Adaptive Time-Frequency Decompositions with Matching Pursuits

Geoffrey Davis, Stéphane Mallat, and Zhifeng Zhang

New York University, Courant Institute
251 Mercer Street, New York, NY 10012

Abstract

Computing the optimal expansion of a signal in a redundant dictionary of waveforms is an NP-hard problem. We introduce a greedy algorithm, called a matching pursuit, which computes a sub-optimal expansion. The dictionary waveforms which best match a signal's structures are chosen iteratively. An orthogonalized version of the matching pursuit is also developed. Matching pursuits are general procedures for computing adaptive signal representations. With a dictionary of Gabor functions, a matching pursuit defines an adaptive time-frequency transform. We derive a signal energy distribution in the time-frequency plane which does not contain interference terms, unlike the Wigner and Cohen class distributions. Matching pursuits are chaotic maps whose attractors define a generic noise with respect to the dictionary. We derive an algorithm that isolates the coherent structures of a signal and describe an application to pattern extraction from noisy signals.

1 Introduction

Flexible decompositions are particularly important for representing signal components whose localizations in time and frequency vary widely. The goal is to expand a signal into waveforms whose time-frequency properties are adapted to the signal's local structures. The waveforms we use for these expansions are called time-frequency atoms. For example, impulses need to be decomposed into functions well-localized in time, while spectral lines are better represented by waveforms which have a narrow frequency support. When the signal includes both of these elements, the time-frequency atoms must be adapted accordingly. We must use a procedure that selects from all the time-frequency atoms of a large dictionary the waveforms that are best adapted to decomposing the signal structures.

Computing the optimal approximation of a signal in a redundant dictionary is an NP-hard problem. We therefore introduce a sub-optimal greedy algorithm, called a matching pursuit, which decomposes any signal into a linear expansion of waveforms belonging to a redundant collection called a dictionary. These waveforms are selected in order to best match the signal's structures. Although matching pursuits are non-linear, they possess an energy conservation relation like an orthogonal expansion which guarantees their convergence. We also introduce an orthogonalized version of a matching pursuit.

The application of matching pursuits to adaptive time-frequency decompositions is described in section 6. The signal is decomposed into a selected set of time-frequency atoms, the dilations, translations, and modulations of a single window function. We derive a time-frequency energy distribution by adding the Wigner distributions of the selected time-frequency atoms. Unlike the Wigner distribution or Cohen's class distributions, this energy distribution does not include interference terms and thus provides a clear picture in the time-frequency plane.

Matching pursuits are chaotic maps whose properties are studied. The approximation error of the pursuit converges to an attractor which corresponds to a class of signals which are not efficiently represented by the waveforms of the dictionary. Measurement of this convergence to the attractor allows us to separate our decomposition into portions which are coherent and incoherent with respect to the dictionary. Isolating the coherent part of a signal enables us to perform denoising.

Notation

The space $\mathbf{L}^2(\mathbf{R})$ is the Hilbert space of complex valued functions such that

$$\|f\|^2 = \int_{-\infty}^{+\infty} |f(t)|^2 dt < +\infty. \tag{1}$$

The inner product of $f(t), g(t) \in \mathbf{L}^2(\mathbf{R})$ is defined by

$$< f, g >= \int_{-\infty}^{+\infty} f(t)\bar{g}(t)dt, \tag{2}$$

where $\bar{g}(t)$ is the complex conjugate of $g(t)$. The Fourier transform of $f(t) \in \mathbf{L}^2(\mathbf{R})$ is written $\hat{f}(\omega)$ and defined by

$$\hat{f}(\omega) = \int_{-\infty}^{+\infty} f(t)e^{-i\omega t}dt. \tag{3}$$

2 Time-Frequency Atoms

A general family of time-frequency atoms can be generated by scaling, translating and modulating a single window function $g(t) \in \mathbf{L}^2(\mathbf{R})$. We suppose that $g(t)$ is real and centered at 0. We also impose that $\|g\| = 1$, that the integral of $g(t)$ is non-zero, and that $g(0) \neq 0$. For any scale $s > 0$, frequency modulation ξ, and translation u, we denote $\gamma = (s, u, \xi)$ and define

$$g_\gamma(t) = \frac{1}{\sqrt{s}} g(\frac{t-u}{s}) e^{i\xi t}. \tag{4}$$

The index γ is an element of the set $\mathbf{\Gamma} = \mathbf{R}^+ \times \mathbf{R^2}$. The factor $\frac{1}{\sqrt{s}}$ normalizes $\|g_\gamma(t)\|$ to 1. The function $g_\gamma(t)$ is centered at the abscissa u and its energy is concentrated in a neighborhood of u of size proportional to s. Its Fourier transform is centered at the frequency $\omega = \xi$ and has its energy concentrated in a neighborhood of ξ, of size proportional to $1/s$. For our numerical examples we use the Gaussian window $g(t) = 2^{1/4}e^{-\pi t^2}$.

The dictionary of time-frequency atoms $\mathcal{D} = (g_\gamma(t))_{\gamma \in \mathbf{\Gamma}}$ is a very redundant set of functions that includes window Fourier frames and wavelet frames [3]. When the signals include time-frequency structures of very different types, one cannot choose *a priori* a frame that is well adapted to performing the expansion. Instead, we need to find the atoms in the dictionary that best match each given signal's structures in order to perform a compact decomposition. In the next section we develop an algorithm for computing such adaptive decompositions in redundant dictionaries.

