GROUP-LEVEL SUPPORT RECOVERY GUARANTEES FOR GROUP LASSO ESTIMATOR

Mojtaba Kadkhodaie Elyaderani, Swayambhoo Jain, Jeffrey Druce, Stefano Gonella, and Jarvis Haupt

University of Minnesota–Twin Cities, Minneapolis, MN 55455
{kadkh004, jainx174, druce001, sgonella, jdhaupt}@umn.edu

ABSTRACT

This paper considers the problem of estimating an unknown high dimensional signal from (typically low-dimensional) noisy linear measurements, where the desired unknown signal is assumed to possess a *group-sparse* structure, i.e. given a (pre-defined) partition of its entries into groups, only a small number of such groups are non-zero. Assuming the unknown group-sparse signal is generated according to a certain statistical model, we provide guarantees under which it can be efficiently estimated via solving the well-known group Lasso problem. In particular, we demonstrate that the set of indices for non-zero groups of the signal (called the group-level support of the signal) can be exactly recovered by solving the proposed group Lasso problem provided that its constituent non-zero groups are small in number and possess enough energy. Our guarantees rely on the well-conditioning of measurement matrix, which is expressed in terms of the block coherence parameter and can be efficiently computed. Our results are non-asymptotic in nature and therefore applicable to practical scenarios.

Index Terms— Group sparsity, structured support recovery, group Lasso, primal-dual witness.

1. INTRODUCTION

In recent years, the recovery of structured signals from a small number of linear measurements as compared to their ambient dimension has been a mainstay of research in the field of signal processing, high-dimensional statistics, and machine learning [1–4]. In this work we focus on the recovery of a group-sparse structured signal $\boldsymbol{\beta}^* \in \mathbb{R}^p$ observed according to the model

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

where $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ (with $n < p$) is the measurement matrix (also called the dictionary) and $\boldsymbol{w} \in \mathbb{R}^p$ is the noise vector. We assume that the signal vector $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is group-sparse structured with G groups defined as

$$\boldsymbol{\beta}^* = \left[(\boldsymbol{\beta}^*_{\mathcal{I}_1})^T(\boldsymbol{\beta}^*_{\mathcal{I}_2})^T \cdots (\boldsymbol{\beta}^*_{\mathcal{I}_G})^T\right]^T, \tag{2}$$

where $\boldsymbol{\beta}^*_{\mathcal{I}_g} \in \mathbb{R}^{d_g}$ and $\mathcal{I}_g$ denote the g^{th} group of $\boldsymbol{\beta}^*$ and its corresponding subset of indices, respectively. A group-sparse structured vector $\boldsymbol{\beta}^*$ with respect to (2) has only a few non-zero groups. Such structure naturally arises in many applications including structural health monitoring [5], bio-medical imaging [6], multi-task compressive sensing [7], multi-task learning [8], among many others. We focus on recovery guarantees of group-sparse signal using the following group Lasso based estimation problem

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \sum_{g=1}^{G} \lambda_g ||\boldsymbol{\beta}_{\mathcal{I}_g}||_2, \tag{3}$$

where $\lambda_g > 0$ is the regularization constant for the g^{th} group.

1.1. Related works

Several works have studied the recovery of group-sparse structured signals. The studies that provide statistical guarantees for the group Lasso problem (3) when the measurements are generated according to (1), are quite diverse in terms of their statistical signal generation assumptions and their requirements for successful recovery. In terms of the statistical model assumptions, a large body of work is focused on the case where the measurement matrix $\boldsymbol{X}$ is generated according to a Gaussian distribution [9, 10]. These results cannot be applied to many practical scenarios where the measurement matrix is not random but structured. Another line of work studies the asymptotic behavior of this recovery procedure when the number of measurements and unknown parameters can grow infinitely [11–13]. Since in many practical applications the dimensions remain finite, the utility of such asymptotic results may be limited. In terms of the studied requirements for successful recovery, various conditions are proposed so far: the group RIP condition of [14] and the restricted group eigenvalue condition of [3, 15] are among the most popular. Since verifying such conditions for structured measurement matrices can be computationally prohibitive, we do not base our analysis on those requirements and instead use the concept of block coherence which is always computable in polynomial time. Finally, we note that a recent effort [16] has analyzed group-sparse estimation methods using measurement matrices similar to what we consider here. However, the focus of that work is on providing regression error guarantees, instead of support recovery guarantees, which comprise our primary focus here.

1.2. Our Contributions

Our main contribution here is that we provide the conditions on the number of non-zero groups in $\boldsymbol{\beta}^*$ as well as the strength of corresponding coefficient vectors so that the group Lasso framework in (3) can successfully recover the groups under the measurement model (1). Our theoretical analysis is based on the primal-dual witness construction approach used in [17] and uses a generalization of that work to the case where there is a predefined grouping over the unknown coefficients.

The paper provides the proof of the main theorem. However, the useful lemmata are stated without proof; detailed proofs appear in a full-length manuscript, which is in preparation [18].

MKE, SJ, and JH are with the Dept. of Electrical and Computer Engineering. JD and SG are with the Dept. of Civil, Environmental, and Geo-Engineering. This work was supported by the DTI grant, NSF Award CCF-1217751, and the DARPA Young Faculty Award N66001-14-1-4047.

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

1.3. Organization

After quick overview of notation used throughout the paper, in Section 2 we present our main theoretical result on group-level support recovery guarantees along with the required assumptions. In Section 3, we provide a proof sketch of the main result. Finally, we provide a few brief conclusions in Section 4.

