SIGNAL PROCESSING WITH THE SPARSENESS CONSTRAINT

Bhaskar D. Rao

Electrical and Computer Engineering dept.
University of California, San Diego
La Jolla, CA 92093-0407
e-mail: brao@ucsd.edu

ABSTRACT

An overview is given of the role of the sparseness constraint in signal processing problems. It is shown that this is a fundamental problem deserving of attention. This is illustrated by describing several applications where sparseness of solution is desired. Lastly, a review is given of the algorithms that are currently available for computing sparse solutions.

1. INTRODUCTION

In many signal processing applications, algorithms based on the ℓ_2 norm criteria have found wide spread use. Least squares problem wherein the ℓ_2 norm of the residual error, $||Ax-b||_2^2$, is minimized are common place. In rank deficient least-squares it is popular to choose the smallest 2-norm solution of all feasible solutions. In signal representation and linear inverse problems involving under-determined system of equations, the non-uniqueness is often resolved by choosing the minimum 2-norm solution. The ℓ_2 norm results in a optimization framework that is attractive both theoretically and computationally. Consequently, this has been a topic of much academic interest with many signal processing curricula including a detailed treatment of this subject.

In this paper, we discuss an alternate paradigm for developing algorithms which also has wide application potential. This is the incorporation of the sparseness constraint on signal processing algorithm development. Sparseness constraint refers to the requirement that the vector being sought or optimized must have as few non-zero entries as possible. Depending on the application, sparseness constraint can be imposed on the residual being minimized, or on the solution being computed. Sparseness constraint on the residual has been explored in many applications e.g. seismic deconvolution [1], speech modeling [2], etc. This is often achieved by using an optimization function such as the ℓ_p norm, $(1 \leq p \leq 2)$, or by using statistical characterization of the residue that is compatible with the sparseness assumption. The under-determined problem has received much attention recently because of its application to the signal representation problem [3, 4, 5, 6, 7, 8, 9, 10, 11], and to linear inverse problems [12, 13, 14, 15, 16, 17]. In this context, the goal is to find a solution x with the least number of non-zero entries. We concentrate on this latter problem in this paper. However, some of these ideas do have relevance to the problem of computing a sparse residual.

Though the paper is written with a view towards providing an overview, the presentation is clearly biased by our experiences and work. Fortunately, there are several other papers on this topic in this session. We are hopeful that collectively they will paint a more complete picture.

2. PROBLEM FORMULATION

Linear Inverse problems or the problem of signal representation can be formulated as a problem of finding a solution to an under-determined system of equations [15, 7, 10],

$$Ax = b. \qquad (1)$$

A is an $m \times n$ matrix with $m < n$, and it is assumed that $\text{rank}(A) = m$. In linear inverse problems, the columns of A are formed from the forward model usually determined based on the physics of the problem. For signal representation, the columns of A are formed using basis vectors chosen from a over-complete dictionary. b is the $m \times 1$ measurement vector or the given signal to be represented. The goal is to solve for x, a $n \times 1$ vector. A requirement on the solution vector x is that it be sparse, i.e. many of its entries be zero. This requirement on the vector x naturally arises out of the application requirements as discussed in section 3.

The under-determined system of equations (1) has many solutions. Any solution can be expressed as

$$x = x_{mn} + v,$$

where x_{mn} is the minimum 2-norm solution (i.e. solution with the smallest ℓ_2 norm defined as $||x||_2^2 = \sum_{i=1}^n x[i]^2$) and is given by $x_{mn} = A^+b$, where A^+ denotes the Moore-Penrose pseudo-inverse. The vector v is any vector that lies in $\mathcal{N}(A)$, the null space of A. In this case A has a nontrivial null-space of dimension $(n - m)$. In many situations, a popular approach has been to set $v = 0$ and to select x_{mn} as the desired solution. However, the minimum 2-norm criteria favors solutions with many small nonzero entries, a property that is contrary to the goal of sparsity/concentration [7, 15, 13]. Consequently there is a need to develop approaches that lead to sparse solutions.

