Subset Selection in Noise Based on Diversity Measure Minimization

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

***Abstract*—In this paper, we develop robust methods for subset selection based on the minimization of diversity measures. A Bayesian framework is used to account for noise in the data and a maximum *a posteriori* (MAP) estimation procedure leads to an iterative procedure which is a regularized version of the FOCal Underdetermined System Solver (FOCUSS) algorithm. The convergence of the regularized FOCUSS algorithm is established and it is shown that the stable fixed points of the algorithm are sparse.**

We investigate three different criteria for choosing the regularization parameter: quality of fit, sparsity criterion, and L-curve. The L-curve method, as applied to the problem of subset selection, is found not to be robust, and we propose a novel modified L-curve procedure that solves this problem. Each of the regularized FOCUSS algorithms is evaluated through simulation of a detection problem, and the results are compared with those obtained using a sequential forward selection algorithm termed orthogonal matching pursuit (OMP). In each case, the regularized FOCUSS algorithm is shown to be superior to the OMP in noisy environments.

***Index Terms*—Diversity measures, linear inverse problems, matching pursuit, regularization, sparsity, subset selection, undetermined systems.**

I. Introduction

SUBSET selection algorithms have received a lot of attention in recent years because of the large number of applications in which they arise [1]. The task of a subset selection algorithm can be viewed, in many instances, as that of selecting a small number of elements or vectors from a large collection of elements (termed a dictionary) that are then used to represent a signal of interest. The subset selection problem has been shown to be NP-hard and many algorithms have been proposed for finding suboptimal solutions to the problem, including algorithms based on forward sequential search or elimination of elements from the full dictionary available [2]. In previous work [3]–[5], an iterative algorithm termed **FOC**al **U**nderdetermined **S**ystem **S**olver (FOCUSS) has been developed based on the minimization of diversity measures. This algorithm essentially removes elements from the dictionary in parallel and has been shown to outperform other subset selection algorithms in low noise environments.

The goal of this paper, which expands on work presented in [6] and [7], is to extend the FOCUSS algorithm so that it can be used in subset selection problems where the signal-to-noise ratio (SNR) is low. A formal methodology is developed for deriving algorithms that can deal with noise in the data. It is shown how a Bayesian framework coupled with priors on the solution components consistent with the $\ell_{(p\leq 1)}$ diversity measure leads to a regularized version of the FOCUSS algorithm. The convergence of the regularized FOCUSS algorithm is established, and it is shown that the stable fixed points of the algorithm are sparse.

In practice, some method must be used in choosing the magnitude of the regularization parameter. Motivated by applications, we explore three different ways of setting this parameter. First, we consider a discrepancy criterion that assures a certain quality of fit in the representation as is typically required in signal representation problems [1]. Next, we consider limiting the size of the selected subset, which is important in compression; we term this a sparsity criterion since the representation obtained uses a small number of vectors from the available dictionary. Finally, we experiment with an L-curve criterion, which seeks to trade off the representation error and the size of the selection subset [8], [9]. This criterion is applicable to the problem of dictionary/frame learning as considered in [10] and [11]. However, as applied to the problem of subset selection, we find that the L-curve method did not provide robust solutions. This leads us to develop a novel modified L-curve procedure to determine the regularization parameter that incorporates a target SNR. This results in a robust procedure for implementing the regularized FOCUSS algorithm when compared with the L-curve method of [8], [9]. A detection problem is used to examine the implementations of these regularized FOCUSS algorithms. The results obtained using each of the regularized FOCUSS algorithms are compared to the results of an improved sequential forward selection algorithm termed orthogonal matching pursuit (OMP) [12], [13]. We conclude that the regularized FOCUSS procedures give much better results than OMP in detecting the correct subset in noisy environments.

The outline of this paper is as follows. In Section II, we outline the subset selection problem. We give some examples of diversity measures and show how minimization of these measures can be used to provide solutions to the subset selection problem. In Section III, we use a Bayesian framework to account for noise in the measured data, and the MAP procedure is used to produce an iterative algorithm to provide solutions to

Manuscript received March 12, 2002; revised October 10, 2002. This work was supported in part, by the National Science Foundation under Grant CCR-9902961. The associate editor coordinating the review of this paper and approving it for publication was Dr. Athina Petropulu.

B. D. Rao, S. F. Cotter, J. Palmer, 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: brao@ece.ucsd.edu; scotter@ece.ucsd.edu; japalmer@ece.ucsd.edu; kreutz@ece.ucsd.edu).

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

Digital Object Identifier 10.1109/TSP.2002.808076

the subset selection problem. This procedure can be viewed as a regularized FOCUSS algorithm. The convergence of this algorithm is established, and it is shown that the stable fixed points of the algorithm are sparse in Section IV. Practical methods for the choice of the regularization parameter are considered in Section V, and a modified L-curve criterion is introduced. The different choices of regularization parameter are examined through simulation of a detection problem in Section VI, and the results obtained using OMP are also included for comparison. We draw some conclusions in Section VII.

II. Minimizing Diversity Measures

The subset selection problem can be written in matrix form and consists of solving an *underdetermined* linear system of equations of the form [1]

$$Ax = b \tag{1}$$

where A is an $m \times n$ matrix with $m \leq n$ (and, usually, $m \ll n$), and rank$(A) = m$. The columns of A are formed from the elements of the dictionary in signal representation problems or derived from the physics of the problem in linear inverse problems [1], [2]. There are *many* solutions to the system of equations in (1) and the subset selection problem corresponds to identifying a few columns of the matrix A, which can be used to represent the data vector b [1], [2], [14]. This corresponds to finding a solution x with few nonzero entries that satisfies (1), and such a solution is said to be sparse.

Finding an optimal solution to this problem generally requires a combinatorial search that is computationally unattractive. Therefore, suboptimal techniques are usually employed [1], [2]. We discuss one such method called FOCUSS, which has been extensively examined in [4] and [5]. The FOCUSS method was motivated by the observation that if a sparse solution is desired then choosing a solution based on the smallest l_2-norm is not appropriate. The minimum l_2-norm criterion favors solutions with many small nonzero entries, which is a property that is contrary to the goal of sparsity [4], [15]. Consequently, there is a need to consider the minimization of alternative measures that promote sparsity. In this context, of particular interest are diversity measures that are functionals that measure the lack of concentration/sparsity and algorithms for minimizing these measures to obtain sparse solutions. There are many measures of diversity [16], [17], but a set of diversity measures that has been found to produce very good results as applied to the subset selection problem is the $\ell_{(p\leq 1)}$ diversity measure given by [5], [18]

$$E^{(p)}(x) = \text{sgn}(p) \sum_{i=1}^{n} |x[i]|^p, \quad p \leq 1. \tag{2}$$

Minimization of this diversity measure leads to the FOCUSS algorithm [4], [5]. The algorithm is iterative and produces intermediate approximate solutions according to

$$x_{k+1} = W_{k+1} \left(AW_{k+1}\right)^{\dagger} b \tag{3}$$

where $W_{k+1} = \text{diag}(|x_k[i]|^{1-(p/2)})$, and $\dagger$ is used to denote the Moore–Penrose pseudoinverse [19]. The properties of this algorithm have been examined in depth in [4], [5], and [18]. Intuitively, the algorithm can be explained by noting that there is competition between the columns of A to represent b. In each iteration, certain columns get emphasized while others are de-emphasized. In the end, a few columns survive to represent b, providing a sparse solution.

