Performance Guarantees of the Thresholding Algorithm for the Co-Sparse Analysis Model

Tomer Peleg *Student Member, IEEE* and Michael Elad, *Fellow, IEEE*

Abstract

The co-sparse analysis model for signals assumes that the signal of interest can be multiplied by an analysis dictionary $\mathbf{\Omega}$, leading to a sparse outcome. This model stands as an interesting alternative to the more classical synthesis based sparse representation model. In this work we propose a theoretical study of the performance guarantee of the thresholding algorithm for the pursuit problem in the presence of noise. Our analysis reveals two significant properties of $\mathbf{\Omega}$, which govern the pursuit performance: The first is the degree of linear dependencies between sets of rows in $\mathbf{\Omega}$, depicted by the co-sparsity level. The second property, termed the Restricted Orthogonal Projection Property (ROPP), is the level of independence between such dependent sets and other rows in $\mathbf{\Omega}$. We show how these dictionary properties are meaningful and useful, both in the theoretical bounds derived, and in a series of experiments that are shown to align well with the theoretical prediction.

Index Terms

Sparse Representations, Analysis Model, Thresholding Algorithm, Probability of Success, Linear Dependencies, Restricted Orthogonal Projection Property (ROPP).

T. Peleg is with the Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel (e-mail: tomerfa@tx.technion.ac.il). M. Elad is with the Computer Science Department, Technion – Israel Institute of Technology, Haifa 32000, Israel (e-mail: elad@cs.technion.ac.il).

This work was supported by the European Commissions FP7-FET program, SMALL project (grant agreement no. 225913).

I. INTRODUCTION

Signal models lie at the core of various processing tasks, such as denoising, solving inverse problems, compression, interpolation, sampling, and more. One approach that has become very popular in the past decade is the synthesis-based sparse representation model. In this model, a signal $\mathbf{x} \in \mathbb{R}^d$ is assumed to be composed as a linear combination of a *few* atoms (columns) from a dictionary $\mathbf{D} \in \mathbb{R}^{d \times n}$ [1], [2]. We typically consider a redundant dictionary with $n > d$. The vector $\boldsymbol{\alpha} \in \mathbb{R}^n$ is the sparse representation of the signal, i.e. $\|\boldsymbol{\alpha}\|_0 = k \ll d$.

Vast work on the synthesis model during the past decade has been invested in an attempt to better understand it, and build practical tools for its use. The main activity concentrated on problems such as how to perform pursuit of the sparse representation from the possibly corrupted signal, deriving theoretical success guarantees for such pursuit algorithms, and techniques to learn the dictionary $\mathbf{D}$ from signal examples. Referring specifically to the theoretical success guarantees, various measures were suggested along the years to formalize the notion of the suitability of a synthesis dictionary $\mathbf{D}$ for sparse estimation. These include mutual coherence [3], [4], the exact recovery condition (ERC) [5], the spark [4] and the restricted isometry property (RIP) [6], [7], the capacity sets [8], the characteristics for "s-goodness" [9], and others.

Using these measures, theoretical performance guarantees were developed for various synthesis pursuit algorithms in different setups. For example, the work presented in [10] provided a coherence-based guarantee on the probability of success for the thresholding algorithm in a noise-free setup, under certain assumptions on the representation coefficients. A later work, [11], suggested coherence-based performance guarantees for a wide range of pursuit algorithms, including the thresholding algorithm, in the presence of white Gaussian random noise. These two contributions are mentioned here since both these papers and the work reported here correspond to the simplest of all pursuit methods – the thresholding algorithm.

While the *synthesis* model has been extensively studied, there is a dual *analysis* viewpoint to sparse representations that has only recently started to attract attention [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22]. The analysis model relies on a linear operator (a matrix) $\boldsymbol{\Omega} \in \mathbb{R}^{p \times d}$, which we will refer to as the *analysis dictionary*, and whose rows constitute *analysis atoms*. The key property of this model is our expectation that the analysis representation vector $\boldsymbol{\Omega}\mathbf{x} \in \mathbb{R}^p$ should be sparse with ℓ zeros. These zeros carve out the low-dimensional subspace that this signal belongs to. We shall assume that the dimension of this subspace, which is

denoted by r is indeed small, namely $r \ll d$.

While this description of the analysis model may seem similar to the synthesis counterpart approach, it is in-fact very different when dealing with a redundant dictionary $p > d$. Until recently, relatively little was known about this model, and little attention has been given to it in the literature, compared to the synthesis counterpart model. Several recent works have already started to treat some of the basic research questions arising from the analysis model, such as how to perform pursuit with this model [16], [20], [22], what are the theoretical performance guarantees for the suggested pursuit algorithms [13], [16], [17], [20], [21] and how to learn an analysis dictionary from a set of signal examples [15], [18], [19], [22]. We shall return to some of these contributions towards the end of this paper, and discuss their relation to our work.

The main goal of this paper is a theoretical study of the analysis thresholding pursuit algorithm, deriving conditions for its success in recovering the co-support in the presence of additive noise. A by-product of this study is an identification of two complementary measures of goodness that characterize the analysis dictionary. The first is the degree of linear dependencies between rows in $\mathbf{\Omega}$, which is depicted by the co-sparsity level. This property has already been noticed and discussed in previous works on the analysis model [20], [22]. The second property, termed the Restricted Orthogonal Projection Property (ROPP), is the level of independence between such dependent sets and other rows taken from the analysis dictionary. To the best of our knowledge, this is the first time that this property has been used in the published literature. In this paper we derive an explicit relation between these properties and the expected performance of analysis pursuit by means of thresholding. We demonstrate the goodness of our theoretical findings by matching them versus empirical performance results.

This paper is organized as follows: In Section II we present the core concept of the analysis-based model, characterize the signals that belong to it, and discuss the notion of linear dependencies within the rows of the analysis dictionary. In Section III we present the analysis pursuit problem of denoising a signal using the analysis model and suggest the thresholding algorithm for solving this problem. We test the performance of this algorithm in a series of synthetic experiments for different types of analysis dictionaries. A theoretical study of the performance of the analysis thresholding algorithm is conducted in Section IV. We begin by developing theoretical success guarantees for the thresholding algorithm and discuss the dictionary properties arising from this theoretical analysis. Then we revisit the empirical results

in light of the developed theoretical guarantees. Section V discusses the relation of this work to existing contributions, and Section VI concludes this paper.

II. The Analysis Model and its Dictionary

A. Basic Properties of the Analysis Model

This section begins with a brief review of the analysis-based model. The analysis model for the signal $\mathbf{x} \in \mathbb{R}^d$ uses the possibly redundant analysis dictionary $\mathbf{\Omega} \in \mathbb{R}^{p\times d}$, where redundancy here implies $p \geq d$. Throughout this paper the jth row in $\mathbf{\Omega}$ will be denoted by $\mathbf{w}_j^T$. A fundamental property of this model is the assumption that the analysis representation vector $\mathbf{\Omega x}$ should be sparse. In this work we consider specifically ℓ_0 sparsity, which implies that $\mathbf{\Omega x}$ contains many zeros. The *co-sparsity* ℓ of the analysis model is defined as the number of zeros in the vector $\mathbf{\Omega x}$,

$$\|\mathbf{\Omega x}\|_0 = p - \ell. \tag{1}$$

In this model we put an emphasis on the zeros of $\mathbf{\Omega x}$, and define the *co-support* Λ of $\mathbf{x}$ as the set of $\ell = |\Lambda|$ rows that are orthogonal to it. In other words, $\mathbf{\Omega}_\Lambda \mathbf{x} = 0$, where $\mathbf{\Omega}_\Lambda$ is a submatrix of $\mathbf{\Omega}$ that contains only the rows indexed in Λ. We also define the *co-rank* of a signal $\mathbf{x}$ with co-support Λ as the rank of $\mathbf{\Omega}_\Lambda$. The signal $\mathbf{x}$ is thus characterized by its co-support, which determines the subspace it is orthogonal to, and consequently the complement space to which it belongs. Just like in the synthesis model, we assume that the dimension of the subspace the signal belongs to, denoted by r, is small, namely $r \ll d$. The co-rank of such an analysis signal is $d - r$. How sparse can the analysis representation vector be? The answer to this question is directly related to the existence of linear dependencies within the rows of the analysis dictionary. This will become more clear in the next subsection where we discuss in detail the effect of having such dependencies on the possible co-sparsity levels.

B. Linear Dependencies in the Analysis Dictionary

To motivate our discussion on the advantage of having linear dependencies within the rows of the analysis dictionary, let us first assume that the rows in $\mathbf{\Omega}$ are in general-position, implying that every subset of d or less rows are necessarily linearly independent. This is equivalent to the claim that the spark of $\mathbf{\Omega}^T$ is full [2]. Naturally, for this case, $\ell < d$, since otherwise there would be d independent rows orthogonal to $\mathbf{x}$, implying $\mathbf{x} = 0$. Thus, in this case the analysis

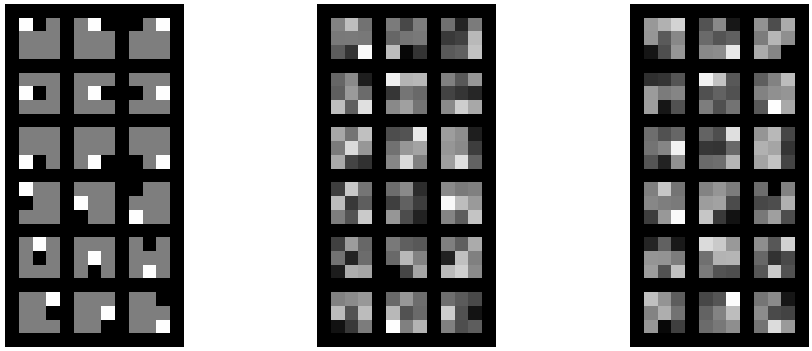

Figure 1. Three types of analysis dictionaries of size 18×9: Left - $\mathbf{\Omega}_{DIF}$, Middle - $\mathbf{\Omega}_{RAND}$, Right - $\mathbf{\Omega}_{MIX}$. Each dictionary atom is displayed as a 2D patch of size 3-by-3.

model leads necessarily to a mild sparsity, $\|\mathbf{\Omega x}\|_0 > p-d$, and for a highly redundant analysis operator, the cardinality of the analysis representation vector $\mathbf{\Omega x}$ is expected to be quite high. In this case, the dimension of the subspace the signal belongs to is $r = d - \ell$. An example for such a dictionary is a Gaussian random one, denoted $\mathbf{\Omega}_{RAND}$, where the rows are drawn identically and independently from a normal distribution.

A more interesting case is when $\mathbf{\Omega}^T$ has *non-full spark*, implying that linear dependencies exist between the dictionary atoms. The immediate implication is that ℓ *could go beyond* d, and yet the signal would not necessarily be nulled. An example of such a dictionary is the set of cyclic horizontal and vertical one-sided derivatives, applied on a 2D signal of size $\sqrt{d}\times\sqrt{d}$. The corresponding analysis dictionary, denoted $\mathbf{\Omega}_{DIF}$, is of size $2d\times d$, thus twice redundant. This dictionary was discussed in detail in [20], showing that its rows exhibit strong linear dependencies.

Note that if we perform right multiplication of an analysis dictionary $\mathbf{B}$ by an invertible square matrix $\mathbf{A}$ then the resulting analysis dictionary $\mathbf{\Omega} \doteq \mathbf{BA}$ exhibits the same linear dependencies between its rows as in $\mathbf{B}$. To see that this is indeed true, let $\Lambda \subseteq \{1,\ldots,p\}$ and suppose that there exists a vector $\boldsymbol{\gamma} \in \mathbb{R}^\ell$ such that $\boldsymbol{\gamma}^T\mathbf{B}_\Lambda = 0$, namely the rows of $\mathbf{B}_\Lambda$ are linearly dependent. Then $\boldsymbol{\gamma}$ also satisfies $\boldsymbol{\gamma}^T\mathbf{\Omega}_\Lambda = \boldsymbol{\gamma}^T\mathbf{B}_\Lambda\mathbf{A} = 0$. For example, the rows of the analysis dictionary that is generated as $\mathbf{\Omega}_{MIX} = \mathbf{\Omega}_{DIF}\mathbf{A}$, where $\mathbf{A}$ is a square matrix consisting of d Gaussian random rows, exhibit the same linear dependencies as $\mathbf{\Omega}_{DIF}$.

