On Rank Awareness, Thresholding, and MUSIC for Joint Sparse Recovery

Jeffrey D. Blanchard[a,1,*], Caleb Leedy[a,2], Yimin Wu[a,2]

[a]*Department of Mathematics and Statistics, Grinnell College, Grinnell, IA 50112*

Abstract

This letter establishes sufficient conditions for the sparse multiple measurement vector (MMV) or row-sparse matrix approximation problem for the Rank Aware Row Thresholding (RART) algorithm. Using the rank aware selection operator to define RART results in discrete MUltiple SIgnal Classification (MUSIC) from array signal processing. When the sensing matrix is drawn from the random Gaussian matrix ensemble, we establish that the rank of the row-sparse measurement matrix in the noiseless row-sparse recovery problem allows RART (MUSIC) to reduce the effect of the $\log(n)$ penalty term that is present in traditional compressed sensing results and simultaneously provides a row-rank deficient recovery result for MUSIC. Empirical evidence shows that Thresholding closes matches RART in successful row-sparse approximation. The theoretical and emprical evidence provides further support for the conjecture that the thresholding operator in more sophisticated greedy algorithms is the source of their observed rank awareness.

Keywords: Compressed Sensing, Multiple Measurement Vectors, Row-sparse Approximation, Rank Aware, Thresholding, MUSIC, Joint Sparsity
2000 MSC: 68Q25, 68R10, 68U05

1. Introduction

Many authors have noted that increasing the number of observed signals, vectors, or columns leads to improved recovery rates for algorithms solving the row-sparse approximation problem, i.e. the multiple measurement vector or joint sparsity problems in compressed sensing. These observations also match intuition that adding more information to a problem should make it easier to solve. Repeatedly sensing the same signal or adding several copies of a single column to a matrix provides no new information, while sensing distinct signals or adding independent columns certainly should. Many sophisticated greedy algorithms for row-sparse approximation exhibit very profound performance gains when solving the row-sparse approximation problem with many independent columns in the row-sparse matrix. These algorithms all employ a thresholding operation to enforce a sparsity constraint on the solution; it was conjectured in [1] that sophisticated greedy algorithms are inherently rank-aware in that the thresholding operator is itself rank aware.

In this letter, we show that the simple hard thresholding algorithm can be slightly modified to be explicitly rank aware. Moreover, for Gaussian sensing matrices we detail the nature of the relationship between the rank of the row-sparse matrix and the number of measurements necessary for successful recovery. We then provide empirical evidence that aligns with the theory for Gaussian matrices followed by observing that the rank aware behavior is typically matched by the original hard thresholding algorithm. Finally, we show that the empirical behavior also applies in a few other scenarios. Together, the theoretical result for Rank Aware Row Thresholding (RART) and the similar observed behavior of standard Thresholding provide new insight to the observed behavior of the more sophisticated algorithms that rely on thresholding in their iterations. Moreover, the rank aware modification of thresholding provides a direct connection to the well studied discrete MUSIC algorithm and fills in a theoretical gap.

*Corresponding Author
Email address: `jeff@math.grinnell.edu` (Jeffrey D. Blanchard)
[1]JDB was supported by NSF DMS 1620390.
[2]CL and YW were supported by the Grinnell College MAP program.

1.1. *Compressed Sensing and Sparse Approximation*

The compressed sensing problem is now well known [2, 3, 4]. Let $x \in \mathbb{R}^n$ be a sparse vector with only $k < n$ nonzero entries. Let $A \in \mathbb{R}^{m\times n}$ be the sensing matrix used to produce the measurements $y = Ax \in \mathbb{R}^m$ with $k < m < n$. There are of course infinitely many solutions to the under-determined system $y = Ax$ so the compressed sensing problem asks for a solution with the fewest nonzero entries, i.e. the sparsest vector $\hat{x}$ with

$$\hat{x} = \arg\min_{z\in\mathbb{R}^n} \|z\|_0 \text{ subject to } y = Az \tag{1}$$

where $\|z\|_0$ counts the number of nonzero entries in z. This problem is known to be NP Hard in general [5]. For most applications with noise in the measurements or with only approximately sparse signals, the problem is better formulated as a *sparse approximation* problem. In this setting, one seeks a sparse vector which produces measurements that are closest to those observed through the sensing acquisition. The sparse approximation problem seeks the vector $\hat{x}$ with

$$\hat{x} = \arg\min_{z\in\mathbb{R}^n} \|y - Az\|_2 \text{ subject to } \|z\|_0 \leq k. \tag{2}$$

Two simple greedy algorithms designed to solve (2) are Orthogonal Matching Pursuit (OMP) and Thresholding (or one-step hard thresholding). OMP first selects the index of the column of A that is most correlated with the measurements y and then performs an orthogonal projection of the measurements onto the 1-sparse vector with this support. In each subsequent iteration of OMP, the index of the column of A most correlated to the remaining residual is added to the potential support set and an orthogonal projection of the measurements onto this support is performed. Thresholding on the other hand selects all k indices of the potential support simultaneously as the indices of the columns of A that are most correlated to the measurement vector y. Thresholding then projects the measurements y onto this proposed support.

OMP and Thresholding have been widely studied as their simplicity offers several avenues for careful analysis. In particular, both of these algorithms have worst case recovery guarantees [6, 7, 8] and average case guarantees [8, 9, 10, 11]. For example, consider our sparse approximation problem with $x \in \mathbb{R}^n$ a k-sparse Gaussian vector (x has exactly k nonzeros drawn i.i.d. from $\mathcal{N}(0, 1)$), $A \in \mathbb{R}^{m\times n}$ is a Gaussian sensing matrix (all entries of A are drawn i.i.d. from $\mathcal{N}(0, m^{-1})$), and $y = Ax$. Then, an average case analysis of OMP [10, 11] tells us that if OMP returns the vector $\hat{x}$, then there exists a constant c such that

$$\mathbb{P}(\hat{x} = x) \geq 1 - n^2 e^{-c\frac{m}{k}}. \tag{3}$$

Thus, with probability exceeding $1 - \epsilon$ OMP will exactly recover x from y and A provided

$$m \geq Ck \log(n/\epsilon) \tag{4}$$

for some universal constant $C = \frac{2}{c}$.

Let $\hat{x}$ be the approximation from Thresholding and let $\mu_2(k) = \max_{S;|S|=k} \max_{j\notin S} \|A_S^T A_j\|_2$ where A_S is the submatrix of A containing only the columns indexed by S and A_j is the jth column of A. An average case analysis of Thresholding [9] provided that the probability of Thresholding exactly recovering x from y and A is given by

$$\mathbb{P}(\hat{x} = x) \geq 1 - 2ne^{-C\left(\frac{|x_{\min}|^2}{|x_{\max}|^2}\right)\frac{1}{\mu_2^2(k)}}, \tag{5}$$

where $x_{\min}$ and $x_{\max}$ denote the nonzero coefficients of x with smallest and largest absolute values, respectively. The number of measurements m in relation to n and k are hidden within the function $\mu_2(A)$. These average case results are supported by empirical observations and show that the number of required measurements for successful recovery is proportional to the number of nonzeros in the vector times a log penalty in the dimension of the vector. The essential difference in the result for thresholding (5) and the result for OMP (3) is the introduction of a quantity measuring the dynamic range of the entries in A^*Ax which has been bounded by $\frac{|x_{\min}|^2}{|x_{\max}|^2}\frac{1}{\mu_2^2(k)}$.

