, \widehat{Z})$ *is a primal–dual optimal solution to the block-regularized problem, with* $\widehat{S}(\widehat{B}) = S$ *by construction. If* $\widehat{\Sigma}_{SS} \succ 0$, *then* $\widehat{B}$ *is the unique optimal primal solution.*

See Appendix B for the proof of this claim. Based on Lemma 2, we proceed to construct the required primal–dual pair $(\widehat{B}, \widehat{Z})$ as follows. First, we set $\widehat{B}_{S^c} = 0$, so that condition (31d) is satisfied. Next, we specify the pair $(\widehat{B}_S, \widehat{Z}_S)$ by solving the following restricted version of the SOCP (12):

$$\widehat{B}_S = \arg \min_{B_S \in \mathbb{R}^{s \times K}} \left\{ \frac{1}{2n} \left\| Y - X \begin{bmatrix} B_S \\ 0_{S^c} \end{bmatrix} \right\|_F^2 + \lambda_n \|B_S\|_{\ell_1/\ell_2} \right\}. \tag{32}$$

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 (32) is strictly convex and therefore has a unique optimum $\widehat{B}_S$. We then choose $\widehat{Z}_S$ to be the solution of equation (31b). Since any such matrix $\widehat{Z}_S$ is also a dual solution to the restricted SOCP (32), 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 (31a) and (31c). In order to satisfy condition (31a), it suffices to show that no row of the solution $\widehat{B}_S$ is identically zero. From equation (31b) and using the invertibility of the empirical covariance matrix $\widehat{\Sigma}_{SS}$, we may solve as follows:

$$(\widehat{B}_S - B_S^*) = (\widehat{\Sigma}_{SS})^{-1} \left[\frac{X_S^T W}{n} - \lambda_n \widehat{Z}_S \right] =: U_S. \tag{33}$$

Note that for any row $i \in S$, by the triangle inequality, we have

$$\|\widehat{\beta}_i\|_2 \geq \|\beta_i^*\|_2 - \|U_S\|_{\ell_\infty/\ell_2}.$$

Therefore, in order to show that no row of $\widehat{B}_S$ is identically zero, it suffices to show that the event

$$\mathcal{E}(U_S) := \left\{\|U_S\|_{\ell_\infty/\ell_2} \leq \tfrac{1}{2} b^*_{\min}\right\} \tag{34}$$

occurs with high probability [recall from equation (14) that the parameter $b^*_{\min}$ measures the minimum ℓ_2-norm of any row of B^*_S]. We establish this result in Section 3.3.

Turning to condition (31c), by substituting expression (33) for the difference $(\widehat{B}_S - B^*_S)$ into equation (31c), we obtain a $(p-s) \times K$ random matrix V_{S^c}, with rows indexed by S^c. For any index $j \in S^c$, the corresponding row vector $V_j \in \mathbb{R}^K$ is given by

$$V_j := X_j^T\left([\Pi_S - I_n]\frac{W}{n} - \lambda_n \frac{X_S}{n}(\widehat{\Sigma}_{SS})^{-1}\widehat{Z}_S\right). \tag{35}$$

In order for condition (31c) to hold, it is necessary and sufficient that the probability of the event

$$\mathcal{E}(V_{S^c}) := \{\|V_{S^c}\|_{\ell_\infty/\ell_2} < \lambda_n\} \tag{36}$$

converges to one as n tends to infinity. Consequently, the remainder (and bulk) of the proof is devoted to showing that the probabilities $\mathbb{P}[\mathcal{E}(U_S)]$ and $\mathbb{P}[\mathcal{E}(V_{S^c})]$ both converge to one under the specified conditions.

3.2. *Analysis of $\mathcal{E}(V_{S^c})$: Correct exclusion of nonsupport.* In this section, we prove the first claim of Theorem 1(a), namely that rows not in the support are always excluded. 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 $\frac{1}{\lambda_n}\|V\|_{\ell_\infty/\ell_2} \leq \sum_{i=1}^3 T'_i$ where

$$T'_1 := \frac{1}{\lambda_n}\|\mathbb{E}[V \mid X_S]\|_{\ell_\infty/\ell_2}, \tag{37a}$$

$$T'_2 := \frac{1}{\lambda_n}\|\mathbb{E}[V \mid X_S, W] - \mathbb{E}[V \mid X_S]\|_{\ell_\infty/\ell_2}, \tag{37b}$$

$$T'_3 := \frac{1}{\lambda_n}\|V - \mathbb{E}[V \mid X_S, W]\|_{\ell_\infty/\ell_2}. \tag{37c}$$

We deal with each of these three terms in turn, showing that with high probability under the specified scaling of (n, p, s), we have $T'_1 \leq (1-\gamma)$, and $T'_2 = o_p(1)$, and $T'_3 < \gamma$, which suffices to show that $\frac{1}{\lambda_n}\|V\|_{\ell_\infty/\ell_2} < 1$ with high probability.

The following lemma is useful in the analysis:

LEMMA 3. *Define the matrix $\Delta \in \mathbb{R}^{s \times K}$ with rows $\Delta_i := U_i/\|\beta^*_i\|_2$. As long as $\|\Delta_i\|_2 \leq 1/2$ for all row indices $i \in S$, we have*

$$\|\widehat{Z}_S - \zeta(B^*_S)\|_{\ell_\infty/\ell_2} \leq 4\|\Delta\|_{\ell_\infty/\ell_2}.$$

See Appendix G for the proof of this claim.

3.2.1. *Analysis of T_1'.* Note that by definition of the regression model (4), we have the conditional independence relations

$$W \perp\!\!\!\perp X_{S^c}|X_S, \qquad \widehat{Z}_S \perp\!\!\!\perp X_{S^c}|X_S \quad \text{and} \quad \widehat{Z}_S \perp\!\!\!\perp X_{S^c}|\{X_S, W\}.$$

Using the two first conditional independencies, we have

$$\mathbb{E}[V|X_S] = \mathbb{E}[X_{S^c}^T|X_S]\left([\Pi_S - I_n]\frac{\mathbb{E}[W|X_S]}{n} - \lambda_n \frac{X_S}{n}(\widehat{\Sigma}_{SS})^{-1}\mathbb{E}[\widehat{Z}_S|X_S]\right).$$

Since $\mathbb{E}[W|X_S] = 0$, the first term vanishes, and using $\mathbb{E}[X_{S^c}^T|X_S] = \Sigma_{S^cS}\Sigma_{SS}^{-1}X_S^T$, we obtain

$$\mathbb{E}[V|X_S] = \lambda_n \Sigma_{S^cS}\Sigma_{SS}^{-1}\mathbb{E}[\widehat{Z}_S|X_S]. \tag{38}$$

Using the matrix-norm inequality (57a) from Appendix E and then Jensen's inequality yields

$$\begin{aligned} T_1' &= \|\Sigma_{S^cS}\Sigma_{SS}^{-1}\mathbb{E}[Z_S|X_S]\|_{\ell_\infty/\ell_2} \\ &\le |||\Sigma_{S^cS}\Sigma_{SS}^{-1}|||_\infty \mathbb{E}[\|Z_S\|_{\ell_\infty/\ell_2}|X_S] \\ &\le (1-\gamma). \end{aligned} \tag{39}$$

3.2.2. *Analysis of T_2'.* Appealing to the conditional independence relationship $\widehat{Z}_S \perp\!\!\!\perp X_{S^c}|\{X_S, W\}$, we have

$$\begin{aligned} &\mathbb{E}[V|X_S, W] \\ &\quad = \mathbb{E}[X_{S^c}^T|X_S, W]\left([\Pi_S - I_n]\frac{W}{n} - \lambda_n\frac{X_S}{n}(\widehat{\Sigma}_{SS})^{-1}\mathbb{E}[\widehat{Z}_S|X_S, W]\right). \end{aligned}$$

Observe that $\mathbb{E}[\widehat{Z}_S|X_S, W] = \widehat{Z}_S$ because (X_S, W) uniquely specifies $\widehat{B}_S$ through the convex program (32), and the triple $(X_S, W, \widehat{B}_S)$ defines $\widehat{Z}_S$ through equation (31b). Moreover, the noise term disappears because the kernel of the orthogonal projection matrix $(I_n - \Pi_S)$ is the same as the range space of X_S, and

$$\begin{aligned} \mathbb{E}[X_{S^c}^T|X_S, W][\Pi_S - I_n] &= \mathbb{E}[X_{S^c}^T|X_S][\Pi_S - I_n] \\ &= \Sigma_{S^cS}\Sigma_{SS}^{-1}X_S^T[\Pi_S - I_n] = 0. \end{aligned}$$

We have thus shown that $\mathbb{E}[V|X_S, W] = -\frac{\lambda_n}{n}\Sigma_{S^cS}\Sigma_{SS}^{-1}\widehat{Z}_S$, so that we can conclude that

$$\begin{aligned} T_2' &\le |||\Sigma_{S^cS}(\Sigma_{SS})^{-1}|||_\infty \|\widehat{Z}_S - \mathbb{E}[\widehat{Z}_S|X_S]\|_{\ell_\infty/\ell_2} \\ &\le (1-\gamma)\mathbb{E}[\|\widehat{Z}_S - Z_S^*\|_{\ell_\infty/\ell_2}] + (1-\gamma)\|\widehat{Z}_S - Z_S^*\|_{\ell_\infty/\ell_2} \\ &\le (1-\gamma)4\{\mathbb{E}[\|\Delta\|_{\ell_\infty/\ell_2}] + \|\Delta\|_{\ell_\infty/\ell_2}\}, \end{aligned} \tag{40}$$

where the final inequality uses Lemma 3. Under the assumptions of Theorem 1, this final term is of order $o_p(1)$, as will be shown in Section 3.3.

