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 guarantee a unique solution. For each constraint, we prove conditions for uniqueness, and suggest a simple method to retrieve the solution. We demonstrate through simulations that our methods can achieve results similar to those of standard compressed sensing, which rely on prior knowledge of the sparsity basis, as long as the signals are sparse enough. 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.**

***Index Terms*—Blind reconstruction, compressed sensing, dictionary learning, sparse representation.**

I. Introduction

SPARSE signal representations have gained popularity in recent years in many theoretical and applied areas [1]–[9]. 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 [7], [10]. Other models for analog sparsity are discussed in detail in [8], [9], [11].

Compressed sensing (CS) [1], [2] 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 [3], [12], [13], 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 [7], [9], [14].

Manuscript received February 16, 2010; May 26, 2011; accepted June 17, 2011. Date of current version October 07, 2011. This work was supported in part by the Israel Science Foundation under Grant 1081/07 and by the European Commission in the framework of the FP7 Network of Excellence in Wireless COMmunications NEWCOM++ under Contract 216715.

Communicated by J. Romberg, Associate Editor for Signal Processing.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2011.2165821

In principle, recovery from compressed measurements is NP-hard. Nonetheless, many suboptimal methods have been proposed to approximate its solution [1], [2], [4]–[6]. These algorithms recover the true value of x when x is sufficiently sparse and the columns of A are incoherent [3], [5], [12], [13]. However, all known recovery approaches use the prior knowledge of the sparsity basis P.

Dictionary learning (DL) [15]–[18] 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 [15], the authors study conditions under which the DL problem yields a unique solution for a 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 sampling or 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.

Since in BCS the sparsity basis is unknown, the uncertainty about the signal x is larger 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.

The goal of this work is to define the new BCS problem, and investigate basic theoretical conditions for uniqueness of its solution. We first prove that BCS is ill-posed in general. Next we analyze its uniqueness under several additional constraints. Due to the relation between BCS and the problems of CS and DL, we base the BCS uniqueness analysis on the prior uniqueness conditions of CS and DL, as presented in [3] and [15]. Finally, under each of the investigated constraints we propose a method to retrieve the solution. To prove the concept of BCS we begin by discussing two simple constraints on the sparsity basis. Under each of these constraints we show that BCS can be viewed as a CS or DL problem, which serves as the basis of our uniqueness analysis and proposed solution method. We then turn to a third constraint which is our main focus, and is inspired by multichannel systems. The resulting BCS problem can no longer be viewed as CS or DL, so that the analysis and algorithms are more involved. We note that many other constraints are possible in order to allow blind recovery of sparse signals. The purpose

of our work is to introduce BCS and prove its concept. Our hope is to inspire future work in this area, leading to additional setups under which BCS may be possible.

The first constraint on the basis that we consider relies on the fact that over the years there have been several bases which have been considered "good" in the sense that they are known to sparsely represent many natural signals. These include, for example, various wavelet representations [19] and the discrete-cosine transform (DCT) [20]. 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 [21]. We compare these two approaches for BCS with a sparse basis. For both classes of constrains we show that a Gaussian random measurement matrix satisfies the uniqueness conditions we develop with probability one.

The third and main constraint we treat 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 [22]–[24], we impose in addition that P is orthonormal. We then choose to measure a set of signals X, all sparse in the same basis, by a measurement matrix A consisting of a union of orthonormal bases. This choice has been used in previous CS and DL works as well [25]–[29]. 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 DL framework and then relying on results obtained in that context. 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 permutation ambiguity, which is inherent in DL. Instead, we propose a simple and direct algorithm for recovery, which we refer to as the orthonormal 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 (SVDs).

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, our simulations show that the number of signals needed is reasonable and much smaller than that used for DL [21], [28]–[30].

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. Sections IV, V, and VI consider the three constrained BCS problems respectively. Simulation results and 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:

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

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

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

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

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

An important question is under what conditions (1)–(3) have a unique solution. In [3] 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 $\sigma(AP) > 2k$, then the solution to (2), or equivalently (3), is unique. Unfortunately, calculating the spark of a matrix is a combinatorial problem. However, it can be bounded by the *mutual coherence* [3], which can be calculated easily. Denoting the ith column of a matrix D by d_i, the mutual coherence of D is given by

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

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

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

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 orthonormal 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 orthonormal basis P.

Proof: Due to the properties of Gaussian random variables, the product AP is also an i.i.d. Gaussian random matrix for any unitary P. Since any n (or less than n) i.i.d. Gaussian vectors in $\mathbb{R}^n$ are linearly independent with probability 1, we have that $\sigma(AP) > n$ 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 $k < n/2$, then the uniqueness of the solution to (2) or (3) is guaranteed with probability 1 for any fixed orthonormal basis P (see also [31]).

Problems (2) and (3) are NP-hard in general. Many suboptimal methods have been proposed to approximate their solutions, such as [1], [2], [4]–[6]. These algorithms can be divided into two main approaches: greedy techniques and convex relaxation methods. Greedy algorithms approximate the solution by selecting indices of the nonzero elements in $\hat{s}$ sequentially. One of the most common methods of this type is orthogonal matching pursuit (OMP) [5]. Convex relaxation approaches change the objective in (2) to a convex problem. The most common of these techniques is basis pursuit (BP) [4], which considers the problem

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

Under suitable conditions on the product AP and the sparsity level of the signals, both the greedy algorithms and the convex relaxation methods recover s. For instance, both OMP and BP recover the true value of s when the number of nonzero elements in s is less than $\frac{1}{2}(1 + \frac{1}{\mu(AP)})$ [3], [5], [12], [13].

B. BCS Problem Formulation

Even when the universality property is achieved in CS, all existing algorithms require knowledge of the sparsity basis P for the reconstruction process. The idea of BCS is to entirely avoid the need of this prior knowledge. That is, perform both sampling and 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 [15], [30] 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 1: Given measurements B and a 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.

III. Uniqueness

We now discuss BCS uniqueness, namely the uniqueness of the signal matrix X which solves Problem 1. Unfortunately, although Problem 1 seems quite natural, its solution is not unique for any choice of measurement matrix A, for any number of signals, and for any sparsity level. We prove this result by first reducing the problem to an equivalent one using the field of DL, and then 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. We then use these results in Section III-B to prove that the BCS problem does not have a unique solution. In Sections IV, V, and VI, we suggest several constraints on the basis P that ensure uniqueness.

A. Dictionary Learning (DL)

DL [15]–[18] 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 interesting question is under what requirements DL has a unique solution. 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 (including sign changes) and permutation of the columns of D and rows of S respectively do not change the product $B = DS$. Therefore, in the context of DL the term uniqueness refers to uniqueness up to scaling and 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 permutation.

Conditions for DL uniqueness when the dictionary D is orthogonal or just square are provided in [23] and [24]. However, in BCS $D = AP$ is in general rectangular. In [15], the authors

prove sufficient conditions on D and S for the uniqueness of a general DL problem. We refer to the condition on D as the *spark condition* and to those on S as the *richness conditions*. The main idea behind these requirements is that D should satisfy CS uniqueness, and the columns of S should be diverse with respect to both the locations and the values of the nonzero elements. More specifically, the requirements for DL uniqueness are:

- the spark condition: $\sigma(D) > 2k$.
- 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, which is the number of columns in S, must be at least $\binom{m}{k}(k+1)$. Nevertheless, it was shown in [15] 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 [15] 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 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 ignore the ambiguity in the scaling and permutation, and assume that applying DL on B provides S and $D = AP$. Therefore, under the DL uniqueness conditions on S and D, the BCS problem is equivalent to the following problem.

Problem 2: Given $D \in \mathbb{R}^{n\times m}$ and $A \in \mathbb{R}^{n\times m}$, where $n < m$, find a basis P such that $D = AP$.

Since $n < m$ the matrix A in Problem 2 has a null space, and therefore without the constraint that P must be a basis, there is obviously no unique P such that $D = AP$. Moreover, as implied from the following proposition, even with the constraint that P is a basis there is still no unique solution to the DL equivalent problem.

Proposition 2: If there is a solution P to Problem 2, then it is necessarily not unique.