1.2. Row-sparse Approximation

In this letter, we consider *row-sparse* approximation where $X \in \mathbb{R}^{n\times r}$ is a desired matrix, $A \in \mathbb{R}^{m\times n}$ is the sensing matrix producing the measurements $Y = AX \in \mathbb{R}^{m\times r}$, and $\|X\|_{R0}$ counts the number of rows in X with at least one nonzero entry. The row-support of X, $S = \text{rowsupp}(X)$, is the index set $S \subset \{1, 2, \dots, n\}$ of rows containing a nonzero entry in any column. X is *row-sparse* if $\|X\|_{R0} = |S| = k \ll n$. The row-sparse approximation problem then searches for the matrix $\hat{X}$ satisfying

$$\hat{X} = \arg\min_{Z\in\mathbb{R}^{n\times r}} \|Y - AZ\|_2 \;\text{ subject to }\; \|Z\|_{R0} \leq k. \tag{6}$$

When X is row-sparse, we can interpret X as a collection of *jointly sparse* column vectors and the measurements Y as a *multiple measurement vector* (MMV). It is clear that when $r = 1$ in (6) we have a single measurement vector (SMV) problem that is equivalent to the standard sparse approximation problem (2). If S is the row support of X with $|S| = k$, we let $X_{(S)} \in \mathbb{R}^{k\times r}$ denote the submatrix of X consisting only of the nonzero rows of X. Therefore, recalling that A_S is the submatrix of A consisting of the columns of A indexed by the set S, we have that $Y = AX = A_S X_{(S)}$.

There has been considerable work on the row-sparse approximation problem under the names of the MMV problem and the joint sparsity problem. The majority of this work considers the row-sparse approximation problem (6) as an extension of the sparse approximation problem (2) rather than initially studying (6) with (2) as a special case. This led to a large focus on extending existing SMV algorithms to the MMV case following the path of the compressed sensing literature. For example, much work has been done on the row-sparse approximation problem via convex relaxations [12, 13, 14, 15, 16] and greedy algorithms [1, 8, 17, 18, 19, 20, 21, 22, 23]. Empirical investigations of row-sparse approximation extensions of sophisticated greedy algorithms for compressed sensing advocate for using Conjugate Gradient Iterative Hard Thresholding (CGIHT) with restarting on row-sparse approximation problems [1, 22, 23].

1.3. Rank Awareness in Row-sparse Approximation

Although the concepts of row-sparse approximation appeared in the context of array-signal processing prior to the popular onset of compressed sensing, e.g. [24, 25], the apparent focus on extending SMV algorithms to the row-sparse approximation problem resulted in the overwhelming majority of algorithms for row-sparse approximation failing to explicitly exploit the additional information contained in multiple columns of the target matrix. The row support of the target matrix X, namely $S = \text{rowsupp}(X)$, ensures that each column, X_j, of X will have $\text{supp}(X_j) \subset S$, so that the measurements Y have several distinct encodings of S when the columns of X are independent. It would be ideal to use this additional information in the reconstruction process.

One approach exploiting rank information was a discrete version of the **MU**ltiple **SI**gnal **C**lassification (MUSIC) algorithm [24, 25, 26] which preceded the wide attention given to compressed sensing techniques. The discrete MUSIC algorithm provides optimal recovery guarantees [26] under mild conditions on the sensing operator A whenever the row-sparse matrix has full row-rank, i.e. whenever $r = \text{rank}(X) = |\text{rowsupp}(X)| = k$. We let $S = \text{rowsupp}(X)$ and observe first that $Y = AX = A_S X_{(S)}$ so that the range or column space of Y must be contained in the column space of A_S, i.e. $\text{col}(Y) \subset \text{col}(A_S)$. In the case with $\text{rank}(Y) = k$, we must have that $\text{col}(Y) = \text{col}(A_S)$. In other words, for all $i \in S$ we have $A_i \in \text{col}(Y)$. Hence, Discrete MUSIC exploits the structure of the column space[3] of Y by finding an orthonormal basis U of $\text{col}(Y)$ and then searching for the columns of A that are most correlated with this basis. As presented in [27], the Discrete MUSIC algorithm will select a threshold Θ_k so that $\{i : \|A_i^* U\|_2 > \Theta_k\}$ will be the index set of k columns of A most correlated with U. In this setting, MUSIC identifies the row support of X and therefore a simple projection of Y onto the support reconstructs the row sparse matrix X. The mild assumptions on A are needed to ensure that whenever $\text{rank}(X) = k$, we can ensure $\text{rank}(Y) = \text{rank}(AX) = k$. In this full row-rank case, MUSIC is guaranteed to reconstruct the row sparse matrix X whenever $m \geq k + 1$. However, in the row-rank deficient case $r < k$, we are unaware of any recovery guarantees given for the discrete MUSIC algorithm.

The lack of rank aware recovery guarantees in the row-rank deficient setting inspired investigations into algorithms whose success in solving (6) is directly related to the rank of the target matrix; this setting was expertly captured as *rank awareness* by Davies and Eldar [27]. In [27], the authors clearly demonstrate the distinction between rank aware

[3] Schmidt's original description of MUSIC [24] focused on determining vectors orthogonal to the noise space (null space) while solving a direction of arrival problem.

and non rank aware algorithms. They establish that the row-sparse approximation extension of OMP is not rank aware, introduce rank aware variants of OMP, and provide sufficient conditions for successful row-sparse recovery using these rank aware versions of OMP. The rank awareness of these algorithms comes from finding an orthonormal basis for the column space of the residual. After iteration $j-1$, with S^{j-1} the current index set for the support, the residual is defined to be $R^{j-1} = Y - AX^{j-1}$ with $X^{j-1}_{(S_{j-1})} = A^{\dagger}_{S^{j-1}}Y$ and $X^{j-1}_{(S^C_{j-1})} = 0$. Rather than selecting the next index based on the correlation of the columns of A with the residual, the selection is based on the correlation of the columns of A with an orthonormal basis of the column space of the residual, namely $U^j = \mathrm{orth}(R^{j-1})$. One then performs the *rank aware selection*

$$S^j = S^{j-1} \cup \arg\max_{i \in S^C} \|A_i^* U^j\|_2 \quad \text{with } U^j = \mathrm{orth}(R^{j-1}).$$

