Basis Selection in the presence of Noise

B. D. Rao and K. Kreutz-Delgado
Electrical and Computer Engineering Dept.
University of California, San Diego
La Jolla, CA 92093-0407
e-mail: {brao, kkreutzd}@ucsd.edu

Abstract

In this paper, we consider procedures to enhance the reliability of basis selection procedures with particular attention being given to methods based on minimizing diversity measures. To deal with noise in the data, basis selection procedures based on a Bayesian framework are considered. An algorithm based on the MAP estimation procedure is developed which leads to a regularized version of the FOCUSS algorithm. Another approach considered is to select basis vectors over multiple measurement vectors thereby achieving an averaging effect and enhancing the reliability. New diversity measures are presented for this purpose, and algorithms are derived for minimizing them.

1 Introduction

The goal of this paper is to develop robust, subset selection methods that have applications to signal representation and to finding sparse solutions to linear inverse problems [1]. Particular attention will be paid to methods that are based on minimizing diversity measures [2], and a formal methodology is developed for deriving algorithms that can deal with noise in the data. In particular, in Section 3 it is shown how a Bayesian framework coupled with priors consistent with the $\ell_{(p\leq 1)}$ diversity measure can lead to a regularized version of the FOCUSS algorithm. Issues relating to the convergence of these procedures are also discussed in detail. In addition, the framework is extended to deal with multiple measurement vectors (MMV), a scenario common in linear inverse problems. This MMV approach also allows for dealing with uncertainty in the data by virtue of the averaging inherent in the framework. New diversity measures are presented for this purpose, and algorithms are derived for minimizing them in Section 4.

2 Minimizing Diversity Measures

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

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

A is an $m \times n$ matrix with $m \leq n$, and $\text{rank}(A) = m$. The columns of A are formed from the elements of the dictionary for signal representation problems or derived from the physics of the problem for linear inverse problems. There are *many* solutions to the system of equations (1) and the best basis selection problem corresponds to identifying a few columns of the matrix A that best represent the data vector b [3]. This corresponds to finding a solution to (1) with few nonzero entries.

Finding an optimal solution to the basis selection problem generally requires a combinatorial search which is computational unattractive. Therefore suboptimal techniques are usually employed and we discuss one such method called FOCUSS, for **FOC**al **U**nderdetermined **S**ystem **S**olver [4]. The FOCUSS method was motivated by the observation that if a sparse solution is desired then choosing a solution based on the smallest 2-norm is not appropriate. The minimum 2-norm criteria favors solutions with many small nonzero entries, a property that is contrary to the goal of sparsity [5, 4]. Consequently there is a need to consider the minimization of alternative measures that promote sparsity. In this context, of particular interest are diversity measures, functionals which measure the lack of concentration/sparsity, and algorithms for minimizing them to obtain sparse solutions. A popular diversity measure is the $\ell_{(p\leq 1)}$ diversity measure given by [2, 6],

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

0-7803-5148-7/98/$10.00©1998 IEEE

Minimizing these measures naturally leads to the algorithm FOCUSS, whose iterations are as follows [4, 2, 6]:

$$x_{k+1} = W_{k+1} (AW_{k+1})^{+} b, \qquad (3)$$

where $W_{k+1} = \mathrm{diag}(|x_k[i]|^{1-\frac{p}{2}})$. 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 deemphasized. In the end a few columns survive to represent b providing a sparse solution.

Interesting insight can be gained into (3), when viewed as a sequence of weighted minimum 2-norm problems [4]. Defining $q \triangleq W_{k+1} x$, 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

$$q_{k+1} = \arg\min_q \|q\|^2 \text{ subject to } AW_{k+1} q = b. \qquad (4)$$

3 Basis Selection in the presence of Noise

The previous work did not explicitly account for noise in the data in the derivation of FOCUSS [4, 2]. It was assumed that there is a perfect match between the data b and a linear combination of a few columns of A, and later reasonable modifications were made to the algorithm to deal with noise [4]. We take a formal approach and extend the FOCUSS method to deal with noise in the measurements using a Bayesian framework. As we will see, the stochastic framework provides theoretical insights and assists in developing robust methods. For this discussion, we assume that

$$b = Ax + v,$$

where v is an additive noise vector. Furthermore, in this formulation x is also assumed to be a random vector independent of v. Under these assumptions, a Maximum A Posteriori (MAP) estimate of x can be obtained,

$$\begin{aligned} x_{map} &= \arg\max_x \ln p(x|b) \\ &= \arg\max_x [\ln p(b|x) + \ln p(x)] \\ &= \arg\max_x [\ln p_v(b - Ax) + \ln p(x)]. \end{aligned}$$