3 Matching Pursuits

Let $\mathbf{H}$ be a signal space. A dictionary for $\mathbf{H}$ is a family $\mathcal{D} = (g_\gamma)_{\gamma \in \mathbf{\Gamma}}$ of vectors in $\mathbf{H}$, such that linear combinations of the g_γ are dense in $\mathbf{H}$ and for which $\|g_\gamma\| = 1$. The smallest possible dictionary is a basis of $\mathbf{H}$; general dictionaries are redundant families of vectors. A signal does not have a unique representation as a sum of elements of a redundant dictionary. Unlike the case of a basis, we have some degrees of freedom in choosing a signal's particular representation. This freedom allows us to choose a subset of the dictionary that is tailored to the signal in question and which provides the most compact representation. We can choose a subset of the dictionary for which the signal energy is concentrated in as few terms as possible. The chosen vectors highlight the predominant signal features.

Let $\mathcal{D}$ be dictionary of vectors in an N-dimensional Hilbert space. For any given $\beta \in (0,1)$, we define an optimal approximation of $f \in \mathbf{H}$ to be an expansion

$$\tilde{f} = \sum_{n=1}^{\beta N} a_n g_{\gamma_n},$$

where the a_n and $g_{\gamma_n} \in \mathcal{D}$ are chosen in order to minimize

$$\|f - \tilde{f}\|.$$

When the dictionary is redundant, we can show that finding an optimal solution is a fundamentally intractable problem. If we restrict the number of bits of a_n to $\Theta(N^j)$ and the number of vectors in $\mathcal{D}$ to $\Theta(N^k)$, for fixed j, k, then we can prove [4] that finding an optimal expansion is NP-hard.

Because of the difficulty of finding optimal solutions, we instead develop a greedy algorithm that computes a good sub-optimal approximation. Let $f \in \mathbf{H}$. We want to compute a linear expansion of f over a set of vectors selected from $\mathcal{D}$ which best matches the inner structures of f. A matching pursuit is a greedy algorithm which successively approximates f with orthogonal projections onto elements of $\mathcal{D}$. Let $g_{\gamma_0} \in \mathcal{D}$. The vector f can be decomposed into

$$f = < f, g_{\gamma_0} > g_{\gamma_0} + Rf, \tag{5}$$

where Rf is the residual vector after approximating f in the direction of g_{γ_0}. Clearly g_{γ_0} is orthogonal to Rf, hence

$$\|f\|^2 = |< f, g_{\gamma_0} >|^2 + \|Rf\|^2. \tag{6}$$

To minimize $\|Rf\|$, we must choose $g_{\gamma_0} \in \mathcal{D}$ such that $|< f, g_{\gamma_0} >|$ is maximal. In some cases, it is only possible to find a vector g_{γ_0} that is close to the maximum in the sense that

$$|< f, g_{\gamma_0} >| \geq \alpha \sup_{\gamma \in \mathbf{\Gamma}} |< f, g_\gamma >|, \tag{7}$$

where $\alpha \in (0,1]$ is an optimality factor.

We sub-decompose the residue Rf by projecting it onto the vector of $\mathcal{D}$ that best matches Rf, as was done for f. This projection of Rf generates a second residue, $R^2 f$, which we again decompose to obtain a third residue, and so on.

We describe the algorithm inductively. Let $R^0 f = f$. We suppose that we have computed the n^{th} order residue $R^n f$, for $n \geq 0$. We choose with a choice function C an element $g_{\gamma_n} \in \mathcal{D}$ which closely matches the residue $R^n f$ in the sense that

$$| < R^n f, g_{\gamma_n} > | \geq \alpha \sup_{\gamma \in \mathbf{\Gamma}} | < R^n f, g_\gamma > |. \tag{8}$$

The residue $R^n f$ is sub-decomposed into

$$R^n f = < R^n f, g_{\gamma_n} > g_{\gamma_n} + R^{n+1} f, \tag{9}$$

which defines the residue at order n+1. Since $R^{n+1} f$ is orthogonal to g_{γ_n}, we have

$$\|R^n f\|^2 = | < R^n f, g_{\gamma_n} > |^2 + \|R^{n+1} f\|^2. \tag{10}$$

Let us carry this decomposition up to order m. We decompose f into the telescoping sum

$$f = \sum_{n=0}^{m-1} \left(R^n f - R^{n+1} f \right) + R^m f. \tag{11}$$

Equation (9) yields

$$f = \sum_{n=0}^{m-1} < R^n f, g_{\gamma_n} > g_{\gamma_n} + R^m f. \tag{12}$$

Similarly, we write $\|f\|^2$ as a telescoping sum

$$\|f\|^2 = \sum_{n=0}^{m-1} \left(\|R^n f\|^2 - \|R^{n+1} f\|^2 \right) + \|R^m f\|^2 \tag{13}$$

which we combine with (10) to obtain an energy conservation equation

$$\|f\|^2 = \sum_{n=0}^{m-1} | < R^n f, g_{\gamma_n} > |^2 + \|R^m f\|^2. \tag{14}$$

Thus, the original vector f is decomposed into a sum of dictionary elements which are chosen to best match its residues. Although this decomposition is non-linear, we maintain an energy conservation as though it were a linear, orthogonal decomposition. An important issue is to understand the behavior of the residue $R^m f$ when m increases. By adapting a result proved by Jones [9] for projection pursuit algorithms [5], one can prove [10] that the matching pursuit algorithm converges, even in infinite dimensional spaces.