This rank aware greedy selection, clearly inspired by the orthogonalization in MUSIC, is the heart of the iterations of Rank Aware Orthogonal Matching Pursuit (RA-OMP) and Rank Aware Order Recursive Matching Pursuit (RA-ORMP) [27]. The selection process in RA-OMP can ultimately cause a reduction in the rank of the residual, i.e. rank degradation. The rank degradation can be overcome in two ways: employing RA-ORMP at an additional computational cost or using RA-OMP for the initial iterations and then employing Subspace Augmented MUSIC (SA-MUSIC) [28, 29]. In both cases, Blanchard and Davies [30] provided an average case analysis for Gaussian sensing matrices which reveals the relationship between the rank of the target matrix and the number of measurements required to ensure exact reconstruction. While the average case analysis for OMP can be easily extended to the SOMP algorithm to establish a sufficient number of measurements $m \geq Ck(\log(n/\epsilon) + r)$ as in (3), it was shown in [30] that the rank aware algorithms RA-OMP+SA-MUSIC and RA-ORMP each have a (distinct) constant in a probabilistic recovery guarantee defining the number of required measurements. Suppose $X \in \mathbb{R}^{n\times r}$ is a k-row-sparse matrix whose matrix of nonzero rows is in general position, $A \in \mathbb{R}^{m\times n}$ is a Gaussian sensing matrix, and $Y = AX$; then with probability exceeding $1-\epsilon$ there exists a constant C so that RA-OMP+SA-MUSIC (RA-ORMP) will exactly recover X from Y and A provided

$$m \geq Ck\left(\frac{\log(n/\epsilon)}{r} + 1\right). \tag{7}$$

Note the relationship between the bound on the number of measurements for OMP given in (4) and that of the rank aware version given in (7): the rank of X dampens the impact of the logarithmic factor in n. In other words, when $\mathrm{rank}(X) \approx \log(n)$ the number of measurements required for accurate reconstruction is linearly proportional to the row-sparsity of the matrix. This is optimal.

The rank aware selection of RA-OMP is clearly related to the support selection technique for the MUSIC algorithm. If one was unfamiliar with MUSIC, they might naturally define Rank Aware Row Thresholding (RART) by extending the rank aware selection of RA-OMP to select the index set for the k largest row-ℓ_2 norms of A^*U where $U = \mathrm{orth}(Y)$. Thus, by avoiding the need to find a basis for the residual and solve an optimization in each of the k iterations as required by RA-OMP and RA-ORMP, Rank Aware Row Thresholding (RART), Algorithm 1, requires substantially less computation than the rank aware versions of OMP at the cost of a lower rate of successful recovery. Moreover, in the full rank case, Rank Aware Row Thresholding *is* the discrete MUSIC algorithm.

Algorithm 1 Rank Aware Row Thresholding / MUSIC

Input: Y the measurements, A the sensing matrix, k the sparsity level
Output: $\hat{X}$ a row-sparse matrix

1: $U = \mathrm{orth}(Y)$ (compute an orthonormal basis of the measurement space)
2: $w = \mathrm{row}\ell_2(A^*U)$ (compute the row ℓ_2 norms of the residual)
3: $T = \mathrm{IndexTopk}(w)$ (identify the residual matrix rows with the largest row-ℓ_2 norms)
4: $\hat{X}_{(T)} = A_T^{\dagger}Y$ and $\hat{X}_{(T^c)} = 0$ (create row-sparse solution with row-support T)

When the rank of the matrix is as large as the row support, i.e. $X_{(S)}$ has full row-rank, and $m \geq k+1$, then RART/MUSIC (Alg. 1) is guaranteed exact recovery [26]. However, when $X_{(S)}$ does not have full row-rank, i.e. if $r < k$, no such guarantee is available. One response in the literature was the introduction of the SA-MUSIC and rank aware OMP algorithms. In this letter, we specifically investigate Rank Aware Row Thresholding to provide a recovery guarantee in the rank deficient case and in turn establish a rank deficient recovery result for MUSIC.

The main result of this letter is the average case analysis of RART which establishes precisely how the rank awareness of the algorithm impacts the required number of measurements. The difference between our result and that for RA-OMP+SA-MUSIC or RA-ORMP taken from [30] is the introduction of a measure of the dynamic range of A^*AX. This difference is analogous to the difference in average-case results for OMP and Thresholding, in that they differ primarily by the introduction of a measure of the dynamic range of A^*Ax in the form of the quantity $\frac{|x_{min}|}{|x_{max}|}\frac{1}{\mu_2^2(A)}$. The new measure of dynamic range is a ratio describing the range of values in the inner products of the sensing matrix A with an orthonormal basis for the span of the columns of the measurements $Y = AX$.

Definition 1.1. *Let $A \in \mathbb{R}^{m\times n}$, $X \in \mathbb{R}^{n\times r}$ with $S = rowsupp(X)$, and $U = orth(AX)$. Define the* orthogonalized coherence of A with respect to X *by*

$$\nu_X(A) = \frac{\min_{i\in S} \|A_i^* U\|_2^2}{\max_{i\in S} \|A_i^* U\|_2^2}.$$

Remark 1.2. *The orthogonalized coherence of A with respect to X in Definition 1.1 is akin to the ratio $\frac{|x_{min}|^2}{|x_{max}|^2} \cdot \frac{1}{\mu_2^2(k)}$ in the average case result on thresholding in* (5)*. Some measurement of the dynamic range of X and the interaction of the columns of A appear in all results on thresholding, in particular [8, 9]. The introduction of an orthonormal basis of the column space of $Y = AX$ presents difficulty in decoupling the entries of X from the columns of A. Rather than introduce other forms of coherence or complicate statements by computing a specific orthonormal basis stored in the matrix U, we employ $\nu_X(A)$ as the precise quantity we need. The quantity $\nu_X(A)$ may be of its own theoretical interest, but this letter will omit further analysis.*

The formal statement developed in the next section is presented as Theorem 2.5. Let $r < k < m < n$ and suppose we have a row-sparse matrix $X \in \mathbb{R}^{n\times r}$ with $S = \text{rowsupp}(X)$, $|S| = k$, and $X_{(S)}$ is in general position (i.e. every set of r rows of $X_{(S)}$ is linearly independent). If $A \in \mathbb{R}^{m\times n}$ is drawn from the family of Gaussian matrices, then Rank Aware Row Thresholding will recover X from the measurements $Y = AX$ with probability at least $(1 - \epsilon)$ provided

$$m \geq \frac{C}{\nu_X(A)} k \left(\frac{\log(n/\epsilon)}{r} + 1 \right). \tag{8}$$

Here, as with all probability statements in this letter, the probability is on the draw of the sensing matrix A from the family of $m \times n$ Gaussian matrices, i.e. $m \times n$ matrices with entries drawn i.i.d. from the normal distribution $\mathcal{N}(0, \frac{1}{m})$.