Fig. 1 shows the three types of dictionaries mentioned above for $p = 18,\ d = 9$. Throughout

this paper we will experiment with these three dictionaries. The reason for such low dimensional matrices is the fact that the study of the properties of the analysis dictionary will require exhaustive computations over all possible 2^p co-supports. In particular, these dictionary properties will appear in the performance guarantees we are about to derive for the analysis thresholding algorithm (see Section IV-A). Towards the end of this paper we will replace the exact dictionary properties by approximate ones, which are obtained from a set of signal examples generated from the dictionary. This will allow us to show theoretical results also for higher dimensions and check how well they predict the empirical results (see the end of Section IV-C).

As mentioned above, when the rows in $\mathbf{\Omega}$ are not in general-position, the co-sparsity ℓ can be greater than d. This behavior is demonstrated in Fig. 2 showing the distributions of ℓ for the three types of $\mathbf{\Omega}$ shown in Fig. 1 and co-rank 7. For each type the exact co-sparsity distribution is computed exhaustively for all possible co-supports corresponding to a co-rank of 7. We also show an empirical normalized histogram, which is computed from $10,000$ analysis signals of co-rank 7 that are generated using the process that will be described in the beginning of Section III-C. As can be seen the distribution for $\mathbf{\Omega}_{DIF}$ and $\mathbf{\Omega}_{MIX}$ coincide, as should be expected from the observation mentioned above (both dictionaries exhibit the same linear dependencies between their rows). In both cases, though the signals have a fixed co-rank 7, their actual co-sparsities are much higher, varying in the range 8 to 14. Interestingly, odd co-sparsity values cannot lead to the chosen co-rank, as indeed seen in Fig. 2. Thus, we see that by allowing linear dependencies between the rows in $\mathbf{\Omega}$, co-sparsities much higher than the signal dimension d can be achieved.

An alternative measure for the linear dependencies between sets of rows in $\mathbf{\Omega}$ is the signature of the analysis dictionary, which is defined as the ratio of linearly independent sets of k rows out of all possible sets of size k – this ratio is denoted by $f(k)$ [23]. Since every set of size at least $d+1$ is necessarily linearly dependent, it is sufficient to compute the ratios mentioned above for $k = 1, \ldots, d$. The spark of $\mathbf{\Omega}^T$ can be readily computed from the signature $f(k)$ – it is the smallest index k such that $f(k) < 1$. The signatures of the three analysis dictionaries that were shown in Fig. 1 are depicted in Fig. 3. Clearly, $\mathbf{\Omega}_{DIF}$ and $\mathbf{\Omega}_{MIX}$ have the same signature, as they exhibit the same linear dependencies. Their signature is much lower than for $\mathbf{\Omega}_{RAND}$ whose signature equals 1 for all $k = 1, \ldots, d$. We observe that the spark of $\mathbf{\Omega}_{DIF}^T$ and $\mathbf{\Omega}_{MIX}^T$ is 3, whereas the spark of $\mathbf{\Omega}_{RAND}^T$ is $d+1 = 10$ (i.e. the spark is full). To conclude this

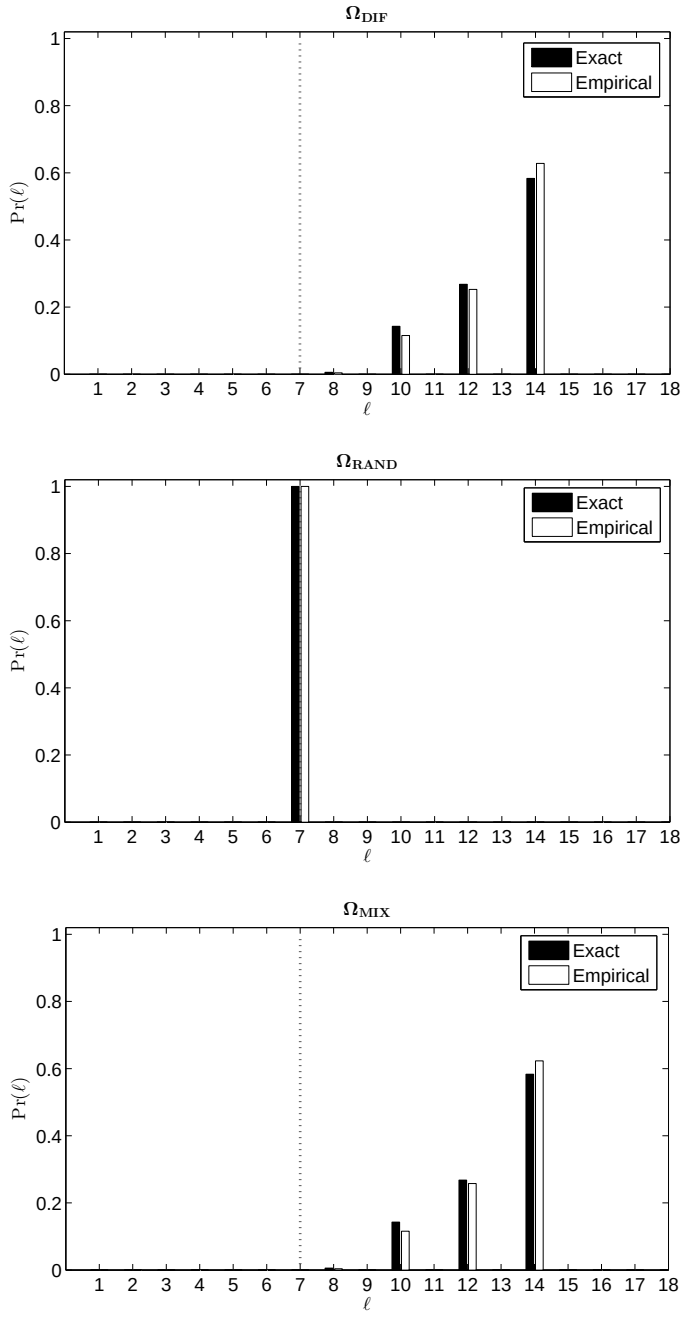

Figure 2. The effective co-sparsities corresponding to each type of analysis dictionary of size 18×9: Top - $\mathbf{\Omega}_{DIF}$, Middle - $\mathbf{\Omega}_{RAND}$, Bottom - $\mathbf{\Omega}_{MIX}$. For each type we show the exact co-sparsity distribution, which is computed exhaustively for all possible co-supports corresponding to a co-rank of 7. We also show an empirical normalized histogram, which is computed from $10,000$ analysis signals of co-rank 7 that were generated using the process described in the beginning of Section III-C. The reference value of $\ell = 7$ is indicated by the vertical dotted line. As can be seen, the effective co-sparsities are all strictly higher for both $\mathbf{\Omega}_{DIF}$ and $\mathbf{\Omega}_{MIX}$.

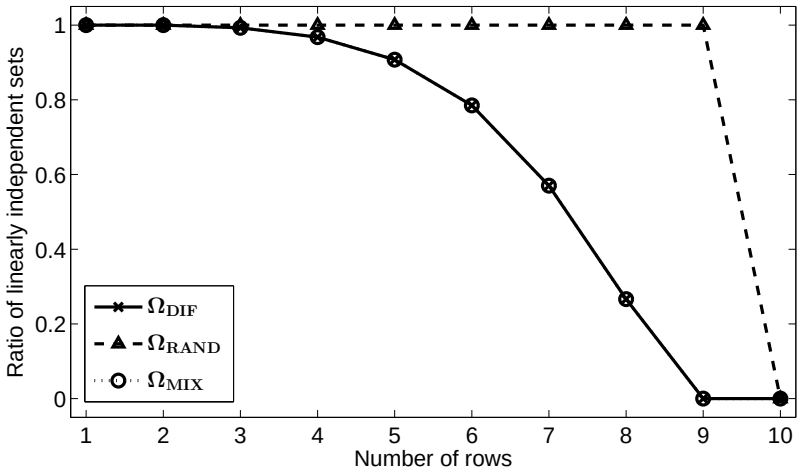

Figure 3. The signatures for three types of analysis dictionary of size 18×9 that were shown in Fig. 1. As can be seen, both $\mathbf{\Omega}_{DIF}$ and $\mathbf{\Omega}_{MIX}$ have the same signature, which is strictly lower than 1 for $k \geq 3$. Therefore the spark of these dictionaries is 3, namely it is non-full. For $\mathbf{\Omega}_{RAND}$ however the signature equals 1 for all $k = 1, \ldots, 9$ and therefore its spark is $d + 1 = 10$.

section, note that a lower dictionary signature indicates that there are more linear dependencies within its rows, and these allow for larger co-sparsity levels.

III. Analysis Thresholding

A. Analysis Pursuit

In this paper we assume that $\mathbf{x}$ is a co-sparse analysis signal with co-rank $d - r$, and this signal is contaminated by additive noise, $\mathbf{y} = \mathbf{x} + \mathbf{e}$. Starting with the *oracle* setup, where the true co-support Λ is known, we can simply recover $\mathbf{x}$ by projecting $\mathbf{y}$ onto the subspace orthogonal to $\mathbf{\Omega}_\Lambda$:

$$\hat{\mathbf{x}} = \left(\mathbf{I} - \mathbf{\Omega}_\Lambda^\dagger \mathbf{\Omega}_\Lambda\right) \mathbf{y}. \tag{2}$$

Assuming a deterministic signal $\mathbf{x}$ residing in a r-dimensional analysis subspace and white and zero-mean Gaussian noise $\mathbf{v}$ with variance σ^2, the mean denoising error in the oracle setup is given by

$$\mathrm{E}\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \mathrm{tr}\left(\mathbf{I} - \mathbf{\Omega}_\Lambda^\dagger \mathbf{\Omega}_\Lambda\right) \sigma^2 = r\sigma^2, \tag{3}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. For more details see [22].

In the general case the correct co-support is unknown and it should be estimated from $\mathbf{y}$. Recovering the noise-free signal $\mathbf{x}$ requires solving a problem of the form

$$\left\{\hat{\mathbf{x}},\ \hat{\Lambda}\right\} = \underset{\mathbf{x},\,\Lambda}{\text{Argmin}}\ \|\mathbf{x}-\mathbf{y}\|_2 \quad \text{Subject To} \quad \begin{array}{l}\boldsymbol{\Omega}_\Lambda \mathbf{x} = 0 \\ \text{Rank}\left(\boldsymbol{\Omega}_\Lambda\right) = d - r\end{array} \tag{4}$$

We refer to this problem as the analysis sparse-coding or analysis-pursuit. This problem can be readily reformulated as a two-step recovery process. To eliminate the dependency on $\mathbf{x}$ we can place the oracle formula of (2) into the problem of (4). We get that recovering the co-support Λ results in solving the problem

$$\hat{\Lambda} = \underset{\Lambda}{\text{Argmin}}\ \|\boldsymbol{\Omega}_\Lambda^\dagger \boldsymbol{\Omega}_\Lambda \mathbf{y}\|_2 \quad \text{Subject To} \quad \text{Rank}\left(\boldsymbol{\Omega}_\Lambda\right) = d - r \tag{5}$$

Once the co-support has been recovered we can project $\mathbf{y}$ onto the orthogonal subspace (using (2)), just as in the oracle setup.

Similar to the synthesis sparse approximation problem, the problem posed in Eq. (4) is combinatorial in nature and can thus only be approximated in general. One approach for approximating the solution is to use a relaxed ℓ_1 penalty function on the coefficients $\boldsymbol{\Omega}\mathbf{x}$, producing

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\text{Argmin}}\ \|\mathbf{x}-\mathbf{y}\|_2 \quad \text{Subject To} \quad \|\boldsymbol{\Omega}\mathbf{x}\|_1 \leq T. \tag{6}$$

This approach is parallel to the basis-pursuit approach for synthesis approximation [24]. A second approach parallels the synthesis greedy pursuit algorithms [25], [26] and suggests selecting rows from $\boldsymbol{\Omega}$ one-by-one in a greedy fashion. The solution can be built by either detecting the rows that correspond to the non-zeros in $\boldsymbol{\Omega}\mathbf{x}$, or by detecting the zeros. The GAP algorithm, described in [20], aims at detecting the non-zeros, whereas the BG and OBG algorithms developed in [22] detect the zeros.

