Presented/published in: Proc Int'l Conf Acoustics Speech Signal Processing (ICASSP), Prague, Czech Republic, May 2011, ©IEEE Sig Proc Society.

SPARSE DECOMPOSITION OF TRANSFORMATION-INVARIANT SIGNALS WITH CONTINUOUS BASIS PURSUIT

Chaitanya Ekanadham, Daniel Tranchina, Eero P. Simoncelli

Courant Institute of Mathematical Sciences
New York University
251 Mercer St, New York, NY 10012

ABSTRACT

Consider the decomposition of a signal into features that undergo transformations drawn from a continuous family. Current methods discretely sample the transformations and apply sparse recovery methods to the resulting finite dictionary. These methods do not exploit the underlying continuous structure, thereby limiting the ability to produce sparse solutions. Instead, we employ interpolation functions which linearly approximate the manifold of scaled and transformed features. Coefficients are interpreted as interpolation weights, and we formulate a convex optimization problem for obtaining them, enforcing both reconstruction accuracy and sparsity. We compare our method, which we call continuous basis pursuit (CBP) with the standard basis pursuit approach on a sparse deconvolution task. CBP yields substantially sparser solutions without sacrificing accuracy, and does so with a smaller dictionary. We conclude that for signals generated by transformation-invariant processes, a representation that explicitly accommodates the transformation(s) can yield sparser and more interpretable decompositions.

Index Terms— sparsity, feature decomposition, basis pursuit, interpolation, invariance

1. INTRODUCTION

Decomposing signals into a sparse linear combination of features is an important and well-studied problem, The observed signal is assumed to be of the form:

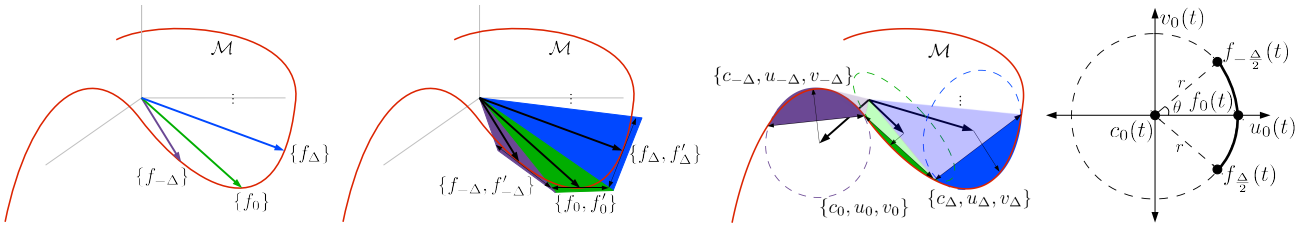
$$y(t) = \sum_{j \in \mathcal{I}} x_j \phi_j(t) + \eta(t) \tag{1}$$

where $\mathcal{I}$ indexes a subset of a known finite dictionary $\Phi = \{\phi_k(t)\}_{k=1}^{d}$, and $\eta(t)$ is noise. One tries to recover $\mathcal{I}$ and $\{x_j\}_{j \in \mathcal{I}}$ by solving

$$\min_{\vec{x} \in \mathbb{R}^d} \|\vec{x}\|_0 \qquad \text{s.t. } \|y(t) - \sum_{j=1}^{d} x_j \phi_j(t)\|_2 \leq \epsilon \tag{2}$$

where $\epsilon := \|\eta(t)\|_2$ is assumed to be known, and $\|\cdot\|_0$ indicates the L_0 pseudonorm (number of nonzero elements). The dictionary may be optimized (so as to best represent an ensemble of signals) or fixed in advance. Minimizing Eq. (2) is NP-hard in general [1]. Approximate methods fall into two broad classes: greedy methods and relaxation methods. Neither make assumptions about dictionary structure. However, many real signals are generated by processes that obey natural invariances (e.g., translation-invariance, rotation-invariance). In the majority of published examples, the dictionary is formed by transforming the features by discrete amounts (e.g., "convolutional" dictionaries for sound processing [2], translated/dilated/rotated features for images [3]). Although this dictionary is generated using the transformation structure of the source, the discretization limits the ability of current methods to approximate the true solution of Eq. (2).

We develop a variant of the well-known *basis pursuit denoising* (BP) method [4] (which is equivalent to the LASSO [5]), that we call *continuous basis pursuit* (CBP). We construct a dictionary from groups of "interpolator" functions that approximate transformed versions of features through continuous variation of their coefficients. We formulate a convex optimization problem to solve for the coefficients with respect to this dictionary that best approximate the signal. The coefficients are constrained to allow only configurations producing transformed features. We impose sparsity across (but not within) the interpolator groups. The amount of transformation and amplitude strength of a feature can be extracted from the coefficients by inverting the interpolation. Our method reduces to BP when one assumes nearest-neighbor interpolation, where each group consists of a transformed copy of the feature. We find through empirical simulations that our method, equipped with two simple types of interpolation, produces sparser representations that approximate the signal just as well (if not better) than BP. There are two additional advantages: (1) we can explicitly recover the locations and amplitudes of features in the signal from the optimal coefficients, and (2) there is usually a decrease in computational complexity.



