Sparse Solutions to Linear Inverse Problems With Multiple Measurement Vectors

Shane F. Cotter, *Member, IEEE*, Bhaskar D. Rao, *Fellow, IEEE*, Kjersti Engan, *Member, IEEE*, and Kenneth Kreutz-Delgado, *Senior Member, IEEE*

***Abstract*—We address the problem of finding sparse solutions to an underdetermined system of equations when there are multiple measurement vectors having the same, but unknown, sparsity structure. The single measurement sparse solution problem has been extensively studied in the past. Although known to be NP-hard, many single–measurement suboptimal algorithms have been formulated that have found utility in many different applications. Here, we consider in depth the extension of two classes of algorithms–Matching Pursuit (MP) and FOCal Underdetermined System Solver (FOCUSS)–to the multiple measurement case so that they may be used in applications such as neuromagnetic imaging, where multiple measurement vectors are available, and solutions with a common sparsity structure must be computed. Cost functions appropriate to the multiple measurement problem are developed, and algorithms are derived based on their minimization. A simulation study is conducted on a test-case dictionary to show how the utilization of more than one measurement vector improves the performance of the MP and FOCUSS classes of algorithm, and their performances are compared.**

I. Introduction

THE problem of computing sparse solutions (i.e., solutions where only a very small number of entries are nonzero) to linear inverse problems arises in a large number of application areas [1]. For instance, these algorithms have been applied to biomagnetic inverse problems [2], [3], bandlimited extrapolation and spectral estimation [4], [5], direction-of-arrival estimation [6], [3], functional approximation [7], [8], channel equalization [9], echo cancellation [10], image restoration [11], and stock market analysis [12]. It has also been argued that overcomplete representations and basis selection have a role in the coding of sensory information in biological systems [13], [14]. In all cases, the underlying linear inverse problem is the same and can be stated as follows: Represent a signal of interest using the minimum number of vectors from an overcomplete dictionary (set of vectors). This problem has been shown to be NP-hard [7], [15]. Much research effort has been invested in finding low complexity algorithms that yield solutions very close, in a chosen metric, to those obtained using an exhaustive search.

Manuscript received June 30, 2003; revised June 23, 2004. This work was supported in part by the National Science Foundation under Grant CCR-9902961. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Trac D. Tran.

S. F. Cotter, B. D. Rao, and K. Kreutz-Delgado are with the Electrical and Computer Engineering Department, University of California, San Diego, La Jolla, CA 92093-0407 USA (e-mail: scotter@ece.ucsd.edu; brao@ece.ucsd.edu; kreutz@ece.ucsd.edu).

K. Engan is with the University of Stavanger, Stavanger, Norway (e-mail: kjersti.engan@tn.his.no).

Digital Object Identifier 10.1109/TSP.2005.849172

A popular search technique for finding a sparse solution/representation is based on a suboptimal *forward* search through the dictionary [7], [8], [16]–[22]. These algorithms, termed Matching Pursuit (MP) [16], proceed by sequentially adding vectors to a set which will be used to represent the signal. Simple procedures were implemented initially [16], [17], while more complex algorithms were developed later which yielded improved results [7], [8], [18]–[22]. Other approaches have also been suggested which are based on the use of optimization techniques to minimize diversity measures and, hence, promote sparsity. In [23] and [24], the ℓ_1 norm of the solution was used as the diversity measure, and consideration of the more general $\ell_{(p\leq 1)}$ norm-like diversity measures led to the development of the FOCUSS (FOCal Underdetermined System Solver) class of algorithms [2], [3], [25]–[27]. A robust version of the FOCUSS algorithm, called Regularized FOCUSS, handles noisy data and can also be used as an efficient representation for compression purposes [28], [29]. Yet another approach was introduced in [30]–[32], where the search is based on a sequential backward elimination of elements from a complete or undercomplete (i.e., *non*overcomplete) dictionary. This has recently been extended in [33] and [34] to the case where the dictionary of elements is overcomplete.

In this paper, we consider in depth an important variation of the sparse linear inverse problem: the computation of sparse solutions when there are *multiple measurement vectors* (MMV) and the solutions are assumed to have a common sparsity profile. This work expands on some of the initial results presented in [35] and [36]. More recently, extensions of the matching pursuit framework to the MMV framework were also introduced and studied in [37]–[39]. It will be shown that we can greatly improve on our ability to provide sparse signal representations by utilizing MMV. As motivation for the study of this problem, we outline some applications in which MMV are at our disposal.

Our initial interest in solving the MMV problem was motivated by the need to solve the neuromagnetic inverse problem that arises in Magnetoencephalography (MEG), which is a modality for imaging the brain [2], [3], [40]. It is assumed that the MEG signal is the result of activity at a small number of possible activation regions in the brain. When several snapshots (measurement vectors) are obtained over a small time period, the assumption is made that the variation in brain activity is such that while the activation magnitudes change, the activation sites themselves do not. This naturally leads to the formulation of the MMV problem studied in this paper (see Section II). The formulation is also useful in array processing where there are multiple snapshots available, in particular, when the number

of snapshots is smaller than the number of sensors [3], [6]. Another important application of this formulation is in nonparametric spectrum analysis of time series where the data is often partitioned into segments for statistical reliability [41]. In this context, each segment corresponds to a measurement vector leading to the MMV problem. Recently, forward sequential search-based methods have been applied to the equalization of sparse channels which are found in some communication environments [9], [42]. In this case, for a fast time-varying channel, oversampling at the receiver leads to the MMV problem. While these applications are meant to highlight the importance of the MMV problem, the framework is quite general, and we are sure that the algorithms developed in Sections IV and V have application in many other areas.

The outline of the paper is as follows. In Section II, we formulate the MMV problem so that our framework is consistent with the applications we have outlined above. We address the issue of uniqueness in Section III. In Section IV, we show how the forward selection algorithms can be extended to solve MMV problems. In Section V, we extend the class of diversity measures used for the single measurement problem to the MMV problem, which leads us to derive a variant of the FOCUSS algorithm for the solution of the MMV problem. In the simulations of Section VI, we consider a test-case dictionary. The effects on the different MMV algorithms of increasing the number of measurement vectors and of varying the SNR are considered. We draw some conclusions in Section VII.

II. Problem Formulation

Noiseless Model: The noiseless MMV problem can be stated as solving the following L underdetermined systems of equations:

$$A\mathbf{x}^{(l)} = \mathbf{b}^{(l)}, \quad l = 1, \ldots, L \tag{1}$$

where $A \in \mathcal{C}^{m \times n}$, $m < n$, and, often, $m \ll n$. It is assumed that A has full row rank (rank$(A) = m$). L is the number of measurement vectors and it is usually assumed that $L < m$. The quantities $\mathbf{b}^{(l)} \in \mathcal{C}^m$, $l = 1, \ldots, L$ are the measurement vectors, and $\mathbf{x}^{(l)} \in \mathcal{C}^n$, $l = 1, \ldots, L$ are the corresponding source vectors. Assumptions on the structure of these source vectors are stated below.

In the past, algorithm development has mainly dealt with the problem of one measurement vector, i.e., $L = 1$ [22], [25]. Here, we concentrate on the case where $L > 1$, as initially considered in [35] and [36]. Since the matrix A is common to each of the L representation problems, we can succinctly rewrite (1) as

$$AX = B \tag{2}$$

where $X = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(L)}]$, and $B = [\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(L)}]$. In formulating the MMV problem, we make the following distinct and important assumptions about the desired solution.

Solution Vector Assumptions:

1) The solution vectors $\mathbf{x}^{(l)}$, $l = 1, \ldots, L$ are sparse, i.e., most of the entries are zero. This requirement is the same as that imposed in the single measurement vector case.
2) The solution vectors $\mathbf{x}^{(l)}$, $l = 1, \ldots, L$ are assumed to have the *same* sparsity profile so that the indices of the nonzero entries are independent of l. This requirement provides informative coupling between the vectors, but it also leads to additional complexity in formulating algorithms to solve the sparse linear inverse problem. The number of nonzero rows is referred to as the diversity of the solution.

Our emphasis in this paper is on Assumption 2. Because of the assumption that A has full row rank, (2) is consistent and always has a solution. The issue is how to find a maximally *sparse* solution from among the infinity of solutions which exist because $m < n$ (and usually $m \ll n$). Unfortunately, it has been shown for $L = 1$ that finding the solution that has the minimum number of nonzero entries is NP-hard [7]. The MMV problem further complicates the problem, particularly in the problem addressed here, where the values in each nonzero position of $\mathbf{x}^{(l)}$, $l = 1, \ldots, L$ can be very different. Coherent combining of the data and reducing the MMV to a single vector problem is not feasible, and new methods are called for. Because of the difficulty in finding the optimally sparse solution to (2), the suboptimal algorithms we develop seek a good compromise between complexity and optimality of solution.

Measurement Noise: The model (2) is noiseless. This is often an oversimplification either because of modeling error or because a nonnegligible level of noise is actually present. The addition of noise terms to the model (2) provides a mechanism for dealing with both situations. A model including additive noise can be written as

$$AX + N = B \tag{3}$$

where $X = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(L)}]$, $B = [\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(L)}]$, and $N = [\mathbf{n}^{(1)}, \ldots, \mathbf{n}^{(L)}]$. $\mathbf{n}^{(l)} \in \mathcal{C}^m$ represent the additive noise. In the presence of noise, an additional complicating factor one has to consider is the tradeoff between quality of fit, e.g., as measured by $\|AX - B\|$ and the sparsity of the solution.