B. The Thresholding Algorithm

In this work we will take the alternative (and simpler) approach of thresholding. This algorithm computes the analysis representation $\boldsymbol{\Omega}\mathbf{y}$ and chooses the smallest entries as the estimated co-support. Thresholding will always obtain a perfect recovery of the co-support in noise-free setups since $\boldsymbol{\Omega}_\Lambda \mathbf{x} = 0$ and $|\mathbf{w}_j^T \mathbf{x}| > 0$ for all $j \in \Lambda^C$. We suggest using it also in

Algorithm 1 ANALYSIS THRESHOLDING ALGORITHM

Input: Analysis dictionary $\mathbf{\Omega} \in \mathbb{R}^{p \times d}$, signal $\mathbf{y} \in \mathbb{R}^d$, and target co-rank $d - r$

Output: Signal $\hat{\mathbf{x}} \in \mathbb{R}^d$ with co-rank $d - r$ approximating the minimization of $\|\mathbf{y} - \hat{\mathbf{x}}\|_2$ and its co-support $\hat{\Lambda}$

Inner Products: $z_k := |\, \mathbf{w}_k^T \mathbf{y} \,|, \forall k = 1, \ldots, p$

Sort: Set Γ to be the index set $\{\, 1, \ldots, p \,\}$ sorted by the value of z_k in increasing order

Initialization: Set $i = 0$, $\Lambda := \emptyset$

while Rank $(\mathbf{\Omega}_\Lambda) < d - r$ **do**

  $i := i + 1$

  Update Co-Support: $\Lambda := \Lambda \cup \{\, \Gamma_i \,\}$

end while

Project: $\hat{\mathbf{x}} = \left(\mathbf{I} - \mathbf{\Omega}_\Lambda^\dagger \mathbf{\Omega}_\Lambda \right) \mathbf{y}$

Refine Co-Support $\hat{\Lambda} = \{k \,|\, 1 \leq k \leq p, \, |\, \mathbf{w}_k^T \hat{\mathbf{x}} \,| \, < \epsilon_0\}$

the presence of noise. A detailed description of the analysis thresholding algorithm is given in Algorithm 1.

The process begins by computing the inner products between all the rows in $\mathbf{\Omega}$ and the signal $\mathbf{y}$ and sorting the index set $\{\, 1, \ldots, p \,\}$ according to the magnitudes of these inner products in increasing order, resulting in a new index set Γ. The co-support is initialized to be an empty set. We then accumulate rows into the co-support, in a row-by-row fashion, according to their order of appearance in the set Γ. This process repeats until the target co-rank is achieved, namely Rank $(\mathbf{\Omega}_\Lambda) = d - r$. The solution $\hat{\mathbf{x}}$ is then computed by projecting $\mathbf{y}$ onto the subspace orthogonal to the selected rows. Finally, the co-support is refined by recalculating the representation vector $\mathbf{\Omega}\hat{\mathbf{x}}$ and finding the additional coefficients that fall below some small threshold ϵ_0. This can reveal additional rows that are orthogonal to the signal estimate, namely the rows that are spanned by the existing set of rows $\mathbf{\Omega}_\Lambda$. Despite the fact that the last step ("Refine Co-Support") has no impact on the signal recovery, it is still significant for our purposes, as our study checks the correctness of the found co-support.

In practice, the above algorithm can be implemented efficiently by accumulating an orthog-

onalized set of the co-support rows using a modified Gram-Schmidt process. This process is applied according to the order of appearance in the set Γ. Denoting by $\{\mathbf{q}_j\}_{j=1}^{J}$ the orthogonal set accumulated so far (as column vectors), the orthogonalization of a new row $\mathbf{w}_{\Gamma_i}^T$ is obtained by

$$\mathbf{q}_i = \mathbf{w}_{\Gamma_i} - \sum_{j=1}^{J} (\mathbf{q}_j^T \mathbf{w}_{\Gamma_i}) \mathbf{q}_j. \tag{7}$$

If $\mathbf{q}_i$ equals zero, it is not added to the orthogonal set, as it is already spanned by the existing one. Otherwise, this vector is normalized, $\mathbf{q}_i = \mathbf{q}_i / \|\mathbf{q}_i\|_2$.

The above-described orthogonalization process allows us first of all to avoid the computation of the rank of the submatrix $\mathbf{\Omega}_\Lambda$, since the number of vectors in the orthogonalized set (J) equals the desired rank. Secondly, the orthogonalized set $\{\mathbf{q}_j\}_{j=1}^{d-r}$ can also be used to avoid the matrix inversion in the "Projection" step, which translates comfortably to

$$\hat{\mathbf{x}}_i = \left(\mathbf{I} - \mathbf{\Omega}_{\Lambda_i}^{\dagger} \mathbf{\Omega}_{\Lambda_i}\right) \mathbf{y} = \left[\mathbf{I} - \sum_{j=1}^{i} \mathbf{q}_j \mathbf{q}_j^T\right] \mathbf{y}. \tag{8}$$

C. Synthetic Experiments

We now demonstrate how the thresholding algorithm (see Algorithm 1) performs through a series of synthetic experiments. Throughout this subsection we shall assume that the analysis signals are generated by the following process: Choose randomly a set of row indices $\Lambda \subseteq \{1, \ldots, p\}$, which will be the signal's co-support. Starting with a random vector $\mathbf{u}$, whose entries are assumed to be drawn independently and identically from a zero-mean Gaussian distribution with variance σ_u^2, project it onto the subspace orthogonal to $\mathbf{\Omega}_\Lambda$:

$$\mathbf{x} = (\mathbf{I} - \mathbf{\Omega}_\Lambda^{\dagger} \mathbf{\Omega}_\Lambda)\mathbf{u}, \tag{9}$$

and $\mathbf{x}$ is an analysis signal that satisfies our co-sparsity assumption. For a general-positioned $\mathbf{\Omega}$ we choose exactly ℓ rows from $\mathbf{\Omega}$ at random. Otherwise we choose $d - r$ linearly independent rows from $\mathbf{\Omega}$. Once a signal $\mathbf{x}$ has been generated, its analysis representation $\mathbf{\Omega x}$ is re-computed, possibly revealing additional rows that are orthogonal to this signal, due to linear dependence on the chosen subset Λ.

We generate $N = 10,000$ analysis signals in $\mathbb{R}^9$ residing in 2-dimensional subspaces for the three types of analysis dictionaries shown in Fig. 1 – normalized histograms of their effective co-supports are depicted in Fig. 2. These signals are contaminated with additive white Gaussian

noise at different noise levels σ, resulting in a set of noisy signals $\{y_j\}_{j=1}^N$ for each dictionary type and noise level. The thresholding algorithm is then applied on these signals with a target co-rank of $d-r=7$. Results are shown in Fig. 4 for various signal-to-noise ratios (SNR) in the range $6dB$ to $74dB$. Each SNR level is related to the ratio σ/σ_u by

$$SNR \doteq 10\log_{10}\left(\frac{\mathrm{E}\|\mathbf{x}\|_2^2}{\mathrm{E}\|\mathbf{y}-\mathbf{x}\|_2^2}\right) = -20\log_{10}\left(\sqrt{\frac{d}{r}}\frac{\sigma}{\sigma_u}\right). \tag{10}$$

where in the last equation we used the equation $\mathrm{E}\|\mathbf{x}\|_2^2 = \mathrm{tr}\left(\mathbf{I}-\mathbf{\Omega}_\Lambda^\dagger\mathbf{\Omega}_\Lambda\right)\sigma_u^2 = r\sigma_u^2$, which holds since $\mathbf{x}$ is a zero-mean Gaussian vector with a covariance matrix $\left(\mathbf{I}-\mathbf{\Omega}_\Lambda^\dagger\mathbf{\Omega}_\Lambda\right)\sigma_u^2$ (exhibiting a similar form as in the oracle error – see Eq. (3)), and $\mathrm{E}\|\mathbf{y}-\mathbf{x}\|_2^2 = d\sigma^2$. At this point we should mention that the SNR levels shown on the right part of the figure are very high ones (for example SNR=$60dB$ means that the signal energy is 1000 times the noise energy). Setups with such high SNR levels can be considered as almost noise-free. Therefore we expect that the thresholding algorithm will obtain a perfect recovery of the co-support in these setups, just like in the noise-free setup.

In Fig. 4 we can see on the top the empirical probability of success for the thresholding algorithm on each of the dictionaries. Note that "success" refers here to an exact recovery of the true co-support. On the bottom we can see the denoising performance, measured as the average SNR improvements (ISNR):

$$ISNR \doteq -10\log_{10}\left(\frac{\|\hat{\mathbf{x}}-\mathbf{x}\|_2^2}{d\sigma^2}\right) \tag{11}$$

These are also compared with the oracle performance, which corresponds to an ISNR of $-10\log_{10}(r/d) = 6.53dB$. We can see at the top right corner of the figure that thresholding succeeds with probability one for all three types of dictionaries, which aligns with our expectations for high SNRs that were mentioned before.

Several important observations can be drawn from the results shown in Fig. 4. First of all, we can see that the probability of success decreases as the SNR deteriorates. This aligns with the simple intuition that the higher the noise, the higher the chance of any pursuit algorithm to make mistakes in the co-support detection. Second, the highest success ratio and ISNR are obtained for $\mathbf{\Omega}_{DIF}$ at all noise levels; the second-best results relate to $\mathbf{\Omega}_{MIX}$ and the worse to $\mathbf{\Omega}_{RAND}$.

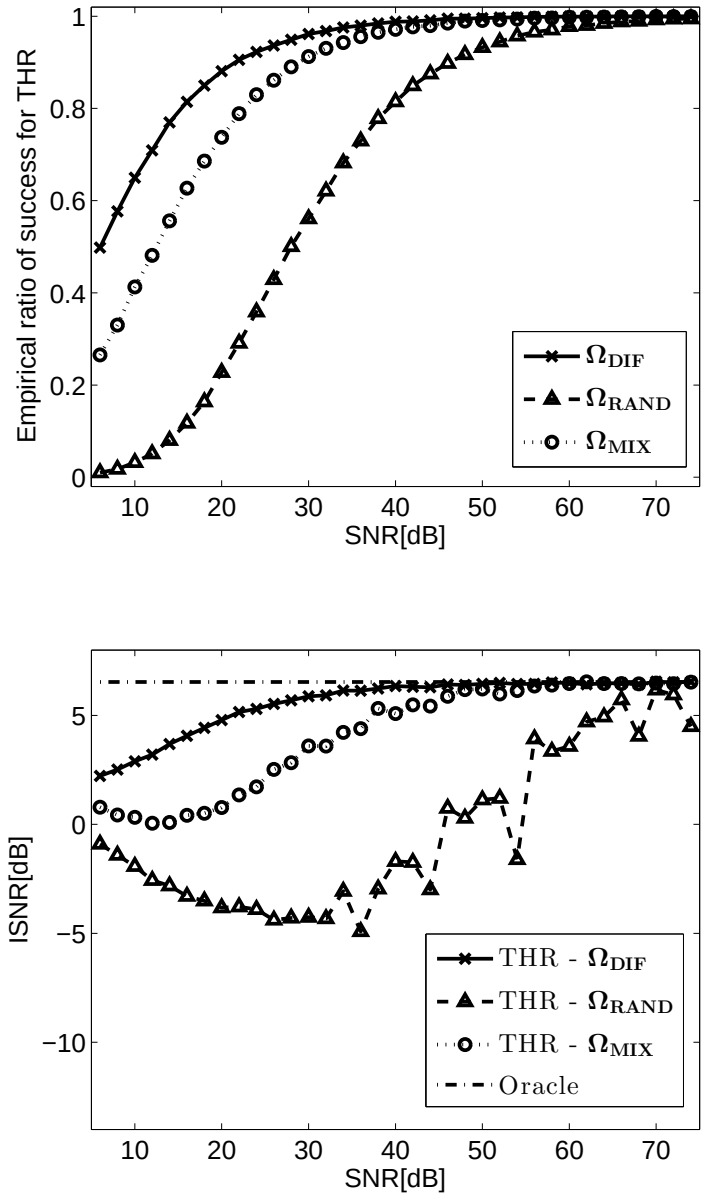

Figure 4. Denoising experiments with analysis signals of co-rank 7 created from the three types of analysis dictionaries of size 18×9 that were shown in Fig. 1. Additive white noise is added to each of these signals for varying noise levels and then the thresholding algorithm (see Algorithm 1) is applied on each signal to obtain a recovery of its co-support and its resulting denoised signal. Top: The empirical probability of success in recovering the true co-support for the thresholding algorithm on each of the dictionary types. Bottom: The noise attenuation performance obtained for the thresholding algorithm on each of the dictionary types. These are compared with the oracle result, where denoising is obtained by projection onto the correct analysis subspace (knowing the true co-support of the signals).

