Basis Pursuit

Shaobing Chen
schen@playfair.stanford.edu

David Donoho
donoho@playfair.stanford.edu

Statistics Department
Stanford University
Stanford, CA 94305

Abstract

The Time-Frequency and Time-Scale communities have recently developed an enormous number of overcomplete signal dictionaries – wavelets, wavelet packets, cosine packets, wilson bases, chirplets, warped bases, and hyperbolic cross bases being a few examples. Basis Pursuit is a technique for decomposing a signal into an "optimal" superposition of dictionary elements. The optimization criterion is the l^1 norm of coefficients. The method has several advantages over Matching Pursuit and Best Ortho Basis, including super-resolution and stability.

1 Introduction

Over the last five years or so, there has been an explosion of awareness of alternatives to traditional signal representations. Instead of just representing objects as superpositions of sinusoids (the traditional Fourier representation) we now have available alternate dictionaries – signal representation schemes – of which the Wavelets dictionary is only the most well-known. Wavelet dictionaries, Gabor dictionaries, Multi-scale Gabor Dictionaries, Wavelet Packets, Cosine Packets, Chirplets, and a wide range of other representations are now available. Each such dictionary $\mathcal{D}$ is a collection of waveforms $(\phi_\gamma)_{\gamma\in\Gamma}$, and we envision a decomposition of a signal s as

$$s = \sum_{\gamma\in\Gamma} \alpha_\gamma \phi_\gamma, \qquad (1)$$

or an approximate decomposition

$$s = \sum_{\gamma\in\Gamma} \alpha_\gamma \phi_\gamma + R, \qquad (2)$$

where R is a residual. Depending on the dictionary, such a decomposition is a decomposition into pure tones (Fourier dictionary), bumps (wavelet dictionary), chirps (chirplet dictionary), etc.

A key point. The dictionaries we are interested in are all *overcomplete*, either because they start out that way, or because we can merge complete dictionaries, obtaining a new mega-dictionary consisting of several types of waveforms (e.g Fourier & Wavelets dictionaries along with Gabor). The decomposition (1) is then nonunique, because some elements in the dictionary have representations in terms of other elements. We assume such nonuniqueness in what follows. It gives us the possibility of adaptation, i.e. of choosing among many representations one which is most suited to our purposes.

2 Goals of Adaptive Representation

We are motivated by the aim of achieving simulatneously the following goals.

- *Speed.* It should be possible to obtain a representation in order $O(n)$ or $O(n\log(n))$ time.
- *Sparsity.* We should obtain the sparsest possible representation of the object – i.e. the one with the fewest significant coefficients.
- *Perfect separation.* When the signal is made up of a superposition of a few very disparate phenomena (e.g. impulses and sinusoids), those should be clearly separated and marked.
- *Superresolution.* We should obtain a resolution of sparse objects that is much higher-resolution than that possible with traditional non-adaptive approaches.
- *Stability.* Small perturbations of s should not seriously degrade the results.

1058-6393/95 $4.00

3 Finding an Adaptive Representation

We briefly mention a few methods that have been proposed to find decompositions and then comment on how well they achieve those goals. In my talk, I will give examples of these methods in action.

3.1 Method of Frames

Imagine that we write out all the vectors of the dictionary as columns of a matrix Φ, and we write out all the coefficients (α_γ) as a column vector. Then the decomposition problem is that of finding a solution $\Phi\alpha = s$. There are many solutions; the method of Frames [4] picks one whose coefficients have minimum l^2 norm.

$$\min \|\alpha\|_2 \text{ subject to } \Phi\alpha = s.$$

This solution can often be computed in $O(n \log(n))$ time. The key problem: the solution is an average of all possible solutions $\Phi\alpha = s$; so it is typically of very poor sparsity, and also does not super-resolve.

3.2 Best Ortho Basis

Coifman and Meyer have invented some special dictionaries, wavelet packets and cosine packets, which have a very special structure. Certain structured subcollections of the elements amount to orthogonal bases; one gets in this way a wide range of orthonormal bases (in fact $\gg 2^n$ such orthogonal bases for signals of length n). Coifman and Wickerhauser [1] have proposed a method of adaptively picking from among these many bases, a single orthogonal basis which is the best one. If $(s[\mathcal{B}]_I)$ denotes the vector of coefficients of s in orthogonal basis $\mathcal{B}$, and if we define the "entropy" $\mathcal{E}(s[\mathcal{B}]) = \sum_I e(s[\mathcal{B}]_I)$, where $e(s)$ is a scalar function of a scalar argument, they give a fast algorithm for solving

$$\min\{\mathcal{E}(s[\mathcal{B}]) : \quad \mathcal{B} \text{ ortho basis} \subset \mathcal{D}\}.$$

The algorithm is fast – it delivers a basis in order $n \log(n)$ time – and in some cases delivers near-optimal sparsity respresentations. Possible problem: when the signal is composed of a very few highly non-orthogonal components, the method may not deliver sparse representations.

