

Blind Compressed Sensing

Sivan Gleichman and Yonina C. Eldar, *Senior Member, IEEE*

Abstract—The fundamental principle underlying compressed sensing is that a signal, which is sparse under some basis representation, can be recovered from a small number of linear measurements. However, prior knowledge of the sparsity basis is essential for the recovery process. This work introduces the concept of blind compressed sensing, which avoids the need to know the sparsity basis in both the sampling and the recovery process. We suggest three possible constraints on the sparsity basis that can be added to the problem in order to make its solution unique. For each constraint we prove conditions for uniqueness, and suggest a simple method to retrieve the solution. Under the uniqueness conditions, and as long as the signals are sparse enough, we demonstrate through simulations that without knowing the sparsity basis our methods can achieve results similar to those of standard compressed sensing, which rely on prior knowledge of the sparsity basis. This offers a general sampling and reconstruction system that fits all sparse signals, regardless of the sparsity basis, under the conditions and constraints presented in this work.

I. INTRODUCTION

Sparse signal representations have gained popularity in recent years in many theoretical and applied areas [1]–[6]. Roughly speaking, the information content of a sparse signal occupies only a small portion of its ambient dimension. For example, a finite dimensional vector is sparse if it contains a small number of nonzero entries. It is sparse under a basis if its representation under a given basis transform is sparse. An analog signal is referred to as sparse if, for example, a large part of its bandwidth is not exploited [4], [7]. Other models for analog sparsity are discussed in detail in [5], [6], [8].

Compressed sensing (CS) [2], [3] focuses on the role of sparsity in reducing the number of measurements needed to represent a finite dimensional vector $x \in \mathbb{R}^m$. The vector x is measured by $b = Ax$, where A is a matrix of size $n \times m$, with $n \ll m$. In this formulation, determining x from the given measurements b is ill posed in general, since A has fewer rows than columns and is therefore non-invertible. However, if x is known to be sparse in a given basis P , then under additional mild conditions on A [9]–[11], the measurements b determine x uniquely as long as n is large enough. This concept was also recently expanded to include sub-Nyquist sampling of structured analog signals [4], [6], [12].

In principle, recovery from compressed measurements is NP-hard. Nonetheless, many suboptimal methods have been proposed to approximate its solution [1]–[3], [13]–[15]. These algorithms recover the true value of x when x is sufficiently sparse and the columns of A are incoherent [1], [9]–[11],

[13]. However, all known recovery approaches use the prior knowledge of the sparsity basis P .

Dictionary learning (DL) [16]–[20] is another application of sparse representations. In DL, we are given a set of training signals, formally the columns of a matrix X . The goal is to find a dictionary P , such that the columns of X are sparsely represented as linear combinations of the columns of P . In [17], the authors study conditions under which the DL problem yields a unique solution for the given training set X .

In this work we introduce the concept of blind compressed sensing (BCS), in which the goal is to recover a high-dimensional vector x from a small number of measurements, where the only prior is that there exists some basis in which x is sparse. We refer to our setting as blind, since we do not require knowledge of the sparsity basis for the sampling or the reconstruction. This is in sharp contrast to CS, in which recovery necessitates this knowledge. Our BCS framework combines elements from both CS and DL. On the one hand, as in CS and in contrast to DL, we obtain only low dimensional measurements of the signal. On the other hand, we do not require prior knowledge of the sparsity basis which is similar to the DL problem. The goal of this work is to investigate the basic conditions under which blind recovery from compressed measurements is possible theoretically, and to propose concrete algorithms for this task.

Since the sparsity basis is unknown, the uncertainty about the signal x is larger in BCS than in CS. A straightforward solution would be to increase the number of measurements. However, we show that no rate increase can be used to determine x , unless the number of measurements is equal the dimension of x . Furthermore, we prove that even if we have multiple signals that share the same (unknown) sparsity basis, as in DL, BCS remains ill-posed. In order for the measurements to determine x uniquely we need an additional constraint on the problem. To prove the concept of BCS we begin by discussing two simple constraints on the sparsity basis, which enable blind recovery of a single vector x . We then turn to our main contribution, which is a BCS framework for structured sparsity bases. In this setting, we show that multiple vectors sharing the same sparsity pattern are needed to ensure recovery. For all of the above formulations we demonstrate via simulations that when the signals are sufficiently sparse the results of our BCS methods are similar to those obtained by standard CS algorithms which use the true, though unknown in practice, sparsity basis. When relying on the structural constraint we require in addition that the number of signals must be large enough. However, the simulations show that the number of signals needed is reasonable and much smaller than that used for DL [21]–[24].

The first constraint on the basis we consider relies on the fact that over the years there have been several bases that

have been considered "good" in the sense that they are known to sparsely represent many natural signals. These include, for example, various wavelet representations [25] and the discrete-cosine transform (DCT) [26]. We therefore treat the setting in which the unknown basis P is one of a finite and known set of bases. We develop uniqueness conditions and a recovery algorithm by treating this formulation as a series of CS problems. To widen the set of possible bases that can be treated, the next constraint allows P to contain any sparse enough combination of the columns of a given dictionary. We show that the resulting CS problem can be viewed within the framework of standard CS, or as DL with a sparse dictionary [23]. We compare these two approaches for BCS with a sparse basis. For both classes of constraints we show that a Gaussian random measurement matrix satisfies the uniqueness conditions we develop with probability one.

Our main contribution is inspired by multichannel systems, where the signals from each channel are sparse under separate bases. In our setting this translates to the requirement that P is block diagonal. For simplicity, and following several previous works [27]–[29], we impose in addition that P is orthogonal. We then choose to measure the set of signals X by a measurement matrix A consisting of a union of orthogonal bases. This choice has been used in previous CS and DL works as well [21], [22], [30]–[32]. For technical reasons we also choose the number of blocks in P as an integer multiple of the number of bases in A . Using this structure we develop uniqueness results as well as a concrete recovery algorithm. The uniqueness condition follows from reformulating the BCS problem within the framework of DL and then relying on results obtained in that context. In particular, we require an ensemble of signals X , all sparse in the same basis. As we show, a suitable choice of random matrix A satisfies the uniqueness conditions with probability 1.

Unfortunately, the reduction to an equivalent DL problem which is used for the uniqueness proof, does not lead to a practical recovery algorithm. This is due to the fact that it necessitates resolving the signed permutation ambiguity, which is inherent in DL. Instead, we propose a simple and direct algorithm for recovery, which we refer to as the orthogonal block diagonal BCS (OBD-BCS) algorithm. This method finds $X = PS$ by computing a basis P and a sparse matrix S using two alternating steps. The first step is sparse coding, in which P is fixed and S is updated using a standard CS algorithm. In the second step S is fixed and P is updated using several singular value decompositions (SVD).

The remainder of the paper is organized as follows. In Section II we review the fundamentals of CS and define the BCS problem. In Section III we prove that BCS is ill posed by showing that it can be interpreted as a certain ill-posed DL problem. In Sections IV, V, VI we consider the three constrained BCS problems respectively. A comparison between the different approaches is provided in Section VII.

II. BCS PROBLEM DEFINITION

A. Compressed Sensing

We start by shortly reviewing the main results in the field of CS needed for our derivations. The goal of CS is to reconstruct

a vector $x \in \mathbb{R}^m$ from measurements $b = Ax$, where $A \in \mathbb{R}^{n \times m}$ and $n \ll m$. This problem is ill posed in general and therefore has infinitely many possible solutions. In CS we seek the sparsest solution:

$$\hat{x} = \arg \min \|x\|_0 \quad \text{s.t.} \quad b = Ax, \quad (1)$$

where $\|\cdot\|_0$ is the ℓ_0 semi-norm which counts the number of nonzero elements of the vector. This idea can be generalized to the case in which x is sparse under a given basis P , so that there is a sparse vector s such that $x = Ps$. Problem (1) then becomes

$$\hat{s} = \arg \min \|s\|_0 \quad \text{s.t.} \quad b = APs, \quad (2)$$

and the reconstructed signal is $\hat{x} = P\hat{s}$. When the maximal number of nonzero elements in s is known to equal k , we may consider the objective

$$\hat{s} = \arg \min \|b - APs\|_2^2 \quad \text{s.t.} \quad \|s\|_0 \leq k. \quad (3)$$

An important question is under what conditions (1)–(3) have a unique solution. In [9] the authors define the *spark* of a matrix, denoted by $\sigma(\cdot)$, which is the smallest possible number of linearly dependent columns. They prove that if s is k -sparse, and $\sigma(AP) \geq 2k$, then the solution to (2), or equivalently (3), is unique. Unfortunately, calculating the spark of a matrix is a combinatorial problem. However, it is often bounded by the *mutual coherence* [9], which can be calculated easily. Denoting the i th column of a matrix D by d_i , the mutual coherence of D is given by

$$\mu(D) = \max_{i \neq j} \frac{|d_i^T d_j|}{\|d_i\|_2 \|d_j\|_2}.$$

It is easy to see that $\sigma(D) \geq 1 + \frac{1}{\mu(D)}$. Therefore, a sufficient condition for the uniqueness of the solutions to (2) or (3) is

$$k \leq \frac{1}{2} \left(1 + \frac{1}{\mu(AP)} \right).$$

Although the uniqueness condition involves the product AP , some CS methods are universal. This means that by constructing a suitable measurement matrix A , uniqueness is guaranteed for any fixed orthogonal basis P . In such cases knowledge of P is not necessary for the sampling process. One way to achieve this universality property with probability 1 relies on the next proposition.

Proposition 1. *If A is an i.i.d. Gaussian random matrix of size $n \times m$, where $n < m$, then $\sigma(AP) = n + 1$ with probability 1 for any fixed orthogonal basis P .*

Proof: Due to the properties of Gaussian random variables and since P is orthogonal, the product AP is also an i.i.d. Gaussian random matrix. Since any n , or less, i.i.d. Gaussian vectors in \mathbb{R}^n are linearly independent with probability 1, $\sigma(AP) > n$ with probability 1. On the other hand, more than n vectors in \mathbb{R}^n are always linearly dependent, therefore $\sigma(AP) = n + 1$. ■

According to Proposition 1 if A is an i.i.d. Gaussian matrix and the number of nonzero elements in s is $k \leq n/2$, then the uniqueness of the solution to (2) or (3) is guaranteed with probability 1 for any fixed orthogonal basis P (see also [33]).

Problems (2) and (3) are NP-hard in general. Many sub-optimal methods have been proposed to approximate their solutions, such as [1]–[3], [13]–[15]. These algorithms can be divided into two main approaches: greedy algorithms and convex relaxation methods. Greedy algorithms approximate the solution by selecting the indices of the nonzero elements in \hat{s} sequentially. One of the most common methods of this type is orthogonal matching pursuit (OMP) [13]. Convex relaxation approaches change the objective in (2) to a convex problem. The most common of these methods is basis pursuit (BP) [15], which considers the problem:

$$\hat{s} = \arg \min \|s\|_1 \quad \text{s.t.} \quad b = APs. \quad (4)$$

Under suitable conditions on the product AP and the sparsity level of the signals, both the greedy algorithms and the convex relaxation methods recover the true value of s . For instance, both OMP and BP recover the true value of s when the number of nonzero elements in s is no more than $\frac{1}{2}(1 + \frac{1}{\mu(AP)})$ [1], [9]–[11], [13].

B. BCS Problem Formulation

Even when the universality property is achieved in CS, all existing algorithms require the knowledge of the sparsity basis P for the reconstruction process. The idea of BCS is to avoid entirely the need of this prior knowledge. That is, perform both the sampling and the reconstruction of the signals without knowing under which basis they are sparse.

This problem seems impossible at first, since every signal is sparse under a basis that contains the signal itself. This would imply that BCS allows reconstruction of any signal from a small number of measurements without any prior knowledge, which is clearly impossible. Our approach then, is to sample an ensemble of signals that are all sparse under the same basis. Later on we revisit problems with only one signal, but with additional constraints.