Theorem 1 *Let $f \in \mathbf{H}$. The residue $R^m f$ defined by the induction equation (9) satisfies*

$$\lim_{m \to +\infty} \|R^m f\| = 0. \tag{15}$$

Hence

$$f = \sum_{n=0}^{+\infty} < R^n f, g_{\gamma_n} > g_{\gamma_n}, \tag{16}$$

and

$$\|f\|^2 = \sum_{n=0}^{+\infty} |< R^n f, g_{\gamma_n} >|^2. \tag{17}$$

When $\mathbf{H}$ is of finite dimension, $\|R^m f\|$ decays exponentially to zero.

4 Implementation of Matching Pursuits

When the dictionary is very redundant, the search for the vectors that best match the signal residues can be limited to a sub-dictionary $\mathcal{D}_\alpha = (g_\gamma)_{\gamma \in \mathbf{\Gamma}_\alpha} \subset \mathcal{D}$. We suppose that $\mathbf{\Gamma}_\alpha$ is a finite set of indices from $\mathbf{\Gamma}$ such that for any $f \in \mathbf{H}$

$$\sup_{\gamma \in \mathbf{\Gamma}_\alpha} |< f, g_\gamma >| \geq \alpha \sup_{\gamma \in \mathbf{\Gamma}} |< f, g_\gamma >|. \tag{18}$$

Depending upon α and the dictionary redundancy, the set $\mathbf{\Gamma}_\alpha$ can be much smaller than $\mathbf{\Gamma}$. The matching pursuit is initialized by computing the inner products $(< f, g_\gamma >)_{\gamma \in \mathbf{\Gamma}_\alpha}$, and continues by induction as follows. Suppose that we have already computed $(< R^n f, g_\gamma >)_{\gamma \in \mathbf{\Gamma}_\alpha}$, for $n \geq 0$. We search in $\mathcal{D}_\alpha$ for an element $g_{\tilde{\gamma}_n}$ for which

$$|< R^n f, g_{\tilde{\gamma}_n} >| = \max_{\gamma \in \mathbf{\Gamma}_\alpha} |< R^n f, g_\gamma >|. \tag{19}$$

We can find a dictionary element that matches f even better than $g_{\tilde{\gamma}_n}$ by using Newton's method to maximize $|< f, g_\gamma >|$ for $\gamma \in \mathbf{\Gamma}$ in a neighborhood of $g_{\tilde{\gamma}_n}$. We then have

$$|< R^n f, g_{\gamma_n} >| \geq |< R^n f, g_{\tilde{\gamma}_n} >| \geq \alpha \sup_{\gamma \in \mathbf{\Gamma}} |< R^n f, g_\gamma >|. \tag{20}$$

The choice function mentioned in section 3 is defined indirectly by this double search strategy. Once the vector g_{γ_n} is selected, we compute the inner product of the new residue $R^{n+1} f$ with any $g_\gamma \in \mathcal{D}_\alpha$, with an updating formula derived from equation (9)

$$< R^{n+1} f, g_\gamma > = < R^n f, g_\gamma > - < R^n f, g_{\gamma_n} > < g_{\gamma_n}, g_\gamma > . \tag{21}$$

Since we have already computed $< R^n f, g_\gamma >$ and $< R^n f, g_{\gamma_n} >$, this update requires only that we compute $< g_{\gamma_n}, g_\gamma >$. Dictionaries are generally built so that this inner product is computed with a small number of operations. We describe in [10] how to compute efficiently the inner product of two discrete Gabor atoms, in $O(1)$ operations.

Let I be the number of operations required to compute $< g_{\gamma_n}, g_\gamma >$ and let Z be the average number of g_γ's in $\mathcal{D}_\alpha$ for which $< g_{\gamma_n}, g_\gamma >$ is nonzero. In [4] we show that p iterations of a pursuit requires $O(pIZ)$ operations.

The number of times we sub-decompose the residues of a given signal f depends upon the desired precision of the approximation, ϵ. The number of iterations we require is the minimum p for which

$$\|R^p f\| = \|f - \sum_{n=0}^{p-1} < R^n f, g_{\gamma_n} > g_{\gamma_n}\| \leq \epsilon \|f\|. \tag{22}$$

From the energy conservation relation (14), we see that

$$\|f\|^2 - \sum_{n=0}^{p-1} |< R^n f, g_{\gamma_n} >|^2 \leq \epsilon^2 \|f\|^2. \tag{23}$$

Since we do not compute the residue $R^n f$ at each iteration, we test condition (23) to determine when to stop the decomposition.