3.2.3. *Analysis of* T_3'. This third term requires a little more care. We begin by noting that conditionally on X_S and W, each vector $V_j \in \mathbb{R}^K$ is normally distributed. Since $\mathrm{Cov}(X^{(j)}|X_S, W) = (\Sigma_{S^c|S})_{jj} I_n$, we have

$$\mathrm{Cov}(V_j|X_S, W) = M_n (\Sigma_{S^c|S})_{jj},$$

where the $K \times K$ random matrix $M_n = M_n(X_S, W)$ is given by

$$M_n := \frac{\lambda_n^2}{n} \widehat{Z}_S^T (\widehat{\Sigma}_{SS})^{-1} \widehat{Z}_S + \frac{1}{n^2} W^T (\Pi_S - I_n) W. \tag{41}$$

We begin by noting that by its definition (31a), the candidate dual matrix $\widehat{Z}_S$ is a function only of W and X_S. Therefore, conditioned on the pair (W, X_S), the matrix M_n is fixed, and we have

$$(\|V_j - \mathbb{E}[V_j|X_S, W]\|_2^2 | W, X_S) \stackrel{d}{=} (\Sigma_{S^c S^c|S})_{jj} \xi_j^T M_n \xi_j, \tag{42}$$

where $\xi_j \sim N(\vec{0}_K, I_K)$. By definition of $\rho_u(\Sigma_{S^c S^c|S}) = \max_j (\Sigma_{S^c S^c|S})_{jj}$, we have $(\Sigma_{S^c S^c|S})_{jj} \le \rho_u(\Sigma_{S^c S^c|S}) \le C_{\max}$ and

$$\max_{j \in S^c} (\Sigma_{S^c S^c|S})_{jj} \xi_j^T M_n \xi_j \le \rho_u(\Sigma_{S^c S^c|S}) |\!|\!| M_n |\!|\!|_2 \max_{j \in S^c} \|\xi_j\|_2^2,$$

where $|\!|\!| M_n |\!|\!|_2$ is the spectral norm.

We now state a result that provides control on this spectral norm, in particular showing that the rescaled random matrix $\frac{n}{\lambda_n^2} M_n$ concentrates around the deterministic matrix $M^* := Z_S^{*T} (\Sigma_{SS})^{-1} Z_S^*$. This concentration establishes the link to the sparsity-overlap function (16), which is given by the spectral norm $|\!|\!| M^* |\!|\!|_2$. For any $\delta \in (0, 1)$, define the event

$$\mathcal{T}(\delta) := \left\{ \frac{\lambda_n^2 \psi(B^*) + \sigma^2}{n} (1 - \delta) \le |\!|\!| M_n |\!|\!|_2 \le \frac{\lambda_n^2 \psi(B^*) + \sigma^2}{n} (1 + \delta) \right\}. \tag{43}$$

Moreover, recall the definition of Δ from Lemma 3. The following result provides sufficient conditions for the event $\mathcal{T}(\delta)$ to hold with high probability.

LEMMA 4. *Suppose that* $\frac{s}{n} = o(1)$ *and* $\|\Delta\|_{\ell_\infty/\ell_2} = o(1)$. *Then for any* $\delta \in (0, 1)$, *there is some* $c_1 = c_1(\delta) > 0$ *such that* $\mathbb{P}[\mathcal{T}(\delta)^c] \le c_1 \exp(-c_0 K \log s) \to 0$.

See Appendix H for the proof of this lemma.

Given the assumptions of Theorem 1 and the bound (46), we observe that the hypotheses of Lemma 4 are satisfied, and we can now complete the proof. For any fixed but arbitrarily small $\delta > 0$, we have

$$\mathbb{P}[T_3' \geq \gamma] \leq \mathbb{P}[T_3' \geq \gamma | \mathcal{T}(\delta)] + \mathbb{P}[\mathcal{T}(\delta)^c].$$

Since $\mathbb{P}[\mathcal{T}(\delta)^c] \to 0$ from Lemma 4, it suffices to deal with the first term. Conditioning on the event $\mathcal{T}(\delta)$, we have

$$\mathbb{P}[T_3' \geq \gamma | \mathcal{T}(\delta)] \leq \mathbb{P}\biggl[\max_{j \in S^c} \|\xi_j\|_2^2 \geq \frac{\gamma^2}{\rho_u(\Sigma_{S^c S^c|S})} \frac{n}{(\psi(B^*) + \sigma^2/\lambda_n^2)(1+\delta)}\biggr].$$

Now define the quantity

$$t^*(n, B^*) := \frac{1}{2} \frac{\gamma^2}{\rho_u(\Sigma_{S^c S^c|S})} \frac{n}{(\psi(B^*) + \sigma^2/\lambda_n^2)(1+\delta)},$$

and note that $t^* \to +\infty$ under the specified scaling of (n, p, s). By applying Lemma 11 from Appendix I on large deviations for χ^2-variates with $t = t^*(n, B^*)$, we obtain

$$\begin{aligned} \mathbb{P}[T_3' \geq \gamma | \mathcal{T}(\delta)] &\leq (p-s)\exp\biggl(-t^*\biggl[1 - 2\sqrt{\frac{K}{t^*}}\biggr]\biggr) \\ &\leq (p-s)\exp\bigl(-t^*(1-\delta)\bigr) \end{aligned} \tag{44}$$

for (n, p, s) sufficiently large. Now denoting $\theta_u := \rho_u(\Sigma_{S^c S^c|S})/\gamma^2$, we have, by assumption, that $n \geq 2(1+\nu)\theta_u \psi(B^*) \log(p-s)$. Given that $\lambda_n^2 = \frac{f(p)\log(p)}{n}$, we have $\frac{\sigma^2}{\lambda_n^2}\log(p-s) \leq \sigma^2 \frac{n}{f(p)} = o(n)$ so that for any $\varepsilon > 0$, we have

$$n \geq \frac{1+\nu}{1+\varepsilon}\biggl(2\theta_u \psi(B^*) \log(p-s) + \frac{2\sigma^2}{\lambda_n^2}\log(p-s)\biggr)$$

once n is sufficiently large. This inequality implies that $(1-\delta)t^*(n, B^*) \geq \frac{(1+\nu)(1-\delta)}{(1+\varepsilon)(1+\delta)}\log(p-s)$. Thus for δ and ε sufficiently small, the bound (44) tends to zero at rate $\mathcal{O}(\exp(-\nu/2\log(p-s)))$ which establishes the claim.

3.3. *Analysis of $\mathcal{E}(U_S)$: Correct inclusion of supporting covariates.* This section is devoted to the analysis of the event $\mathcal{E}(U_S)$ from equation (34), and in particular showing that its probability converges to one under the specified scaling. This allows us to establish the ℓ_2/ℓ_∞ bound in Theorem 1(a), as well as the correct support recovery claim in part (b).

If we define the noise matrix $\widetilde{W} := \frac{1}{\sqrt{n}}(\widehat{\Sigma}_{SS})^{-1/2} X_S^T W$, then we have

$$U_S = \widehat{\Sigma}_{SS}^{-1/2} \frac{\widetilde{W}}{\sqrt{n}} - \lambda_n (\widehat{\Sigma}_{SS})^{-1} \widehat{Z}_S.$$

Using this representation and the triangle inequality, we obtain

$$\begin{aligned}\|U_S\|_{\ell_\infty/\ell_2} &\le \left\|(\widehat{\Sigma}_{SS})^{-1/2}\frac{\widetilde{W}}{\sqrt{n}}\right\|_{\ell_\infty/\ell_2} + \lambda_n\|(\widehat{\Sigma}_{SS})^{-1}\widehat{Z}_S\|_{\ell_\infty/\ell_2}\\ &\le \underbrace{\left\|(\widehat{\Sigma}_{SS})^{-1/2}\frac{\widetilde{W}}{\sqrt{n}}\right\|_{\ell_\infty/\ell_2}}_{T_1} + \underbrace{\lambda_n|||(\widehat{\Sigma}_{SS})^{-1}|||_\infty}_{T_2},\end{aligned}$$

where the form of T_2 in the second line uses a standard matrix norm bound [see equation (57a) in Appendix E], and the fact that $\|\widehat{Z}_S\|_{\ell_\infty/\ell_2} \le 1$.

Using the triangle inequality, we bound T_2 as follows:

$$\begin{aligned}T_2 &\le \lambda_n\{|||(\Sigma_{SS})^{-1}|||_\infty + |||(\widehat{\Sigma}_{SS})^{-1} - (\Sigma_{SS})^{-1}|||_\infty\}\\ &\le \lambda_n\{D_{\max} + \sqrt{s}|||(\widehat{\Sigma}_{SS})^{-1} - (\Sigma_{SS})^{-1}|||_2\}\\ &\le \lambda_n\{D_{\max} + \sqrt{s}|||(\Sigma_{SS})^{-1}|||_2|||(\widetilde{X}_S^T\widetilde{X}_S/n)^{-1} - I_s|||_2\}\\ &\le \lambda_n\left\{D_{\max} + \frac{\sqrt{s}}{C_{\min}}|||(\widetilde{X}_S^T\widetilde{X}_S/n)^{-1} - I_s|||_2\right\},\end{aligned}$$

which defines $\widetilde{X}_S$ as a random matrix with i.i.d. standard Gaussian entries. From concentration results in random matrix theory (see Appendix F), for $s/n \to 0$, we have $|||(\widetilde{X}_S^T\widetilde{X}_S/n)^{-1} - I_s|||_2 \le 6\sqrt{\frac{s}{n}}$ with probability $1 - 2\exp(-s/2) - \exp(-\Theta(n))$. Overall, we conclude that

$$T_2 \le \lambda_n\left\{D_{\max} + \frac{6}{C_{\min}}\sqrt{\frac{s^2}{n}}\right\}$$

with probability $1 - 2\exp(-s/2) - \exp(-\Theta(n))$.

