An Algebraic Approach to the Subset Selection Problem

Ahmed Tewfik and Mohammed Nafie
Dept. of Electrical Engineering, University of Minnesota
Minneapolis, MN 55455
e-mail: tewfik,mnafie@ece.umn.edu

ABSTRACT

The need for decomposing a signal into its *optimal* representation arises in many applications. In such applications, one can usually represent the signal as a combination of an over-complete dictionary elements. The non-uniqueness of signal representation, in such dictionaries, provides us with the opportunity to adapt the signal representation to the signal. The adaptation is based on sparsity, resolution and stability of the signal representation. In this paper, we propose an algebraic approach for identifying the sparsest representation of a given signal in terms of a given over-complete dictionary. Unlike other current techniques, our approach is guaranteed to find the solution, given that certain conditions apply. We explain these conditions.

1 Introduction

In many applications one needs to identify the sparsest representation of the given signal in terms of the elements of an overcomplete set of vectors or signals. Such applications include signal coding for compression, chemical analysis of compounds and direction finding. In signal coding for example, certain parts of the signal might be most compactly represented by certain types of known dictionaries while other parts might need other dictionaries. For example in audio coding, there are tonal parts which are best represented by Fourier dictionaries, and there are edges which are best represented by wavelets. Some other parts might be represented by some other dictionary. If we form a large dictionary that contains all these dictionaries, and choose the sparsest representation of the signal in terms of the elements of this overcomplete dictionary, we will achieve a higher compression rate. In direction of arrival estimation, we have an array of sensors which has a certain response of a unit strength signal for each arrival angle. If we form a matrix containing all these responses for all possible arrival angles, we can obtain an estimate for the arrival angles. Another application for such a problem, is when we have several actuators with each of them producing a certain signal, and we are trying to fit a received signal to the minimum possible number of actuator signals. This has been used for example in neuroimaging.

This problem is referred to in the literature as the "subset selection problem". Other examples of such are the decomposition of signals using over-complete dictionaries such as wavelets, wave packets, cosine packets etc.. In this paper we will present an algebraic approach for solving this problem.

Notice that, while in the case of, for example, lossless compression of signals, we have an exact representation of the signal in terms of the dictionary, there are cases where this does not apply. If the model representing the signal has discrepancies, or noise added to it, then we have what we call "stochastic subset selection", where the signal cannot be represented exactly in terms of the dictionary elements, but rather we try to find the best sparse representation such as to minimize a certain cost function in the difference between this representation and the signal. Notice that this can be used in blind equalization of communication channels [1].

Several algorithms for solving the subset selection problem have been proposed in the literature [2][3][4][5][6][7]. But, as the problem is NP-complete [9], none of these methods always finds the true global solution. Subset Selection algorithms use several optimization criteria. We shall present a brief description of several of these methods in the next section. Our algebraic technique attempts to solve the problem through generating other vectors that span the space of the solution. Our technique is gauaranteed to reach the true global solution if certain conditions apply. These conditions might be on the size of the overcomplete dictionary, or they might be on the structure of the dictionary as will be explained later. This summary will be organized as follows: in the next section we formally explain the subset selection problem, and we give a brief description of the current available techniques to arrive at the sparsest solution. In section 3 we present our technique, explaining the condition in which it will apply. In section 4 we give some simulation results, and then we end with the conclusion.

2 Subset Selection Problem

The subset selection problem may be described as follows in a finite dimensional space. We have N discrete waveforms $\vec{\phi}_i, 1 \leq i \leq N$. Each of these waveforms is a vector of size $M \times 1$ with its energy normalized to unity. These waveforms are collected in a $M \times N$ matrix $\mathbf{\Phi}$. Each column of $\mathbf{\Phi}$ therefore represents a particular waveform. The signal to be analyzed, $\vec{s}$, is a vector of size $M \times 1$. The signal decomposition problem is then equivalent to solving

$$\mathbf{\Phi}\vec{\alpha} = \vec{s}. \tag{1}$$

If the dictionary forms a basis ($M = N$ and all the waveforms are independent), then there is a unique solution given by $\vec{\alpha} = \mathbf{\Phi}^{-1}\vec{s}$. Further, if the waveforms are mutually or-

thonormal (orthonormal transform), then $\mathbf{\Phi}^{-1} = \mathbf{\Phi}^T$, and computing the decomposition is simple.