Measures of Algorithm Performance: For evaluation purposes, we denote the actual sparse generating matrix by X_0 (which produces the observations B) and the resulting solution found by a subset selection algorithm by X. We propose two measures for measuring the performance of the subset selection algorithm:

- the number (or percentage) of generating columns from A, which are used in forming B in (2) or (3), that are correctly identified;
- the relative mean squared error (MSE) between the true and the estimated solution, which is calculated as

$$\text{MSE} = E\left(\frac{\|X - X_0\|_F^2}{\|X_0\|_F^2}\right). \tag{4}$$

The true generating X_0 is not usually known and can be replaced by a solution obtained using an exhaustive search method. An alternative, and an approach we use, is to use synthetic data wherein the generating sparse matrix X_0 is known and can be used for performance evaluation.

III. MMV and Sparsity

In this section, we develop uniqueness results in the noiseless case which can be helpful in identifying the optimality of the

obtained solution. Several results have been recently developed for the $L = 1$ case [43]–[49]. The presence of multiple measurements will be shown to be helpful in this regard. In the case where the diversity r of the solution is known to be bounded from above by the value r_u, we look for an exact solution X to (2), which has only $r \leq r_u$ nonzero rows. The following lemma, which is an extension of the results for $L = 1$ in [3], proves that a unique solution exists subject to certain conditions on the dictionary and the measurement vectors.

Lemma 1: Consider the MMV problem of (2). With the assumptions that any m columns of A are linearly independent and $\text{rank}(B) = L \leq m$*, a solution with number of nonzero entries r, where $r \leq r_u = \lceil (m+L)/2 \rceil - 1$, is unique (where $\lceil \cdot \rceil$ denotes the ceiling operation).*

Proof: We will prove the lemma by showing that all other sparse solutions must have diversity greater than r. Let $\mathbf{x}_i^{(l)}, i = 1, 2; l = 1, \ldots, L$, contain the amplitudes of the nonzero entries of two solutions to (2). Let r_1 and r_2 represent the diversity of the two solutions. Then, $A_1 \mathbf{x}_1^{(l)} = \mathbf{b}^{(l)}$, $l = 1, \ldots, L$ and $A_2 \mathbf{x}_2^{(l)} = \mathbf{b}^{(l)}$, $l = 1, \ldots, L$, where A_1 and A_2 have r_1 and r_2 columns, respectively. We rewrite this as

$$[A_1 \; A_2]\mathbf{y}^{(l)} = 0, \text{ where } \mathbf{y}^{(l)} = \begin{bmatrix} \mathbf{x}_1^{(l)} \\ -\mathbf{x}_2^{(l)} \end{bmatrix}, \; l = 1, \ldots, L. \quad (5)$$

We assume that A_1 and A_2 share no common columns, i.e., matrix $[A_1 \; A_2]$ has $r_1 + r_2$ columns.[1] Using the assumption that $\text{rank}(B) = L, \mathbf{y}^{(l)}, l = 1, \ldots, L$ are linearly independent. It follows from (5) that matrix $[A_1 \; A_2]$ has a null space of dimension at least L. This implies that $r_1 + r_2 \geq m + L$ or, equivalently, $r_2 \geq m + L - r_1$. If $r_1 \leq \lceil (m+L)/2 \rceil - 1$, then $r_2 > r_1$, which gives the uniqueness result. □

Lemma 1 indicates that any other solution to (2) must have a diversity $r > \lceil (m+L)/2 \rceil - 1$. If we know *a priori* that $r \leq \lceil (m+L)/2 \rceil - 1$, this lemma justifies termination of computational algorithms once a solution with less than or equal to $\lceil (m+L)/2 \rceil - 1$ nonzero rows has been obtained. Furthermore, we can expect the generated solution to be equal to the true solution. In addition, given the improvement in the bound because of the multiple measurement vectors, the ability to find the true solution also should improve.

IV. Forward Sequential Selection Methods

The methods described in this section find a sparse solution by sequentially building up a small subset of column vectors selected from A to represent B. *Selection of a column of A corresponds to selecting a nonzero row of X.* The algorithms described use different criteria to select a column vector and, consequently, have different performance and computational complexity. Each of the methods presented in the following subsections are motivated by and are extensions of methods developed for $L = 1$. The algorithm descriptions will be kept brief, and the reader is referred to [22] for a more detailed description of the algorithms for $L = 1$, as well as an examination of their complexity. For completeness, though not considered in this paper, we would like to mention MMV algorithms that are based on Backward Elimination algorithms. These eliminate vectors sequentially from the available dictionary until a sparse solution is obtained and can also be developed by extending the corresponding $L = 1$ algorithms [32]–[34]. Now, we introduce some notation to facilitate the presentation which is summarized in Table I. Without loss of generality, it is assumed that the columns of the matrix A are of unit norm.

[1]The case where they share columns can be dealt with in a similar manner.

TABLE I
Algorithm Notation Used in the Description of the Forward Sequential Selection Methods

- $\|\cdot\|$ - denotes the 2-norm.
- B_p - the residual vectors after the pth iteration, where $B_0 = B$. The lth column of B_p is denoted by $\mathbf{b}_p^{(l)}$.
- $I_p = \{k_1, k_2, \cdots, k_p\}$, $I_0 = \emptyset$. This set stores the indices k_i of the p vectors selected.
- $S_p = [\mathbf{a}_{k_1}, \mathbf{a}_{k_2}, \cdots, \mathbf{a}_{k_p}]$, $S_0 = \emptyset$. This matrix stores the selected vectors as columns.
- P_{S_p} - the orthogonal projection matrix onto the range space of S_p. Its orthogonal complement $P_{S_p}^{\perp} = (I - P_{S_p})$, $P_{S_0} = 0$, $P_{S_0}^{\perp} = I$.
- $P_{\mathbf{a}_l} = \mathbf{a}_l \mathbf{a}_l^H$ - the projection matrix onto the space spanned by a single unit norm vector $\mathbf{a}_l$ is denoted by $P_{\mathbf{a}_l}$.

A. MMV Basic Matching Pursuit (M-BMP)

In the M-BMP algorithm, we first find the column in the matrix A, which is best aligned with the measurement vectors comprising the columns of $B_0 = B$, and this is denoted $\mathbf{a}_{k_1}$. Then, the projection of B_0 along the direction $\mathbf{a}_{k_1}$ is removed from B_0, and the residual B_1 is obtained. Next, the column $\mathbf{a}_{k_2}$ in A, which is best aligned with B_1, is found, and a new residual B_2 is formed. Thus, the algorithm proceeds by sequentially choosing the column that best matches the residual matrix. We now detail the M-BMP algorithm by looking at the pth iteration.

In the pth iteration, we find the vector most closely aligned with the residual B_{p-1} by examining the residual $E_{p,k} = P_{\mathbf{a}_k}^{\perp} B_{p-1}$ for each column vector $\mathbf{a}_k$, $k = 1, \ldots, n$ in A. The vector which minimizes the Frobenius norm of the error is selected, i.e.,

$$\begin{aligned} \|E_{p,k}\|_F^2 &= \text{Tr}(E_{p,k}^H E_{p,k}) = \text{Tr}(B_{p-1}^H P_{\mathbf{a}_k}^{\perp} B_{p-1}) \\ &= \|B_{p-1}\|_F^2 - \text{Tr}(B_{p-1}^H P_{\mathbf{a}_k} B_{p-1}). \end{aligned}$$

The minimization is achieved by maximizing the second term $\text{Tr}(B_{p-1}^H P_{\mathbf{a}_k} B_{p-1})$ in the above expression. Using the fact that $P_{\mathbf{a}_k} = \mathbf{a}_k \mathbf{a}_k^H$, we select the column as

$$k_p = \arg\max_k \|\mathbf{z}_k\|^2, \text{ where } \mathbf{z}_k^H = \mathbf{a}_k^H B_{p-1}, k = 1, \ldots, n. \quad (6)$$

If $k_p \notin I_{p-1}$, the index and basis sets are updated, i.e., $I_p = I_{p-1} \cup \{k_p\}$ and $S_p = [S_{p-1}, \mathbf{a}_{k_p}]$. Otherwise, we set $I_p = I_{p-1}$ and $S_p = S_{p-1}$. The new residual vector is computed as $B_p = P_{\mathbf{a}_k}^{\perp} B_{p-1}$ or, more explicitly

$$B_p = P_{\mathbf{a}_{k_p}}^{\perp} B_{p-1} = B_{p-1} - \mathbf{a}_{k_p} \mathbf{z}_{k_p}^H. \quad (7)$$

Equations (6) and (7) give the M-BMP algorithm (with $B_0 = B$). There are two possibilities for termination of the algorithm. We may terminate the algorithm when $\|B_p\|_F \leq \epsilon$ (for specified

error ϵ), which gives a good approximate solution but no control over the sparsity of the solution generated. Alternatively, we terminate when a prespecified number (r) of distinct columns have been chosen, which means that we have a solution of the required sparsity. In the low noise case, a good approximation to the true solution using r columns from A should still be obtainable, although an exact solution is no longer possible. The actual solution matrix X can be obtained by solving a Least Squares (LS) problem using the chosen subset of dictionary vectors and the measurement vectors similar to the $L = 1$ case [22].

