Average Case Analysis of High-Dimensional Block-Sparse Recovery and Regression for Arbitrary Designs

Waheed U. Bajwa[†] **Marco F. Duarte**[‡] **Robert Calderbank**[♯]

[†]Dept. of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08816
[‡]Dept. of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 01003
[♯]Dept. of Electrical and Computer Engineering, Duke University, Durham, NC 27708

Abstract

This paper studies conditions for high-dimensional inference when the set of observations is given by a linear combination of a small number of groups of columns of a design matrix, termed the "block-sparse" case. In this regard, it first specifies conditions on the design matrix under which most of its block submatrices are well conditioned. It then leverages this result for average-case analysis of high-dimensional block-sparse recovery and regression. In contrast to earlier works: (i) this paper provides conditions on arbitrary designs that can be explicitly computed in polynomial time, (ii) the provided conditions translate into near-optimal scaling of the number of observations with the number of active blocks of the design matrix, and (iii) the conditions suggest that the spectral norm, rather than the column/block coherences, of the design matrix fundamentally limits the performance of computational methods in high-dimensional settings.

1 Introduction

Consider the linear model $y = X\beta$, which relates a parameter vector $\beta \in \mathbb{R}^p$ to observations $y \in \mathbb{R}^n$ through a design matrix (henceforth referred to as a *dictionary*) $X \in \mathbb{R}^{n\times p}$. Statistical inference using this linear model requires understanding conditions under which the inference problem is well posed. For instance, inferring anything about β will be a moot point if the nullspace of X were to contain β. Thus, a large part of the literature on linear models is devoted to characterizing conditions on X and β for reliable inference.

Traditionally, inference using linear models proceeds under the assumption that the number of observations n equals or exceeds the number of parameters p. In this setting, *explicitly verifiable conditions* such as X being full column rank or XX^T being well conditioned are common in the literature [1–3]. In contrast, there has recently been a growing interest to study *high-dimensional inference* under linear models, corresponding to n being much smaller than p. This setting is the hallmark of high-dimensional statistics [4], arises frequently in many application areas [5], and forms the cornerstone of the philosophy behind compressed sensing [6, 7]. It of course follows from simple linear algebra that inferring about every possible β from $y = X\beta$ is impossible in this setting; instead, the high-dimensional inference literature commonly operates under the assumption that β has only a few nonzero parameters—typically on the order of n—and characterizes corresponding conditions on X for reliable inference. Some notable conditions in this regard include the spark [8], the restricted isometry property [9], the irrepresentable condition [10], the incoherence condition [11], the restricted eigenvalue assumption [12], and the nullspace property [13].

While these and other conditions in the literature differ from each other in one way or the other, they all share one simple fact: *requiring that X satisfies one of these conditions implies that one or more column submatrices (subdictionaries) of X must be full column rank and/or well conditioned.* Explicitly verifying that X satisfies one of these properties is computationally daunting (NP-hard in some cases [14]), while indirect means of verifying these conditions provide rather pessimistic bounds on the dimensions of subdictionaries of X that are well conditioned [15]. In a recent series of seminal works, these pessimistic bounds associated with *verifiable conditions* on X have been circumvented through an *average-case analysis* [16–21].

Our Contributions: Our focus in here is on high-dimensional inference for the case when β not only has a few nonzero parameters, but also its nonzero

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33. Copyright 2014 by the authors.

parameters exhibit a certain *block* (or *group*) structure. Specifically, we have $\beta = [\beta_1^T \ \beta_2^T \ \dots \ \beta_r^T]^T$ with $\beta_i \in \mathbb{R}^m$ for $m, r \in \mathbb{Z}_+$, $p = rm$, and only $k \ll r$ of the β_i's are nonzero (sub)vectors. Such setups are referred to as *block sparse* (or *group sparse*) and arise in various contexts in a number of applications [22–26]. The challenge for inference in this block-sparse setting then becomes specifying conditions under which one or more *block subdictionaries* of X are full column rank and/or well conditioned. A number of researchers have made progress in this regard recently, reporting conditions on X in the block setting that mirror many of the ones reported in [8–13] for the classical setup; see, e.g., [22, 23, 27–52]. However, just like in the classical setup, verifying that X satisfies one of these properties in the block setting ends up being either computationally intractable or results in rather pessimistic bounds on the dimensions of block subdictionaries of X that are well conditioned. In contrast to these works, and in much the same way [16–21] reasoned in the classical case, we are interested in overcoming the pessimistic bounds associated with verifiable conditions on X for high-dimensional inference in the block-sparse setting by resorting to an average-case analysis.

Our first contribution in this regard is a generalization of [18, 19] that establishes that *most block subdictionaries* of X having unit ℓ_2-norm columns are guaranteed to be well conditioned with the number of *blocks* in the subdictionary proportional to $r/(\|X\|_2^2 \log p)$ provided X satisfies a polynomial-time verifiable condition, termed the *block incoherence condition* (BIC). This result also implies that if X is a unit norm tight frame [15], corresponding to $\|X\|_2^2 = p/n$, then it can be explicitly verified that most *block subdictionaries* of X of dimension $n \times O(n/\log p)$ are well conditioned.

While our ability to guarantee that most block subdictionaries of a dictionary that satisfies the BIC are well conditioned makes us optimistic about the use of such designs in inference problems, there remains an analytical gap in going from conditioning of block subdictionaries to performance of inference tasks. Our second contribution in this regard is the application of the result concerning the conditioning of block subdictionaries to provide tighter verifiable conditions for average-case guarantees in block-sparse recovery (i.e., obtaining block-sparse β from $y = X\beta$) and block-sparse regression (i.e., estimating $X\beta$ from $y = X\beta +$ noise with β being block sparse).

Finally, we present numerical experiments to highlight an aspect of inference under the linear model that is seldom discussed in the related literature: *the spectral norm of the dictionary $\|X\|_2$ influences the inference performance much more than any of its other measures.* Specifically, our numerical experiments show that the performances of block-sparse recovery and regression are inversely proportional to $\|X\|_2^2$ and are for the most part independent of correlations between the columns of X—an outcome that also hints at the possible (orderwise) tightness of our results.

Notational Convention: We use uppercase and lowercase Roman/Greek letters for matrices and vectors/scalars, respectively. Given a matrix A, $\|A\|_2$ and A^T denote the spectral norm ($\sigma_{\max}(A)$) and the adjoint operator of A, respectively. Given a vector v, we use $\|v\|_q$ and v^T to denote the usual ℓ_q norm and transpose of v, respectively. Given a set $\mathcal{S}$, we use $A_\mathcal{S}$ (resp. $v_\mathcal{S}$) to denote the submatrix (resp. subvector) obtained by retaining the columns of A (resp. entries of v) corresponding to the indices in $\mathcal{S}$. Given a random variable R, we use $\mathbb{E}_q[R]$ to denote $\big(\mathbb{E}[R^q]\big)^{1/q}$. Finally, Id, $\otimes$, and $\langle \cdot, \cdot \rangle$ denote the identity operator, Kronecker product and inner product, respectively.