Let $X \in \mathbb{R}^{m \times N}$ denote a matrix whose columns are the original signals, and let $S \in \mathbb{R}^{m \times N}$ denote the matrix whose columns are the corresponding sparse vectors, such that $X = PS$ for some basis $P \in \mathbb{R}^{m \times m}$. The signals are all sampled using a measurement matrix $A \in \mathbb{R}^{n \times m}$, producing the matrix $B = AX$. For the measurements to be compressed the dimensions should satisfy $n < m$, where the compression ratio is $L = m/n$. Following [17], [24] we assume the maximal number of nonzero elements in each of the columns of S , is known to equal k . We refer to such a matrix S as a k -sparse matrix. The BCS problem can be formulated as follows.

Problem 2. *Given the measurements B and the measurement matrix A find the signal matrix X such that $B = AX$ where $X = PS$ for some basis P and k -sparse matrix S .*

Note that our goal is not to find the basis P and the sparse matrix S . We are only interested in the product $X = PS$. In fact, for a given matrix X there is more than one pair of matrices P and S such that $X = PS$. Here we focus on the question of whether X can be recovered given the knowledge that such a pair exists for X .

III. UNIQUENESS

We now discuss BCS uniqueness, namely the uniqueness of the signal matrix X which solves Problem 2. Unfortunately, although Problem 2 seems quite natural, its solution is not unique for any choice of measurement matrix A , for any number of signals and any sparsity level. We prove this result by reducing the problem to an equivalent one, using the field of DL, and proving that the solution to the equivalent problem is not unique.

In Section III-A we review results in the field of DL needed for our derivation. In Section III-B we use these results to prove that the BCS problem does not have a unique solution. In Sections IV, V, VI we suggest several constraints on the basis P that ensure uniqueness.

A. Dictionary Learning (DL)

The field of DL [16]–[20] focuses on finding a sparse matrix $S \in \mathbb{R}^{m \times N}$ and a dictionary $D \in \mathbb{R}^{n \times m}$ such that $B = DS$ where only $B \in \mathbb{R}^{n \times N}$ is given. Usually in DL the dimensions satisfy $n \ll m$. BCS can be viewed as a DL problem with $D = AP$ where A is known and P is an unknown basis. Thus, one may view BCS as a DL problem with a constrained dictionary. However, there is an important difference in the output of DL and BCS. DL provides the dictionary $D = AP$ and the sparse matrix S . On the other hand, in BCS we are interested in recovering the unknown signals $X = PS$. Therefore, after performing DL some postprocessing is needed to retrieve P from D . This is an important distinction which, as we show in Section VI-B, makes it hard to directly apply DL algorithms.

An important question is the uniqueness of the DL factorization. That is, given a matrix $B \in \mathbb{R}^{n \times N}$ what are the conditions for the uniqueness of the pair of matrices $D \in \mathbb{R}^{n \times m}$ and $S \in \mathbb{R}^{m \times N}$ such that $B = DS$ where S is k -sparse. Note that if some pair D, S satisfies $B = DS$, then scaling and signed permutation of the columns of D and rows of S respectively do not change the product $B = DS$. Therefore, there cannot be a unique pair D, S . In the context of DL the term uniqueness refers to uniqueness up to scaling and signed permutation. In fact in most cases without loss of generality we can assume the columns of the dictionary have unit norm, such that there is no ambiguity in the scaling, but only in the signed permutation.

Conditions for DL uniqueness when the dictionary D is orthogonal or just square are provided in [28] and [29]. However, in BCS $D = AP$ is in general rectangular. In [17] the authors prove sufficient conditions on D and S for the uniqueness of a general DL. We refer to the condition on D as the *spark condition* and to the conditions on S as the *richness conditions*. The main idea behind these conditions is that D should satisfy the condition for CS uniqueness, and that the columns of S should be diverse regarding both the locations and the values of the nonzero elements. More specifically, the conditions for DL uniqueness are:

- The spark condition: $\sigma(D) \geq 2k$.
- The richness conditions:
 - 1) All the columns of S have exactly k nonzero elements.

- 2) For each possible k -length support there are at least $k + 1$ columns in S .
- 3) Any $k + 1$ columns in S , which have the same support, span a k -dimensional space.
- 4) Any $k + 1$ columns in S , which have different supports, span a $(k + 1)$ -dimensional space.

According to the second of the richness conditions the number of signals, that is the number of columns in S , must be at least $\binom{m}{k}(k + 1)$. Nevertheless, it was shown in [17] that in practice far fewer signals are needed. Heuristically, the number of signals should grow at least linearly with the length of the signals. It was also shown in [17] that DL algorithms perform well even when there are at most k nonzero elements in the columns of S instead of exactly k .

B. BCS Uniqueness

Under the conditions above the DL solution given the measurements B is unique. That is, up to scaling and signed permutations there is a unique pair D, S such that $B = DS$ and S is k -sparse. Since we are interested in the product PS and not in P or S themselves, without loss of generality we can always assume that the columns of P are scaled so that the columns of $D = AP$ have unit norm. This way there is no ambiguity in the scaling of D and S , but only in their signed permutation. That is, applying DL on B provides $\tilde{D} = APQ$ and $\tilde{S} = Q^T S$ for some unknown signed permutation matrix Q . A signed permutation matrix is a column (or row) permutation of the identity matrix, where the sign of each column (or row) can change separately. In other words, it has only one nonzero element, equal ± 1 , in each column and each row. Any signed permutation matrix is obviously orthogonal.

If we can find the basis $\tilde{P} = PQ$ out of \tilde{D} , then we can recover the correct signal matrix by:

$$\tilde{P}\tilde{S} = PQQ^T S = PS = X.$$

Therefore, under the uniqueness conditions for DL on S and $D = AP$ Problem 2 is equivalent to the following problem.

Problem 3. Given $\tilde{D} \in \mathbb{R}^{n \times m}$ and $A \in \mathbb{R}^{n \times m}$, where $n < m$, find a basis \tilde{P} such that $\tilde{D} = A\tilde{P}$.

We therefore focus on the uniqueness of Problem 3. Since $n < m$ the matrix A has a null space. As we now show, even with the constraint that \tilde{P} is a basis there is still no unique solution.

To see that assume \tilde{P}_1 is a basis, i.e., has full rank, and satisfies $\tilde{D} = A\tilde{P}_1$. Decompose \tilde{P}_1 as $\tilde{P}_1 = P_{N^\perp} + P_N$ where the columns of P_N are in $N(A)$, the null space of A , and those of P_{N^\perp} are in its orthogonal complement $N(A)^\perp$. Note that necessarily $P_N \neq 0$, otherwise the matrix $\tilde{P}_1 = P_{N^\perp}$ is in $N(A)^\perp$ and has full rank. However, since the dimension of $N(A)^\perp$ is at most $n < m$, it contains at most n linearly independent vectors. Therefore, there is no $m \times m$ full rank matrix whose columns are all in $N(A)^\perp$.

Next define the matrix $\tilde{P}_2 = P_{N^\perp} - P_N$ which is different from \tilde{P}_1 , but it is easy to see that $\tilde{D} = A\tilde{P}_2$. Moreover, since

the columns of P_N are perpendicular to the columns of P_{N^\perp} ,

$$\tilde{P}_1^T \tilde{P}_1 = \tilde{P}_2^T \tilde{P}_2 = \|P_{N^\perp}\|_F^2 + \|P_N\|_F^2.$$

A square matrix P has full rank if and only if $P^T P$ has full rank. Therefore, since \tilde{P}_1 has full rank and $\tilde{P}_2^T \tilde{P}_2 = \tilde{P}_1^T \tilde{P}_1$, \tilde{P}_2 also has full rank. So that both \tilde{P}_1 and \tilde{P}_2 are solutions to Problem 3. In fact there are many more solutions; some of them can be found by changing the signs of only part of the columns of P_N .

We now return to the original BCS problem, as defined in Problem 2. We just proved that when the DL solution given B is unique, Problem 2 is equivalent to Problem 3 which has no unique solution. Obviously if the DL solution given B is not unique, then BCS will not be unique. Therefore, Problem 2 has no unique solution for any choice of parameters.

In order to guarantee a unique solution we need an additional constraint. We next discuss constraints on P that can render the solution to Problem 3 unique, and therefore in addition to the richness conditions on S and the spark condition on AP they guarantee the uniqueness of the solution to Problem 2. Although there are many possible constraints, we focus below on the following.

- 1) P is one of a finite and known set of bases.
- 2) P is sparse under some known dictionary.
- 3) P is orthogonal and has a block diagonal structure.

The motivation for these constraints comes from the uniqueness of Problem 3. Nonetheless, we provide conditions under which the solution to Problem 2 with constraints 1 or 2 is unique even without DL uniqueness. In fact, under these conditions the solution to Problem 2 is unique even when $N = 1$, so that there is only one signal.

In the next sections we consider each one of the constraints, prove conditions for the uniqueness of the constrained BCS solution, and suggest a method to retrieve the solution. Table I summarizes these three approaches.

IV. FINITE SET OF BASES

One way to guarantee a unique solution to Problem 3 is to limit the number of possible bases \tilde{P} to a finite set of bases, and require that these bases are different from one another under the measurement matrix A . Since \tilde{P} in Problem 3 is a column signed permutation of P in Problem 2, by limiting P to a finite set we also limit the possible \tilde{P} to a finite set. The new constrained BCS, instead of Problem 2, is then:

Problem 4. Given the measurements B , the measurement matrix A and a finite set of bases Ψ , find the signal matrix X such that $B = AX$ and $X = PS$ for some basis $P \in \Psi$ and k -sparse matrix S .

The motivation behind Problem 4 is that over the years a variety of bases were proven to lead to sparse representations of many natural signals, such as wavelet [25] and DCT [26]. These bases have fast implementations and are known to fit many types of signals. Therefore, when the basis is unknown it is natural to try one of these choices.

TABLE I
SUMMARY OF CONSTRAINTS ON P

The constraint	Conditions for uniqueness	Algorithm
Finite Set - Section IV P is in a given finite set of possible bases Ψ .	<ul style="list-style-type: none"> • $\sigma(AP) \leq 2k$ for any $P \in \Psi$. • A is k-rank preserving of Ψ (Definition 5). 	<ul style="list-style-type: none"> • F-BCS - Solving (6) or (7) for each $P \in \Psi$ using a standard CS algorithm, and choosing the best solution.
Sparse Basis - Section V P is k_P -sparse under a given dictionary Φ .	<ul style="list-style-type: none"> • $\sigma(A\Phi) \geq 2k_P k$. 	<ul style="list-style-type: none"> • Direct method - Solving (9) or (10) using a standard CS algorithm, where the recovery is $X = \Phi C$. • Sparse K-SVD - Using sparse K-SVD algorithm [23] to retrieve S, Z, where the recovery is $X = \Phi Z S$.
Structure - Section VI P is orthogonal $2L$ -block diagonal.	<ul style="list-style-type: none"> • The richness conditions on S. • A is a union of L orthogonal bases. • $\sigma(AP) = n + 1$. • A is not inter-block diagonal (Definition 10). 	<ul style="list-style-type: none"> • OBD-BCS - Updating S and P alternately according to the algorithm in Table IV, where the recovery is $X = PS$.

A. Uniqueness Conditions

We now show that under proper conditions the solution to Problem 4 is unique even when there is only one signal, namely $N = 1$. In this case instead of the matrices X, S, B we deal with the vectors x, s, b respectively.

Assume x is a solution to Problem 4. That is, x is k -sparse under $P \in \Psi$ and satisfies $b = Ax$. Uniqueness is achieved if there is no $\bar{x} \neq x$ which is k -sparse under a basis $\bar{P} \in \Psi$ and also satisfies $b = A\bar{x}$. We first require that $\sigma(AP) \geq 2k$; otherwise even if $\bar{P} = P$ there is no unique solution [9]. Since the real sparsity basis P is unknown we require that $\sigma(AP) \geq 2k$ for any $P \in \Psi$.