Turning now to T_1, let us introduce the notation $\mathrm{vec}(A)$ to denote the vectorized version of a matrix A, obtained by stacking all of its rows into a single vector. Conditioning on X_S, we have $(\mathrm{vec}(\widetilde{W})|X_S) \sim N(\vec{0}_{s\times K}, I_s \otimes I_K)$. Combined with the definition of the block ℓ_∞/ℓ_2 norm, we obtain

$$T_1 = \max_{i\in S}\left\|e_i^T(\widehat{\Sigma}_{SS})^{-1/2}\frac{\widetilde{W}}{\sqrt{n}}\right\|_2 \le |||(\widehat{\Sigma}_{SS})^{-1}|||_2^{1/2}\left[\frac{1}{n}\max_{i\in S}\zeta_i^2\right]^{1/2},$$

where the variates $\{\zeta_i^2\}$ are an i.i.d. sequence of χ^2-variates with K degrees of freedom. Using the tail bound in Lemma 11 (see Appendix I) with $t = 2K\log s > K$, we have

$$\mathbb{P}\left[\frac{1}{n}\max_{i\in S}\zeta_i^2 \ge \frac{4K\log s}{n}\right] \le \exp\Big(-2K\log s(1 - 2(2\log s)^{-1/2})\Big) \to 0.$$

Define the event $\mathcal{T} := \{|||(\widehat{\Sigma}_{SS})^{-1}|||_2 \le \frac{2}{C_{\min}}\}$; the bound $\mathbb{P}[\mathcal{T}] \ge 1 - \exp(-\Theta(n))$ then follows from known concentration results in random matrix theory (see Appendix F). Thus, we obtain

$$\begin{aligned}\mathbb{P}\biggl[T_1 \ge \sqrt{\frac{8K\log s}{C_{\min}n}}\biggr] &\le \mathbb{P}\biggl[T_1 \ge \sqrt{\frac{8K\log s}{C_{\min}n}}\Big|\mathcal{T}\biggr] + \mathbb{P}[\mathcal{T}^c]\\ &\le \mathbb{P}\biggl[\frac{1}{n}\max_{i\in S}\zeta_i^2 \ge \frac{4K\log s}{n}\biggr] + \exp\biggl\{-n\biggl(\frac{1}{2} - \sqrt{\frac{s}{n}}\biggr)\biggr\} \qquad (45)\\ &= \mathcal{O}(\exp(-c_0K\log s)) \to 0,\end{aligned}$$

where $c_0 > 0$ is a universal constant. Combining the pieces, we conclude with probability $1 - \exp(-c_0K\log s)$, we have

$$\begin{aligned}\|U_S\|_{\ell_\infty/\ell_2} \le \frac{1}{b^*_{\min}}[T_1 + T_2] &\le \biggl[\sqrt{\frac{8K\log s}{C_{\min}n}} + \lambda_n\biggl(D_{\max} + \frac{6}{C_{\min}}\sqrt{\frac{s^2}{n}}\biggr)\biggr]\\ &= \rho(n, s, \lambda_n),\end{aligned}$$

which establishes the bound (20) from Theorem 1(a).

Moreover, under the assumptions of Theorem 1(b), we can conclude that

$$\frac{\|U_S\|_{\ell_\infty/\ell_2}}{b^*_{\min}} \le \frac{\rho(n, s, \lambda_n)}{b^*_{\min}} = o(1), \qquad (46)$$

with probability greater than $1-\Theta(\exp(-c_0K\log s)) \to 1$. Consequently, the conditions of Theorem 1(b) are sufficient to ensure that the event $\mathcal{E}(U_S)$ holds with high probability as claimed.

REMARK. As we noted following the statement of Theorem 1, the fact that the claims hold with probability converging to one only if $s \to +\infty$ might appear counter-intuitive and does not allow the result to cover problems with fixed sizes s of the row support. Here we discuss how this condition can be weakened. Note that our assumptions imply that $p - s \to \infty$ and that $\frac{s}{n} = o(1)$. Consequently, for any $a > 0$, we have $\frac{\log s}{n^a} = \frac{\log s}{s^a}\frac{s^a}{n^a} = o(1)$, so that we may use a slightly weaker bound on T_1 in equation (45). Indeed, with the same notation as in that equation, we have

$$\begin{aligned}&\mathbb{P}\biggl[T_1 \ge \sqrt{\frac{4(K + \log s + n^a)}{C_{\min}n}}\Big|\mathcal{T}\biggr]\\ &\quad\le \mathbb{P}\biggl[\frac{1}{n}\max_{i\in S}\zeta_i^2 \ge \frac{2}{n}(K + \log s + n^a)\Big|\mathcal{T}\biggr]\\ &\quad\le \exp\biggl\{-n^a\biggl(1 - 2\biggl(1 + \frac{\log s}{n^a}\biggr)\sqrt{\frac{K}{K + n^a}}\biggr)\biggr\} \to 0,\end{aligned}$$

where the last inequality is obtained by setting $t = K + \log s + n^a$ in Lemma 11 of Appendix I.

4. Proof of Theorem 2. In this section, we prove the necessary conditions stated in Theorem 2. We begin by noting that we may assume without loss of generality that $s < n$, since it is otherwise impossible to recover the support (even in the absence of noise). In order to develop some intuition for the argument to follow, recall the definition (36) of the event $\mathcal{E}(V_{S^c})$. The proof of Theorem 2 is based on the fact that if $\mathcal{E}(V_{S^c})$ does not hold, then no solution of the multivariate group Lasso has the correct row support.

Again, to lighten notation, we write V for the quantity V_{S^c}. Recall the definitions (37) of the quantities T_i' for $i = 1$, 2 and 3. By the triangle inequality, we have

$$\frac{1}{\lambda_n}\|V\|_{\ell_\infty/\ell_2} \geq T_3' - T_2' - T_1'. \tag{47}$$

From our earlier argument [see equation (39)], we know that $T_1' \leq (1-\gamma)$. From the bound (40), in order to show that $T_2' = o(1)$ with high probability, it suffices to show that $\|\widehat{Z}_S - Z_S^*\|_{\ell_\infty/\ell_2} = o(1)$. We reason by contradiction and assume that in the regime considered in Theorem 2, there is a solution of the multivariate group Lasso which satisfies $\|\widehat{B} - B^*\|_{\ell_\infty/\ell_2} = o(b^*_{\min})$ with high probability. Note that this condition implies that $\max_{i\in S} \frac{\|\widehat{B}_i - B_i^*\|_2}{\|B_i^*\|_2} = o(1)$, so that we may apply Lemma 3 to conclude that $\|\widehat{Z}_S - Z_S^*\|_{\ell_\infty/\ell_2} = o(1)$ as well. Consequently, we conclude that $T_2' = o(1)$.

Considering the decomposition (47), we obtain that

$$T_3' - T_2' - T_1' = \frac{1}{\lambda_n}\|V - \mathbb{E}[V|X_S, W]\|_{\ell_\infty/\ell_2} - (1-\gamma) - o(1). \tag{48}$$

Therefore, it suffices to prove that $T_3' > 2 - \gamma$. The remainder of the proof is devoted to establishing this claim.

In order to analyze T_3', let us recall the notation $\widetilde{V}_j = V_j - \mathbb{E}[V_j|X_S, W]$, where for each $j \in S^c$, the quantity $V_j \in \mathbb{R}^K$ denotes the jth row of the matrix V. As shown earlier in Section 3.2, we can write

$$(\|\widetilde{V}_j\|_2^2 | W, X_S) \stackrel{d}{=} \Sigma_{jj|S}\xi_j^T M_n \xi_j,$$

where for each $j \in S^c$, the random vector $\xi_j \sim N(0, I_K)$. The random vectors $(\xi_j, j \in S^c)$ are not i.i.d. in general, since for each pair $i, j \in S^c$, we have $\operatorname{cov}(\xi_i, \xi_j) = \frac{\Sigma_{ij|S}}{\sqrt{\Sigma_{ii|S}\Sigma_{jj|S}}} I_K$.

The next part of the proof is devoted to analyzing the behavior of the random variable

$$V_{\max} := \max_{j\in S^c} \|\widetilde{V}_j\|_2 = \max_{j\in S^c} \sqrt{\Sigma_{jj|S}\xi_j^T M_n \xi_j}, \tag{49}$$

with our goal in particular being to show that $\frac{V_{\max}}{\lambda_n} \geq 2 - \gamma$ with high probability. In order to lower bound the random variable $V_{\max}$, our first step is to show that it is sharply concentrated around its expectation.

LEMMA 5. *For any* $\delta > 0$, *we have*

$$\mathbb{P}\big[|V_{\max} - \mathbb{E}[V_{\max}]| \geq \delta | X_S, W\big] \leq 4 \exp\left\{-\frac{1}{2}\frac{\delta^2}{\rho_u(\Sigma_{S^c S^c|S}) |||M_n|||_2}\right\}, \tag{50}$$

where $\rho_u(\Sigma_{S^c S^c|S}) = \max_{j \in S^c} \Sigma_{jj|S}$.