5 Back-projection and Orthogonal Pursuits

After m iterations, a matching pursuit decomposes a signal f into a sum of m dictionary vectors and an error term. We have

$$f = \sum_{n=0}^{m-1} < R^n f, g_{\gamma_n} > g_{\gamma_n} + R^m f. \tag{24}$$

Suppose that $< g_{\gamma_n}, g_{\gamma_{n+1}} >= \beta \neq 0$. $R^{n+1} f$ is obtained by removing the component of $R^n f$ in the direction of g_{γ_n},

$$R^{n+1} f = R^n f - < R^n f, g_{\gamma_n} > g_{\gamma_n}. \tag{25}$$

Similarly, $R^{n+2} f$ is obtained from $R^{n+1} f$ by removing the component of $R^{n+1} f$ in the direction of $g_{\gamma_{n+1}}$. Because $< g_{\gamma_n}, g_{\gamma_{n+1}} >\neq 0$, we find that the component of $R^{n+2} f$ in the g_{γ_n} direction is no longer zero.

$$\begin{aligned} < R^{n+2} f, g_{\gamma_n} > &= < R^{n+1} f, g_{\gamma_n} > - < R^{n+1} f, g_{\gamma_{n+1}} >< g_{\gamma_{n+1}}, g_{\gamma_n} > \\ &= -\beta < R^{n+1} f, g_{\gamma_{n+1}} > . \end{aligned} \tag{26}$$

In removing $g_{\gamma_{n+1}}$, we have replaced a small portion of the g_{γ_n} component which we previously removed.

Let $\mathbf{V}_m$ be the space generated by $(g_{\gamma_n})_{0 \leq n < m}$ and $\mathbf{P}_{\mathbf{V}_m}$ be the orthogonal projector onto $\mathbf{V}_m$. For any $f \in \mathbf{H}$, $\mathbf{P}_{\mathbf{V}_m} f$ is the closest vector to f that can be written as linear expansion of the m vectors $(g_{\gamma_n})_{0 \leq n < m}$. We obtain from (24) that

$$\mathbf{P}_{\mathbf{V}_m} f = \sum_{n=0}^{m-1} < R^n f, g_{\gamma_n} > g_{\gamma_n} + \mathbf{P}_{\mathbf{V}_m} R^m f. \tag{27}$$

When the family of vectors $(g_{\gamma_n})_{0 \leq n < m}$ is not orthogonal, which is generally the case, then $\mathbf{P}_{\mathbf{V}_m} R^m f \neq 0$. The computation of

$$\mathbf{P}_{\mathbf{V}_m} R^m f = \sum_{n=0}^{m-1} x_n g_{\gamma_n}, \tag{28}$$

is called a back-projection. The values x_n give corrections to the coefficients $< R^n f, g_{\gamma_n} >$ which improve the linear expansion approximation to f. The approximation error for the corrected sum,

$$\mathbf{P}_{\mathbf{W}_m} f = f - \mathbf{P}_{\mathbf{V}_m} f \tag{29}$$

is the orthogonal projection of f on the space $\mathbf{W}_m$, the orthogonal complement of $\mathbf{V}_m$ in $\mathbf{H}$.

The calculation of the coefficients $(x_n)_{0 \leq n < m}$ requires that we solve the following linear system. For any g_{γ_k}, $0 \leq k < m$,

$$< \mathbf{P}_{\mathbf{V}_m} R^m f, g_{\gamma_k} > = < R^m f, g_{\gamma_k} > = \sum_{n=0}^{m-1} x_n < g_{\gamma_n}, g_{\gamma_k} > . \tag{30}$$

Let us denote $X = (x_n)_{0 \leq n < m}$ and $Y = (< R^m f, g_{\gamma_k} >)_{0 \leq k < m}$. Let $G = (< g_{\gamma_n}, g_{\gamma_k} >)_{0 \leq k < m, 0 \leq n < m}$ be the Gram matrix of the selected vectors. The linear system of equations (30) can be written $Y = GX$. A solution of this system is computed efficiently with a conjugate gradient algorithm [10]. If $\mathbf{H}$ is of finite dimension N, there are many classes of dictionaries for which any collection of N distinct dictionary vectors is a basis of $\mathbf{H}$. This is the case for the Gabor dictionary used for time-frequency decompositions. Hence, after selecting N different vectors with a matching pursuit, the back-projection reduces to 0 the remaining residue.

Instead of recovering the orthogonal projection $\mathbf{P}_{\mathbf{V}_m} f$ at the end of the matching pursuit, we can modify the pursuit algorithm to prevent the replacement of components which were previously removed, which we saw in (26). We accomplish this by orthogonalizing the set of dictionary vectors as we proceed with the decomposition. This iterative orthogonalization is equivalent to performing the back-projection described above at each step of the decomposition, but is much more efficient. This type of algorithm was first introduced for control applications [1] and also studied independently from this work by Pati et al. [11]. It has the advantage of providing better approximations than the matching pursuit algorithm, but it requires much more computation and can introduce numerical instabilities into the expansions. We describe by induction this orthogonal pursuit.