However, in over-complete dictionaries, we have $N > M$ (usually $N \gg M$). In such cases, we do not have a unique solution. In other words, we may represent the signal $\vec{s}$ using the waveforms in the dictionary $\mathbf{\Phi}$ in infinitely many ways. Given any M independent columns from $\mathbf{\Phi}$, we may uniquely determine the decomposition. However, there may exist a better or more sparse signal representation. The problem therefore is to find the "optimal" representation of the signal with as few elements from the dictionary as possible, i.e., with the maximal number of zero components. This problem has been proven to be NP-complete [9].

Currently, there are several methods which attempt to solve this problem. The Method of Frames (MOF) [3] picks the solution that has the minimum l_2 norm of $\vec{\alpha}$. MOF thus selects the solution that is closest to the origin. However, the MOF is not sparsity preserving. Even if the signal has an underlying sparse representation in terms of the elements of the dictionary, the MOF representation is usually very dense. Thus MOF representation unnecessarily uses many elements (sometimes much more than M) for signal representation.

The Matching Pursuit (MP) algorithm [4] is an iterative algorithm that picks the element that best correlates with the present residual. This algorithm is greedy and myopic, and in certain cases it chooses the wrong elements in the first few iterations and further iterations are used in correcting the initial mistakes. Several examples have been reported where MP performs very poorly [2].

The Best Orthogonal Basis (BOB) [7] method designed for wavelet packets and cosine packets, picks a single orthogonal basis out of all possible orthogonal bases in the dictionary based on the minimum entropy criterion. This algorithm is fast and delivers near optimal sparsity if the columns of $\mathbf{\Phi}$ are near-orthogonal. But for highly non-orthogonal elements in $\mathbf{\Phi}$, this algorithm often fails. Further, this method applies to certain structured dictionaries only.

The Basis Pursuit (BP) [2] minimizes the l_1 norm of the solution vector $\vec{\alpha}$ by converting the decomposition problem to a linear programming problem. The advantages of such solution are that linear programming gives the global solution instead of a local one, and such a solution is usually sparse but not necessarily so. However, the complexity of linear programming is much higher than the other techniques. The numerical implementation used in [2] to reduce this complexity does not always guarantee a solution. For non-exact representation, this method requires quadratic programming [2].

An iterative technique we proposed in [6] divides the dictionary into two sets: the "active" set consisting of the dictionary elements that would be used to describe the signal and the "inactive" set consisting of the rest of the elements. In each iteration, the procedure swaps a single vector between the "active" and "inactive" sets. This swapping is such that the elements in the current signal representation corresponding to the "inactive" set are smaller than those corresponding to the "active" set. Although this technique has the flexibility of choosing the starting point, and hence by starting from the end point of any of the other algorithms and by iterating on that, this algorithm is guaranteed to give a better solution [1], like the other techniques, this method may converge to local minima of the norm of the error between the signal and its approximation.

Several other suboptimal approaches have been developed for optimal subset selection for use in least squares regression in Statistics. They include backward elimination and sequential replacement [10].

A true global optimization technique is to use brute force search to search through all linearly independent subsets of a given size from the dictionary and decompose the signal in each such basis. A numerical measure may be applied to select the "optimality" criterion. This is an NP-complete problem, with complexity increasing combinatorially with the size of dictionary. Global optimization is, therefore, computationally prohibitive.

3 Proposed Technique

The technique that we are proposing is an exact technique in the sense that, under certain conditions on the dictionary, or equivalently on the matrix $\mathbf{\Phi}$, it will find the true optimal solution. This technique relies on being able to derive, from the given signal vector $\vec{s}$, several other vectors that all lie in the same minimal subspace as $\vec{s}$. If we were able to derive a number of these vectors equal to the dimension of the subspace, then we can identify the columns of the matrix $\mathbf{\Phi}$ corresponding to non-zero elements in $\vec{\alpha}$ as we will shortly see. We will also show that for certain kinds of matrices, among which are Vandermonde matrices, we can easily generate these vectors whatever the size of the matrix $\mathbf{\Phi}$ is. For general matrices $\mathbf{\Phi}$ however, this can only be done given certain conditions on the size of the matrix. We will derive these conditions. To explain our technique, we will first present an example using a Vandermonde matrix, then explain how this technique can be generalized. Assume that we have the following equation

$$\begin{bmatrix} a_1 & a_2 & a_3 & . & . & a_{10} \\ {a_1}^2 & {a_2}^2 & {a_3}^2 & . & . & {a_{10}}^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ {a_1}^6 & {a_2}^6 & {a_3}^6 & . & . & {a_{10}}^6 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ . \\ . \\ . \\ \alpha_{10} \end{bmatrix} = \begin{bmatrix} s_1 \\ s_2 \\ . \\ . \\ s_6 \end{bmatrix}. \tag{2}$$