Similar to the average case analysis in [8, 9, 10, 11], we focus the analysis on RART with Gaussian sensing matrices A. For clarity, our analysis also assumes there is no noise in the measurements just as in [30]. It is possible that a much longer analysis could provide results for other sensing matrix ensembles and in the presence of noise. This letter is focused on clearly demonstrating the role of rank in determining the required number of measurements for RART. As seen in (8), if the sensing matrix is Gaussian and the measurements are uncorrupted, the rank of the row-sparse matrix dampens the logarithmic penalty term essentially eliminating it for modest ranks of $r \approx \log(n)$.

2. Average Case Analysis of RART

In this section, we establish that to recover a row-sparse matrix of rank r, the logarithmic penalty in the number of Gaussian sensing measurements is reduced by the factor r^{-1}. Before proving this result, we recall or establish a series of intermediate results. First we translate the question of exact recovery into a statement about the ℓ_2 norms of the interaction of the sensing matrix A with an orthonormalized version of the measurements $Y = AX$. From the definition of Rank Aware Row Thresholding, Alg. 1, Step 3, the following lemma is immediate.

Lemma 2.1. *Suppose $A \in \mathbb{R}^{m\times n}$, $X \in \mathbb{R}^{n\times r}$ with $S = rowsupp(X)$, $|S| = k$, $Y = AX \in \mathbb{R}^{m\times r}$, and $U = orth(Y) = orth(AX)$. Let $\hat{X} = RART(A, Y, k)$ be the output of Rank Aware Row Thresholding given the inputs A, Y, and k. Then $\hat{X} = X$ if and only if $\max_{i'\in S^c} \|A_{i'}^* U\|_2 < \min_{i\in S} \|A_i^* U\|_2$.*

Proof. If the output $\hat{X}$ has the same support set as the original row-sparse matrix X, the projection step ensures that $\hat{X} = X$. RART will select the correct support set S exactly when $\max_{i'\in S^c} \|A_{i'}^* U\|_2 < \min_{i\in S} \|A_i^* U\|_2$. □

We now recall two important results from related work on the rank aware orthogonal matching pursuit algorithms presented in [30]. First we bound the inner products of any subset of the columns of a sensing matrix A with any matrix with normalized columns.

Lemma 2.2 ([30], Lem. 2)**.** *Let $A \in \mathbb{R}^{m\times n}$ and $S \subset \{1,\dots,n\}$ with $|S| = k < m$. Suppose $U \in \mathbb{R}^{m\times r}$ with $\|U_i\|_2 = 1$ and $U_i \in range(A_S)$ for all $i = 1,\dots,r$. Let $\alpha = \sigma_{min}(A_S)$ be the smallest singular value of A_S. Then*

$$\max_{i\in S} \|A_i^* U\|_2 \geq \alpha\sqrt{\frac{r}{k}}. \tag{9}$$

This lower bound involves the smallest singular value of the sub-matrix A_S where S contains k indices; it holds for any matrix A. We now need to control the size of these vector-matrix products for the columns of A that are not indexed by S. When A is a Gaussian sensing matrix and $i \notin S$, the Laplace transform method was employed to establish a lower bound on the probability that the maximum ℓ_2 norm of the vector-matrix product exceeds a value β^2.

Lemma 2.3 ([30], Lem. 3)**.** *If $A \in \mathbb{R}^{m\times n}$ with entries drawn i.i.d. from $\mathcal{N}(0, m^{-1})$, $S \subset \{1,\dots,n\}$ is an index set with $|S| = k$, and $U \in \mathbb{R}^{m\times r}$ is a matrix with orthonormal columns, $rank(U) = r$, and $span(U) \subset span(A_S)$, then*

$$\mathbb{P}\left\{\max_{i\in S^c} \|A_i^* U\|_2^2 < \beta^2\right\} \geq 1 - (n-k)e^{-(m\beta^2-2r)/4}. \tag{10}$$

With these results in hand from [30], we can extend the analysis to Rank Aware Row Thresholding. To analyze the average case behavior of RART, the bound in Lemma 2.2 is insufficient. Instead, we must provide a lower bound on the minimum inner product between the subset of the columns of A and the normalized matrix U. This is the role of our new measure $\nu_X(A)$ given in Definition 1.1.

Corollary 2.4. *Let $A \in \mathbb{R}^{m\times n}$ and $S \subset \{1,\dots,n\}$ with $|S| = k < m$. Suppose $U \in \mathbb{R}^{m\times r}$ with $\|U_i\|_2 = 1$ and $U_i \in range(A_S)$ for all $i = 1,\dots,r$. Let $\alpha = \sigma_{min}(A_S)$ be the smallest singular value of A_S. Then*

$$\min_{i\in S} \|A_i^* U\|_2^2 \geq \nu_X(A)\alpha^2\frac{r}{k}. \tag{11}$$

Proof. We can simply employ Definition 1.1 to see that

$$\min_{i\in S} \|A_i^* U\|_2^2 = \nu_X(A)\max_{i\in S} \|A_i^* U\|_2^2. \tag{12}$$

Therefore, applying the lower bound from Lemma 2.2, we have (11). □

With Lemma 2.1 and Corollary 2.4, we are now ready to establish a rank aware recovery guarantee for Rank Aware Row Thresholding. As mentioned previously, this also serves as a rank deficient recovery guarantee for discrete MUSIC. We consider the case of $k > r$ since the results of Feng and Bressler [26] in the full row rank case, $k = r$, only require that $m \geq k + 1$.

Theorem 2.5 (RART Recovery)**.** *Let $X \in \mathbb{R}^{n\times r}$, $S = rowsupp(X)$ with $|S| = k > r$, and assume $X_{(S)}$ is in general position. Let $A \in \mathbb{R}^{m\times n}$ be a random matrix independent of X with i.i.d. entries $A_{ij} \sim \mathcal{N}(0, m^{-1})$. Then for some constant C and with probability at least $1-\epsilon$, Rank Aware Row Thresholding will exactly recover X from $Y = AX$ if*

$$m \geq \frac{C}{\nu_X(A)}k\left(\frac{\log(n/\epsilon)}{r} + 1\right). \tag{13}$$

Proof. Let $\hat{X}$ be the output of RART given Y, A, k and let $U = \text{orth}(Y) = \text{orth}(AX)$. Combining Lemma 2.1 and Corollary 2.4 we see that

$$\begin{aligned}\mathbb{P}\left\{\hat{X} = X\right\} &= \mathbb{P}\left\{\max_{i\in S^c} \|A_i^* U\|_2^2 < \min_{i\in S} \|A_i^* U\|_2^2\right\} \\ &\geq \mathbb{P}\left\{\max_{i\in S^c} \|A_i^* U\|_2^2 < \nu_X(A)\alpha^2\frac{r}{k}\right\}. \end{aligned} \tag{14}$$