For $n = 0$, we set $R^0 f = f$. Like in a matching pursuit, we define an optimality factor α, with $0 < \alpha \leq 1$, and choose $g_{\gamma_0} \in \mathcal{D}$ which satisfies

$$| < f, g_{\gamma_0} > | \geq \alpha \sup_{\gamma \in \mathbf{\Gamma}} | < f, g_\gamma > |. \tag{31}$$

The space $\mathbf{V}_1$ is generated by the single vector g_{γ_0}. We use a Gram-Schmidt orthogonalization to generate a basis for $\mathbf{V}_1$. The first orthogonal basis vector for $\mathbf{V}_1$ is $u_0 = g_{\gamma_0}$. The next residue is defined by

$$Rf = f - \mathbf{P}_{\mathbf{V}_1} f = f - < f, g_{\gamma_0} > g_{\gamma_0}. \tag{32}$$

We explain by induction how to compute the orthogonal residue $R^{n+1}f$ from $R^n f$. We suppose that we have already selected n vectors $(g_{\gamma_p})_{0\leq p<n}$ that are linearly independent and that we have computed the corresponding Gram-Schmidt orthogonal basis $(u_p)_{0\leq p<n}$ of the space $\mathbf{V}_n$ spanned by these n vectors. We have

$$R^n f = f - \mathbf{P}_{\mathbf{V}_n} f. \tag{33}$$

We choose a vector $g_{\gamma_n} \in \mathcal{D}$ which satisfies

$$|< R^n f, g_{\gamma_n} >| \geq \alpha \sup_{\gamma\in\mathbf{\Gamma}} |< R^n f, g_\gamma >|. \tag{34}$$

If $< R^n f, g_{\gamma_n} > \neq 0$, then the vector g_{γ_n} cannot belong to the space $\mathbf{V}_n$ since $R^n f$ is orthogonal to $\mathbf{V}_n$. Hence the vectors $(g_{\gamma_p})_{0\leq p\leq n}$ are linearly independent. The next vector u_n of the Gram-Schmidt basis is obtained by subtracting from g_{γ_n} its projection on the space $\mathbf{V}_n$

$$u_n = g_{\gamma_n} - \sum_{p=0}^{n-1} \frac{< g_{\gamma_n}, u_p >}{\|u_p\|^2} u_p. \tag{35}$$

The family $(u_p)_{0\leq p\leq n}$ is an orthogonal basis of $\mathbf{V}_{n+1}$. The residue $R^{n+1}f$ is defined by

$$R^{n+1} f = f - \mathbf{P}_{\mathbf{V}_{n+1}} f = f - \sum_{p=0}^{n} \frac{< f, u_p >}{\|u_p\|^2} u_p. \tag{36}$$

This can also be rewritten

$$R^{n+1} f = R^n f - \frac{< R^n f, u_n >}{\|u_n\|^2} u_n. \tag{37}$$

Since $R^n f$ is orthogonal to the vectors $(g_{\gamma_p})_{0\leq p<n}$, equation (35) implies that $< R^n f, u_n >=< R^n f, g_{\gamma_n} >$ and thus

$$R^{n+1} f = R^n f - \frac{< R^n f, g_{\gamma_n} >}{\|u_n\|^2} u_n. \tag{38}$$

This equation is similar to the residue updating equation (9) of a matching pursuit, but instead of subtracting a vector in the direction of g_{γ_n}, we subtract a component in a direction orthogonal to all vectors previously selected. Since $R^{n+1}f$ and u_n are orthogonal,

$$\|R^{n+1} f\|^2 = \|R^n f\|^2 - \frac{|< R^n f, g_{\gamma_n} >|^2}{\|u_n\|^2}. \tag{39}$$

An orthogonal pursuit guarantees that the selected vectors $(g_{\gamma_n})_{0\leq n\leq m}$ are linearly independent, and computes the best possible approximation of f from these vectors. Since $R^0 f = f$, we derive from equations (38) and (39) that for any $m > 0$

$$f = \sum_{0\leq n<m} \frac{< R^n f, g_{\gamma_n} >}{\|u_n\|^2} u_n + R^m f, \tag{40}$$

and

$$\|f\|^2 = \sum_{0 \le n < m} \frac{| < R^n f, g_{\gamma_n} > |^2}{\|u_n\|^2} + \|R^m f\|^2. \tag{41}$$

The derivations are similar to those for equations (12) and (14). The next theorem is similar to Theorem 1 and guarantees the convergence of the orthogonal pursuit [4].

Theorem 2 *Let $f \in \mathbf{H}$. Let N be the dimension of $\mathbf{H}$ (N may be infinite). The orthogonal matching pursuit converges in $M \le N$ iterations (M may be infinite if N is infinite). The residue $R^n f$ defined inductively by equation (38) satisfies*

$$\lim_{n \to M} \|R^n f\| = 0, \tag{42}$$

$$f = \sum_{0 \le n < M} \frac{< R^n f, g_{\gamma_n} >}{\|u_n\|^2} u_n, \tag{43}$$

and

$$\|f\|^2 = \sum_{0 \le n < M} \frac{| < R^n f, g_{\gamma_n} > |^2}{\|u_n\|^2}. \tag{44}$$

If $\mathbf{H}$ is of finite dimension, the orthogonal pursuit converges in a finite number of iterations.

Although the orthogonalized version of the pursuit has better convergence properties, it has several disadvantages. Because the updating relation (38) involves the orthogonalized dictionary vectors, u_n, instead of the g_{γ_n}, updating the inner products $< R^n f, g_\gamma >$ requires much more work. In [4] we show that computing p iterations of an orthogonalized pursuit requires $O(p^2 IZ)$ operations, as compared to $O(pIZ)$ for a non-orthogonal pursuit. Here I is the number of operations required to compute $< g_{\gamma_n}, g_\gamma >$ and Z is the average number of g_γ's in $\mathcal{D}_\alpha$ for which $< g_{\gamma_n}, g_\gamma >$ is nonzero. So for p iterations, the orthogonalized pursuit is $O(p)$ times slower. Unless we are interested in decomposing signals into a very small number of elements, the orthogonalized pursuit will be much slower.