Next we write $x = Ps = P_T s_T$, where T is the index set of the nonzero elements in s with $|T| \leq k$, s_T is the vector of nonzero elements in s , and P_T is the sub-matrix of P containing only the columns with indices in T . If \bar{x} is also a solution to Problem 4 then $\bar{x} = \bar{P}\bar{s} = \bar{P}_J \bar{s}_J$, where J is the index set of the nonzero elements in \bar{s} , and $|J| \leq k$. Moreover, $b = A\bar{P}_J \bar{s}_J = AP_T s_T$, which implies that the matrix $A[P_T, \bar{P}_J]$ has a null space. This null space contains the null space of $[P_T, \bar{P}_J]$. By requiring

$$\text{rank}(A[P_T, \bar{P}_J]) = \text{rank}[P_T, \bar{P}_J], \quad (5)$$

we guarantee that the null space of $A[P_T, \bar{P}_J]$ equals the null space of $[P_T, \bar{P}_J]$. Therefore, under (5), $A\bar{P}_J \bar{s}_J = AP_T s_T$ if and only if $\bar{P}_J \bar{s}_J = P_T s_T$, which implies $\bar{x} = x$.

Therefore, in order to guarantee the uniqueness of the solution to Problem 4 in addition to the requirement that $\sigma(AP) \geq 2k$ for any $P \in \Psi$, we require that any two index sets T, J of size k and any two bases $P, \bar{P} \in \Psi$ satisfy (5).

Definition 5. A measurement matrix A is k -rank preserving of the bases set Ψ if any two index sets T, J of size k and any two bases $P, \bar{P} \in \Psi$ satisfy (5).

The conditions for the uniqueness of the solution to Problem 4 are therefore: $\sigma(AP) \geq 2k$ for any $P \in \Psi$, and A is k -rank preserving of the set Ψ . In order to satisfy the first condition with probability 1, according to Section II-A we can require all $P \in \Psi$ to be orthogonal and generate A from an i.i.d. Gaussian distribution. However, since the number of bases is finite, we can instead verify the first condition is satisfied by checking the spark of all the products AP .

Alternatively, one can bound the spark of these matrices using their mutual coherence.

It is easy to see that any full column rank matrix A is k -rank preserving for any k and any set Ψ . However, in our case A is rectangular and therefore does not have full column rank. In order to guarantee that A is k -rank preserving with probability 1 we rely on the following proposition:

Proposition 6. An i.i.d Gaussian matrix A of size $n \times m$ is with probability 1 k -rank preserving of any fixed finite set of bases and any $k \leq n/2$.

Proof: If $n \geq m$ then A has full column rank with probability 1, and is therefore k -rank preserving with probability 1. We therefore focus on the case where $n < m$. Assume T, J are index sets of size k , and $P, \bar{P} \in \Psi$. Denote $r = \text{rank}[P_T, \bar{P}_J]$. We then need to prove that $\text{rank}(A[P_T, \bar{P}_J]) = r$.

Perform a Gram Schmidt process on the columns of $[P_T, \bar{P}_J]$ and denote the resulting matrix by G . G is then an $m \times r$ matrix with orthonormal columns, with $\text{rank}(G) = r$ and $\text{rank}(AG) = \text{rank}(A[P_T, \bar{P}_J])$. Next we complete G to an orthogonal matrix G_u by adding columns. According to Proposition 1 since A is an i.i.d Gaussian matrix and G_u is orthogonal $\sigma(AG_u) = n + 1$ with probability 1. Therefore, with probability 1 any t columns of AG_u are linearly independent, with $t \leq n$. In particular, with probability 1 the columns of AG are linearly independent, so that $\text{rank}(AG) = r$, completing the proof. ■

Until now we proved conditions for the uniqueness of Problem 4 when there is only one signal $N = 1$. The same conditions are true for $N > 1$ since we can look at every signal separately. However, since all the signals are sparse under the same basis, if $N > 1$ then the condition that A must be k -rank preserving can be relaxed.

For instance, consider the case where there are only two index sets T, J and two bases $P, \bar{P} \in \Psi$ (P is the real sparsity basis) that do not satisfy (5). In this case if we have many signals with different sparsity patterns, then only a small portion of them fall in the problematic index set, and therefore might falsely indicate that \bar{P} is the sparsity basis. However, most of the signals correspond to index sets that satisfy (5), and therefore these signals indicate the correct basis. The selection of the sparsity bases is done according to the majority of signals and therefore the correct basis is selected.

Another example is the case where there are enough diverse signals such that the richness conditions on S are satisfied. In this case it is enough to require that for any two bases $P, \bar{P} \in \Psi$ the matrices AP and $A\bar{P}$ are different from one another even under scaling and signed permutation of the columns. This way we guarantee that the problem equivalent to Problem 4 under the richness and spark conditions has a unique solution, and therefore Problem 4 also has a unique solution.

Problem 4 can also be viewed as a CS problem with a block sparsity constraint [34], [35]. That is, if $\Psi = \{P_1, P_2, \dots\}$ then the desired signal matrix can be written as

$$X = [P_1, P_2, \dots] \begin{bmatrix} S_1 \\ S_2 \\ \vdots \end{bmatrix},$$

where only one of the submatrices S_i is not all zeros. In contrast to the usual block sparsity constraint here the sub-matrix S_i which is not zero is itself sparse. However, the uniqueness conditions which are implied from this block sparsity CS approach are too strong comparing to our BCS approach. For instance, they require all $P_j \in \Psi$, to be incoherent, whereas the BCS uniqueness is not disturbed by coherent bases. In fact the solution is unique even if the bases in Ψ equal one another. This is because here we are not interested in recovering S_i but rather $P_i S_i$.

B. The F-BCS Method

The uniqueness conditions we discussed lead to a straightforward method for solving Problem 4. We refer to this method as F-BCS which stands for finite BCS. When $N = 1$, F-BCS solves a CS problem for each $P \in \Psi$

$$\hat{s} = \arg \min_s \|s\|_0 \text{ s.t. } b = APs, \quad (6)$$

and chooses the sparsest \hat{s} . Under the uniqueness conditions it is the only one with no more than k nonzero elements. Therefore if we know the sparsity level k we can stop the search when we found a sparse enough \hat{s} . The recovered signal is $x = P\hat{s}$ where P is the basis corresponding to the \hat{s} we chose. When k is known an alternative method is to solve for each $P \in \Psi$

$$\hat{s} = \arg \min_s \|b - APs\|_2^2 \text{ s.t. } \|s\|_0 < k, \quad (7)$$

and choose \hat{s} that minimizes $\|b - AP\hat{s}\|_2^2$. In the noiseless case this minimum is zero for the correct basis P .

When $N > 1$ we can solve either (6) or (7) for each of the signals and select the sparsity basis according to the majority.

The solution to problems (6) and (7) can be approximated using one of the standard CS algorithms. Since these algorithms are suboptimal, there is no guarantee that they provide the correct solution x , even for the correct basis P . In general, when k is small enough relative to n these algorithms are known to perform very well. Moreover, when $N > 1$, P is selected according to the majority of signals, and therefore if the CS algorithm did not work well on a few of the signals it will not effect the recovery of the rest of the signals.

TABLE II
F-BCS SIMULATION RESULTS

SNR	Miss Detected	Average Error
∞	0%	$10^{-14}\%$
30dB	0%	1.3%
25dB	0%	2.7%
20dB	0%	5.4%
15dB	1%	11.6%
10dB	12%	22.5%
5dB	25%	40.1%

C. F-BCS Simulation Results

We now demonstrate the F-BCS method in simulation. We chose the set of bases Ψ to contain 5 bases of size 64×64 : the identity, DCT [26], Haar wavelet, Symlet wavelet and Biorthogonal wavelet [25]. 100 signals of length 64 were created randomly by generating random sparse vectors and multiplying them by the Biorthogonal wavelet basis in Ψ . Each sparse vector contained up to 6 nonzero elements in uniformly random locations, and values from a normal distribution.

The measurement matrix A was an i.i.d Gaussian matrix of size 32×64 . The measurements were calculated first without noise, that is $B = AX$, and then with additive Gaussian noise with varying SNR from 30dB to 5dB. For each noise level the F-BCS method was performed, where the CS algorithm we used was OMP [13].

Table II summarizes the results. For all noise levels the basis selection according to the majority was correct. The miss detected column in the table contains the percentage of signals that indicated a false basis. The average error column contains the average reconstruction error, calculated as the average of

$$e_i = \frac{\|x_i - \hat{x}_i\|_2}{\|x_i\|_2} \quad (8)$$

where x_i, \hat{x}_i are the columns of the real signal matrix X and the reconstructed signal matrix \hat{X} respectively. The average is performed only on the signals that indicated the correct basis. The reconstruction of the rest of the signals obviously failed. As can be seen from Table II in the noiseless case the recovery is perfect and the error grows with the noise level. For high SNR there are no false reconstructions, but as the SNR decreases beyond 15dB the percentage of false reconstructions increases. In these cases, one should use more than one signal, such that if one of the signals failed there will be an indication for this through the rest of the signals.

Another simulation we performed investigated the influence of the sparsity level k , which is the number of nonzero elements in S . The settings of this simulation were the same as those of the first simulation, only this time there was no noise added to the measurements, and k was gradually increased from 1 to 32. For each sparsity level new signals were generated with the same sparsity basis and measured by the same measurement matrix. For $k < 8$ the recovery of the signal was perfect, but as expected, for higher values of k the number of false reconstructed signals and the average error grew. The reason for this is that the OMP algorithm works well with small values of k , for higher values of k , even if the

uniqueness conditions are still satisfied, the OMP algorithm may not find the correct solution.

V. SPARSE BASIS

A different constraint that can be added to Problem 2 in order to reduce the number of solutions is the sparsity of the basis P . That is, we assume that the columns of the basis P are sparse under some known dictionary Φ , so that there exists some unknown sparse matrix Z such that $P = \Phi Z$. We assume the number of nonzero elements in each column of Z is known to equal k_p . We refer to Φ as a dictionary since it does not have to be square. Note that in order for P to be a basis Φ must have full row rank, and Z must have full column rank.

The constrained BCS in this case is then:

Problem 7. *Given the measurements B , the measurement matrix A and the dictionary Φ , which has full row rank, find the signal matrix X such that $B = AX$ where $X = \Phi ZS$ for some k -sparse matrix S and k_p -sparse and full column rank matrix Z .*

This problem is similar to that studied in [23] in the context of sparse DL. The difference is that [23] finds the matrices Z, S , while we are only interested in their product. The motivation behind Problem 7 is to overcome the disadvantage of the previously discussed Problem 4 in which the bases are fixed. When using a sparse basis we can choose a dictionary Φ with fast implementation, but enhance its adaptability to different signals by allowing any sparse enough combination of the columns of Φ . Note that we can solve the problem separately for several different dictionaries Φ , and choose the best solution. This way we can combine the sparse basis constraint and the constraint of a finite set of bases. Another possible combination between these two approaches is to define the basic dictionary as $\Phi = [P_1, P_2, \dots]$, where the finite set of bases is $\Psi = \{P_1, P_2, \dots\}$. This way we allow any sparse enough combination of columns from all the bases in Ψ .

A. Uniqueness Conditions

As we now show, here too under appropriate conditions the constrained problem has a unique solution even when there is only one signal $N = 1$. Therefore, instead of matrices X, S, B we deal with vectors x, s, b respectively. Since $\|s\|_0 \leq k$ and Z is k_p -sparse, the vector $c = Zs$ necessarily satisfies $\|c\|_0 \leq k_p k$. Therefore, Problem 7 as

$$\hat{c} = \arg \min_c \|c\|_0 \quad \text{s.t. } b = A\Phi c, \quad (9)$$

or equivalently:

$$\hat{c} = \arg \min_c \|b - A\Phi c\|_2^2 \quad \text{s.t. } \|c\|_0 \leq k_p k, \quad (10)$$

where the recovery is $x = \Phi \hat{c}$. The solutions to (9) and (10) are unique if $\sigma(A\Phi) \geq 2k_p k$. If there is more than one signal, $N > 1$, then one can solve (9) and (10) for each signal separately.