Assume also that only three α_k of $\vec{\alpha}$ are non-zero. Define three new signals $\vec{s_1}$, $\vec{s_2}$ and $\vec{s_3}$ as $\begin{bmatrix} s_1 & s_2 & s_3 & s_4 \end{bmatrix}^T$, $\begin{bmatrix} s_2 & s_3 & s_4 & s_5 \end{bmatrix}^T$, and $\begin{bmatrix} s_3 & s_4 & s_5 & s_6 \end{bmatrix}^T$ respectively. We can write the above equation as

$$\begin{bmatrix} a_1 & a_2 & a_3 & . & . & a_{10} \\ {a_1}^2 & {a_2}^2 & {a_3}^2 & . & . & {a_{10}}^2 \\ {a_1}^3 & {a_2}^3 & {a_3}^3 & . & . & {a_{10}}^3 \\ {a_1}^4 & {a_2}^4 & {a_3}^4 & . & . & {a_{10}}^4 \end{bmatrix} \begin{bmatrix} \alpha_1 & \beta_1 & \gamma_1 \\ \alpha_2 & \beta_2 & \gamma_2 \\ . & . & . \\ . & . & . \\ . & . & . \\ \alpha_{10} & \beta_{10} & \gamma_{10} \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 \\ s_2 & s_3 & s_4 \\ s_3 & s_4 & s_5 \\ s_4 & s_5 & s_6 \end{bmatrix}. \tag{3}$$

where $\beta_i = a_i * \alpha_i$, $\gamma_i = {a_i}^2 * \alpha_i$. Note that β_k and γ_k are non-zero only if α_k is non-zero. Therefore the matrix on the

[1] In the worst case, it won't improve the solution any further

right hand side is a full representation of the subspace which the vectors corresponding to non-zero α's form. We form the left null space matrix of the right hand side matrix, and multiply this null space matrix by the matrix $\mathbf{\Phi}$. If the matrix composed of the coefficients is full rank, we will get zeros in the columns corresponding to the non-zero coefficients, and hence we can easily identify them.

In the general case, let us assume that we have a maximum of q non-zero components in the vector $\vec{\alpha}$, and assume there exists $p > q$ such that we can find matrices ψ_i of size $p * N$, a diagonal matrix D_i of size $M * M$ and W of size $p * M$, such that

$$\psi_i \mathbf{\Phi} = W D_i. \tag{4}$$

Notice that this equation is equivalent to

$$\psi_i \vec{s} = W D_i \vec{\alpha}, \tag{5}$$

for all $\vec{s}$ and corresponding $\vec{\alpha}$ vectors.

Given the positions of the non-zero elements of α we can write

$$\psi_i \vec{s} = W^{'} c_i^{'}. \tag{6}$$

If we can find k such matrices, where $k \geq q$, then

$$\left[\begin{array}{cccccc} \psi_1 \vec{s} & \psi_2 \vec{s} & . & . & \psi_k \vec{s} \end{array} \right] = W^{'} \left[\begin{array}{cccccc} c_1^{'} & c_2^{'} & . & . & c_k^{'} \end{array} \right]. \tag{7}$$

Now, the rank of the left hand side matrix cannot be larger than q. Since $p > q$, therefore a left null space matrix exists. Call it w^T. Assuming that $\left[\begin{array}{cccccc} c_1^{'} & c_2^{'} & . & . & c_k^{'} \end{array} \right]$ is full rank, i.e. of rank q, then

$$w^T W^{'} = 0. \tag{8}$$

Therefore by multiplying w^T by the matrix W, we can identify the columns corresponding to non-zero components, and we are done.