3.3 Matching Pursuit

Mallat and Zhang [5] have proposed the use of a greedy algorithm which builds up a sequence of sparse approximations starting from $s^{(0)} = 0$ and $R^{(0)} = s$; adding to the approximation at stage k that multiple of that element of the dictionary which best correlates with the residual, so that $s^{(k)} = s^{(k-1)} + \alpha_k \phi_{\gamma_k}$ and $R^{(k)} = s - s^{(k)}$. After N steps, one has a representation of the form (2), with residual $R = R^{(N)}$.

Possible problem: because the algorithm is greedy, when run for many iterations, it might spending most of its time correcting for any mistakes made in the first few terms. One can give examples of dictionaries and signals where the method gives a solution which is badly sub-optimal in terms of sparsity.

3.4 Massive Optimization

To avoid the limitations of greedy optimization, one might consider replacing matching pursuit by a true global optimization employing exhaustive enumeration. One would sift through all possible subsets of size $\leq n$ of the dictionary and fit each subset to the signal by least squares. That subset optimizing a tradeoff of complexity and lack-of-fit would be chosen to generate the fit $\hat{s}$ and residual. The key problem: sifting through $\gg 2^n$ least-squares fits is not computationally feasible.

4 Basis Pursuit

Chen and Donoho [2] have suggested a method of decomposition based on a true global optimization which is at least theoretically feasible, due to recent advances in linear programming. Among the many possible solutions to $\Phi\alpha = s$, they pick one whose coefficients have minimum l^1 norm.

$$\min \|\alpha\|_1 \text{ subject to } \Phi\alpha = s. \tag{3}$$

For dealing with data at noise level $\sigma > 0$, they propose approximate decomposition as in (2), solving

$$\min \|\Phi\alpha - s\|_2^2 + \lambda_n \|\alpha\|_1, \tag{4}$$

with $\lambda_n = \sigma\sqrt{2\log(\#\mathcal{D})}$ depending on the number $\#\mathcal{D}$ of distinct vectors in the dictionary.

4.1 Comparisons

In comparison to the method of frames, the l^2 norm is replaced by the l^1 norm. Special properties of the l^1 norm force Basis Pursuit to be nonlinear and to therefore exhibit potentially very different behavior than the method of frames.

The comparison to the Best Ortho Basis method is interesting in the case (4). When Best Ortho Basis is run with the l^1 norm as an entropy, the two methods compare as follows: BOB finds the *orthogonal* basis which optimizes the l^1 norm, while BP finds the optimum l^1 norm among *all bases*, not just orthogonal ones.

An understanding of this last comment comes from noting that problem (3) is equivalent to a linear programming problem. From the theory of linear programming, we know that a solution is obtained at a basis. We also know that linear programming solutions tend to be sparse, and this helps us understand why BP may tend to give sparse solutions.

In comparison to Matching Pursuit, suppose we solve the linear program underlying BP via the simplex method. Then MP works by starting with an empty model, building up a new model in greedy fashion term by term. BP starts from an initial basis (for example, the Best Ortho Basis for l^1-entropy) and iteratively improves the basis by swapping atoms not in the basis for atoms in the basis. Both algorithms are greedy, but the theory of linear programming says that simplex must converge to a global optimum; in contrast, global optimality of MP is not guaranteed.

4.2 Computation

BP is only thinkable because of recent advances in linear programming via "interior point" methods. Indeed, suppose we need to do BP on an 8192 point long signal, and decompose the object in the Wavelet Packets Dictionary. The dictionary will have $n \log_2(n) = 8192*13 = 106496$ elements. The corresponding linear program has 8192 constraints and more than 212,992 variables. Moreover the matrix of the linear program is not sparse.

Until ten years ago, it was a major effort to solve a linear program with more than 1000 constraints and 1000 variables. Today, problems with more than 50,000 variables and constraints are being solved in practical work.

This increase in size of problems treated is in large part due to the explosion of interest following Karmarkar's work on interior point methods. Modern interior point methods have evolved far beyond Karmarkar's original proposal; we are using primal-dual log-barrier methods. The key point of such methods is the solution of a system of equations $ADA^T = A^T v$ about ten or twenty times, where the diagonal matrix D changes from iteration to iteration, and A is the matrix in the definition of the linear program. Our approach is oriented around the special properties of A and A^T in BP. We use pre-conditioned conjugate gradient solvers so that we never have to store the matrix A of the underlying linear program; we only have to know how to apply A and A^T rapidly. This is possible since our signal dictionaries possess fast transforms.

4.3 Example

Figure 1 shows the use of BP, Frames, and MP to super-resolve a superposition of two sinusoids with the two frequencies spaced closer than the Rayleigh distance. The signal-to-noise ratio is 5/1. BP resolves the presence of two specific frequencies; the other methods do not. Figure 2 shows the use of BP and BOB to de-noise a signal at 1/1 signal/noise ratio. Both methods are working from the same dictionary. BOB is applied as in [3]. BP is evidently both accurate and stable. Figure 3 shows a phase plane obtained by analyzing the signal "Greasy" [5] using BOB with l^1 entropy and wavelet packet dictionary. Figure 4 shows the phase palne from BP. Note the enhanced resolution.