Note that in Problem 7 the matrix Z necessarily has full column rank, while this constraint is dropped in (9) and (10).

However, if the solution without this constraint is unique then obviously the solution with this constraint is also unique. Therefore, a sufficient condition for the uniqueness of Problem 7 is $\sigma(A\Phi) \geq 2k_p k$.

B. Algorithms For Sparse BCS

1) *Direct Method:* When there is only one signal, according to the uniqueness discussion, the solution to Problem 7 can be found by solving either (9) or (10) using a standard CS algorithm. When there are more signals the same process can be performed for each signal separately. Since we use a standard CS algorithm, for this method to succeed we require the product $k_p k$ to be small relative to n .

2) *Sparse K-SVD:* The sparse K-SVD algorithm [23] is a DL algorithm that seeks a sparse dictionary. That is, given the measurements B and a base dictionary D it finds k_p -sparse Z and k -sparse S , such that $B = DZS$. In our case we can run sparse K-SVD on B with $D = A\Phi$ in order to find Z and S , and then recover the signals by $X = \Phi ZS$. The sparse K-SVD algorithm is a variation of the K-SVD algorithm [24], which is a popular DL algorithm. Sparse K-SVD consists of two alternating steps. The first is sparse coding, in which Z is fixed and S is updated using a standard CS algorithm. The second step is dictionary update, in which the support of S is fixed and Z is updated together with the value of the nonzero elements in S . The difference between sparse K-SVD and K-SVD is only in the dictionary update step. Since the sparse K-SVD is a DL algorithm, it requires a large number of diverse signals. Moreover, the required diversity of the signals can prevent the algorithm from working, for instance in cases of block sparsity.

In general, BCS cannot be solved using DL methods. However, under the sparse basis constraint BCS is reduced to a problem that can be viewed as constrained DL, and therefore solved using sparse K-SVD. Nevertheless, Problem 7 is not exactly constrained DL, since in DL we seek the matrices S and Z themselves, whereas here we are interested only in their product $X = \Phi ZS$. Moreover, as in any DL algorithm, for sparse K-SVD to perform well it requires many diverse signals. However, for the uniqueness of Problem 7 or for the direct method of solution, there is no need for such a requirement. The sparse K-SVD algorithm is also much more complicated than the direct method.

Nonetheless, sparse K-SVD has one advantage over the direct method in solving Problem 7. The direct method uses a standard CS algorithm in order to find $C = ZS$ which is $k_p k$ -sparse. This algorithm provides the correct result only if the product $k_p k$ is small enough relative to n . On the other hand, the standard CS algorithms used in sparse K-SVD attempt to find separately S which is k -sparse and Z which is k_p -sparse, and therefore require k and k_p themselves to be small instead of the product $k_p k$. Thus, when there are few signals, or even just one, and when $k_p k$ is small relative to n , then Problem 7 should be solved using the direct method. If $k_p k$ is large but still satisfies $\sigma(A\Phi) \geq 2k_p k$, and if there are enough diverse signals, then sparse K-SVD should be used.

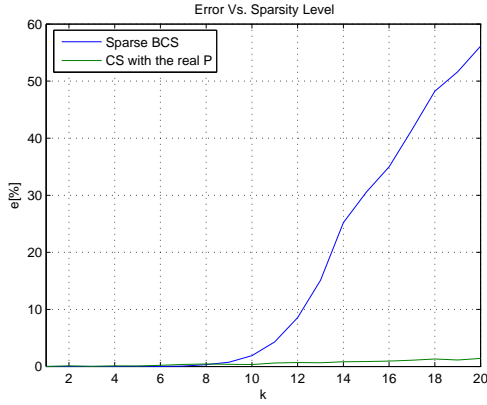


Fig. 1. Reconstruction error as a function of the sparsity level

C. Simulation Results

Simulation results for sparse K-SVD can be found in [23]. Here we present simulation results for the direct method. First of all we tested the influence of the sparsity level of the basis. We generated a random sparse matrix - Z , of size 256×256 with up to $k_p = 6$ nonzero elements in each column. The value of k - the number of nonzero elements in S , was gradually increased from 1 to 20. For each k we generated S as a random k -sparse matrix of size 256×100 , and created the signal matrix $X = \Phi Z S$, where Φ was the DCT basis. X was measured using a random Gaussian matrix A of size 128×256 , resulting in $B = AX$.

We solved Problem 7 given A and B using the direct method, where again the CS algorithm we used was OMP. For comparison we also performed OMP with the real basis P , which is unknown in practice. Fig 1 summarizes the results. For every value of k the error of each of the graphs is an average over the reconstruction errors of all the signals, calculated as in (8). Both the errors are similar for $k \leq 8$, but for larger k 's the error of the blind method is much higher.

Since A is an i.i.d Gaussian matrix and the DCT matrix is orthogonal with probability 1, $\sigma(A\Phi) = 129$. Therefore with probability 1 the uniqueness of the sparse BCS method is achieved as long as $k_p k \leq 64$, or $k \leq 10$. The error began to grow before this sparsity level because OMP is a suboptimal algorithm that is not guaranteed to find the solution even when it is unique, but works well on sparse enough signals. The reconstruction error of the OMP which used the real P grows much less for the same values of k . That is since in this case k itself, instead of $k_p k$, should be small relative to n .

Sparse K-SVD can improve the results for high value of k , assuming of course it is small enough for the solution to be unique. However, in this simulation the number of signals is even less than the length of the vectors, and sparse K-SVD does not work well with such a small number of signals. In the sparse K-SVD simulations which are presented in [23] the number of signals is at least 100 times the length of the signals.

We also investigated the influence of noise on the algorithm. The setting of this simulations were the same as in the previous

TABLE III
RECONSTRUCTION ERROR FOR DIFFERENT NOISE LEVELS

SNR	CS	sparse BCS
∞	$10^{-14}\%$	$10^{-14}\%$
30dB	1.2%	2.8%
25dB	1.5%	5.8%
20dB	3.3%	11.9%
15dB	7.1%	23.5%

simulation only this time we fixed $k = 3$ and added Gaussian noise to the measurements B . We looked at different noise levels, and for each level we ran the direct method for sparse BCS, and also for comparison we ran an OMP algorithm which used the real basis P . Table III summarizes the average errors of each of the methods. In the noiseless case there is a perfect recovery in both cases. As the SNR decreases both errors increases, but as can be expected, the one of the BCS grows faster. The reason for the big difference in the low SNR cases is again the fact that in the CS case the OMP algorithm is performed on sparser signals, relative to the sparse BCS case.

VI. STRUCTURAL CONSTRAINT

The last constraint we discuss is a structural constraint on the basis P . We require P to be block diagonal and orthogonal. The motivation for the block diagonal constraint comes from Problem 3, which looks for \tilde{P} such that $\tilde{D} = A\tilde{P}$. Assume for the moment that \tilde{P} is block diagonal, such that:

$$\tilde{P} = \begin{bmatrix} \tilde{P}_1 & & \\ & \ddots & \\ & & \tilde{P}_L \end{bmatrix},$$

and A is chosen to be a union of orthonormal bases, as in [21], [22], [30]–[32]. That is, $A = [A_1, \dots, A_L]$ where A_1, \dots, A_L are all orthonormal matrices. In this case

$$D = [D_1, \dots, D_L] = [A_1 P_1, \dots, A_L P_L],$$

and we can simply recover \tilde{P} by:

$$\tilde{P} = \begin{bmatrix} A_1^T D_1 & & \\ & \ddots & \\ & & A_L^T D_L \end{bmatrix}. \quad (11)$$

Therefore, the solution to Problem 3 under the constraint that \tilde{P} is block diagonal is very simple.

Under the richness and spark conditions the BCS problem, as defined in Problem 2, is equivalent to Problem 3, where the basis \tilde{P} in Problem 3 is a column signed permutation of the basis P in Problem 2. Since we are interested in the solution to Problem 2, the constraint should be on the basis P instead of \tilde{P} . However, if we constrain P to be block diagonal, then the solution to the equivalent Problem 3 is not as simple as in (11). In Problem 3 we look for $\tilde{P} = PQ$, for some unknown signed permutation matrix Q . Under the block diagonal constraint on P the matrix $\tilde{P} = PQ$ is not necessarily block diagonal, and therefore we cannot use (11) to recover it.

We can guarantee that \tilde{P} is block diagonal only if we can guarantee that Q is block diagonal. That is, Q permutes only the columns inside each block of P , and does not mix the blocks or change the outer order of them. As we prove below in the uniqueness discussion, this can be guaranteed if we require P to have more blocks than A . Specifically, we require P to have $2L$ blocks, which is twice the number of blocks in A . Such a basis P is called *2L-block diagonal*. In fact, the number of blocks in P can be ML for any integer $M \geq 2$. We use $M = 2$ for simplicity; the expansion to $M > 2$ is trivial.

We also constraint P to be orthogonal. The motivation for this is the spark condition. In order to be able to solve Problem 3 instead of Problem 2, we need to satisfy $\sigma(AP) \geq 2k$. By constraining P to be orthogonal we can use results similar to Proposition 1 in order to achieve this requirement with probability 1.

The constrained BCS problem is then:

Problem 8. *Given the measurements B and the measurement matrix $A \in \mathbb{R}^{n \times nL}$ find the signal matrix X such that $B = AX$ where $X = PS$ for some orthogonal $2L$ -block diagonal matrix P and k -sparse matrix S .*

In this new settings the size of the measurement matrix A is $n \times nL$, where n is the number of measurements and L is the number of $n \times n$ blocks in A , which equals the compression ratio. Moreover, The length of the signals is $m = nL$, and the size of the basis P is $nL \times nL$. Since P is $2L$ -block diagonal, the size of its blocks is $\frac{n}{2} \times \frac{n}{2}$. Therefore, n must be even.

This constrained problem can be useful for instance in multichannel systems, where the signals from each channel are sparse under separate bases. In such systems we can construct X by concatenating signals from several different channels, and compressively sampling them. For example, in microphone arrays [36] or antenna arrays [37], we can divide the samples from each microphone / antenna into time intervals in order to obtain the ensemble of sampled signals B . Each column of B is a concatenation of the signals from all the microphones / antennas over the same time interval.

A. Uniqueness Conditions

To ensure a unique solution to Problem 8, we need the DL solution given B to be unique. Therefore, we assume that the richness conditions on S and the spark condition on AP are satisfied. Then, Problem 8 is equivalent to the following problem:

Problem 9. *Given the matrices \tilde{D} and A , which have more columns than rows, find an orthogonal \tilde{P} such that $\tilde{D} = A\tilde{P}$, and $\tilde{P} = PQ$ for some signed permutation matrix Q and orthogonal $2L$ -block diagonal matrix P .*

In order to discuss conditions for uniqueness of the solution to Problem 9 we introduce the following definition.

Definition 10. *Denote $A = [A_1, \dots, A_L]$, such that $A_i \in \mathbb{R}^{n \times n}$ for any $1 \leq i \leq L$. A is called inter-block diagonal if there*

are two indices $i \neq j$ for which the product:

$$A_i^T A_j = \begin{bmatrix} R_1 & R_2 \\ R_3 & R_4 \end{bmatrix},$$

satisfies:

$$\begin{aligned} \text{rank}(R_1) &= \text{rank}(R_4) \\ \text{rank}(R_2) &= \text{rank}(R_3) = \frac{n}{2} - \text{rank}(R_1). \end{aligned}$$

In particular if the product $A_i^T A_j$ is 2-block diagonal then A is inter-block diagonal.

With this definition in hand we can now define the conditions for the uniqueness of Problem 9.

Theorem 11. *If $A \in \mathbb{R}^{n \times nL}$ is a union of L orthogonal bases, which is not inter-block diagonal, and $\sigma(AP) = n + 1$, then the solution to Problem 9 is unique.*

The proof of this theorem uses the next lemma.