Proof: Assume P_1 solves Problem 2, so that it has full rank and satisfies $D = AP_1$. Decompose P_1 as $P_1 = P_{N^\perp} + P_N$ 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 $P_1 = P_{N^\perp}$ 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 $P_2 = P_{N^\perp} - P_N$ which is different from P_1, but it is easy to see that $D = AP_2$. Moreover, since the columns of P_N are perpendicular to the columns of $P_{N^\perp}$

$$P_1^T P_1 = P_2^T P_2 = P_{N^\perp}^T P_{N^\perp} + P_N^T P_N. \tag{5}$$

A square matrix P has full rank if and only if $P^T P$ has full rank. Therefore, since P_1 has full rank, (5) implies that P_2 also has full rank, so that both P_1 and P_2 are solutions to Problem 2. ■

According to Proposition 2 if there is a solution to Problem 2, then there is at least one more solution. However, it is easy to see from the proof that in fact there are many more solutions. For instance, 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 1. We just proved that when the DL solution given B is unique, Problem 1 is equivalent to Problem 2 which has no unique solution. Obviously if the DL solution given B is not unique, then BCS will not be unique. Therefore, Problem 1 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 2 unique. Therefore, in addition to the richness conditions on S and the spark condition on AP these constraints guarantee the uniqueness of the solution to Problem 1. 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 orthonormal and has a block diagonal structure.

The motivation for these constraints comes from the uniqueness of Problem 2. Nonetheless, we provide conditions under which the solution to Problem 1 with constraints 1 or 2 is unique even without DL uniqueness. In fact, under these conditions the solution to Problem 1 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 corresponding constrained BCS solution, and suggest a method to retrieve the solution. Table I summarizes these three approaches.

IV. Finite Set of Bases

Over the years a variety of bases were proven to lead to sparse representations of many natural signals, such as wavelet [19] and DCT [20]. 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. Motivated by this intuition we discuss in this section the BCS problem with an additional constraint which limits the possible sparsity basis to a finite and known set of bases. The constrained BCS is then:

Problem 3: Given measurements B, a 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.

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.	• $\sigma(AP) > 2k$ for any $P \in \Psi$. • A is k-rank preserving of Ψ (Definition 3).	• F-BCS - Solving (7) or (8) 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.	• $\sigma(A\Phi) > 2k_P k$.	• 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 [21] to retrieve S, Z, where the recovery is $X = \Phi ZS$.
Structure - Section VI P is orthonormal $2L$-block diagonal.	• The richness conditions on S. • A is a union of L orthonormal bases. • $\sigma(AP) = n + 1$. • A is not inter-block diagonal (Definition 6).	• OBD-BCS - Updating S and P alternately according to the algorithm in Table II, where the recovery is $X = PS$.

A. Uniqueness Conditions

We now show that under proper conditions the solution to Problem 3 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 vectors x, s, b, respectively.

Assume x is a solution to Problem 3. 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 $\sigma(AP) > 2k$; otherwise even if $\bar{P} = P$ there is no unique solution [3]. Since the real sparsity basis P is unknown, we require that $\sigma(AP) > 2k$ 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 submatrix of P containing only the columns with indices in T. If $\bar{x}$ is also a solution to Problem 3 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

$$\operatorname{rank}(A[P_T, \bar{P}_J]) = \operatorname{rank}[P_T, \bar{P}_J] \tag{6}$$

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 (6), $A\bar{P}_J \bar{s}_J = AP_T s_T$ 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 3 in addition to the requirement that $\sigma(AP) > 2k$ 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 (6).

Definition 3: 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 (6).

The conditions for the uniqueness of the solution to Problem 3 can therefore be summarized by the following theorem:

Theorem 4: If $\sigma(AP) > 2k$ for any $P \in \Psi$, and A is k-rank preserving of the set Ψ, then the solution to Problem 3 is unique.

In order to satisfy the condition on the spark with probability 1, according to Section II.A we may require all $P \in \Psi$ to be orthonormal and generate A from an i.i.d. Gaussian distribution. However, since the number of bases is finite, we can instead verify this 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.

In order to satisfy the condition on A with probability 1, we can generate it at random, as incorporated in the following proposition.

Proposition 5: An i.i.d. Gaussian matrix A of size $n \times m$ is k-rank preserving of any fixed finite set of bases and any $k \leq n/2$, with probability 1.

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 $r = \operatorname{rank}[P_T, \bar{P}_J]$. We now need to prove that $\operatorname{rank}(A[P_T, \bar{P}_J]) = r$.

Perform a Gram Schmidt process on the columns of $[P_T, \bar{P}_J]$ and denote the resulting matrix by G. It follows that G is an $m \times r$ matrix with orthonormal columns, with $\operatorname{rank}(G) = r$ and $\operatorname{rank}(AG) = \operatorname{rank}(A[P_T, \bar{P}_J])$. Next we complete G to an orthonormal matrix G_u by adding columns. According to Proposition 1 since A is an i.i.d. Gaussian matrix and G_u is orthonormal $\sigma(AG_u) = n+1$ 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 $\operatorname{rank}(AG) = r$, completing the proof. ∎

Until now we proved conditions for the uniqueness of Problem 3 assuming there is only one signal $N = 1$. The same conditions hold for $N > 1$ as 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 example, 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 (6). 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 (6), 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 when there are enough diverse signals such that the richness conditions on S are satisfied. In this case it is sufficient 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 permutation of the columns. This way we guarantee

that the problem equivalent to Problem 3 under the richness and spark conditions has a unique solution, and therefore the solution to Problem 3 is also unique.

Problem 3 can also be viewed as CS with a block sparsity constraint [11], [32]. That is, if $\Psi = \{P_1, P_2, \cdots\}$ then the desired signal matrix can be written as

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

where only one of the submatrices S_i is not all zeros. In contrast to the usual block sparsity, as in [33], in our case the nonzero submatrix S_i is itself sparse. However, the uniqueness conditions which are implied from this block sparsity CS approach are too strong compared to our BCS results. For example, they require all $P_j \in \Psi$ to be incoherent, whereas 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 3. We refer to this method as F-BCS which stands for finite BCS. When $N = 1$, F-BCS solves a CS problem for each $P \in \Psi$

$$\hat{s} = \arg\min_s ||s||_0 \quad \text{s.t. } b = APs \tag{7}$$

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 find 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 $P \in \Psi$

$$\hat{s} = \arg\min_s ||b - APs||_2^2 \quad \text{s.t.} \quad ||s||_0 < k \tag{8}$$

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 (7) or (8) for each of the signals. The solution for each signal indicates a basis P. We select the sparsity basis to be the one which is chosen by the largest number of signals.

To solve problems (7) and (8), we can use any of the standard CS algorithms. Since these techniques are suboptimal in general, there is no guarantee that they provide the correct solution x, even for the true 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.

Simulation results are presented in Section VII-A, where we test the method on signals with additive noise and different number of nonzero elements. As we show the proposed algorithm preforms very well as long as the number of nonzero elements and the noise level are small enough.

V. Sparse Basis

A different constraint that can be added to Problem 1 in order to reduce the number of solutions is sparsity of the basis P. That is, the columns of P are assumed to be sparse under some known dictionary Φ, so that there exists an 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 problem then become:

Problem 4: Given measurements B, a 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 [21] in the context of sparse DL. The difference is that [21] finds the matrices Z, S, while we are only interested in their product. The motivation behind Problem 4 is to overcome the disadvantage of the previously discussed Problem 3 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 $\Phi = [P_1, P_2, \ldots]$, where the finite set of bases is $\Psi = \{P_1, P_2, \ldots\}$. 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 $||c||_0 \leq k_p k$. We can then write Problem 4 as

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

or equivalently:

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

The recovery is $x = \Phi\hat{c}$. From CS we know that the solutions to (9) and (10) are unique if $\sigma(A\Phi) > 2k_p k$. If there is more then one signal, $N > 1$, then (9) and (10) can be solved for each signal separately.

Note that in Problem 4 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 4 is $\sigma(A\Phi) > 2k_p k$.

B. Algorithms for Sparse BCS

1) Direct Method: When there is only one signal the solution to Problem 4 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: Sparse K-SVD [21] 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 $X = \Phi ZS$. 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 method. 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.