A second disadvantage of the orthogonalized pursuit is that we obtain an expansion of our signal in the orthogonalized basis u_n rather than in the g_{γ_n}'s. The process of transforming the sum of the u_n's into a sum of g_{γ_n}'s can introduce instabilities into the expansion. We are looking for coefficients $(\beta_n)_{0 \le n < M}$ such that

$$f = \sum_{0 \le n < M} \beta_n g_{\gamma_n}. \tag{45}$$

Since $u_n \in \mathbf{V}_n$ and $(g_{\gamma_p})_{0 \le p \le n}$ is a basis of $\mathbf{V}_n$, we can decompose u_n into

$$u_n = \sum_{p=0}^{n} b_{p,n} g_{\gamma_p}. \tag{46}$$

These coefficients are computed during the pursuit. Inserting expression (46) into (43) yields

$$f = \sum_{0 \le n < M} \frac{< R^n f, g_{\gamma_n} >}{\|u_n\|^2} \sum_{p=0}^{n} b_{p,n} g_{\gamma_p}. \tag{47}$$

One could naively try to rearrange the terms of this double summation to obtain

$$f = \sum_{0 \leq p < M} g_{\gamma_p} \sum_{p \leq n < M} b_{p,n} \frac{< R^n f, g_{\gamma_n} >}{\|u_n\|^2}. \tag{48}$$

However, when $M = +\infty$ the infinite sum over n that defines each coefficient β_p may not converge. Such a situation arises when the family $(g_{\gamma_n})_{0 \leq n < M}$ is not a Riesz basis of the closed space $\mathbf{V}_M$ that it generates. For such a case, we cannot obtain an expansion of the form (45) from the orthogonal matching pursuit. If the signal space $\mathbf{H}$ has a finite dimension N, then M is finite, so we can always invert the two sums of (47) to obtain (48). The basis $(g_{\gamma_n})_{0 \leq n < M}$ may, however, be very badly conditioned in which case we can have numerical instabilities, in which case

$$\sum_{0 \leq n < M} |\beta_n|^2 >> \|f\|^2. \tag{49}$$

We find that the orthogonalized pursuit converges more quickly, but is much harder to compute and potentially less numerically stable than the non-orthogonalized pursuit. Figure 1 compares the convergence rates for an orthogonal matching pursuit and a non-orthogonal matching pursuit for a synthetic signal of 512 samples. The benefits of the orthogonalization do not become evident until after roughly 150 iterations, but in the tail of the expansion the convergence of the orthogonalized pursuit is much faster. For our denoising algorithm, we are uninterested in the tail of the expansion, so for such an application the orthogonalized pursuit gives little advantage.

6 Matching Pursuits With Time-Frequency Dictionaries

For dictionaries of time-frequency atoms, a matching pursuit yields an adaptive time-frequency transform. It decomposes any function $f(t) \in \mathbf{L}^2(\mathbf{R})$ into a sum of complex time-frequency atoms that best match its residues. This section studies the properties of this particular matching pursuit decomposition. We derive a new type of time-frequency energy distribution by summing the Wigner distributions of each time-frequency atom.

Since a time-frequency atom dictionary is complete in $\mathbf{L}^2(\mathbf{R})$, Theorem 1 implies that a matching pursuit decomposes any function $f \in \mathbf{L}^2(\mathbf{R})$ into

$$f = \sum_{n=0}^{+\infty} < R^n f, g_{\gamma_n} > g_{\gamma_n}, \tag{50}$$

where $\gamma_n = (s_n, u_n, \xi_n)$ and

$$g_{\gamma_n}(t) = \frac{1}{\sqrt{s_n}} g(\frac{t - u_n}{s_n}) e^{i \xi_n t}. \tag{51}$$

These atoms are chosen to best match the residues of f.

We derive a new time-frequency energy distribution from the decomposition of a function $f(t)$ within a time-frequency dictionary by adding the Wigner distributions of each selected atom. The cross Wigner distribution of two functions $f(t)$ and $h(t)$ is defined by

$$W[f,h](t,\omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} f(t + \frac{\tau}{2}) \overline{h}(t - \frac{\tau}{2}) e^{-i\omega\tau} d\tau. \tag{52}$$

The Wigner distribution of $f(t)$ is $Wf(t,\omega) = W[f,f](t,\omega)$. Since the Wigner distribution is quadratic, we derive from the atomic decomposition (50) of $f(t)$ that

$$\begin{aligned} Wf(t,\omega) &= \sum_{n=0}^{+\infty} |< R^n f, g_{\gamma_n} >|^2 W g_{\gamma_n}(t,\omega) \\ &+ \sum_{n=0}^{+\infty} \sum_{m=0, m\neq n}^{+\infty} < R^n f, g_{\gamma_n} > \overline{< R^m f, g_{\gamma_m} >} W[g_{\gamma_n}, g_{\gamma_m}](t,\omega). \end{aligned} \tag{53}$$