Lemma 12. *Assume P and \hat{P} are both orthogonal $2L$ -block diagonal matrices, and A satisfies the conditions of Theorem 11. If $A\hat{P} = APQ$ for some signed permutation matrix Q , then $\hat{P} = PQ$.*

In general since A has a null space, if the matrices A, P, \hat{P} did not have their special structures, then the equality $A\hat{P} = APQ$ would not imply $\hat{P} = PQ$. However, according to Lemma 12 under the constraints on A, P, \hat{P} this is guaranteed. The full proof of Lemma 12 appears in Appendix A. Here we present only the proof sketch.

Proof sketch: It is easy to see that due to the orthogonality of the blocks of A , if Q is block diagonal then $A\hat{P} = APQ$ implies $\hat{P} = PQ$. Therefore, we need to prove that Q is necessarily block diagonal. Denote $D = AP$. In general the multiplication DQ can yield three types of changes in D . It can mix the blocks of D , permute the order of the blocks of D , and permute the columns inside each block. Q is block diagonal if and only if it permutes only the columns inside each block, but does not mix the blocks or change their outer order. First we prove that Q cannot mix the blocks of D . For this we use the condition on the spark of D , and the orthogonality of the blocks. Next we prove that Q cannot change the outer order of the blocks. This time we use the fact that both P and \hat{P} have $2L$ blocks and that A is not inter-block diagonal. Therefore, Q can only permute the columns inside each block, which implies it is block diagonal. ■

If P and \hat{P} have only L blocks instead of $2L$, then Q can change the outer order of the blocks of D , such that it does not have to be block diagonal. Therefore, if the constraint on P was that it has L blocks instead of $2L$, then Lemma 12 would be incorrect, such that the solution to the Problem 9, and therefore to Problem 8, would not be unique. On the other hand the extension of the proof of Lemma 12 to ML blocks where $M > 2$ is trivial.

Proof of Theorem 11: The proof we provide for Theorem 11 is constructive, although far from being a practical method to deploy in practice. Denote the desired solution of Problem 9

by $\tilde{P} = PQ$, and denote:

$$A = [A_1, \dots, A_L], \quad P = \begin{bmatrix} P^1 & & \\ & \ddots & \\ & & P^{2L} \end{bmatrix},$$

where A_i for $i = 1, \dots, L$ and P^j for $j = 1, \dots, 2L$ are all orthogonal matrices.

We first find a permutation matrix Q_D such that $\hat{D} = \tilde{D}Q_D = A\hat{P}$, where \hat{P} is an orthogonal $2L$ -block diagonal matrix. There is always at least one such permutation. For instance, we can choose Q_D to equal the absolute value of Q^T . In this case \hat{P} equals P up to the signs, and therefore it is necessarily orthogonal $2L$ -block diagonal.

Denote the blocks of \hat{P} by \hat{P}^j for $j = 1, \dots, 2L$, and note that

$$\hat{D} = [\hat{D}_1, \dots, \hat{D}_L] = \left[A_1 \begin{pmatrix} \hat{P}^1 & \\ & \hat{P}^2 \end{pmatrix}, \dots, A_L \begin{pmatrix} \hat{P}^{2L-1} & \\ & \hat{P}^{2L} \end{pmatrix} \right].$$

Since A_i are orthogonal for all $i = 1, \dots, L$, we can recover the blocks of \hat{P} by

$$\begin{bmatrix} \hat{P}^{2i-1} & \\ & \hat{P}^{2i} \end{bmatrix} = A_i^T \hat{D}_i,$$

such that

$$\hat{P} = \begin{bmatrix} A_1^T \hat{D}_1 & & \\ & \ddots & \\ & & A_L^T \hat{D}_L \end{bmatrix}.$$

Since both P and \hat{P} are orthogonal $2L$ -block diagonal, according to Lemma 12 the equality $\hat{D} = A\hat{P} = APQQ_D$ implies $\hat{P} = PQQ_D$. Therefore, we can recover \hat{P} by $\hat{P} = PQ = \hat{P}Q_D^T$. ■

The conclusion from Theorem 11 is that if the richness conditions on S are satisfied and A satisfies the conditions of Theorem 11, then the solution to Problem 8 is unique.

As proven in Appendix B one way to guarantee that A satisfies the conditions of Theorem 11 with probability 1 is to generate it randomly from an i.i.d Gaussian distribution and perform a Gram Schmidt process on each block in order to make it orthogonal. This claim is similar to Proposition 1 except that the statistics of A is a bit different due to the Gram Schmidt process.

B. The OBD-BCS Algorithm

Although the uniqueness proof is constructive it is far from being practical. In order to solve Problem 8 by following the uniqueness proof one needs to perform a DL algorithm on B , resulting in \tilde{D}, \tilde{S} . Then go over all the permutations $\hat{D} = \tilde{D}Q_D$, and look for Q_D such that the matrices $A_i^T \hat{D}_i$, for all $i = 1, \dots, L$, are 2-block diagonal. After finding such a permutation the recovery of X is

$$X = \begin{bmatrix} A_1^T \hat{D}_1 & & \\ & \ddots & \\ & & A_L^T \hat{D}_L \end{bmatrix} Q_D^T \tilde{S}.$$

The problem with this method is the search for the permutation Q_D . There are $m!$ different permutations of the columns of D , where $m = nL$ is the length of the signals, while only $[(\frac{m}{2L})!]^{2L}$ of them satisfy the requirement (see Appendix C). As m and L grow the relative fraction of the desirable permutations decreases. For instance, for signals of length $m = 16$ and a compression ratio of $L = 2$ only $1.58 \cdot 10^{-6}\%$ of the permutations satisfy the requirement. For the same signals but a higher compression ratio of $L = 4$ only $1.22 \cdot 10^{-9}\%$ satisfy the condition, and for longer signals of length $m = 64$ and $L = 2$ only $1.51 \cdot 10^{-34}\%$ satisfy the requirement.

Therefore, a systematic search is not practical, even for short signals. Moreover, in practice the output of the DL algorithm contains some error, so that even for the correct permutation the matrices $A_i^{-1} \hat{D}_i$ are not exactly 2-block diagonal, which renders the search even more complicated. Although there exist suboptimal methods for permutation problems such as [38], these techniques are still computationally extensive and are sensitive to noise.

Instead we present the orthogonal block diagonal BCS (OBD-BCS) algorithm for the solution of Problem 8, which is, in theory, equivalent to DL followed by the above post-processing. However, it is much more practical and simple. This algorithm is a variation of the DL algorithm in [21], [22], which learns a dictionary under the constraint that the dictionary is a union of orthogonal bases. Given B the algorithm in [21], [22] aims to solve

$$\min_{D, S} \|B - DS\|_F^2 \quad (12)$$

s.t. S is k -sparse and D is a union of orthogonal bases.

In the BCS case P is orthogonal $2L$ -block diagonal and A is a union of L orthogonal bases. Therefore, the equivalent dictionary is:

$$D = AP = \left[A_1 \begin{pmatrix} P^1 & \\ & P^2 \end{pmatrix}, \dots, A_L \begin{pmatrix} P^{2L-1} & \\ & P^{2L} \end{pmatrix} \right].$$

Since all A_i and P^i are orthogonal, here too D is a union of orthogonal bases. The measurement matrix A is known and we are looking for an orthogonal $2L$ -block diagonal matrix P and a sparse matrix S such that $B = APS$. This leads to the following variant of (12):

$$\min_{P, S} \|B - APS\|_F^2 \quad (13)$$

s.t. S is k -sparse and P is orthogonal $2L$ -block diagonal.

The algorithm in [21], [22] consists of two alternating steps. The first step is sparse coding, in which the dictionary D is fixed and the sparse matrix S is updated. The second step is dictionary update, in which S is fixed and D is updated. This algorithm finds the dictionary $D = AP$ and the sparse matrix S but not the basis P , and consequently, not the signal matrix $X = PS$.

In OBD-BCS we follow similar steps. The first step is again sparse coding, in which P is fixed and S is updated. The second step is basis update, in which S is fixed and P is updated. The difference between OBD-BCS and the algorithm

in [21], [22] is mainly in the second step, where we add the prior knowledge of the measurement matrix A and the block diagonal structure of P . In addition, we use a different CS algorithm in the sparse coding step.

We now discuss in detail the two steps of OBD-BCS.

1) *Sparse Coding*: In this step P is fixed so that the optimization in (13) becomes:

$$\min_S \|B - APS\|_F^2 \quad \text{s.t. } S \text{ is } k\text{-sparse.} \quad (14)$$

It is easy to see that (14) is separable in the columns of S . Therefore, for each column of B and S we need to solve

$$\min_s \|b - APs\|_2^2 \quad \text{s.t. } \|s\|_0 \leq k, \quad (15)$$

where s, b are the appropriate columns of S, B respectively. This is a standard CS problem, as in (3), with the additional property that the combined measurement matrix $D = AP$ is a union of orthogonal bases. This property is used by the block coordinate relaxation (BCR) algorithm [21], [22], [39]. The idea behind this algorithm is to divide the elements of s into blocks corresponding to the orthogonal blocks of D . In each iteration all the blocks of s are fixed except one, which is updated using soft thresholding. The DL algorithm proposed by [21], [22] is a variation of the BCR algorithm, which aims to improve its convergence rate. In OBD-BCS we can also use this variation. However, experiments showed that the results are about the same as the results with OMP. Therefore, we use OMP in order to update the sparse matrix S , when the basis P is fixed.

2) *Basis Update*: In this step the sparse matrix S is fixed and P is updated. Divide each of the $nL \times N$ matrices S and X into $2L$ submatrices of size $\frac{n}{2} \times N$ such that:

$$S = \begin{bmatrix} S^1 \\ \vdots \\ S^{2L} \end{bmatrix}, \quad X = \begin{bmatrix} X^1 \\ \vdots \\ X^{2L} \end{bmatrix}.$$

Divide each orthogonal block of A into two blocks: $A_i = [A^{2i-1}, A^{2i}]$ for $i = 1, \dots, L$, such that:

$$A = [A_1, \dots, A_L] = [A^1, A^2, \dots, A^{2L-1}, A^{2L}].$$

With this notation $X^i = P^i S^i$, and $B = \sum_{i=1}^{2L} A^i P^i S^i$. Therefore, (13) becomes:

$$\min_{P^1, \dots, P^{2L}} \|B - \sum_{j=1}^{2L} A^j P^j S^j\|_F^2 \quad (16)$$

s.t. P^1, \dots, P^{2L} are orthogonal.

To minimize (16), we iteratively fix all the blocks P^j for $j = 1, \dots, 2L$ except one, denoted by P^i , and solve

$$\min_{P^i} \|B^i - A^i P^i S^i\|_F^2 \quad \text{s.t. } P^i \text{ is orthogonal} \quad (17)$$

where $B^i = B - \sum_{j \neq i} A^j P^j S^j$. With slight abuse of notation, from now on we abandon the index i .

Since P is orthogonal and A is constructed of columns from an orthogonal matrix, $P^T A^T A P = I$, and $\|APS\|_F^2 = \|S\|_F^2$. Thus, (17) reduces to

$$\max_P \{\text{Tr} [B^T APS]\} \quad \text{s.t. } P \text{ is orthogonal.} \quad (18)$$

TABLE IV
THE OBD-BCS ALGORITHM

Inputs:
• $B \in \mathbb{R}^{n \times N}$ - measurements
• $A \in \mathbb{R}^{n \times nL}$ - measurement matrix (union of L orthogonal bases)
Outputs:
• $\hat{X} \in \mathbb{R}^{nL \times N}$ - reconstructed signal matrix
Algorithm:
• Initiate $\hat{P} = I$ (the identity).
• Repeat until a stopping criteria is reached:
◦ <i>Sparse coding</i> : find the sparsest \hat{S} such that $B = A\hat{P}\hat{S}$, for instance using OMP.
◦ <i>Basis update</i> : for all $i = 1, \dots, 2L$:
Calculate $B^i = B - \sum_{j \neq i} A^j \hat{P}^j \hat{S}^j$.
Use SVD: $\hat{S}^i (B^i)^T A^i = U \Sigma V^T$.
Update: $\hat{P}^i = V U^T$.
• Calculate: $\hat{X} = \hat{P} \hat{S}$.