2 Conditioning of Random Block Subdictionaries

In this section, we state and discuss the main result of this paper concerning the conditioning of block subdictionaries of the $n \times p$ dictionary X. Here, and in the following, we assume X has a block structure that comprises $r = p/m$ blocks of dimensions $n \times m$ each; in particular, we can write without loss of generality that $X = [X_1 \ X_2 \ \dots \ X_r]$, where each block $X_i = [X_{i,1} \ \dots \ X_{i,m}]$ is an $n \times m$ matrix. We also assume throughout this paper that the columns of X are normalized: $\|X_{i,j}\|_2 = 1$ for all $i = 1, \dots, r$, $j = 1, \dots, m$. The problem we are interested in addressing in this section is the following. Let $\mathcal{S} \subset \{1, \dots, r\}$ with $|\mathcal{S}| = k$ and define an $n \times km$ block subdictionary $X_\mathcal{S} = [X_i : i \in \mathcal{S}]$. Then what are the conditions on X that will guarantee that the singular values of $X_\mathcal{S}$ concentrate around unity? Since addressing this question for an *arbitrary* subset $\mathcal{S}$ is known to lead to either nonverifiable conditions or pessimistic bounds on k, our focus here is on a subset $\mathcal{S}$ that is drawn uniformly at random from all $\binom{r}{k}$ possible k-subsets of $\{1, \dots, r\}$. In words, such a model for $\mathcal{S}$ in the context of high-dimensional inference asserts that no one block in $\beta = [\beta_1^T \ \beta_2^T \ \dots \ \beta_r^T]^T$ is more likely to be nonzero than the others. Our main result for the conditioning of random block subdictionaries relies on a condition that we term the *block incoherence condition* (BIC).

Definition 1. Define the intra-block coherence of the dictionary X as $\mu_I := \max_{1 \le i \le r} \|X_i^T X_i - \mathrm{Id}_m\|_2$, and define the inter-block coherence[1] of the dictionary X as $\mu_B := \max_{1 \le i \ne j \le r} \|X_i^T X_j\|_2$. We say that X satisfies the *block incoherence condition* (BIC) with parameters (c_1, c_2) if $\mu_I \le c_1$ and $\mu_B \le c_2/\log p$ for some positive numerical constants c_1, c_2.

[1] See [38] for a related measure that is given by μ_B/m.

Note that μ_I measures the deviation of individual blocks $\{X_i\}$ from being orthonormal and is identically equal to zero for the case of orthonormal blocks. In contrast, μ_B measures the similarity between different blocks and cannot be zero in the n smaller than p setting. Informally, the BIC dictates that individual blocks of X do not diverge from being orthonormal in an unbounded fashion and the dissimilarity between different blocks scales as $O(1/\log p)$. *Note also that the BIC can be verified in polynomial time.* We are now ready to state our first result, proven in [53].

Theorem 1. *Suppose that the $n \times p$ dictionary $X = [X_1\ X_2\ \ldots\ X_r]$ satisfies the BIC with parameters (c_1, c_2). Let $\mathcal{S}$ be a k-subset drawn uniformly at random from all $\binom{r}{k}$ possible k-subsets of $\{1, \ldots, r\}$. Then, as long as $k \leq c_0 r/(\|X\|_2^2 \log p)$ for some positive numerical constant c_0 that depends only on (c_1, c_2), the singular values of the block subdictionary $X_{\mathcal{S}} = [X_i : i \in \mathcal{S}]$ satisfy $\sigma_i(X_{\mathcal{S}}) \in [\sqrt{1/2}, \sqrt{3/2}\,]$, $i = 1, \ldots, km$, with probability (with respect to the random choice of the subset $\mathcal{S}$) of at least $1 - 2p^{-4\log 2}$.*

Note that Theorem 1 does require (c_1, c_2) to be sufficiently small, with c_0 decreasing as c_1 and c_2 increase. In words, Theorem 1 states that if X satisfies the BIC then most of its block subdictionaries of dimensions $n \times km$ act as isometries on $\mathbb{R}^{km}$ for $k = O(r/(\|X\|_2^2 \log p))$. To better understand the bound $k = O(r/(\|X\|_2^2 \log p))$, notice that $\|X\|_2^2 \geq p/n$ for the case of a normalized dictionary [18], implying $r/(\|X\|_2^2 \log p) = O(n/(m \log p))$. The equality $\|X\|_2^2 = p/n$ is achievable by an X with orthogonal rows, implying Theorem 1 allows optimal scaling of the dimensions of well-conditioned block subdictionaries. Perhaps the most surprising aspect of this theorem, which sets it apart from other works in block settings [28–30, 34, 38, 48, 50], is the assertion it makes about the effects of different measures of X on the conditioning of random block subdictionaries. Roughly, Theorem 1 suggests that as soon as the BIC is satisfied for sufficiently small (c_1, c_2), both μ_I and μ_B stop playing a role in determining the dimensions of well-conditioned subdictionaries; rather, it is the spectral norm $\|X\|_2$ that plays a primary role in this regard. Such an assertion of course needs to be carefully examined, given that Theorem 1 is only concerned with sufficient conditions. Nevertheless, numerical experiments carried out in the context of block-sparse recovery (cf. Section 3) and block-sparse regression (cf. Section 4) lend credence to this assertion.

Discussion: Among existing works focusing on the conditioning of random (non-block) subdictionaries [16–21], [19] and [20] are the ones with the strongest results. Specifically, [16, 17] deal with the case of X being a concatenation of two orthonormal bases, while [21] studies the case of X being a disjoint union of orthonormal bases. The results in [19] and [20] are related to each other in the sense that [20] extends [19] to the case when the subdictionaries of X are not necessarily selected uniformly at random. Theorem 1 is inspired by [19] and is rather tight in the sense that it reduces to the result of [19] for $m = 1$.

3 Average-Case Analysis of Block-Sparse Recovery

We now shift our focus to the applicability of Theorem 1 in the context of inference problems. In this section, we begin with the problem of recovery of block-sparse β from $y = X\beta$. Because of the relevance of block sparsity in many applications, significant efforts have been made toward development of block-sparse recovery methods and matching guarantees on the number of observations required for successful recovery [27–30, 34–44, 49–51, 54]. However, the results reported in some of these works are only applicable in the case of random designs [35–37, 44, 51], while those reported in other works rely on conditions that either cannot be explicitly verified in polynomial time [27–29, 36, 41–44, 49, 50, 54] or result in a suboptimal scaling of the number of observations due to their focus on the worst-case performance [28–30, 38–41, 50].