Interesting insight can be gained into (3) when it is viewed as a sequence of weighted minimum l_2-norm problems [1]. Defining $q \triangleq W_{k+1}^{-1} x$, in each iteration of the FOCUSS algorithm, the solution x_{k+1} is computed as $x_{k+1} = W_{k+1} q_{k+1}$, where

$$q_{k+1} = \arg\min_{q} \|q\|^2 \text{ subject to } AW_{k+1}q = b. \tag{4}$$

Therefore, the FOCUSS iteration is obtained as the minimum norm solution to an underdetermined set of linear constraints. Imposing the equality constraint in (4) is equivalent to assuming the absence of noise. As we will see in Sections III–VI, accounting for noise means that an exact minimum norm solution of the form (4) is not sought, and instead, we find a solution at each iteration step that minimizes $\|q\|^2$ and *approximately* satisfies the set of constraints.

III. Subset Selection in Noisy Environments

The derivation of FOCUSS in [4], [5] was based on the assumption that there was no noise in the data, i.e., the data vector b in (1) is formed as an *exact* linear combination of a few columns from A. Later, reasonable modifications to the algorithm were suggested to deal heuristically with noise [1], [4]. Here, we take a formal approach and extend the FOCUSS method to deal with noise in the measurements using a Bayesian framework. This stochastic framework provides theoretical insights and assists in developing robust methods.

A. Bayesian Formulation

For this discussion, we assume that each of the measured data vectors b consists of a linear combination of a small number of columns from A together with additive noise v:

$$b = b' + v = Ax + v. \tag{5}$$

It is assumed in this formulation that x is a random vector that is sparse and independent of v. Under these assumptions, a maximum *a posteriori* (MAP) estimate of x can be obtained as

$$\begin{aligned} x_{\text{MAP}} &= \arg\max_{x} \ln p(x|b) \\ &= \arg\max_{x} \left[\ln p(b|x) + \ln p(x)\right] \\ &= \arg\max_{x} \left[\ln p_v(b - Ax) + \ln p(x)\right]. \end{aligned}$$

This formulation is general and offers considerable flexibility. In order to proceed further, however, some assumptions must be made on the distributions of the noise components in v and the components of the solution vector x.

B. Generalized Gaussian Priors

Because, here, we are interested in a sparse x, the distribution of v is not very critical to the approach except for analytical and computational tractability. We assume that v is a Gaussian

random vector with independent identically distributed (i.i.d.) elements,[1] i.e., each component $v[i], i = 1, \ldots, m$ is distributed as $p_{v[i]}(u) = c_1 e^{-(u^2/2\sigma^2)}$, where $c_1 = 1/\sqrt{2\pi}\sigma$, and σ^2 is the noise variance. The distribution of x is important for the generation of sparse solutions. Probability density functions (pdf's) that are concentrated near zero but also have heavy tails are appropriate for this purpose [5], [17]. The elements $x[i]$ are assumed to be i.i.d. random variables with a generalized Gaussian distribution. The pdf of the generalized Gaussian distribution family is defined as [20], [21]

$$f(u;p,\beta) = \frac{p}{2\sqrt[p]{2}\beta\Gamma\left(\frac{1}{p}\right)} e^{-(|u|^p/2\beta^p)}, \quad p > 0 \tag{6}$$

where $\Gamma(\cdot)$ is the standard gamma function. The factor p controls the shape, and β is a generalized variance. For instance, setting $p = 1$ reduces this generalized form to that of a Laplacian distribution that has been assumed as the prior distribution of x in [15] and [22]. If we set $p = 2$ and $\beta = 1$, this distribution reduces to the standard normal distribution. If a unit variance distribution is desired, i.e., $\sigma^2 = 1$, then β becomes a function of p as given by

$$\beta^2 = 2^{-(2/p)} \frac{\Gamma\left(\frac{1}{p}\right)}{\Gamma\left(\frac{3}{p}\right)}. \tag{7}$$

Therefore, only one parameter characterizes the distribution, and Fig. 1 plots the pdf for different values of p when $\sigma^2 = 1$. From the figure, it can be seen that the pdf moves toward a uniform distribution as $p \to \infty$ and toward a very peaky distribution as $p \to 0$.

A vector $x \in \mathcal{R}^n$ with elements that are distributed as generalized Gaussian and are independent has the following pdf:

$$\begin{aligned} p_x(x) = & p_x(x[1], \ldots, x[n]) = \left(\frac{p}{2\sqrt[p]{2}\beta\Gamma\left(\frac{1}{p}\right)} \right)^n \\ & \times \exp\left(-\frac{1}{2\beta^p} \mathrm{sgn}(p) \sum_{i=1}^{n} |x[i]|^p \right) \end{aligned} \tag{8}$$

where, for consistency with the $l_{(p\leq 1)}$ diversity measure, $\mathrm{sgn}(p)$ is added to allow for $p < 0$.

C. Algorithm Development Based on Gradient Factorization

With the densities of the noise v and the solution x chosen as in the previous section, we can now proceed to find the MAP estimate that is found from

$$\begin{aligned} x_{\mathrm{MAP}} = & \arg\min_x J(x) \\ \text{where } J(x) = & \left[||Ax - b||^2 + \gamma E^{(p)}(x) \right] \\ \text{with} \quad \gamma = & \frac{\sigma^2}{\beta^p} \end{aligned} \tag{9}$$

and $E^{(p)}(x)$ as defined in (2). We note that the substitution of $p = 2$, which is consistent with a Gaussian distribution of the components in x, gives rise to the standard regularized least squares problem. With $p \leq 1$, it will be shown in Section IV that the minima of $J(x)$ are sparse. γ controls the tradeoff between quality of fit $||Ax - b||$ and the degree of sparsity. Large values of γ lead to sparser solutions, and small values lead to better fit and, hence, lower error $||Ax - b||$.

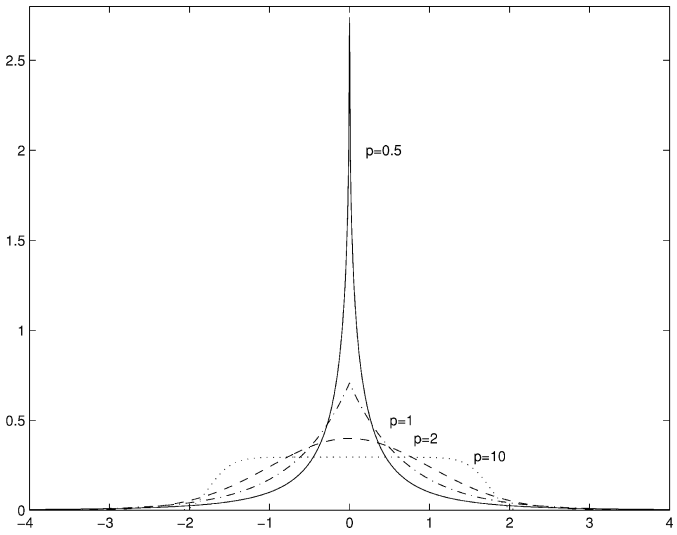

Fig. 1. Pdf of the generalized Gaussian distribution ($\sigma^2 = 1$) for different values of p: $p = 10(\cdots)$, $p = 2(--)$ (standard normal distribution), $p = 1(- \cdot -)$, $p = 0.5$ (solid).