PROOF. By standard Gaussian concentration theorems [e.g., Theorem 3.8 of Massart (2003)], if X has a standard Gaussian measure on $\mathbb{R}^m$ and f is a Lipschitz function with Lipschitz constant L, then

$$\mathbb{P}\big[|\mathbb{E}[f(X)] - f(X)| \geq x\big] \leq 4 \exp\big(-x^2/(2L^2)\big). \tag{51}$$

In order to exploit this result in application to $V_{\max}$, we consider the function $f : \mathbb{R}^{(p-s)\times K} \to \mathbb{R}$ defined by

$$f(\xi_j, j \in S^c) := \max_{j \in S^c} \sqrt{\Sigma_{jj|S}} \big\|\sqrt{M_n}\xi_j\big\|_2,$$

which is equal to $V_{\max}$ by construction. Let $u = (u_j, j \in S^c)$ and $v = (v_j, j \in S^c)$ be two collections of vectors. We have

$$\begin{aligned}
|f(u) - f(v)| &= \max_{j \in S^c} \sqrt{\Sigma_{jj|S}} \big\|\sqrt{M_n} u_j\big\|_2 - \max_{k \in S^c} \sqrt{\Sigma_{kk|S}} \big\|\sqrt{M_n} v_k\big\|_2 \\
&\leq \max_{j \in S^c} \sqrt{\Sigma_{jj|S}} \big\|\sqrt{M_n}(u_j - v_j)\big\|_2 \\
&\leq \sqrt{\rho_u(\Sigma_{S^c S^c|S})} \sqrt{|||M_n|||_2} \|u - v\|_2.
\end{aligned}$$

We may therefore apply the bound (51) with $L^2 = |||M_n|||_2 \rho_u(\Sigma_{S^c S^c|S})$ to obtain the claim. $\square$

The second key ingredient in our proof is a lower bound on the expected value of $V_{\max}$:

LEMMA 6. *For any fixed* $\delta' > 0$, *with probability* $1 - o(1)$ *as* $(p - s) \to +\infty$, *we have*

$$\mathbb{E}[V_{\max} | X_S, W] \geq \sqrt{|||M_n|||_2} \sqrt{2(1 - \delta')\rho_\ell(\Sigma_{S^c S^c|S}) \log(p - s)}. \tag{52}$$

PROOF. We may diagonalize M_n, writing $M_n = U^T D U$, where $U \in \mathbb{R}^{K \times K}$ is orthogonal, and $D = \text{diag}\{d_1, \ldots, d_K\}$ is diagonal with $d_1 = |||M_n|||_2$. Since the distribution of the K-dimensional normal vector $\xi_j \sim N(0, I)$ remains invariant under orthogonal transformations, for each $j \in S^c$, we can write

$$\sqrt{\Sigma_{jj|S}\xi_j^T M_n \xi_j} \stackrel{d}{=} \sqrt{\Sigma_{jj|S}\eta_j^T D\eta_j} \geq \sqrt{\Sigma_{jj|S}|||M_n|||_2}|\eta_{j,1}|,$$

where $\eta_{j,1} \sim N(0, \Sigma_{jj|S})$. Overall, we have

$$\mathbb{E}[V_{\max}|X_S, W] = \mathbb{E}\Big[\max_{j\in S^c}\sqrt{\Sigma_{jj|S}\xi_j^T M_n\xi_j}|X_S, W\Big] \geq \sqrt{|||M_n|||_2}\mathbb{E}\Big[\max_{j\in S^c}|\eta_{j,1}|\Big],$$

where the vector $\eta = (\eta_{j,1}, j \in S^c)$ is zero-mean Gaussian with covariance $\Sigma_{S^cS^c|S}$.

Our next step is to lower bound the expectation $\mathbb{E}[\max_{j\in S^c}|\eta_{j,1}|]$ by a Gaussian comparison argument, in particular exploiting the Sudakov–Fernique inequality [Ledoux and Talagrand (1991)]. Let $\widetilde{\eta} \in \mathbb{R}^{p-s}$ be a Gaussian random vector with i.i.d. $N(0, 1)$ entries. By the definition (18a) of $\rho_\ell(\cdot)$, we have

$$\begin{aligned}\mathbb{E}[(\eta_i - \eta_j)^2] &= \Sigma_{ii|S} - 2\Sigma_{ij|S} + \Sigma_{jj|S}\\ &\geq \rho_\ell(\Sigma_{S^cS^c|S})\mathbb{E}[(\widetilde{\eta}_j - \widetilde{\eta}_i)^2] \qquad \text{for all } i, j.\end{aligned}$$

Consequently, the Sudakov–Fernique inequality implies that

$$\mathbb{E}\Big[\max_{j\in S^c}|\eta_j|\Big] \geq \sqrt{\rho_\ell(\Sigma_{S^cS^c|S})}\mathbb{E}\Big[\max_{j\in S^c}|\widetilde{\eta}_j|\Big].$$

From standard results on Gaussian extrema [Ledoux and Talagrand (1991)], for any fixed $\delta' \in (0, 1)$, we have $\mathbb{E}[\max_{j\in S^c}|\widetilde{\eta}_j|] \geq \sqrt{2(1-\delta')\log(p-s)}$ once $(p-s)$ is sufficiently large, which completes the proof. $\square$

It remains to show that the random matrix $|||M_n|||_2$ previously defined (41) is suitably concentrated. Our approach is to show that unless the hypotheses of Lemma 4—namely, $s/n = o(1)$ and $\|\widehat{Z} - Z^*\|_{\ell_\infty/\ell_2} = o(1)$—are both satisfied, then the multivariate group Lasso fails. We have shown previously that the latter condition is satisfied, so it remains to show that the condition $s/n = o(1)$ must hold. Note that

$$|||M_n|||_2 \geq \frac{\lambda_n^2}{n}|||(\widehat{Z}_S)^T(\widehat{\Sigma}_{SS})^{-1}\widehat{Z}_S|||_2.$$

By definition of the sub-differential of the ℓ_1/ℓ_2 norm, we have $|||\widehat{Z}_S|||_F^2 = s$, so that there must be at least one column of $\widehat{Z}_S$ with squared ℓ_2 norm greater than s/K. Without loss of generality, let us assume that it is the first column $\widehat{Z}_1 \in \mathbb{R}^s$. We then have

$$\begin{aligned}|||M_n|||_2 &\geq \frac{\lambda_n^2}{n}\widehat{Z}_1^T(\widehat{\Sigma}_{SS})^{-1}\widehat{Z}_1\\ &\geq \frac{\lambda_n^2 s}{nK}\lambda_{\min}((\widehat{\Sigma}_{SS})^{-1})\\ &\geq \frac{\lambda_n^2 s}{nK}\frac{1}{\lambda_{\max}(\widehat{\Sigma}_{SS})}.\end{aligned}$$

From concentration of random matrix eigenvalues [see equation (60) in Appendix F], we have $\lambda_{\max}(\widehat{\Sigma}_{SS}) \le 2\lambda_{\max}(\Sigma_{SS})$ with probability greater than $1 - \exp(-\Theta(n))$, so that we conclude that the lower bound $|||M_n|||_2 \ge \frac{\lambda_n^2 s}{2Kn}$ holds with high probability (w.h.p.).

Substituting this lower bound into the lower bound (52) from Lemma 6, we obtain that w.h.p. for any $\delta' \in (0,1)$,

$$\frac{1}{\lambda_n}\mathbb{E}[V_{\max}|X_S, W] \ge \sqrt{\frac{s}{2Kn}}\sqrt{2(1-\delta')\rho_\ell(\Sigma_{S^cS^c|S})\log(p-s)}, \tag{53}$$

which tends to infinity unless $s/n = o(1)$. By the concentration around this expected value from Lemma 5, this fact implies that the multivariate group Lasso fails w.h.p. unless $s/n = o(1)$.

We have thus shown that the conditions of Lemma 4 are necessary conditions for the multivariate group Lasso to succeed, and given that these conditions are satisfied, the quantity $|||M_n|||_2$ is concentrated. Recalling the definition of the event $\mathcal{T}(\delta)$ from equation (43), we can write

$$\mathbb{P}\left[\frac{V_{\max}}{\lambda_n} \le 2-\gamma\right] \le \mathbb{P}[T_3' \le 2-\gamma|\mathcal{T}(\delta)] + \mathbb{P}[\mathcal{T}(\delta)^c],$$

where we are guaranteed that $\mathbb{P}[\mathcal{T}(\delta)^c] \to 0$ by Lemma 4.

Recall that we have established that $\frac{s}{n} = o(1)$. Conditioned on the event $\mathcal{T}(\delta)$, the inequality $|||M_n|||_2 \ge \lambda_n^2 \frac{\psi(B^*)}{n}(1-\delta)$ holds; combined with the lower bound (52), for any $\delta' \in (0,1)$, we have for $(p-s)$ sufficiently large and if $\frac{s}{n} = o(1)$ that

$$\begin{aligned}&\frac{1}{\lambda_n}\mathbb{E}[V_{\max}|\mathcal{T}(\delta), X_S, W]\\ &\quad \ge \sqrt{\frac{\psi(B^*)}{n}(1-\delta)}\sqrt{2(1-\delta')\rho_\ell(\Sigma_{S^cS^c|S})\log(p-s)}.\end{aligned}$$

Consequently, if the lower bound (21) holds strictly, then for $(p-s)$ sufficiently large, denoting $\theta_\ell := \rho_\ell(\Sigma_{S^cS^c|S})/(2-\gamma)^2$ and $\delta'' := \sqrt{(1-\delta')(1-\delta)} - 1$ we have

$$\begin{aligned}&\frac{1}{\lambda_n}\mathbb{E}[V_{\max}|\mathcal{T}(\delta), X_S, W]\\ &\quad \ge (2-\gamma)\sqrt{\frac{2\theta_\ell\psi(B^*)\log(p-s)}{n}}(1-\delta'')\\ &\quad \ge \frac{2-\gamma}{\sqrt{1-\nu}}(1-\delta'') \ge (2-\gamma)\left(1+\frac{\nu}{2}\right)(1-\delta'') \ge 2-\gamma+\varepsilon\end{aligned}$$

with[3] $\varepsilon = (2-\gamma)\frac{\nu}{3}$.