B. MMV Orthogonal Matching Pursuit (M-OMP)

This procedure, which is also referred to as Modified Matching Pursuit (MMP) in our earlier work [22], is a modification of the BMP method and seeks to improve the computation of the residual matrix B_{p-1} [22]. The index selection is computed as in (6), but the residual matrix B_p is computed as $P^{\perp}_{S_p} B_{p-1}$ as opposed to $P^{\perp}_{\mathbf{a}_{k_p}} B_{p-1}$. To obtain the new residual matrix, we need to first form the projection matrix $P_{S_p} = P_{[S_{p-1}, \mathbf{a}_{k_p}]}$. Once the column index k_p has been selected, a Modified Gram–Schmidt [50] type of procedure is used on the vector $\mathbf{a}_{k_p}$. With the initialization $\hat{\mathbf{a}}^{(0)}_{k_p} = \mathbf{a}_{k_p}$, $\mathbf{q}_0 = 0$, we have $P_{S_p} = P_{S_{p-1}} + \mathbf{q}_p \mathbf{q}_p^H$, where

$$\begin{aligned} \hat{\mathbf{a}}^{(\ell)}_{k_p} &= \hat{\mathbf{a}}^{(\ell-1)}_{k_p} - (\mathbf{q}^H_{\ell-1} \hat{\mathbf{a}}^{(\ell-1)}_{k_p}) \mathbf{q}_{\ell-1}, \ \ell = 1, \ldots, p \\ \mathbf{q}_p &= \frac{\hat{\mathbf{a}}^{(p)}_{k_p}}{\|\hat{\mathbf{a}}^{(p)}_{k_p}\|}. \end{aligned} \tag{8}$$

The residual B_p is updated via

$$B_p = P^{\perp}_{S_p} B_{p-1} = B_{p-1} - \mathbf{q}_p (\mathbf{q}_p^H B_{p-1}). \tag{9}$$

Equations (6), (8), and (9) define the M-OMP algorithm, and similar stopping rules to those used for M-BMP are used. We have described how the generating vectors are identified, but to obtain the solution matrix X, a backsolve is necessary, as in the $L = 1$ formulation [22].

C. MMV Order Recursive Matching Pursuit (M-ORMP)

This method is an extension of the methodology developed in [7], [18], and [21]. In this method, the pursuit of the matching pth basis vector conceptually involves solving $(n-p+1)$ order recursive least squares problems of the type $\min_Y \|S_p^{(k)} Y - B\|_F^2$, where we use the notation $S_p^{(k)} = [S_{p-1}, \mathbf{a}_k]$. The vector $\mathbf{a}_k \notin S_{p-1}$ that reduces the residual the most is selected and added to S_{p-1} to form S_p.

With the initialization $\mathbf{a}_k^{(0)} = \mathbf{a}_k, k = 1, \ldots, n$, and $B_0 = B$, the index selection criterion in the pth iteration is given by

$$k_p = \arg\max_k \frac{\|\mathbf{z}_k\|^2}{\|\mathbf{a}_k^{(p-1)}\|^2} \ \ k \notin I_{p-1} \quad \text{where } \mathbf{z}_k^H = \mathbf{a}_k^H B_{p-1}, \ k = 1, \ldots, n. \tag{10}$$

The form of (10) is similar to (6), but there are a number of important differences. First, as in M-OMP, the residual is calculated by projecting B onto $S^{\perp}_{p-1}$, i.e., $B_{p-1} = P^{\perp}_{S_{p-1}} B$. Second, we have a denominator term $\|\mathbf{a}_k^{(p-1)}\|^2$, where $\mathbf{a}_k^{(p-1)} = P^{\perp}_{S_{p-1}} \mathbf{a}_k$, which must be computed for each vector $\mathbf{a}_k, k = 1, \ldots, n$.

Once the column $\mathbf{a}_{k_p}$ has been selected, we update $I_p = I_{p-1} \cup \{k_p\}$, $S_p = [S_{p-1}, \mathbf{a}_{k_p}]$ and then, using (8), compute $\mathbf{q}_p$ such that $P_{S_p} = P_{S_{p-1}} + \mathbf{q}_p \mathbf{q}_p^H$. This enables us to update recursively the norms $\|\mathbf{a}_l^{(p)}\|, l = 1, \ldots, n$ required in the denominator of (10) as follows [22]:

$$\begin{aligned} \|\mathbf{a}_l^{(p)}\|^2 &= \mathbf{a}_l^H (I - P_{S_p}) \mathbf{a}_l = \mathbf{a}_l^H (I - P_{S_{p-1}}) \mathbf{a}_l - \mathbf{a}_l^H \mathbf{q}_p \mathbf{q}_p^H \mathbf{a}_l \\ &= \|\mathbf{a}_l^{(p-1)}\|^2 - |\mathbf{q}_p^H \mathbf{a}_l|^2. \end{aligned} \tag{11}$$

The pth iteration is completed by computing B_p as in (9). Equations (8)–(11) constitute the M-ORMP algorithm. The termination procedure is the same as that used in the M-BMP and M-OMP algorithms. If the solution matrix X is required, a backsolve must be performed as in the M-OMP [22].

D. Convergence of Matching Pursuit Algorithms

There are no new complications with regard to the convergence of the class of forward sequential algorithms developed in the previous subsections. The algorithms can be used for noiseless and noisy data with no change, except for some modification in the termination criteria. It has been shown with $L = 1$ that the residual vector will monotonically decrease in magnitude [16], and this is easily extended to $L > 1$. Certain cases for M-BMP exist where there may be cycling among several elements of the dictionary [24], and therefore, the convergence may be very slow. Anticycling rules may be used to deal with such situations. In most cases, the selection of m elements from the dictionary will yield a linearly independent set of vectors. Due to the suboptimality of the residual computed in the M-BMP algorithm, the residual is not reduced to zero even after m iterations of this algorithm [16]. A more complete study of convergence and rate of convergence can be found in [39].

V. Diversity Minimization Methods

Now, we consider another class of algorithms based on minimization of diversity measures, which, for $L = 1$, has been found to be promising [26], [27]. Of particular interest is the FOCUSS algorithm, which is an alternative and complementary approach to the forward sequential methods [25]. In this section, we extend the FOCUSS algorithm to incorporate MMV and expand on the work presented in [35] and [36].

A. Background

In this approach, all the vectors are initially selected, and an iterative procedure is employed to asymptotically eliminate column vectors until only a small number of columns remain [3]. In developing this methodology, we start with the noiseless problem and assume that an exact solution of diversity r exists which satisfies (2). Any solution can be expressed as

$$X = X_{mn} + V \tag{12}$$

where X_{mn} is the minimum Frobenius norm solution and is given by $X_{mn} = A^{\dagger} B$. $A^{\dagger} = A^H (A A^H)^{-1}$ denotes the Moore–Penrose pseudo-inverse, and the lth column of X_{mn}

is the minimum 2-norm solution to the system of equations $A\mathbf{x}^{(l)} = \mathbf{b}^{(l)}$. The matrix V is a matrix whose column vectors $\mathbf{v}^{(l)}, l = 1, \ldots, n$ lie in the null space of A so that $AV = 0$..

In many situations, a popular approach has been to set $V = 0$ and to select X_{mn} as the desired solution. This has two main drawbacks. First, the minimum 2-norm solutions which make up X_{mn} are based on a criterion that favors solutions with many small nonzero entries, which is contrary to the goal of sparsity/concentration [3], [23]. Second, the solution for each of the measurement vectors is computed independently so that a common sparsity structure is *not* enforced across the solutions. The first problem (i.e., the use of the 2-norm criterion) has been addressed in [3] and [25], but the second problem, called common sparsity enforcement, has not been addressed. This is dealt with in the next subsection.

B. Diversity Measures for the MMV Problem

Since the minimum 2-norm solutions are nonsparse, we need to consider alternate functionals, referred to as diversity measures, which lead to sparse solutions when minimized. A popular diversity measure for vectors ($L = 1$) is $E^{(p)}(\mathbf{x})$ [11], [25], [31], [51]–[53], where

$$E^{(p)}(\mathbf{x}) = \sum_{i=1}^{n} |x[i]|^p, \quad 0 \le p \le 1.$$

Due to the close connection to ℓ_p norms, these measures are referred to as "$\ell_{(p\le 1)}$ diversity measures" or "p-norm-like diversity measures." The diversity measure for $p = 0$ (or, equivalently, the *numerosity* discussed in [52]) is of special interest because it is a *direct* measure of sparsity. It provides a count of the number of nonzero components in $\mathbf{x}$:

$$E^{(0)}(\mathbf{x}) = \#\{i : x[i] \neq 0\}.$$

Finding a global minimum to the numerosity measure requires an enumerative search that is NP-hard [7]. Consequently, alternate diversity measures that are more amenable to optimization techniques are of interest. The $E^{(p)}(\mathbf{x})$ measures for $0 < p \le 1$ are useful candidate measures in this context [11], [51]–[53].

All the above measures are relevant to the single measurement case ($L = 1$), and not much work is available for the MMV problem. To extend the measures to the MMV scenario, a good starting point is to consider suitably extending diversity measures developed for $L = 1$ such as the Gaussian or Shannon entropy measures, among others [53]. Extension of the measures to the MMV problem were also considered in the matching pursuit context in [38]. A general and comprehensive study of diversity measures for the MMV problem is outside the scope of this work. Instead, we present one measure which our study has shown to hold much promise. It is an extension of the $\ell_{(p\le 1)}$ diversity measure, which has often been found to produce better results than other diversity measures for $L = 1$ [25]. The modified measure is given by

$$J^{(p,q)}(X) = \sum_{i=1}^{n} (\|\mathbf{x}[i]\|_q)^p, \; 0 \le p \le 1, q \ge 1 \tag{13}$$

where $\mathbf{x}[i] = [x^{(1)}[i], x^{(2)}[i], \ldots, x^{(L)}[i]]$ is the ith row of X, and the row norm is given by $\|\mathbf{x}[i]\|_q = \left(\sum_{l=1}^{L} |x^{(l)}[i]|^q\right)^{1/q}$. For simplicity, we consider the case $q = 2$ in the rest of this paper and denote $J^{(p,2)}(X)$ by $J^{(p)}(X)$, i.e.,

$$J^{(p)}(X) = \sum_{i=1}^{n} (\|\mathbf{x}[i]\|_2)^p = \sum_{i=1}^{n} \left(\sum_{l=1}^{L} |x^{(l)}[i]|^2\right)^{p/2}. \tag{14}$$

This choice of cost function may be motivated in two ways. First, it may be seen that as p approaches zero, it provides a count of the number of nonzero *rows* in X. A nonzero row gets penalized as p is reduced, which promotes a common sparsity profile across the columns of X. Second, pragmatic considerations such as computational complexity also favor its utilization. The minimization of the $J^{(p)}(X)$ measure (14) will be found to lead to a low complexity computational algorithm.