The problem now is to find the matrices ψ_i, D_i and W. We notice that equation 4 can be written as

$$(\mathbf{\Phi}^T \otimes I_p) vec(\psi_i) = vec(W D_i). \tag{9}$$

Or

$$\left[\begin{array}{cccc} & w_1 & & \\ & & w_2 & \\ \mathbf{\Phi}^T \otimes I_p & & & \ddots \\ & & & w_M \end{array} \right] \left[\begin{array}{c} vec(\psi_i) \\ \left[\begin{array}{c} d_{i1} \\ d_{i2} \\ . \\ . \\ d_{iM} \end{array} \right] \end{array} \right] = 0, \tag{10}$$

where w_i is the ith column of W, and d_{il} is the lth diagonal element of D_i. Let us right the above equation as

$$R\mu = 0. \tag{11}$$

R has size $Mp * (Np + M)$. To find a solution to the previous equation, R has to be of deficient row rank. As we showed before, this works for any size Vandermonde matrix, and there might be other special matrices where this can work, especially highly structured matrices such as Toeplitz or Hankel matrices. We will investigate this matter further. But for general matrices with no structure a sufficient condition would hence be

$$Mp < Np + M, \tag{12}$$

or

$$p < \frac{M}{M - N}. \tag{13}$$

Obviously, this would be useful if M is close to N, but if M is larger than $2N$, this technique wouldn't work for general matrices. Fig. 1 shows how p changes with $\frac{M}{N}$. Notice that the maximum number of non-zero elements that can be found using this technique for general matrices has to be less than p. Say $q = p - 1$. We have to find at least q solutions to 11. Therefore the condition on the number of non-zero elements is

$$Nq + N - Mq >= q, \tag{14}$$

or

$$q <= \frac{N}{M - N + 1}. \tag{15}$$

Therefore for a minimum of 2 non-zero elements, M cannot be bigger than $1.5 * N$ for the limit as $N \to \infty$, and less than that for smaller N. Fig. 2 shows the maximum ratio between M and N vs N, the length of the vectors in the dictionary.

3.1 Simulation Results

Here, we present simulation results to the "stochastic subset selection" problem, were we assume some noise (or model inaccuracies) to be added. Therefore the problem can be written as

$$\mathbf{\Phi} \vec{\alpha} + noise = \vec{s}. \tag{16}$$

We start by an example on a random Vandermonde matrix. A Vandermonde matrix of size 6x12 was generated whose vectors have unit norm. This matrix was multiplied by a vector which has only 2 non-zero elements in random places. White Gaussian noise was added to the resultant vector. Notice that we have 2 options here, either use a $p = 4$ and generate 3 $\vec{\psi_i}$'s or use $p = 3$ and and generate 4 $\vec{\psi_i}$'s. Since we have noise added, we will choose the columns corresponding to the least norm of $w^T W$. We ran several simulations and compared the probability of error using the 2 values of p, and the results are shown in Fig. 3. We did the same for a random 10x12 matrix. We also used $p = 3$ and $p = 4$. The results are shown in Fig. 4.

4 Conclusion

In this paper we presented a new algebraic technique for solving the subset selection problem. We have shown cases were our technique is guaranteed of finding the global solution. We have shown that for special types of matrices, from which we were able to identify the Vandermonde matrix, the technique works.

We are currently investigating other types of matrices for which the proposed technique will work, and also how our algorithm behaves if two representations of the signal vector yield the same sparsity.

References

[1] M. Nafie and A. Tewfik, "Reduced Complexity M-ary Hypotheses Testing in Wireless Communications," ICASSP 1998.

[2] S. Chen, and D. Donoho, "Atomic Decomposition by Basis Pursuit," Preprint, 1995.

[3] I. Daubechies, "Time-Frequency Localization operators: a Geometric Phase Space Approach," IEEE Trans. on Inf. Theory, vol. 34, no. 4, pp. 605-612, July 1988.

[4] S. Mallat, and Z. Zhang, "Matching Pursuits with Time-Frequency Dictionaries," IEEE Trans. on Signal Processing, vol. 41, no. 12, pp. 3397-3415, Dec. 1993.

[5] I. Gorodnitsky, and B. Rao, "Sparse Signal Reconstruction from Limited Data Using FOCUSS: A Re-weighted Minimum Norm Algorithm," IEEE Trans. on Sig. Proc., Vol. 45, No. 3, March 1997.

[6] M. Nafie, M. Ali,and A. Tewfik, "Optimal Subset Selection for Adaptive Signal Representation," ICASSP 1996

[7] R. Coifman, and M. Wickerhauser, "Entropy-based Algorithms for Best Basis Selection," IEEE Trans. on Inf. Theory, vol. 38, no. 2, pp. 713-718, March 1992.