[3]Here we have used the fact that for δ, δ' sufficiently small, we have $(1-\delta'')(1+\frac{\nu}{2}) \ge (1+\frac{\nu}{3})$.

Combining this lower bound with the concentration statement from Lemma 5, we obtain

$$\begin{aligned}\mathbb{P}\left[\frac{V_{\max}}{\lambda_n} \le 2-\gamma \,\middle|\, \mathcal{T}(\delta)\right] &\le 4\exp\left\{-\frac{1}{2}\left(\frac{\varepsilon^2}{\rho_u(\Sigma_{S^cS^c|S})}\frac{n}{\psi(B^*)(1-\delta)}\right)\right\}\\ &\le 4\exp\left\{-\frac{1}{2}\left(\frac{\varepsilon^2 C_{\max}}{\rho_u(\Sigma_{S^cS^c|S})}\frac{Kn}{s(1-\delta)}\right)\right\}\\ &\le 4\exp\left\{-c'\frac{Kn}{s}\right\},\end{aligned}$$

where we have defined the constant $c' := \frac{(2-\gamma)^2 C_{\max}}{18\gamma^2\theta_u(1-\delta)}$, and used the facts that $\varepsilon = (2-\gamma)\frac{\nu}{3}$ and $\theta_u := \rho_u(\Sigma_{S^cS^c|S})/\gamma^2$. Therefore, the probability vanishes, since the condition $s/n = o(1)$ is equivalent to $n/s \to +\infty$.

5. Discussion. In this paper, we have analyzed the high-dimensional behavior of block-regularization for multivariate regression problems. Our main result is to show that that its behavior is governed by the sample complexity parameter,

$$\theta_{\ell_1/\ell_2}(n, p, s) := n/[2\psi(B^*)\log(p-s)],$$

where n is the sample size, p is the ambient dimension and $\psi(\cdot)$ is a sparsity-overlap function that measures a combination of the sparsity and overlap properties of the true regression matrix B^*. In particular, Theorems 1 and 2 show that the multivariate group Lasso either succeeds (or fails) depending on whether this sample complexity parameter is larger (or smaller) than a threshold parameter depending in the design covariance matrix Σ.

Our results were obtained under high-dimensional scaling, in particular, assuming the quantities n, $p-s$ and s all were tending to infinity. As have discussed, the hypothesis that $s \to +\infty$ can be relaxed at the expense of slightly weaker guarantees on the ℓ_2/ℓ_∞ norm of the solution. One could also imagine relaxing the constraint $p-s \to +\infty$, but for the high-dimensional problems that motivate our analysis, this is not as interesting, since in such a case, either the true model is nonsparse (and hence variable selection is of questionable relevance), or we fall back in the low-dimensional setting.

There are a number of open questions associated with this work. The current work applies to the "hard"-sparsity model, in which a subset S of the regressors are nonzero, and the remaining coefficients are zero. As with the ordinary Lasso, it would also be interesting to study block-regularization under soft sparsity models (e.g., ℓ_q "balls" for coefficients, with $q < 1$). It is also interesting to consider alternative loss functions such as ℓ_2 error or prediction error, as opposed to the exact support recovery criterion considered here. We note that since this work was first posted, other researchers have provided related results on consistency in ℓ_2 error [Lounici et al. (2009); Huang and Zhang (2009)], again under hard sparsity constraints.

APPENDIX A: PROOF OF COROLLARY 2

Let $\mathcal{F}$ (resp., $\mathcal{F}_0$) be the event that the thresholded ROLS method fails to recover the individual supports when applied to the estimated row set $\widehat{S}$ (resp., true row set S). By a union bound, the overall probability of failure in the multi-stage procedure is upper bounded as $\mathbb{P}[\mathcal{F}] \le \mathbb{P}[\widehat{S} \neq S] + \mathbb{P}[\mathcal{F}_0 \mid \widehat{S} = S]$. Under the conditions of Theorem 1, the row support is recovered with probability greater than $1 - \Theta(\exp(-c_0 K \log s))$, so that $\mathbb{P}[\widehat{S} \neq S] \to 0$. As for the remaining term, we have $\mathbb{P}[\mathcal{F}_0 \mid \widehat{S} = S] \le \frac{\mathbb{P}[\mathcal{F}_0]}{\mathbb{P}[\widehat{S}=S]}$, which is less than $2\mathbb{P}[\mathcal{F}_0]$ for (n, s) large enough, since $\mathbb{P}[\widehat{S} = S] \to 1$.

Consequently, it suffices to upper bound the unconditional probability that the ROLS estimate applied to the true support fails to recover the individual supports. Introducing the shorthand $\widehat{\Sigma}_{SS} := \frac{1}{n} X_S^T X_S$, some straightforward linear algebra shows that the ROLS estimate of B_S^* takes the form $\widehat{B}_S = B_S^* + \widetilde{U}$, where $\widetilde{U} := (\widehat{\Sigma}_{SS})^{-1/2}\widetilde{W}/\sqrt{n}$, and $\widetilde{W} := (\widehat{\Sigma}_{SS})^{-1/2} X_S^T W/\sqrt{n}$ is an $s \times K$ noise matrix with i.i.d. standard Gaussian entries.

Let $\widetilde{W}^{(j)}$ denote the jth column of $\widetilde{W}$, and let e_i denote the ith canonical basis vector in $\mathbb{R}^s$. We then have

$$\begin{aligned}\max_{i,j} |\widetilde{U}_{i,j}| = \max_{i,j} \frac{1}{\sqrt{n}} |e_i^T (\widehat{\Sigma}_{SS})^{-1/2} \widetilde{W}^{(j)}| &\le \frac{1}{\sqrt{n}} \max_i \Big[\|(\widehat{\Sigma}_{SS})^{-1/2} e_i\| \max_j |\xi_{i,j}| \Big] \\ &\le \frac{1}{\sqrt{n}} |||(\widehat{\Sigma}_{SS})^{-1/2}|||_2 \max_{i,j} |\xi_{i,j}|,\end{aligned}$$

where $(\xi_{i,j})$ forms a sequence of identically distributed standard Gaussian variables (which are dependent in general). Using a union bound and standard Gaussian tail bounds, for all $\nu > 0$, we have

$$\mathbb{P}\Big[\max_{i,j} |\xi_{i,j}| \ge (1+\nu)\sqrt{2\log(Ks)}\Big] \le 2\exp(-\nu \log(Ks)) \to 0.$$

A concentration bound for random matrices (see Appendix F) yields $|||\widehat{\Sigma}_{SS}^{-1/2}|||_2 \le \sqrt{2} C_{\min}^{-1/2}$ with probability greater than $1 - \exp(-\Theta(n))$, so that we obtain

$$\mathbb{P}\Bigg[\max_{i,j} |\widetilde{U}_{i,j}| \ge (1+\nu)\sqrt{\frac{4\log Ks}{C_{\min} n}}\Bigg] = \mathcal{O}(\exp(-\Theta(\log s))).$$

This result, together with the lower bound on the smallest absolute value of the nonzero coefficients of B^*, shows that the threshold procedure in step 3 will retain all nonzero coefficients of B^* while correctly setting to zero all entries for which B^* is actually zero.

APPENDIX B: PROOF OF LEMMA 2

Using the notation β_i to denote a row of B and denoting by

$$\mathcal{K} := \{(w, v) \in \mathbb{R}^K \times \mathbb{R} \mid \|w\|_2 \le v\} \tag{54}$$

the usual second-order cone (SOC), we can rewrite the original convex program (12) with $q = 2$ as

$$\min_{\substack{B\in\mathbb{R}^{p\times K}\\ b\in\mathbb{R}^p}} \frac{1}{2n}|||Y - XB|||_F^2 + \lambda_n \sum_{i=1}^{p} b_i \tag{55}$$
$$\text{s.t. } (\beta_i, b_i) \in \mathcal{K},\ 1 \le i \le p.$$

We now dualize the conic constraints [Boyd and Vandenberghe (2004)], using conic Lagrange multipliers belonging to the dual cone $\mathcal{K}^* = \{(z,t) \in \mathbb{R}^{K+1} | z^T w + vt \ge 0, (w, v) \in \mathcal{K}\}$. The second-order cone $\mathcal{K}$ is self-dual [Boyd and Vandenberghe (2004)], so that the convex program (55) is equivalent to

$$\min_{\substack{B\in\mathbb{R}^{p\times K}\\ b\in\mathbb{R}^p}} \max_{\substack{Z\in\mathbb{R}^{p\times K}\\ t\in\mathbb{R}^p}} \frac{1}{2n}|||Y - XB|||_F^2 + \lambda_n \sum_{i=1}^{p} b_i - \lambda_n \sum_{i=1}^{p}(-z_i^T\beta_i + t_i b_i)$$
$$\text{s.t. } (z_i, t_i) \in \mathcal{K},\ 1 \le i \le p,$$

where Z is the matrix whose ith row is z_i.