The observation that $\mathbf{\Omega}_{RAND}$ exhibits the worst performance does not come as a surprise to us. The fact that having many linear dependencies in an analysis dictionary $\mathbf{\Omega}$ leads to better denoising results has already been observed in a previous work [22]. However, the performance gap between $\mathbf{\Omega}_{DIF}$ and $\mathbf{\Omega}_{MIX}$ is not obvious at all, if we recall that both exhibit the same linear dependencies between their rows (and hence the same co-sparsity distribution). This calls for a deeper theoretical study of the thresholding algorithm, which is the topic of the next section.

IV. Theoretical Study of Analysis Thresholding

This section consists of the main contribution of this paper: A theoretical analysis of the capability of the thresholding algorithm to recover the true analysis co-support in the presence of additive noise, and the implications of this analysis. We start in Section IV-A with the derivation of our main result – a lower-bound on the probability of successfully recovering the co-support by the analysis thresholding algorithm. Section IV-B discusses the obtained results and specifically the meaning of the measures proposed for the analysis dictionary. In Section IV-C we revisit these results in an attempt to explain them further, and contrast them with the empirical evidence we have just created. As this work focuses on the probability of the analysis thresholding algorithm to recover the exact co-support, the relative denoising performance will not be further explored in this paper and remains a topic for future research.

A. Theoretical Guarantees for Analysis Thresholding

Before we turn to the development of the theoretical guarantees for the analysis thresholding algorithm, we would like to set some basic assumptions and notations. First, we assume that all the rows in $\mathbf{\Omega}$ have unit-norm. Secondly, we denote an index set of $d-r$ linearly independent rows taken from Λ by $\widetilde{\Lambda} \subseteq \Lambda$, namely $\text{Span}\left\{\mathbf{\Omega}_{\widetilde{\Lambda}}\right\} = \text{Span}\left\{\mathbf{\Omega}_{\Lambda}\right\}$. Finally, given a noise-free signal $\mathbf{x}$ and an analysis dictionary $\mathbf{\Omega}$, let us define

$$z_{min} \doteq \underset{j \in \Lambda^C}{\text{Min}} \left|\mathbf{w}_j^T \mathbf{x}\right|, \tag{12}$$

where Λ is the co-support of $\mathbf{\Omega x}$ and Λ^C is the complementary index set. For the co-sparse analysis signal $\mathbf{x}$ we have that $\mathbf{\Omega}_{\Lambda}\mathbf{x} = 0$, implying that $\mathbf{\Omega}_{\Lambda^C}\mathbf{x} \neq 0$. The value of z_{min} is the smallest of those non-zero inner-products with Ω_{Λ^C}, and it plays a major role in the ability of the thresholding algorithm to tell the right co-support rows from the rest in the noisy case. We begin our performance study of this algorithm with a sufficient condition on z_{min} for success.

Lemma 1. *Let $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{e}$, where $\boldsymbol{x}$ is a co-sparse analysis signal with co-support Λ on $\boldsymbol{\Omega}$. If $\boldsymbol{x}$ and $\boldsymbol{\Omega}$ satisfy $z_{min} \geq 2$ Max$_{j\in\widetilde{\Lambda}\cup\Lambda^C}$ $\left|\boldsymbol{w}_j^T\boldsymbol{e}\right|$, then the thresholding algorithm succeeds in recovering the true co-support Λ of $\boldsymbol{x}$ from $\boldsymbol{y}$.*

Proof: We begin with the simple observation that the thresholding algorithm succeeds in recovering the true co-support Λ of $\mathbf{x}$ when

$$\underset{j\in\widetilde{\Lambda}}{\text{Max}} \quad \left|\mathbf{w}_j^T\mathbf{y}\right| < \underset{j\in\Lambda^C}{\text{Min}} \quad \left|\mathbf{w}_j^T\mathbf{y}\right|. \tag{13}$$

Since $\mathbf{w}_j^T\mathbf{x} = 0$ for all $j \in \widetilde{\Lambda}$ the left-hand side of (13) translates to

$$\underset{j\in\widetilde{\Lambda}}{\text{Max}} \quad \left|\mathbf{w}_j^T\mathbf{y}\right| = \underset{j\in\widetilde{\Lambda}}{\text{Max}} \quad \left|\mathbf{w}_j^T\mathbf{e}\right|. \tag{14}$$

For the right-hand side of (13) we derive a lower bound

$$\underset{j\in\Lambda^C}{\text{Min}} \quad \left|\mathbf{w}_j^T\mathbf{y}\right| \geq \underset{j\in\Lambda^C}{\text{Min}} \quad \left|\mathbf{w}_j^T\mathbf{x}\right| - \left|\mathbf{w}_j^T\mathbf{e}\right| \geq z_{min} - \underset{j\in\Lambda^C}{\text{Max}} \quad \left|\mathbf{w}_j^T\mathbf{e}\right|, \tag{15}$$

where the first inequality holds from the triangle inequality and the second holds from the properties of the minimum and maximum operators,

$$\text{Min } (f-g) \geq \text{Min } f + \text{Min } (-g) = \text{Min } f - \text{Max } g. \tag{16}$$

From (13)-(15) we get that a sufficient condition for success of the thresholding algorithm is:

$$\underset{j\in\widetilde{\Lambda}}{\text{Max}} \quad \left|\mathbf{w}_j^T\mathbf{e}\right| < z_{min} - \underset{j\in\Lambda^C}{\text{Max}} \quad \left|\mathbf{w}_j^T\mathbf{e}\right|, \tag{17}$$

which can be comfortably replaced by the sufficient condition

$$z_{min} > 2 \underset{j\in\widetilde{\Lambda}\cup\Lambda^C}{\text{Max}} \quad \left|\mathbf{w}_j^T\mathbf{e}\right|, \tag{18}$$

since

$$2 \underset{j\in\widetilde{\Lambda}\cup\Lambda^C}{\text{Max}} \quad \left|\mathbf{w}_j^T\mathbf{e}\right| \geq \underset{j\in\widetilde{\Lambda}}{\text{Max}} \quad \left|\mathbf{w}_j^T\mathbf{e}\right| + \underset{j\in\Lambda^C}{\text{Max}} \quad \left|\mathbf{w}_j^T\mathbf{e}\right|. \tag{19}$$

■

Note that so far we have made no specific assumptions on the signal generative model or the noise. The only assumption is on the inner products between the signal $\mathbf{x}$ and rows in $\boldsymbol{\Omega}$ that are not indexed in the true co-support. An immediate observation arising from the above lemma appears in the following corollary. Using the Cauchy-Schwarz inequality and the fact that all rows in $\boldsymbol{\Omega}$ are normalized, we get that $\left|\mathbf{w}_j^T\mathbf{e}\right| \leq \|\mathbf{e}\|_2$. Thus,

Corollary 1. *Let $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{e}$, where $\boldsymbol{x}$ is a co-sparse analysis signal with co-support Λ on $\boldsymbol{\Omega}$ and $\|\boldsymbol{e}\|_2 \leq \epsilon$. If $\boldsymbol{x}$ and $\boldsymbol{\Omega}$ satisfy $z_{min} \geq 2\epsilon$, then the thresholding algorithm succeeds in recovering the true co-support Λ of $\boldsymbol{x}$ from $\boldsymbol{y}$.*

Note that we have referred to the noise as deterministic and bounded. This results in a very pessimistic success condition, as should be expected for a worst-case performance analysis like the one practiced here, in which an estimator must perform well even when the noise maximally damages the measurements (the noise in this case is thus called adversarial). This should remind the reader of the theoretical guarantees derived for synthesis-based pursuit algorithms under adversarial noise [1], [2], [3], [4], [5].

To improve the theoretical guarantees, we turn to a setup where the noise is assumed to be random. Specifically, we assume white and zero-mean Gaussian noise with variance σ^2, and derive a lower bound on the probability of success under a sufficient condition on z_{min}.

Theorem 1. *Let $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{e}$ and $\boldsymbol{e} \sim N(0, \sigma^2 \boldsymbol{I})$. If $\boldsymbol{x}$ is a co-sparse analysis signal with co-support Λ on $\boldsymbol{\Omega}$, co-sparsity ℓ, and co-rank $d-r$, and $\boldsymbol{\Omega}$ and $\boldsymbol{x}$ satisfy $z_{min} \geq \beta\sigma$, then the thresholding algorithm succeeds in recovering the true co-support Λ of $\boldsymbol{x}$ from $\boldsymbol{y}$ with probability at least* $\left(\text{Max } \left\{0, 1 - \sqrt{\frac{8}{\pi\beta^2}} \exp\left\{-\frac{\beta^2}{8}\right\}\right\}\right)^{p-\ell+d-r}$.

Before turning to prove this result, a short discussion is in order. This theorem provides a lower bound on the conditional probability of success given that $z_{min} \geq \beta\sigma$. The derived expression has an exponential form with a base in the range $[0, 1]$ depending on β and a power $p - \ell + d - r$. The observant reader might ask at this stage: Why is the performance guarantee of Theorem 1 better than the result of Corollary 1? To answer this question we explore the dependence of this performance guarantee on β. The bound on this probability increases exponentially from zero to one as β grows, but at the same time the condition on z_{min} becomes stricter. This bound is shown in Fig. 5 for a setup with $d = 9$, $p = 18$, $r = 2$ and $\ell = 14$. First, we can see that the exact co-support is recvovered with overwhelming probability (i.e. near one) for $z_{min} \geq 6\sigma$. This aligns with the guarantee of Corollary 1 requiring $z_{min} \geq 2\epsilon$, where ϵ is of order $\sqrt{d}\sigma = 3\sigma$. More importantly, Theorem 1 provides probabilistic success guarantees for weaker conditions on z_{min}, for which Corollary 1 cannot make any guarantee.

Next, we explore the dependence of the obtained lower bound on the number of atoms p and the co-sparsity ℓ and the co-rank $d - r$. Clealry, the probability of success of the thresholding

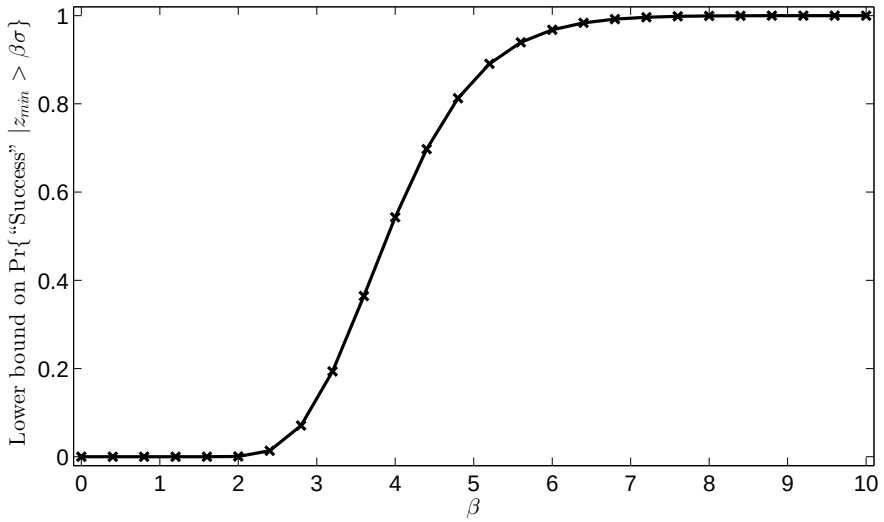

Figure 5. The dependence on β of the lower bound on the conditional probability of success given that $z_{min} \geq \beta\sigma$ (see Theorem 1) for a setup with $d = 9$, $p = 18$, $r = 2$ and $\ell = 14$.

algorithm improves (grows) when $p - \ell + d - r$ gets smaller. Such is the case, for example, when the dictionary size (p, d) is kept fixed, the co-rank $d - r$ is chosen as well, and the level of dependencies, as depicted in ℓ, grows. This manifests the surprising fact that strong linear-dependencies within $\mathbf{\Omega}$ lead to better performance. Adopting a different point of view, when p (the dictionary's redundancy) grows, the level of performance may remain the same as long as ℓ grows with it such that their difference remains unchanged.