C. M-FOCUSS Algorithm

Starting from this measure (13), the factored-gradient approach of [25] and [53] is used to develop an algorithm to minimize it, subject to the constraint (2). This algorithm, which is useful in the noiseless case, represents an extension of the FOCUSS class of algorithms developed for $L = 1$ to the MMV case. Therefore, it is referred to as M-FOCUSS. The details are given in Appendix A, and the algorithm is summarized as follows:

$$\begin{aligned} W_{k+1} &= \text{diag}(c_k[i]^{1-p/2}), \\ &\qquad \text{where } c_k[i] = \|\mathbf{x}_k[i]\| = \left(\sum_{l=1}^{L} (x_k^{(l)}[i])^2\right)^{1/2}, \; p \in [0,1] \\ Q_{k+1} &= A_{k+1}^{\dagger} B, \text{ where } A_{k+1} = AW_{k+1} \\ X_{k+1} &= W_{k+1} Q_{k+1}. \end{aligned} \tag{15}$$

As mentioned in Section V-B, we choose $p \in [0, 1]$ in order to encourage sparsity. The algorithm is terminated once a convergence criterion has been satisfied, e.g.,

$$\frac{\|X_{k+1} - X_k\|_F}{\|X_k\|_F} < \delta$$

where δ is a user-selected parameter.[2]This algorithm can be proven to reduce $J^{(p)}(X)$ in each iteration (Section V-E).

D. Regularized M-FOCUSS Algorithm

Now, we generalize the algorithm to deal with additive noise by developing the Regularized M-FOCUSS algorithm, which is a generalization of Regularized FOCUSS [28], [29]. The algorithm is summarized as follows:

$$\begin{aligned} W_{k+1} &= \text{diag}(c_k[i]^{1-p/2}) \\ &\qquad \text{where } c_k[i] = \left(\sum_{l=1}^{L} (x_k^{(l)}[i])^2\right)^{1/2}, \; p \in [0,2] \\ Q_{k+1} &= A_{k+1}^H (A_{k+1} A_{k+1}^H + \lambda I)^{-1} B \\ &\qquad \text{where } A_{k+1} = AW_{k+1} \text{ with } \lambda \ge 0 \\ X_{k+1} &= W_{k+1} Q_{k+1}. \end{aligned} \tag{16}$$

[2]In our experiments, δ was chosen as 0.01.

Note that the M-FOCUSS algorithm corresponds to setting λ to zero in Regularized M-FOCUSS. There are two useful ways to view the Regularized M-FOCUSS algorithm. One is by viewing the algorithms (Regularized M-FOCUSS and M-FOCUSS) as solving at each iteration a weighted least squares (WLS) problem with the Regularized M-FOCUSS algorithm, providing a more robust solution. This can be seen by examining the difference between the two underlying WLS problems in (15) and(16). The iterative step provided in (16) can be regarded as a solution to the following Tikhonov regularization problem:

$$Q_{k+1} = \arg\min_{Q} \left(\|AW_{k+1}Q - B\|_F^2 + \lambda \|Q\|_F^2 \right)$$

where $\|.\|_F$ is the Frobenius norm. The identity $(A^HA + \lambda I)^{-1}A^H = A^H(AA^H + \lambda I)^{-1}$ is useful in establishing this result. Alternately

$$X_{k+1} = \arg\min_{X} G_{k+1}(X)$$
$$\text{where } G_{k+1}(X) = \|AX - B\|_F^2 + \lambda \|W_{k+1}^{-1}X\|_F^2. \quad (17)$$

If $X_{k+1} \neq X_k$, then

$$G_{k+1}(X_{k+1}) < G_{k+1}(X_k). \quad (18)$$

Insight into the algorithm can also be obtained using this viewpoint. In weighted least squares, the weights applied to the columns play a role in determining the contribution of the columns to the final solution. A small weight usually results in a smaller contribution and vice versa. Since the column weighting matrix W_{k+1} is computed from the row norms of the solution obtained in the previous iteration, columns corresponding to rows with smaller norm are likely to be de-emphasized if they are not relevant in fitting the data and vice versa.

A second interpretation of Regularized M-FOCUSS is as an iterative algorithm designed to minimize a regularized cost function given by

$$C(X) = \|AX - B\|_F^2 + \gamma J^{(p)}(X), \text{ with } \gamma = \lambda \frac{2}{|p|} \geq 0. \quad (19)$$

This can be shown by adapting the factored gradient approach shown in Appendix A to minimize this regularized cost function. We omit the derivation as it is an extension of the result for $L = 1$ [29]. An interesting consequence of this interpretation is that the tradeoff between quality of fit and sparsity made by the algorithm becomes readily evident. A larger γ emphasizes sparsity over quality of fit and vice versa.

The Regularized M-FOCUSS algorithm given by (16), similar to the $L = 1$ counterpart, can be shown to reduce the regularized cost function given by (19), indicating that the algorithm will likely converge to a local minimum. This result is summarized in the following theorem.

Theorem 1: For the Regularized M-Focuss algorithm given by (16), if $X_{k+1} \neq X_k$, then the regularized cost function $C(X)$ given by (19) decreases, i.e., $C(X_{k+1}) < C(X_k)$.

Proof: The result is shown using the concavity of the $\ell_{(p \leq 1)}$ diversity measure, and the details are in Appendix B.

Experimentally, the algorithm has always converged to a sparse solution for $p \in [0, 1]$. However, unlike the $L = 1$ case, no rigorous proof of such a property appears feasible.

E. Parameter Selection in the Regularized M-FOCUSS Algorithm

The parameters in the regularized M-FOCUSS algorithm impacting its performance are the regularization parameter λ, the parameter p, and the initial condition. These parameters play the same role as in the $L = 1$ case, and the key observations are summarized next.

The challenge in the Regularized M-FOCUSS algorithm, as in the Regularized FOCUSS [28], [29], is finding the regularization parameter λ. This parameter has to be found for every iteration of the algorithm to ensure that the algorithm does a reasonable tradeoff between finding a solution as sparse as possible and with as small an error as possible. Fortunately, the *modified l-curve method* described in [28] and [29] as a method of choosing the regularization parameter also performs well in this context. The modified l-curve method is based on the l-curve method introduced in [54] and [55] as a method for finding the parameter λ, and more details can be found in [28] and [29].

In the M-FOCUSS algorithm, the parameter p has to be chosen. The choice of p is dictated by the speed of convergence and the sparsity of the solution generated. Letting $p = 2$ gives the l_2 norm solution. Values of $p \leq 1$ give sparse solutions, but the order of convergence [56] is given by (2-p) [3], which implies that the algorithm converges more quickly for small values of p. However, for small values of p, it has been found to have a higher likelihood of getting trapped in a local minima. In practice, values of p between 0.8 and 1.0 have been found to represent a good compromise between speed of convergence and quality of the generated sparse solution.

Another parameter to be chosen is the initial condition. There is flexibility in the choice of the initial starting point in the M-FOCUSS algorithm. Often in engineering applications, good initial solutions can be postulated using domain-specific knowledge and should be used for initialization. If no good starting points are available, then the minimum Frobenius norm solution is a good initializer. It has been shown through simulation on the single measurement case in [57] that approximately the same success in identifying the generating subset is obtained from any random starting point. However, convergence to a solution was fastest when the minimum norm solution was used as the starting point. Therefore, for the case $L > 1$, the minimum Frobenius norm solution is used for initialization. If the true sparsity of the optimal solution is known or if the conditions of Lemma 1 are known to be satisfied, then should the algorithm fail to yield a solution of desired diversity r, the algorithm can be reinitialized with a random starting point and the algorithm run again. This led to 100% success in identifying the generating vector subsets for the case $L = 1$ in [25].

VI. SIMULATIONS

In this section, we detail computer simulations which were conducted to evaluate the performance of the algorithms developed in Sections IV and V. In order to evaluate the methods, the true sparse solution has to be known, and this is often hard to know in real data. An exhaustive search technique can be employed and the resulting solution used as reference. However,

this is computationally prohibitive, and furthermore, for the inferences to be reliable, this has to be carried out over several data sets. Another approach (the one used in our evaluations) is to generate synthetic data wherein the true sparse solution is known and stored as a reference for comparison with the solutions obtained by the algorithms. In addition to simplicity, an advantage of this approach is that it can facilitate exhaustive testing enabling one to draw reliable conclusions. We now describe the data-generation process.

A. Data Generation

A random $m \times n$ matrix A is created whose entries are each Gaussian random variables with mean 0 and variance 1. This matrix will generically satisfy the condition used in Lemma 1 that any m columns are linearly independent. Then, a known sparse matrix X_0 with L columns and only r rows with nonzero entries is created. The indices of the r nonzero rows are chosen randomly from a discrete uniform distribution, and the amplitudes of the row entries are chosen randomly from a standard Gaussian distribution. The MMV matrix B is computed by first forming the product $\hat{B} = AX_0$ according to

$$\hat{\mathbf{b}}^{(\ell)} = A\mathbf{x}_0^{(\ell)}, \quad \ell = 1, \ldots, L.$$

To determine the robustness of the MMV algorithms to the presence of measurement noise, the columns of the measurement vector B are then computed as

$$\mathbf{b}^{(\ell)} = \hat{\mathbf{b}}^{(\ell)} + \mathbf{n}^{(\ell)}$$

where the components of the independent noise sequence $\mathbf{n}^{(\ell)}$, $\ell = 1, \ldots, L$ are i.i.d and Gaussian with variance σ^2 determined from a specified SNR level as

$$\sigma^2 = \frac{1}{m}\|\mathbf{b}\|^2 10^{-\text{SNR}/10}. \tag{20}$$

The methods are evaluated by using the known generating sparse matrix X_0.