This formulation is general with considerable flexibility. In order to proceed further, some assumptions about the noise v and the solution vector x have to be made. 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 i.i.d elements[1], i.e. $p_v(v) = c_1 e^{-\frac{\|v\|^2}{2\sigma^2}}$. The distribution of x is quite important for the generation of sparse solutions. For this purpose, the elements $x[i]$ are assumed to be i.i.d. random variables with density function concentrated near the origin to promote sparsity [7]. A distribution consistent with the $\ell_{(p\leq 1)}$ diversity measures is $p_x(x) = c_2 e^{-\frac{\beta}{2}\mathrm{sgn}(p)\sum_{i=1}^{n}|x[i]|^p}$, where $\beta > 0$ and c_2 is the normalizing constant. Substituting these densities in the expression for the MAP estimate results in

$$\begin{aligned} x_{map} &= \arg\min_x J(x), \\ \text{where } J(x) &= \left[\|Ax - b\|^2 + \mathrm{sgn}(p)\sigma^2\beta \sum_{i=1}^{n} |x[i]|^p\right]. \end{aligned}$$

Note that $p = 2$ gives rise to the standard regularized least squares problem. For $p \leq 1$ it can be shown that the minima of $J(x)$ are sparse. Following the factored-gradient approach in [2, 6], an iterative algorithm can be derived to minimize $J(x)$ which has the form[2] [8]:

$$x_{k+1} = W_{k+1} \left(A_{k+1}^T A_{k+1} + \lambda I\right)^{-1} A_{k+1}^T b, \qquad (5)$$

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

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

algorithm (5) can be expressed as

$$x_{k+1} = W_{k+1} A_{k+1}^T \left(A_{k+1} A_{k+1}^T + \lambda I\right)^{-1} b. \qquad (6)$$

When the noise level is reduced, $\sigma \to 0$, then $\lambda \to 0$ and the algorithm reduces to the original FOCUSS algorithm (3). The algorithm (6) has an interesting interpretation as Tikhonov regularization applied to (4). This can be readily seen by rewriting (6) as a solution to a regularized least squares problem. Then we have $x_{k+1} = W_{k+1} q_{k+1}$, where

$$q_{k+1} = \arg\min_q \|AW_{k+1} q - b\|^2 + \lambda \|q\|^2$$

Interestingly, this results in an algorithm identical to that suggested in [4]. In [4], the algorithm was arrived at as a way to make the 2-norm minimization problem of (4) more robust to noise. The derivation provided here provides formal support to the approach.

[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

3.1 Convergence Results

Before we present convergence results concerning the regularized algorithms, we first present some preparatory 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, m being the number of rows.

Lemma 1 *In each iteration of the regularized FOCUSS algorithm (6), if* $\|q_{k+1}\|^2 \leq \sum_{i=1}^{n} |x_k[i]|^p$, *then the algorithm converges and the stable fixed points are sparse solutions.*

Proof: The proof follows readily from the convergence proof of FOCUSS presented in [2].

Lemma 2 *In each iteration of the regularized FOCUSS algorithm (6), if* $\|q_{k+1}\|^2 \leq n$, *then the algorithm converges and the stable fixed points are sparse solutions.*

Proof: The proof follows readily from the results presented in [9]. In [9], the proof for the case $p = 0$ is presented and it can be readily generalized.

Lemma 3 *If the regularized FOCUSS algorithm (6) converges for a given* $\lambda > 0$, *then the algorithm converges to a sparse solution.*

Proof: The proof is by contradiction. Let the algorithm converge to a non sparse solution, and let that be a stable fixed point. Then $q^* = \arg\min_q \|AW_* q - b\|^2 + \lambda\|q\|^2$. Let $AW_* q^* = b + e$, then q^* is also an optimum solution to the problem $\arg\min_q \|q\|^2$ subject to $AW_* q = b + e$ This is the unregularized FOCUSS algorithm (3) with b replaced by $b+e$. From [4, 2], we know that the stable fixed points of this algorithm are sparse solutions. This leads to a contradiction.