To the best of our knowledge, the only work that does not have the aforementioned limitations is [34]. However, the focus in [34] is only on the restrictive multiple measurement vector (MMV) problem, rather than the general block-sparse recovery problem. In addition, the guarantees provided in [34] rely on the nonzero entries of β following either Gaussian or spherical distributions. In contrast, we make use of Theorem 1 in the following to state a result for average-case block-sparse recovery that suffers from none of these and earlier limitations. Our result depends primarily on the spectral norm of X, while it has a mild dependence on the intra- and inter-block coherence through the BIC; all three of these quantities can be explicitly computed in polynomial time. It further requires only weak assumptions on the distribution of the nonzero entries of β. Equally important, the forthcoming result does not suffer from the so-called "square-root bottleneck" [20]; specifically, it allows near-optimal scaling of the sparsity level km as a function of n for dictionaries X with small spectral norms (e.g., tight frames).

Problem Formulation: Our exposition throughout this section will be based upon the following formulation. We are interested in recovering a block-sparse $\beta \in \mathbb{R}^p$ from observations $y = X\beta$, where X is an $n \times p$ observation matrix with $n \ll p$ and $y \in \mathbb{R}^n$ denotes the observations. We assume β comprises a total of r blocks, each of size m (yielding $p = rm$), and represent it without loss of generality as $\beta = [\beta_1^T\ \beta_2^T\ \ldots\ \beta_r^T]^T$ with each block $\beta_i \in \mathbb{R}^m$. In order to make this problem well posed, we require that β is k-block sparse with $\#\{i : \beta_i \neq \mathbf{0}\} = k \ll r$. Finally, we impose a mild statistical prior on β, as described below.

M1) The *block support* of β, $\mathcal{S} = \{i : \beta_i \neq \mathbf{0}\}$, has a uniform distribution over all k-subsets of $\{1, \ldots, r\}$,

M2) Entries in β have zero median (i.e., its nonzero entries are equally likely to be positive and negative): $\mathbb{E}(\text{sign}(\beta)) = \mathbf{0}$, where $\text{sign}(\cdot)$ denotes the entry-wise sign operator, and

M3) Nonzero blocks of β have statistically independent "directions." Specifically, defining the block-wise sign operator $\overline{\text{sign}}(\beta_i) = \beta_i/\|\beta_i\|_2$ to be the unit-norm vector pointing in the direction of β_i in $\mathbb{R}^m$, we require $\mathbb{P}\left(\bigcap_{i\in\mathcal{S}}\left(\overline{\text{sign}}(\beta_i) \in \mathcal{A}_i\right)\right) = \prod_{i\in\mathcal{S}}\mathbb{P}\left(\overline{\text{sign}}(\beta_i) \in \mathcal{A}_i\right)$, where $\mathcal{A}_i \subset \mathbb{S}^{m-1}$ and $\mathbb{S}^{m-1}$ is the unit sphere in $\mathbb{R}^m$.

Note that M2 and M3 are trivially satisfied in the case of the nonzero blocks of β drawn independently from either Gaussian or spherical distributions. However, it is easy to convince oneself that many other distributions—including those that are not absolutely continuous—will satisfy these two conditions.

Main Result and Discussion: We are interested in understanding the average-case performance of the following mixed-norm convex optimization program for recovery of block-sparse β satisfying M1–M3:

$$\widehat{\beta} = \arg\min_{\bar{\beta}\in\mathbb{R}^p} \|\bar{\beta}\|_{2,1} \text{ such that } y = X\bar{\beta}, \qquad (1)$$

where the $\ell_{2,1}$ norm of a vector $\beta \in \mathbb{R}^p$ containing r blocks of m entries each is defined as $\|\beta\|_{2,1} := \sum_{i=1}^{r}\|\beta_i\|_2$. While (1) has been utilized in the past for block-sparse recovery (see, e.g., [34, 35, 44]), an average-case analysis result along the following lines is novel. The following theorem is proven in [53].

Theorem 2. *Suppose that $\beta \in \mathbb{R}^p$ is k-block sparse and it is drawn according to the statistical model M1, M2, and M3. Further, assume that β is observed according to the linear model $y = X\beta$, where the $n \times p$ matrix X satisfies the BIC with some parameters (c_1, c_2). Then, as long as $k \leq c_0 r/\|X\|_2^2 \log p$ for some positive numerical constant $c_0 := c_0(c_1, c_2)$, the minimization (1) results in $\widehat{\beta} = \beta$ with probability at least $1 - 4p^{-4\log 2}$.*

Interestingly, Theorem 2 specialized to the non-block sparse case (by setting $m = 1$ and $r = p$) gives us an average-case analysis result for sparse recovery that has never been explicitly stated in prior works. In the interest of space, we forgo a formal statement of that corollary of Theorem 2 (see [53] for details), but we do elaborate on the similarities and differences between the two results. In terms of similarities, the results for both non-block sparse and block-sparse settings allow for the same scaling of the *total number of nonzero entries* in β. However, while the guarantee for non-block sparsity requires that the inner product of any two columns in X be $O(1/\log p)$, Theorem 2 allows for less restrictive inner products of columns *within* blocks as long as $\mu_I = O(1)$. Similarly, while in non-block sparsity it is required that the signs of the nonzero entries in β be independent, Theorem 2 allows for correlations among the signs of entries *within* nonzero blocks. With the caveat that in both cases the guarantees only specify sufficient conditions, this seems to suggest that explicitly accounting for block structures allows one to expand the classes of sparse β *and* dictionaries X under which successful (average-case) recovery can be guaranteed.

Next, we comment on the tightness of the scaling on the number of nonzero entries in both non-block sparse and block-sparse settings. Assuming appropriate conditions on β and (intra-/inter-block) coherence of X are satisfied, both results allow for the number of nonzero entries to scale like $O(n/\log p)$ for dictionaries X that are "approximately" tight frames [15]: $\|X\|_2^2 \approx p/n$. This suggests a near-optimal nature of both results (modulo perhaps log factors) as one cannot expect better than linear scaling of the number of nonzero entries as a function of the number of observations. In particular, literature on frame theory [55] can be leveraged to specialize these results for oft-used designs (e.g., Gaussian, random partial Fourier) and to establish that in such cases the scaling of our guarantee matches that obtained using nonverifiable conditions such as the restricted isometry property [9, 56].