1.4. Notation

For any integer n, $[n] = \{1, 2, \cdots, n\}$. The maximum of two numbers a, b is denoted as $a \vee b = \max\{a, b\}$. Vectors and matrices are denoted by bold-face lowercase and uppercase, respectively. Given a vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|_p$ denotes its standard ℓ_p norm. For a given matrix $\boldsymbol{X}$: $\|\boldsymbol{X}\|_{2\to 2}$, $\|\boldsymbol{X}\|_F$, $\|\boldsymbol{X}\|_1$, and $\|\boldsymbol{X}\|_\infty$ denote the spectral norm, Frobenius norm, sum of absolute values of the entries, and the maximum absolute value of entries, respectively. For a column-wise block partitioned matrix $\boldsymbol{X} = [\boldsymbol{X}_{\mathcal{I}_1}\, \boldsymbol{X}_{\mathcal{I}_2} \cdots \boldsymbol{X}_{\mathcal{I}_G}]$ the inter-block coherence constant $\mu_B(\boldsymbol{X})$ is defined as

$$\mu_B(\boldsymbol{X}) := \max_{1 \le i \ne j \le G} \|\boldsymbol{X}_{\mathcal{I}_i}^H \boldsymbol{X}_{\mathcal{I}_j}\|_{2\to 2}, \tag{4}$$

and the intra-block coherence parameter $\mu_I(\boldsymbol{X})$ is defined as

$$\mu_I(\boldsymbol{X}) := \max_{g \in [G]} \|\boldsymbol{X}_{\mathcal{I}_g}^H \boldsymbol{X}_{\mathcal{I}_g} - \boldsymbol{I}_{d_g \times d_g}\|_{2\to 2}. \tag{5}$$

If p denotes the length of $\boldsymbol{\beta}$ and the number of columns of $\boldsymbol{X}$, then for the index set $\mathcal{I}_g \subset [p]$, $\boldsymbol{\beta}_{\mathcal{I}_g}$ will denote the group of entries of $\boldsymbol{\beta}$ whose indices belong to this set and $\boldsymbol{X}_{\mathcal{I}_g}$ will denote the columns of $\boldsymbol{X}$ indexed by $\mathcal{I}_g$. Throughout the paper, we will use different notions of support defined as in following:

- $\mathcal{G}(\boldsymbol{\beta}) := \{g \in [G] : \boldsymbol{\beta}_{\mathcal{I}_g} \ne \boldsymbol{0}\}$ will denote the set that contains the indices of the nonzero groups of $\boldsymbol{\beta}$, where G is the total number of groups.
- $\mathcal{S}_\mathcal{G}(\boldsymbol{\beta}) := \cup_{g \in \mathcal{G}(\boldsymbol{\beta})} \mathcal{I}_g$. In words, $\mathcal{S}_\mathcal{G}(\boldsymbol{\beta})$ will denote the set that contains all indices comprising groups that are nonzero.

Moreover, we let $d_{\min} := \min_{g \in [G]} d_g$ and $d_{\max} := \max_{g \in [G]} d_g$ denote the minimum and maximum group sizes, respectively, and $d_{\mathcal{G}(\boldsymbol{\beta})} := \sum_{g \in \mathcal{G}(\boldsymbol{\beta})} d_g$ be the total number of entries in the group-level support $\mathcal{G}(\boldsymbol{\beta})$ of $\boldsymbol{\beta}$. In order to not overly complicate the notation, we will always use $\mathcal{G}^*$, $\mathcal{S}_\mathcal{G}^*$, and $d_\mathcal{G}^*$ as abbreviations for $\mathcal{G}(\boldsymbol{\beta}^*)$, $\mathcal{S}_\mathcal{G}(\boldsymbol{\beta}^*)$, and $d_{\mathcal{G}(\boldsymbol{\beta}^*)}$, respectively.

2. RECOVERY GUARANTEES

The recovery guarantees presented in this paper are under specific statistical data model assumptions. We assume that the group-sparse vector $\boldsymbol{\beta}^* \in \mathbb{R}^p$ with groups as defined in (2) is randomly generated as described below:

1. The group-level support of $\boldsymbol{\beta}^*$, which we denote by $\mathcal{G}^* \subseteq [G]$, comprises s non-zero blocks whose indices are selected uniformly at random from all subsets of $[G]$ of size s.
2. The non-zero entries of $\boldsymbol{\beta}^*$ are equally likely to be positive or negative: $\mathbb{E}\,\mathrm{sign}(\boldsymbol{\beta}_j^*) = 0$ for $j \in [p]$.
3. The non-zero blocks of $\boldsymbol{\beta}^*$ have statistically independent "directions." Specifically, it is assumed that

$$\Pr\left\{ \bigcap_{g \in \mathcal{G}^*} \frac{\boldsymbol{\beta}_{\mathcal{I}_g}^*}{\|\boldsymbol{\beta}_{\mathcal{I}_g}^*\|_2} \in \mathcal{A}_g \right\} = \prod_{g \in \mathcal{G}^*} \Pr\left\{ \frac{\boldsymbol{\beta}_{\mathcal{I}_g}^*}{\|\boldsymbol{\beta}_{\mathcal{I}_g}^*\|_2} \in \mathcal{A}_g \right\},$$

where each $\mathcal{A}_g$ is any subset of the unit sphere in $\mathbb{R}^{d_g}$.

Under the above statistical model assumptions the main theoretical result of the paper can be stated as follows.