Using the factored gradient approach developed in [5], an iterative algorithm can be derived to minimize $J(x)$. A necessary condition for the optimum solution x_* is that it satisfies

$$\nabla_x J(x_*) = 2A^T A x_* - 2A^T b + 2\lambda \Pi(x_*) x_* = 0 \tag{10}$$

where

$$\lambda = \frac{|p|}{2}\gamma = \frac{|p|}{2}\frac{\sigma^2}{\beta^p}$$

and $\Pi(x) = \mathrm{diag}(|x[i]|^{p-2})$. For convenience, we define the scaling matrix $W(x) = \mathrm{diag}(|x[i]|^{1-(p/2)})$. Substituting $\Pi(x) = W^{-2}(x)$ in (10) and performing some simple manipulations, we are left with

$$\left((AW(x_*))^T (AW(x_*) + \lambda I) \right) W^{-1}(x_*) x_* = (AW(x_*))^T b. \tag{11}$$

Hence, the optimum solution satisfies

$$x_* = W(x_*) \left((AW(x_*))^T (AW(x_*) + \lambda I \right)^{-1} (AW(x_*))^T b. \tag{12}$$

This suggests the following iterative relaxation algorithm:[2]

$$x_{k+1} = W_{k+1} \left(A_{k+1}^T A_{k+1} + \lambda I \right)^{-1} A_{k+1}^T b \tag{13}$$

where $A_{k+1} = AW_{k+1}$ with $W_{k+1} = \mathrm{diag}(|x_k[i]|^{1-(p/2)})$ and $\lambda = (|p|/2)(\sigma^2/\beta^p)$. Using the fact that

$$A_{k+1}^T \left(A_{k+1} A_{k+1}^T + \lambda I \right)^{-1} = \left(A_{k+1}^T A_{k+1} + \lambda I \right)^{-1} A_{k+1}^T$$

[1] More general Gaussian distributions can be also easily dealt with.

[2] When the elements of A and b are complex, the transpose operation has to be replaced by the Hermitian transpose

algorithm (13) can be expressed as

$$x_{k+1} = W_{k+1}A_{k+1}^T \left(A_{k+1}A_{k+1}^T + \lambda I\right)^{-1} b. \qquad (14)$$

When the noise level is reduced, i.e., $\sigma \to 0$, this implies that $\lambda \to 0$, and the algorithm reduces to the original FOCUSS algorithm given in (3). Note that the algorithm (14) provides a solution to the problem (9), which is well-posed for the underdetermined case as well as the overdetermined case. Therefore, even though we concentrate on the underdetermined case in this paper, the algorithm is also useful in the overdetermined context.

D. Interpretation as Regularized FOCUSS

The algorithm given in (14) has an interesting interpretation as Tikhonov regularization [23] applied to (4). This can be readily seen by rewriting (14) as $x_{k+1} = W_{k+1}q_{k+1}$, where q_{k+1} is obtained as

$$q_{k+1} = \arg\min_q \|AW_{k+1}q - b\|^2 + \lambda\|q\|^2. \qquad (15)$$

Alternately and equivalently, x_{k+1} can be shown to be the solution to the following optimization problem:

$$\begin{aligned} x_{k+1} &= \arg\min_x Q_{k+1}(x), \text{ where} \\ Q_{k+1}(x) &= \|Ax - b\|^2 + \lambda \left\|W_{k+1}^{-1}x\right\|^2. \end{aligned} \qquad (16)$$

By the uniqueness of the minima of $Q_{k+1}(x)$, we have $Q_{k+1}(x_{k+1}) < Q_{k+1}(x_k)$ for $x_{k+1} \neq x_k$.

Interestingly, this results in algorithm identical to that suggested in [4]. In [4], this algorithm was proposed as a method of making the 2-norm minimization problem of (4) more robust to noise. The derivation given here provides formal support to this approach.

IV. Convergence Results

Throughout this discussion, a sparse solution refers to a basic or degenerate basic solution, i.e., a solution with less than or equal to m nonzero entries where m is the dimension of the data vector b. Another assumption we make is that any m columns of A are linearly independent. We present two key results in this section. First, we show that the local minima of the regularized cost function $J(x)$ [(9), cf. Section III-C] are sparse. This justifies minimization of the regularized cost function to achieve sparsity. Second, we show that the regularized FOCUSS algorithm does indeed reduce $J(x)$ at each step and that the stable fixed points of the algorithm are sparse.

Theorem 1: If x^* is a local minima of $J(x)$, where $J(x)$ is the regularized cost function $J(x) = \left[\|Ax - b\|^2 + \gamma E^{(p)}(x)\right]$ with $E^{(p)}(x) = \text{sgn}(p)\sum_{i=1}^n |x[i]|^p$, $p \leq 1$, and $\gamma \geq 0$, then x^* is sparse.

Proof: Let $Ax^* - b = e^*$ or $Ax^* = b + e^*$. Since x^* is a local minima of $J(x)$, it is also a local minima to the optimization problem

$$\begin{aligned} \min_x E^{(p)}(x) &= \text{sgn}(p)\sum_{i=1}^n |x[i]|^p, \quad p \leq 1 \\ \text{subject to} \quad Ax &= e^* + b. \end{aligned}$$

It has been shown in [17] and [18] that the local minima of the above optimization problem are necessarily sparse. Hence, the local minima of $J(x)$ are sparse. □

The above proof and theorem also indicate how, in general, inclusion of proper diversity measures as a regularizing component can facilitate sparsity. Now, we show that the regularized FOCUSS algorithm does indeed achieve the desired goal by showing that $J(x)$ is a descent function for the algorithm. Before we prove that, we need a preparatory result that helps connect $J(x)$ to the quadratic cost function being minimized at each iteration. This is presented next in Lemma 1, which specializes more general results to be found in [24] and [25].

Lemma 1:

$$\begin{aligned} E^{(p)}(x_2) - E^{(p)}(x_1) \leq & \frac{|p|}{2}\left(x_2^T\Pi(x_1)x_2 - x_1^T\Pi(x_1)x_1\right) \\ & p \leq 1 \end{aligned} \qquad (17)$$

where $\Pi(x) = W^{-2}(x) = \text{diag}(|x[k]|^{p-2})$.

Proof: Consider the scalar function $f(y) = \text{sgn}(p)|y|^{p/2}$, $y \geq 0$, and $p \leq 1$. Since it is concave [26]

$$f(y_2) - f(y_1) \leq \nabla f(y_1)(y_2 - y_1).$$

Hence

$$\text{sgn}(p)|y_2|^{p/2} - \text{sgn}(p)|y_1|^{p/2} \leq \frac{|p|}{2}|y_1|^{(p/2)-2}y_1(y_2 - y_1).$$

Substituting $y_2 = z^2$ and $y_1 = w^2$, we have

$$\text{sgn}(p)|z|^p - \text{sgn}(p)|w|^p \leq \frac{|p|}{2}|w|^{p-2}(z^2 - w^2).$$

The above inequality applies to each of the components of $E^{(p)}(x)$ leading to

$$\begin{aligned} E^{(p)}(x_2) - E^{(p)}(x_1) &\leq \sum_{l=1}^n \frac{|p|}{2}|x_1[l]|^{p-2}(|x_2[l]|^2 - |x_1[l]|^2) \\ &= \frac{|p|}{2}\left(x_2^T\Pi(x_1)x_2 - x_1^T\Pi(x_1)x_1\right). \quad \square \end{aligned}$$

Now, we present the main convergence result.