3. APPLICATIONS

3.1. Signal Representation

In this application, A is formed from the basis vectors used in the expansion. Often, A is square and an orthogonal matrix, making it easy to transition between the signal b and its transform x.

Recently, there has been a great deal of interest in finding efficient representations of signals using an over-complete dictionary [3, 4, 5, 6, 7]. The motivation for such an approach is that a minimal spanning set of basis vectors is usually only adequate to efficiently represent a small class of signals, while forming an over-complete dictionary using a carefully chosen set of redundant basis vectors can represent a larger class of signals compactly. The problem is commonly referred to as basis selection. Finding a succinct representation requires that most of the coefficients of the representation are zero. Developing algorithms for optimal basis selection is a subject of current research.

Though the optimal basis selection problem makes the requirement of sparsity evident, similar situations arise in routine operations and often go unnoticed. An example is the common practice of zero padding and using the FFT to densely sample the Fourier transform of a sequence. If we have a sequence of $y[l]$ of duration m, and we use a n point FFT ($n \geq m$), then

$$y[l] = \frac{1}{n}\sum_{k=0}^{n-1} Y[k] e^{j\frac{2\pi}{n}kl},\ 0 \leq l \leq (m-1).$$

The problem of computing $Y[k]$ can be readily expressed as solving an under-determined system of equations. The signal vector b is formed from the given sequence $y[l]$. The A matrix is the DFT matrix, i.e. the kth column is $\frac{1}{n}[1, e^{j\omega_k} e^{j2\omega_k} ... e^{j(m-1)\omega_k}]^T$, where $\omega_k = (k-1)\frac{2\pi}{n}$. The solution vector x contains the Fourier coefficients $Y[k]$. The expansion coefficients $Y[k]$ are computed as

$$Y[k] = \sum_{l=0}^{n-1} y[l] e^{-j\frac{2\pi}{n}kl} = \sum_{l=0}^{m-1} y[l] e^{-j\frac{2\pi}{n}kl}, 0 \leq k \leq (n-1).$$

The second equality exploits the fact that the sequence is zero padded. In matrix form, the solution vector obtained via the FFT is,

$$x_{fft} = n\, A^H b = A^+ b.$$

The pseudo-inverse is equal to the Hermitian transpose scaled by n because of the orthogonality of the rows of A. In summary, zero padding and computing the FFT corresponds to choosing a representation with the *smallest 2-norm* solution from among many possible choices. This may be adequate for most purposes. However, being aware of alternative solutions, and methods for computing them, is beneficial. An example to demonstrate the usefulness of the sparseness constraint is now given.

Consider a sequence $y[l]$ which consists of a single complex exponential, i.e. $y[l] = e^{j\frac{2\pi}{n}k_0 l}, 0 \leq l \leq (m-1)$, where $n = 128$, $m = 64$, and the frequency of the exponential is $k_0 = 33$. The magnitude of the Fourier transform computed using a 64 point FFT and a 128 point FFT are shown in figure 4.3. The second figure corresponds to frequency domain interpolation. When a 128 point FFT is computed, one of the basis vectors in the set, the 34th column of A, has the same frequency as the data itself. Under such conditions, a desirable and intuitive solution would be one that has all zero entries except for the one coefficient corresponding to column number 34. Unfortunately, it is evident that the FFT solution, which is the minimum 2-norm solution, does not possess representational simplicity. However, algorithms that employ the sparseness constraint can be used to obtain high resolution nonparametric spectrum estimates [15, 26, 16, 7].