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, the sparse BCS problem is not exactly a constrained DL problem, 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 a large number of diverse signals. However, such diversity is not needed for the uniqueness of the solution to the sparse BCS problem or for the direct method of solution. In some cases this required diversity of the signals can prevent sparse K-SVD from working, for instance when the signals are jointly sparse (have similar sparse patterns). Sparse K-SVD is also much more complicated than the direct method.

Simulation results for sparse K-SVD can be found in [21]; simulation results of the direct method are presented in Section VII-B. As we will see, this algorithm preforms very well when the number of nonzero elements and the noise level are small enough. This method can also be used in cases sparse K-SVD cannot be used, such as when there is a small number of signals or when the signals share similar sparsity patterns.

VI. Structural Constraint

In this section, we discuss a structural constraint on the basis P, which is motivated by multichannel systems, where the signals from each channel are sparse under separate bases. In such systems we can construct the set of signals X by concatenating signals from several different channels. In this setting, the sparsity basis is block diagonal, where the number of blocks equals the number of channels, and each block is the sparsity basis of the corresponding channel.

For example, in microphone arrays [34] or antenna arrays [35], we can divide the samples from each microphone/antenna into time intervals in order to obtain the ensemble of signals X. Each column of X is a concatenation of the signals from all the microphones/antennas over the same time interval. Alternatively we can look at large images that can be divided into patches such that each patch is sparse under a separate basis. In this case every column of X is a concatenation of the patches in the same locations in different images. This partition into patches is used, for example, in JPEG compression [36].

The advantage of the block structure of P is that with the right choice of A we can look at the problem as a set of separate simple problems. For instance, Problem 2 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 [25]–[29]. 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 recover $\tilde{P}$ simply by

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

Therefore, the solution to Problem 2 under the constraint that $\tilde{P}$ is block diagonal is straightforward.

Eventually we are interested in the BCS problem, as defined in Problem 1, therefore the constraint should be added to this problem and not to Problem 2. The BCS problem with an additional structural constraint on P is not equivalent to Problem 2 with the same constraint. Therefore, the solution to this structural BCS problem is not as simple as in (11). The reason for this asymmetry between the problems is that in Section III we reduce Problem 1 into Problem 2 by applying DL on the measurements B. In this reduction we ignore the ambiguity in the scaling and permutation of the DL output. However, here we can no longer ignore the ambiguity in the permutation, since a permutation of P can destroy its block diagonal structure. The ambiguity in the scaling can still be ignored following the same reasoning as in Section III.

We conclude that we cannot refer to Problem 2 with a block diagonal constraint on P and solve it using (11). Instead the additional constraint in Problem 2 should be that the basis is an unknown column permutation of a block diagonal matrix. We can solve this new problem using (11) only if we can guarantee that this unknown column permutation keeps the block diagonal structure. That is, it permutes only the columns inside each block of P, and does not mix the blocks or change their outer order. As we prove in the uniqueness discussion below, this can be guaranteed if the number of blocks in P is an integer multiple of the number of blocks in A, which we denote by L. That is, we assume that P has ML blocks for $M \geq 2$. For simplicity we will use $M = 2$ in our derivations below; the extension to $M > 2$ is trivial. Following many works in CS and DL, we further assume that P is orthonormal [22]–[24].

Summarizing our discussion, the constrained BCS problem is:

Problem 5: Given measurements B and a measurement matrix $A \in \mathbb{R}^{n \times nL}$ find the signal matrix X such that $B = AX$

where $X = PS$ for some orthonormal $2L$-block diagonal matrix P and k-sparse matrix S.

In this new setting 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. The signal length is $m = nL$, and the size of the basis P is $nL \times nL$. The $2L$-block diagonal structure of P implies that the size of its blocks is $\frac{n}{2} \times \frac{n}{2}$, so that n must be even.

A. Uniqueness Conditions

The uniqueness discussion in this section uses the definition of a permutation matrix, which is a column (or row) permutation of the identity matrix. In other words, it has only one nonzero element, which equals 1, in each column and row. If Q is a permutation matrix, then for any matrix D the product DQ is a column permutation of D, and QD is a row permutation of D. Obviously, any permutation matrix is orthonormal, and a product of any two permutation matrices is also a permutation matrix.

To prove uniqueness of the solution to Problem 5, we note that we can solve this problem by first applying DL on the measurements B, which provides $\tilde{D} = APQ$ and $\tilde{S} = Q^T S$ for some unknown permutation matrix Q. We then extract $\tilde{P} = PQ$ out of $\tilde{D}$, and recover $\hat{X} = \tilde{P}\tilde{S} = PQQ^T S = X$. As we show in Section VI-B, this method is not practical. However, it is useful for the uniqueness proof. For this method to recover the original signal X, first we need the solution of the DL in the first step to be unique (up to scaling and permutation). Therefore, we assume that the richness conditions on S and the spark condition on AP are satisfied. Under this assumption, the uniqueness of the solution to Problem 5 is achieved if the solution of the second step in the above method is unique. That is, under the richness and spark conditions the structural BCS problem is equivalent to the following problem:

Problem 6: Given matrices $\tilde{D}$ and A, which have more columns than rows, find an orthonormal $\tilde{P}$ such that $\tilde{D} = A\tilde{P}$, and $\tilde{P} = PQ$ for some permutation matrix Q and orthonormal $2L$-block diagonal matrix P.

In order to discuss conditions for uniqueness of the solution to Problem 6 we introduce the following definition.

Definition 6: Denote $A = [A_1, \ldots, A_L]$, 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

$$A_i^T A_j = \begin{bmatrix} R_1 & R_2 \\ R_3 & R_4 \end{bmatrix} \tag{12}$$

where $R_1, R_2, R_3, R_4 \in \mathbb{R}^{\frac{n}{2} \times \frac{n}{2}}$, satisfies:

$$\begin{aligned} &\operatorname{rank}(R_1) = \operatorname{rank}(R_4) \\ &\operatorname{rank}(R_2) = \operatorname{rank}(R_3) = \frac{n}{2} - \operatorname{rank}(R_1). \end{aligned} \tag{13}$$

An example of an inter-block diagonal matrix is presented in the next proposition:

Proposition 7: Assume $A = [A_1, \ldots, A_L]$ where $A_i \in \mathbb{R}^{n \times n}$ for any $1 \leq i \leq L$ and $\sigma(A) = n - 1$. If the product $A_i^T A_j$ for some $i \neq j$ is 2-block diagonal then A is inter-block diagonal.

Proof: Since $\sigma(A) = n - 1$ for any $1 \leq i \leq L$ the block A_i has full rank, so that $A_i^T A_j$ also has full rank. Therefore if for some $i \neq j$ the product $A_i^T A_j$ is 2-block diagonal then the relevant submatrices, as defined in (12), satisfy $R_2 = R_3 = 0$ and R_1, R_4 have full ranks, so that (13) holds. ∎

With Definition 6 in hand we can now derive conditions for uniqueness of Problem 6.

Theorem 8: If $A \in \mathbb{R}^{n \times nL}$ is a union of L orthonormal bases, which is not inter-block diagonal, and $\sigma(AP) = n+1$, then the solution to Problem 6 is unique.

The proof of this theorem relies on the following lemma.

Lemma 9: Assume P and $\hat{P}$ are both orthonormal $2L$-block diagonal matrices, and A satisfies the conditions of Theorem 8. If $A\hat{P} = APQ$ for some permutation matrix Q, then $\hat{P} = PQ$.

In general since A has a null space, if the matrices $A, P, \hat{P}$ do not have special structure, then the equality $A\hat{P} = APQ$ does not imply $\hat{P} = PQ$. However, according to Lemma 9 under the constraints on $A, P, \hat{P}$ this is guaranteed. The full proof of Lemma 9 appears in Appendix A. Here we present only the proof sketch.

Proof Sketch: It is easy to see that due to the orthonormality 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. The matrix 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 orthonrmality 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 that it is block diagonal. ∎

The extension of the proof of Lemma 9 to ML blocks where $M > 2$ is trivial. It is also clear from the proof that if P and $\tilde{P}$ have only L blocks instead of $2L$, then Q is no longer guaranteed to be block diagonal. Therefore, we cannot change the number of diagonal blocks in Lemma 9 to L instead of $2L$. This implies that the number of blocks in the structural BCS problem also cannot be only L.