We conclude by noting that the main difference between our non-block sparse average-case result (Theorem 2 using $m = 1$) and existing literature (e.g., [18, Theorem 14]) is the role that the coherence $\mu(X)$ plays in the guarantees. In [18, Theorem 14], the maximum allowable sparsity k is inversely proportional to $\mu^2(X)$. In contrast, we assert that the maximum allowable sparsity is not fundamentally determined by the coherence. Numerical experiments reported in the following verify that this is indeed the case.

Numerical Experiments: One of the fundamental takeaways of this section is that the spectral norm of X, rather than its (intra-/inter-block) coherence, determines the maximum allowable sparsity in (block)-sparse recovery problems. In order to experimentally verify this insight, we performed a set of block-sparse recovery experiments with custom-designed dictionaries having varying spectral norms and coherence values. Throughout our experiments, we set $p = 5000$, the block size and the number of blocks to $m = 10$ and $r = 500$, respectively, and the number of observations to $n = 858$ (computed from the bound in [57] for $k = 20$ nonzero blocks). In order to design our dictio-

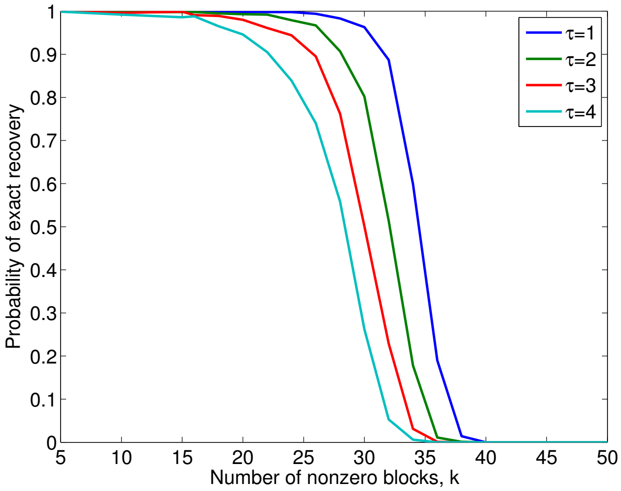

Figure 1: Performances of dictionaries X with varying spectral norms and roughly equal coherences (cf. Table 1) in block-sparse recovery as a function of the number of nonzero blocks k; $\tau \in \mathcal{T}$ denotes the spectral norm multiplier used to generate the dictionary.

τ	1	2	3	4
$\|X_\tau\|_2$	3.3963	6.7503	10.0547	13.2034
$\mu(X_\tau)$	0.1992	0.2026	0.2000	0.2207
$\mu_B(X_\tau)$	0.2973	0.3431	0.5573	0.8490
$\mu_I(X_\tau)$	0.1992	0.2026	0.2177	0.3787

Table 1: Spectral norms and coherences for the dictionaries used in the experiments of Figure 1.

naries, we first used Matlab's random number generator to obtain 2000 standard normal matrices, followed by normalization of the columns. Next, we manipulated the singular values of each of these matrices to increase their spectral norms by a set of integer multipliers $\mathcal{T}$. Finally, for each of the $2000 \cdot |\mathcal{T}|$ resulting matrices, we normalized their columns to obtain our dictionaries and recorded their spectral norms $\|X\|_2$, coherences $\mu(X)$, inter-block coherences $\mu_B(X)$, and intra-block coherences $\mu_I(X)$.

We evaluate the block-sparse recovery performance of each resulting X using Monte Carlo trials, corresponding to the generation of 1000 block-sparse β's with k nonzero blocks. Each β has block support selected uniformly at random according to M1 and nonzero entries drawn independently from the standard Gaussian distribution $\mathcal{N}(0, \mathrm{Id})$. We then obtain the observations $y = X\beta$ using the dictionary X under study for each one of these β and perform recovery using the minimization (1).[2] We define successful recovery to be the case when the block support of $\widehat{\beta}$ matches the block support of β and the submatrix of X with columns corresponding to the block support of β has full rank.

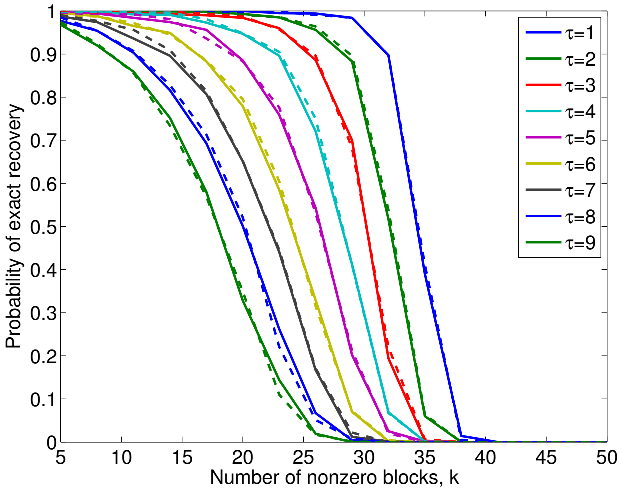

Figure 2: Performances of dictionaries X with varying spectral norms and extremal coherences (cf. Table 2) in block-sparse recovery as a function of the number of nonzero blocks k; $\tau \in \mathcal{T}$ denotes the value of the spectral norm multiplier used. Solid lines correspond to X's with minimum coherence, while dashed lines correspond to X's with maximum coherence.

Figure 1 shows the performances of dictionaries X of increasing spectral norms ($\mathcal{T} = \{1, 2, 3, 4\}$), where we choose the dictionary (among the 2000 available options) whose coherence value is closest to 0.2. The spectral norms, coherences, inter-block coherences, and intra-block coherences for each one of these four chosen (and fixed) dictionaries are collected in Table 1. The performance is shown as a function of the number of nonzero blocks k in β. The figure shows a consistent improvement in the values of k for which successful recovery is achieved as the spectral norm of X decays, even though $\mu(X)$ does not significantly change among the dictionaries.

To further emphasize strong dependence of sparse recovery on spectral norm and weak dependence on (intra-/inter-block) coherences, Figure 2 shows the performance of dictionaries X with increasing spectral norms ($\mathcal{T} = \{1, \dots, 9\}$), where we choose dictionaries with the largest and smallest coherence values for each $\tau \in \mathcal{T}$ (among the 2000 available options). The spectral norms, coherences, inter-block coherences, and intra-block coherences for these 18 chosen (and fixed) dictionaries are collected in Table 2. The figure shows not only the same consistent improvement as the spectral norm of the dictionary decays, but also that significant changes in the values of the (intra-/inter-block) coherences do not significantly affect the recovery performance. This behavior agrees with our expectation from Theorem 2 that the role of the intra-/inter-block coherences in performance guarantees is limited to the BIC and decoupled from the number of nonzero blocks k (equivalently, number of nonzero entries km) in β.

[2] We used the SPGL1 package [58] for these simulations.