Proof: Let us first define the event

$$B = \left\{ \mathbf{e} \mid \underset{j \in \widetilde{\Lambda} \cup \Lambda^C}{\text{Max}} \quad \left| \mathbf{w}_j^T \mathbf{e} \right| < \tau \right\}. \tag{20}$$

A similar event was defined in [11] when developing success guarantees for the synthesis-based thresholding and OMP algorithms. We start by deriving a lower bound on the probability of this event:

$$\begin{aligned} \Pr\{B\} &\geq \prod_{j \in \widetilde{\Lambda} \cup \Lambda^C} \Pr\left\{ \left| \mathbf{w}_j^T \mathbf{e} \right| < \tau \right\} = \left[1 - 2Q\left(\frac{\tau}{\sigma} \right) \right]^{p-\ell+d-r} \\ &\geq \left[1 - \sqrt{\frac{2\sigma^2}{\pi\tau^2}} \exp\left\{ -\frac{\tau^2}{2\sigma^2} \right\} \right]^{p-\ell+d-r}, \end{aligned} \tag{21}$$

where $Q(\cdot)$ is the Gaussian distribution tail,

$$Q(t) = \frac{1}{\sqrt{2\pi}} \int_{t}^{\infty} \exp\left\{-\frac{z^2}{2}\right\} \mathrm{d}z. \tag{22}$$

The first inequality holds due to Ŝidák's lemma [27] for a set of jointly Gaussian random variables. The next equality holds due to the fact that $\widetilde{\Lambda}$ and Λ^C are disjoint sets of sizes $d-r$ and $p-\ell$ respectively. In the last inequality we use a well-known upper bound on the Gaussian distribution tail,

$$Q(t) \leq \frac{1}{t\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\}. \tag{23}$$

We set $\tau = \frac{1}{2}\beta\sigma$, and thus the event B corresponds to all the noise vectors $\mathbf{e}$ satisfying $2\text{Max}_{j\in\widetilde{\Lambda}\cup\Lambda^C} \left|\mathbf{w}_j^T\mathbf{e}\right| < \beta\sigma$. Therefore, if $z_{min} > \beta\sigma$ as this theorem states, then necessarily z_{min} also satisfies the condition of Lemma 1, namely $z_{min} > \beta\sigma > 2\text{Max}_{j\in\widetilde{\Lambda}\cup\Lambda^C} \left|\mathbf{w}_j^T\mathbf{e}\right|$, which guarantees the success of the analysis thresholding algorithm. The probability for this to happen is bounded from below by the expression we have derived in Eq. (21), as claimed[1].

■

Next, we would like to eliminate the dependence on z_{min} and derive a theoretical guarantee in terms of the analysis subspace dimension r, the co-sparsity ℓ and possibly some internal properties of the dictionary $\mathbf{\Omega}$. This will help to reveal what makes an analysis dictionary more suitable for co-sparse estimation. To initiate such an analysis, we make an additional assumption on the signal generative model. Given a dictionary $\mathbf{\Omega}$, a co-support Λ and a random Gaussian vector $\mathbf{u} \sim N\left(0, \sigma_u^2\mathbf{I}\right)$, $\mathbf{x}$ is generated by projecting $\mathbf{u}$ onto the subspace orthogonal to $\mathbf{\Omega}_\Lambda$, as described in Section III-C (see (9)). We further assume that $\mathbf{u}$ and $\mathbf{e}$ are statistically independent. Using this generative model for $\mathbf{x}$, we shall derive a theoretical guarantee for success of the thresholding algorithm, based on a new property of $\mathbf{\Omega}$ we shall refer to as ROPP:

Definition 1. *Given an analysis dictionary $\mathbf{\Omega}$, the Restricted Orthogonal Projection Property (ROPP) of this dictionary with a constant α_r is defined as*

$$\alpha_r = \underset{\Lambda, j | \mathit{Rank}(\mathbf{\Omega}_\Lambda)=d-r, j\in\Lambda^C}{\mathit{Min}} \left\|(\boldsymbol{I} - \mathbf{\Omega}_\Lambda^\dagger\mathbf{\Omega}_\Lambda)\boldsymbol{w}_j\right\|_2. \tag{24}$$

[1]For values of β that lead to a negative argument in this expression we replace Eq. (21) by a trivial zero lower bound on the probability.

More on the meaning of this constant is brought in Section IV-B. Armed with this definition, we now turn to improve Theorem 1, by removing the dependency on z_{min}.

Theorem 2. *Let $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{e}$, where $\boldsymbol{u} \sim N(0, \sigma_u^2 \boldsymbol{I})$, $\boldsymbol{x}$ is a co-sparse analysis signal with co-support Λ on $\boldsymbol{\Omega}$, obtained by $\boldsymbol{x} = (\boldsymbol{I} - \boldsymbol{\Omega}_\Lambda^\dagger \boldsymbol{\Omega}_\Lambda)\boldsymbol{u}$, and $\boldsymbol{e} \sim N(0, \sigma^2 \boldsymbol{I})$ is the additive noise statistically independent of $\boldsymbol{u}$. If $\boldsymbol{\Omega}$ satisfies the ROPP with a constant α_r and $\boldsymbol{x}$ has co-rank $d - r$ and co-sparsity ℓ on $\boldsymbol{\Omega}$, then the thresholding algorithm succeeds in recovering the true co-support Λ of $\boldsymbol{x}$ from $\boldsymbol{y}$ with probability at least* $\left(\text{Max } \left\{0, 1 - \sqrt{\frac{8}{\pi\beta^2}} \exp\left\{-\frac{\beta^2}{8}\right\}\right\}\right)^{p-\ell+d-r} \left(2Q\left(\frac{\beta\sigma}{\alpha_r \sigma_u}\right)\right)^{p-\ell}$ *for any constant $\beta > 0$.*

Note that $Q(\cdot)$ appearing in this theorem is the Gaussian distribution tail (see (22)).

Just as we did for the conditional probability os success of Theorem 1, we start by exploring the dependence of the resulting bound with respect to β. This is shown in Fig. 6 for a setup with $d = 9$, $p = 18$, $r = 2$, $\ell = 14$ (same as before – see Fig. 5), $\alpha_r = 0.75$ and $\sigma/\sigma_u = 0.01$. We can see that the choice of β is crucial for the strictness of the resulting lower bound on the probability of success. For the setup considred here the optimal value of β is 6, which results in a lower bound of 0.744. The lower bound appearing in this theorem is a product of two exponential terms. The first is the bound on the conditional probability that appeared in Theorem 1 and the second terms is a bound on the probability that the condition $z_{min} \geq \beta\sigma$ holds (this bound will be derived in the proof that follows). The first terms grows with β, while the second decreases, thus explaining the peak between 0 and infinity.

Next, we explore the dependence of the obtained lower bound on the number of atoms p and the co-sparsity ℓ, fixing the noise ratio σ/σ_u, the signal dimension d and the analysis subspace dimension r, and assuming that the dictionary satisfies the ROPP with a constant α_r. Since both the bases of the exponential terms are in the range $[0, 1]$, we can see that the probability of success of the thresholding algorithm improves when the difference $p - \ell$ becomes smaller. This means that the same observations made before on p and ℓ for the conditional probability also hold here: For a given dictionary of size (p, d) performance improves as ℓ grows, and when the redundancy of the dictionary is increased the performance remain the same as long as the difference $p - \ell$ remains unchanged. Finally, we observe that since $Q(\cdot)$ is monotonic decreasing, the performance improves as the noise ratio σ/σ_u decreases or the ROPP constant α_r grows.

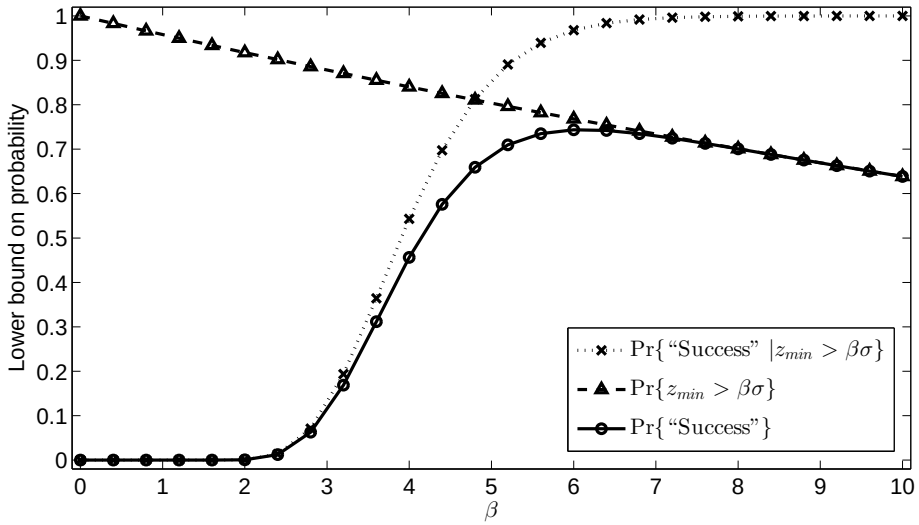

Figure 6. The dependence on β of the lower bound on the probability of success of Theorem 2 for a setup with $d = 9$, $p = 18$, $r = 2$, $\ell = 14$, $\sigma/\sigma_u = 0.01$ and $\alpha_r = 0.75$. For this setup the optimal value of β is 6, which results in a lower bound of 0.744 on the probabilty of success. For each value of β we also show the lower bounds on the conditional probability of success of Theorem 1 and on the probability that the condition $z_{min} \geq \beta\sigma$ holds (see Eq. (25)). The final bound of Theorem 2 is a product of these two bounds.

Proof: We begin by observing that a signal $\mathbf{x}$ generated as an orthogonal projection of a Gaussian i.i.d vector $\mathbf{u}$ is also Gaussian, $\mathbf{x} \sim N\left(0, \sigma_u^2(\mathbf{I} - \mathbf{\Omega}_\Lambda^\dagger \mathbf{\Omega}_\Lambda)\right)$ and so is any inner product with $\mathbf{x}$, $\mathbf{w}_j^T\mathbf{x} \sim N\left(0, \|(\mathbf{I} - \mathbf{\Omega}_\Lambda^\dagger \mathbf{\Omega}_\Lambda)\mathbf{w}_j\|_2^2\sigma_u^2\right)$. Using this observation, we now derive a lower bound on the probability that the condition for success of Theorem 1 holds:

$$\begin{aligned}\Pr\{z_{min} > \beta\sigma\} &= \Pr\left\{\underset{j\in\Lambda^C}{\text{Min}} \ \left|\mathbf{w}_j^T\mathbf{x}\right| > \beta\sigma\right\} \geq \prod_{j\in\Lambda^C} \Pr\left\{\left|\mathbf{w}_j^T\mathbf{x}\right| > \beta\sigma\right\} \\ &= \prod_{j\in\Lambda^C} 2Q\left(\frac{\beta\sigma}{\|(\mathbf{I} - \mathbf{\Omega}_\Lambda^\dagger \mathbf{\Omega}_\Lambda)\mathbf{w}_j\|_2\sigma_u}\right) \geq \left[2Q\left(\frac{\beta\sigma}{\alpha_r\sigma_u}\right)\right]^{p-\ell}. \qquad (25)\end{aligned}$$

The first inequality relies on Ŝidák's lemma, as before[2]. In the next equality we use the fact that $\mathbf{w}_j^T\mathbf{x}$ is Gaussian with the variance mentioned above. The last inequality holds from the definition of the ROPP in (24) and since $Q(\cdot)$ is monotonic decreasing. The power $p-\ell$ comes

[2]In fact, we are not explicitly using Ŝidák's lemma, but a related inequality resulting from this lemma. Let $\{v_j\}_{j=1}^M$ be a set of jointly Gaussian random vectors. Then according to Ŝidák's lemma, $\Pr\{\text{Max}_{1\leq j\leq M} \ |v_j| < \tau\} \geq \prod_{j=1}^M \Pr\{|v_j| < \tau\}$. Thus, turning to our expression, we observe that $\Pr\{\text{Min}_{1\leq j\leq M} \ |v_j| > \tau\} = \Pr\{-\text{Max}_{1\leq j\leq M} \ (-|v_j|) > \tau\} = \Pr\{\text{Max}_{1\leq j\leq M} \ (-|v_j|) < -\tau\} \geq \prod_{j=1}^M \Pr\{-|v_j| < -\tau\} = \prod_{j=1}^M \Pr\{|v_j| > \tau\}$, leading to the relation we used.

from the cardinality of the set Λ^C.