Theorem 2.1. *Given the noisy linear measurement model* (1) *where* $\boldsymbol{\beta}^*$ *is generated according to the statistical assumptions listed above and* $\boldsymbol{w} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_{n\times n})$ *assume that*

$$A_1 : \; \mu_I(\boldsymbol{X}) \le c_0 \textit{ and } \mu_B(\boldsymbol{X}) \le \sqrt{\frac{d_{\min}}{d_{\max}^2}\,\frac{c_1}{\log p}},$$

$$A_2 : \; |\mathcal{G}^*| \le \min\left\{ \frac{c_2\, G}{\|\boldsymbol{X}\|_{2\to 2}^2 \log p}, \frac{d_{\min}}{d_{\max}^2}\,\frac{c_3\, \mu_B^{-2}(\boldsymbol{X})}{\log p} \right\},$$

$$A_3 : \; \|\boldsymbol{\beta}_{\mathcal{I}_g}^*\|_2 \ge 10\sigma(1+\epsilon)\left(\sqrt{d_{\mathcal{G}^*}} + \sqrt{d_g}\right)\left(1 \vee \sqrt{\frac{s}{d_{\max}\, \log p}}\right)$$
for every $g \in \mathcal{G}^*$,

all hold for some non-negative constants c_0, $c_1 \le 0.004$, $c_2 \le \frac{1}{14}(\frac{1}{4} - 3\,c_0 - 48\,c_1)$, $c_3 = \min\{c_2, 0.0004\}$, *and some*

$$\epsilon \ge \sqrt{\frac{(1 + \mu_I(\boldsymbol{X}))\, \log(p\, G)}{d_{\min}}}. \tag{6}$$

Then, with probability at least $1 - 14\, p^{-2\log 2}$, *the solution* $\widehat{\boldsymbol{\beta}}$ *of the problem* (3), *with* $\lambda_g = 4\sigma(1+\epsilon)\sqrt{d_g}$ *for every* $g \in [G]$, *is unique with the same group-level support as* $\boldsymbol{\beta}^*$, *i.e.* $\mathcal{G}(\boldsymbol{\beta}^*) = \mathcal{G}(\widehat{\boldsymbol{\beta}})$, *and satisfies the error bound*

$$\left\| \widehat{\boldsymbol{\beta}}_{\mathcal{I}_g} - \boldsymbol{\beta}_{\mathcal{I}_g}^* \right\|_2 \le 5\sigma(1+\epsilon)\left(\sqrt{d_g} + \sqrt{d_{\mathcal{G}^*}}\right), \tag{7}$$

for every $g \in \mathcal{G}(\boldsymbol{\beta}^*)$.

Here, we note on several aspects of the main theorem, which turn it into a practically appealing result. First, notice that the support recovery guarantee relies on the well-conditioning of the dictionary $\boldsymbol{X}$ as required by assumption A1. We measure the well-conditioning of the dictionary in terms of its block coherence constant $\mu_B(\boldsymbol{X})$. Fortunately, $\mu_B(\boldsymbol{X})$ can be computed in polynomial time for a given column-wise partitioned dictionary (unlike other quantities such as restricted isometry constant, which are widely used in proving similar recovery guarantees but can be NP-hard to compute [19]).

The second condition A_2 specifies the requirement on the maximum number of allowable non-zero groups in the group-level support of $\boldsymbol{\beta}^*$ that can be recovered. Unlike some earlier coherence-based recovery results for group-sparse structured signals [20], which verify pessimistic bounds on the number of recoverable groups, the condition provided here is stronger in the sense that it allows for a linear scaling between the number of non-zero groups $|\mathcal{G}^*|$ and the total number of measurements n. Moreover the block coherence parameter appears in the upper-bound in the form of $\mu_B^{-2}(\boldsymbol{X})$, which is a significant improvement over similar results, e.g. in [20], that require $|\mathcal{G}^*|$ be bounded by functions of $\mu_B^{-1}(\boldsymbol{X})$. We bring an example here to make this argument more clear. Assume the dictionary $\boldsymbol{X}$ is the concatenation of two orthonormal bases, i.e. $\boldsymbol{X} := \left[\boldsymbol{X}_{(1)} \middle| \boldsymbol{X}_{(2)}\right] \in \mathbb{R}^{n\times 2n}$, where $\boldsymbol{X}_{(1)} \in \mathbb{R}^{n\times n}$ is the discrete cosine transform (DCT) matrix and $\boldsymbol{X}_{(2)} \in \mathbb{R}^{n\times n}$ is the identity matrix. The authors leveraged this widely-studied dictionary in the context of structural anomaly detection using propagating wave-field measurements [21], where $\boldsymbol{X}_{(2)}$ was column-wise partitioned into groups of size $d_g = d$ and $\boldsymbol{X}_{(1)}$ was divided into singleton groups of size $d_g = 1$. For such $\boldsymbol{X}$ with the specified partition, it can be shown that $\mu_B(\boldsymbol{X}) \le \sqrt{4d/n}$, $\mu_I(\boldsymbol{X}) = 0$, $\|\boldsymbol{X}\|_{2\to 2}^2 = 2$, and $G = n\,(1 + 1/d)$. Substituting these in A_2, it can be shown [18] that for $|\mathcal{G}^*| \le c\,(n/(d^3 \cdot \log n))$, successful support recovery can be guaranteed, where $c > 0$ is a universal constant. However, if the bound on $|\mathcal{G}^*|$ were in terms of $\mu_B^{-1}(\boldsymbol{X})$, then successful recovery would have been possible only for $|\mathcal{G}^*| = \mathcal{O}(\sqrt{n})$.

The third assumption A_3 is on the energy of the non-zero groups, which requires their Euclidean norms to be above a certain threshold depending on the noise variance σ.