Proof of Theorem 8: The proof we provide for Theorem 8 is constructive, although far from being a practical method to deploy in practice. Denote the desired solution to Problem 6 by $\tilde{P} = PQ$, and denote

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

where A_i for $i = 1, \ldots, L$ and P^j for $j = 1, \ldots, 2L$ are all orthonormal 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 orthonormal $2L$-block diagonal matrix. There is always at least one such permutation. For instance, we can choose $Q_D = Q^T$ so that $\hat{P} = P$, and therefore it is necessarily orthonormal $2L$-block diagonal.

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

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

Since A_i are orthonormal for all $i = 1, \ldots, L$, 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 orthonormal $2L$-block diagonal, and since the product QQ_D is itself a permutation matrix, according to Lemma 9 the equality $\hat{D} = A\hat{P} = APQQ_D$ implies $\hat{P} = PQQ_D$. Therefore, we can recover $\tilde{P}$ by $\tilde{P} = PQ = \hat{P}Q_D^T$. ■

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

One way to guarantee that A satisfies the conditions of Theorem 8 with probability 1 is to generate it randomly according to the following proposition, which is proved in Appendix B.

Proposition 10: If $A = [A_1, \ldots, A_L] \in \mathbb{R}^{n \times nL}$ is a union of L orthonormal bases, where each block is generated randomly from an i.i.d. Gaussian distribution followed by a Gram Schmidt process, then with probability 1 A satisfies the conditions of Theorem 8.

The conditions of Theorem 8 guarantee the uniqueness of the solution to Problem 6, which is equivalent to Problem 5 under the DL uniqueness condition. Therefore, the uniqueness conditions of Problem 5 are formulated as follows:

Theorem 11: If $A \in \mathbb{R}^{n \times nL}$ is a union of L orthonormal bases, which is not inter-block diagonal, and if $\sigma(AP) = n + 1$ and S satisfies the richness conditions, then the solution to Problem 5 is unique.

B. The OBD-BCS Algorithm

Although the proof of Theorem 8 is constructive it is far from being practical. In order to solve Problem 5 by following this proof one needs to first perform a DL algorithm on B, resulting in $\tilde{D}, \tilde{S}$. Then, it is necessary to try 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 $i = 1, \ldots, L$, 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 [37], these techniques are still computationally extensive and are sensitive to noise.

Instead, we present here the orthonormal block diagonal BCS (OBD-BCS) algorithm for the solution of Problem 5, which is, in theory, equivalent to DL followed by the above postprocessing. However, it is much more practical and simple. This algorithm is a variation of the DL algorithm in [28], [29], which learns a dictionary under the constraint that the dictionary is a union of orthonormal bases. Given B, the algorithm in [28] and [29] aims to solve

$$\min_{D,S} \|B - DS\|_F^2 \tag{14}$$

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

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

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

Since all A_i and P^i are orthonormal, here too D is a union of orthonormal bases. The measurement matrix A is known and we are looking for an orthonormal $2L$-block diagonal matrix P and a sparse matrix S such that $B = APS$. This leads to the following variant of (14):

$$\min_{P,S} \|B - APS\|_F^2 \tag{15}$$

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

The algorithm in [28], [29] 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 [28], [29] 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 (15) becomes

$$\min_S \|B - APS\|_F^2 \quad \text{s.t. } S \text{ is } k-\text{sparse}. \tag{16}$$

It is easy to see that (16) is separable in the columns of S. Therefore, for each column of B and S, we need to solve

$$\min_s \|b - APs\|_2^2 \quad \text{s.t. } \|s\|_0 \leq k \tag{17}$$

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 orthonormal bases. This property is used by the block coordinate relaxation (BCR) algorithm [28], [29], [38]. The idea behind this technique is to divide the elements of s into blocks corresponding to the orthonormal blocks of D. In each iteration, all the blocks of s are fixed except one, which is updated using soft thresholding. The DL method proposed by [28], [29] 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:

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

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

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

With this notation $X^i = P^i S^i$, and $B = \sum_{i=1}^{2L} A^i P^i S^i$. Therefore, (15) becomes

$$\min_{P^1,\ldots,P^{2L}} \left\| B - \sum_{j=1}^{2L} A^j P^j S^j \right\|_F^2 \quad \text{s.t. } P^1, \ldots, P^{2L} \text{ are orthonormal}. \tag{18}$$

To minimize (18), we iteratively fix all the blocks P^j for $j = 1, \ldots, 2L$ except one, denoted by P^i, and solve

$$\min_{P^i} \|B^i - A^i P^i S^i\|_F^2 \quad \text{s.t. } P^i \text{ is orthonormal} \tag{19}$$

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 orthonormal and A consists of columns from an orthonormal matrix, $P^T A^T A P = I$, and $\|APS\|_F^2 = \|S\|_F^2$. Thus, (19) reduces to

$$\max_P \{\text{Tr}[B^T APS]\} \quad \text{s.t. } P \text{ is orthonormal}. \tag{20}$$

Let the singular value decomposition (SVD) of the matrix $R = SB^T A$ be $R = U\Sigma V^T$, where U, V are orthonormal matrices and Σ is a diagonal matrix. Using this notation we can manipulate the trace in (20) as follows:

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

The matrix $Z = V^T PU$ is orthonormal if and only if P is orthonormal. Therefore, (20) is equivalent to

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

If the matrix $R = SB^T A$ has full rank then Σ is invertible. In this case the maximization is achieved by $Z = I$, and therefore $P^i = VU^T$ is the unique minimum of (19). Even if R does not have full rank $P^i = VU^T$ achieves a minimum of (19).

TABLE II
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 orthonormal bases)
Outputs:
• $\hat{X} \in \mathbb{R}^{nL \times N}$ - reconstructed signal matrix
Algorithm:
• Initiate $\hat{P} = I$ (the identity).
• Repeat until a stoping criteria is reached:
◦ *Sparse coding*: find the sparsest $\hat{S}$ such that $B = A\hat{P}\hat{S}$, for instance using OMP.
◦ *Basis update*: for all $i = 1, ..., 2L$:
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: $\hat{P}^i = VU^T$.
• Calculate: $\hat{X} = \hat{P}\hat{S}$.

Table II 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 is simple to implement. This algorithm is much simpler then following the uniqueness proof, which requires a combinatorial permutation search. Each iteration of OBD-BCS employs a standard CS algorithm and $2L$ SVDs.

As we discussed in the Section VI-A there is more than one possible pair P, S that follows the structural BCS conditions and satisfies $X = PS$. Therefore, although OBD-BCS finds a pair P, S this is not necessarily the original pair. However, under the uniqueness conditions P and S are unique up to scaling and permutation. Consequently, in this case OBD-BCS can be used in order to find the original matrices P, S up to scaling and permutation.

An important question that arises is whether the OBD-BCS algorithm converges. To answer this question we consider each step separately. If the sparse coding step is performed perfectly it solves (16) for the current P. That is, the objective of (15) is reduced or at least stays the same. In practice, for small enough k the CS algorithm converges to the solution of (16). However, in order to guarantee that the objective of (15) 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 one leading to the smallest objective.

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 stage is divided into $2L$ steps. In each, all the blocks of P are fixed except one, which is updated to minimize (19). Therefore, the objective of (19) is reduced or at least stays the same in each of the $2L$ steps constructing the basis update step. Therefore, the objective of (18), which is equivalent to (15) with fixed S, is reduced or not increased during the basis update step.

As in the algorithm we are based on [28], [29] and as in other DL techniques such as [18] and [30], we cannot prove that OBD-BCS converges to the unique minimum of (15). 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. Furthermore, as can be seen in Section VII-C the OBD-BCS algorithm performs very well in simulations on synthetic data.

VII. Simulation Results

In this section, we present simulation results on synthetic data of all the algorithms developed in this paper. In Section VII-A we show simulation results of the F-BCS method for the solution of Problem 3. In Section VII-B, we present the direct method for the solution of Problem 4. The OBD-BCS algorithm for the solution of Problem 5 is demonstrated in Section VII-C. In Section VII-D, we compare the results of all three methods.

A. F-BCS Simulation Results

For this simulation, we chose the set of bases Ψ to contain five bases of size 64×64: the identity, DCT [20], Haar wavelet, Symlet wavelet and Biorthogonal wavelet [19]. One-hundred 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 six 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 30 to 5 dB. For each noise level the F-BCS method was performed, where the CS algorithm we used was OMP [5].