Let the singular value decomposition (SVD) of the matrix $R = SB^T A$ be $R = U \Sigma V^T$, where U, V are orthogonal matrices and Σ is a diagonal matrix. Using this notation we can manipulate the trace in (18) as follows:

$$\text{Tr}[B^T APS] = \text{Tr}[SB^T AP] = \text{Tr}[\Sigma V^T P U].$$

The matrix $Z = V^T P U$ is orthogonal if and only if P is orthogonal. Therefore, (18) is equivalent to

$$\max_Z \{\text{Tr} [\Sigma Z]\} \quad \text{s.t. } Z \text{ is orthogonal.}$$

If the matrix $R = SB^T A$ has full rank then Σ is invertible. In this case the maximization is achieved only for $Z = I$, and therefore $P^i = V U^T$ is the unique minimum of (17). Even if R does not have full rank $P^i = V U^T$ achieves a minimum of (17).

Table IV summarize the OBD-BCS algorithm. Note that the initiation can be any $2L$ -block diagonal matrix, not necessarily the identity matrix as written in the table; however, the identity matrix is simple to implement. This algorithm is much simpler than following the uniqueness proof, which requires a combinatorial permutation search. Each iteration of the OBD-BCS algorithm uses a standard CS algorithm and $2L$ SVDs.

An important question that arises is whether the OBD-BCS algorithm converges. To answer this question we look at each step separately. If the sparse coding step is performed perfectly it solves (14) for the current P . That is, the objective of (13) is reduced or at least stays the same. In practice, for small enough k the CS algorithm converges to the solution of (14). However, in order to guarantee the objective of (13) is reduced or at least not increased in this step, we can always compare the new solution after this step with the one from the previous iteration and chose the best of them.

Note that this step is performed separately on each column of S . That is, we can choose to keep only some of the columns from the previous iteration, while the rest are updated. If at least part of the columns are updated then the next basis update step changes the basis P , so that in the following sparse coding step we can get a whole new matrix S . Therefore, the decision to keep the results from the previous iteration does not imply we keep getting the same results in all the

next iterations. Another possibility is to keep only the support of the previous solution and update the values of the nonzero elements using least-squares. In practice, in our simulations the algorithm converges even without any comparison to the previous iteration.

The basis update step is divided into $2L$ steps. In each, all the blocks of P are fixed except one, which is updated to minimize (17). Therefore, the objective of (17) is reduced or at least stays the same in each of the $2L$ steps constructing the basis update step. Therefore, the objective of (16), which is equivalent to (13) with fixed S , is reduced or not increased in the basis update step.

Thus, as in [21], [22], the algorithm we are based on, and as in other DL algorithms such as [20], [24], we cannot prove the OBD-BCS algorithm converges to the unique minimum of (13). However, we can guarantee that under specific conditions there is a unique minimum and that the objective function is reduced or at least stays the same in each step of the algorithm. Furthermore, as can be seen in the next section the OBD-BCS algorithm performs very well in simulations on synthetic data.

C. OBD-BCS Simulations

As in the first two constraints we evaluated the algorithm performance on synthetic data. The signal matrix X had 64 rows and was generated as a product of a random sparse matrix - S and a random orthogonal 4-block diagonal matrix - P . The value of the nonzero elements in S was generated randomly from a normal distribution, and the four orthogonal blocks of P were generated from a normal distribution followed by a Gram Schmidt process. The measurement matrix A was constructed of two random 32×32 orthogonal matrices, that were generated from a normal distribution followed by a Gram Schmidt process. The number of signals and the sparsity level were gradually changed in order to investigate their influence.

The stopping rule of the algorithm was based on a maximal number of iterations and the amount of change in the matrices S and P . If the change from the last iteration was too small, or if the maximal number of iterations was reached, then the algorithm stopped. In most cases the algorithm stopped due to small change between iterations after about 30 iterations.

First we examined the influence of two parameters, N - the number of signals needed for the reconstruction, and k - the sparsity level. Fig. 2 considers the influence of N where the sparsity level is set to $k = 4$. For each value of N from 150 to 2500 the error presented in the upper graph is an average over 20 simulations of the OBD-BCS algorithm. In each simulation the sparse vectors and the orthogonal matrix were generated independently, but the measurement matrix was not changed. The error of each signal was calculated according to (8).

For comparison, the lower graph in Fig. 2 is the average error of a standard CS algorithm that was performed on the same data, and used the real basis P , which is unknown in practice. The CS algorithm we used was again OMP. As expected, the results of the CS algorithm are independent of the number of signals, since it is performed separately and independently on each signal. The average error of this algorithm is 0.08%. The reason for this nonzero error, although

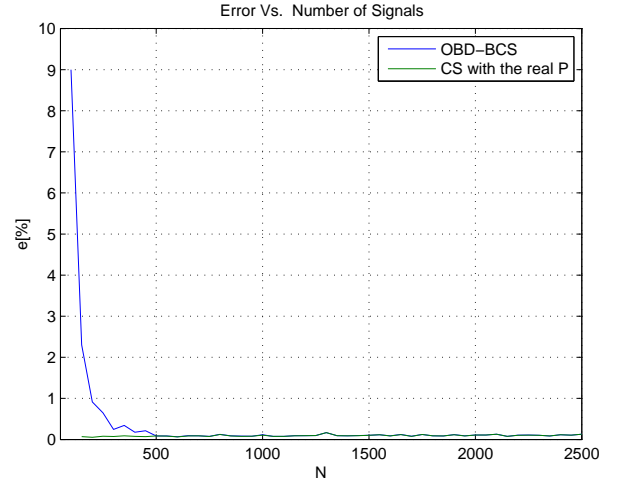


Fig. 2. Reconstruction error as a function of the number of signals, for sparsity level of $k = 4$.

P is known, is that for a small portion of the signals the OMP algorithm fails.

It is clear from Fig. 2 that for $N > 500$ the reconstruction results of the proposed algorithm are successful and similar to those obtained when P is known. Similarly to the conclusion in [17], the reconstruction is successful even for n much smaller than the number needed in order to satisfy the sufficient richness conditions, which is $\binom{m}{k}(k+1) \approx 3 \cdot 10^6$. As in most DL algorithms, the algorithm in [21], [22] was evaluated by counting the number of columns of the dictionary that are detected correctly. The conclusions of [21], [22] are that their algorithm can find about 80% of the columns when the number of signals is at least $20n = 640$, and can find all the columns when the number of signals is at least $50n = 1600$. Using the same measurement matrix dimensions as in [21], [22], the minimal number of signals the OBD-BCS algorithm requires is only 500.

In order to examine the influence of k we performed the same experiment as before but for different values of $k \leq 10$. The results are presented in Fig. 3. It can be seen that for all values of k the graph has the same basic shape: the error decreases with N until a critical N , after which the error is almost constant. As k grows this critical N increases and so does the value of the constant error. The graphs for $k = 1$, $k = 2$, $k = 3$ follow the same pattern; they are not in the figure since they are not visible on the same scale as the rest.

Next we investigated the influence of noise on the algorithm. In this simulation the noisy measurements B were calculated as $B = APS + W$, where the elements of W were white Gaussian noise. For each noise level 20 simulations were performed and the average error was calculated. In all simulations $k = 4$ and $N = 800$. Table V summarizes the results of the OBD-BCS algorithm and those of OMP algorithm which uses the real P . It is clear from the table that in the noiseless case the error of both algorithms is similar, therefore in this case the prior knowledge of the basis P can be avoided. As the SNR decreases both error increase, but the error of OBD-BCS

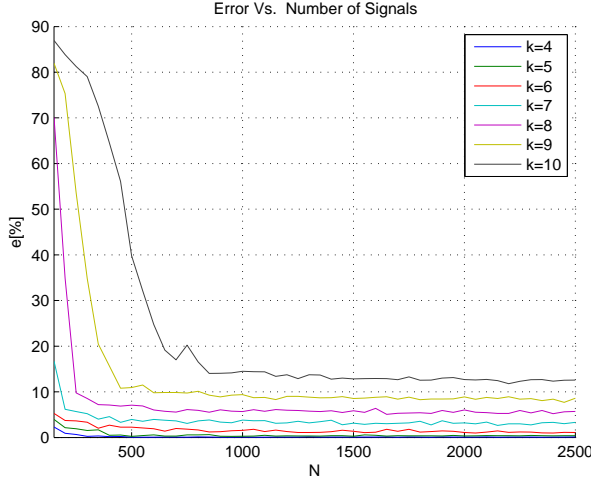


Fig. 3. Reconstruction error as a function of the number of signals for different values of k .

TABLE V
RECONSTRUCTION ERROR FOR DIFFERENT NOISE LEVELS

SNR	CS	OBD-BCS
∞	0.008%	0.008%
35dB	0.82%	0.88%
30dB	1.54%	1.64%
25dB	2.95%	3.23%
20dB	5.81%	6.10%
15dB	12.03%	12.58%
10dB	25.11%	26.04%

algorithm increases a bit faster than that of the CS algorithm. However, the difference is not very big.

VII. COMPARATIVE SIMULATION

The following simulation illustrates the difference between the three BCS methods presented in this work. In this simulation the length of the signals was $m = 128$, the sparsity level was $k = 6$, the number of signals was $N = 2000$, and the compression ratio was $L = 2$. The synthetic data was generated as in Section VI-C, but this time the instead of generating $P \in \mathbb{R}^{128 \times 128}$ randomly we used

$$P = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 & & & \\ & 1 & 1 & & \\ & & & \ddots & \\ & & & & 1 & -1 \\ & & & & 1 & 1 \end{bmatrix},$$

which can be viewed as an orthogonal 4-block diagonal matrix (each block is 16-block diagonal by itself).

We used five different methods for the reconstruction of these signals.

- 1) CS algorithm with the real basis P .
- 2) CS algorithm with an estimated basis P_{DL} .
- 3) The F-BCS method.
- 4) The direct method for sparse BCS.
- 5) The OBD-BCS algorithm.

TABLE VI
DL ALGORITHM FOR ORTHOGONAL DICTIONARY

Inputs
<ul style="list-style-type: none"> • X - training set • k - sparsity level
Outputs
<ul style="list-style-type: none"> • P - orthogonal dictionary • S - sparse matrix
Algorithm
<ul style="list-style-type: none"> • Initiate $P = I$. • Repeat until a stopping criteria is reached: <ul style="list-style-type: none"> ◦ Fix P and calculate $S = P^T X$. ◦ Keep only the k highest (absolute value) elements in each column of S. ◦ Fix S, and calculate the SVD: $SX^T = U\Sigma V^T$. ◦ Update $P = VU^T$.

In all the methods above we used OMP as the standard CS algorithm. The first method, came as a reference for the rest. It used the real basis P , whose knowledge we are trying to avoid. The second method is an intuitive way to reconstruct the signals. Since the basis P is unknown one can estimate it first and then perform a CS algorithm which uses the pre-estimated basis. We performed the estimation using a training set of 2000 signals and a DL algorithm. The estimated basis is denoted by P_{DL} . There are several different DL algorithms, eg. [20]–[22], [24], [40]. However, in this case we have important prior knowledge that the basis P is orthogonal 4-block diagonal. One way of using this knowledge is dividing the signals X into 4 blocks corresponding to the 4 blocks of P , and estimating each block of P from the relevant block of X using the algorithm in Table VI, which is designed for learning an orthogonal basis.

Due to this structure of P and the sparsity of S in each column of X there are up to 12 nonzero elements. Therefore, the identity matrix I was one of the bases in the finite set Ψ that we used. Specifically, we used the same set Ψ as in the simulations in Section IV. X had about twice as many nonzero elements in each column compared to the real sparse matrix S , such that X is $2k$ -sparse under I . Therefore, we ran the F-BCS method with sparsity level of $2k$ instead of k . Moreover, since P is sparse itself we used $\Phi = I$ as the base dictionary in the sparse BCS method. It is easy to see that $k_p = 2$.