Applying Lemma 2.3 with $\beta^2 = \nu_X(A)\alpha^2\frac{r}{k}$ to (14), we then have

$$\mathbb{P}\left\{\hat{X} = X\right\} \geq 1 - (n-k)e^{-(m\nu_X(A)\alpha^2\frac{r}{k}-2r)/4}. \tag{15}$$

From [10, 31] we know that there exists a constant $c > 0$ such that $\mathbb{P}\{\alpha > \frac{1}{\sqrt{2}}\} \geq 1 - e^{-cm}$ since A is Gaussian. Applying this result to (15), we arrive at the following lower bound:

$$\begin{aligned}
\mathbb{P}\left\{\hat{X} = X\right\} &\geq \left(1 - (n-k)e^{-(m\nu_X(A)(\frac{1}{2})\frac{r}{k}-2r)/4}\right)\cdot(1 - e^{-cm}) \\
&\geq 1 - (n-k)e^{-(m\nu_X(A)\frac{r}{k}-4r)/8} - e^{-cm} \\
&\geq 1 - ne^{-\bar{C}(m\nu_X(A)\frac{r}{k}-4r)} \qquad (16)
\end{aligned}$$

for an appropriately chosen constant $\bar{C}$.

Now letting $\epsilon \geq ne^{-\bar{C}(m\nu_X(A)\frac{r}{k}-4r)}$ and isolating m while letting $C = \min(4, \frac{1}{\bar{C}})$, we arrive at (13).

□

The clear difference between the bound on rank aware OMP algorithms and RART is the inclusion of $\nu_X(A)$. However, this is analogous to such differences for nearly all comparable statements regarding OMP and thresholding. While the orthogonalization of the columns of the measurements Y inhibits our ability to decouple X and A as was possible in the average case analysis of Thresholding (5) [9], it is intuitively clear that the dynamic range of the nonzeros entries in X will have an impact on Row Thresholding and Rank Aware Row Thresholding. The relationship between (7) and (13) is unsurprising when considering Remark 1.2.

3. Average-case Performance of RART

The numerical experiments in this section were conducted on a machine equipped with 4 Intel Xeon CPU E5-2637 v2 @ 3.50 GHz running MATLAB on Debian 8 (jessie).

3.1. Observed Average Case Performance of RART

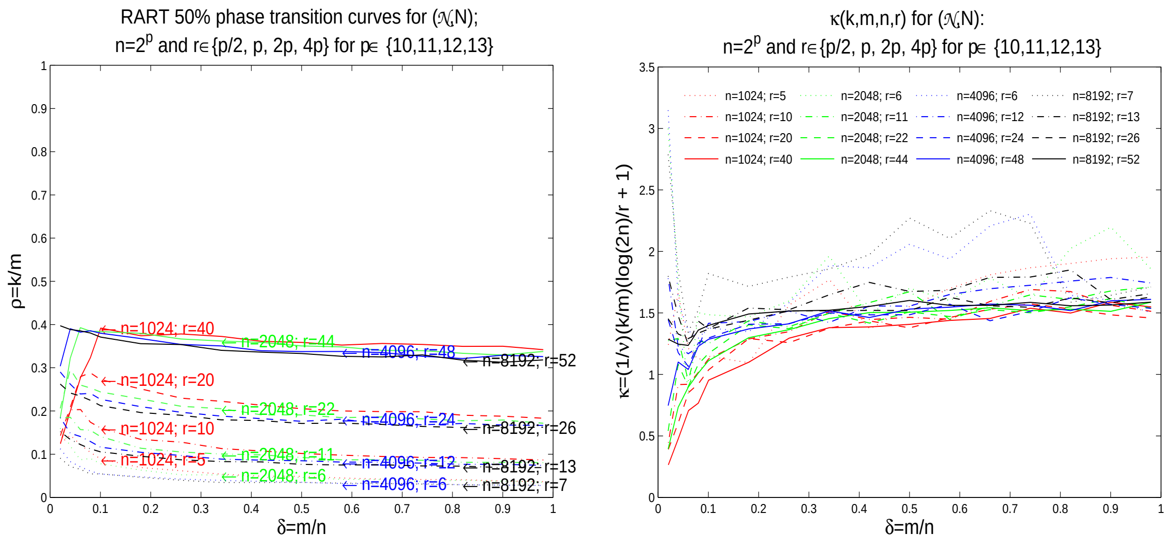
For comparison across problem dimensions and to existing literature, we describe algorithm success in terms of the phase transition framework. Consider the unit square as the $(\delta, \rho) = (\frac{m}{n}, \frac{k}{m})$ phase space where a phase transition curve for an algorithm and given rank will divide the phase space into two regions of success (below the curve) and failure (above the curve) in the row-sparse approximation problem (6). To match the theoretical analysis in this letter we present empirical testing on the problem class $(\mathcal{N}, N)$ of Gaussian sensing matrices A and Gaussian target matrices X: the sensing matrix A has it entries drawn i.i.d. from the normal distribution $\mathcal{N}(0, m^{-1})$ and the target matrix X has its row support selected uniformly at random and the nonzero entries are drawn i.i.d. from the normal distribution $\mathcal{N}(0, 1)$. For a given problem instance, $Y = AX$ we define a successful recovery to be an exact identification of the row-support[4], rowsupp(X).

For testing the row-sparse approximation problem (6), the parameter $\delta \in (0, 1)$ takes on fifteen values

$$\delta \in \{0.02, 0.04, 0.06, 0.08, 0.1, 0.18, 0.26, \ldots, 0.98\} \tag{17}$$

with 7 additional uniformly spaced values of δ between 0.26 and 0.98. For a given value of n, we define $m = \lceil \delta n \rceil$ for each of the 15 values of delta. We are then in search of the constant $\rho = \frac{k}{m}$ which approximates the location in the phase space where the algorithm abruptly switches from always succeeding to always failing.

Given (n, r, δ), the empirical recovery phase transitions are identified by first determining a phase transition region $[k_{min}, k_{max}]$ where an algorithm successfully recovers each of 10 random vectors at k_{min} but fails to recover any of 10 random vectors at k_{max}. The phase transition region is then extensively tested at $min\{50, k_{max} - k_{min}\}$ values of k uniformly spaced in $[k_{min}, k_{max}]$. For each problem size (n, m, k, r), 10 tests were conducted and the *weak recovery phase transition*, $\hat{\rho}(n, r, \delta)$, is the curve defined by a logistic regression for a 50% success rate of exact support identification.



(a) (b)

Figure 1: RART for $n = 2^p$ and $r \in \{\lceil \frac{p}{2} \rceil$ (dotted), p(dot-dashed), $2p$(dashed), $4p$(solid)$\}$ with $p \in \{10$(red), 11(green), 12(blue), 13(black)$\}$: (a) The observed 50% phase transition curves exhibiting rank awareness; (b) the nearly constant behavior of $\kappa(n, r, \delta)$ as predicted by the analysis leading to (18).