3. PROOF SKETCH

Our analysis uses a basic result for characterizing the optimal solutions of the group Lasso problem (3). We state the result as a lemma; its proof follows what are, by now, fairly standard methods in convex analysis so we omit it here [12, 22].

Lemma 3.1. *A vector $\widehat{\boldsymbol{\beta}}$ solves problem* (3) *if and only if*

$$\boldsymbol{X}_{\mathcal{I}_g}^T \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) - \boldsymbol{X}_{\mathcal{I}_g}^T \boldsymbol{w} + \lambda_g \widehat{\boldsymbol{z}}_{\mathcal{I}_g} = \boldsymbol{0}, \;\; \forall\, g \in [G] \tag{8}$$

holds for some vector $\widehat{\boldsymbol{z}}$, whose elements satisfy

$$\begin{array}{rl} \widehat{\boldsymbol{z}}_{\mathcal{I}_g} = \frac{\widehat{\boldsymbol{\beta}}_{\mathcal{I}_g}}{\|\widehat{\boldsymbol{\beta}}_{\mathcal{I}_g}\|_2}, & \textit{if } \widehat{\boldsymbol{\beta}}_{\mathcal{I}_g} \neq \boldsymbol{0} \\ \|\widehat{\boldsymbol{z}}_{\mathcal{I}_g}\|_2 \leq 1, & \textit{otherwise} \end{array} . \tag{9}$$

If $\|\widehat{\boldsymbol{z}}_{\mathcal{I}_g}\|_2 < 1$ for all $g \notin \mathcal{G}(\widehat{\boldsymbol{\beta}})$ then any optimal solution $\widehat{\boldsymbol{\beta}}$ to (3) *satisfies $\widehat{\boldsymbol{\beta}}_{\mathcal{I}_g} = \boldsymbol{0}$ for all $g \notin \mathcal{G}(\widehat{\boldsymbol{\beta}})$; if in addition, the matrix $\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}(\widehat{\boldsymbol{\beta}})}^T \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}(\widehat{\boldsymbol{\beta}})}$ is invertible, then $\widehat{\boldsymbol{\beta}}$ is the unique solution to* (3).

The optimality condition (8) can be written in matrix form, as

$$\boldsymbol{X}^T \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) - \boldsymbol{X}^T \boldsymbol{w} + \boldsymbol{\Lambda}\, \widehat{\boldsymbol{z}} = \boldsymbol{0}, \tag{10}$$

where $\boldsymbol{\Lambda}$ is the $p \times p$ diagonal matrix whose j-th diagonal entry is $\Lambda_{j,j} = \lambda_{g(j)}$, where $g(j) = \{g \in [G] : j \in \mathcal{I}_g\}$.

Our proof follows the so-called *Primal-Dual Witness* (PDW) technique utilized in [17] for the analysis of the Lasso problem and also in [9, 12] for the analysis of the group Lasso problem arising in the context of multivariate regression with Gaussian-distributed $\boldsymbol{X}$.

To construct the primal-dual certificate pair $(\check{\boldsymbol{\beta}}, \check{\boldsymbol{z}})$, we first identify the solution of a *restricted* group Lasso problem over the true group-level support $\mathcal{G}^*$. Specifically, we construct $\check{\boldsymbol{\beta}}_{\mathcal{S}_{\mathcal{G}}^*} \in \mathbb{R}^{d_{\mathcal{G}^*}}$ as

$$\check{\boldsymbol{\beta}}_{\mathcal{S}_{\mathcal{G}}^*} = \arg \min_{\boldsymbol{\beta}_{\mathcal{S}_{\mathcal{G}}^*} \in \mathbb{R}^{d_{\mathcal{G}^*}}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*} \boldsymbol{\beta}_{\mathcal{S}_{\mathcal{G}}^*}\|_2^2 + \sum_{g \in \mathcal{G}^*} \lambda_g \|\boldsymbol{\beta}_{\mathcal{I}_g}\|_2. \tag{11}$$

Second, we choose the restricted dual vector $\check{\boldsymbol{z}}_{\mathcal{S}_{\mathcal{G}}^*} \in \mathbb{R}^{d_{\mathcal{G}^*}}$ such that the primal-dual pair $(\check{\boldsymbol{\beta}}_{\mathcal{S}_{\mathcal{G}}^*}, \check{\boldsymbol{z}}_{\mathcal{S}_{\mathcal{G}}^*})$ satisfies the following optimality condition of the restricted problem:

$$\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*} (\check{\boldsymbol{\beta}}_{\mathcal{S}_{\mathcal{G}}^*} - \boldsymbol{\beta}_{\mathcal{S}_{\mathcal{G}}^*}^*) - \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T \boldsymbol{w} + \boldsymbol{\Lambda}_{\mathcal{S}_{\mathcal{G}}^*} \check{\boldsymbol{z}}_{\mathcal{S}_{\mathcal{G}}^*} = \boldsymbol{0}, \tag{12}$$

where $\boldsymbol{\Lambda}_{\mathcal{S}_{\mathcal{G}}^*}$ denotes the sub-matrix of $\boldsymbol{\Lambda}$ obtained by sampling rows and columns at the locations in $\mathcal{S}_{\mathcal{G}}^*$; and also $\check{\boldsymbol{z}}_{\mathcal{S}_{\mathcal{G}}^*}$ satisfies the subgradient condition (9). Using Lemma 3.2 we can argue that under our statistical data model assumptions, in addition to the assumptions A_1 and A_2, the matrix $\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}$ is full column-rank, with high probability, which implies that $\check{\boldsymbol{\beta}}_{\mathcal{S}_{\mathcal{G}}^*}$ is the unique solution to (12) given by

$$\boldsymbol{\beta}_{\mathcal{S}_{\mathcal{G}}^*}^* - \check{\boldsymbol{\beta}}_{\mathcal{S}_{\mathcal{G}}^*} = (\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*})^{-1} (\boldsymbol{\Lambda}_{\mathcal{S}_{\mathcal{G}}^*} \check{\boldsymbol{z}}_{\mathcal{S}_{\mathcal{G}}^*} - \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T \boldsymbol{w}). \tag{13}$$