Table III 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

$$e_i = \frac{\|x_i - \hat{x}_i\|_2}{\|x_i\|_2} \tag{21}$$

TABLE III
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%

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. Reconstruction of the rest of the signals obviously failed. As can be seen from Table III in the noiseless case the recovery is perfect and the error grows with the noise level. For high SNR there are no miss detections, but as the SNR decreases beyond 15 dB the percentage of miss detections increases. In these cases, one should use more than one signal, so that if one of the signals fails 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.

B. Sparse Basis Simulation Results

We now present simulation results for the direct method. First, 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. The matrix X was measured using a random Gaussian matrix A of size 128×256, resulting in $B = AX$.

We solved Problem 4 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 summaries 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 (21). 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 orthonormal with probability 1, $\sigma(A\Phi) = 129$. Therefore, with probability 1 the uniqueness of the sparse BCS method is

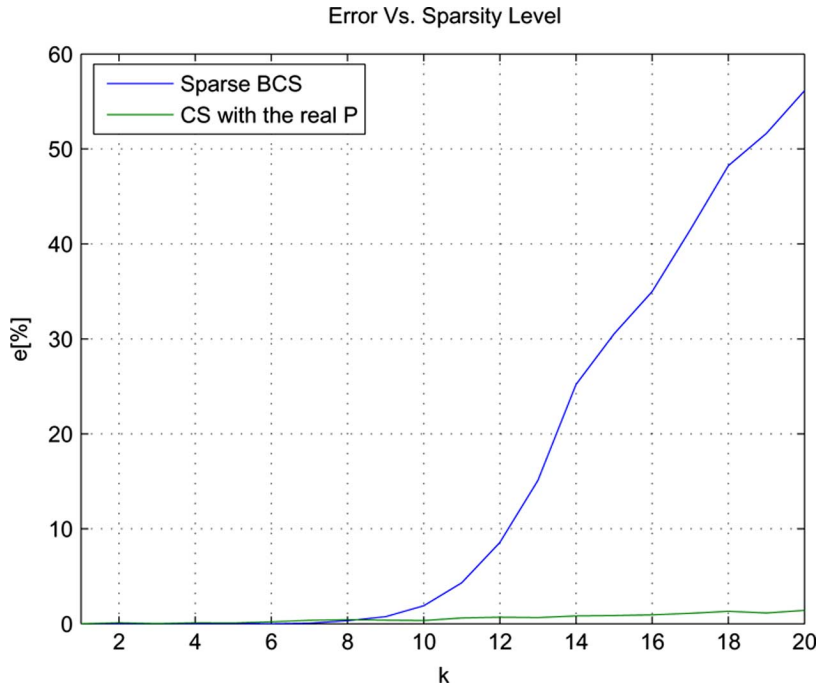

Fig. 1. Reconstruction error as a function of the sparsity level.

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 values 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 then 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 [21], 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 simulation was the same as in the previous simulation only this time we fixed $k = 3$ and added Gaussian noise to the measurements B. We looked at different noise levels, where for each level we ran the direct method for sparse BCS, and also for comparison an OMP algorithm which used the real basis P. Table IV summarizes the average errors of each of the methods. In the noiseless case both approaches lead to perfect recovery. As the SNR decreases both errors increase, but as can be expected, the one of the BCS grows faster. The reason for the big difference in the low SNR regime is again the fact that in standard CS the OMP algorithm is performed on sparser signals, relative to the sparse BCS setting.

TABLE IV
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%

C. OBD-BCS Simulation Results

In these simulations, the signal matrix X had 64 rows and was generated as a product of a random sparse matrix S and a random orthonormal 4-block diagonal matrix P. The value of the nonzero elements in S were generated randomly from a normal distribution, and the four orthonormal blocks of P were generated from a normal distribution followed by a Gram Schmidt process. The measurement matrix A was constructed from two random 32×32 orthonormal 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. That is, the algorithm stopped when either the change from the last iteration was too small, or when the maximal number of iterations was reached. In most cases the algorithm stopped due to small change between iterations after about 30 iterations.

First we examined the influence of N—the number of signals needed for the reconstruction, and k—the sparsity level. Fig. 2 considers the behavior as a function 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 blue (upper) graph is an average over 20 simulations of the OBD-BCS algorithm. In each simulation the sparse vectors and the orthonormal matrix where generated independently, but the measurement matrix was not changed. The error of each signal was calculated according to (21).

For comparison, the green (lower) graph in Fig. 2 is the average error of a standard CS algorithm that was performed on the same data, using the real basis P, which is unknown in practice.

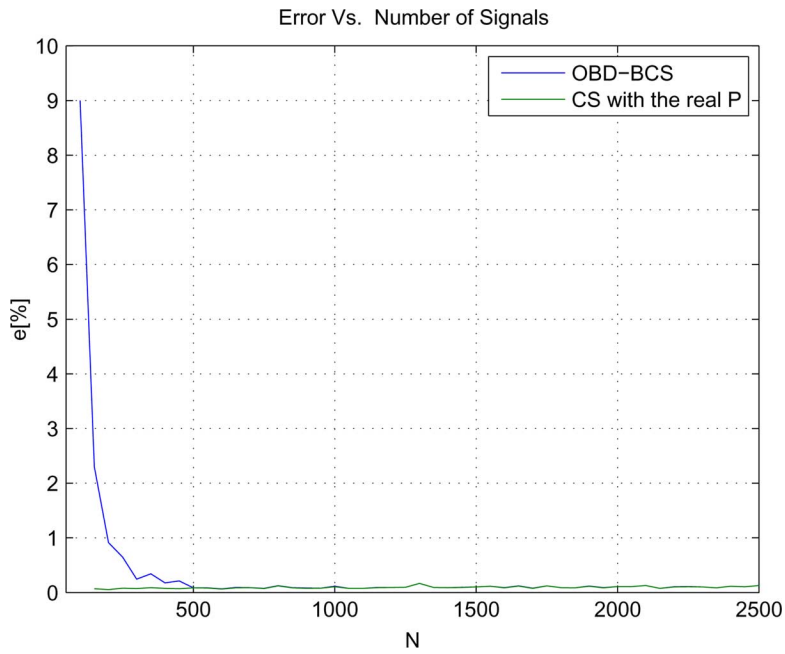

Fig. 2. Reconstruction error as a function of the number of signals, for sparsity level $k = 4$.

The CS algorithm we used was again OMP. As expected, the results of CS 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 P is known, is that for a small portion of the signals the OMP algorithm fails.

It is clear from Fig. 2 that for $N > 500$ reconstruction of the proposed algorithm is successful and similar to that obtained when P is known. Similarly to the conclusion in [15], the reconstruction is successful even for n much smaller then 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 methods, the technique in [28], [29] was evaluated by counting the number of columns of the dictionary that are detected correctly. The conclusions of [28], [29] 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 [28], [29], 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 $k = 1$, $k = 2$, and $k = 3$ 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 are calculated by $B = APS + W$, 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 which uses the real P. It is clear from the table that in the noiseless case the error of both algorithms is similar, therefore prior knowledge of the basis P can be avoided. As the SNR decreases both errors increase, but the error of OBD-BCS increases a bit faster than that of the CS algorithm. However, the difference is not very large.

D. Comparative Simulation

We conclude by illustrating 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 $k = 6$, the number of signals $N = 2000$, and the compression ratio $L = 2$. The syntectic data was generated as in Section VII-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 orthonormal 4-block diagonal matrix (each block is 16-block diagonal by itself).