Theorem 2: The regularized cost function $J(x) = \left[\|Ax - b\|^2 + \gamma E^{(p)}(x)\right]$, $p \leq 1$, with $\gamma = \sigma^2/\beta^p$ ((9), cf. Section III-C) is a descent function for the regularized FOCUSS algorithm (cf. (14)). Furthermore, the stable fixed points of the algorithm are sparse.

Proof: To show that $J(x)$ is a descent function for the regularized FOCUSS algorithm, we need to show that $J(x_{k+1}) < J(x_k)$, for x_{k+1} computed using (14) and $x_{k+1} \neq x_k$

$$\begin{aligned} J(x_{k+1}) - J(x_k) &= \left[\|Ax_{k+1} - b\|^2 + \gamma E^{(p)}(x_{k+1})\right] \\ &\quad - \left[\|Ax_k - b\|^2 + \gamma E^{(p)}(x_k)\right] \\ &\leq \left[\|Ax_{k+1} - b\|^2 + \lambda x_{k+1}^T\Pi(x_k)x_{k+1}\right] \\ &\quad - \left[\|Ax_k - b\|^2 + \lambda x_k^T\Pi(x_k)x_k\right] \\ \text{with} \quad \lambda &= \frac{\gamma|p|}{2} \\ &= Q_{k+1}(x_{k+1}) - Q_{k+1}(x_k) < 0. \end{aligned} \qquad (18)$$

The first inequality follows from Lemma 1 and the last inequality from (16). Thus, $J(x)$ is decreased at every iteration of the algorithm as desired.

Let x^* be a fixed point of the algorithm and, therefore, necessarily a solution of (12). If x^* is not sparse, then from Theorem 1, it is not a local minima of $J(x)$. Using the notation of Theorem 1, it is therefore not a local minima of $E^{(p)}(x)$ subject to $Ax = b + e^*$. By the concavity property of $E^{(p)}(x)$ on the positive orthant [17], [18], it can be shown that there are points arbitrarily close to x^* that can reduce $J(x)$ [5]. Hence, nonsparse x^* are not stable fixed points, and only sparse solutions can be stable fixed points.

Note that taking $\lambda \rightarrow 0$ provides an alternate proof of the convergence of the unregularized FOCUSS algorithm. In addition, note the key role that Lemma 1 has in proving the descent aspect of the algorithm. In particular, the RHS of (17) (and the RHS of its consequent (18)) shows why the FOCUSS algorithm formulated as a sequence of two-norm optimization problems is capable of minimizing the more complex objective function [cf. the noiseless optimization case of (4) or the noisy case of (15) and (16)]. More general results on related FOCUSS-like algorithms and their convergence can be found in [24] and [25]. The rate of convergence, the number of local minima, and their basins of attraction is a complex function of p. Some discussion of these issues in the noiseless case can be found in [4].

V. Methods for Choosing the Regularization Parameter λ

The sparse solution obtained via the regularized version of FOCUSS is governed by the choice of λ, and there remains the implementation-level problem of determining a proper value for λ. In addition, there appears to be no practical reason to limit the choice of λ to a fixed value for all the iterations; therefore, a value that is dependent on the iteration may be more appropriate. With this in mind, we suggest three approaches motivated by three different scenarios. In the first approach, we ensure a certain quality of fit in the signal representation. This may potentially be motivated by the availability of some information on the perturbations. The second approach ensures a certain degree of sparsity in the solution, as would be required in applications like compression. Finally, in the third approach, we seek stable sparse solutions without the need for much prior information, and a tradeoff is made between the sparsity of the solution and the representation error. Note that once the columns to be used are identified, then finding a least square solution to the resulting problem can avoid any regularization bias. This is the approach used in the simulations. A drawback is the penalty that is incurred when the wrong columns are chosen.

A. Quality-of-Fit Criterion/Discrepancy Principle

A potentially useful approach is to seek a sparse solution that ensures a certain quality in the nature of the representation, i.e., $||Ax - b|| \leq \epsilon$. For instance, in a signal representation problem, this a very commonly used criterion [1]. This is termed the *discrepancy principle* in [8]. Algorithmically, this reduces to solving the optimization problem

$$\min_x E^{(p)}(x) \text{ subject to } ||Ax - b|| \leq \epsilon.$$

Assuming that the inequality constraint is active, which is usually true, and following the approach used in deriving the regularized solution, an iterative algorithm can be derived that at each iteration computes $x_{k+1} = W_{k+1} q_{k+1}$, where

$$q_{k+1} = \arg\min_q ||q||^2 \text{ subject to } ||Aq_k - b|| \leq \epsilon.$$

This is in the form of a standard regularization problem, and the solution is given by [19]

$$q_{k+1} = \sum_{i=1}^{\rho} \left(\frac{\sigma_i u_i^T b}{\sigma_i^2 + \lambda_k} \right) v_i$$

where ρ is the rank of the matrix A_k. σ_i, $i = 1, \ldots, \rho$ are the dominant ρ singular values of A_k, and u_i, v_i are the left and right singular values, respectively. $\lambda \geq 0$ is the regularization parameter that satisfies the equation [19]

$$\sum_{i=1}^{\rho} \left(\frac{\lambda u_i^T b}{\sigma_i^2 + \lambda} \right)^2 = \epsilon^2.$$

B. Sparsity Criterion

In some applications, we may have prior knowledge of the number of vectors from A that were used to produce the data vector. As an example, in a compression algorithm, the number of vectors used in the representation of a data vector would be fixed [10], [11]. Therefore, another option is to choose λ so that the solution produced has a predetermined number of nonzero entries r. Note that upon convergence, the rank of AW_{k+1} is equal to r, i.e., $\lim_{k \rightarrow \infty} \text{rank}(AW_{k+1}) = r$. Therefore, a desirable approach would be to use a sequence λ_k that satisfies this limiting rank property while providing the best possible fit. Unfortunately, a reliable procedure for doing this is not yet available. However, one practical approach is to use a sequential basis selection method like the OMP to first select r columns from A [12]. Then, based on the representation obtained using these columns, a value for the error ϵ can be obtained, and this value can be used in running the FOCUSS algorithm in the manner suggested in Section V-A. If the FOCUSS procedure returns more columns than desired, one can prune the selected subset using OMP or a backward elimination procedure [27]. At this stage, we can choose to proceed using either the OMP or FOCUSS generated solution, depending on which is better.

C. Modified L-Curve Method

A final possibility is that the number of dictionary vectors used in forming the data vectors is variable, and some variability in the representation error must also be allowed. Therefore, the sparse nature of the solution must be controlled so that a tradeoff between quality of fit and sparsity is made. In particular, this formulation of the problem is applicable to dictionary/frame learning [10]. From our development above, this translates to

finding a regularization parameter that makes a compromise between minimizing the norm $||q||^2$ and the error in the representation $||AW_{k+1}q - b||^2$. The use of such an approach was first suggested in [4]. The L-curve was introduced in [8] and [9] as a method for finding the parameter λ in the regularization problem

$$\min_x \{||Ax - b||^2 + \lambda||x||^2\}. \tag{19}$$

The regularization problem encountered in (15) can easily be translated to this form.