3.2. Neuromagnetic Imaging

Linear inverse problems with the sparsity requirement on the solution arise naturally in Magnetoencephalography (MEG) [18, 19, 20, 21, 14]. In MEG, one is interested in solving the neuromagnetic inverse problem which is to estimate the cerebral current sources underlying a measured distribution of the magnetic field. Measurement of the external magnetic field is made with an array of super-conducting quantum interference device (SQUID) detectors. The current field is solved by inversion of the Biot-Savart law which relates the continuous vector current field and the induced magnetic field. According to the Biot-Savart law, the current density $\vec{J}$ as a function of position $\vec{r}^{'}$ relates to the magnetic induction $\vec{B}$ at a given point of observation $\vec{r}$ as

$$\vec{B}(\vec{r}) = k \int \vec{J}(\vec{r}^{'}) \times \frac{\vec{r} - \vec{r}^{'}}{|\vec{r} - \vec{r}'|^3} d\vec{r}^{'}, \qquad (2)$$

where k is a constant if magnetic permeability assumed constant throughout the volume. For numerical purposes, discretization of the equation is carried out to get a linear equation. During the discretization process, the reconstruction volume is divided into N voxels (VOlume ELements) and the putative continuous current is approximated in each voxel by a point dipole $Q(\vec{r}_n)$,

$$\vec{B}(\vec{r}_m) = k \sum_{n=1}^{N} Q(\vec{r}_n) \times \frac{\vec{r}_m - \vec{r}_n}{|\vec{r}_m - \vec{r}_n|^3}. \qquad (3)$$

This leads to a linear inverse problem which requires solving an under-determined system of equations as the number of voxels N is larger than the number of measurements. However, *a priori* (from physiological evidence) it is known that the currents are limited in spatial extent. This suggests the use of the sparseness constraint for solving the linear inverse problem.

3.3. Speech Coding

Considerable work on sparsity has been done in the area of speech coding, particularly in the computation of the excitation sequence [22, 23, 24]. In speech coding, analysis by synthesis (ABS) are popular approaches for coding with the multi-pulse excited linear predictive coder (MPELPC) being most relevant one for exposing the sparseness aspect. The overall approach in MPELPC coders can be divided into finding a vocal tract model and the appropriate excitation sequence. The vocal tract model is an all-pole filter whose parameters are estimated using linear prediction methods. They are usually computed over segments/frames of speech, typically, 20 msec long. The excitation is computed for each subframe with the excitation consisting of a few non zero pulses strategically placed. Typically there are four sub-frames per frame. Finding the location of the pulses, and their amplitude, results in solving a linear inverse problem with the sparseness constraint.

3.4. Other Applications

The linear inverse problems with the sparseness constraint arises in many other applications, and researchers in several areas have independently attempted to solve this problem. Applications include band-limited extrapolation and spectral estimation [25, 26], direction of arrival estimation [15], functional approximation [27, 28, 29], failure diagnosis [30], sparse coding [31], and pattern recognition for medical diagnosis [32]. It is clear that an effective solution to this problem has wide ranging consequences.

4. ALGORITHMS

Algorithms for computing sparse solutions are discussed in the framework of signal representation/ basis selection. Without loss of generality, it is assumed that the vectors in set/dictionary $\{a_k\}_{k=1}^n$ are of unit norm. The basis selection can be stated as follows. Given a signal vector $b \in R^m$, and a preset error tolerance, ϵ, find the most compact representation of b to within the given tolerance using the basis vectors $\{a_k\}_{k=1}^n$. This involves determining the number r (the *sparsity index*) and the set of vectors $\{a_{k_i}\}_{i=1}^r$ that best model b.

Finding an solution with optimal sparsity index r is NP hard and requires an combinatorial search [28, 7]. For example, if we were interested in selecting p vectors that best represented the data, this would require searching over the $\binom{n}{p}$ possible ways in which the basis sets can be chosen to find the best solution. Though there exist efficient techniques for such a search, the cost of such searches is prohibitive for even moderate size problems making finding an optimal solution using an exhaustive search infeasible. Suboptimal methods have been developed to deal with this problem and some of them are discussed next.