Next, we set the "off group-level support" primal variable $\check{\boldsymbol{\beta}}_{(\mathcal{S}_{\mathcal{G}}^*)^c} \in \mathbb{R}^{n - d_{\mathcal{G}}^*}$ to be zero and solve for an off group-level support dual variable $\check{\boldsymbol{z}}_{(\mathcal{S}_{\mathcal{G}}^*)^c} \in \mathbb{R}^{n - d_{\mathcal{G}}^*}$ such that the optimality conditions for the full (unrestricted) group Lasso problem are satisfied. Using the result of Lemma 3.1, this translates to the following condition:

$$\boldsymbol{X}_{(\mathcal{S}_{\mathcal{G}}^*)^c}^T \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*} (\check{\boldsymbol{\beta}}_{\mathcal{S}_{\mathcal{G}}^*} - \boldsymbol{\beta}_{\mathcal{S}_{\mathcal{G}}^*}^*) - \boldsymbol{X}_{(\mathcal{S}_{\mathcal{G}}^*)^c}^T \boldsymbol{w} + \boldsymbol{\Lambda}_{(\mathcal{S}_{\mathcal{G}}^*)^c} \check{\boldsymbol{z}}_{(\mathcal{S}_{\mathcal{G}}^*)^c} = \boldsymbol{0}, \tag{14}$$

where $\boldsymbol{\Lambda}_{(\mathcal{S}_{\mathcal{G}}^*)^c}$ denotes the sub-matrix of $\boldsymbol{\Lambda}$ obtained by sampling rows and columns at the locations in $(\mathcal{S}_{\mathcal{G}}^*)^c$. Using (14) along with (13), it is straightforward to show that for each $g \notin \mathcal{G}^*$, the corresponding block $\check{\boldsymbol{z}}_{\mathcal{I}_g}$ of the dual vector can be expressed as

$$\begin{aligned} \check{\boldsymbol{z}}_{\mathcal{I}_g} &= \frac{1}{\lambda_g} \boldsymbol{X}_{\mathcal{I}_g}^T \left[\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*} (\boldsymbol{\beta}_{\mathcal{S}_{\mathcal{G}}^*}^* - \check{\boldsymbol{\beta}}_{\mathcal{S}_{\mathcal{G}}^*}) + \boldsymbol{w} \right] \\ &= \frac{1}{\lambda_g} \boldsymbol{X}_{\mathcal{I}_g}^T \left[\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*} (\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*})^{-1} \boldsymbol{\Lambda}_{\mathcal{S}_{\mathcal{G}}^*} \check{\boldsymbol{z}}_{\mathcal{S}_{\mathcal{G}}^*} + \Pi_{(\mathcal{S}_{\mathcal{G}}^*)^\perp}(\boldsymbol{w}) \right], \end{aligned}$$

where $\Pi_{(\mathcal{S}_{\mathcal{G}}^*)^\perp}(\boldsymbol{w}) := (\boldsymbol{I} - \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*} (\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*})^{-1} \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T) \boldsymbol{w}$.

In order to ensure the off group-level support primal vector $\check{\boldsymbol{\beta}}_{(\mathcal{S}_{\mathcal{G}}^*)^c}$ is zero, we impose that $\|\check{\boldsymbol{z}}_{\mathcal{I}_g}\|_2 < 1$ for every $g \notin \mathcal{G}^*$, which is known as the "strict dual feasibility" condition [17]. As a result of this condition, no "spurious" nonzero groups will be present in the support of $\check{\boldsymbol{\beta}}$. Notice that by the triangle inequality we will have that for any $g \notin \mathcal{G}^*$

$$\begin{aligned} \|\check{\boldsymbol{z}}_{\mathcal{I}_g}\|_2 \leq & \left\| \frac{1}{\lambda_g} \boldsymbol{X}_{\mathcal{I}_g}^T \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*} (\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*})^{-1} \boldsymbol{\Lambda}_{\mathcal{S}_{\mathcal{G}}^*} \check{\boldsymbol{z}}_{\mathcal{S}_{\mathcal{G}}^*} \right\|_2 \\ & + \left\| \frac{1}{\lambda_g} \boldsymbol{X}_{\mathcal{I}_g}^T \Pi_{(\mathcal{S}_{\mathcal{G}}^*)^\perp}(\boldsymbol{w}) \right\|_2 . \end{aligned} \tag{15}$$