B. Experimental Details

Two quantities are varied in this experiment: SNR and the number of measurement vectors L. In a Monte Carlo simulation, 500 trials are run with the dimensions set to $m = 20$, $n = 30$, and the diversity to $r = 7$. In each trial, a different realization of the generating dictionary A, the solution matrix X_0, and the noise vectors are used. The performance of the M-OMP, M-ORMP, M-FOCUSS, and Regularized M-FOCUSS algorithms is evaluated based on these trials. Even though M-FOCUSS was derived assuming no noise, we test it on noisy data to get an indication of its robustness and to better understand the improvements afforded by Regularized M-FOCUSS. The M-FOCUSS and Regularized M-FOCUSS algorithms have an extra degree of freedom in the choice of the parameter p. A value of $p = 0.8$, which in general gives results closer to the best obtainable while converging more slowly, was used.

The same criterion cannot be used to terminate each of the basis selection algorithms due to the noise in the data. In the case of the MP algorithms, the number of iterations was set equal to the number of dictionary vectors r used to form each of the observations. With low noise, terminating the algorithm after r steps has been found to give good results, as in the case $L = 1$ [22]. The situation for the M-FOCUSS and the Regularized M-FOCUSS is different. After terminating the algorithms, three situations may occur; exactly r vectors (columns) from the dictionary are selected, less than r vectors are selected, or more than r vectors are selected. These vectors correspond to the selected nonzero rows of X. In the first situation, we use the r selected vectors. If less than r vectors are selected, we continue choosing vectors using M-ORMP on the residual until we have exactly r selected vectors. This situation is not likely to occur using the M-FOCUSS algorithm but is more likely to happen for the Regularized M-FOCUSS algorithm. A more frequent situation for both the M-FOCUSS and the Regularized M-FOCUSS is that more than r vectors are selected. In this case, the r rows in X which yield the largest magnitude row norms are selected.

In the M-ORMP algorithm, the solution X is obtained directly as the algorithm sequentially obtains the Least Squares (LS) solution. For the M-OMP, the M-FOCUSS, and the Regularized M-FOCUSS algorithms, a further processing step is required to obtain the solution matrix X. This is simply done using the r selected columns $\{k_1, k_2, \ldots, k_r\}$ from the dictionary A. X is obtained as

$$X = A_0^{\dagger} B, \text{ where } A_0 = [a_{k_1}, a_{k_2}, \ldots, a_{k_r}]. \tag{21}$$

C. Measurement of Algorithm Performance

The algorithms are run over a large number of trials, and their performance is measured in the following two ways.

- The number (or percentage) of generating columns from A are used in forming B in (2) or (3) that are correctly identified. This is done by tracking the number of trials in which *all* of the r columns used to generate $\hat{B}$, and only those r columns were correctly identified by the MMV algorithm. These results are plotted for different values of SNR and L. This gives us a performance comparison among the algorithms when used in a component detection problem, i.e., here we are trying to identify the sparsity pattern.
- While the actual generating components may not be detected by the algorithm, it may still provide a solution that is very close to the actual solution. We measure this by using the MSE as given in (4), where X is the solution matrix found using each of the selection algorithms (see above), and X_0 is the true sparse matrix used to generate the vectors of observations. The expectation is replaced by an average over the number of trials run.

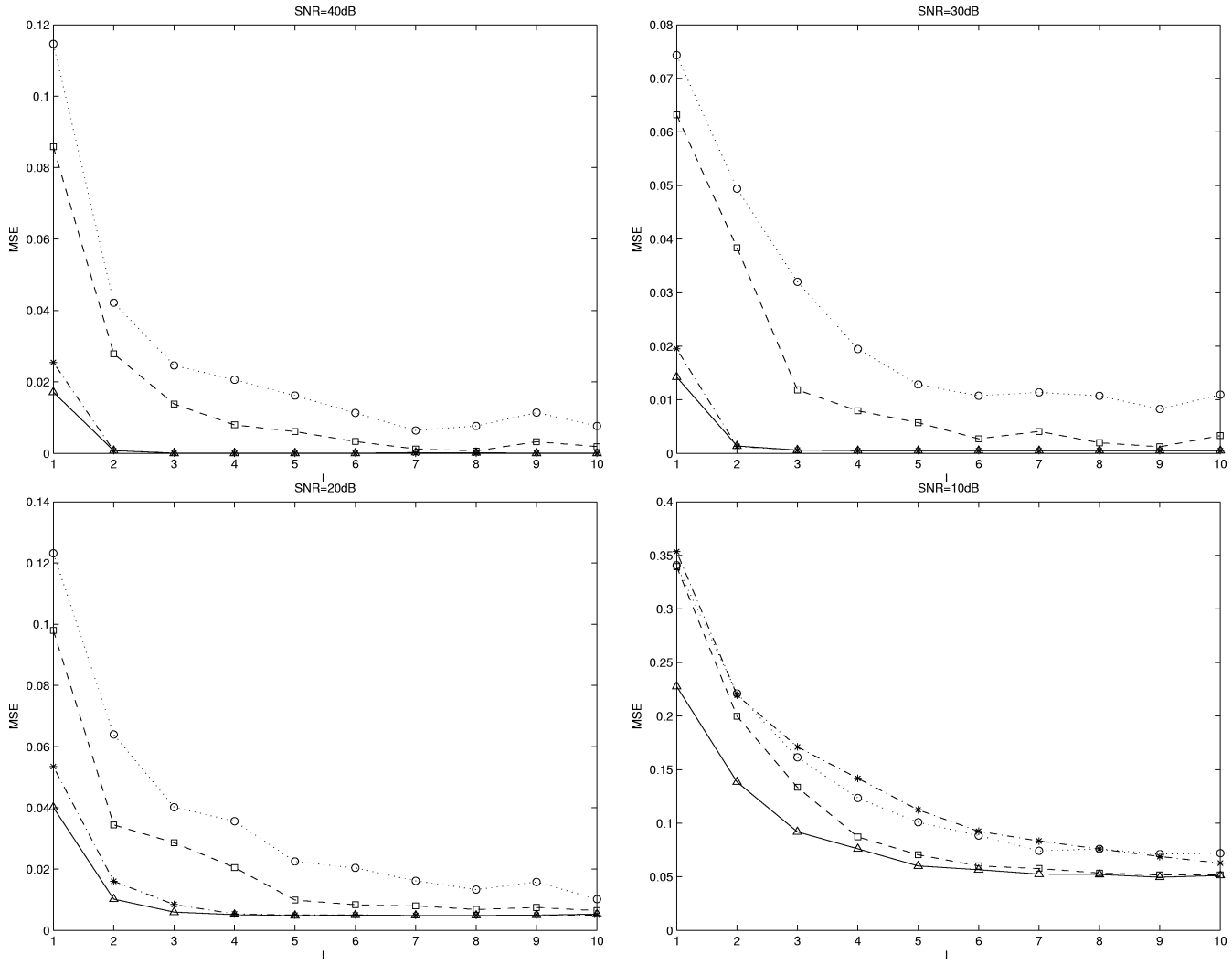

Fig. 1. $m = 20$, $n = 30$, and $r = 7$. MSE obtained using M-OMP($\circ$), M-ORMP($\square$), M-FOCUSS($*$), and Regularized M-FOCUSS($\triangle$) with $p = 0.8$ as L is varied for SNR $= 40$, 30, 20, and 10 dB.

D. Results

In Fig. 1, the SNR is held fixed in each set of trials, and the MSE is calculated for M-OMP, M-ORMP, M-FOCUSS, and Regularized M-FOCUSS as L is varied. We can also extract from Fig. 1 the MSE obtained using each algorithm with L held fixed while the SNR is varied. These results are plotted in Fig. 2 for $L = 1$, 3, and 5. In Fig. 3, we plot the percentage of trials in which *all* $r = 7$ generating vectors are identified by each algorithm in its selected set of 7 vectors with the number of observation vectors set to $L = 1$, 3, and 5. As expected, there is a strong correlation between the results of Fig. 2 and those of Fig. 3.

We first note that using a value of $L > 1$ results in an improvement in performance for all algorithms. For M-FOCUSS and Regularized M-FOCUSS, the largest performance gain, as seen from all plots, is obtained by increasing the number of observation vectors from $L = 1$ to $L = 2$. For instance, at 20 dB, this increase in L reduces the MSE by a factor of 3–4, while at 30 dB, the MSE is reduced by a factor of 10–15. Similar performance gains can be seen for the other algorithms in Fig. 1. The M-OMP algorithm performs slightly worse than the M-ORMP algorithm, as seen in [22] for $L = 1$. However, the M-ORMP algorithm requires more computation than the M-OMP algorithm.

In terms of MSE, the Regularized M-FOCUSS perform best in all the tests. For the low noise case, the difference between the M-FOCUSS and the Regularized M-FOCUSS is only present for small values of L. When $L > 3$, the M-FOCUSS and Regularized M-FOCUSS performs the same. When the SNR falls below 20 db, the Regularized M-FOCUSS performs significantly better. This is expected since the Regularized M-FOCUSS algorithm is developed for noisy data.