The *recovery region* is the region below the recovery phase transition curve, $\hat{\rho}(n, r, \delta)$, and identifies the region of the phase space where an algorithm is typically successful when given a problem from the problem class $(\mathcal{N}, N)$.

In Figure 1(a) we see the empirical 50% phase transition curves for $n = 2^p$ with $p \in \{10, 11, 12, 13\}$ and four distinct ranks, $r = \lceil \alpha p \rceil$ for each $\alpha \in \{\frac{1}{2}, 1, 2, 4\}$. We can observe the clear increase in the size of the phase transition region as the rank increases. For the problem parameters $(n, r) = (2^p, \lceil \alpha p \rceil)$ for each fixed $\alpha \in \{\frac{1}{2}, 1, 2, 4\}$ we observe a general consistency across the varying values of $p = 10, 11, 12, 13$. However, for each rank the empirical 50% phase transition curve $\hat{\rho}(n, r, \delta)$ is decreasing in n. This behavior aligns with the dependence of RART on the orthogonalized coherence of the sensing matrix with respect to the target matrix.

For an interpretation related to the full row rank guarantees of MUSIC, consider a row-sparse matrix X with 520 nonzero rows of the 8192 rows so that $(n, k) = (8192, 520)$. If X also has 520 independent columns, then MUSIC (RART) is guaranteed to recover X as long as $m \geq 521$. From Figure 1(a), we observe that if X has only 52 linearly independent columns and we have taken 1733 measurements, then we can recover X using RART (MUSIC); in this case we have $(k, m, n) = (520, 1733, 8192)$ so that $(\delta, \rho) = (.21, .30)$ which lies clearly below the phase transition curve for this problem.

3.2. *Observed Accuracy of the Main Result*

The fact that Theorem 2.5 is valid does not necessarily imply that it is enlightening. However, if our observed average-case phase transitions occur roughly where our bound on m is enforced, then we can be confident that (13) is in fact descriptive of the rank awareness of RART. Enforcing the bound on the measurements in (13), Theorem 2.5 tells us RART should successfully recovery the row-sparse matrix with probability $1 - \epsilon$ whenever the number of measurements m satisfies

$$m \approx \frac{Ck}{\nu} \left(\frac{\log(n/\epsilon)}{r} + 1 \right) \quad \text{or} \quad \tilde{C} \approx \frac{1}{\nu} \frac{k}{m} \left(\frac{\log(n/\epsilon)}{r} + 1 \right) \tag{18}$$

[4]While successful recovery is typically defined in terms of mismatch error measured in Frobenius norm, this stricter support recovery definition is more aligned with the single iteration of thresholding which ignores the method for projecting the measurements onto the support.

where $\tilde{C} = \frac{1}{C}$. This implies for a given success probability $1 - \epsilon$, the right hand quantity in (18) should be essentially constant.

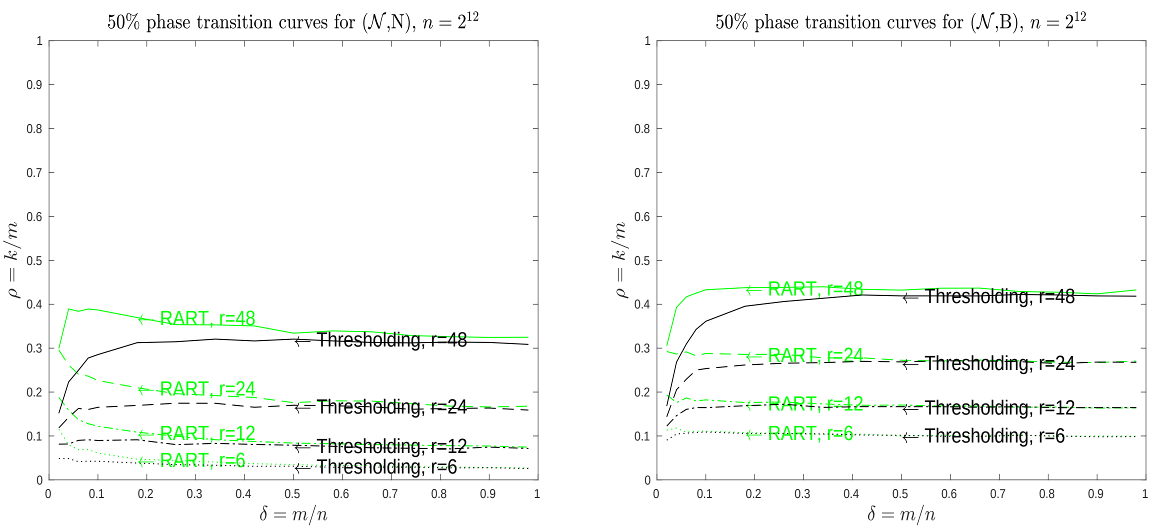
In Section 3.1 we have identified some observed, average-case 50% phase transition curves, $\hat{\rho}(n, r, \delta)$, for a selection of dimensions and ranks (n, r). For a 50% phase transition curve, we have $\epsilon = \frac{1}{2}$ so that $n/\epsilon = 2n$. Furthermore, we need a function $\nu(k, m, n, r)$ which will return the expected value of the orthogonalized coherence $\nu_X(A)$ for $X \in \mathbb{R}^{n\times r}$ and $A \in \mathbb{R}^{m\times n}$. While this may be of interest theoretically, its analysis is not necessary in this letter and we instead approximate this numerically with a function $\hat{\nu}(n, r, \delta)$. For each pair (n, r) we compute the mean value of $\nu_X(A)$ over 100 randomly generated problem instances in the mesh $\{(0.02i, 0.05 + .025j) \mid i = 1, \ldots, 49; j = 0, \ldots, 36\}$. The function $\hat{\nu}(n, r, \delta)$ will then return the empirical expected value of $\nu_X(A)$ from this data for (n, r) that is closest in ℓ_1 norm to the point $(\delta, \hat{\rho}(n, r, \delta))$ for the given value of δ. With this numerical approximation for $\nu_X(A)$ we can define a function that approximates the expected value of the right hand side of (18). Given (n, r) and the functions $\hat{\nu}$ and $\hat{\rho}$, the function

$$\kappa(n, r, \delta) = \frac{\hat{\rho}(n, r, \delta)}{\hat{\nu}(n, r, \delta)} \left(\frac{\log(2n)}{r} + 1 \right) \tag{19}$$

approximates the expected value of the right hand side of (18).

In Figure 1(b) we observe that for varying ranks, r, across various values of n, the function $\kappa(n, r, \delta)$ is roughly constant[5]. The constant behavior is enhanced for fixed n as r increases, and for fixed (n, r) as δ increases. Also, the range of values of $\kappa(n, r, \delta)$ for larger values of n is related to the larger range of values in the orthogonalized coherence, $1/\nu_X(A)$. The curves furthest from the concentration of curves come from the smallest tested ranks, $r = \left\lceil \frac{\log(n)}{2} \right\rceil$. Moreover, we observe that the constant, C, in (13) is likely rather small; from our experiments it appears to be less than 1 since Figure 1(b) depicts $\kappa \approx \tilde{C} = \frac{1}{C} > 1$ for most δ.