Next, we bound the first term on the right-hand side using our statistical assumptions, i.e. that the "directions" $\boldsymbol{\beta}_{\mathcal{I}_g}^* / \|\boldsymbol{\beta}_{\mathcal{I}_g}^*\|_2$ associated with each nonzero block of $\boldsymbol{\beta}^*$ are random, and statistically independent. To do this, we need to express the elements of the vector $\check{\boldsymbol{z}}_{\mathcal{S}_{\mathcal{G}}^*}$ (or more specifically, its individual blocks) in terms of the direction vectors associated with the corresponding nonzero blocks of the true vector $\boldsymbol{\beta}_{\mathcal{S}_{\mathcal{G}}^*}^*$. In fact, it can be shown that we can write

$$\check{\boldsymbol{z}}_{\mathcal{S}_{\mathcal{G}}^*} = \widetilde{\boldsymbol{\beta}}_{\mathcal{S}_{\mathcal{G}}^*}^* + \boldsymbol{u}_{\mathcal{S}_{\mathcal{G}}^*},$$

where $\widetilde{\boldsymbol{\beta}}_{\mathcal{S}_{\mathcal{G}}^*}^*$ is obtained by concatenating the direction vectors $\boldsymbol{\beta}_{\mathcal{I}_g}^* / \|\boldsymbol{\beta}_{\mathcal{I}_g}^*\|_2$ for all $g \in \mathcal{G}^*$ and $\boldsymbol{u}_{\mathcal{S}_{\mathcal{G}}^*}$ is a perturbation vector, whose norm can be controlled using a similar argument as in the proof of Lemma 3 in [9]. Equipped with this decomposition of $\check{\boldsymbol{z}}_{\mathcal{S}_{\mathcal{G}}^*}$ and applying triangle inequality in (15) for each $g \notin \mathcal{G}^*$ we have

$$\begin{aligned} \|\check{\boldsymbol{z}}_{\mathcal{I}_g}\|_2 \;\leq\; & \frac{1}{\lambda_g} \left\| \boldsymbol{X}_{\mathcal{I}_g}^T \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*} (\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*})^{-1} \boldsymbol{\Lambda}_{\mathcal{S}_{\mathcal{G}}^*} \widetilde{\boldsymbol{\beta}}_{\mathcal{S}_{\mathcal{G}}^*}^* \right\|_2 \\ + \; & \frac{1}{\lambda_g} \left\| \boldsymbol{X}_{\mathcal{I}_g}^T \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*} (\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*})^{-1} \boldsymbol{\Lambda}_{\mathcal{S}_{\mathcal{G}}^*} \boldsymbol{u}_{\mathcal{S}_{\mathcal{G}}^*} \right\|_2 \\ + \; & \frac{1}{\lambda_g} \left\| \boldsymbol{X}_{\mathcal{I}_g}^T \Pi_{(\mathcal{S}_{\mathcal{G}}^*)^\perp}(\boldsymbol{w}) \right\|_2 . \end{aligned} \tag{16}$$

Using the Lemmata 3.3, 3.4, and 3.5 on the following page, we can argue that, with high probability, the right-hand side of the above inequality (16) is strictly less than 1 as long as the conditions of the main Theorem 2.1 are met. This will ensure $\mathcal{G}(\check{\boldsymbol{\beta}}) \subseteq \mathcal{G}^*$.

Further, if the following condition holds true

$$\|\boldsymbol{\beta}_{\mathcal{I}_g}^* - \check{\boldsymbol{\beta}}_{\mathcal{I}_g}\|_2 < \|\boldsymbol{\beta}_{\mathcal{I}_g}^*\|_2 \;\; \text{for all } g \in \mathcal{G}^*, \tag{17}$$

then it follows, (essentially, by the triangle inequality) that $\check{\boldsymbol{\beta}}_{\mathcal{I}_g} \neq \mathbf{0}$, whenever $\boldsymbol{\beta}^*_{\mathcal{I}_g} \neq \mathbf{0}$, which is equivalent to $\mathcal{G}^* \subseteq \mathcal{G}(\check{\boldsymbol{\beta}})$. Using assumption A_3 in the main theorem, and in addition assuming $\|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}} - \boldsymbol{I}_{d^*_{\mathcal{G}} \times d^*_{\mathcal{G}}}\|_{2\to 2} \leq \frac{1}{2}$ which in turn implies (along with (13)) that for each $g \in \mathcal{G}^*$ we have

$$\|\check{\boldsymbol{\beta}}_{\mathcal{I}_g} - \boldsymbol{\beta}^*_{\mathcal{I}_g}\|_2 \leq \|\boldsymbol{X}^T_{\mathcal{I}_g}\boldsymbol{w}\|_2 + \lambda_g + \|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{w}\|_2 + \|\boldsymbol{\lambda}_{\mathcal{G}^*}\|_2, \quad (18)$$

guarantees (17) is true.

Therefore, having established $\|\check{\boldsymbol{z}}_{\mathcal{I}_g}\|_2 < 1$ for every $g \notin \mathcal{G}^*$ *in addition to* a guarantee of the form (17) will ensure that $\mathcal{G}(\check{\boldsymbol{\beta}}) = \mathcal{G}^*$. Finally, by using the Hanson-Wright inequality [23] together with the choice $\lambda_g = 4\sigma(1+\epsilon)\sqrt{d_g}$ for every $g \in [G]$, we can prove the estimation error bound (7) stated by the Theorem.

So far we have shown recovery gaurantees under a series of assumptions we argued hold with high probability. These assumptions are listed as events E_1 to E_4 given below:

$$E_1 := \left\{\|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}} - \boldsymbol{I}_{d^*_{\mathcal{G}} \times d^*_{\mathcal{G}}}\|_{2\to 2} \leq \frac{1}{2}\right\}$$

$$E_2 := \left\{\left\|\boldsymbol{X}^T_{\mathcal{I}_g}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\widetilde{\boldsymbol{\beta}}^*_{\mathcal{S}^*_{\mathcal{G}}}\right\|_2 \leq \frac{\lambda_g}{4}, \ \forall g \notin \mathcal{G}^*\right\}$$

$$E_3 := \left\{\left\|\boldsymbol{X}^T_{\mathcal{I}_g}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{u}_{\mathcal{S}^*_{\mathcal{G}}}\right\|_2 \leq \frac{\lambda_g}{4}, \ \forall g \notin \mathcal{G}^*\right\}$$

$$E_4 := \left\{\|\boldsymbol{X}^T_{\mathcal{I}_g}\Pi_{(\mathcal{S}^*_{\mathcal{G}})^\perp}(\boldsymbol{w})\|_2 \leq \frac{\lambda_g}{4}, \ \forall g \notin \mathcal{G}^*\right\}$$