τ	1	2	3	4	5	6	7	8	9
$\|X_{\tau,\min}\|_2$	3.4064	6.7726	10.0536	13.2034	16.3421	19.2980	22.1413	24.6710	27.2951
$\|X_{\tau,\max}\|_2$	3.3963	6.7503	10.0543	13.2250	16.2747	19.1506	21.9975	24.7026	27.3199
$\mu(X_{\tau,\min})$	0.1230	0.1198	0.1500	0.2207	0.2964	0.3760	0.4583	0.5337	0.6000
$\mu(X_{\tau,\max})$	0.1992	0.2026	0.2698	0.3816	0.4863	0.5778	0.6566	0.7225	0.7758
$\mu_B(X_{\tau,\min})$	0.2887	0.3177	0.5357	0.8490	1.2917	1.7372	2.2263	2.5989	3.1204
$\mu_B(X_{\tau,\max})$	0.2973	0.3431	0.6516	1.0287	1.4419	1.6616	2.0230	2.4479	2.8737
$\mu_I(X_{\tau,\min})$	0.1487	0.2002	0.3368	0.3787	0.3472	0.4385	0.5462	1.0551	1.3095
$\mu_I(X_{\tau,\max})$	0.1992	0.2026	0.2698	0.3816	0.4863	0.5778	0.8273	1.0415	1.2723

Table 2: Metrics for the dictionaries used in the experiments of Figure 2 and Figure 3.

4 Average-Case Analysis of Block-Sparse Linear Regression

In this section, we leverage Theorem 1 to obtain average-case results for block-sparse linear regression, defined as estimating $X\beta$ from $y = X\beta + \text{noise}$ when β has a block-sparse structure. In particular, we focus on two popular convex optimization-based methods, the lasso [59] and the group lasso [22], for characterizing results for block-sparse regression. Empirical evidence in the literature suggests that an appropriately regularized group lasso can outperform the lasso whenever there is a natural grouping of the regression variables in terms of their contributions to the observations [22, 23]. We analytically characterize the block-sparse regression performances of both the lasso and the group lasso, which helps us highlight one of the ways in which the group lasso might outperform the lasso for regression problems.

Note that the analytical characterization of the group lasso using ℓ_1/ℓ_2 regularization in the high-dimensional ($n \ll p$) setting has received attention recently in the literature [23, 31–33, 45–48, 52]. However, prior work on the performance of the group lasso either studies an asymptotic regime [23, 31–33], focuses on random designs [23, 32, 45, 46], and/or relies on conditions that are either computationally prohibitive to verify [31, 33, 47, 52] or that do not allow for near-optimal scaling of the number of observations with the number of active blocks of regression variables k [48]. In contrast, our forthcoming analysis circumvents these shortcomings by adopting a probabilistic model for the blocks of regression coefficients in β. Our probabilistic model, described by the conditions M1–M3 in Section 3, is motivated by that of Candès and Plan [60] for non-block linear regression, which helped them overcome somewhat similar analytical hurdles in relation to non-block regression performance of the lasso. To the best of our knowledge, the results stated in the sequel concerning the block-sparse regression performances of the lasso and the group lasso[3] are the first ones that are non-asymptotic in nature and applicable to arbitrary designs through verifiable conditions, while still allowing for near-optimal scaling of the number of observations with the number of nonzero blocks of β.

Problem Formulation: We are once again in the high-dimensional ($n \ll p$) setting with $y = X\beta + z$, where X is the design matrix containing one regression variable per column, $\beta \in \mathbb{R}^p = [\beta_1^T\ \beta_2^T\ \dots\ \beta_r^T]^T$ is the k-block sparse vector of regression coefficients corresponding to these variables (i.e., $\#\{i : \beta_i \neq \mathbf{0}\} = k \ll r$), and $z \in \mathbb{R}^n$ is the modeling error. Here, we assume without loss of generality that X has unit-norm columns, while we assume the modeling error z to be an independent and identically distributed (i.i.d.) Gaussian vector with variance σ^2. Finally, in keeping with the earlier discussion, we impose a mild statistical prior on β that is given by the conditions M1, M2, and M3 in Section 3. The fundamental goal in here then is to obtain an estimate $\widehat{\beta}$ from y such that $X\widehat{\beta}$ is as close to $X\beta$ as possible, where the closeness is measured in terms of the ℓ_2 regression error, $\|X\beta - X\widehat{\beta}\|_2$.

Main Results and Discussion: In this section, we are interested in understanding the average-case block-sparse regression performance of two methods. The first one of these methods is the lasso [59], which ignores any grouping of the regression variables and estimates the vector of regression coefficients as

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + 2\lambda\sigma\|\beta\|_1, \tag{2}$$

where $\lambda > 0$ is a tuning parameter. In terms of a baseline result for the lasso, we can extend the probabilistic model of Candès and Plan [60] for non-block linear regression to the block setting and state the following theorem that follows trivially from Theorem 1 in this paper and the proof of [60, Theorem 1.2].

Theorem 3 ([60, Theorem 1.2] and Theorem 1)**.** *Suppose that the vector of regression coefficients $\beta \in \mathbb{R}^p$ is k-block sparse and that the observation vector can be modeled as $y = X\beta + z$ with the modeling error z being i.i.d. Gaussian with variance σ^2. Further, assume that β is drawn according to the statistical model M1 and M2 with the signs of its nonzero entries being i.i.d., and the $n \times p$ matrix X satisfies (i) $\mu(X) = O(1/\log p)$*

[3] We refer to the group lasso using ℓ_1/ℓ_2 regularization as simply "group lasso" in the following for brevity.

and (ii) the BIC with some parameters (c_1', c_2'). Then, as long as $k \leq c_0' r/\|X\|_2^2 \log p$ for some positive numerical constant $c_0' := c_0'(c_1', c_2')$, the lasso estimate $\widehat{\beta}$ in (2) *computed with $\lambda = \sqrt{2\log p}$ obeys*

$$\|X\beta - X\widehat{\beta}\|_2^2 \leq C' mk\sigma^2 \log p$$

with probability at least $1 - O(p^{-1})$, where $C' > 0$ is a constant independent of the problem parameters.

While this theorem suggests that the lasso solution in the block setting enjoys many of the optimality properties of the lasso solution in the non-block setting (see, e.g., the discussion in [60]), it fails to extend to the case when the independence assumption on the signs of the nonzero regression coefficients is replaced by the less restrictive condition M3. In particular, one expects that allowing for arbitrary correlations within the blocks of regression coefficients will limit the usefulness of the lasso for linear regression in the presence of large blocks. While such an insight can be difficult to confirm in the case of arbitrary design matrices and average-case analysis, we provide an extension of Theorem 3 in the following that highlights the challenges for the lasso in the case of regression of block-sparse vectors with arbitrarily correlated blocks. The following theorem is proven in [53].