Since the original program is convex and strictly feasible, strong duality holds and any pair of primal $(B^\star, b^\star)$ and dual $(Z^\star, t^\star)$ solutions has to satisfy the Karush–Kuhn–Tucker conditions:

$$\|\beta_i^\star\|_2 \le b_i^\star, \qquad 1 < i < p, \tag{56a}$$

$$\|z_i^\star\|_2 \le t_i^\star, \qquad 1 < i < p, \tag{56b}$$

$$z_i^{\star T}\beta_i^\star - t_i^\star b_i^\star = 0, \qquad 1 < i < p, \tag{56c}$$

$$\nabla_B\left[\frac{1}{2n}|||Y - XB|||_F^2\right]\Big|_{B=B^\star} + \lambda_n Z^\star = 0, \tag{56d}$$

$$\lambda_n(1 - t_i^\star) = 0. \tag{56e}$$

Since equations (56c) and (56e) impose the constraints $t_i^\star = 1$ and $b_i^\star = \|\beta_i^\star\|_2$, a primal–dual solution to this conic program is determined by $(B^\star, Z^\star)$.

Any solution satisfying the conditions in Lemma 2 also satisfies these KKT conditions, since equation (31b) and the definition (31c) are equivalent to equation (56d), and equation (31a) and the combination of conditions (31d) and (31c) imply that the complementary slackness equations (56c) hold for each primal–dual conic pair (β_i, z_i).

Now consider some other primal solution $\widetilde{B}$; when combined with the optimal dual solution $\widehat{Z}$, the pair $(\widetilde{B}, \widehat{Z})$ must satisfy the KKT conditions [Bertsekas (1995)]. But since for $j \in S^c$, we have $\|\hat{z}_j\|_2 < 1$, then the complementary slackness condition (56c) implies that for all $j \in S^c$, $\widetilde{\beta}_j = 0$. This fact in turn implies that the primal solution $\widetilde{B}$ must also be a solution to the restricted convex program (32), obtained by only considering the covariates in the set S or equivalently by setting $B_{S^c} = 0_{S^c}$. But since $s < n$ by assumption, the matrix $X_S^T X_S$ is strictly positive definite with probability one, and therefore the restricted convex program (32) has a unique solution $B_S^\star = \widehat{B}_S$. We have thus shown that a solution $(\widehat{B}, \widehat{Z})$ to the program (12) that satisfies the conditions of Lemma 2, if it exists, must be unique.

APPENDIX C: CHARACTERIZATION OF THE SPARSITY-OVERLAP FUNCTION

In this appendix, we prove Lemma 1. (a) To verify this claim, we first set $Z_S^* = \zeta(B_S^*)$, and use $Z_S^{(k)*}$ to denote the kth column of Z_S^*. Since the spectral norm is upper bounded by the sum of eigenvalues, and lower bounded by the average eigenvalue, we have

$$\frac{1}{K}\operatorname{tr}(Z_S^{*T}\Sigma_{SS}^{-1}Z_S^*) \le \psi(B^*) \le \operatorname{tr}(Z_S^{*T}\Sigma_{SS}^{-1}Z_S^*).$$

Given our assumption (A1) on Σ_{SS}, we have

$$\operatorname{tr}(Z_S^{*T}\Sigma_{SS}^{-1}Z_S^*) = \sum_{k=1}^K Z_S^{(k)*T}\Sigma_{SS}^{-1}Z_S^{(k)*} \ge \frac{1}{C_{\max}}\sum_{k=1}^K \|Z_S^{(k)*}\|^2 = \frac{s}{C_{\max}},$$

using the fact that $\sum_{k=1}^K \|Z_S^{(k)*}\|^2 = \sum_{i=1}^s \|Z_i^*\|^2 = s$. Similarly, in the other direction, we have

$$\operatorname{tr}(Z_S^{*T}\Sigma_{SS}^{-1}Z_S^*) = \sum_{k=1}^K Z_S^{(k)*T}\Sigma_{SS}^{-1}Z_S^{(k)*} \le \frac{1}{C_{\min}}\sum_{k=1}^K \|Z_S^{(k)*}\|^2 = \frac{s}{C_{\min}},$$

which completes the proof.

(b) Under the assumed orthogonality, the matrix $Z^{*T}Z^*$ is diagonal with $\|Z^{(k)*}\|^2$ as the diagonal elements, so that the largest $\|Z^{(k)*}\|^2$ is then the largest eigenvalue of the matrix.

APPENDIX D: GROUP LASSO VERSUS ORDINARY LASSO

In this appendix, we provide the proof of Corollary 3 which characterizes the relative efficiency of the group versus the ordinary Lasso. From the discussion preceding the statement of Corollary 3, we know that the quantity

$$\max_{k=1,\dots,K}\psi\bigl(\beta_S^{*(k)}\bigr)\log(p-s_k) = \max_{k=1,\dots,K} s_k\log(p-s_k) \ge \max_{k=1,\dots,K} s_k\log(p-s)$$

governs the performance of the ordinary Lasso procedure for row selection. It remains to show then that $\psi(B_S^*) \le \max_k s_k$.

As before, we use the notation $Z_S^* = \zeta(B_S^*)$, and Z_i^* for the ith row of Z_S^*. Since $\Sigma_{SS} = I_{s\times s}$, we have $\psi(B^*) = \|Z_S^*\|^2$. Consequently, by the variational representation of the ℓ_2-norm, we have

$$\psi(B^*) = \max_{x\in\mathbb{R}^K : \|x\|\le 1}\|Z_S^*x\|^2 \le \max_{x\in\mathbb{R}^K : \|x\|\le 1}\sum_{i=1}^s (Z_i^{*T}x)^2.$$

Let $|Z_i^*| = (|Z_{i1}^*|, \dots, |Z_{ik}^*|)^T$ and $y_i = (x_1 \operatorname{sign}(Z_{i1}^*), \dots, x_K \operatorname{sign}(Z_{iK}^*))^T$. By the Cauchy–Schwarz inequality,

$$(Z_i^{*T} x)^2 = (|Z_i^*|^T y_i)^2 \le \| |Z_i^*| \|^2 \|y_i\|^2 = \|Z_i^*\|^2 \sum_k x_k^2 \operatorname{sign}(Z_{ik}^*)^2$$

so that

$$\sum_{i=1}^s (Z_i^{*T} x)^2 \le \sum_{i=1}^s \|Z_i^*\|^2 \sum_{k=1}^K x_k^2 \operatorname{sign}(Z_{ik}^*)^2 = \sum_{k=1}^K x_k^2 \sum_{i=1}^s \operatorname{sign}(Z_{ik}^*)^2 = \sum_{k=1}^K x_k^2 s_k,$$

and if $\|x\| \le 1$, we have $\sum_{k=1}^K x_k^2 s_k \le \max_{1 \le k \le K} s_k$ thereby establishing the claim.

APPENDIX E: INEQUALITIES WITH BLOCK-MATRIX NORMS

In general, the two families of matrix norms that we have introduced, $||| \cdot |||_{p,q}$ and $\| \cdot \|_{\ell_a/\ell_b}$, are distinct, but they coincide in the following useful special case:

LEMMA 7. *For $1 \le p \le \infty$ and for r defined by $1/r + 1/p = 1$ we have*

$$\| \cdot \|_{\ell_\infty/\ell_p} = ||| \cdot |||_{\infty,r}.$$

PROOF. Indeed, if a_i denotes the ith row of A, then

$$\begin{aligned} \|A\|_{\ell_\infty/\ell_p} &= \max_i \|a_i\|_p = \max_i \max_{\|y_i\|_r \le 1} y_i^T a_i \\ &= \max_{\|y\|_r \le 1} \max_i |y^T a_i| = \max_{\|y\|_r \le 1} \|Ay\|_\infty. \end{aligned}$$

□

We conclude by stating some useful bounds and relations:

LEMMA 8. *Consider matrices $A \in \mathbb{R}^{m \times n}$ and $Z \in \mathbb{R}^{n \times \ell}$ and $p, r > 0$ with $\frac{1}{p} + \frac{1}{r} = 1$, we have*

$$\|AZ\|_{\ell_\infty/\ell_p} = |||AZ|||_{\infty,r} \le |||A|||_{\infty,\infty} |||Z|||_{\infty,r} = |||A|||_{\infty,\infty} \|Z\|_{\ell_\infty/\ell_p}, \tag{57a}$$

$$|||A|||_r \le |||I_m|||_{r,\infty} |||A|||_{\infty,r} = s^{1/r} \|A\|_{\ell_\infty/\ell_p}. \tag{57b}$$

APPENDIX F: SOME CONCENTRATION INEQUALITIES FOR RANDOM MATRICES

In this appendix, we state some known concentration inequalities for the extreme eigenvalues of Gaussian random matrices. Although these results hold more generally, our interest here is on scalings (n, s) such that $s/n \to 0$. The following result is from Davidson and Szarek (2001).