[8] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C*, Cambridge University Press, 1992

[9] B. Natarajan, "Sparse Approximate Solutions to Linear Systems," SIAM J. Comput., vol. 24, pp.227-234, Apr. 1995.

[10] A. J. Miller, *Subset Selection in Regression*, London, UK: Chapman and Hall, 1990.

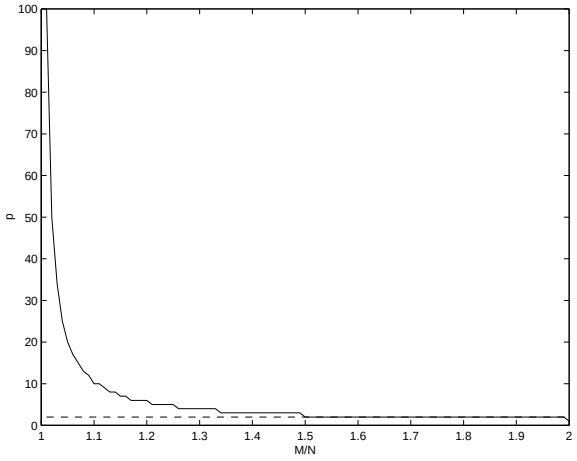

Figure 1: Change of p with the ratio between the 2 dimensions of the dictionary

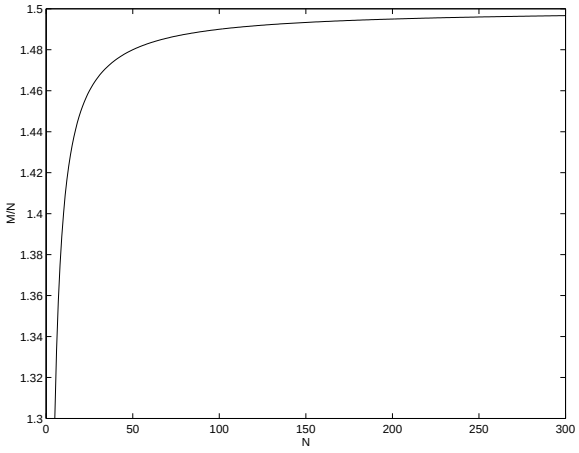

Figure 2: The maximum ratio between the 2 dimensions of the dictionary vs the length of the vectors

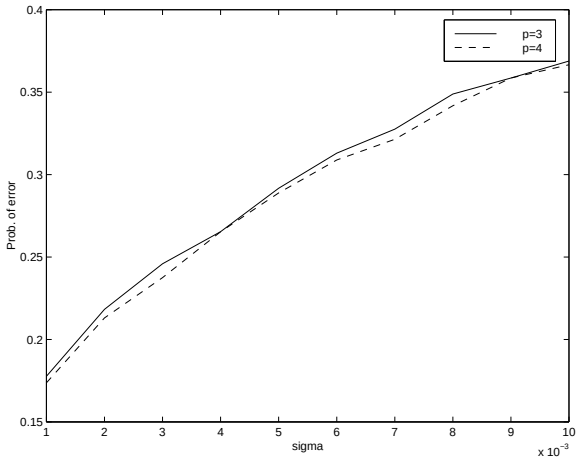

Figure 3: Probability of not finding the correct vectors at various noise standard deviations for 2 values of p for a Vandermonde matrix

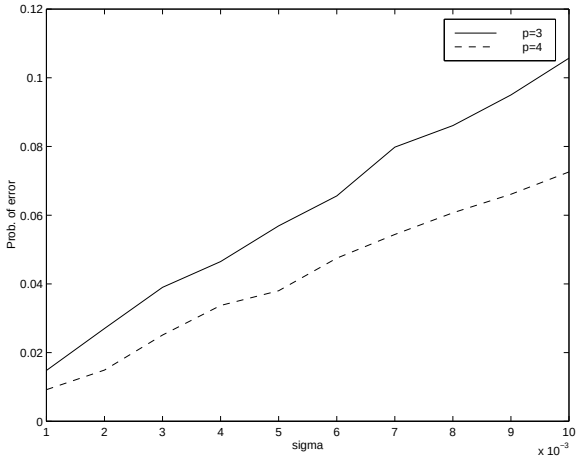

Figure 4: Probability of not finding the correct vectors at various noise standard deviations for 2 values of p for a random matrix