In the low noise case, i.e., SNR $\geq$ 30 dB, the FOCUSS algorithms gives a clear improvement over the other algorithms if MMV are available. For instance, with $L = 3$, we are able to find the generating vectors 100% of the time using FOCUSS; the MP algorithms are not able to achieve this. For SNR $\geq$ 20 dB, there is little improvement obtained in the FOCUSS solutions by using more than three measurement vectors. In contrast, we need many more measurement vectors with the MP algorithms to achieve performance equivalent to that of FOCUSS.

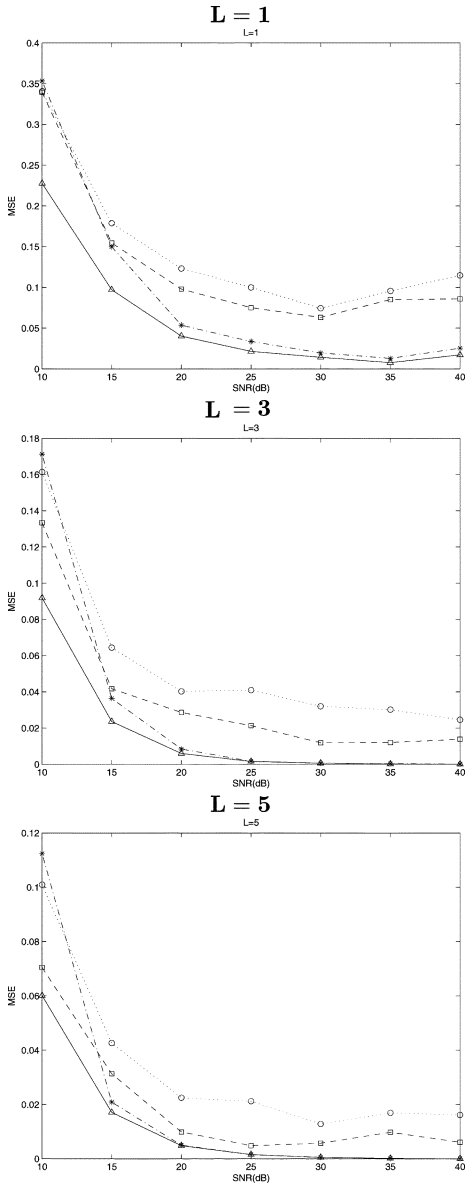

Fig. 2. $m = 20$, $n = 30$, and $r = 7$. Number of observation vectors is set to $L = 1$, 3, and 5, and the MSE obtained using M-OMP($\circ$), M-ORMP($\square$), M-FOCUSS($*$) with $p = 0.8$, and Regularized M-FOCUSS($\triangle$) with $p = 0.8$ is plotted as SNR is varied.

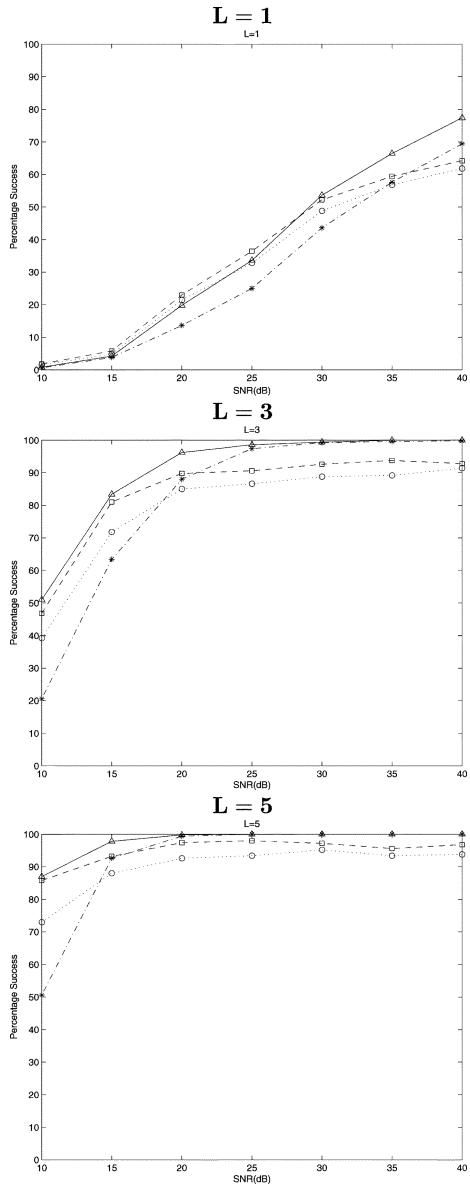

Fig. 3. $m = 20$, $n = 30$, and $r = 7$. Number of observation vectors is set to $L = 1$, 3, 5 and the percentage of trials in which we successfully obtain *all* $r = 7$ of the generating vectors in the solution set is plotted for M-OMP($\circ$), M-ORMP($\square$), M-FOCUSS($*$), and Regularized M-FOCUSS($\triangle$) as SNR is varied.

From Fig. 1, we note that the MSE curves obtained for the FOCUSS algorithms are uniformly lower than the curves obtained for the MP algorithms when SNR $\geq$ 20 dB. This can be explained by the fact that the FOCUSS algorithms for large L achieves 100% success in identifying the generating vectors. This is not true for the MP algorithms, and the cases in which the generating vectors are not identified dominate the MSE calculation. This can also be seen in Fig. 2. As the percentage success of the MP algorithms approximately levels off below 100% for a fixed L, the MSE also approximately levels off. As the SNR is increased, the successful trials result in solutions that are very close to the generating matrix X_0. However,the percentage of unsuccessful trials remains approximately the same or may increase slightly as seen for both MP algorithms, and correspondingly, the MSE also increases.

As the SNR is lowered, it is found that the generating vectors are not always those that give the lowest MSE, especially for $L = 1$. This can be seen by comparing Fig. 3 with Fig. 2. From Fig. 2, we note that the Regularized M-FOCUSS performs best in terms of MSE all the time. However, in Fig. 3 for $L = 1$, the MP algorithms perform best in terms of percentage success for noisy data.

In order to obtain a measure of the computational complexity, the cpu times for each of the method averaged over ten trials is tabulated in Table II. As can be seen, the FOCUSS class of methods are computationally more demanding.

In summary, all the algorithms developed are able to make use of and benefit from the presence of multiple measurement vectors. For the test cases considered, the methods based on

TABLE II
ALL THE CPUTIMES ARE AVERAGED OVER 10 TRIALS. THE SIMULATIONS ARE DONE IN MATLAB 6.0 ON A PENTIUM 4, 2.4-GHZ, 1-GBYTE RAM PC

method	m	n	r	L	SNR	CPUtime (sec)
m-omp	20	30	7	3	20	0.0046
m-ormp	20	30	7	3	20	0.0189
m-focuss	20	30	7	3	20	0.0219
m-regfoc	20	30	7	3	20	0.2375
m-omp	100	200	10	3	20	0.0469
m-ormp	100	200	10	3	20	0.2656
m-focuss	100	200	10	3	20	1.0876
m-regfoc	100	200	10	3	20	1.0048
m-omp	100	1000	10	3	20	0.9922
m-ormp	100	1000	10	3	20	2.2046
m-focuss	100	1000	10	3	20	6.2624
m-regfoc	100	1000	10	3	20	24.9219
m-omp	200	400	10	3	20	0.3236
m-ormp	200	400	10	3	20	1.9657
m-focuss	200	400	10	3	20	8.167
m-regfoc	200	400	10	3	20	4.8610
m-omp	200	1000	10	3	20	2.2810
m-ormp	200	1000	10	3	20	6.4375
m-focuss	200	1000	10	3	20	15.2983
m-regfoc	200	1000	10	3	20	22.3125

minimizing diversity measures, i.e., M-FOCUSS and Regularized M-FOCUSS, perform better than the forward sequential methods. This indicates the potential of the diversity minimization framework, although general claims are hard to make given the various choices of m, n, and r that are possible. From a computational complexity point of view, the forward sequential methods are superior. However, there is no significant increase in the computational complexity of M-FOCUSS with L as there is still only one pseudo-inverse computation per iteration [cf. (15)]. Developing efficient and reliable algorithms, particularly to deal with large size problems, is an interesting topic for further study.

VII. CONCLUSION

We have extended two classes of algorithms [Matching Pursuit (MP) and FOCUSS], which are used for obtaining single–measurement sparse solutions to linear inverse problems in low–noise environments — to allow for applications in which we have access to multiple measurement vectors (MMV) with a common, but unknown, sparsity structure. The convergence of these algorithms was considered. In a series of simulations on a test case dictionary and with additive Gaussian measurement noise, we showed how the performance of the different algorithms varied with SNR and the number of measurement vectors available. As expected, the availability of MMV led to a performance improvement for both the extensions to the MP and FOCUSS algorithms. The Regularized M-FOCUSS algorithm gave better performance at all SNR levels. However, at very low noise with $L \geq 1$, the less computationally expensive M-FOCUSS algorithm is preferable. Further research is required to develop a more computationally efficient version of the Regularized M-FOCUSS as well as to better analytically characterize the performance of the algorithms.