LEMMA 9. *Let $U \in \mathbb{R}^{n\times s}$ be a random matrix from the standard Gaussian ensemble* [*i.e.*, $U_{ij} \sim N(0,1)$, *i.i.d.*]. *Then if we denote by $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ the smallest and largest singular value of U, respectively, we have*

$$\mathbb{P}\left[1-\lambda_{\min}\left(\frac{U}{\sqrt{n}}\right) \geq \sqrt{\frac{s}{n}} + t\right] \leq \exp\left(-\frac{nt^2}{2}\right), \tag{58}$$

$$\mathbb{P}\left[\lambda_{\max}\left(\frac{U}{\sqrt{n}}\right) - 1 \geq \sqrt{\frac{s}{n}} + t\right] \leq \exp\left(-\frac{nt^2}{2}\right). \tag{59}$$

As a consequence, for $s/n \to 0$ we obtain the two following inequalities:

LEMMA 10.

$$\mathbb{P}\left[\left\|\frac{1}{n}U^T U\right\|_2 \leq \frac{1}{2}\right] \leq \exp\left\{-\frac{n}{2}\left(\frac{1}{4} - \sqrt{\frac{s}{n}}\right)_+^2\right\}, \tag{60}$$

$$\mathbb{P}\left[\left\|\frac{1}{n}U^T U - I_{s\times s}\right\|_2 \geq 6\sqrt{\frac{s}{n}}\right] \leq 2\exp\left(-\frac{s}{2}\right) + \exp(-\Theta(n)) \to 0. \tag{61}$$

PROOF. For simplicity, we write $\lambda_{\min}$ for $\lambda_{\min}(\frac{U}{\sqrt{n}})$ and $\lambda_{\max}$ for $\lambda_{\max}(\frac{U}{\sqrt{n}})$. For equation (60), we have

$$\begin{aligned}\mathbb{P}\left[\left\|\frac{1}{n}U^T U\right\|_2 \leq \frac{1}{2}\right] &\leq \mathbb{P}\left[\lambda_{\min} \leq \frac{1}{\sqrt{2}}\right] \leq \mathbb{P}\left[1-\lambda_{\min} \geq \sqrt{\frac{s}{n}} + \left(\frac{1}{4} - \sqrt{\frac{s}{n}}\right)\right] \\ &\leq \exp\left\{-\frac{n}{2}\left(\frac{1}{4} - \sqrt{\frac{s}{n}}\right)_+^2\right\}.\end{aligned}$$

For equation (61),

$$\begin{aligned}&\mathbb{P}\left[\left\|\frac{1}{n}U^T U - I_{s\times s}\right\|_2 \geq 6\sqrt{\frac{s}{n}}\right] \\ &\quad = \mathbb{P}\left[\max(\lambda_{\max}^2 - 1, 1 - \lambda_{\min}^2) \geq 6\sqrt{\frac{s}{n}}\right] \\ &\quad \leq \mathbb{P}\left[\lambda_{\max} - 1 \geq 2\sqrt{\frac{s}{n}}\right] + \mathbb{P}\left[1 - \lambda_{\min} \geq 2\sqrt{\frac{s}{n}}\right] \\ &\qquad + \mathbb{P}[\lambda_{\max} + 1 \geq 3] + \mathbb{P}[\lambda_{\min} + 1 \geq 3] \\ &\quad \leq 2\exp\left\{-\frac{n}{2}\left(\sqrt{\frac{s}{n}}\right)^2\right\} + 2\exp\left\{-\frac{n}{2}\left(\frac{1}{4} - \sqrt{\frac{s}{n}}\right)_+^2\right\},\end{aligned}$$

where we used that $\{\lambda^2 - 1 \geq x\} \subset \{\lambda - 1 \geq \frac{x}{3}\} \cup \{\lambda + 1 \geq 3\}$ to obtain the first inequality. $\square$

These results are easily adapted to more general Gaussian ensembles. Letting $X = U\sqrt{\Lambda}$, we obtain an $n \times s$ matrix with i.i.d. rows, $X_i \sim N(0, \Lambda)$. If the covariance matrix Λ has maximum eigenvalue $C_{\max} < +\infty$, then we have

$$\|\!|n^{-1}X^T X - \Lambda\|\!|_2 = \left\|\!\left|\sqrt{\Lambda}[n^{-1}U^T U - I]\sqrt{\Lambda}\right|\!\right\|_2 \leq C_{\max}\|\!|n^{-1}U^T U - I\|\!|_2 \tag{62}$$

so that the bound (61) immediately yields an analogous bound on different constants.

The final type of bound that we require is on the difference

$$\|\!|(X^T X/n)^{-1} - \Lambda^{-1}\|\!|_2,$$

assuming that $X^T X$ is invertible. We note that

$$\begin{aligned}\|\!|(X^T X/n)^{-1} - \Lambda^{-1}\|\!|_2 &= \|\!|(X^T X/n)^{-1}[\Lambda - (X^T X/n)]\Lambda^{-1}\|\!|_2 \\ &\leq \|\!|(X^T X/n)^{-1}\|\!|_2 \|\!|\Lambda - (X^T X/n)\|\!|_2 \|\!|\Lambda^{-1}\|\!|_2.\end{aligned}$$

As long as the eigenvalues of Λ are bounded below by $C_{\min} > 0$, then $\|\!|\Lambda^{-1}\|\!|_2 \leq 1/C_{\min}$. Moreover, since $s/n \to 0$, we have [from equation (60)] that $\|\!|(X^T X/n)^{-1}\|\!|_2 \leq 2/C_{\min}$ with probability converging to one exponentially in n. Thus, equation (62) implies the desired bound.

APPENDIX G: PROOF OF LEMMA 3

The analysis in Section 3.3 shows that the condition $\|\Delta_i\|_2 \leq 1/2$ implies that $\widehat{\beta}_i \neq \vec{0}$ and hence $\widehat{Z}_i = \widehat{\beta}_i / \|\widehat{\beta}_i\|_2$ for all rows $i \in S$. Therefore, using the notation $Z_i^* = \beta_i^* / \|\beta_i^*\|_2$ we have

$$\begin{aligned}\widehat{Z}_i - Z_i^* &= \frac{\widehat{\beta}_i}{\|\widehat{\beta}_i\|_2} - Z_i^* = \frac{Z_i^* + \Delta_i}{\|Z_i^* + \Delta_i\|_2} - Z_i^* \\ &= Z_i^*\left(\frac{1}{\|Z_i^* + \Delta_i\|_2} - 1\right) + \frac{\Delta_i}{\|Z_i^* + \Delta_i\|_2}.\end{aligned}$$

Note that, for $z \neq 0$, the function $g(z, \delta) = \frac{1}{\|z+\delta\|_2}$ is differentiable with respect to δ, with gradient $\nabla_\delta g(z, \delta) = -\frac{z+\delta}{2\|z+\delta\|_2^3}$. By the mean-value theorem, there exists $h \in [0, 1]$ such that

$$\frac{1}{\|z + \delta\|_2} - 1 = g(z, \delta) - g(z, 0) = \nabla_\delta g(z, h\delta)^T \delta = -\frac{(z + h\delta)^T \delta}{2\|z + h\delta\|_2^3},$$

which implies that there exists $h_i \in [0, 1]$ such that

$$\begin{aligned}\|\widehat{Z}_i - Z_i^*\|_2 &\le \|Z_i^*\|_2 \frac{|(Z_i^* + h_i\Delta_i)^T \Delta_i|}{2\|Z_i^* + h_i\Delta_i\|_2^3} + \frac{\|\Delta_i\|_2}{\|Z_i^* + \Delta_i\|_2} \\ &\le \frac{\|\Delta_i\|_2}{2\|Z_i^* + h_i\Delta_i\|_2^2} + \frac{\|\Delta_i\|_2}{\|Z_i^* + \Delta_i\|_2}.\end{aligned} \tag{63}$$

We note that $\|Z_i^*\|_2 = 1$ and $\|\Delta_i\|_2 \le \frac{1}{2}$ imply that $\|Z_i^* + h_i\Delta_i\|_2 \ge \frac{1}{2}$. Combined with inequality (63), we obtain $\|\widehat{Z}_i - Z_i^*\|_2 \le 4\|\Delta_i\|_2$, which proves the lemma.

APPENDIX H: PROOF OF LEMMA 4

With $Z_S^* = \zeta(B_S^*)$, define the $K \times K$ random matrix

$$M_n^* := \frac{\lambda_n^2}{n}(Z_S^*)^T(\widehat{\Sigma}_{SS})^{-1}Z_S^* + \frac{1}{n^2}W^T(I_n - \Pi_S)W$$

and note that (using standard results on Wishart matrices [Anderson (1984)])

$$\mathbb{E}[M_n^*] = \frac{\lambda_n^2}{n-s-1}(Z_S^*)^T(\Sigma_{SS})^{-1}Z_S^* + \sigma^2\frac{n-s}{n^2}I_K. \tag{64}$$

To bound M_n in spectral norm, we use the triangle inequality,

$$\begin{aligned}\big|\|\!|M_n\|\!|_2 - \|\!|\mathbb{E}[M_n^*]\|\!|_2\big| &\le \|\!|M_n - \mathbb{E}[M_n^*]\|\!|_2 \\ &\le \underbrace{\|\!|M_n - M_n^*\|\!|_2}_{A_1} + \underbrace{\|\!|M_n^* - \mathbb{E}[M_n^*]\|\!|_2}_{A_2}.\end{aligned} \tag{65}$$

Considering the term A_1 in the decomposition (65), we have

$$\begin{aligned}&\|\!|M_n^* - M_n\|\!|_2 \\ &\quad= \frac{\lambda_n^2}{n}\|\!|Z_S^*\widehat{\Sigma}_{SS}^{-1}Z_S^* - \widehat{Z}_S\widehat{\Sigma}_{SS}^{-1}\widehat{Z}_S\|\!|_2 \\ &\quad= \frac{\lambda_n^2}{n}\big\|\!\big|Z_S^*\widehat{\Sigma}_{SS}^{-1}(Z_S^* - \widehat{Z}_S) + (Z_S^* - \widehat{Z}_S)\widehat{\Sigma}_{SS}^{-1}\big(Z_S^* + (\widehat{Z}_S - Z_S^*)\big)\big\|\!\big|_2 \\ &\quad\le \frac{\lambda_n^2}{n}\|\!|\widehat{\Sigma}_{SS}^{-1}\|\!|_2\|\!|Z_S^* - \widehat{Z}_S\|\!|_2(2\|\!|Z_S^*\|\!|_2 + \|\!|Z_S^* - \widehat{Z}_S\|\!|_2).\end{aligned} \tag{66}$$