We used five different reconstruction algorithms:

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

In all the methods above we used OMP as the standard CS algorithm. The first method, serves as a reference point since it uses the real basis P, whose knowledge we are trying to avoid. The second method is an intuitive way to reconstruct the signals when training is possible. Since the basis P is unknown one can estimate it first and then perform CS using the pre-estimated basis. Note however that this requires access to training data,

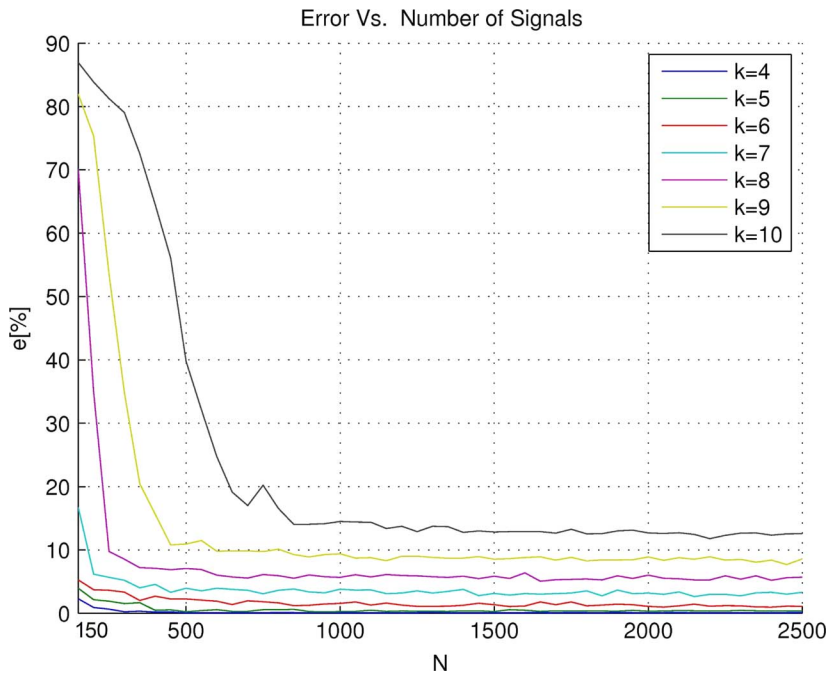

Fig. 3. Reconstruction error as a function of the number of signals for different values of k.

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%

which we do not assume in BCS. We performed the estimation using a training set of 2000 signals and DL. The estimated basis is denoted by P_{DL}. There are several different DL algorithms, e.g., [18], [28]–[30], [39]. However, in this case we have important prior knowledge that the basis P is orthonormal 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 orthonormal basis.

Due to this structure of P and due to 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 VII-A. The signal matrix 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 (21). As can be seen, the results of F-BCS are much worse than all the others. This can be expected since in this case OMP is operated with sparsity level of $2k$ instead of k. The error of the sparse BCS is also larger than the rest. The reason for this

TABLE VI
DL ALGORITHM FOR ORTHONORMAL DICTIONARY

Inputs
- X - training set
- k - sparsity level

Outputs
- P - orthonormal dictionary
- S - sparse matrix

Algorithm
- Initiate $P = I$.
- Repeat until a stoping criteria is reached:
  - 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 $P = VU^T$.

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}%

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 it is not small enough. However, note that though higher from the rest, the errors of the sparse BCS and F-BCS are quite small. We performed the same simulations with $k = 3$ in which case both the errors of sparse BCS and F-BCS were reduced to the level of the rest.

Both OBD-BCS and CS with an estimated basis, do not use the knowledge of the basis P. Nevertheless, the results of these algorithms are similar to those of the algorithm which uses this

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

VIII. Conclusion

We presented the problem of BCS which aims to solve CS problems without 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 its solution. Under each of these constraints we proved uniqueness conditions and proposed simple methods to retrieve the solution. The first two constraints reduce the problem to a CS or DL problem, which can be solved using methods based on existing algorithms. The third case of structural constraints required more extensive analysis. In particular, we developed a new algorithm for this setting, OBD-BCS.

The OBD-BCS algorithm reconstructs the signal using alternating steps of CS and basis update, where the latter is based on the structure of the basis. This method can be useful in multichannel systems, since the structure of such systems translates to block diagonal structure of P as required for OBD-BCS. This approach serves as a proof of the concept of BCS, showing that BCS can be solved efficiently under an applicable constraint which does not reduce the problem to a trivial one.

All the methods presented in this paper perform very well in simulations on synthetic data. In fact, the performance of our methods is similar to those of standard CS which uses the real, though unknown in practice, sparsity basis, as long as k is small enough and sufficiently many signals are measured (relevant only for the structural constraint case).

We also demonstrated through simulations the advantage of BCS over CS with an estimated sparsity basis. BCS does not require any training set, and therefore can be used in applications where there is no access to full signals but only to their measurements.

An interesting direction for future research is to examine more ways to assure uniqueness, besides the three presented here, and weaken the constraint on the basis. For example, one direction is using different measurement matrices on different signals without additional constraints on the sparsity basis.

Appendix A
Proof of Lemma 9

We prove the Lemma by showing that under its conditions Q is necessarily block diagonal. The completion of the proof is then straightforward.

For any $D = [D_1, \ldots, D_L] \in \mathbb{R}^{n \times nL}$ such that $D_1, \ldots, D_L \in \mathbb{R}^{n \times n}$, 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. The matrix 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 block permutation matrices, namely the permutation matrices that keep all blocks together. That is, if $Q \in Q_B$ then when multiplying Q by D to form DQ only the order of the blocks $D_1, \ldots, D_L$ and the order of the columns inside the blocks change, but there is no mixing between the blocks. To this end we rely on the following three lemmas.

Lemma A.1: If $D = [D_1, \ldots, D_L] \in \mathbb{R}^{n \times nL}$ is a union of L orthonormal 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, for any column d in D_1, which is not in Γ, the set of columns $\Gamma_2 \cup d$ is either linearly dependent or empty. However, the set $\Gamma_2 \cup d$ contains at most n columns, so that $\sigma(D) = n + 1$ implies that this set cannot be linearly dependent. Therefore, Γ_2 is necessarily empty, and all the columns of Γ are from the same block of D. ∎

Lemma A.2: Assume $D = [D_1, \ldots, D_L] \in \mathbb{R}^{n \times nL}$ is a union of L orthonormal bases, with $\sigma(D) = n + 1$, and $\hat{D} = DQ$ for some permutation matrix Q. Then, 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 orthonormal blocks of $\hat{D}$. However, according to Lemma A.1 any n orthogonal columns must be from the same block, and therefore $Q \in Q_B$. ∎

Lemma A.3: If $Q \in \mathbb{R}^{n \times n}$ is a permutation matrix and $R_1, R_2, R_3, R_4 \in \mathbb{R}^{\frac{n}{2} \times \frac{n}{2}}$ are the following submatrices:

$$Q = \begin{bmatrix} R_1 & R_2 \\ R_3 & R_4 \end{bmatrix} \tag{A-1}$$

then the ranks of these submatrices satisfy:

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

Proof: Since Q is a permutation matrix it has only one nonzero in each column and row. Therefore, if there is a nonzero in the ith row and jth column of R_1, then the ith row of R_2 and the jth column of R_3 are all zero. Consequently

$$\begin{aligned} \operatorname{rank}(R_1) + \operatorname{rank}(R_2) &= \frac{n}{2} \\ \operatorname{rank}(R_1) + \operatorname{rank}(R_3) &= \frac{n}{2}. \end{aligned} \tag{A-3}$$

The same consideration on R_4 implies that

$$\begin{aligned} \operatorname{rank}(R_4) + \operatorname{rank}(R_2) &= \frac{n}{2} \\ \operatorname{rank}(R_4) + \operatorname{rank}(R_3) &= \frac{n}{2}. \end{aligned} \tag{A-4}$$

Combining (A-3) and (A-4) results in (A-2). ∎

Denote the orthonormal blocks of A by A_i for $i = 1, \ldots, L$ and the orthonormal blocks of P and $\hat{P}$ by P^j and $\hat{P}^j$ respectively for $j = 1, \ldots, 2L$. Also denote

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

which are both unions of L orthonormal bases since A_i, P^j and $\hat{P}^j$ are all orthonormal. Therefore, if $A\hat{P} = APQ$, then 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 to the contrary that Q changes the outer order of the blocks of D. Without loss of generality we assume this change is a switch between the first two blocks of D. That is

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

where Q_1, Q_2 are the corresponding submatrices 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

$$\hat{D}_1 = A_1 \begin{bmatrix} \hat{P}^1 & \\ & \hat{P}^2 \end{bmatrix} = A_2 \begin{bmatrix} P^3 & \\ & P^4 \end{bmatrix} Q_2$$
$$\hat{D}_2 = A_2 \begin{bmatrix} \hat{P}^3 & \\ & \hat{P}^4 \end{bmatrix} = A_1 \begin{bmatrix} P^1 & \\ & P^2 \end{bmatrix} Q_1. \quad \text{(A-5)}$$