The double sum corresponds to the cross terms of the Wigner distribution. It contains the terms that one usually tries to remove in order to obtain a clear picture of the energy distribution of $f(t)$ in the time-frequency plane. We therefore only keep the first sum and define

$$Ef(t,\omega) = \sum_{n=0}^{+\infty} |< R^n f, g_{\gamma_n} >|^2 W g_{\gamma_n}(t,\omega). \tag{54}$$

A similar decomposition algorithm over time-frequency atoms was derived independently by Qian and Chen [12], in order to define this energy distribution in the time-frequency plane. From the dilation and translation properties of the Wigner distribution and the expression (51) of a time-frequency atom, we derive that for $\gamma = (s, \xi, u)$

$$W g_\gamma(t,\omega) = W g(\frac{t-u}{s}, s(\omega - \xi)), \tag{55}$$

and hence

$$Ef(t,\omega) = \sum_{n=0}^{+\infty} |< R^n f, g_{\gamma_n} >|^2 W g(\frac{t-u_n}{s_n}, s_n(\omega - \xi_n)). \tag{56}$$

The Wigner distribution also satisfies

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} W g(t,\omega) dt\, d\omega = \|g\|^2 = 1, \tag{57}$$

so the energy conservation equation (17) implies that

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} Ef(t,\omega) dt\, d\omega = \|f\|^2. \tag{58}$$

We can thus interpret $Ef(t,\omega)$ as an energy density of f in the time-frequency plane (t,ω). Unlike the Wigner and the Cohen class distributions, it does not include cross terms. It also remains positive if $Wg(t,\omega)$ is positive, which the case when $g(t)$ is Gaussian. On the other hand, the energy density $Ef(t,\omega)$ does not satisfy marginal properties, as opposed to certain Cohen class distributions [2]. However, the importance of these marginal properties for signal processing is not clear. If $g(t)$ is the Gaussian window

$$g(t) = 2^{1/4} e^{-\pi t^2}, \tag{59}$$

then

$$Wg(t,\omega) = 2e^{-2\pi(t^2+(\frac{\omega}{2\pi})^2)}. \tag{60}$$

The time-frequency atoms $g_\gamma(t)$ are Gabor functions, and the time-frequency energy distribution $Ef(t,\omega)$ is a sum of Gaussian blobs whose locations and variances along the time and frequency axes depend upon the parameters (s_n, u_n, ξ_n).

Figure 2 is a signal f of 512 samples that has been built by adding waveforms of different time-frequency localizations. It is the sum of $cos(b(1-cos(ax)))$, two truncated sinusoids, two dirac functions, and $cos(cx)$. Figure 3 shows the time-frequency energy distribution $Ef(t,\omega)$. Since $Ef(t,\omega) = Ef(t,-\omega)$, we only display its values for $\omega \geq 0$. Each Gabor time-frequency atom selected by the matching pursuit is an elongated Gaussian blob in the time-frequency plane. We see clearly the arch of the the $cos(b(1-cos(ax)))$. The truncated sinusoids are in the center and upper left-hand corner of the graph. The horizontal line corresponds to $cos(cx)$ and the two vertical lines are the Dirac functions.

Figure 4 is the graph of a digitized recording of a female speaker pronouncing the word, "wavelets," sampled at 11.125 kHz. From the time-frequency energy density shown in Figure 5, we can see the initial low-frequency onset of the "w" followed by the long "a" and its harmonics. The central cluster of energy corresponds to the "l" and the short "e" of the second syllable. The last third of the time-frequency plane is the "s," which resemebles a band-limited white noise. Most of the signal energy is characterized by few time-frequency atoms. For $n = 300$ atoms, $\frac{\|R^n f\|}{\|f\|} = 0.073$, although the signal has 8192 samples, and the sound recovered from these atoms is of excellent quality.

Figure 6 shows a signal obtained by adding a Gaussian white noise to the speech recording given in Figure 4, with a signal to noise ratio of 2 dB. Figure 7 is the time-frequency energy distribution of this noisy signal. The white noise generates time-frequency atoms spread across the whole time-frequency plane, but we can still distinguish the time-frequency structures of the original signal because their energy is more concentrated than the noise.

7 Chaos in Matching Pursuits

For signal spaces with finite dimension, the energy of the residue converges exponentially to zero. We renormalize the signal residues in order to study their properties when the number of iterations increases, and define

$$\tilde{R}^n f = \frac{R^n f}{\|R^n f\|}. \tag{61}$$

The renormalized matching pursuit map M maps the n^{th} renormalized residue of a matching pursuit to the $n+1^{st}$.

$$M(\tilde{R}^n f) = \tilde{R}^{n+1} f. \tag{62}$$

At each iteration the renormalized matching pursuit map removes the largest dictionary component of the residue and renormalizes the new residue. This action is much like that of a left-shift operator acting on a base-N decimal: the shift operator removes the most significant digit of the

expansion and then multiplies the decimal by N, which is analogous to a renormalization. Let Σ_N be the set of all base N decimals. The left-shift map $L_N : \Sigma_N \to \Sigma_N$ is formally defined by

$$L_N(0.s_1 s_2 s_3 \ldots) = 0.s_2 s_3 s_4 \ldots . \tag{63}$$

where $0.s_1 s_2 \ldots$ is the base-N decimal $\sum_{k=1}^{\infty} \frac{s_k}{N^k}$. The left shift map is known to be a chaotic map. The topological properties of the renormalized matching pursuit map are similar to those of the left shift map, at least locally. We proved [4] that for a particular dictionary in $\mathbf{R^3}$, the renormalized matching pursuit map is topologically equivalent to a shift map, which proves that this renormalized matching pursuit map is a chaotic map with well-understood properties.