4. Observed Rank Awareness of Thresholding



(a) (b)

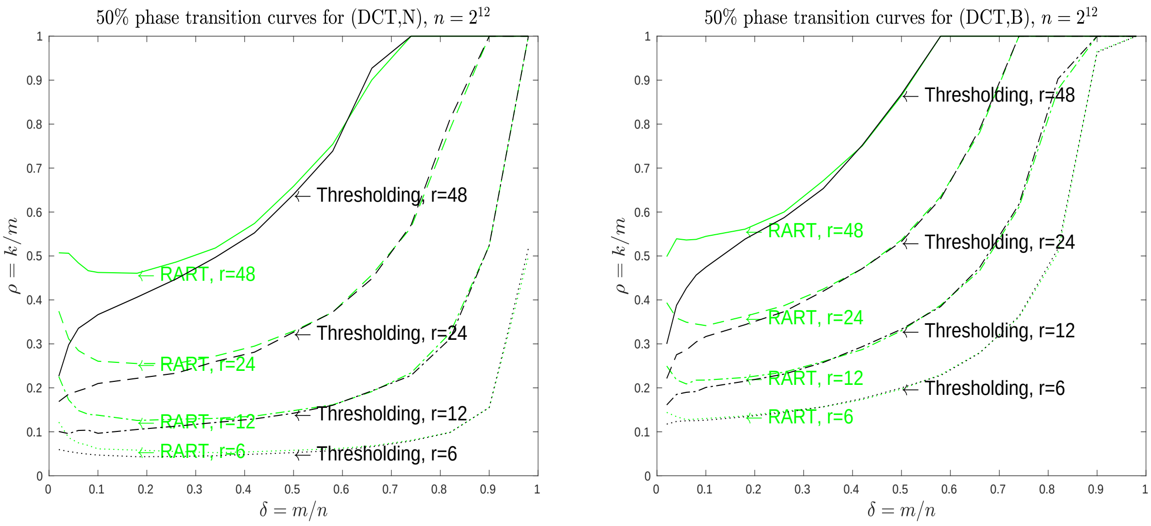
Figure 2: Thresholding and Rank Aware Row Thresholding: the observed rank awareness of thresholding when compared to RART for Gaussian sensing matrices. (a) The nonzeros of X are drawn i.i.d. from $\mathcal{N}(0, 1)$; (b) The nonzeros of X are drawn with equal probability from $\{-1, 1\}$.

In [1], the observed rank awareness of several more sophisticated greedy algorithms was conjectured to be related to an inherent rank awareness of the row thresholding operator. In a given iteration of those greedy algorithms, there

[5]The inclusion of $\hat{\nu}(n, r, \delta) \approx \nu_X(A)$ in $\kappa(n, r, \delta)$ is critical in this observed, nearly constant behavior.

is no explicit orthogonalization of the residual and therefore no specific rank aware step. However, an early analysis of Row Thresholding by Gribonval et al. [8] established an average case result [8, Theorem 4] which is dependent on the rank, r, of the target matrix $X \in \mathbb{R}^{n\times r}$. At the time, the authors of [8] were not investigating rank awareness and instead framed the appearance of the rank in the result as an obstacle to achieving an average case analysis of Thresholding for $r = 1$. While [8, Theorem 4] requires a rather large rank[6], it does in fact indicate thresholding is rank aware under their specific Gaussian signal model.

At the same time, the observed behavior shows similar rank awareness of Row Thresholding to that of RART. In Figure 2, the sensing matrices, A, are again all drawn from the matrix family $\mathcal{N}$, namely the entries are selected i.i.d from $\mathcal{N}(0, m^{-1})$. In the left panel, Figure 2(a), the target matrices are again taken from the matrix family N where their row support is uniformly selected and the nonzero entries are drawn i.i.d. from $\mathcal{N}(0, 1)$. In the right panel, Figure 2(b), the target matrices are taken from the binary family, B, where the uniformly chosen row support is populated with ± 1 with equal probability. Note that in both cases Thresholding exhibits clear rank aware behavior and its 50% phase transition curve closely tracks that of RART. The impact of the dynamic range of the entries on RART and Thresholding can be seen in the increased phase transitions for the problem class $(\mathcal{N}, B)$ in Figure 2(b) over $(\mathcal{N}, N)$ in Figure 2(a); all nonzero entries in a target matrix from the binary family have the same magnitude.



(a) (b)

Figure 3: Thresholding and Rank Aware Row Thresholding: the observed rank awareness of thresholding when compared to RART for sensing matrices formed by uniformly selecting m rows from an $n \times n$ Discrete Cosine Transform. (a) The nonzeros of X are drawn i.i.d. from $\mathcal{N}(0, 1)$; (b) The nonzeros of X are drawn with equal probability from $\{-1, 1\}$.

Figure 3 shows the performance of RART and Thresholding on two additional problem classes, (DCT,N) and (DCT,B). As above, the phase transitions are computed in precisely the same fashion though the sensing matrices, $A \in \mathbb{R}^{m\times n}$, are formed by uniformly selecting m rows from a discrete cosine transform. The target matrices are again drawn from N and B in panels (a) and (b), respectively. As shown in previous empirical studies [22, 23, 32], when used with these structured sensing matrices, greedy algorithms have improved phase transitions over greedy algorithms with Gaussian sensing matrices. Notice the apparent rank awareness of both RART and Thresholding persists for these structured sensing matrices. Though the theory of Section 2 does not apply to the DCT matrices, we can be confident in the average case rank awareness of RART and Thresholding with sensing matrices from this family. The improvement in performance for equal magnitude nonzeros is also readily apparent when comparing Figure 3(b) to 3(a).

[6] It is not as large as $r = k$ required by MUSIC.

In all four problem classes depicted in Figures 2 and 3, we see that the empirical 50% phase transition curves for Thresholding nearly match those of RART. While the phase transitions for RART are generally superior to those of thresholding, the advantage of the rank aware algorithm seems to increase as the rank increases and is most pronounced in the areas of extreme undersampling ($\delta \to 0$).

5. Conclusion

From Section 3 we observe that both RART and Row Thresholding perform surprisingly well on the row-sparse approximation problem in the row rank-deficient scenario. More sophisticated algorithms such as CoSaMP [33] and Conjugate Gradient Iterative Hard Thresholding [22] have been shown to be even more successful on the row-sparse approximation problem and are rather stable to additive noise [1, 22, 23]. The empirical performance of these greedy algorithms clearly exhibit rank aware behavior. In this letter we have provided evidence in support of the conjecture that the thresholding operator is rank aware and have provided an average case analysis detailing how the rank impacts a rank aware version of row thresholding.