APPENDIX A
DERIVATION OF M-FOCUSS

For simplicity, we consider the noiseless case. A similar approach can be used to derive Regularized M-FOCUSS. To minimize the $J^{(p)}(X)$ diversity measure subject to the equality constraints (2), we start with the standard method of Lagrange multipliers. Define the Lagrangian $L(X, \Lambda)$ as

$$L(X,\Lambda) = J^{(p)}(X) + \sum_{l=1}^{L} \boldsymbol{\lambda}_l^T (A\mathbf{x}^{(l)} - \mathbf{b}^{(l)})$$

where $\boldsymbol{\lambda}_l$, $l = 1, \ldots, n$ are the vectors of Lagrange multipliers. A necessary condition for a minimizing solution X_* to exist is that (X_*, Λ_*) be stationary points of the Lagrangian function, i.e., for $l = 1, \ldots, L$

$$\begin{aligned} \nabla_{\mathbf{x}^{(l)}} L(X_*, \Lambda_*) &= \nabla_{\mathbf{x}^{(l)}} J^{(p)}(X_*) + A^T \boldsymbol{\lambda}_{l,*} = 0 \\ \nabla_{\boldsymbol{\lambda}_l} L(X_*, \Lambda_*) &= A\mathbf{x}_*^{(l)} - \mathbf{b}^{(l)} = 0 \end{aligned} \tag{22}$$

where $\nabla_{\mathbf{x}^{(l)}} L(X,\Lambda) = [\partial L(X,\Lambda)/\partial x[1,l], \ldots, \partial L(X,\Lambda)/\partial x[n,l]]^T$, and $\nabla_{\boldsymbol{\lambda}_l} L(X,\Lambda)$ is defined similarly. The partial derivative of the diversity measure $J^{(p)}(X)$ with respect to element $x[i,l]$ can be readily shown to be

$$\nabla_{x[i,l]} J^{(p)}(X) = |p|\ ||\mathbf{x}[i]||^{p-2}\, x[i,l].$$

For tractability purposes, as in [25], we use a *factored representation* for the gradient vector of the diversity measure

$$\nabla_{\mathbf{x}^{(l)}} J^{(p)}(X) = |p| \Pi(X) \mathbf{x}^{(l)} \tag{23}$$

where $\Pi(X) = \text{diag}(||\mathbf{x}[i]||^{p-2})$. At this point, it is useful to note that the $\Pi(X)$ matrix is independent of the column index l, which leads to considerable simplicity in the algorithm. This is a consequence of the choice $q = 2$ in (13), and other choices do not lead to such tractability. From (22) and (23), the stationary points satisfy

$$|p|\Pi(X_*)X_* + A^T\Lambda_* = 0 \text{ and } AX_* - B = 0. \tag{24}$$

From (24), carrying out some simple manipulations as in [25], it can be show that

$$X_* = \Pi^{-1}(X_*)A^T(A\Pi^{-1}(X_*)A^T)^{-1}B \tag{25}$$

which suggests the following iterative procedure for computing X_*:

$$X_{k+1} = \Pi^{-1}(X_k)A^T(A\Pi^{-1}(X_k)A^T)^{-1}B. \tag{26}$$

The computation of $\Pi^{-1}(X_k) = \text{diag}(||\mathbf{x}_k[i]||^{(2-p)})$ for $p \leq 1$ does not pose any implementation problems, even as entries converge to zero (as is desired, the goal being a *sparse* stationary point X_*).

Letting $W_{k+1}^2 = \Pi^{-1}(X_k)$ and $A_{k+1} = AW_{k+1}$ allows us to write the M-FOCUSS algorithm in (26), as given in (15). Note that the algorithm can be used for $p \in [0, 2]$, but for obtaining sparse solutions the range [0, 1] is appropriate.

APPENDIX B
DESCENT PROPERTY OF REGULARIZED M-FOCUSS

To show that (19) is a descent function for the regularized M-FOCUSS algorithm (16), we first developed some notation helpful for the proof. We define a vector

$\mathbf{c}_k = [c_k[1], c_k[2], \ldots, c_k[L]]^T$, where $c_k[i] = ||\mathbf{x}_k[i]||$ with $\mathbf{x}_k[i]$ the ith row of matrix X_k. Then, it is easy to see that

$$J^{(p)}(X_k) = \sum_{i=1}^{n} (||\mathbf{x}_k[i]||_2)^p = E^{(p)}(\mathbf{c}_k)$$

and hence

$$C(X_k) = ||AX_k - B||_F^2 + \gamma J^{(p)}(X_k) = ||AX_k - B||_F^2 + \gamma E^{(p)}(\mathbf{c}_k). \tag{27}$$

In addition, since $W_{k+1} = \text{diag}(c_k[i]^{1-p/2})$, we have from (17)

$$\begin{aligned} G_{k+1}(X_{k+1}) &= ||AX_{k+1} - B||_F^2 + \lambda ||W_{k+1}^{-1} X_{k+1}||_F^2 \\ &= ||AX_{k+1} - B||_F^2 + \lambda \mathbf{c}_{k+1}^T \Pi(\mathbf{c}_k) \mathbf{c}_{k+1} \end{aligned} \tag{28}$$

where $\Pi(\mathbf{c}_k) = W_{k+1}^{-2} = \text{diag}(|c_k[i]|^{p-2})$. Similarly, $G_{k+1}(X_k) = ||AX_k - B||_F^2 + \lambda \mathbf{c}_k^T \Pi(\mathbf{c}_k) \mathbf{c}_k$. With these preliminaries, we can establish the descent property.

Theorem 1: For the Regularized M-FOCUSS algorithm given by (16), if $X_{k+1} \neq X_k$, then the regularized cost function $C(X)$ given by (19) decreases, i.e., $C(X_{k+1}) < C(X_k)$.

Proof: From the concavity of the $\ell_{(p \leq 1)}$ diversity measure (Lemma 1 in [29] and [58]), we have

$$\begin{aligned} &E^{(p)}(\mathbf{c}_{k+1}) - E^{(p)}(\mathbf{c}_k) \\ &\quad \leq \frac{|p|}{2} \left(\mathbf{c}_{k+1}^T \Pi(\mathbf{c}_k) \mathbf{c}_{k+1} - \mathbf{c}_k^T \Pi(\mathbf{c}_k) \mathbf{c}_k \right), \; p \leq 1. \end{aligned} \tag{29}$$

Then, using (27)

$$\begin{aligned} &C(X_{k+1}) - C(X_k) \\ &\quad = \left[||AX_{k+1} - B||^2 + \gamma E^{(p)}(\mathbf{c}_{k+1}) \right] \\ &\qquad - \left[||AX_k - B||^2 + \gamma E^{(p)}(\mathbf{c}_k) \right] \\ &\quad \leq \left[||AX_{k+1} - B||^2 + \lambda \mathbf{c}_{k+1}^T \Pi(\mathbf{c}_k) \mathbf{c}_{k+1} \right] \\ &\qquad - \left[||AX_k - B||^2 + \lambda \mathbf{c}_k^T \Pi(\mathbf{c}_k) \mathbf{c}_k \right], \text{ with } \lambda = \frac{\gamma |p|}{2} \\ &\quad = G_{k+1}(X_{k+1}) - G_{k+1}(X_k) < 0. \end{aligned} \tag{30}$$

The first inequality follows from (29), the last equality from (28), and the last inequality from (18). Thus, $C(X)$ is decreased at every iteration of the algorithm as desired.

REFERENCES

[1] B. D. Rao, "Signal processing with the sparseness constraint," in *Proc. ICASSP*, vol. III, Seattle, WA, May 1998, pp. 1861–4.

[2] I. F. Gorodnitsky, J. S. George, and B. D. Rao, "Neuromagnetic source imaging with FOCUSS: A recursive weighted minimum norm algorithm," *J. Electroencephalog. Clinical Neurophysiol.*, vol. 95, no. 4, pp. 231–251, Oct. 1995.

[3] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstructions from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.

[4] H. Lee, D. P. Sullivan, and T. H. Huang, "Improvement of discrete band-limited signal extrapolation by iterative subspace modification," in *Proc. ICASSP*, vol. 3, Dallas, TX, Apr. 1987, pp. 1569–1572.

[5] S. D. Cabrera and T. W. Parks, "Extrapolation and spectral estimation with iterative weighted norm modification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 39, no. 4, pp. 842–851, Apr. 1991.

[6] B. D. Jeffs, "Sparse inverse solution methods for signal and image processing applications," in *Proc. ICASSP*, vol. III, Seattle, WA, May 1998, pp. 1885–1888.

[7] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, Apr. 1995.

[8] E. S. Cheng, S. Chen, and B. Mulgrew, "Efficient computational schemes for the orthogonal least squares learning algorithm," *IEEE Trans. Signal Process.*, vol. 43, no. 1, pp. 373–376, Jan. 1995.

[9] I. J. Fevrier, S. B. Gelfand, and M. P. Fitz, "Reduced complexity decision feedback equalization for multipath channels with large delay spreads," *IEEE Trans. Commun.*, vol. 47, no. 6, pp. 927–937, Jun. 1999.

[10] D. L. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancelers," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 5, pp. 508–518, Sep. 2000.

[11] B. Jeffs and M. Gunsay, "Restoration of blurred star field images by maximally sparse optimization," *IEEE Trans. Image Process.*, vol. 2, no. 2, pp. 202–211, Mar. 1993.

[12] J. B. Ramsey and Z. Zhang, "The application of waveform dictionaries to stock market index data," in *Predictablility of Complex Dynamical Systems*, J. Kadtke and A. Kravtsov, Eds. New York: Springer-Verlag, 1996.

[13] D. J. Field, "What is the goal of sensory coding," *Neural Comput.*, vol. 6, pp. 559–601, 1994.

[14] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, Jun. 1996.

[15] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *Constructive Approx.*, vol. 13, no. 1, pp. 57–98, 1997.

[16] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[17] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *Proc. ICASSP*, Paris, France, May 1982, pp. 614–17.

[18] S. Chen and J. Wigger, "Fast orthogonal least squares algorithm for efficient subset model selection," *IEEE Trans. Signal Process.*, vol. 43, no. 7, pp. 1713–1715, Jul. 1995.

[19] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Asilomar Conf. Signal, Syst., Comput.*, Nov. 1993, pp. 40–44.