Since A_1 and A_2 are orthonormal, (A-5) implies that

$$\begin{bmatrix} \hat{P}^1 & \\ & \hat{P}^2 \end{bmatrix} = A_1^T A_2 \begin{bmatrix} P^3 & \\ & P^4 \end{bmatrix} Q_2$$
$$\begin{bmatrix} \hat{P}^3 & \\ & \hat{P}^4 \end{bmatrix} = A_2^T A_1 \begin{bmatrix} P^1 & \\ & P^2 \end{bmatrix} Q_1. \quad \text{(A-6)}$$

From (A-6)

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

where $R_1, R_2, R_3, R_4 \in \mathbb{R}^{\frac{n}{2} \times \frac{n}{2}}$. Since $\hat{P}^1, \hat{P}^2, P^3, P^4$ are all orthonormal, it is easy to see that the ranks of the submatrices R_1, R_2, R_3, R_4 equal the ranks of the corresponding submatrices of Q_2. Since Q_2 is a permutation matrix, Lemma A.3 implies that these ranks satisfy (A-2). Therefore, according to Definition 6, A is necessarily inter-block diagonal. However, according to the conditions of Theorem 8 A is not inter-block diagonal, so that Q cannot change the outer order of the blocks, and it must be L-block diagonal.

Denote the diagonal blocks of Q by Q_i for $i = 1, \ldots, L$. Then

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

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

$$\begin{bmatrix} \hat{P}^{2i-1} & \\ & \hat{P}^{2i} \end{bmatrix} = \begin{bmatrix} P^{2i-1} & \\ & P^{2i} \end{bmatrix} Q_i, \quad \text{(A-7)}$$

and since Q is block diagonal (A-7) implies $\hat{P} = PQ$, which concludes the proof of Lemma 9.

In fact, in this appendix we prove 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 $\hat{P}$ have L blocks instead of $2L$, then the proof fails. This is a result of the fact that in this proof in order to eliminate solutions of the form of (A-6) we use the 2-block diagonal structure of the matrices. If there are only L blocks in $P, \hat{P}$, then the matrices in (A-6) are no longer 2-block diagonal. In this case, besides the solution $\hat{P} = PQ$ there is another possibility:

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

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

Appendix B
Proof of Proposition 10

In order to prove Proposition 10, we show that under its requirements with probability 1 $\sigma(A) = n+1$ and A is not inter-block diagonal (Definition 6). In fact, the proposition claims that $\sigma(AP) = n+1$, however this is implied from $\sigma(A) = n+1$ and from the orthonormality of P. Namely, if $\sigma(A) = n+1$ then the rows of A are necessarily linearly independent, and since P is orthonormal so are the rows of AP.

In Proposition 10, each block of A is generated from a Gaussian distribution followed by a Gram Schmidt process. Before we begin the proof, we note that this random generation of each block can be viewed as the following equivalent process.

- The first column of the block, a_1, is generated randomly from a Gaussian distribution on $\mathbb{R}^n$.
- The second column, a_2, is generated randomly from the space G_2, which is the space orthogonal to a_1. The dimension of G_2 is $n-1$.
- Similarly, any column a_i is generated randomly from G_i, which is the space orthogonal to the span of all previous columns, and its dimension is $n-i+1$.
- Finally, all the columns are normalized.

While a_1 is generated randomly from a Gaussian distribution, the distributions from which the rest of the a_i's are generated are more complicated, but these distributions are necessarily continuous.

In order to prove that $\sigma(A) = n+1$ with probability 1, we use the following lemma.

Lemma B.1: 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 subspace of R^n with dimension d. If $d < n$ then with probability 1 none of the columns of G are in U.

Proof: Denote the columns of G by g_i for $i = 1, \ldots, n$. We start by showing that $g_1 \notin U$ with probability 1. According to the discussion above g_1 is generated randomly with a continuous (Gaussian) distribution on $\mathbb{R}^n$. The continuity of the distribution implies that the probability of g_1 to fall in any subspace of $\mathbb{R}^n$ with dimension smaller than n, is zero. Following the definition of Tropp in [13] such a subspace has zero volume in $\mathbb{R}^n$. Therefore, since $d < n$ the space U has zero volume in $\mathbb{R}^n$, so that $g_1 \notin U$ with probability 1. Similarly since the dimension of $U^\perp$ (the orthogonal complement space of U) is $n - d < n$ then $g_1 \notin U^\perp$ with probability 1.

The column g_2 is generated randomly with a continuous distribution on G_2—the space orthogonal to g_1, whose dimension is $d_2 = n-1$. The probability that $g_2 \in U$ equals the probability that $g_2 \in U \cap G_2$. If the dimension of $U \cap G_2$ is less than d_2, then $U \cap G_2$ has zero volume in G_2 so that $g_2 \notin U$ with probability 1. The dimension of $U \cap G_2$ is not smaller then d_2 only if $G_2 \subseteq U$, which implies that $U^\perp \subseteq (G_2)^\perp = \mathrm{span}(g_1)$. However, $g_1 \notin U^\perp$ with probability 1, therefore with probability 1 $U^\perp \subseteq (G_2)^\perp$ and $g_2 \notin U$. Similarly, $g_2 \notin U^\perp$ with probability 1.

We continue sequentially to the rest of the columns of G: For $2 < i \leq n$, the column g_i is generated randomly with a continuous distribution on G_i—the space orthogonal to the $i-1$ previous columns in G. Note that the dimension of G_i is $d_i = n - i + 1$. The probability that $g_i \in U$ equals the probability that $g_i \in U \cap G_i$. As before, this requires that $U^\perp \subseteq (G_i)^\perp$, however $(G_i)^\perp = \mathrm{span}(g_1, g_2, \ldots, g_{i-1})$ while $g_1, g_2, \ldots, g_{i-1} \notin U^\perp$ with probability 1. Therefore $g_i \notin U$ and $g_i \notin U^\perp$ with probability 1 for any $1 \leq i \leq n$, so that all the columns of G are not in U with probability 1. ∎

With Lemma B.1 in hand we can now prove that $\sigma(A) = n+1$ with probability 1. To that end we assume Γ is a set of $\sigma(A)$ linearly dependent columns from A. We will prove that $|\Gamma| = n+1$ with probability 1. Denote $\Gamma = \Gamma_1 \cup \Gamma_2$, where Γ_1 is the subset of Γ that contains 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. Note that since A_1 is orthonormal Γ_1 is also orthonormal, therefore Γ_2 cannot be empty, otherwise Γ is not linearly dependent.

Any $n+1$ columns from A are linearly dependent so that $\sigma(A) = |\Gamma| \leq n+1$. We assume by contradiction that $\sigma(A) = |\Gamma| \leq n$ and show that the probability for this is zero. Note that $|\Gamma| \leq n$ implies that $|\Gamma_1| < n$ and $|\Gamma_2| \leq n - |\Gamma_1|$. If $|\Gamma_1|$ contains only one column, denoted by γ_1, then this column must also be in the span of Γ_2, otherwise Γ is not linearly dependent. However, the dimension of this space is at most $|\Gamma_2| \leq n - |\Gamma_1| = n - 1$, and according to Lemma B.1 the probability for a column of A_1 to be in this space is zero. Similarly, 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 $|\Gamma_2| + 1 \leq n - 1$, so that again according to Lemma B.1 the probability for this is 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, $\sigma(A) = |\Gamma| = n+1$ with probability 1.

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

$$A_i^T A_j = \begin{bmatrix} R_1 & R_2 \\ R_3 & R_4 \end{bmatrix}. \tag{B-1}$$

For A to be inter-block diagonal there should be a pair $i \neq j$ for which (13) is satisfied. However, since A_i and A_j are orthonormal with probability 1 the rank of all the blocks R_1, R_2, R_3, R_4 is $\frac{n}{2}$ so that (13) is not satisfied, and A is not inter-block diagonal with probability 1.