Table VII reports the average error of all five methods, calculated as in (8). As can be seen, the results of F-BCS are much worse than all the others. This can be expected since in this case X is $2k$ -sparse, so that the OMP reconstruction is not as good. The error of the sparse BCS is also higher than the rest. The reason for this is that in order for the direct method of sparse BCS to work well the product $k_p k$ should be small relative to n . In this case this product is not small enough. Note that though higher from the rest the errors of the sparse BCS and F-BCS are quite small. We performed the same simulation with $k = 3$ and then the error of sparse BCS was reduced to the level of the rest, but the error of F-BCS was still high.

The results of both the OBD-BCS algorithm and the CS with the estimated basis, which both did not use the knowledge

TABLE VII
RECONSTRUCTION ERROR OF DIFFERENT RECONSTRUCTION ALGORITHMS

Algorithm	Error
CS with the real P	$10^{-5}\%$
CS with $\hat{P} = P_{DL}$	$10^{-5}\%$
F-BCS	0.522%
Sparse BCS	0.084%
OBD-BCS	$10^{-5}\%$

of the basis P , are similar to those of the algorithm which used this knowledge. Thus, the prior knowledge of P can be avoided. The advantage of OBD-BCS over the CS with the estimated basis is that it does not require any training set, and therefore can be used in applications where there is no access to any full signals but only to their measurements.

VIII. CONCLUSIONS

We presented the problem of BCS which aims to solve CS problems without the prior knowledge of the sparsity basis of the signals. Therefore, this work renders CS universal not only from the measurement process point of view, but also from the recovery point of view.

We presented three different constraints on the sparsity basis, that can be added to the BCS problem in order to guarantee the uniqueness of the solution to the BCS problem. Under each of these constraints we proved uniqueness conditions and proposed simple methods to retrieve the solution. All the proposed methods perform very well in simulations on synthetic data. In fact, when k is small enough and when enough signals are measured (only for the structural constraint case), the performance of our methods is similar to those of a standard CS which uses the real, though unknown in practice, sparsity basis. We also demonstrated through simulations the advantage of BCS over CS with an estimated sparsity basis. The advantage of BCS is that it does not require any training set, and therefore can be used in applications where there is no access to any full signals but only to their measurements.

An interesting direction for future research is to examine more ways to assure uniqueness, beside the three presented here, and weaken the constraint on the basis.

IX. ACKNOWLEDGMENTS

The authors would like to thank Prof. David Malah and Mr. Moshe Mishali for fruitful discussions and helpful advice.

APPENDIX A

The following proves Lemma 12. That is, if P and \hat{P} are both $2L$ -block diagonal matrices, A satisfies the conditions of Theorem 11, and Q is a permutation matrix, then $A\hat{P} = APQ$ implies $\hat{P} = PQ$.

We begin this proof by proving that under the lemma's conditions Q is necessarily block diagonal, after this is done the completion of the proof is straight forward. For any $D = [D_1, \dots, D_L] \in \mathbb{R}^{n \times nL}$ such that $D_1, \dots, D_L \in \mathbb{R}^{n \times n}$ the permutation DQ can yield three types of changes in D . It can mix the blocks of D , permute the order of the blocks of

D , and permute the columns inside each block. Q is L -block diagonal if and only if it permutes only the columns inside each block, but does not mix the blocks or change their outer order.

First we prove that Q cannot mix the blocks of D . We denote by Q_B the group of all block permutation matrices, which is the group of all the permutation matrices that keep all blocks together. That is, if $Q \in Q_B$ then when multiplying DQ only the order of the blocks D_1, \dots, D_L and the order of the columns inside the blocks change, but there is no mixture between the blocks. After we prove that $Q \in Q_B$ we prove that Q also cannot change the outer order of the blocks, and therefore must be block diagonal. In order to prove that necessarily $Q \in Q_B$, we use the next two lemmas.

Lemma A.1. *If $D = [D_1, \dots, D_L] \in \mathbb{R}^{n \times nL}$ is a union of L orthogonal bases, and $\sigma(D) = n + 1$, then any set of n orthogonal columns of D are necessarily all from the same block of D .*

Proof: Assume Γ is a set of n orthogonal columns from D . Denote $\Gamma = \Gamma_1 \cup \Gamma_2$, where Γ_1 is the set of columns taken from D_1 , and Γ_2 contains the rest of the columns in Γ . Without loss of generality assume the set Γ_1 is not empty. Since both D_1 and Γ are orthogonal bases of \mathbb{R}^n , the span of Γ_2 equals the span of the columns of D_1 which are not in Γ . Therefore, the set of columns $\Gamma_2 \cup d$, where d is any column from D_1 which is not in Γ , is either linearly dependent or empty. However, the set $\Gamma_2 \cup d$ contains at most n columns, so that since $\sigma(D) = n + 1$ this set cannot be linearly dependent. Therefore, Γ_2 is necessarily empty, such that all the columns of Γ are from the same block of D . ■

Lemma A.2. *Assume $D = [D_1, \dots, D_L] \in \mathbb{R}^{n \times nL}$ is a union of L orthonormal bases, with $\sigma(D) = n + 1$, and $\hat{D} = DQ$ for some permutation matrix Q . If \hat{D} is also a union of L orthonormal bases, then $Q \in Q_B$.*

Proof: If there was a permutation $Q \notin Q_B$ such that $\hat{D} = DQ$, it would imply that n columns of D , not all from the same block, form one of the orthogonal blocks of \hat{D} . However, according to Lemma A.1 any n orthogonal columns must be from the same block, and therefore $Q \in Q_B$. ■

We need to prove that the equality $A\hat{P} = APQ$ implies $\hat{P} = PQ$. Denote the orthogonal blocks of A by A_i for $i = 1, \dots, L$ and the orthogonal blocks of P and \hat{P} by P^j and \hat{P}^j respectively for $j = 1, \dots, 2L$. Also denote:

$$D = AP = \left[A_1 \begin{pmatrix} P^1 & & \\ & P^2 & \\ & & \ddots \end{pmatrix}, \dots, A_L \begin{pmatrix} & & P^{2L-1} \\ & & P^{2L} \end{pmatrix} \right]$$

$$\hat{D} = A\hat{P} = \left[A_1 \begin{pmatrix} \hat{P}^1 & & \\ & \hat{P}^2 & \\ & & \ddots \end{pmatrix}, \dots, A_L \begin{pmatrix} & & \hat{P}^{2L-1} \\ & & \hat{P}^{2L} \end{pmatrix} \right]$$

which are both unions of L orthogonal bases since A_i , P^j and \hat{P}^j are all orthogonal. Therefore, according to Lemma A.2 $Q \in Q_B$.

Next we prove that Q also cannot change the outer order of the blocks, and therefore must be L -block diagonal. Assume by contradictions that Q changes the outer order of the blocks of D . Without loss of generality we can assume this change

is a switch between the first two blocks of D . That is,

$$\begin{aligned}\hat{D}_1 &= D_2 Q_2 = A_2 \begin{bmatrix} P^3 & \\ & P^4 \end{bmatrix} Q_2 \\ \hat{D}_2 &= D_1 Q_1 = A_1 \begin{bmatrix} P^1 & \\ & P^2 \end{bmatrix} Q_1\end{aligned}$$

where Q_1, Q_2 are the corresponding sub-matrices of Q which permute the columns inside the blocks D_1, D_2 . In order to satisfy $\hat{D} = A\hat{P}$ we must have

$$\begin{aligned}\hat{D}_1 &= A_1 \begin{bmatrix} \hat{P}^1 & \\ & \hat{P}^2 \end{bmatrix} = A_2 \begin{bmatrix} P^3 & \\ & P^4 \end{bmatrix} Q_2 \\ \hat{D}_2 &= A_2 \begin{bmatrix} \hat{P}^3 & \\ & \hat{P}^4 \end{bmatrix} = A_1 \begin{bmatrix} P^1 & \\ & P^2 \end{bmatrix} Q_1.\end{aligned}$$

Since A_1 and A_2 are orthogonal the above implies

$$\begin{aligned}\begin{bmatrix} \hat{P}^1 & \\ & \hat{P}^2 \end{bmatrix} &= A_1^T A_2 \begin{bmatrix} P^3 & \\ & P^4 \end{bmatrix} Q_2 \\ \begin{bmatrix} \hat{P}^3 & \\ & \hat{P}^4 \end{bmatrix} &= A_2^T A_1 \begin{bmatrix} P^1 & \\ & P^2 \end{bmatrix} Q_1.\end{aligned}\tag{A-1}$$

If there is an orthogonal $2L$ -block diagonal matrix \hat{P} that satisfies (A-1), then in contradiction to Lemma 12 $\hat{P} \neq PQ$. However, (A-1) implies:

$$A_1^T A_2 = \begin{bmatrix} \hat{P}^1 & \\ & \hat{P}^2 \end{bmatrix} Q_2^T \begin{bmatrix} P^{3^T} & \\ & P^{4^T} \end{bmatrix} = \begin{bmatrix} R_1 & R_2 \\ R_3 & R_4 \end{bmatrix}.$$

Due to the structure of the permutation matrix Q_2 and due to the orthogonality of the blocks of P and \hat{P} , the ranks of R_1, R_2, R_3, R_4 must satisfy:

$$\begin{aligned}\text{rank}(R_1) &= \text{rank}(R_4) \\ \text{rank}(R_2) &= \text{rank}(R_3) = \frac{n}{2} - \text{rank}(R_1).\end{aligned}$$

Therefore, A is necessarily inter block diagonal. However, according to the conditions of Theorem 11 A is not inter block diagonal, so that the contradictions assumption is incorrect and Q cannot change the outer order of the blocks, such that Q must be L -block diagonal.

Denote the diagonal blocks of Q by Q_i for $i = 1, \dots, L$, such that:

$$\begin{aligned}\hat{D} &= \left[A_1 \begin{pmatrix} \hat{P}^1 & \\ & \hat{P}^2 \end{pmatrix}, \dots, A_L \begin{pmatrix} \hat{P}^{2L-1} & \\ & \hat{P}^{2L} \end{pmatrix} \right] = \\ &= \left[A_1 \begin{pmatrix} P^1 & \\ & P^2 \end{pmatrix} Q_1, \dots, A_L \begin{pmatrix} P^{2L-1} & \\ & P^{2L} \end{pmatrix} Q_L \right].\end{aligned}$$

Since all A_i are orthogonal the above implies that for all $i = 1, \dots, L$

$$\begin{bmatrix} \hat{P}^{2i-1} & \\ & \hat{P}^{2i} \end{bmatrix} = \begin{bmatrix} P^{2i-1} & \\ & P^{2i} \end{bmatrix} Q_i,$$

such that $\hat{P} = PQ$. \blacksquare

In fact the above proves not only that Q is L -block diagonal, it is also $2L$ -block diagonal. Note that the extension of this proof to the case where P and \hat{P} have ML blocks, for $M > 2$, is trivial. However, if P and \hat{P} had L blocks instead of $2L$, this proof would not work. That is since in this proof in order

to eliminate solutions of the form of (A-1) we use the 2-block diagonal structure of the matrices. If there were only L blocks, then beside the solution $\hat{P} = PQ$ there would have been another possibility, which is:

$$\hat{P} = \begin{bmatrix} A_1^T A_2 P_2 Q_2 & & & \\ & A_2^T A_1 P_1 Q_1 & & \\ & & P_3 Q_3 & \\ & & & \ddots \\ & & & & P_L Q_L \end{bmatrix},$$

where P_1, \dots, P_L are the L blocks of P and Q_1, \dots, Q_L the corresponding blocks of Q . Obviously in this case $\hat{P} \neq PQ$.

APPENDIX B

The following proves that if $A = [A_1, \dots, A_L] \in \mathbb{R}^{n \times nL}$ is a union of L orthogonal bases, where each block is generated randomly from an i.i.d Gaussian distribution followed by a Gram-Schmidt process, then with probability 1 $\sigma(A) = n + 1$ and A is not inter-block diagonal (Definition 10). Multiplication by an orthogonal P does not change the statistics, therefore if $\sigma(A) = n + 1$ with probability 1, then also $\sigma(AP) = n + 1$ with probability 1. Therefore, such an A satisfies the conditions of Theorem 11 with probability 1.