Using the concentration results on random matrices in Appendix F, we have the bound $\|\!|\widehat{\Sigma}_{SS}^{-1}\|\!|_2 \le 2/C_{\min}$ with probability greater than $1 - \exp(-\Theta(n))$, and we have $\|\!|Z_S^*\|\!|_2 = \mathcal{O}(\sqrt{s})$ by definition. Moreover, from equation (57b) in Lemma 7, we have $\|\!|Z_S^* - \widehat{Z}_S\|\!|_2 \le \sqrt{s}\|Z_S^* - \widehat{Z}_S\|_{\ell_\infty/\ell_2}$. Using the bound (46) and Lemma 3, we have $\|Z_S^* - \widehat{Z}_S\|_{\ell_\infty/\ell_2} = o(1)$ with probability greater than $1 - c_1\exp(-c_0 K\log s)$, so that from equation (66), we conclude that

$$A_1 = \|\!|M_n^* - M_n\|\!|_2 = o\left(\frac{\lambda_n^2 s}{n}\right) \qquad \text{w.h.p.} \tag{67}$$

Turning to term A_2, we have the upper bound $A_2 \le T_1^\dagger + T_2^\dagger$, where

$$T_1^\dagger := \frac{\lambda_n^2}{n} |||Z_S^*|||_2^2 \left|\left|\left| \frac{n}{n-s-1}(\Sigma_{SS})^{-1} - (\widehat{\Sigma}_{SS})^{-1} \right|\right|\right|_2$$

and

$$T_2^\dagger := \frac{1}{n^2} |||W^T(I_n - \Pi_S)W - \sigma^2(n-s)I_K|||_2.$$

Since $|||Z_S^*|||_2^2 \le s$, and $|||\frac{n}{n-s-1}(\Sigma_{SS})^{-1} - (\widehat{\Sigma}_{SS})^{-1}|||_2 = o(1)$ with high probability (see Appendix F), we have $T_1^\dagger = o(\frac{\lambda_n^2 s}{n})$ with probability greater than $1 - 2\exp(-\Theta(n))$.

Turning to $T_2^\dagger$, we have with probability greater than $1 - 2\exp(-s/2) - \exp(-\Theta(n))$,

$$T_2^\dagger = \mathcal{O}\left(\frac{\sqrt{s}}{n\sqrt{n}}\right) = o\left(\frac{1}{n}\right),$$

using the random matrix bound (61) once again. Overall, we conclude that

$$A_2 = |||M_n^* - \mathbb{E}[M_n^*]|||_2 = o\left(\frac{\lambda_n^2 s + 1}{n}\right) \qquad \text{w.h.p.} \tag{68}$$

Finally, turning to $|||\mathbb{E}[M_n^*]|||_2$, from equation (64), we have

$$\begin{aligned} |||\mathbb{E}[M_n^*]|||_2 &= \frac{\lambda_n^2 \psi(B^*)}{n} \frac{n}{n-s-1} + \frac{\sigma^2}{n}\left(1 - \frac{s}{n}\right) \\ &= (1 + o(1))\left[\frac{\lambda_n^2 \psi(B^*) + \sigma^2}{n}\right]. \end{aligned} \tag{69}$$

Finally, we combine bounds (67), (68) and (69) in the decomposition (65), and apply Lemma 1(a) to obtain that $\psi(B^*) = \Theta(s)$; combining these facts yields that

$$(1-\delta)\left[\frac{\lambda_n^2 \psi(B^*) + \sigma^2}{n}\right] \le |||M_n|||_2 \le (1+\delta)\left[\frac{\lambda_n^2 \psi(B^*) + \sigma^2}{n}\right]$$

with probability greater than $1 - c_1 \exp(-c_0 K \log s)$, which establishes the claim.

APPENDIX I: LARGE DEVIATIONS FOR χ^2-VARIATES

LEMMA 11. *Let $Z_1, \dots, Z_m$ be i.i.d. χ^2-variates with d degrees of freedom. Then for all $t > d$, we have*

$$\mathbb{P}\left[\max_{i=1,\dots,m} Z_i \ge 2t\right] \le m \exp\left(-t\left[1 - 2\sqrt{\frac{d}{t}}\right]\right). \tag{70}$$

PROOF. Given a central χ^2-variate X with d degrees of freedom, Laurent and Massart (2000) prove that $\mathbb{P}[X - d \geq 2\sqrt{dx} + 2x] \leq \exp(-x)$, or equivalently

$$\mathbb{P}[X \geq x + (\sqrt{x} + \sqrt{d})^2] \leq \exp(-x),$$

valid for all $x > 0$. Setting $\sqrt{x} + \sqrt{d} = \sqrt{t}$, we have

$$\begin{aligned}
\mathbb{P}[X \geq 2t] &\overset{\text{(a)}}{\leq} \mathbb{P}[X \geq (\sqrt{t} - \sqrt{d})^2 + t] \leq \exp(-(\sqrt{t} - \sqrt{d})^2) \\
&\leq \exp(-t + 2\sqrt{td}) \\
&= \exp\left(-t\left[1 - 2\sqrt{\frac{d}{t}}\right]\right),
\end{aligned}$$

where inequality (a) follows since $\sqrt{t} \geq \sqrt{d}$ by assumption. Thus, the claim (70) follows by the union bound. □

REFERENCES

ANDERSON, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York. MR0771294

ARGYRIOU, A., EVGENIOU, T. and PONTIL, M. (2006). Multi-task feature learning. In *Advances in Neural Information Processing Systems* **19** 41–48. MIT Press, Cambridge, MA.

BACH, F. (2008). Consistency of the group Lasso and multiple kernel learning. *J. Mach. Learn. Res.* **9** 1179–1225. MR2417268

BACH, F., LANCKRIET, G. and JORDAN, M. I. (2004). Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the 21st International Conference in Machine Learning* 41–48. ACM, New York.

BERTSEKAS, D. P. (1995). *Nonlinear Programming*. Athena Scientific, Belmont, MA.

BICKEL, P., RITOV, Y. and TSYBAKOV, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge, UK. MR2061575

CHEN, S., DONOHO, D. L. and SAUNDERS, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20** 33–61. MR1639094

DAVIDSON, K. R. and SZAREK, S. J. (2001). Local operator theory, random matrices, and Banach spaces. In *Handbook of Banach Spaces* **1** 317–336. Elsevier, Amsterdam. MR1863696

DONOHO, D. and HUO, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory* **47** 2845–2862. MR1872845

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1361. MR1946581

FRANK, I. E. and FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 109–135.

HUANG, J., HOROWITZ, J. L. and MA, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36** 587–613. MR2396808

HUANG, J. and ZHANG, T. (2009). The benefit of group sparsity. Technical report, Rutgers University. Available at arXiv:0901.2962.

KNIGHT, K. and FU, W. J. (2000). Asymptotics for Lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

LAURENT, B. and MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** 1303–1338. MR1805785

LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, New York.

LIU, H. and ZHANG, J. (2008). On the $\ell_1 - \ell_q$ regularized regression. Technical report, Carnegie Mellon University. Available at arXiv:0802.1517v1.

LOUNICI, K., TSYBAKOV, A. B., PONTIL, M. and VAN DE GEER, S. A. (2009). Taking advantage of sparsity in multi-task learning. In *Proceedings of the 22nd Conference on Learning Theory*. Montreal.

MASSART, P. (2003). *Concentration Inequalities and Model Selection*. *Lecture Notes in Math.* **1896**. Springer, New York. MR2319879

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2008). The group Lasso for logistic regression. *J. Roy. Statist. Soc. Ser. B* **70** 53–71. MR2412631

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. MR2488351

NEGAHBAN, S. and WAINWRIGHT, M. (2008). Joint support recovery under high-dimensional scaling: Benefits and perils of ℓ_1/ℓ_∞ regularization. In *Advances in Neural Information Processing Systems* **21** 1161–1168. MIT Press, Cambridge, MA.

OBOZINSKI, G., TASKAR, B. and JORDAN, M. I. (2010). Joint covariate selection and joint subspace selection for multiple classification problems. *Statist. Comput.* **20** 231–252.

OSBORNE, M. R., PRESNELL, B. and TURLACH, B. A. (2000). A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.* **20** 389–403. MR1773265

RAVIKUMAR, P., LIU, H., LAFFERTY, J. and WASSERMAN, L. (2009). SpAM: Sparse additive models. *J. Roy. Statist. Soc. Ser. B* **71** 1009–1030.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TROPP, J. A. (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theory* **52** 1030–1051. MR2238069

TURLACH, B., VENABLES, W. and WRIGHT, S. (2005). Simultaneous variable selection. *Technometrics* **27** 349–363. MR2164706

WAINWRIGHT, M. J. (2009a). Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory* **55** 5728–5741. MR2597190

WAINWRIGHT, M. J. (2009b). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202.

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68** 49–67. MR2212574

ZHANG, H., LIU, H., WU, Y. and ZHU, J. (2008). Variable selection for the multi-category SVM via adaptive sup-norm regularization. *Electron. J. Statist.* **2** 1149–1167.

ZHAO, P., ROCHA, G. and YU, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.* **37** 3468–3497. MR2549566

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2567. MR2274449

G. OBOZINSKI
M. J. WAINWRIGHT
M. I. JORDAN
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
367 EVANS HALL
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: guillaume.obozinski@ens.fr
wainwrig@stat.berkeley.edu
jordan@stat.berkeley.edu