Theorem 1 *For* $\lambda = 0$, *and for* $\lambda > \frac{\|b\|^2}{n}$, *the algorithm converges and the stable fixed points are sparse solutions.*

Proof: The results for $\lambda = 0$ follows readily from the fact that this corresponds to the unregularized FOCUSS algorithm (3) which has the desired properties [4, 2]. For the case $\lambda > \frac{\|b\|^2}{n}$, note that for each iteration $\min_q \|AW_{k+1}q - b\|^2 + \lambda\|q\|^2 < \|AW_{k+1}0 - b\|^2 + \lambda\|0\|^2 = \|b\|^2$. Since q_{k+1} is the minimum to this function, we have $\lambda\|q_{k+1}\|^2 \leq \|b\|^2$. For $\lambda > \frac{\|b\|^2}{n}$, this implies $\|q_{k+1}\|^2 < n$ at each iteration. The convergence then follows from lemma 2.

Though the above convergence result is limited, in practice we have not experienced any problem with convergence for all $\lambda > 0$.

3.2 Practical Considerations

The quality of the sparse solution obtained via the regularized version of FOCUSS is governed by the choice of λ, and there remains the problem of determining a proper value for λ. Also, there appears to be no practical reason to limit the choice of λ to a fixed value for all the iterations. 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. The first by the desire to produce stable sparse solutions without the need for much prior information. The second by the desire to ensure a certain quality of representation potentially motivated by the availability of some information on the perturbations. The third by the need to ensure a certain degree of sparsity on the solution as would be required in applications like compression.

3.2.1 L-Curve Criteria

The regularizing parameter is found by striking a compromise between minimizing the norm of the solution vector, $\|q\|^2$, versus the error in the representation, $\|AW_{k+1}q - b\|$. In this context, this choice also translates into controlling the sparse nature of the solution. As one varies λ, one obtains regularized solution q_λ whose norm varies continuously and decreases monotonically as λ increases from zero to infinity. More precisely for $\lambda \in [0, \infty]$, the norm of q_λ lies between $\|(AW_{k+1})^+ b\|$ and zero, i.e. $\|q_\lambda\| \in [0, \|(AW_{k+1})^+ b\|]$. On the other hand $\|AW_{k+1}q_\lambda - b\|$ increases monotonically as λ increases from 0 to infinity. A plot of $\|q_\lambda\|$ versus $\|AW_{k+1}q_\lambda - b\|$ has a characteristic L-shaped appearance and the regularization parameter is chosen near the L-shaped "corner" of the L-curve [10]. Several methods for doing so are investigated and discussed in [10]. The use of such an approach was first suggested in [4].

3.2.2 Quality of Fit criteria

Another potentially useful approach is to try to seek a sparse solution that assures a certain quality in the nature of the representation, i.e. $\|Ax - b\| \leq \epsilon$. 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 to derive the regularized solution, an iterative algorithm can be derived which at each iteration computes

$x_{k+1} = W_{k+1} q_{k+1}$, where

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

An algorithmic for computing q_{k+1} is given in [11]. The convergence of the algorithm to a sparse solution can be shown based on lemma 1, because it is possible to show that in each iteration $\|q_{k+1}\|^2 \leq \sum_{i=1}^{n} |x_k[i]|^p$.

3.2.3 Sparsity Criteria

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. $\lim_{k\to\infty} \text{rank}(AW_{k+1}) = r$. So a desirable approach would be to use a sequence λ_k to satisfy this limiting rank property, while providing the best fit possible. A reliable procedure for doing this is not yet available. One practical approach is to use a sequential basis selection method like the Order Recursive Matching Pursuit (ORMP) to select r columns [12], and to determine a value for the error ϵ in the representation. This ϵ can be the basis of FOCUSS along the lines suggested in section 3.2.2. If the procedure returns more columns than desired, one can either prune the selected subset or go with ORMP solution whichever is better.

4 Basis Selection using Multiple Measurement Vectors (MMV)

In the MMV case, more than one measurement vector is available enhancing the ability of suboptimal procedure to find the proper sparse solution and also offers potential robustness to noise. Mathematically the MMV problem can be stated as solving the following system of equations[3] [13]:

$$Ax^{(l)} = b^{(l)},\ l = 1, .., L, \text{ or } AX = B, \qquad (7)$$

where $X = [x^{(1)}, .., x^{(L)}]$, and $B = [b^{(1)}, .., b^{(L)}]$. L is the number of measurement vectors usually assumed to be much less than m. $b^{(l)}$ denotes the lth measurement vector, and $x^{(l)}$ the corresponding solution. Furthermore, $x^{(l)}[i]$ denotes the (i,l)th entry of X, and we denote the ith row of X by $x[i]$. A is a $m \times n$ matrix which is assumed known, and it also assumed that $m << n$ and $\text{rank}(A) = m$.