Theorem 4. *Suppose that the vector of regression coefficients $\beta \in \mathbb{R}^p$ is k-block sparse and it is drawn according to the statistical model M1, M2, and M3. Further, assume that the observation vector can be modeled as $y = X\beta + z$, where the $n \times p$ matrix X satisfies $\mu_I(X) \leq c_1''$ and $\mu_B(X) \leq c_2''/(\sqrt{m}\log p)$ for some positive numerical constants c_1'', c_2'', and the modeling error z is i.i.d. Gaussian with variance σ^2. Then, as long as $k \leq c_0'' r/\|X\|_2^2 m \log p$ for some positive numerical constant $c_0'' := c_0''(c_1'', c_2'')$, the lasso estimate $\widehat{\beta}$ in* (2) *computed with $\lambda = \sqrt{2\log p}$ obeys*

$$\|X\beta - X\widehat{\beta}\|_2^2 \leq C'' mk\sigma^2 \log p$$

with probability at least $1 - p^{-1}(2\pi \log p)^{-1/2} - 8p^{-4\log 2}$, where $C'' > 0$ is a constant independent of the problem parameters.

While both Theorems 3 and 4 guarantee same scaling of the regression error, the scalings of the maximum number of allowable nonzero blocks and the block coherence in Theorem 4 match the ones in Theorem 3 only for the case of $m = O(1)$; otherwise, Theorem 4 with correlated blocks results in less-desirable scalings of k and $\mu_B(X)$. The proof of Theorem 4 in [53] shows that this dependence upon m—the size of the blocks—is a direct consequence of allowing for arbitrary correlations within blocks. A natural question to ask then is whether it is possible to return to the scalings of Theorem 3 *without* sacrificing intra-block correlations. The answer to this is in the affirmative as long as one explicitly accounts for the block structure of β.

Specifically, the group lasso explicitly accounts for the grouping of the regression variables in its formulation and estimates the vector of regression coefficients as

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + 2\lambda\sigma\sqrt{m}\|\beta\|_{2,1}, \qquad (3)$$

where $\lambda > 0$ is once again a tuning parameter. The following theorem shows that the group lasso can achieve the same scaling results as the lasso for block-sparse vectors (cf. [60]), while allowing for arbitrary correlations among the regression coefficients within blocks. The following theorem is proven in [53].

Theorem 5. *Suppose that the vector of regression coefficients $\beta \in \mathbb{R}^p$ is k-block sparse and it is drawn according to the statistical model M1, M2, and M3. Further, assume that the observation vector can be modeled as $y = X\beta + z$, where the $n \times p$ matrix X satisfies the BIC with some parameters (c_1, c_2), and the modeling error z is i.i.d. Gaussian with variance σ^2. Then, as long as $k \leq c_0 r/\|X\|_2^2 \log p$ for some positive numerical constant $c_0 := c_0(c_1, c_2)$, the group lasso estimate $\widehat{\beta}$ in* (3) *computed with $\lambda = \sqrt{2\log p}$ obeys*

$$\|X\beta - X\widehat{\beta}\|_2^2 \leq C mk\sigma^2 \log p$$

with probability at least $1 - p^{-1}(2\pi \log p)^{-1/2} - 8p^{-4\log 2}$, where $C > 0$ is a constant independent of the problem parameters.

We note in passing that when $m = 1$, block sparsity reduces to the canonical sparsity, block coherence $\mu_B(X)$ reduces to the coherence $\mu(X)$, (3) reduces to (2), and Theorem 5 essentially reduces to [60, Theorem 1.2].

With the caveat that both Theorems 4 and 5 are concerned with sufficient conditions for average-case regression, we now comment on the strengths and weaknesses of these two results. Assuming appropriate conditions are satisfied, we have that both the lasso and the group lasso result in the same scaling of the regression error, $\|X\beta - X\widehat{\beta}\|_2^2 = O(mk\sigma^2 \log p)$, in the presence of intra-block correlations. This scaling of the regression error is indeed the best that any method can achieve, modulo the logarithmic factor, since we are assuming that the observations are described by a total of mk regression variables. Unlike the lasso, however, the group lasso also allows for a more favorable scaling of the maximum number of regression variables contributing to the observations, $km = O(p/\|X\|_2^2 \log p)$, even when arbitrary intra-block correlations are permitted. In fact, similar to the discussion in Section 3, it is easy to conclude that this scaling of the number of nonzero regression coefficients is near-optimal since it leads to a linear relationship (modulo logarithmic factors) between the number of observations n and the number of active regression variables km for the case of design matrices that are approximately tight frames: $\|X\|_2^2 \approx p/n$. The other main difference between Theorems 4 and 5 is the role that the inter-block coherence $\mu_B(X)$ plays in guarantees for the lasso and the group lasso. Specifically, Theorem 4 requires the inter-block coherence to be smaller, $\mu_B(X) = O(1/\sqrt{m}\log p)$, than Theorem 5 for the lasso to yield near-optimal regression error in the case of intra-block correlations. This discussion suggests that reliable linear regression of block-sparse vectors can be carried out using the group lasso for a larger class of regression vectors and design matrices than the lasso. We plan to provide a more rigorous mathematical understanding of these and other subtle but important differences between the lasso and the group lasso in future works.

Numerical Experiments: One of the most important implications of this section is that, similar to the case of block-sparse recovery, the number of maximum allowable active regression variables in regression of block-sparse vectors is fundamentally a function of the spectral norm of X, provided its inter- and intra-block coherences are not too large. However, such a claim needs to be carefully investigated since our results are only concerned with sufficient conditions. To this end, we resort to numerical experiments that help us evaluate the regression performance of the group lasso for a range of design matrices with varying spectral norms, coherences, inter-block coherences, and intra-block coherences. In order to generate these design matrices, we reuse the experimental setup described in Section 3 (corresponding to $n = 858$, $m = 10$, and $r = 500$).

For the sake of brevity, we focus only on the performance of the group lasso (3) for regression of block-sparse vectors.[4] This performance is evaluated for different design matrices using Monte Carlo trials, corresponding to generation of 1000 block-sparse β with k nonzero blocks. Each vector of regression coefficients has block support selected uniformly at random according to M1 in Section 3 and nonzero entries drawn independently from the Gaussian distribution. We then obtain the observations $y = X\beta + z$ using X under study for each one of the block-sparse β, where the variance σ^2 of the modeling error z is selected such that $\|\beta\|_2^2/n\sigma^2 \approx 0.84$. Finally, we carry out linear regression using the group lasso by setting $\lambda \approx 1.4592$ and we then record the regression error $\|X\beta - X\widehat{\beta}\|_2^2$.