As λ is increased, one obtains regularized solutions $\{q_\lambda\}$ whose norms vary continuously and decrease monotonically. If λ is varied from 0 to ∞, $||q||^2$ decreases monotonically from $||(AW_{k+1})^\dagger b||^2$ to zero, and $||AW_{k+1}q - b||^2$ increases monotonically. The theory of the L-curve proposes that a plot of $||q||^2$ versus $||AW_{k+1}q - b||^2$ for different values of λ be shaped like an L and that a good choice of value for λ is the one corresponding to the corner in the L. Furthermore, it is suggested that the corner of the L-shaped curve can be found by finding the maximum curvature [8], [9], [28]. The plot of $||q||^2$ versus $||AW_{k+1}q - b||^2$ can be shown to be convex [9], and the point of maximum curvature represents a tradeoff point between sparsity and accuracy. The curvature can be computed by means of the formula

$$K(\lambda) = \frac{X'(\lambda)Y''(\lambda) - X''(\lambda)Y'(\lambda)}{\{[X'(\lambda)]^2 + [Y'(\lambda)]^2\}^{3/2}} \tag{20}$$

where, in our problem, $X(\lambda) = ||AW_{k+1}q - b||^2$, $Y(\lambda) = ||q||^2$, and $'$ and $''$ denote the first and second derivatives respectively. Alternatively, as in [9] and [28], the curvature computation may be done in the log–log scale, that is, $X(\lambda) = \log\{||AW_{k+1}q - b||^2\}, Y(\lambda) = \log\{||q||^2\}$. The argument made for the adoption of this scale in [9] is that the corner is found to be more distinct in the log–log scale. However, a problem pointed out in [29] is that the L-curve in the log–log scale is, in general, no longer convex. In [30], a linear scale L-curve is used, and in [31], both linear and log–log scale L-curves are mentioned. In fact, experiments have shown that the log–log curve often has several corners, and finding the maximum curvature in this scale does not necessarily correspond to a λ with a good tradeoff between sparsity and accuracy.

We implemented our procedure in both the linear and the log–log scales and found that for this application, the log–log scale does not give good results. In fact, the algorithm ended up emphasizing the quality of fit at the expense of the sparsity of the solution. L-curve experiments using the linear scale showed that the regularized FOCUSS algorithm can perform better in noise than greedy algorithms such as the OMP, but it failed completely for some data vectors. The variance of the error was found to be large, which indicates that the procedure is not very robust. Further exploration of the results showed that the L-curve approach failed because the data does *not* produce an L-curve in each iteration of the FOCUSS algorithm.

This led us to develop a novel solution to the regularization problem that uses a combination of the discrepancy principle and the linear scale L-curve method. We call this the *modified L-curve method*. When using the basic L-curve method to determine λ, there is neither direct control of how many vectors are selected (i.e., the sparsity of the solution) nor a limit on the representation error. The L-curve method seeks the value of λ that best minimizes *both* these terms, i.e., it finds the best trade off between accuracy and sparsity. In our proposed modified L-curve method, we assume that we have some knowledge of the variance of the noise or, alternatively, an approximate target SNR that a representation must satisfy. From this knowledge an upper and a lower target can be set on the residual norm $\epsilon^2 = ||Ax - b||^2$. Then, for every iteration in FOCUSS, the upper and lower targets for ϵ^2 are used to find upper and lower bounds on the value of λ that are denoted by $\lambda_{\max}$ and $\lambda_{\min}$, respectively. The L-curve parameter λ corresponding to the maximum curvature in the linear scale λ_c is also calculated in every FOCUSS iteration. λ_c is then compared with the limits established, and if $\lambda_c < \lambda_{\min}$, then $\lambda_{\min}$ is used, and if $\lambda_c > \lambda_{\max}$, then $\lambda_{\max}$ is used. Otherwise, the calculated value of λ_c may be used. *This adjustment of the value of λ ensures that λ will always produce an acceptable representation even if there is no distinct corner in the L-curve.*

VI. Experiments and Results

We now conduct a series of simulations where we examine the different methods of choosing the regularization parameter that we have outlined in Section V.

The matrix A is generated as a 20×30 matrix. The entries are first chosen randomly from a standard normal distribution, and then, each column is normalized to give the matrix A. Each vector b', as given in (5), is obtained as a linear combination of r vectors from the matrix A, and the vectors are randomly selected and are equiprobable. In our experiments, r is set to 7, and the coefficients associated with these vectors are drawn from a standard normal distribution. The vector b' is then normalized, and finally, a noise vector n is added to b' to produce the final data vector b. The noise vector is generated from a Gaussian distribution with zero mean and variance determined by the SNR of the experiment. Two values of SNR (10 and 20 dB) are used in the experiments. Each experiment is carried out using 100 different data vectors.

Two error measures were utilized in evaluating the success of the different algorithms. The first error measure compares the representation obtained using the algorithm, which is given by $\hat{b} = Ax_{alg}$, to the data vector b and is denoted by

$$\epsilon^2 = ||Ax_{alg} - b||^2 = ||\hat{b} - b||^2.$$

This measures the representation error and is the most important measure when we are concerned with representing the data vector without trying to denoise the signal (as is the case in the compression of data signals). However, in the case of interest here, we are trying to get to the underlying (denoised) signal; therefore, an error measure that compares $\hat{b}$ to the underlying information signal in the data vector b' (i.e., the signal uncorrupted by noise) is more informative. The error measure we consider is

$$\epsilon_1^2 = ||\hat{b} - b'||^2.$$

Of course, in practice, this measure is not readily computable, but in the artificial simulations of Section VI-A, we know the

TABLE I
$r = 7$: RESULTS OBTAINED USING THE DISCREPANCY PRINCIPLE IN THE FOCUSS (F) ALGORITHM AND THE RESULTS FROM THE OMP (O) ALGORITHM

C	p	#	$\#r$		mean ϵ_1^2 (10^{-3})		mean ϵ^2 (10^{-3})		ϵ_1^2 %		ϵ^2 %	
			F	O	F	O	F	O	F	O	F	O
SNR=20dB												
0.8	0	9.3	5.86	5.58	8.0	12.3	4.4	7.4	57	29	24	62
1.2	0.5	6.7	5.44	5.37	9.6	16.9	9.8	14.4	30	33	25	38
1.5	0.8	7.9	5.84	5.58	8.0	12.5	6.3	9.4	45	41	23	63
SNR=10dB												
0.8	0	7.0	4.21	4.10	78.2	84.0	49.6	42.7	55	32	22	65
1.0	0.5	5.3	3.99	3.78	84.4	90.9	73.9	70.0	38	31	50	38
1.0	0.8	6.4	4.28	4.09	79.1	84.1	58.7	51.3	41	41	19	63

denoised signal. Therefore, we can evaluate this measure that will then indicate how the algorithm will perform using more realistic data.

The three methods of choosing the regularization parameter that have been discussed in Section V were experimented with, and the results are given in Sections VI-A and B. In addition, results were obtained using the OMP algorithm [12], [13] on the same data sets so that the performance of this algorithm could be compared with that of the regularized FOCUSS algorithm.