Acknowledgements

Michael Saunders, Department of Operations Research, Stanford University provided considerable advice, software, and wisdom about large-scale optimization problems.

This research was supported by NSF DMS-92-09130 and by the NASA Astrophysical Data Program.

References

[1] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best-basis selection", *IEEE Trans. Info. Theory* **38** (1992) p. 713-718.

[2] S. Chen and D. Donoho. *Basis Pursuit.* Technical Report, Department of Statistics, Stanford University.

[3] D.L. Donoho and I.M. Johnstone, *Ideal Time-Frequency Denoising.* Technical Report, Department of Statistics, Stanford University.

[4] I. Daubechies, *Ten Lectures on Wavelets,* SIAM, (1992).

[5] S. Mallat and Z. Zhang. Matching Pursuit with Time Frequency Dictionaries. *IEEE Trans. Sign. Proc.* (1993).

eject

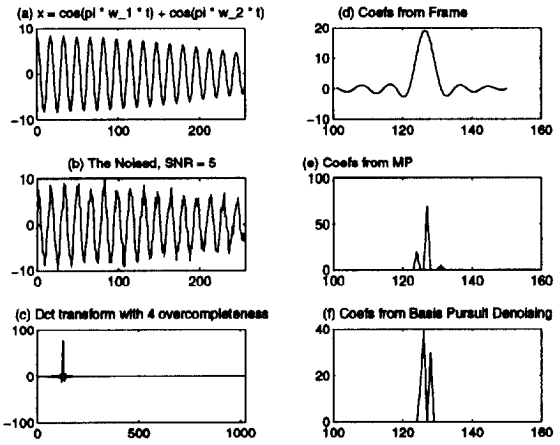

Figure 1: Super-resolve Two Close Cosines

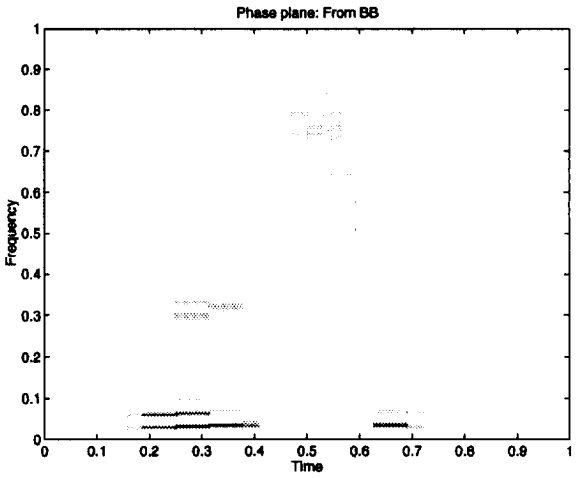

Figure 3: BOB on Greasy using Wavelet Packet

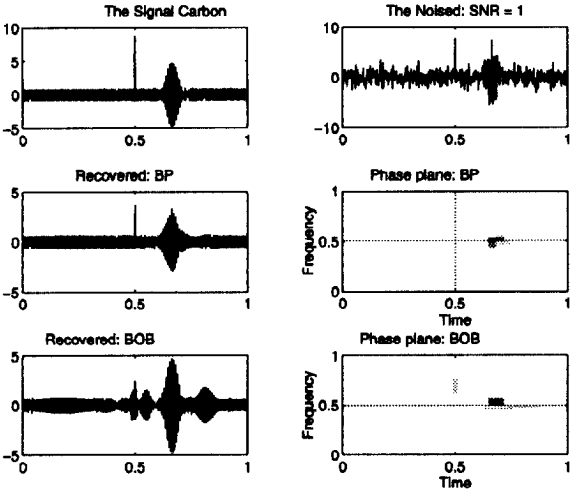

Figure 2: Denoising on Carbon

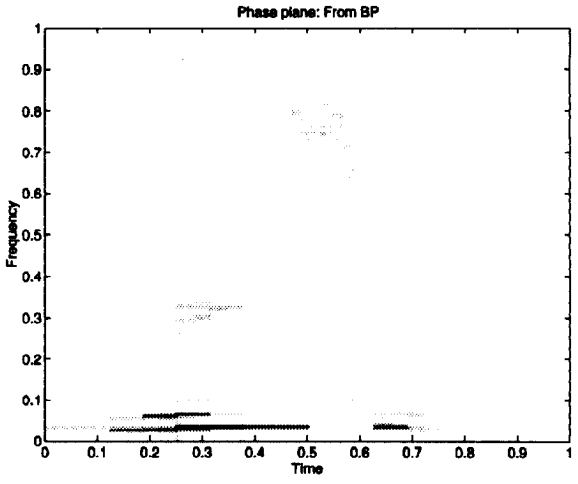

Figure 4: BP on Greasy using Wavelet Packet