Figure 3 shows the regression performance of the group lasso for designs with increasing spectral norms ($\tau = \{3, \dots, 7\}$), where we once again choose matrices with the largest and smallest coherence values for each τ (among the 2000 available options). The spectral norms, coherences, inter-block coherences, and intra-block coherences for these 10 chosen design matrices are still given by Table 2 in Section 3. Similar to the case of block-sparse recovery, we not only observe a consistent increase in the range of values of k for which the regression error exhibits linearity as the spectral norm of X decreases, but also see that significant changes in the (intra-/inter-block) coherences do not significantly affect the regression performance. This is in agreement with our expectation from Theorem 5 that the role of (intra-/inter-block) coherences in regression is limited to the BIC and is decoupled from the number of nonzero blocks k.

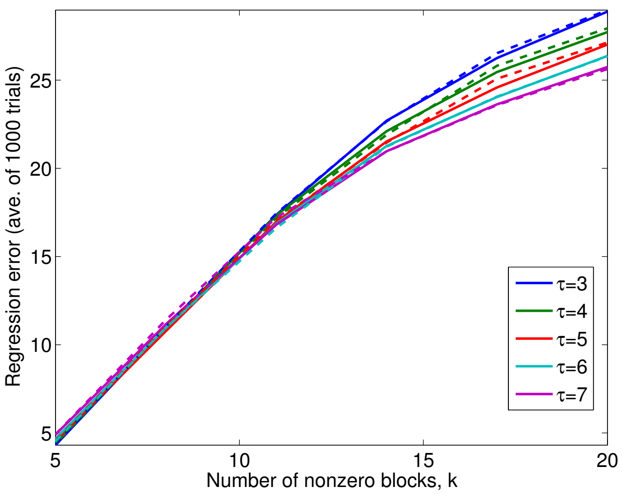

Figure 3: Performances of the group lasso for designs with varying spectral norms and extremal coherence values (cf. Table 2) in regression of block-sparse vectors as a function of the number of nonzero regression blocks k; τ denotes the value of the spectral norm multiplier used. Solid lines correspond to X's with minimum coherence, while dashed lines correspond to X's with maximum coherence.

5 Conclusion

We have provided conditions under which most block subdictionaries of a dictionary are well conditioned, and utilized these conditions for average-case analysis of block-sparse recovery and regression. Our results are based on verifiable conditions, they lead to near-optimal scaling of the number of observations with the number of active blocks, and they suggest that the spectral norm plays a far important role than the (inter-/intra-block) coherences in statistical inference.

Acknowledgements: WUB was supported in part by NSF grant CCF-1218942. MFD was supported in part by NSF grant ECCS-1201835 and by NSF Supplemental Funding DMS-0439872 to UCLA-IPAM, P.I. R. Caflisch. RC was supported in part by AFOSR under the Complex Networks Program.

[4] We used the `SpaRSA` Matlab package [61] with `debias` option turned on in all simulations in this section.

References

[1] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory.* Upper Saddle River, NJ: Prentice Hall, 1993.

[2] K. Knight and W. Fu, "Asymptotics for lasso-type estimators," *Ann. Statist.*, vol. 28, pp. 1356–1378, 2000.

[3] H. Zou, "The adaptive lasso and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1418–1429, Dec. 2006.

[4] D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," in *Proc. AMS Conf. Math Challenges of the 21st Century*, Los Angeles, CA, Aug. 2000. [Online]. Available: http://www-stat.stanford.edu/~donoho/Lectures/AMS2000/Curses.pdf

[5] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint." *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[6] D. L. Donoho, "Compressed sensing," *IEEE Trans. Info. Theory*, vol. 52, no. 4, pp. 1289–1306, Sept. 2006.

[7] E. J. Candès, "Compressive sampling," in *Proc. International Congress of Mathematicians*, vol. 3, Madrid, Spain, 2006, pp. 1433–1452.

[8] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," *Proc. Natl. Acad. Sci.*, vol. 100, no. 5, pp. 2197–2202, Mar. 2003.

[9] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," in *Compte Rendus de l'Academie des Sciences, Paris, Series I*, vol. 346, 2008, pp. 589–592.

[10] P. Zhao and B. Yu, "On model selection consistency of lasso," *J. Machine Learning Res.*, vol. 7, pp. 2541–2563, Nov. 2006.

[11] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy recovery of sparsity," in *Proc. 44th Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Sept. 2006.

[12] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of lasso and Dantzig selector," *Ann. Statist.*, vol. 37, no. 4, pp. 1705–1732, Aug. 2009.

[13] A. Cohen, W. Dahmen, and R. A. DeVore, "Compressed sensing and best k-term approximation," *Journal of the American Mathematical Society*, vol. 22, no. 1, pp. 211–231, Jan. 2009.

[14] A. S. Bandeira, E. Dobriban, D. G. Mixon, and W. F. Sawin, "Certifying the restricted isometry property is hard," *IEEE Trans. Inform. Theory*, vol. 59, no. 6, pp. 3448–3450, June 2013.

[15] W. U. Bajwa and A. Pezeshki, "Finite frames for sparse signal processing," in *Finite Frames*, P. Casazza and G. Kutyniok, Eds. Cambridge, MA: Birkhuser Boston, 2012, ch. 10, pp. 303–335.

[16] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[17] E. J. Candès and J. Romberg, "Quantitative robust uncertainty principles and optimally sparse decompositions," *Foundations of Computational Mathematics*, vol. 6, no. 2, pp. 227–254, April 2006.

[18] J. A. Tropp, "On the conditioning of random subdictionaries," *Appl. Comput. Harmon. Anal.*, vol. 25, pp. 1–24, 2008.

[19] ——, "Norms of random submatrices and sparse approximation," *C. R. Acad. Sci. Paris, Ser. I*, vol. 346, no. 23–24, pp. 1271–1274, 2008.

[20] P. Kuppinger, G. Durisi, and H. Bölcskei, "Uncertainty relations and sparse signal recovery for pairs of general signal sets," *IEEE Trans. Inform. Theory*, vol. 58, no. 1, pp. 263–277, Jan. 2012.

[21] S. Gurevich and R. Hadani, "The statistical restricted isometry property and the Wigner semicircle distribution of incoherent dictionaries," Mar. 2009, unpublished manuscript. [Online]. Available: http://arxiv.org/abs/0903.3627

[22] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Statist. Soc. B*, vol. 68, no. 1, pp. 49–67, Feb. 2006.

[23] F. Bach, "Consistency of the group lasso and multiple kernel learning," *J. Machine Learning Research*, vol. 9, no. 6, pp. 1179–1225, June 2008.