Finally, let E denote the event that the group-level support G^* is exactly recovered and the estimation error bound holds. The event E happens when the event E_1 is true and conditioned on that, events E_2, E_3, E_4 also occur. Based on this argument and by using the union bound, the probability of E^c can be upper bounded as

$$\Pr(E^c) \leq \Pr(E_1^c) + \Pr\left(E_2^c|E_1\right) + \Pr\left(E_3^c|E_1\right) + \Pr\left(E_4^c|E_1\right).$$

The rest of the proof briefly reviews conditions under which the probability terms on the right-hand side of the above inequality are bounded. First, by Lemma 3.2 we know that whenever $\mu_I(\boldsymbol{X}) \leq c_0$, $\mu_B(\boldsymbol{X}) \leq \frac{c_1}{\log p}$, and (19) hold, then $\Pr(E_1^c) \leq 2p^{-4\log 2}$. Second, utilizing Lemma 3.3 implies $\Pr\left(E_2^c|E_1\right) \leq 4p^{-4\log 2}$ under the stated conditions of this Lemma. Third, Lemma 3.4, with $\lambda_g = 4\sigma(1+\epsilon)\sqrt{d_g}$, implies that as long as ϵ meets the condition in (6) and $\|\boldsymbol{\beta}^*_{\mathcal{I}_g}\|_2$ satisfies A_3 for every $g \in \mathcal{G}^*$, then $\Pr\left(E_4^c|E_1\right) \leq 6\,p^{-2\log 2}$. Finally, by Lemma 3.5, we have that $\Pr\left(E_5^c|E_1\right) \leq 2p^{-4\log 2}$ whenever $\lambda_g = 4\sigma(1+\epsilon)\sqrt{d_g}$ for all $g \notin \mathcal{G}^*$. Therefore, under stated conditions of the theorem we have

$$\Pr(E^c) \leq 8p^{-4\log 2} + 6p^{-2\log 2} \leq 14\,p^{-2\log 2}.$$

3.1. Useful Lemmata

Throughout the proof, we were proceeding under the assumption that $\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}$ is invertible. This condition is established here:

Lemma 3.2 (Theorem 1 of [16])**.** *Suppose the dictionary $\boldsymbol{X}$ satisfies $\mu_I(\boldsymbol{X}) \leq c_0$ and $\mu_B(\boldsymbol{X}) \leq c_1/\log p$, for some universal positive constants c_0 and c_1. Assume further that $\mathcal{G}^*$ is a subset of size $|\mathcal{G}^*|$ of the set $[G] = \{1, 2, \cdots, G\}$, which is drawn uniformly at random. Then, as long as*

$$s \leq \min\left\{\frac{c_2\,G}{\|\boldsymbol{X}\|^2_{2\to 2}\log p}, \frac{c_3}{\mu_B^2(\boldsymbol{X})\log p}\right\} \quad (19)$$

for positive constants c_2 and c_3 that only depend on c_0 and c_1, it holds that $\left\|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}} - \boldsymbol{I}_{d^*_{\mathcal{G}} \times d^*_{\mathcal{G}}}\right\|_{2\to 2} \leq \frac{1}{2}$, with probability at least $1 - 2p^{-4\log 2}$ with respect to the random choice of $\mathcal{G}^*$.

The above lemma is essentially identical to Theorem 1 of [16], with the difference that in (19) we used $\mu_B(\boldsymbol{X})$ instead of $\overline{\mu}_B(\boldsymbol{X})$, where the latter is called quadratic-mean block coherence in [16]. This yields a slightly more restrictive condition as $\mu_B(\boldsymbol{X}) \geq \overline{\mu}_B(\boldsymbol{X})$.

The Lemmata 3.3, 3.4, and 3.5 establish the boundedness of each of the three terms in (16), with high probability, as long as the conditions of the main Theorem 2.1 are met.

Lemma 3.3. *Suppose the group-level support $\mathcal{G}^*$ is given such that the event E_1 holds true. Moreover, let $\mu_B(\boldsymbol{X}) \leq \sqrt{\frac{d_{\min}}{d^2_{\max}}} \cdot \frac{c_1}{\log p}$, and $s \leq \frac{d_{\min}}{d^2_{\max}} \frac{c_3}{\mu_B^2(\boldsymbol{X})\cdot\log p}$. Then assuming $\boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}}$ is a random vector generated according to our statistical model, we will have that*

$$\Pr\left(\bigcup_{g \notin \mathcal{G}^*}\left\|\boldsymbol{X}^T_{\mathcal{I}_g}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\,\widetilde{\boldsymbol{\beta}}^*_{\mathcal{S}^*_{\mathcal{G}}}\right\|_2 > \frac{\lambda_g}{4}\right) \leq \eta$$

holds for $\eta = 4p^{-4\ \log 2}$.

The above lemma is a consequence of our statistical model assumptions together with Hoeffding's inequality; whereas the proofs of the following two lemmata rely on the Hanson-Wright inequality.