For the MMV problem, the desired solution requires not only that the individual columns of X, $x^{(l)}$, have a sparse structure but that they share the structure and have a common sparsity profile, i.e. the indices of the nonzero entries are independent of l. In terms of the matrix solution X, this implies having a a few nonzero rows as opposed to a few non zero entries. To deal with this problem we develop a new diversity measure that reflects the requirement of few nonzero rows as opposed to simply few nonzero entries and has the form

$$J^{(p,q)}(X) = \text{sgn}(p) \sum_{i=1}^{n} (\|x[i]\|_q)^p,\ p \leq 1, q \geq 1, \quad (8)$$

where $\|x[i]\|_q = \left(\sum_{l=1}^{L} |x^{(l)}[i]|^q\right)^{\frac{1}{q}}$. This is an extension of the "$\ell_{(p\leq 1)}$ diversity measures" and as p approaches zero, it provides a count of the number of nonzero rows in X. For simplicity, we only consider the case of $q = 2$ in the remainder of the paper, and for notational simplicity denote $J^{(p,2)}(X)$ by $J^{(p)}(X)$.

To minimize the $J^{(p)}(X)$ diversity measure subject to the equality constraints (7), we start with the standard method of Lagrange multipliers. Define the Lagrangian $L(X, \Lambda)$,

$$L(X, \Lambda) = J^{(p)}(X) + \sum_{l=1}^{L} \lambda_l^T (Ax^{(l)} - b^{(l)}),$$

where $\lambda_l, l = 1, .., n$ 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 $l = 1, .., L$

$$\begin{aligned} \nabla_{x^{(l)}} L(X_*, \Lambda_*) &= \nabla_{x^{(l)}} J^{(p)}(X_*) + A^T \lambda_{l,*} = 0, \\ \nabla_{\lambda_l} L(X_*, \Lambda_*) &= Ax_*^{(l)} - b^{(l)} = 0, . \end{aligned} \qquad (9)$$

The gradient of the diversity measure $J^{(p)}(X)$ with respect to element $x[i,l]$ can be readily shown to be

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

For tractability purposes, as suggested in [2], we use a *factored representation* for the gradient vector of the diversity measure, i.e.

$$\nabla_{x^{(l)}} J^{(p)}(X) = \alpha(X) \Pi(X) x^{(l)}, \qquad (10)$$

where $\alpha(X) = |p|$, and $\Pi(X) = \text{diag}(\|x[i]\|^{p-2})$. At this point it is useful to note that the $\Pi(X)$ matrix is independent of the column index l which leads to considerably simplicity in the algorithm. This is a consequence of the choice $q = 2$ in (8), and other choices do not lead to such tractability. From (9) and (10), the stationary points satisfy

$$\alpha(X_*) \Pi(X_*) X_* + A^T \Lambda_* = 0 \text{ and } AX_* - B = 0. \quad (11)$$

[3] As in the single measurement vector case, for tractability we first address the case of no noise and incorporate noise considerations later.

From (11), carrying out some simple manipulations as in [2], it can be show that

$$X_* = \Pi^{-1}(X_*)A^T(A\Pi^{-1}(X_*)A^T)^{-1}B. \qquad (12)$$

Eq. (12) suggests the following iterative procedure for computing X_*,

$$X_{k+1} = \Pi^{-1}(X_k)A^T(A\Pi^{-1}(X_k)A^T)^{-1}B. \qquad (13)$$

The computation of $\Pi^{-1}(X_k) = \mathrm{diag}(\|x_k[i]\|^{(2-p)})$ 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_*).

As in the case of the original FOCUSS algorithm, the algorithm can be viewed as solving a sequence of weighted Frobenius norm minimization problems. This results in the following form of the algorithm:

$$\begin{aligned} W_{k+1} &= \mathrm{diag}(\|x_k[i]\|^{1-\frac{p}{2}}), \\ Q_{k+1} &= A^+_{k+1}B, \text{ where } A_{k+1} = AW_{k+1} \qquad (14) \\ X_{k+1} &= W_{k+1}Q_{k+1}. \end{aligned}$$

The algorithm is a generalization of the FOCUSS class of algorithms and can be initialized by using the minimum Frobenius norm solution or any other suitable solution.

Some simulations supporting the usefulness of the algorithm (14) can be found in [13]. To deal with noise, one can use Tikhonov regularization or Truncated SVD based regularization. The Tikhonov regularization procedure can be derived formally using the Bayesian formulation as in Section 3.