Appendix C

Assume $A \in \mathbb{R}^{\frac{m}{L}\times m}$ is a union of L random orthonormal bases and $P \in \mathbb{R}^{m\times m}$ is an orthonrmal $2L$-block diagonal matrix. Denote $\tilde{D} = APQ$, where Q is some unknown 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 orthonormal $2L$-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 9 this implies $PQ_D = \hat{P}$. Since both P and $\hat{P}$ are $2L$-block diagonal Q_D must also be too $2L$-block diagonal, and the size of its blocks is $\frac{m}{2L} \times \frac{m}{2L}$. Since Q_D is a permutation matrix, 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 $[(\frac{m}{2L})!]^{2L}$.

Acknowledgment

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

References

[1] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[2] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[3] D. L. Donoho and M. Elad, "Maximal sparsity representation via l_1 minimization," *Proc. Nat. Acad. Sci.*, vol. 100, pp. 2197–2202, Mar. 2003.

[4] 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.

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

[6] I. Daubechies, M. Defrise, and C. De-Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. on Pure and Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, Aug. 2004.

[7] M. Mishali and Y. C. Eldar, "Blind multi-band signal reconstruction: Compressed sensing for analog signals," *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 993–1009, Mar. 2009.

[8] Y. C. Eldar, "Compressed sensing of analog signals in shift-invariant spaces," *IEEE Trans. Signal Process.*, vol. 57, no. 8, pp. 2986–2997, Aug. 2009.

[9] K. Gedalyahu and Y. C. Eldar, "Time delay estimation from low rate samples: A union of subspaces approach," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3017–3031, Jun. 2010.

[10] M. Mishali and Y. C. Eldar, "From theory to practice: Sub-Nyquist sampling of sparse wideband analog signals," *IEEE J. Select. Topics Signal Process.*, vol. 4, no. 2, pp. 375–391, Feb. 2010.

[11] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 5302–5316, Nov. 2009.
[12] E. Candes and T. Tao, "Decoding by linear programing," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.
[13] J. A. Tropp, "On the conditioning of random subdictionaries," *Appl. and Comput. Harmonic Anal.*, vol. 25, no. 1, 2008.
[14] M. Mishali, Y. C. Eldar, O. Dounaevsky, and E. Shoshan, "Xampling: Analog to digital at sub-Nyquist rates," *IET Circuits Devices and Syst.*, vol. 5, no. 1, pp. 8–20, Jan. 2011.
[15] M. Aharon, M. Elad, and M. Bruckstein, "On the uniqueness of overcomplete dictionaries, and practical way to retrieve them," *Linear Alg. and Its Applic.*, vol. 416, no. 1, pp. 48–67, 2006.
[16] M. S. Lewicki and T. J. Senowski, "Learning overcomplete representations," *Neural Comput.*, vol. 12, no. 2, pp. 337–365, 2000.
[17] 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 Comput.*, vol. 15, no. 2, pp. 349–396, 2003.
[18] K. Engan, S. O. Aase, and J. H. Husoy, "Frame based signal compression using method of optimal directions (MOD)," in *Proc. IEEE Int. Symp. Circuits and Systems*, Jul. 1999, vol. 4, pp. 1–4.
[19] S. Mallat, *A Wavelet Tour of Signal Processing*. New York: Academic Press, 1999.
[20] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transfom," *IEEE Trans. Comput.*, vol. 23, no. 1, pp. 90–93, Jan. 1974.
[21] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1553–1564, Mar. 2010.
[22] M. Mishali and Y. C. Eldar, "Sparse sourse seperation from orthogonal mixtures," *Proc. ICASSP*, pp. 3145–3148, Apr. 2009.
[23] R. Gribonval and K. Schnass, "Dictionary identification sparse matrix-factorization via l_1 minimization," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3523–3539, Jul. 2010.
[24] R. Gribonval and K. Schnass, "Dictionary identification from few training samples," in *Proc. 16th EUSIPCO*, Aug. 2008.
[25] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2845–2862, Nov. 2001.
[26] M. Elad and A. M. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," *IEEE Trans. Inf. Theory*, vol. 48, no. 9, pp. 2558–2567, Sep. 2002.
[27] R. Gribonval and M. Nielsen, "Sparse decompositions in unions of bases," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3320–3325, Dec. 2003.
[28] S. Lesage, R. Gribonval, F. Bimbot, and L. Benaroya, "Learning unions of orthonormal bases with thresholded singular value decomposition," in *Proc. ICASSP*, 2005, vol. 5, pp. 293–296.
[29] S. Lesage, R. Gribonval, F. Bimbot, and L. Benaroya, Learning Unions of Orthonormal Bases With Thresholded Singular Value Decomposition Tech. Rep. IRISA, 2004.
[30] M. Aharon, M. Elad, A. Bruckstein, and Y. Kats, "K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
[31] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Construct. Approx.*, vol. 28, no. 3, pp. 253–263, Dec. 2008.
[32] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3042–3054, Jun. 2010.
[33] P. Sprechmann, I. Ramirez, G. Sapiro, and Y. C. Eldar, "C-hilasso: A collaborative hierarchical sparse modeling framework," *IEEE Trans. Signal Process.*, to be published.
[34] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. New York: Springer, 2001.
[35] 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 Pers. Commun.*, vol. 5, no. 1, pp. 10–22, Feb. 1998.
[36] W. B. Pennebaker and J. L. Mitchell, *JPEG: Still Image Data Compression Standard*. New York: Van Nostrand Reinhold, 1993.
[37] H. F. Wang and K. Y. Wu, "Hybrid genetic algorithm for optimization problems with permutation property," *Comput. and Oper. Res.*, vol. 31, pp. 2453–2471, 2004.
[38] S. Sardy, A. G. Brouce, and P. Tseng, "Block coordinate relaxation methods for nonparametric wavelet denoising," *Comput. and Graph. Statist.*, vol. 9, no. 2, pp. 361–379, June 2000.
[39] M. Yaghoobi, T. Blumensath, and M. E. Davies, "Dictionary learning for sparse approximations with majorization method," *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2178–2191, Jun. 2009.

Sivan Gleichman received the B.Sc. degree in electrical engineering and the B.A. degree in physics in 2008, and the M.Sc. degree in electrical engineering in 2010, all *summa cum laude* , from the Technion—Israel Institute of Technology, Haifa, Israel. Currently she is a researcher with IBM Research Laboratory, Haifa, Israel.

Yonina C. Eldar (S'98–M'02–SM'07) received the B.Sc. degree in physics in 1995 and the B.Sc. degree in electrical engineering in 1996, both from Tel-Aviv University (TAU), Tel-Aviv, Israel, and the Ph.D. degree in electrical engineering and computer science in 2001 from the Massachusetts Institute of Technology (MIT), Cambridge.

From January 2002 to July 2002, she was a Postdoctoral Fellow with the Digital Signal Processing Group, MIT. She is currently a Professor with the Department of Electrical Engineering, the Technion–Israel Institute of Technology, Haifa, Israel. She is also a Research Affiliate with the Research Laboratory of Electronics, MIT, and a Visiting Professor with the Electrical Engineering and Statistics Departments, Stanford University, Stanford, CA. Her research interests are in the general areas of sampling theory, statistical signal processing, optimization methods, and their applications to biology and optics.

Dr. Eldar was in the program for outstanding students at TAU from 1992 to 1996. In 1998, she held the Rosenblith Fellowship for study in electrical engineering, MIT, and in 2000, she held an IBM Research Fellowship. From 2002 to 2005, she was a Horev Fellow of the Leaders in Science and Technology program, Technion, and an Alon Fellow. In 2004, she was awarded the Wolf Foundation Krill Prize for Excellence in Scientific Research, in 2005, the Andre and Bella Meyer Lectureship, in 2007, the Henry Taub Prize for Excellence in Research, in 2008, the Hershel Rich Innovation Award, the Award for Women with Distinguished Contributions, the Muriel and David Jacknow Award for Excellence in Teaching, and the Technion Outstanding Lecture Award, and in 2009, the Technion's Award for Excellence in Teaching. She is a member of the IEEE Signal Processing Theory and Methods Technical Committee and the Bio Imaging Signal Processing Technical Committee, an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the *EURASIP Journal of Signal Processing*, the *SIAM Journal on Matrix Analysis and Applications*, and the *SIAM Journal on Imaging Sciences*, and on the Editorial Board of *Foundations and Trends in Signal Processing*.