Lemma 3.4. *Suppose the group-level support $\mathcal{G}^*$ is given such that the event E_1 holds. Moreover, assume $\mu_B(\boldsymbol{X}) \leq \sqrt{\frac{d_{\min}}{d^2_{\max}}} \cdot \frac{c_1}{\log p}$, $s \leq \frac{d_{\min}}{d^2_{\max}} \frac{c_3}{\mu_B^2(\boldsymbol{X})\cdot\log p}$, $\boldsymbol{w} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I}_{n\times n})$, and that the condition A_3 of the main theorem is met, then*

$$\Pr\left(\bigcup_{g \notin \mathcal{G}^*}\left\|\boldsymbol{X}^T_{\mathcal{I}_g}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{u}_{\mathcal{S}^*_{\mathcal{G}}}\right\|_2 > \frac{\lambda_g}{4}\right) \leq \eta$$

holds for $\eta = 6\,p^{-2\log 2}$.

Lemma 3.5. *Suppose the group-level support $\mathcal{G}^*$ is given such that event E_1 holds true. Moreover, let $\boldsymbol{w} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I}_{n\times n})$. Then the following holds for ϵ satisfying* (6) *and $\lambda_g = 4\sigma(1+\epsilon)\sqrt{d_g}$,*

$$\Pr\left(\bigcup_{g \notin \mathcal{G}^*}\left\|\boldsymbol{X}^T_{\mathcal{I}_g}\Pi_{(\mathcal{S}^*_{\mathcal{G}})^\perp}(\boldsymbol{w})\right\|_2 > \frac{\lambda_g}{4}\right) \leq 2\,p^{-4\log 2}.$$

4. CONCLUSION

In this paper, we consider the recovery of group-sparse signals from low-dimensional noisy linear measurements using the group Lasso estimation procedure. We establish practically appealing group-level support recovery guarantees for non-asymptotic regimes in terms of the efficiently computable block coherence parameter.

5. REFERENCES

[1] David L Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[2] Emmanuel J Candès and Benjamin Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[3] Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar, "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers," in *Advances in Neural Information Processing Systems*, 2009, pp. 1348–1356.

[4] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational mathematics*, vol. 12, no. 6, pp. 805–849, 2012.

[5] Mojtaba Kadkhodaie, Swayambhoo Jain, Jarvis Haupt, Jeff Druce, and Stefano Gonella, "Locating rare and weak material anomalies by convex demixing of propagating wavefields," in *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2015, pp. 373–376.

[6] Shaoting Zhang, Junzhou Huang, Yuchi Huang, Yang Yu, Hongsheng Li, and Dimitris N Metaxas, "Automatic image annotation using group sparsity," in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 3312–3319.

[7] Shihao Ji, David Dunson, and Lawrence Carin, "Multitask compressive sensing," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 92–106, 2009.

[8] Abhishek Kumar and Hal Daume III, "Learning task grouping and overlap in multi-task learning," *arXiv preprint arXiv:1206.6417*, 2012.

[9] Guillaume Obozinski, Martin J Wainwright, and Michael I Jordan, "Support union recovery in high-dimensional multivariate regression," *The Annals of Statistics*, pp. 1–47, 2011.

[10] Nikhil S Rao, Ben Recht, and Robert D Nowak, "Universal measurement bounds for structured sparse signal recovery," in *International Conference on Artificial Intelligence and Statistics*, 2012, pp. 942–950.

[11] Han Liu and Jian Zhang, "Estimation consistency of the group lasso and its applications," in *International Conference on Artificial Intelligence and Statistics*, 2009, pp. 376–383.

[12] Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman, "Sparse additive models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 5, pp. 1009–1030, 2009.

[13] Yuval Nardi and Alessandro Rinaldo, "On the asymptotic properties of the group lasso estimator for linear models," *Electronic Journal of Statistics*, vol. 2, pp. 605–633, 2008.

[14] Junzhou Huang and Tong Zhang, "The benefit of group sparsity," *The Annals of Statistics*, vol. 38, no. 4, pp. 1978–2004, 2010.

[15] Karim Lounici, Massimiliano Pontil, Sara Van De Geer, and Alexandre B Tsybakov, "Oracle inequalities and optimal inference under group sparsity," *The Annals of Statistics*, pp. 2164–2204, 2011.

[16] Waheed U Bajwa, Marco F Duarte, and Robert Calderbank, "Conditioning of random block subdictionaries with applications to block-sparse recovery and regression," *IEEE Transactions on Information Theory*, vol. 61, no. 7, pp. 4060–4079, 2015.

[17] Martin J Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso)," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[18] Mojtaba Kadkhodaie Elyaderani, Swayambhoo Jain, Jeff Druce, Stefano Gonella, and Jarvis D. Haupt, "Support recovery guarantees for group lasso estimator with applications to structural health monitoring," *In Preparation*.

[19] Afonso S Bandeira, Edgar Dobriban, Dustin G Mixon, and William F Sawin, "Certifying the restricted isometry property is hard," *arXiv preprint arXiv:1204.1580*, 2012.

[20] Yonina C Eldar, Patrick Kuppinger, and Helmut Bölcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, 2010.

[21] Mojtaba Kadkhodaie, Swayambhoo Jain, Jarvis Haupt, Jeff Druce, and Stefano Gonella, "Locating rare and weak material anomalies by convex demixing of propagating wavefields," in *Proceedings of IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, 2015, to appear.

[22] Francis R Bach, "Consistency of the group lasso and multiple kernel learning," *The Journal of Machine Learning Research*, vol. 9, pp. 1179–1225, 2008.

[23] Mark Rudelson and Roman Vershynin, "Hanson-wright inequality and sub-gaussian concentration," *Electron. Commun. Probab*, vol. 18, no. 0, 2013.