4.1. Sequential Basis Selection Methods

The methods described in this section select the basis vectors sequentially, i.e. the basis set is built up one vector at a time.

Basic Matching Pursuit (BMP): This method was suggested in [3] and independently for speech coding [22, 23]. In this basis selection method, in the pth iteration the vector most closely aligned with the residual b_{p-1} is chosen, where b_{p-1} denotes the residual vector after the $(p-1)$ th iteration. The computation involved for the selection is

$$k_p = \arg\max_l |a_l^H b_{p-1}|. \tag{4}$$

The new residual vector is then computed as

$$b_p = P_{a_{k_p}}^{\perp} b_{p-1} = b_{p-1} - (a_{k_p}^H b_{p-1}) a_{k_p}. \tag{5}$$

Equations (4) and (5) give the Basic Matching Pursuit (BMP) algorithm. The procedure terminates when either $p = r$ (for specified sparsity index r) or $||b_p|| \leq \epsilon$ (for specified ϵ).

Order Recursive Matching Pursuit (ORMP): This method was developed in [24, 27, 28]. In this method, the pursuit of the matching pth basis vector conceptually involves solving $(n-p+1)$ order recursive least squares problems of the type $\min_y ||[S_{p-1}, a_l]y - b||$, and selecting the vector a_l that reduces the residual the most. $S_{p-1} = \left[a_{k_1}, a_{k_2}, ..., a_{k_{p-1}}\right]$, and is the matrix formed with the basis vectors chosen in the previous iterations. With the notation $S_{p,l} = [S_{p-1}, a_l]$, the index of the next basis vector is given by

$$k_p = \arg\min_l ||P_{S_{p,l}}^{\perp} b|| \,. \tag{6}$$

The residual is then updated, i.e. $b_k = P_{S_p}^{\perp} b$, where $S_p = [S_{p-1}, a_{k_p}]$. Note that the projection operator $P_{S_{p,l}}$ can be recursively updated, and efficient computational algorithms developed [27, 10]

Compared to the BMP, the ORMP is computationally more demanding. However, it has been found to yield more compact representations. A modification to the BMP, called the Modified Matching Pursuit (MMP), was recently suggested to overcome some of its limitations [10].

4.2. Parallel Basis Selection

In these methods, *all* the vectors of the dictionary are initially selected, and processed and vectors are asymptotically eliminated until a requisite number remain.

ℓ_1 **norm minimization** [7, 12]: In this method, instead of finding a minimum 2-norm solution to (1), a solution is found that minimizes the l_1 norm $\sum_{k=1}^n |x[k]|$. The attractiveness of this solution stems from the fact that it leads to sparse solutions, and efficient linear programming techniques can be utilized to compute the solution [7].

FOCUSS: The algorithm FOCUSS, for **FOC**al **U**nderdetermined **S**ystem **S**olver, was recently developed [15, 25, 26, 21]. The iterations of the algorithm are as follows [15]:

$$x_{k+1} = W_{k+1}\,(AW_{k+1})^{+}\,b,\ \text{ where } W_{k+1} = \text{diag}(|x_k[i]|^{1-\frac{p}{2}}).$$

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. Studies have shown that FOCUSS computes sparse solutions [15].

Recently, it has been shown that the method can be derived within a unified framework based on majorization theory, starting from diversity measures, functionals which measure the lack of concentration/sparsity, and minimizing them to obtain sparse solutions [17, 33]. The diversity measure minimized by FOCUSS is the $\ell_{(p\leq 1)}$ diversity measure given by

$$E^{(p)}(x) = \text{sgn}(p) \sum_{i=1}^{n} |x[i]|^p, \;\; p \leq 1. \tag{7}$$

The diversity measures $E^{(p)}(x)$ for $0 \leq p \leq 1$ are the general family of entropy-like measures defined in [5, 6], and also discussed in [12, 13], for computing sparse solutions. Other diversity measures, and generalization of the methodology can be found in [17, 33].