Combining Theorem 1 and Eq. (18) we get that the final lower bound on the probability of success for the thresholding algorithm is a direct multiplication of the two probability expressions, leading to the claimed lower-bound probability posed in terms of the ROPP constant α_r and the co-sparsity ℓ. ■

B. Discussion on the Properties of the Analysis Dictionary

We begin this subsection by taking a closer look at the ROPP. This is an internal property of the analysis dictionary, indicating for a set of $d-r+1$ linearly independent rows from the dictionary how much each row is spread away from the subspace spanned by the rest. At the special case of a unitary dictionary $\mathbf{\Omega}$ we have $\alpha_r = 1$ for all values of r since each row is orthogonal to the subspace spanned by every possible set of rows not including it. How does the ROPP compares to other dictionary properties? Starting with the RIP [6], [7],

$$(1-\delta_k)\left\|\mathbf{v}\right\|_2^2 \leq \left\|\mathbf{D}\mathbf{v}\right\|_2^2 \leq (1+\delta_k)\left\|\mathbf{v}\right\|_2^2, \tag{26}$$

which holds for all k-sparse vectors $\mathbf{v} \in \mathbf{R}^n$, the ROPP also bounds an ℓ_2 norm related to the dictionary. However, the ROPP looks at projection matrices constructed from the dictionary instead of the dictionary itself as in the RIP, and applies these matrices on dictionary atoms not used for the matrix construction instead of looking at all possible signals with a certain sparsity as in the RIP. This should remind the reader of the ERC [5], which has a similar flavor. Turning to the ERC [5], for a better comparison let us replace the ROPP by the sufficient condition

$$\underset{j\in\Lambda^C}{\text{Max}} \left\|\mathbf{\Omega}_\Lambda^\dagger \mathbf{\Omega}_\Lambda \mathbf{w}_j\right\|_2 \leq 1-\alpha_r \tag{27}$$

for the same co-supports Λ as in (24). To see that this is indeed a sufficient condition, we assume that (27) holds and show that

$$\left\| \left(\mathbf{I} - \mathbf{\Omega}_\Lambda^\dagger \mathbf{\Omega}_\Lambda\right) \mathbf{w}_j \right\|_2 \geq \left\|\mathbf{w}_j\right\|_2 - \left\|\mathbf{\Omega}_\Lambda^\dagger \mathbf{\Omega}_\Lambda \mathbf{w}_j\right\|_2 \geq \alpha_r, \tag{28}$$

where in the first inequality we used the well-known relation, $\left\|\mathbf{v}_1 - \mathbf{v}_2\right\|_2 \geq \left|\left\|\mathbf{v}_1\right\|_2 - \left\|\mathbf{v}_2\right\|_2\right|$, which holds for any pair of vectors $\mathbf{v}_1, \mathbf{v}_2$, and in the second inequality we used the fact that $\left\|\mathbf{w}_j\right\|_2 = 1$ and the assumption of (27). The condition appearing in (27) has a similar structure to the ERC,

$$\underset{j\notin s}{\text{Max}} \left\|\mathbf{D}_s^\dagger \mathbf{d}_j\right\|_1 \leq 1. \tag{29}$$

However, there are two inherent differences: The pseudoinverse of the submatrix $\mathbf{D}_s$ is replaced by a projection matrix onto the null space of $\mathbf{\Omega}_\Lambda$ and the ℓ_1 norm is replaced by ℓ_2. Consequently, an upper bound of 1 is a trivial one and it is replaced by the stricter bound $1-\alpha_r$ for some constant α_r.

Next, we turn to the theoretical guarantee of Theorem 2 and observe that it gives rise to two dictionary properties, which serve as two distinct forces dictating the ability to recover the co-supports of analysis signals over the given dictionary. The first property, emanating from the signature or the co-sparsity of $\mathbf{\Omega}$, determines which sets of rows and how many of them are linearly dependent. However, this measure by itself does not provide us with any quantitative relation between these sets and the rows that are linearly independent on them. The second property focuses exactly on these missing relations, telling us how much a row is spread away from the others, provided that it is linearly independent on them.

Are these two dictionary properties somehow related to each other? To provide an answer to this question we explore the joint distribution of the two. For this purpose, we replace α_r by α_r^Λ which has a similar definition, apart from a delicate modification: This is the largest value sasifying (24) for a *single* co-support Λ corresponding to a co-rank $d-r$, rather than for all possible co-supports leading to this co-rank, as in the definition of α_r (see Definition 1). This means that α_r can be obtained by taking the minimal value of α_r^Λ over all of these co-supports. Since α_r^Λ is a continuous measure in the range $[0,1]$, and since we are about to create histograms of possible values, we perform a uniform quantization of α_r^Λ to $T=100$ discrete levels. The joint distribution of ℓ and α_r^Λ is represented by a p-by-T matrix with entries

$$P_{km}^{(r)} = \Pr\left\{\ell = k, \frac{m-1}{T} \leq \alpha_r^\Lambda < \frac{m}{T}\right\}, \tag{30}$$

Obtaining the entries of the matrix $\mathbf{P}^{(r)}$ requires an exhaustive computation over all possible co-supports with co-rank $d-r$. The joint distributions for the three dictionaries (shown in Fig. 1) and a co-rank of 7 (i.e. $r=2$) are depicted in Fig. 7. We can see that increasing the co-sparsity level typically spreads α_r^Λ towards higher values. This makes sense since the minimization appearing in (24) is performed over smaller index sets.

C. Results of the Analysis Thresholding Revisited

We revisit the results shown in Section III-C and try to explain them in light of the theoretical guarantees derived in Section IV-A. Note that the setup considered in Theorem 2 (projection

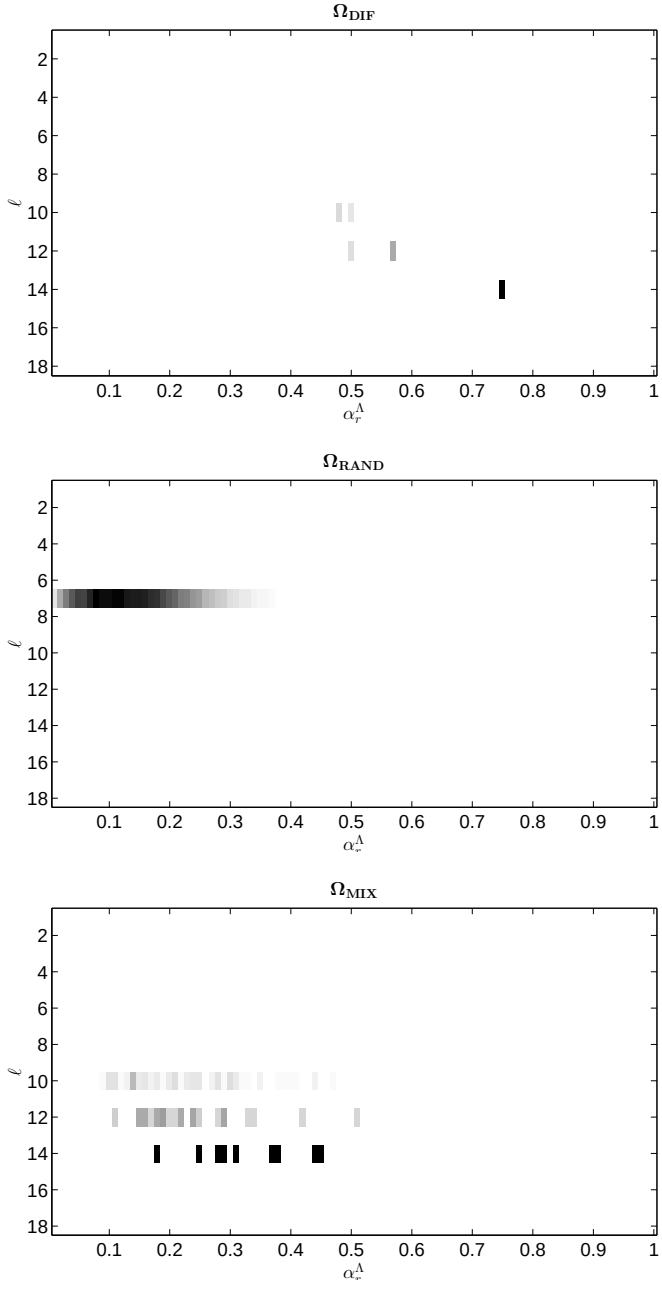

Figure 7. The joint distribution of ℓ and α_r^{Λ} for each type of the analysis dictionaries of size 18×9 that were shown in Fig. 1 and for $r = 2$. Each of these distributions is obtained by an exhaustive computation over all possible subsets of rows from the analysis dictionary with co-rank 7, and is displayed in the form of a matrix $\mathbf{P}^{(2)}$, whose entries where defined in Eq. (30). A darker bin corresponds to a higher value in the joint distribution.

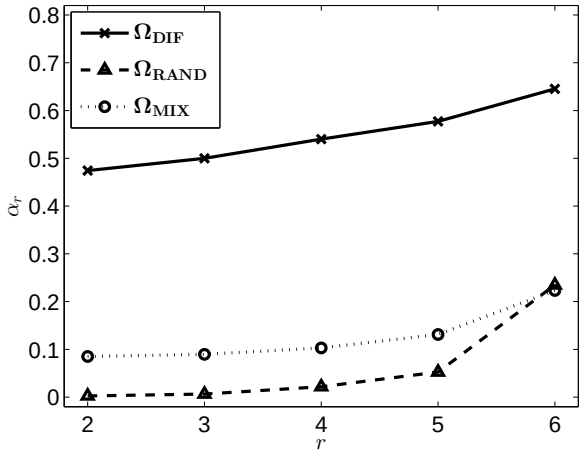

Figure 8. The values of the ROPP constant for each type of the analysis dictionaries of size 18×9 that were shown in Fig. 1 and for varying analysis subspace dimensions r. Each of these values is obtained by an exhaustive minimization over all possible subsets of rows from the analysis dictionary with co-rank $9 - r$.

of a white Gaussian vector $\mathbf{u}$, additive white Gaussian noise) matches completely the one used for the experiments of Section III-C. This will allow us to make the desired connections between the empirical results and the theoretical guarantee. An immediate observation arising from Theorem 2 is that the higher the co-sparsity level ℓ of $\mathbf{x}$ with respect to $\mathbf{\Omega}$, the better the thresholding algorithm is expected to perform in recovering the true co-support. This implies that linear dependencies within $\mathbf{\Omega}$ are highly desired. This stands as a complete contradiction to the intuition gained for the synthesis-based sparsity model, where such dependencies between the atoms lead to a collapse of pursuit algorithms. We also observe that the results of the analysis threshodling algorithm improve as α_r grows. This is closer in spirit to the ERC/RIP rationale, where independencies are encouraged.

Returning to the empirical results of Section III-C, we have already seen in Fig. 2 that $\mathbf{\Omega}_{DIF}$ and $\mathbf{\Omega}_{MIX}$ have the same co-sparsity distribution, where the co-sparsity can be much higher than the co-rank $d-r$. This can explain, at least in part, their superior performance over $\mathbf{\Omega}_{RAND}$, which allows only a constant co-sparsity level $\ell = d - r$. We now turn to examine the value of the ROPP constant for each type of dictionary, with a hope to reveal an additional inherent difference between the dictionaries. These values are shown in Fig. 8 for the three dictionary types and for varying analysis subspace dimensions r. To obtain each of these values we performed an exhaustive minimization over all possible subsets Λ of rows from $\mathbf{\Omega}$ such that

Rank $\{\mathbf{\Omega}_\Lambda\} = d - r$. We can see that $\mathbf{\Omega}_{DIF}$ corresponds to a much higher ROPP constant for all the examined co-ranks, when compared to $\mathbf{\Omega}_{MIX}$ and $\mathbf{\Omega}_{RAND}$. The two latter dictionaries have very low ROPP constants (below 0.14 for $r \leq 5$). Specifically, at a subspace dimension of $r = 2$ that was considered in the experiments of Section III-C, the ROPP constant is 5.6 times higher for $\mathbf{\Omega}_{DIF}$ compared to $\mathbf{\Omega}_{MIX}$ and 202(!) times higher compared to $\mathbf{\Omega}_{RAND}$. We can conclude that the value of the ROPP constant explains the superior behavior of the thresholding algorithm with $\mathbf{\Omega}_{DIF}$ when compared to $\mathbf{\Omega}_{MIX}$, as observed in Fig. 4. This dictionary property also provides additional grounds for the inferior behavior with $\mathbf{\Omega}_{RAND}$.