A. Discrepancy Principle and Sparsity Criterion

We first evaluate the performance of the discrepancy principle and sparsity criterion on a generated data set. In using the discrepancy principle to select a value for λ, we assume that we know something about the variance of the noise. This allows us to set a bound on the norm of the representation error as a function of the noise variance. Letting the variance of each noise component $n[i]$, $i = 1,2 \ldots m$ be σ^2, $E\{||n||^2\} = m\sigma^2$, the error bound is set to $Cm\sigma^2$, where C is a parameter chosen in the experiment. When using this approach the number of vectors r chosen from the matrix A to approximate a data vector b will vary for different data vectors. In order to compare the results obtained using FOCUSS with those obtained using OMP, we have to either fix the error and compare the number of vectors used or fix the number of vectors used and compare the error for each trial. It is not possible to obtain exactly the same error using regularized FOCUSS and OMP. However, we can run an experiment in which the number of vectors selected by each algorithm is the same. The FOCUSS algorithm is run with an upper bound set for the representation error, and the number of vectors selected r_{FOC} is found. Then, the OMP algorithm is run, which selects vectors sequentially from A. This algorithm is terminated once r_{FOC} vectors have been chosen. Thus, each algorithm has chosen the same number of vectors to represent the data vector, and the representation errors can be compared.

When the sparsity criterion is used in determining the regularization parameter λ, the number of vectors used in representing the data vector is fixed, i.e., the number of nonzero entries in the solution vector x that determines the representation $\hat{b} = Ax$ is fixed. In this experiment, we assume that the number of vectors used in forming the data vector b is known to be r. The goal is to find the best possible approximation as measured by the error ϵ^2 using a linear combination of r columns from the matrix A. Unfortunately, it is not trivial to control the number of vectors selected by the FOCUSS algorithm, but we now describe the method we used to do this. For a given data vector b, the OMP algorithm is first run, and the number of vectors used in approximating b is easily controlled so that the algorithm is terminated after r vectors have been selected. The representation error ϵ^2 is calculated and used as the upper bound on the error for the FOCUSS algorithm. Once the representation error falls below this bound, the number of vectors used by the FOCUSS algorithm r_{FOC} is obtained. If $r_{\mathrm{FOC}} > r$, the representation is pruned down to r by using OMP to select r of the r_{FOC} vectors. If $r_{\mathrm{FOC}} < r$, extra vectors are added using OMP until a total of r vectors are again used in the representation. This means that each algorithm will have a representation that utilizes r vectors from the matrix A.

Results:

Description of Table Parameters: The results obtained using the discrepancy principle and sparsity criterion are tabulated in Tables I and II for different values of SNR. In these tables, p is an additional factor used in the FOCUSS algorithm as given in (14) and can be used to trade off convergence speed against sparsity [4], [5]. C is the user chosen factor that determines the error bound used when running FOCUSS and using the discrepancy principle to determine λ. The column headed by # gives the average number of vectors selected in representing the data vectors, whereas $\#r$ gives the average number of vectors selected for representing b that are identical to the vectors used in generating b. The mean values obtained for the errors ϵ^2 and ϵ_1^2 using FOCUSS(F) and OMP(O) are tabulated. In addition, $\epsilon^2\%$ and $\epsilon_1^2\%$ give the percentage of trials in which FOCUSS performs better than OMP or vice-versa (note that the total is not 100% as there are some trials in which the algorithms perform identically, as measured to an accuracy of 10^{-4}).

Comment on ϵ_1^2 and ϵ^2 Results: From Tables I and II, we find that mean ϵ_1^2 is lower in all cases for the FOCUSS algorithm than the OMP algorithm. This is in keeping with the theory that we have presented since FOCUSS tries to denoise the data vector, whereas OMP does not. Looking at the figures given for $\epsilon_1^2\%$, we note that again, FOCUSS performs better than OMP, as measured by the number of trials in which it achieves a lower value of ϵ_1^2. In some cases, there is a very noticeable difference, as observed in the first line of Table I. However, mean ϵ_1^2 is in general a better indication of performance.

Examining the results for mean ϵ^2, we see that with SNR $=$ 20 dB, the FOCUSS algorithm does better than OMP. This is despite the fact that the OMP does better in most trials than FO-

TABLE II
$r = 7$: RESULTS OBTAINED USING THE SPARSITY CRITERION IN THE REGULARIZED FOCUSS (F) ALGORITHM AND THE RESULTS FROM THE OMP (O) ALGORITHM

p	#	$\#r$		mean ϵ_1^2 (10^{-3})		mean ϵ^2 (10^{-3})		ϵ_1^2 %		ϵ^2 %	
		F	O	F	O	F	O	F	O	F	O
SNR=20dB											
0	7	5.47	5.29	15.0	17.9	13.5	16.2	38	35	30	43
0.5	7	5.67	5.57	9.8	12.4	8.6	10.4	35	37	27	45
0.8	7	5.70	5.40	12.6	16.3	11.4	14.6	33	26	24	35
SNR=10dB											
0	7	4.26	4.13	83.9	88.1	47.0	42.8	56	34	32	58
0.5	7	4.46	4.30	78.5	82.4	44.4	40.8	15	54	26	62
0.8	7	4.43	4.18	79.8	82.9	45.2	44.9	39	32	29	42

CUSS as measured by $\epsilon^2\%$. The reason for this is explained by Fig. 2, which shows a histogram of $\epsilon^2_{\rm FOC} - \epsilon^2_{OMP}$ corresponding to the data used in generating the first two rows of Table I, where the discrepancy principle is used in choosing λ, and SNR $= 20$ dB, $r = 7$, $p = 0$, and in (a) $C = 0.8$ and (b) $C = 1.2$. From the skewed nature of the plots in Fig. 2(a) and (b), it is noted that when OMP performs better, it only performs *marginally* better, but when FOCUSS performs better, it sometimes performs *significantly* better. Thus, the mean values of ϵ^2, which are given in Table I, favor the FOCUSS algorithm. A similar assessment can be made of the results presented in Table II.

When the SNR is reduced to 10 dB, as given in Tables I and II, the FOCUSS algorithm still does better than the OMP algorithm, as measured by mean ϵ_1^2. However, it no longer gives a lower value for mean ϵ^2. This is expected since the regularized FOCUSS is acting to denoise the data vector and represent the underlying denoised vector b', whereas the OMP does nothing to remove noise and represents the data vector b. This is also reflected in the figures given for $\epsilon^2\%$ and $\epsilon_1^2\%$.

Comment on $\#r$ Results: Finally, we look at the results given in the $\#r$ column of each of the tables. In Table I (discrepancy principle) and Table II (sparsity criterion), it is observed that the values for FOCUSS are better in all cases than those for OMP. This shows that the regularized FOCUSS algorithm is more successfully selecting the true underlying generating vectors than the OMP.

B. Modified L-Curve Method

The modified L-curve method requires some knowledge of the noise level or the target SNR for the representation that is then used to find $\lambda_{\rm max}$ and $\lambda_{\rm min}$, as described in Section V-C. The values of λ_{max}, λ_{min}, and λ_c are found in each FOCUSS iteration.