4.3. Other Algorithms

Alternate algorithms exist for computing sparse solutions based on statistical approaches [8, 9], concave cost function minimization [13], efficient search methods that exploit structure [4, 5], etc. Similar computational needs arise in the neural network pruning problem which can be considered a generalization of this problem to nonlinear mappings. Relevant algorithms can be found in the pruning literature of neural networks [34].

The above list of methods is only partial. Unfortunately, many signal processing algorithms with the sparseness constraint have been explored in application specific contexts and so are not readily accessible. There is a pressing need for the documentation of the state of the art in algorithms for computing sparse solutions.

5. REFERENCES

[1] J. M. Mendel. *Optimal seismic deconvolution : an estimation-based approach*. Academic Press, 1983.

[2] C-H. Lee. “On Robust Linear Prediction of Speech”. *IEEE Trans. ASSP*, ASSP-36(5):642–650, May 1988.

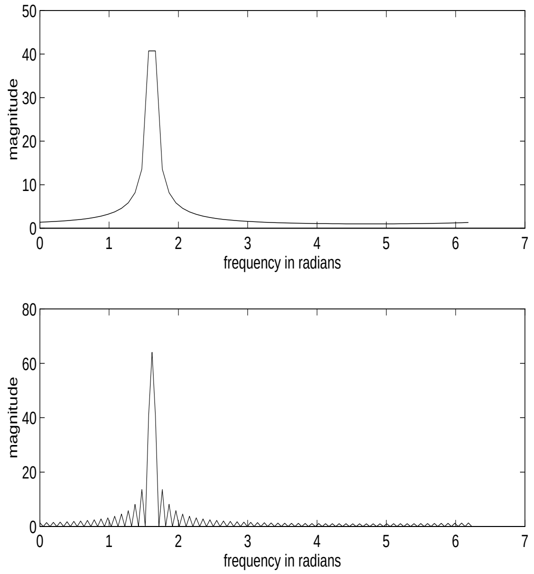

Figure 1: The magnitude of Fourier Transform of a complex exponential of duration 64 obtained using a 64 and 128 point FFT.

[3] S. G. Mallat and Z. Zhang. "Matching Pursuits with Time-Frequency Dictionaries". *IEEE Trans. ASSP*, Dec. 1993.

[4] R. R. Coifman and M. V. Wickerhauser. "Entropy-Based Algorithms for Best Basis Selection". *IEEE Trans. Inform. Theory*, IT-38(2):713–718, March 1992.

[5] M. V. Wickerhauser. *Adapted Wavelet Analysis from Theory to Software*. A. K. Peters, Wellesley, MA, 1994.

[6] D. Donoho. "On Minimum Entropy Segmentation". In *Wavelets: Theory, Algorithms, and Applications, edited by C. K. Chui et al*, pp. 233–269. Academic Press, 1994.

[7] S. Chen and D. Donoho. "Basis Pursuit". In *Twenty-Eighth Asilomar Conference, Vol. I*, Nov. 1994.

[8] H. Krim, S. Mallat, D. Donoho, and A. S. Willsky. "Best Basis Algorithm for Signal Enhancement". In *Proc. ICASSP 1995*, pages 1561–1564, Detroit, Michigan, May 1995.

[9] H. Krim. "On the Distributions of Optimized Multi-Scale Representations". In *Proc. ICASSP* , April 1997.

[10] S. F. Cotter, M. N. Murthi, and B. D. Rao. "Fast Basis Selection Methods". In *Proc. of the 31st Asil. Conf.*, Nov. 1997.

[11] C. S. Burrus, R. A. Gopinath, and H. Guo. *Introduction to Wavelets and Wavelet Transforms*. Prentice Hall, 1998.