Next, we turn to examine the theoretical success guarantee provided in Theorem 2. Fig. 9 (top) displays this lower bound on the probability of success for the thresholding algorithm for each of the dictionaries and for varying SNR levels in the range $6dB$ to $74dB$ [3]. To obtain each of the lower bounds that are shown in this figure, we find for each co-sparsity ℓ and each noise ratio σ/σ_u a value of β such that the lower bound for the probability of success provided in Theorem 2 is as tight (i.e. high) as possible. An example of how to choose an optimal value of β was depicted in Fig. 6. Finally, we perform a weighted average of these lower bounds, where the weights are simply the values of the co-sparsity distribution. This process can be described by the following equation:

$$\begin{aligned}
&\Pr\{\text{“Success”}\} = \sum_{k=1}^{p} \Pr\{\ell = k\} \Pr\{\text{“Success”}|\ell = k\} \\
&\geq \sum_{k=1}^{p} \Pr\{\ell = k\} \left[\text{Max} \left\{0, 1 - \sqrt{\frac{8}{\pi\beta_k^2}} \exp\left\{-\frac{\beta_k^2}{8}\right\}\right\}\right]^{p-k+d-r} \left[2Q\left(\frac{\beta_k \sigma}{\alpha_r \sigma_u}\right)\right]^{p-k}, \quad (31)
\end{aligned}$$

where β_k is the value of β that is set for co-sparsity $\ell = k$. These values are chosen such that the arguments inside the sum are maximized for each k separately.

We can see that the resulting lower bounds can provide some insight into the actual performance. They are capable of predicting success with high probability at high SNR levels for $\mathbf{\Omega}_{DIF}$ and $\mathbf{\Omega}_{MIX}$. Another useful property of these bounds is that they clearly predict which dictionary the thresholding algorithm is expected to perform better with and which would probably lead to failure. Note that in our quest for theoretical guarantees we have lost much tightness with respect to the empirical results. This is typical for a theoretical analysis, but as we shall see in a moment, the tightness of the derived bounds can be considerably improved

[3] See Eq. (10) for the definition of SNR and its dependence on σ/σ_u.

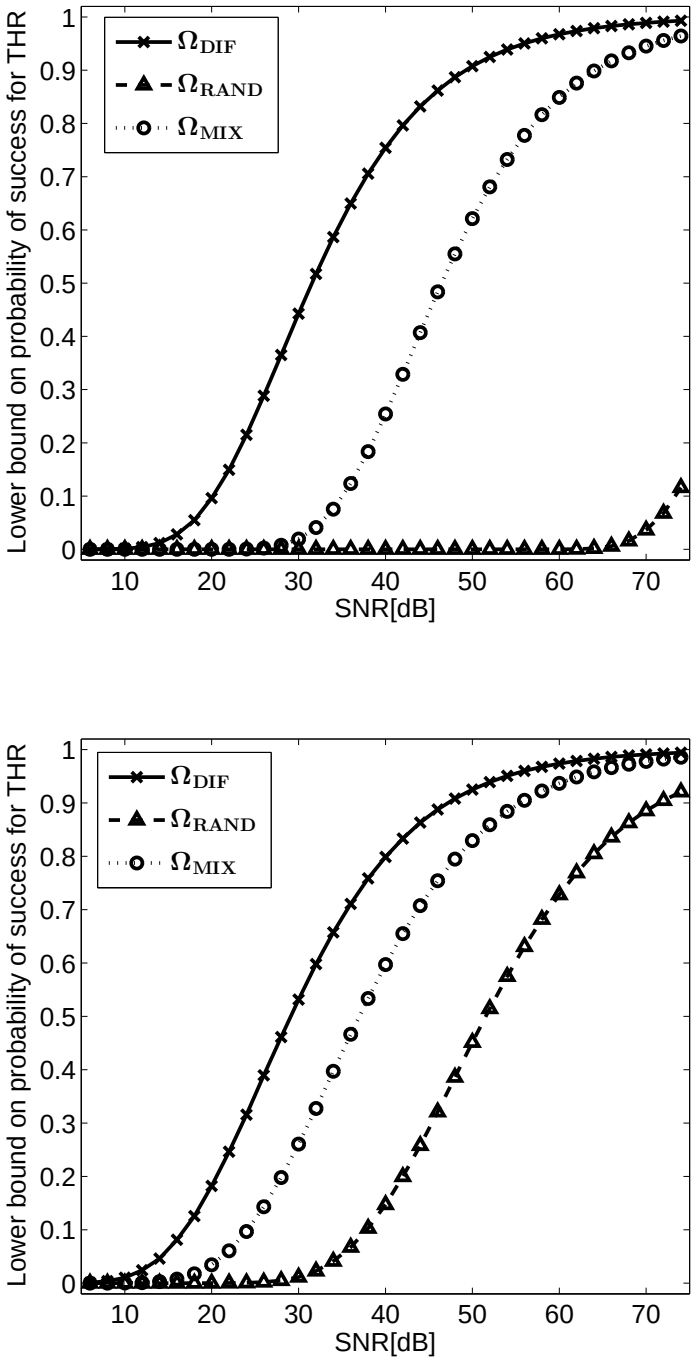

Figure 9. Lower bounds on the probability of success for the thresholding algorithm on the three types of analysis dictionaries of size 18×9 that were shown in Fig. 1 and for varying SNR levels. Top: For each ratio σ/σ_u a lower bound is computed using Eq. (31), where for each co-sparsity level ℓ we choose a value for β such that the resulting bound will be as tight as possible. Bottom: For each ratio σ/σ_u a lower bound is computed using Eq. (32), where an optimal value for β is set for each pair $\ell,\ \alpha_r^{\Lambda}$. As can be seen, the bounds appearing on the right are tighter than those shown on the left.

if we take into account the fact that α_r^{Λ} varies as a function of the co-support, and has a spread of values. Specifically, we can modify the process described in Eq. (31) by replacing the distribution of ℓ and the fixed worst-case value of α_r with the joint distribution of ℓ and α_r^{Λ}, as depicted in Fig. 7. For each such pair and for each noise ratio σ/σ_u we set an optimal value of β as described before, and use the values of the joint distribution as weights for the final average. This means that the process of (31) is replaced by

$$\begin{aligned}\Pr\{\text{“Success”}\} &= \sum_{k=1}^{p}\sum_{m=1}^{T} P_{km}^{(r)} \Pr\left\{\text{“Success”}|\ell=k, \frac{m-1}{T} \le \alpha_r^{\Lambda} < \frac{m}{T}\right\} \\ &\ge \sum_{k=1}^{p}\sum_{m=1}^{T} P_{km}^{(r)} \left[\text{Max}\left\{0, 1-\sqrt{\frac{8}{\pi\beta_{km}^2}}\exp\left\{-\frac{\beta_{km}^2}{8}\right\}\right\}\right]^{p-k+d-r} \left[2Q\left(\frac{\beta_{km}T\sigma}{(m-1)\sigma_u}\right)\right]^{p-k}.\end{aligned} \tag{32}$$

The resulting lower bounds are shown on the bottom of Fig. 9 and as can be seen, they are much tighter than the previous ones appearing in this figure on the top.

Before concluding this section, we bring several additional experiments, this time with higher dimensional signals, in order to demonstrate the behavior of the thresholding algorithm, and the comparison between empirical performance and the theoretical forecasts. We consider signals of dimension $d = 100$ and three types of analysis dictionaries (same as before), each with $p = 200$ atoms. We test denoising setups where the true analysis subspace dimension r varies in the range $[2, 25]$ and the SNR in the range $6dB$ to $75dB$. For each pair of r and noise level σ we generate $N = 1000$ signals. When evaluating the theoretical bounds, we cannot use the value of α_r as exhaustive search for its value is unfeasible. We therefore use the expression given in Eq. (32), where we plug into it an empirical distribution of the values of ℓ and α_r^{Λ} that is computed from the signal examples, instead of the exact one we have used for the low dimensional setups. The empirical ratios of success and their theoretical lower bounds are shown in Fig. 10 for the three types of analysis dictionaries of size 200-by-100. Each of these ratios is displayed as a matrix where white corresponds to one and black corresponds to zero.

Several observations can be made from Fig. 10. First, the general behavior of the three dictionary types remain as before: The performance is best for $\mathbf{\Omega}_{DIF}$, second best for $\mathbf{\Omega}_{MIX}$ and the worse for $\mathbf{\Omega}_{RAND}$, both in terms of the empirical and the theoretical success rates. Secondly, for $\mathbf{\Omega}_{DIF}$ and $\mathbf{\Omega}_{MIX}$ the best performance is obtained for low SNR levels and low subspace dimensions r (the top left corner of the matrix). This is a desired behavior due

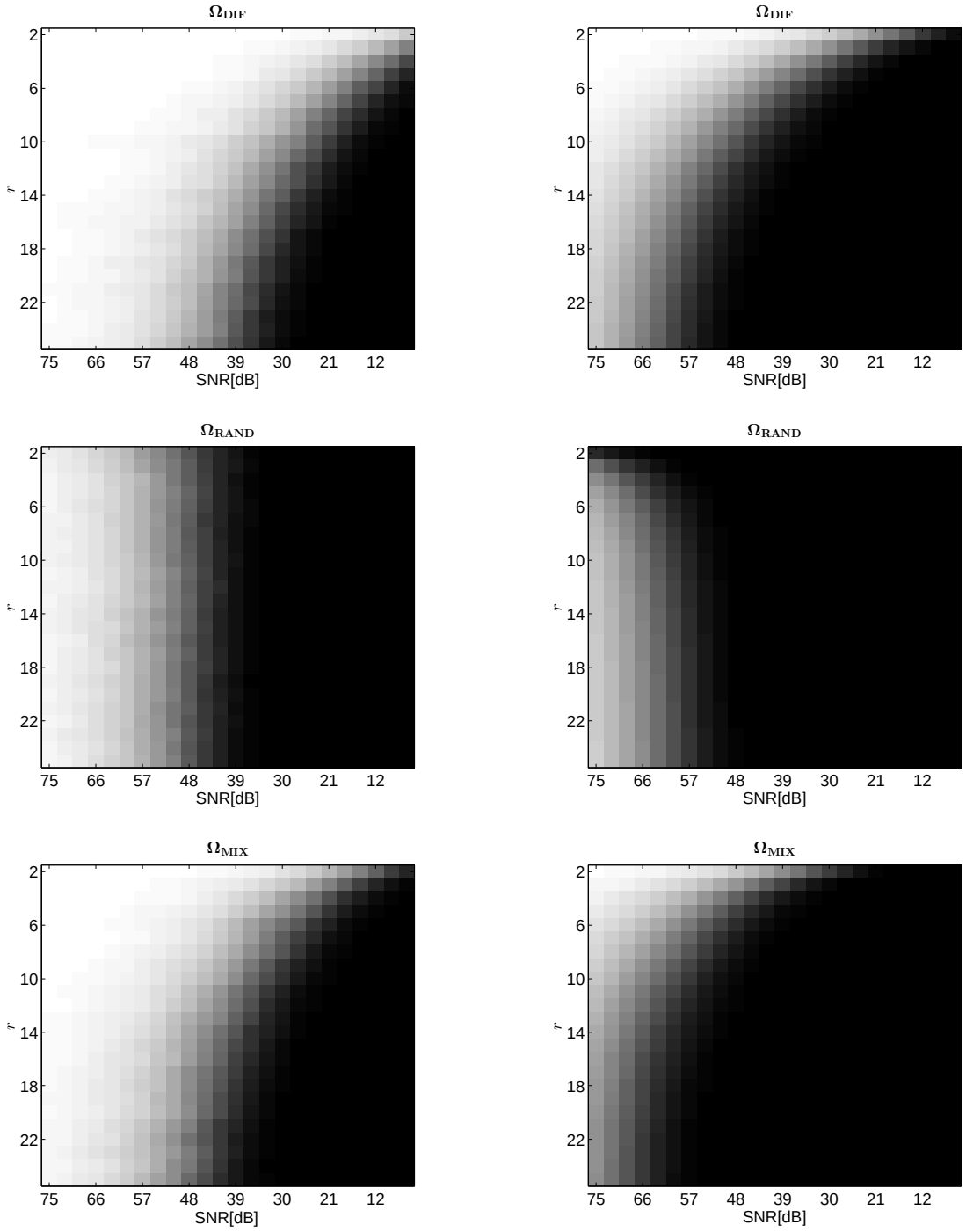

Figure 10. Empirical ratios of success and their theoretical lower bounds for the thresholding algorithm on three types of analysis dictionaries of size 200×100 for varying analysis subspace dimensions r and SNR levels. For each pair of r and SNR we generate $N = 1000$ signals. The theoretical bounds are computed using Eq. (32) by plugging into it the empirical distribution of ℓ and α_r^{Λ}, which is computed from these signals. Left: The empirical ratios of success. Right: The theoretical bounds.

to the fact that we typically want a low subspace dimension, which improves the denoising performance. For $\mathbf{\Omega}_{RAND}$ however, the best theoretical results are obtained for low SNR levels and high values of r (the bottom left corner). The theoretical predictions for this dictionary are less reliable, as we can see that the actual performance is quite similar for all values of r.