The good news for practitioners is that the initialization phase of these algorithms executes Row Thresholding. If one were faced with a row-sparse approximation problem, they need not choose between Row Thresholding or CGIHT; they can simply run CGIHT Restarted knowing that the algorithm will succeed if Thresholding were to succeed and in fact will exactly execute Row Thresholding with a conjugate gradient projection. However, if the problem were one where Thresholding would fail, the ability of CGIHT to change the potential support set provides the opportunity to correct errors in the initial iteration. The rank awareness of Thresholding is at the very least crucial to providing CGIHT with a initial approximation that improves with rank.

These results bridge the divide between the success of MUSIC in the full rank case and the success of CGIHT and CoSaMP in the rank deficient setting. The relationship between the average case analysis of OMP and thresholding also extends to the rank aware versions of those algorithms in the form of RA-OMP + SA-MUSIC and RART. In both cases, the rank of the target matrix directly reduces the logarithmic penalty essentially admitting a linear relation between the number of measurements and the (row) sparsity of the target matrix for very modest ranks on the order of $\log(n)$.

References

[1] J. Blanchard, M. Cermak, D. Hanle, Y. Jing, Greedy algorithms for joint sparse recovery, IEEE Trans. Sig. Proc. 62 (7) (2014) 1694–1704.
[2] E. J. Candès, Compressive sampling, in: International Congress of Mathematicians. Vol. III, Eur. Math. Soc., Zürich, 2006, pp. 1433–1452.
[3] E. J. Candes, T. Tao, Decoding by linear programming, IEEE Trans. Inform. Theory 51 (12) (2005) 4203–4215.
[4] D. L. Donoho, Compressed sensing, IEEE Trans. Inform. Theory 52 (4) (2006) 1289–1306.
[5] B. K. Natarajan, Sparse approximate solutions to linear systems, SIAM J. Computing 24 (2) (1995) 227–234.
[6] M. Davenport, M. Wakin, Analysis of orthogonal matching pursuit using the restricted isometry property, IEEE Trans. Inform. Theory 56 (9) (2010) 4395–4401.
[7] J. A. Tropp, Greed is good: algorithmic results for sparse approximation, IEEE Trans. Inform. Theory 50 (10) (2004) 2231–2242.
[8] R. Gribonval, H. Rauhut, K. Schnass, P. Vandergheynst, Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms, J. Fourier Analysis and Applications 14 (2008) 655–687.
[9] K. Schnass, P. Vandergheynst, Average performance analysis for thresholding, IEEE Signal Processing Letters 14 (11) (2007) 828–831.
[10] J. A. Tropp, A. C. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit, IEEE Trans. Inform. Theory 53 (12) (2007) 4655–4666.
[11] A. K. Fletcher, S. Rangan, Orthogonal matching pursuit from noisy measurements: A new analysis, in: Proceedings of the 22Nd International Conference on Neural Information Processing Systems, NIPS'09, Curran Associates Inc., USA, 2009, pp. 540–548.
[12] J. Chen, X. Huo, Theoretical results on sparse representations of multiple-measurement vectors, IEEE Trans. Sig. Proc. 54 (12) (2006) 4634–4643.
[13] Y. Eldar, H. Rauhut, Average case analysis of multichannel sparse recovery using convex relaxation, IEEE Trans. Inform. Theory 56 (1) (2010) 505–519.
[14] M.-J. Lai, Y. Liu, The null space property for sparse recovery from multiple measurement vectors, Appl. Comp. Harmon. Anal. 30 (3) (2011) 402–406.
[15] J. A. Tropp, Algorithms for simultaneous sparse approximation. Part II: Convex relaxation, Signal Processing 86 (2006) 589–602.
[16] E. van den Berg, M. Friedlander, Theoretical and empirical results for recovery from multiple measurements, IEEE Trans. Inform. Theory 56 (5) (2010) 2516–2527.
[17] S. Foucart, Recovering jointly sparse vectors via hard thresholding pursuit, in: Proc. of SAMPTA, 2011, Online.
[18] D. Leviatan, V. N. Temlyakov, Simultaneous approximation by greedy algorithms, Adv. Comput. Math. 25 (1-3) (2006) 73–90.
[19] A. Lutoborski, V. N. Temlyakov, Vector greedy algorithms, J. Complexity 19 (4) (2003) 458–473.

[20] V. N. Temlyakov, A remark on simultaneous greedy approximation, East J. Approx. 10 (1-2) (2004) 17–25.
[21] J. A. Tropp, A. C. Gilbert, M. J. Strauss, Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit, Signal Processing 86 (2006) 572–588.
[22] J. Blanchard, J. Tanner, K. We, CGIHT: conjugate gradient iterative hard thresholding for compressed sensing and matrix completion, Information and Inference: a journal of the IMA 4 (4) (2015) 289–327.
[23] J. Blanchard, J. Tanner, K. We, Conjugate gradient iterative hard thresholding: observed noise stability for compressed sensing, IEEE Trans. Sig. Proc. 63 (2) (2015) 528–537.
[24] R. O. Schmidt, Multiple emitter location and signal parameter estimation, Proceedings of RADC Spectral Estimation Workshop (1979) 243–258.
[25] I. Gorodnitsky, B. Rao, Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm, IEEE Trans. Sig. Proc. 45 (3) (1997) 600–616.
[26] P. Feng, Y. Bresler, Spectrum-blind minimum-rate sampling and reconstruction of multiband signals, Proceedings of IEEE Intl. Conf. Acoustics, Speech, and Signal Processing (1996) 1688–1691.
[27] M. E. Davies, Y. C. Eldar, Rank awareness in joint sparse recovery, IEEE Trans. Inform. Theory 58 (2) (2012) 1135–1146.
[28] K. Lee, Y. Bresler, M. Junge, Subspace methods for joint sparse recovery, IEEE Trans. Inform. Theory 58 (6) (2012) 3613–3641.
[29] J. M. Kim, O. K. Lee, J. C. Ye, Compressive music: Revisiting the link between compressive sensing and array signal processing, IEEE Trans. Inform. Theory 58 (1) (2012) 278–301.
[30] J. Blanchard, M. Davies, Recovery guarantees for rank aware pursuits, IEEE Signal Processing Letters 19 (7) (2012) 427–430.
[31] R. Baraniuk, M. Davenport, R. DeVore, M. Wakin, A simple proof of the restricted isometry property for random matrices, Constructive Approximation 28 (3) (2008) 53–263.
[32] J. D. Blanchard, J. Tanner, Performance comparisons of greedy algorithms for compressed sensing, Numerical Linear Algebra with Applications 22 (2) (2015) 254–282.
[33] D. Needell, J. Tropp, CoSaMP: Iterative signal recovery from incomplete and inaccurate samples, Appl. Comp. Harmon. Anal. 26 (3) (2009) 301–321.