[24] M. Mishali and Y. Eldar, "Blind multiband signal reconstruction: Compressed sensing for analog signals," *IEEE Trans. Signal Proc.*, vol. 57, no. 3, pp. 993–1009, Mar. 2009.

[25] A. Bolstad, B. Van Veen, and R. Nowak, "Causal network inference via group sparse regularization," *IEEE Trans. Signa*, vol. 59, no. 6, pp. 2628–2641, Jun. 2011.

[26] M. F. Duarte and Y. C. Eldar, "Structured compressed sensing: From theory to applications," *IEEE Trans. Signal Processing*, vol. 59, no. 9, pp. 4053–4085, Sep. 2011.

[27] S. Cotter, B. Rao, E. Kjersti, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Processing*, vol. 53, no. 7, pp. 2477–2488, July 2005.

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

[29] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst, "Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 655–687, 2008.

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

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

[32] L. Meier, S. van de Geer, and P. Bühlmann, "The group lasso for logistic regression," *J. Royal Statist. Soc. B*, vol. 70, no. 1, pp. 53–71, Jan. 2008.

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

[34] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Trans. Info. Theory*, vol. 6, no. 1, pp. 505–519, Jan. 2010.

[35] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," *IEEE Trans. Signal Processing*, vol. 57, no. 8, pp. 3075–3085, Aug. 2009.

[36] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Info. Theory*, vol. 56, no. 4, pp. 1982–2001, Apr. 2010.

[37] M. Stojnic, "ℓ_2/ℓ_1-optimization in block-sparse compressed sensing and its strong thresholds," *IEEE J. Select. Top. Signal Processing*, vol. 4, no. 2, pp. 350–357, Apr. 2010.

[38] Y. C. Eldar, P. Kuppinger, and H. Bölcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. Signal Processing*, vol. 58, no. 6, pp. 3042–3054, June 2010.

[39] J. Fang and H. Li, "Recovery of block-sparse representations from noisy observations via orthogonal matching pursuit," Aug. 2011, preprint. Available at http://arxiv.org/pdf/1109.5430.

[40] Z. Ben-Haim and Y. C. Eldar, "Near-oracle performance of greedy block-sparse estimation techniques from noisy measurements," *IEEE J. Select. Top. Signal Processing*, vol. 5, no. 5, pp. 1032–1047, Sep. 2011.

[41] P. T. Boufounos, G. Kutyniok, and H. Rauhut, "Sparse recovery from combined fusion frame measurements," *IEEE Trans. Info. Theory*, vol. 57, no. 6, pp. 3864–3876, June 2011.

[42] J. M. Kim, O. K. Lee, and J. C. Ye, "Compressive MUSIC: Revisiting the link between compressive sensing and array signal processing," *IEEE Trans. Inform. Theory*, vol. 58, no. 1, pp. 278–301, Jan. 2012.

[43] M. Davies and Y. Eldar, "Rank awareness in joint sparse recovery," *IEEE Trans. Inform. Theory*, vol. 58, no. 2, pp. 1135–1146, Feb. 2012.

[44] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Info. Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

[45] G. Obozinski, M. J. Wainwright, and M. I. Jordan, "Support union recovery in high-dimensional multivariate regression," *Annals of Statistics*, vol. 39, no. 1, pp. 1–47, Jan. 2011.

[46] M. Kolar, J. Lafferty, and L. Wasserman, "Union support recovery in multi-task learning," *J. Machine Learning Res.*, vol. 12, no. 7, pp. 2415–2435, July 2011.

[47] Z. Fang, "Sparse group selection through co-adaptive penalties," Nov. 2011, preprint. Available at http://arxiv.org/pdf/1111.4416.

[48] K. Lounici, M. Pontil, S. van de Geer, and A. B. Tsybakov, "Oracle inequalities and optimal inference under group sparsity," *Ann. Statist.*, vol. 39, no. 4, pp. 2164–2204, 2011.

[49] K. Lee, Y. Bresler, and M. Junge, "Subspace methods for joint sparse recovery," *IEEE Trans. Inform. Theory*, vol. 58, no. 6, pp. 3613–3641, Jun. 2012.

[50] E. Elhamifar and R. Vidal, "Block-sparse recovery via convex optimization," *IEEE Trans. Signal Processing*, vol. 60, no. 8, pp. 4094–4107, Aug. 2012.

[51] M. F. Duarte, M. B. Wakin, D. Baron, S. Sarvotham, and R. G. Baraniuk, "Measurement bounds for sparse signal ensembles via graphical models," *IEEE Trans. Info. Theory*, vol. 59, no. 7, pp. 4280–4289, July 2013.

[52] J. Huang and T. Zhang, "The benefit of group sparsity," *Annals of Statistics*, vol. 38, no. 4, pp. 1978–2004, Aug. 2010.

[53] W. U. Bajwa, M. F. Duarte, and R. Calderbank, "Conditioning of random block subdictionaries with applications to block-sparse recovery and regression," *submitted to IEEE Trans. Inform. Theory*, Sep. 2013. [Online]. Available: http://arxiv.org/abs/1309.5310

[54] M. Mishali and Y. C. Eldar, "Reduce and boost: Recovering arbitrary sets of jointly sparse vectors," *IEEE Trans. Signal Processing*, vol. 56, no. 10, pp. 4692–4702, Oct. 2008.

[55] W. U. Bajwa, R. Calderbank, and D. Mixon, "Two are better than one: Fundamental parameters of frame coherence," *Appl. Comput. Harmon. Anal.*, vol. 33, no. 1, pp. 58–78, July 2012.

[56] M. Rudelson and R. Vershynin, "On sparse reconstruction from Fourier and Gaussian measurements," *Commun. Pure Appl. Math.*, vol. 61, no. 8, pp. 1025–1045, Aug. 2008.

[57] N. Rao, B. Recht, and R. Nowak, "Universal measurement bounds for structured sparse signal recovery," in *Int. Conf. Artificiall Intelligence and Statistics (AISTATS)*, La Palma, Spain, Apr. 2012, pp. 942–950.

[58] E. van den Berg and M. P. Friedlander, "SPGL1: A solver for large-scale sparse reconstruction," June 2007, http://www.cs.ubc.ca/labs/scl/spgl1.

[59] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Statist. Soc B*, vol. 58, no. 1, pp. 267–288, 1996.

[60] E. J. Candès and Y. Plan, "Near-ideal model selection by ℓ_1 minimization," *Annals of Statistics*, vol. 37, no. 5A, pp. 2145–2177, Oct. 2009.

[61] S. Wright, R. Nowak, and M. A. T. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. Signal Processing*, vol. 57, no. 7, pp. 2479–2493, July 2009.