The noise vector $n \in \mathcal{R}^m$ has Gaussian random entries, and each component of the vector has variance σ^2. $||n||^2$ has a χ^2 distribution, and the limits on ϵ^2 are chosen such that $P(||n||^2 \leq \epsilon^2_{\rm min}) = P(||n||^2 \geq \epsilon^2_{\rm max}) = T$ for some threshold T, which was set to 0.1 in these experiments. The values of ϵ^2 are obtained by using the SNR values, and for SNR $= 20$ dB, the limits on ϵ^2 are found to be $\epsilon^2_{\rm min} = 0.0062$ and $\epsilon^2_{\rm max} = 0.0142$. These limits are increased by a factor of 10 for SNR $= 10$ dB. If the true SNR of the data is unknown, targets for the SNR can be used to decide the error limits. If the desired SNR is approximately X dB, an upper error limit can be set using $(X - \Delta_1)$ dB as an SNR target and a lower limit using $(X + \Delta_2)$ dB leading to the limits

$$\epsilon^2_{\rm upper} = 10^{-(X-\Delta_1)/10}||b||^2 \quad (21)$$

$$\epsilon^2_{\rm lower} = 10^{-(X+\Delta_2)/10}||b||^2. \quad (22)$$

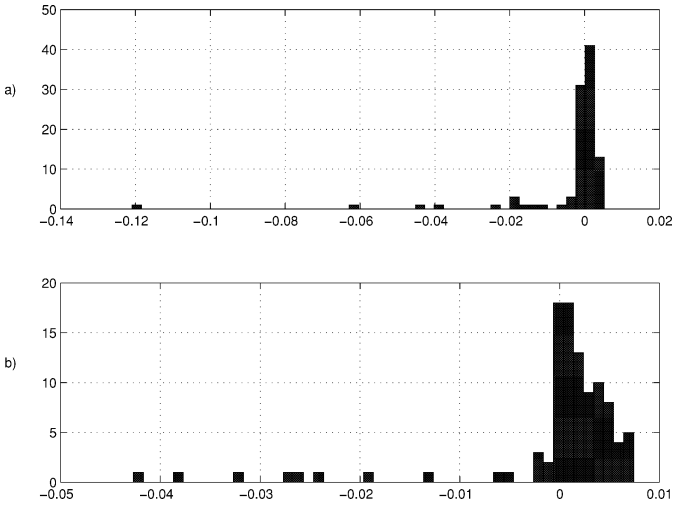

Fig. 2. Histogram of $\epsilon^2_{\rm FOC} - \epsilon^2_{OMP}$ with SNR $= 20$ dB, $r = 7$, $p = 0$. (a) $C = 0.8$. (b) $C = 1.2$.

For each data vector b, the FOCUSS algorithm is first run, and $r_{\rm FOC}$ is found. Then, the OMP algorithm is run on the same data vector and is terminated after it has selected $r_{\rm FOC}$ vectors.

Results: In the first simulations, the SNR is set to known values of 10 and 20 dB, and the algorithm has *precise* knowledge of the SNR. In contrast, in the second simulation, although the true SNR is 20 dB, we assume that *a priori* we are only able to put upper and lower bounds on the SNR; the lower limit is set to 15 dB and the upper limit to 25 dB from which values of $\epsilon^2_{\rm lower}$ and $\epsilon^2_{\rm upper}$ can be obtained. For each case, 100 different data vectors were generated, and the results are given in Tables III and IV, respectively.

Comment on ϵ_1^2 and ϵ^2 Results: From Table III, we note that the mean value of ϵ_1^2 obtained using the FOCUSS algorithm is lower in every instance than the value obtained using the OMP algorithm. This result is further emphasized by the values of $\epsilon_1^2\%$: For instance, with $p = 0.8$, FOCUSS gives a lower value of ϵ_1^2 in over 70% of the trials. In addition, it is noted that $\#r$ is greater for FOCUSS than OMP in all rows of the table, which

TABLE III
SNR IS ASSUMED KNOWN AND $r = 7$: RESULTS OBTAINED USING THE MODIFIED L-CURVE CRITERION IN REGULARIZED FOCUSS (F) WITH PRECISE KNOWLEDGE OF SNR AND OMP (O) ALGORITHMS

p	#	$\#r$		mean ϵ_1^2 (10^{-3})		mean ϵ^2 (10^{-3})		ϵ_1^2 %		ϵ^2 %	
SNR=20dB											
		F	O	F	O	F	O	F	O	F	O
0	7.04	5.35	5.05	9.7	19.2	10.3	17.6	53	47	42	51
0.5	6.86	5.32	5.05	10.2	20.0	9.4	17.8	45	55	36	55
0.8	10.69	5.97	5.68	8.3	11.7	3.6	4.9	74	26	26	74
SNR=10dB											
		F	O	F	O	F	O	F	O	F	O
0	4.08	3.46	3.06	117.1	128.3	115.2	118.6	52	39	39	48
0.5	4.34	3.58	3.22	99.1	108.7	93.8	116.8	59	34	41	46
0.8	8.38	4.57	4.14	82.4	93.9	39.3	29.5	76	24	21	77

TABLE IV
TRUE SNR IS 20 DB AND $r = 7$: RESULTS OBTAINED USING THE MODIFIED L-CURVE CRITERION, WHERE THE TARGET SNR IS TAKEN TO BE BETWEEN 15 AND 25 DB

p	#	$\#r$		mean ϵ_1^2 (10^{-3})		mean ϵ^2 (10^{-3})		ϵ_1^2 %		ϵ^2 %	
		F	O	F	O	F	O	F	O	F	O
0	5.48	4.69	4.67	20.4	22.4	21.1	23.1	44	50	42	46
0.5	5.51	4.93	4.71	16.3	21.8	17.4	21.5	50	45	38	41

shows that we are correctly identifying more of the generating vectors using FOCUSS rather than OMP. In common with the results of Section VI-A, with SNR $= 20$ dB, mean ϵ^2 obtained for FOCUSS is lower than that obtained for OMP. However, with SNR $= 10$ dB, the OMP achieves a lower value. As we have previously stated, this is due to the fact that the regularized FOCUSS tries to represent the underlying denoised signal b' rather than the data signal b.

In Table IV, mean ϵ_1^2 and mean ϵ^2 are both lower for FOCUSS than OMP. However, the gap is not as large as that observed in the top half of Table III. This can be attributed to the less accurate *a priori* knowledge in these simulations. The achieved SNR can be calculated from mean ϵ^2. For $p = 0$, SNR_{FOC} is 16.8 dB, and for $p = 0.5$, it is 17.6 dB. The results give a lower SNR than the true SNR, but the number of selected vectors is approximately 5.5, whereas $r = 7$ vectors were used in generating the data vectors.

Comment on ϵ^2 versus Number of Selected Vectors: In Fig. 3(a), we provide a plot of the number of vectors selected in each trial (the average over 100 trials gives # in the table); this is found to vary between 3 and 11. We plot the corresponding value of ϵ^2 obtained in each trial in Fig. 3(b). First, it is seen that the variance in the error is small, and this means that the variance in the approximation quality for the different trials is also small. The achieved SNR for each trial varies between 15 and 25 dB, which corresponds to the predetermined limits on the SNR. Comparing Fig. 3(a) and (b), it can be seen that the error is, in general, not smaller for the trials where the number of selected vectors is large. This observation, together with the small variance of ϵ^2, indicates that the method we have developed of combining the target SNR with the linear scale L-curve works well. The problem in the original L-curve method that made no attempt to control the quality of the representation and often led to the choice of a regularization parameter that overemphasized either sparsity or representation error has been remedied by our algorithm, which produces more robust results.