Experimental data suggest that the residues of a normalized matching pursuit converge to a chaotic attractor in high dimensional spaces as well. This is proved [4] for a simple dictionary composed of Diracs and complex exponentials, and the similar behavior is observed for more complicated dictionaries such as the one composed of Gabor functions. The residues converge to realizations of a specific process that we call dictionary noise. If the dictionary is invariant when we translate its elements or multiply them by a complex exponential, one can then prove [4] that this process is white and stationary. This is the case for a Gabor dictionary. Realizations of a dictionary noise are signals whose inner products with elements of the dictionary are uniformly small. In other words, such signals have no structure that is particularly coherent with respect to the dictionary.

The correlation ratio, defined by

$$\lambda(\tilde{R}^m f) = \sup_{\gamma \in \mathbf{\Gamma}} | < \tilde{R}^m f, g_\gamma > | \tag{64}$$

is an important measure of the degree to which structures in the residue $\tilde{R}^m f$ resemble dictionary elements. A signal f which possess structures which are well-represented by dictionary elements will have large values of $\lambda(f)$. As the matching pursuit proceeds, these structures are removed, and $\lambda(R^m f)$ decreases. As the residues approach the attractor, we find that $\lambda(\tilde{R}^m f)$ approaches a dictionary-dependent constant, λ_e.

To determine whether a given residue $R^m f$ is close to the attractor, we compute $\lambda(\tilde{R}^m f)$ and compare it to λ_e. If $\lambda(\tilde{R}^m f) \leq \lambda_e$, then $R^m f$ does not include any more coherent component with respect to the dictionary. The coherent components of f, then, are the first m selected dictionary vectors $(g_{\gamma_n})_{0 \leq n < m}$.

The expected value λ_e has been measured numerically for a Gabor dictionary [10]. The values of $\lambda(\tilde{R}^m f)$ as a function of m for the "wavelets" signal and the noisy "wavelets" signal are shown in Figure 8. The coherence ratio for the noisy signal converges much more quickly to λ_e because the noise has diluted the coherent structures. We find that the noisy speech signal in Figure 6 has $m = 119$ coherent structures, whose time-frequency distributions are shown in Figure 10. Figure 9 is the signal reconstructed from these time-frequency atoms. The SNR of the reconstructed signal is 8.7 dB. The white noise has been removed and this signal has a good auditory quality because the main time-frequency structures of the original speech signal have been retained.

8 Conclusion

Matching pursuits provide extremely flexible signal representations because the decomposition is adapted to the structures of the signal, and because the set of waveforms over which we decompose is not limited to any single basis. By using a dictionary of time-frequency atoms, we have obtained an adaptive time-frequency energy distribution for signals. The convergence properties of the pursuit enable us to define a notion of coherence with respect to a dictionary, and this enables us to perform denoising of signals.

9 Acknowledgements

This work was supported by the AFOSR grant F49620-93-1-0102, ONR grant N00014-91-J-1967 and the Alfred Sloan Foundation. Geoffrey Davis is supported by an ONR/ASEE graduate fellowship.

References

[1] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification", *International Journal of Control*, vol. 50, No. 5, pp. 1873-1896, 1989.

[2] L. Cohen, "Time-frequency distributions: a review" Proceedings of the IEEE, Vol. 77, No. 7, pp. 941-979, July 1989.

[3] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Series in Appl. Math., SIAM, 1991.

[4] G. Davis, S. Mallat, and M. Avellaneda, "Chaos in Adaptive Approximations", Technical Report, Computer Science, NYU, April 1994.

[5] J. H. Friedman and W. Stuetzle, "Projection pursuit regression," Journal of the American Statistical Association, Vol. 76, pp. 817-823, 1981.

[6] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman and Co., New York, 1979.

[7] R. Gray, "Vector quantization", *IEEE Acoustic Speech and Signal Processing Magazine*, April 1984.

[8] P. J. Huber, "Projection Pursuit", *The Annals of Statistics*, vol. 13, No. 2, p. 435-475, 1985.

[9] L. K. Jones, "On a conjecture of Huber concerning the convergence of projection pursuit regression", *The Annals of Statistics*, vol. 15, No. 2, p. 880-882, 1987.

[10] S. Mallat and Z. Zhang "Matching Pursuit with Time-Frequency Dictionaries", *IEEE Trans. on Signal Processing*, Dec. 1993.

[11] Y. C. Pati R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition," *Proceedings of the* 27^{th} *Annual Asilomar Conference on Signals, Systems, and Computers*, Nov. 1993.

[12] S. Qian and D. Chen, "Signal Representation via Adaptive Normalized Gaussian Functions," *IEEE Trans. on Signal Processing*, vol. 36, no. 1, Jan. 1994.