[12] B. Jeffs and M. Gunsay. "Restoration of Blurred Star Field Images by Maximally Sparse Optimization". *IEEE Trans. on Image Processing.*, 2(2):202–211, 1993.

[13] G. Harikumar and Y. Bresler. " A New Algorithm for Computing Sparse Solutions to Linear Inverse Problems". In *Proc. ICASSP*, volume III, May 1996.

[14] I.F. Gorodnitsky, J.S. George, and B.D. Rao. "Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm". *Journal of Electroencephalography and Clinical Neurophysiology*, Oct. 1995.

[15] I.F. Gorodnitsky and B.D. Rao. "Sparse Signal Reconstructions from Limited Data using FOCUSS: A Re-weighted Minimum Norm Algorithm". *IEEE Trans. on Signal Processing*, 45:600–616, March 1997.

[16] D. L. Donoho. "Superresolution via Sparsity Constraints". *SIAM Journal Math Analysis*, 23:1309–1331, Sept. 1992.

[17] B. D. Rao and K. Kreutz-Delgado. "Deriving Algorithms for Computing Sparse Solutions to Linear Inverse Problems". In *Proc. of the 31st Asilomar Conference*, Nov. 1997.

[18] M. Hamalainen et al. "Magnetoencephalography: theory, instrumentation, and applications to noninvasive studies of the working human brain". *Rev. Mod. Physics*, April 1993.

[19] J. P. Wickswo Jr. "SQUID Magnetometers for Biomagnetism and Nondestructive Testing: Important Questions and Initial Answers". *IEEE Trans. on Applied Superconductivity*, Vol. 5(2):1–47, June 1995.

[20] B. Jeffs, R. Leahy, and M. Singh. "An evaluation of methods for neuromagnetic image reconstruction". *IEEE Biomed.*, BME-34:713–723, 1987.

[21] A. A. Ioannides, J. P. R. Bolton, and C. J. S. Clarke. "Continuous Probabilistic Solutions to the Biomagnetic Inverse problem". *Inverse Problems*, pages 523–542, 1990.

[22] A. M. Kondoz. *Digital Speech: Low bit rate Coding for Communication Systems*. Wiley, 1996.

[23] B. Kleijn and K. Paliwal Editors. *Speech Coding and Synthesis*. Elsevier Press, 1995.

[24] S. Singhal and B. S. Atal. "Amplitude Optimization and pitch prediction in multipulse coders". *Trans. ASSP*, Mar. 1989.

[25] H. Lee, D. P. Sullivan, and T. H. Huang. "Improvement of discrete band-limited signal extrapolation by iterative subspace modification". *Proc. ICASSP*, vol. 3, April 1987.

[26] S. D. Cabrera and T. W. Parks. "Extrapolation and Spectral Estimation with Iterative weighted norm modification". *IEEE Trans. on ASSP*, 39(4):842–851, April 1991.

[27] S. Chen and J. Wigger. "Fast Orthogonal Least Squares Algorithm for Efficient Subset Model Selection". *IEEE Trans. on Signal Processing*, 43(7):1713–1715, July 1995.

[28] B. K. Natarajan. "Sparse Approximate Solutions to Linear Systems". *SIAM Journal on Computing*, April 1995.

[29] R. E. Carlson and B. K. Natarajan. "Sparse Approximate Multiquadric Interpolation". *Computer Math and Applications*, 27(6):99–108, 1994.

[30] P. Duhamel and J. C. Rault. "Automatic Test Generation Techniques for Analog Circuits and Systems: A Review". *IEEE Trans. on Circuits and Systems*, July 1979.

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

[32] P. S. Bradley and O. L. Mangasarian. "Feature Selection via Mathematical Programming". *Univ. of Wisconsin Mathematical Programming technical report 95-21*.

[33] K. Kreutz-Delgado and B. D. Rao. "Measures and Algorithms for Best Basis Selection". In *Proc. ICASSP 1998*.

[34] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Macmillan, New York, 1994.