4.1 Convergence Analysis

In algorithm (14), denoting the entries of Q_k by $q_k[i,l]$ and the rows by $q_k[i]$ and using the fact that in each iteration the Frobenius norm of Q is minimized, it can be shown that when the algorithm has not converged ($X_{k+1} \neq X_k$)

$$\|Q_{k+1}\|_F^2 = \sum_{i=1}^{n} \|q_{k+1}[i]\|^2 < \sum_{i=1}^{n} \|x_k[i]\|^p. \qquad (15)$$

The entries of $X_{k+1} = W_{k+1}Q_{k+1}$ can be written as $x_{k+1}[i] = \|x_k[i]\|^{\frac{2-p}{2}} q_{k+1}[i]$. Hence

$$J^{(p)}(X_{k+1}) = \sum_{i=1}^{n} \|x_{k+1}[i]\|^p = \sum_{i=1}^{n} \|x_k[i]\|^{\frac{(2-p)p}{2}} \|q_{k+1}[i]\|^p.$$

Let $r = \frac{2}{p}$ and define s by $\frac{1}{r} + \frac{1}{s} = 1$. So we have $s = \frac{2}{2-p}$. Applying the Hölder inequality to the above equation we have

$$\begin{aligned} J^{(p)}(X_{k+1}) &\leq \left(\sum_{i=1}^{n} \|x_k[i]\|^{\frac{(2-p)ps}{2}}\right)^{\frac{1}{s}} \left(\sum_{i=1}^{n} \|q_{k+1}[i]\|^{pr}\right)^{\frac{1}{r}} \\ &= \left(\sum_{i=1}^{n} \|x_k[i]\|^p\right)^{\frac{2-p}{2}} \left(\sum_{i=1}^{n} \|q_{k+1}[i]\|^2\right)^{\frac{p}{2}} \\ &< \left(\sum_{i=1}^{n} \|x_k[i]\|^p\right)^{\frac{2-p}{2}} \left(\sum_{i=1}^{n} \|x_k[i]\|^p\right)^{\frac{p}{2}} \\ &= \sum_{i=1}^{n} \|x_k[i]\|^p = J^{(p)}(X_k) \end{aligned}$$

where the second strict inequality follows from (15). Thus we have shown that the diversity measure $J^{(p)}(X_k)$ is reduced in each iteration for $0 < p \leq 1$. Similarly, the proof can be carried out for $p < 0$.

References

[1] B.D. Rao. "Signal Processing with the Sparseness Constraint". In *Proc. ICASSP* 1998.

[2] B. D. Rao and K. Kreutz-Delgado. "An Affine Scaling Methodology for Best Basis Selection". *to appear in the IEEE Trans. on Signal Processing*, Jan. 1999.

[3] S. G. Mallat and Z. Zhang. "Matching Pursuits with Time-Frequency Dictionaries". *IEEE Trans. ASSP*, 41(12):3397–3415, Dec. 1993.

[4] I.F. Gorodnitsky and B.D. Rao. *IEEE Trans. on Signal Processing*, 45:600–616, March 1997.

[5] S. Chen and D. Donoho. "Application of Basis Pursuit in Spectrum Estimation". In *Proc. ICASSP* 1998.

[6] K. Kreutz-Delgado and B. D. Rao. "Measures and Algorithms for Best Basis Selection". In *ICASSP 98.*

[7] B. A. Olshausen and D. J. Field. "Sparse Coding with an Overcomplete Basis Set: A strategy employed in V1". *In Press*, 1997.

[8] B.D. Rao, K. Kreutz-Delgado, and S. Dharanipragada. "Improving Spectral Resolution using Basis Selection". In *SSAP Workshop*, Sept. 1998.

[9] B. D. Rao and I. Gorodnitsky. "Affine Scaling Transformation Based Methods for Computing Low Complexity Sparse Solutions". In *Proc. ICASSP 96.*

[10] P. C. Hansen. "Analysis of the Discrete Ill-Posed Problems by Means of the L-Curve". *SIAM Review*, 1992.

[11] B. D. Rao. "Analysis and Extensions of the FOCUSS Algorithm". In *Asilomar* 1996.

[12] S. F. Cotter, M. N. Murthi, and B. D. Rao. "Fast Basis Selection Methods". In *Asilomar*, 1997.

[13] B.D. Rao and K. Kreutz-Delgado. "Sparse Solutions to Linear Inverse Problems with Multiple Measurement vectors". In *IEEE DSP Workshop*, Aug. 1998.