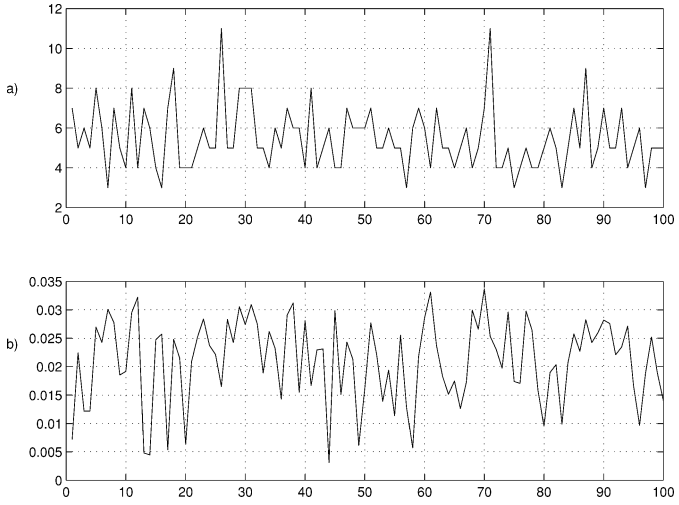

Fig. 3. Modified L-curve FOCUSS algorithm for simulated data with true SNR $= 20$ dB and the target SNR bounds set to 15 and 25 dB. (a) Number of selected vectors in each trial. (b) ϵ_{FOC}^2 for each trial.

VII. CONCLUSION

In this paper, we have tackled the problem of subset selection in noisy environments. A formal methodology was developed using a Bayesian framework that led to the derivation of a regularized FOCUSS algorithm to solve this problem. The convergence of the regularized FOCUSS algorithm is established, and it is shown that the stable fixed points of the algorithm are sparse. We then considered the practical implementation of the algorithm that involves the choice of the regularization parameter. Motivated by different applications, three methods were examined for setting this parameter: The discrepancy principle assures a certain quality of fit in the representation, the sparsity criterion enforces a certain subset size, and the L-curve criterion seeks a tradeoff between representation error and the size of the selection subset. We proposed a novel modified L-curve

procedure, incorporating a target SNR, to determine the regularization parameter that was able to overcome the robustness problems we encountered in applying the L-curve method directly to our application.

Through simulations, we showed that the regularized FOCUSS algorithm can better identify the generating vectors than an algorithm based on a forward sequential selection of vectors such as OMP. It must be stated that the OMP performance is still good and is adequate for many applications. The much greater complexity of the regularized FOCUSS algorithm means that it is not suitable for real-time processing; therefore, the OMP algorithm would be preferred. However, when the detection of the true underlying vectors is of foremost importance rather than the processing time, especially as encountered in some medical applications, the improved detection ability of the regularized FOCUSS algorithm makes the utilization of the regularized FOCUSS algorithm developed here attractive.

REFERENCES

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

[2] S. F. Cotter, "Subset selection algorithms with applications," Ph.D. dissertation, Univ. California, San Diego, La Jolla, CA, 2001.

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

[4] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstructions from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Processing*, vol. 45, pp. 600–616, Mar. 1997.

[5] B. D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Trans. Signal Processing*, vol. 47, pp. 187–200, Jan. 1999.

[6] ——, "Basis selection in the presence of noise," in *Proc. 32nd Asilomar Conf. Signals, Syst., Comput.*, Monterey, CA, Nov. 1998.

[7] K. Engan, B. D. Rao, and K. Kreutz-Delgado, "Regularized FOCUSS for subset selection in noise," in *Proc. Nordic Signal Processi. Symp.*, Linkoping, Sweden, June 2000, pp. 247–50.

[8] P. C. Hansen, "Analysis of discrete ill-posed problems by means of the L-curve," *SIAM Rev.*, vol. 34, pp. 561–80, Dec. 1992.

[9] P. C. Hansen and D. P. O'Leary, "The use of the L-curve in the regularization of discrete ill-posed problems," *SIAM J. Scientific Comput.*, vol. 14, pp. 1487–1503, Nov. 1993.

[10] K. Engan, "Frame based signal representation and compression," Ph.D. dissertation, Norges Teknisk Naturvitenskapelige Univ., Stavanger, Norway, 2001.

[11] S. O. Aase, J. H. Husoy, J. Skretting, and K. Engan, "Optimized signal expansions for sparse representation," *IEEE Trans. Signal Processing*, vol. 49, pp. 1087–96, May 2001.

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

[13] 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.

[14] S. G. M. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, pp. 3397–3415, Dec. 1993.

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

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

[17] ——, (1997) A general approach to sparse basis selection: Majorization, concavity, and affine scaling. Elect. Comput. Eng. Dept. Univ. Calif. San Diego, La Jolla. [Online]. Available: http://raman.ucsd.edu.

[18] ——, "Measures and algorithms for best basis selection," in *Proc. ICASSP*, vol. III, Seattle, WA, May 1998, pp. 1881–1884.

[19] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins Univ. Press, 1989.

[20] W. Weldon, F. Galton, and K. Pearson, *Biometrika*, 1953.

[21] M. K. Varanasi and B. Aazhang, "Parametric generalized Gaussian density estimation," *J. Acoust. Soc. Amer.*, vol. 86, no. 4, pp. 1404–15, Oct. 1989.

[22] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Comput.*, vol. 12, pp. 337–65, Nov 2001.

[23] M. Foster, "An application of the Wiener-Kolmogorov smoothing theory to matrix inversion," *J. SIAM*, vol. 9, pp. 387–92, 1961.

[24] 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.

[25] J. Palmer, "Function curvature, relative concavity, and a new criterion for sub- and super-gaussianity, Tech. Rep.," Elect. Comput. Eng. Dept., La Jolla, CA, 2002.

[26] D. Luenberger, *Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley, 1989.

[27] S. Reeves, "An efficient implementation of the backward greedy algorithm for sparse signal reconstruction," *IEEE Signal Processing Lett.*, vol. 6, pp. 266–268, Oct. 1999.

[28] M. Hanke, "Limitations of the L-curve method in discrete ill-posed problems," *BIT*, 1996.

[29] T. Reginska, "A regularization parameter in discrete ill-posed problems," *SIAM J. Sci. Comput.*, vol. 17, pp. 740–9, May 1996.

[30] C. M. Leung and W. S. Lu, "An L-curve approach to optimal determination of regularization parameter in image restoration," in *Proc. CCECE*, Vancouver, BC, Canada, Sept. 1993, pp. 1021–1024.

[31] H. W. Engl and W. Grever, "Using the L-curve for determining optimal regularization parameters," *Numer. Math.*, vol. 69, pp. 25–31, 1994.

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 Procesing Theory and Methods Technical Committee.

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

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

Shane F. Cotter 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 design engineer with Nokia Mobile Phones, San Diego. His main research interests are statistical signal processing, signal and image representations, optimization, and speech recognition.

Jason Palmer received the B.A. degree in philosophy from the University of Chicago, Chicago, IL, in 1996 and the B.S. and M.S. degrees in electrical and computer engineering from Illinois Institute of Technology, Chicago, in 1999 and the University of California at San Diego (UCSD), La Jolla, in 2001. He is currently pursuing the Ph.D. degree in electrical and computer engineering at UCSD.

His interests are in convex analysis and optimization and feature extraction.

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 he is currently a Professor with the Department of Electrical and Computer Engineering, UCSD. Prior to joining the Faculty of UCSD, he was a researcher at the NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, where he worked on the development of intelligent robotic systems. His current research interests include applications of statistical data analysis and learning theory, nonlinear signal processing, and computational intelligence to communication systems, bioinformatics, predictive failure analysis, machine intelligence, and robotics.