[20] G. Davis, S. Mallat, and Z. Zhang, "Adaptive time-frequency decompositions," *Opt. Eng.*, vol. 33, no. 7, pp. 2183–2191, 1994.

[21] S. Singhal and B. S. Atal, "Amplitude optimization and pitch prediction in multipulse coders," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-37, no. 3, pp. 317–327, Mar. 1989.

[22] S. F. Cotter, J. Adler, B. D. Rao, and K. Kreutz-Delgado, "Forward sequential algorithms for best basis selection," *Proc. Inst. Elect. Eng. Vision, Image, Signal Process.*, vol. 146, no. 5, pp. 235–244, Oct. 1999.

[23] S. Chen and D. Donoho, "Basis pursuit," in *Proc. Twenty-Eighth Asilomar Conf. Signals, Syst., Comput.*, vol. I, Monterey, CA, Nov. 1994, pp. 41–44.

[24] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[25] B. D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Trans. Signal Process.*, vol. 47, no. 1, pp. 187–200, Jan. 1999.

[26] K. Kreutz-Delgado and B. D. Rao, "Sparse basis selection, ICA, and majorization : Toward a unified perspective," in *Proc. ICASSP*, Phoenix, AZ, Mar. 1999.

[27] ——, "Convex/schur-convex (CSC) log-priors and sparse coding," in *Proc. 6th Joint Symp. Neural Comput.*, Pasadena, CA, May 1999.

[28] K. Engan, "Frame Based Signal Representation and Compression," Ph.D. dissertation, Univ. Stavanger/NTNU, Stavenger, Norway, Sep. 2000, [Online] Available http://www.ux.uis.no/~kjersti.

[29] B. D. Rao, K. Engan, S. F. Cotter, J. Palmer, and K. Kreutz-Delgado, "Subset selection in noise based on diversity measure minimization," *IEEE Trans. Signal Process.*, vol. 51, no. 3, pp. 760–770, Mar. 2003.

[30] C. Couvreur and Y. Bresler, "On the optimality of the backward greedy algorithm for the subset selection problem," *SIAM J. Matrix Anal. Applicat.*, vol. 21, no. 3, pp. 797–808, 1999.

[31] G. Harikumar, C. Couvreur, and Y. Bresler, "Fast optimal and suboptimal algorithms for sparse solutions to linear inverse problems," in *Proc. ICASSP*, vol. III, Seattle, WA, May 1998, pp. 1877–80.

[32] S. Reeves, "An efficient implementation of the backward greedy algorithm for sparse signal reconstruction," *IEEE Signal Process. Lett.*, vol. 6, pp. 266–8, Oct. 1999.

[33] S. F. Cotter, K. Kreutz-Delgado, and B. D. Rao, "Backward sequential elimination for sparse vector subset selection," *Signal Process.*, vol. 81, pp. 1849–64, Sep. 2002.
[34] ——, "Efficient backward elimination algorithm for sparse signal representation using overcomplete dictionaries," *IEEE Signal Process. Lett.*, vol. 9, no. 5, pp. 145–7, May 2002.
[35] B. D. Rao and K. Kreutz-Delgado, "Basis selection in the presence of noise," in *Proc. 32st Asilomar Conf. Signals, Syst., Comput.*, Monterey, CA, Nov. 1998.
[36] ——, "Sparse solutions to linear inverse problems with multiple measurement vectors," in *Proc. IEEE Digital Signal Process. Workshop*, Bryce Canyon, UT, Aug. 1998.
[37] R. Gribonval, "Sparse decomposition of stereo signals with matching pursuit and application to blind separation of more than two sources from a stereo mixture," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2002.
[38] ——, "Piecewise linear separation," in *Proc. SPIE, Wavelets: Applications in Signal and Image Processing X*, 2003.
[39] D. Leviatan and V. N. Temlyakov, "Simultaneous Approximation by Greedy Algorithms," Univ. South Carolina, Dept. Math., Columbia, SC, Tech. Rep. 0302, 2003.
[40] J. W. Phillips, R. M. Leahy, and J. C. Mosher, "Meg-based imaging of focal neuronal current sources," *IEEE Trans. Med. Imag.*, vol. 16, no. 3, pp. 338–348, Mar. 1997.
[41] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Upper Saddle River, NJ: Prentice-Hall, 1997.
[42] S. F. Cotter and B. D. Rao, "Sparse channel estimation via matching pursuit with application to equalization," *IEEE Trans. Commun.*, vol. 50, no. 3, pp. 374–377, Mar. 2002.
[43] D. L. Donoho and H. Xiaoming, "Uncertainty principles and ideal atomic decompositions," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2845–2862, Nov. 2001.
[44] M. Elad and A. M. Bruckstein, "A generalized uncertainty principle and sparse representations in pairs of bases," *IEEE Trans. Inf. Theory*, vol. 48, no. 9, pp. 2558–2567, Sep. 2002.
[45] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3320–3325, Dec. 2003.
[46] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," *Proc. Nat. Aca. Sci.*, vol. 100, no. 5, pp. 2197–2202, 2003.
[47] J. J. Fuchs, "On Sparse Representations in Arbitrary Redundant Bases. Tech. Rep.," IRISA, Paris, France, 2002.
[48] J. Tropp, "Greed is Good : Algorithmic Results for Sparse Approximation: Tech. Rep.," Texas Inst. Comput. Eng. Sci., Univ. Texas, Austin, TX, 2003.
[49] R. Gribonval and M. Nielsen, "Highly Sparse Representations from Dictionaries are Unique and Independent of the Sparseness Measure," Aalborg Univ., Dept. Math. Sci., Aalborg, Denmark, Tech. Rep. R-2003–16, 2003.
[50] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD: John Hopkins Univ. Press, 1989.
[51] M. V. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software*. Wellesley, MA: A. K. Peters, 1994.
[52] D. Donoho, "On minimum entropy segmentation," in *Wavelets: Theory, Algorithms, and Applications*, L. Montefusco, C. K. Chui, and L. Puccio, Eds. New York: Academic, 1994, pp. 233–269.
[53] K. Kreutz-Delgado and B. D. Rao, "Measures and algorithms for best basis selection," in *Proc. ICASSP*, vol. III, Seattle, WA, May 1998, pp. 1881–1884.
[54] P. C. Hansen, "Analysis of discrete Ill-posed problems by means of the L-curve," *SIAM Rev.*, vol. 34, pp. 561–580, Dec. 1992.
[55] P. C. Hansen and D. P. O'Leary, "The use of the L-Curve in the regularization of discrete ill-posed problems," *SIAM J. Sci. Comput.*, vol. 14, pp. 1487–1503, Nov. 1993.
[56] D. Luenberger, *Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley, 1989.
[57] S. F. Cotter, "Subset Selection Algorithms with Applications," Ph.D. dissertation, Univ. Calif. San Diego, La Jolla, CA, 2001.
[58] J. Palmer and K. Kreutz-Delgado, "A globally convergent algorithm for maximum likelihood estimation in the Bayesian linear model with non-Gaussian source and noise priors," in *Proc. 36th Asilomar Conf. Signals, Syst., Comput,*, Monterey, CA, Nov. 2002.

Shane F. Cotter (M'02) was born in Tralee, Ireland, in 1973. He received the B.E. degree from University College Dublin, Dublin, Ireland, in 1994 and the M.S. and Ph.D. degrees in electrical engineering in 1998 and 2001, respectively, from the University of California at San Diego, La Jolla.

He is currently a senior design engineer with Nokia Mobile Phones, San Diego. His main research interests are statistical signal processing, signal and image representations, optimization, and speech recognition.

Bhaskar D. Rao (F'00) received the B. Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, India, in 1979 and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1981 and 1983, respectively.

Since 1983, he has been with the University of California at San Diego, La Jolla, where he is currently a Professor with the Electrical and Computer Engineering Department. His interests are in the areas of digital signal processing, estimation theory, and optimization theory, with applications to digital communications, speech signal processing, and human–computer interactions.

Dr. Rao has been a member of the Statistical Signal and Array Processing Technical Committee of the IEEE Signal Processing Society. He is currently a member of the Signal Processing Theory and Methods Technical Committee.

Kjersti Engan (M'01) was born in Bergen, Norway, in 1971. She received the Ing. (B.E.) degree in electrical engineering from Bergen University College in 1994 and the Siv.Ing. (M.S.) and Dr.Ing. (Ph.D.) degrees in 1996 and 2000, respectively, both in electrical engineering from Stavanger University College, Stavanger, Norway.

She was a visiting scholar with Professor B. Rao at the University of California at San Diego, La Jolla, from September 1998 to May 1999. She is currently an Associate Professor with the Department of Electrical and Computer Engineering, Stavanger University College. Her research interests include signal and image representation and compression, image analysis, denoising and watermarking, and, especially, medical image segmentation and classification.

Kenneth Kreutz-Delgado (SM'93) received the M.S. degree in physics and the Ph.D. degree in engineering systems science from the University of California at San Diego (UCSD), La Jolla.

He is a Professor with the department of Electrical and Computer Engineering, UCSD. Before joining the faculty at UCSD, he was a researcher at the NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, where he worked on the development of intelligent telerobotic systems for use in space exploration and satellite servicing and repair. His interest in autonomous intelligent systems that can sense, reason, and function in unstructured and nonstationary environments is the basis for research activities in the areas of adaptive and statistical signal processing; statistical learning theory and pattern recognition; nonlinear dynamics and control of articulated multibody systems; computational vision; and biological inspired robotic systems.

Dr. Kreutz-Delgado is a member of the AAAS and of the IEEE Societies on Signal Processing; Robotics and Automation; Information Theory; and Controls.