V. Relation to Existing Results

There are several exiting contributions in the published literature on developing pursuit algorithms for the co-sparse analysis model and studying their performance from a theoretical stand-point. Here we mention several papers that are of relevance to this work. We provide a brief review of their content, followed by a discussion on the relation to our results.

The first work we briefly refer to is [22], which concentrates on the analysis dictionary learning problem. Two greedy analysis pursuit algorithms are developed for the denoising problem, as part of the overall learning paradigm – these algorithms are the Backward Greedy (BG) and the Optimized BG (OBG). Both these algorithms are constrcuted by imitating synthesis based pursuit methods, and brought without a theoretical justification of any sort. Interestingly, the work in [22] provides an empirical evidence for the positive effect that strong linear dependencies within the analysis dictionary have on the success of pursuit algorithms.

The work of [16], [20] considers a noise-free measurement setup where the co-sparse analysis signal is measured by $\mathbf{y} = \mathbf{M}\mathbf{x}$, from which we would like to recover $\mathbf{x}$. The authors of [16], [20] explore various uniqueness properties of this problem setup and suggest using either an analysis ℓ_1-norm minimization or a Greedy-Analysis-Pursuit (GAP) algorithm (note that GAP is different from the above mentioned BG and OBG - see more in [22]) for recovering the signal. They analyze the performance of these pursuit algorithms for the noise-free setup, deriving a sufficient condition for success of both algorithms in terms of the analysis dictionary $\mathbf{\Omega}$, the true co-support Λ of $\mathbf{x}$ and the null-space of $\mathbf{M}$. Due to its apparent similarity to the ERC for the synthesis model, the derived condition is termed *analysis ERC*.

The theoretical study of analysis ℓ_1-norm based pursuit in a measurement setup is also the main focus of another recent work [21]. This includes the derivation of conditions for noiseless identifiability and robustness to bounded noise, in terms of the sign pattern of $\mathbf{\Omega x}$ and assuming that the null spaces of the measurement matrix $\mathbf{M}$ and the analysis dictionary $\mathbf{\Omega}$ intersect only at the zero vector. Note that all of the resulting conditions in [16], [20], [21] are somewhat

implicit, especially in the latter work, where the condition involves an inner optimization stage for a given sign pattern. This makes the derived conditions hard to interpret.

A different work altogether is proposed in [13]. The authors [13] suggest a hybrid viewpoint to the synthesis and analysis models, where the signal of interest is a synthesis-and-analysis signal, constructed as $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}$ with a sparse synthesis representation $\boldsymbol{\alpha}$. However, this signal is also characterized as an analysis signal in the sense that it has a small ℓ_1 energy in the tail of the analysis representation $\mathbf{D}^T\boldsymbol{\alpha}$. They suggest using an analysis-based approach for recovering the signal from its undersampled and noisy measurements $\mathbf{y} = \mathbf{M}\mathbf{x}+\mathbf{e}$. Their approach is based on ℓ_1-norm sparsity of $\mathbf{D}^T\mathbf{x}$ deriving a theoretical upper bound on the denoising error obtained by ℓ_1 analysis pursuit in this setup. To obtain the desired bound they require the measurement matrix $\mathbf{M}$ to satisfy a certain property adapted to $\mathbf{D}$, termed D-RIP, which is similar to the well-known RIP aside from a delicate modification – instead of bounding the ℓ_2 norm of $\mathbf{M}\mathbf{v}$ for all k-sparse vectors $\mathbf{v}$, the norm of $\mathbf{M}\mathbf{v}$ is bounded for all vectors $\mathbf{v}$ that can be expressed as a linear combination of k columns of $\mathbf{D}$.

The work of [17] suggests a family of new pursuit algorithms for recovering co-sparse analysis signals from their undersampled measurements. These algorithms are analogous to the synthesis-based iterative hard thresholding algorithm, with a modification of the projection step intended for adapting this framework to the analysis model. The authors of [17] present theoretical recovery guarantees for these analysis pursuit algorithms in the noiseless setup, assuming that the measurement matrix satisfies the $\mathbf{\Omega}$-RIP (an analysis counterpart for the D-RIP of [13]).

In this paper we focus on a denoising setup, similar to [22] and assume no measurement matrix. Our focus is the most simple analysis pursuit algorithm – the thresholding. This allows us to remove some of the ambiguities that are present in previous works, where the resulting theoretical conditions mix both the measurement matrix $\mathbf{M}$ and the analysis dictionary $\mathbf{\Omega}$; we focus on internal properties of $\mathbf{\Omega}$ only. Indeed, our derived theoretical guarantees are expressed in terms of the noise level, the co-sparsity ℓ of the signal over $\mathbf{\Omega}$ and internal properties of $\mathbf{\Omega}$. Instead of using dictionary measures that mimick the synthesis counterpart model, as practiced in [20], which uses analysis ERC, or [13], [17], which use RIP-like properties, we suggest a novel measure, termed *Restricted Orthogonal Projection Property* (ROPP), which seems to be more relevant to analysis dictionaries. This property is much more explicit than the one arising

from the theoretical analysis of [21]. Our derived results are simple to interpret, and specifically we see that strong linear dependencies improve the pursuit algorithm's success rate.

VI. Conclusions

In this work we have made an initial attempt at addressing the question of what makes an analysis dictionary suitable for co-sparse estimation. We have concentrated on a denoising setup and considered the use of a thresholding algorithm for the corresponding analysis pursuit problem. Our experiments show that this simple algorithm can perform quite well for certain analysis dictionaries, while failing on others. To better understand this behavior we further explored the performance of this algorithm in the presence of white Gaussian random noise, developing theoretical guarantees for the ability of the algorithm to recover the true underlying co-support. This study reveals two significant properties of an analysis dictionary that are key in dictating whether the pursuit will succeed of fail: The degree of linear dependencies between rows of Ω and the level of independence between subsets of rows and other atoms, a property we termed ROPP. We have found that it is desired to have many linear dependencies, as they increase the co-sparsity level. Similarly, the ROPP constant should be as high as possible. Finally, we have shown how the developed theoretical guarantees can explain our empirical results and predict them quite well. This work gives rise to various open questions that will be the topics of future research. These include topics such as these:

1) While this work concentrated on the thresholding algorithm, a similar theoretical study should be given to other pursuit algorithms. Perhaps the quality measures we identified in this work could be of help in such study.
2) This work defines the success of the pursuit algorithm by the complete identification of the co-support. However, this algorithm may perform rather well (in denoising terms) even in situations where only part of the support has been found. Extending this work to cover such cases would improve our prediction for the range of success of the thresholding algorithm.
3) How could we incorporate the proposed quality measures for Ω directly into the dictionary learning process? By doing so we may design better analysis dictionaries, which will ultimately lead to performance improvement and make the analysis model and its learned dictionary suitable for a wide range of processing applications.

REFERENCES

[1] A.M. Bruckstein, D.L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009.

[2] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, Springer, 2010.

[3] D.L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. on Inf. Theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

[4] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization," *Proc. Nat. Aca. Sci.*, vol. 100, pp. 2197–2202, 2003.

[5] J.A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. on Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[6] E.J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. on Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[7] E.J. Candes, J.K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[8] J. Shtok and M. Elad, "Analysis of the basis pursuit via the capacity sets," *J. Fourier Anal. App.*, vol. 14, no. 5–6, pp. 688–711, 2008.

[9] A. Juditsky and A. Nemirovski, "On verifiable sufficient conditions for sparse signal recovery via l1 minimization," *Mathematical Programming: Series A and B - Special Issue on Optimization and Machine learning*, vol. 127, no. 1, pp. 57–88, 2011.

[10] K. Schnass and P. Vandergheynst, "Average performance analysis for thresholding," *IEEE Signal Process. Lett.*, vol. 14, no. 11, pp. 828–831, 2007.

[11] Z. Ben-Haim, Y.C. Eldar, and M. Elad, "Coherence-based performance guarantees for estimating a sparse vector under random noise," *IEEE Trans. on Signal Processing*, vol. 58, no. 10, pp. 5030–5043, 2010.

[12] M. Elad, P. Milanfar, and R. Rubinstein, "Analysis versus synthesis in signal priors," *Inverse Problems*, vol. 23, no. 3, pp. 947–968, 2007.

[13] E.J. Candes, Y.C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Applied Computational Harmonic Analysis*, vol. 31, no. 1, pp. 59–73, 2011.

[14] S. Nam, M.E. Davies, M. Elad, and R. Gribonval, "Cosparse analysis modeling," in *Proceedings of SAMPTA*, 2011.

[15] G. Peyré and J. Fadili, "Learning analysis sparsity priors," in *Proceedings of SAMPTA*, 2011.

[16] S. Nam, M. Davies, M. Elad, and R. Gribonval, "Cosaprse analysis modeling – uniqueness and algorithms," in *Proceedings of ICASSP*, 2011, pp. 5804–5807.

[17] R. Giryes, S. Nam, R. Gribonval, and M.E. Davies, "Iterative cosparse projection algorithms for the recovery of cosparse vectors," in *Proceedings of EUSIPCO*, 2011.

[18] B. Ophir, M. Elad, N. Bertin, and M.D. Plumbley, "Sequential minimal eigenvalues – an approach to analysis dictionary learning," in *Proceedings of EUSIPCO*, 2011.

[19] R. Gribonval M. Yaghoobi, S. Nam and M.E. Davies, "Analysis operator learning for overcomplete cosparse representations," in *Proceedings of EUSIPCO*, 2011.

[20] S. Nam, M. Davies, M. Elad, and R. Gribonval, "The cosaprse analysis model and algorithms," submitted to *Applied Computational Harmonic Analysis*.

[21] S. Vaiter, G. Peyré, C. Dossal, and J. Fadili, "Robust sparse analysis regulatization," Technical Report HAL-00627452.

[22] R. Rubinstein, T.Faktor, and M. Elad, "Analysis K-SVD: A dictionary-learning algorithm for the analysis sparse model," submitted to *IEEE Trans. on Signal Processing*.

[23] M. Elad, "Sparse representations are most likely to be the sparsest possible," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–12, 2006.

[24] S.S. Chen, D.L. Donoho, and M.A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, pp. 33–61, 1998.

[25] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, pp. 3397–3415, 1999.

[26] Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proceedings of the 27th Asilomar Conference on Signals, Systems and Computers*, 1993, vol. 1, pp. 40–44.

[27] Z. Ŝidák, "Rectangular confidence regions for the means of multivariate normal distributions," *J. Amer. Statist. Assoc.*, vol. 62, no. 318, pp. 626–633, 1967.