We begin the proof by noting that we can look at the generation of each block of A as follows. The first column a_1 is generated randomly from \mathbb{R}^n . The second column a_2 is generated randomly from the $n - 1$ dimensional space orthogonal to a_1 . the column a_3 is generated randomly from the $n - 2$ dimensional space orthogonal to the span of $\{a_1, a_2\}$, and similarly any a_i is generated randomly from the space orthogonal to the span of all previous columns, whose dimension is $n - i + 1$. We start by proving $\sigma(A) = n + 1$. This proof uses the next lemma.

Lemma B.3. Assume $G \in \mathbb{R}^{n \times n}$ is generated as an i.i.d Gaussian matrix followed by a Gram-Schmidt process, and U is a given space of dimension d . If $d < n$ then with probability 1 non of the columns of G are in U .

Proof: Denote the columns of G by g_i for $i = 1, \dots, n$. Since $d < n$ the space U has zero volume in \mathbb{R}^n . g_1 is generated randomly from \mathbb{R}^n , and therefore with probability 1 g_1 is not in U . For any other $1 < i \leq n$, g_i is generated randomly from G_i , which is the space orthogonal to the $i - 1$ previous columns in G . G_i dimension is $d_i = n - i + 1$. In this case we need to look at the probability to generate g_i in the intersection $U \cap G_i$. If $d < d_i$ then obviously this intersection has zero volume in G_i , so that g_i is not in U with probability 1. Furthermore, if $d \geq d_i$ then due to the randomness of the columns of G , G_i is not entirely contained in U with probability 1. Therefore, here too $U \cap G_i$ has zero volume in G_i , such that g_i is not in U with probability 1. \blacksquare

Assume Γ is a set of $\sigma(A)$ linearly dependent columns from A . Denote $\Gamma = \Gamma_1 \cup \Gamma_2$, where Γ_1 is the subset of Γ which contains only the columns taken from the block A_1 , and Γ_2 are the rest of the columns in Γ . Without loss of generality assume Γ_1 is not empty. Moreover, since A_1 is orthogonal

Γ_1 is also orthogonal, such that in order for Γ to be linearly dependent Γ_2 also cannot be empty.

Any $n+1$ columns from A are linearly dependent such that $\sigma(A) \leq n+1$. Therefore, $|\Gamma| \leq n+1$ so that $|\Gamma_1|, |\Gamma_2| \leq n$. If $|\Gamma_1| = n$ or $|\Gamma_2| = n$ then necessarily $\sigma(A) = |\Gamma| = n+1$. Assume by contradiction that $\sigma(A) = |\Gamma| \leq n$, such that $|\Gamma_1| < n$ and $|\Gamma_2| \leq n - |\Gamma_1|$. If $|\Gamma_1|$ contains only one column, denoted by γ_1 , then since Γ is linearly dependent γ_1 must be in the span of Γ_2 . However, the dimension of this span is at most $|\Gamma_2| \leq n-1$, such that according to Lemma B.3 the probability for this is zero. If Γ_1 contains only two columns, denoted by γ_1, γ_2 , then γ_2 must be in the span of $\Gamma_2 \cup \gamma_1$. However, the dimension of this space is at most $|\Gamma_2| + 1 \leq n-1$, such that according to Lemma B.3 the probability for this is again zero. We can keep increasing the cardinality of Γ_1 and as long as $|\Gamma| \leq n$ the probability for Γ to be linearly dependent will be zero. Therefore, the contradiction assumption is incorrect with probability 1, so that $\sigma(A) = |\Gamma| = n+1$ with probability 1.

Next we need to prove that A is not inter-block diagonal. Denote for any pair of indices $i \neq j$:

$$A_i^T A_j = \begin{bmatrix} R_1 & R_2 \\ R_3 & R_4 \end{bmatrix}. \quad (\text{B-2})$$

For A to be inter block diagonal there should be a pair $i \neq j$ for which:

$$\begin{aligned} \text{rank}(R_1) &= \text{rank}(R_4) \\ \text{rank}(R_2) &= \text{rank}(R_3) = \frac{n}{2} - \text{rank}(R_1). \end{aligned} \quad (\text{B-3})$$

However, due to the randomness of A_i, A_j the blocks R_1, R_2, R_3, R_4 all have full rank with probability 1. So that $\text{rank}(R_1) = \text{rank}(R_2) = \frac{n}{2}$ and $\text{rank}(R_2) \neq \frac{n}{2} - \text{rank}(R_1)$. Therefore, A is not inter block diagonal with probability 1.

APPENDIX C

Assume $A \in \mathbb{R}^{\frac{m}{2L} \times m}$ is a union of L random orthogonal bases and $P \in \mathbb{R}^{m \times m}$ is an orthogonal $2L$ -block diagonal matrix. Denote $\tilde{D} = APQ$ where Q is some unknown signed permutation matrix. We prove here that there are $[(\frac{m}{2L})!]^{2L}$ different permutation matrices Q_D such that $\tilde{D}Q_D = A\hat{P}$, where \hat{P} is an orthogonal $2L$ -block diagonal matrix. Without loss of generality we can assume $Q = I$, therefore we need to refer to $APQ_D = A\hat{P}$. According to Lemma 12 this implies $PQ_D = \hat{P}$. Since both P and \hat{P} are $2L$ -block diagonal Q_D must be too, and the size of its blocks is $\frac{m}{2L} \times \frac{m}{2L}$. Q_D is a permutation matrix, therefore each of its blocks is a permutation of the identity matrix of size $\frac{m}{2L}$. Thus, there are only $(\frac{m}{2L})!$ different possibilities for each block. There are $2L$ blocks such that the total number of possible Q_D 's is $[(\frac{m}{2L})!]^{2L}$.

REFERENCES

- [1] A. M. Brucksteiny, 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] D. L. Donoho, "Compressed sensing," *IEEE Trans. Info. Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.
- [3] E. Candes, J. Romberg and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Info. Theory*, vol. 52, pp. 1289–1306, April 2006.
- [4] M. Mishali and Y. C. Eldar, "Blind multi-band signal reconstruction: Compressed sensing for analog signals," *IEEE Trans. on Signal Processing*, vol. 57, no. 3, pp. 993–1009, March 2009.
- [5] Y. C. Eldar, "Compressed sensing of analog signals in shift-invariant spaces," *IEEE Trans. on Signal Processing*, vol. 57, no. 8, pp. 2986–2997.
- [6] K. Gedalyahu and Y. C. Eldar, "Time delay estimation from low rate samples: A union of subspaces approach," to appear in *IEEE Transactions on Signal Processing*.
- [7] M. Mishali and Y. C. Eldar, "From theory to practice: Sub-Nyquist sampling of sparse wideband analog signals," arXiv.org 0902.4291; to appear in *IEEE J. Selected Topics in Signal Processing*.
- [8] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 5302–5316, Nov. 2009.
- [9] D. L. Donoho and M. Elad, "Maximal sparsity representation via l_1 minimization," *Proc. Nat. Acad. Sci.*, vol. 100, pp. 2197–2202, March 2003.
- [10] E. Candes and T. Tao, "Decoding by linear programing," *IEEE Trans. Info. Theory*, vol. 51, no. 12, pp. 4203–4215, December 2005.
- [11] J. A. Tropp, "On the conditioning of random subdictionaries," *Applied and Computational Harmonic Analysis*, vol. 25, no. 1, 2008.
- [12] M. Mishali, Y. C. Eldar, O. Dounaevsky, and E. Shoshan, "Xampling: Analog to digital at sub-nyquist rates," CIT Report #751 Dec-09, EE Pub No. 1708, EE Dept., Technion - Israel Institute of Technology; [Online] arXiv 0912.2495, Dec. 2009.
- [13] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Info. Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.
- [14] I. Daubechies, M. Defrise and C. De-Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, August 2004.
- [15] S.S. Chen, D. L. Donoho and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, 2001.
- [16] R. Rubinstein, A. M. Bruckstein and M. Elad, "Dictionaries for sparse representation modeling," *Submitted to IEEE Proceedings Special Issue on Applications of Compressive Sensing and Sparse Representation*.
- [17] M. Aharon, M. Elad and M. Bruckstein, "On the uniqueness of overcomplete dictionaries, and practical way to retrieve them," *Linear Algebra and Its Applications*, vol. 416, no. 1, pp. 48–67, 2006.
- [18] M. S. Lewicki and T. J. Senowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.
- [19] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T. W. Lee and T. J. Senowski, "Dictionary learning algorithms for sparse representation," *Neural Computation*, vol. 15, no. 2, pp. 349–396, 2003.
- [20] K. Engan, S. O. Aase and J. H. Husoy, "Frame based signal compression using method of optimal directions (MOD)," *IEEE Intern. Symp. Circ. Syst.*, vol. 4, pp. 1–4, July 1999.
- [21] S. Lesage, R. Gribonval, F. Bimbot and L. Benaroya, "Learning unions of orthonormal bases with thresholded singular value decomposition," *ICASSP*, vol. 5, pp. 293–296, 2005.
- [22] S. Lesage, R. Gribonval, F. Bimbot and L. Benaroya, "Learning unions of orthonormal bases with thresholded singular value decomposition," *Tech. Rep.*, IRISA, 2004.
- [23] R. Rubinstein, M. Zibulevsky and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," to appear in *IEEE Trans. on Signal Processing*.
- [24] M. Aharon, M. Elad, A. Bruckstein and Y. Kats, "K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Trans. on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.
- [25] S. Mallat, *A wavelet tour of signal processing*, Academic Press, 1999.
- [26] N. Ahmed, T. Natarajan and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. 23, no. 1, pp. 90–93, January 1974.
- [27] M. Mishali and Y. C. Eldar, "Sparse source separation from orthogonal mixtures," *ICASSP*, pp. 3145–3148, April 2009.
- [28] R. Gribonval, and K. Schnass, "Dictionary identification via l_1 minimization," *submitted to IEEE Trans. Inf. Theory*, 2009.
- [29] R. Gribonval and K. Schnass, "Dictionary identification from few training samples," *Proc. 16th EUSIPCO08*, August 2008.
- [30] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Info. Theory*, vol. 47, no. 7, pp. 2845–2862, November 2001.
- [31] M. Elad and A. M. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," *IEEE Trans. Info. Theory*, vol. 48, no. 9, pp. 2558–2567, September 2002.

- [32] R. Gribonval and M. Nielsen, "Sparse decompositions in unions of bases," *IEEE Trans. Info. Theory*, vol. 49, no. 12, pp. 3320-3325, December 2003.
- [33] R. Baraniuk, M. Davenport, R. DeVore and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation Springer*, vol. 28, no. 3, pp. 253-263, December 2008.
- [34] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 5302-5316, November 2009.
- [35] Y. C. Eldar, P. Kuppinger and H. Bolcskei, "Compressed sensing of block-sparse signals: Uncertainty relations and efficient recovery," *submitted to IEEE Transactions on Signal Processing*, 2009.
- [36] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques And Applications*, Springer, 2001.
- [37] R. B. Ertel, P. Cardieri, K.W. Sowerby, T. S. Rappaport and J. H. Reed, "Overview of spatial channel models for antenna array communication systems," *IEEE Personal Communications*, vol. 5, no. 1, pp. 10-22, February 1998.
- [38] H. F. Wang and K. Y. Wu, "Hybrid genetic algorithm for optimization problems with permutation property," *Computers and Operations Research Elsevier*, vol. 31, pp. 2453-2471, 2004.
- [39] S. Sardy, A. G. Brouce and P. Tseng, "Block coordinate relaxation methods for nonparametric wavelet denoising," *Computational and Graphical Statistics*, vol. 9, no. 2, pp. 361-379, June 2000.
- [40] M. Yaghoobi, T. Blumensath and M. E. Davies, "Dictionary learning for sparse approximations with majorization method," *IEEE Trans. on Signal Processing*, vol. 57, no. 6, pp. 2178 - 2191, June 2009.