(a) (b) (c) (d)

Fig. 1. Illustration of three approximations of a translational manifold $\mathcal{M}$ (red curve is a single level set of $\mathcal{M}$, corresponding to amplitude 1). f_τ is shorthand for the function $f(t-\tau)$. (a) Basis pursuit (BP) dictionary consists of discretely shifted f's. (b) Continuous basis pursuit with first-order Taylor interpolator (CBP-T). Each pair $\{f_{k\Delta}, f'_{k\Delta}\}$, with properly constrained coefficients, represents a triangular region of the space. (c) CBP with polar interpolation (CBP-P). Each triplet, $\{c_{k\Delta}, u_{k\Delta}, v_{k\Delta}\}$, represents a wedge of a cone. (d) Polar interpolator as a circular arc through $\{f_{-\frac{\Delta}{2}}, f_0, f_{\frac{\Delta}{2}}\}$ with radius r (dashed green circle in (c)), approximating a segment of $\mathcal{M}$: $f_\tau(t) \approx c_0(t) + r\cos(\frac{2\theta\tau}{\Delta})u_0(t) + r\sin(\frac{2\theta\tau}{\Delta})v_0(t)$ for $|\tau| < \frac{\Delta}{2}$.

2. PROBLEM FORMULATION

Assume we observe a 1D signal on a finite interval, $y(t) \in L_2([0,T])$, of the form:

$$y(t) = \sum_{j=1}^{N} a_j f(t-\tau_j) + \eta(t) \tag{3}$$

where $f(t)$ is known.[1] Assume also that $\epsilon = \|\eta(t)\|_2$ is known. The goal is to recover the *event times* $\{\tau_j\}_1^N$, and *event amplitudes* $\{a_j\}_1^N$. In general, there are many solutions consistent with $y(t)$, so we focus on obtaining the "sparsest" solution i.e. the one with fewest events:

$$\min_{\{\tau_j\},\{a_j\}} N \qquad \text{s.t. } \|y(t) - \sum_{j=1}^{N} a_j f(t-\tau_j)\|_2 \le \epsilon \tag{4}$$

This problem is intractable because the number of events is unknown and the constraint is nonlinear in the event times. A now-standard approach is to convert (4) into a sparse linear inverse problem by discretizing the time interval $[0,T]$ with some spacing Δ, and solving:

$$\min_{\vec{x}\in\mathbb{R}^{\lceil T/\Delta\rceil}} \|\vec{x}\|_0 \tag{5}$$

$$\text{s.t. } \|y(t) - \sum_{j=1}^{\lceil T/\Delta\rceil} x_j f(t-j\Delta)\|_2 \le \epsilon \tag{6}$$

which is a sparse inverse problem as expressed in (2). Although the constraint is now quadratic, this problem is NP-hard due to the L_0 psuedonorm objective, and so relaxed or greedy versions are solved instead. Furthermore, we need Δ small to accurately represent arbitrary translates of $f(t)$, resulting in a very large dictionary, and a correspondingly large computational cost for solving (5). Relaxation methods such as basis pursuit denoising ([4, 5]) replace the L_0 term with the L_1-norm, making the problem a quadratic program. However, the validity of this relaxation requires limited correlations among dictionary elements ([6]), an assumption which does not hold for small Δ and smooth $f(t)$. On the other hand, greedy approaches (e.g., matching pursuit [7]) do not suffer from a small Δ, but are susceptible to suboptimal minima in cases of superimposed features.

2.1. Continuous basis pursuit

We assume without loss of generality that $\|f(t)\|_2 = 1$ and that event amplitudes are nonnegative.[2]. The advantage of using (5) in place of (4) is that it uses a linear basis for the class of signals which we are trying to model: superpositions of scaled translates of $f(t)$, which we denote by $\mathcal{M} = \{af(t-\tau) : a \ge 0, \tau \in \mathbb{R}\}$ in $L_2([0,T])$. This discretized basis consists of "rays" (see Fig. 1(a)). The approximation is only good when the discretization time step Δ is very small, but in this regime the dictionary is ill-conditioned, making it difficult to infer sparse coefficients. However, there are other linear representations of $\mathcal{M}$. For example, one could augment the discrete basis to include the derivatives $\{f'(t-j\Delta), a \in \mathbb{R}, j \in \mathbb{Z}\}$, so that arbitrary small timeshifts are well-approximated via first-order Taylor expansions (see Fig. 1(b)). More generally, suppose we have an orthogonal set $\{\phi_j(t)\}_1^M$ in $L_2([0,T])$, called the *interpolator group* and an *interpolation map* $D : [-\frac{\Delta}{2}, \frac{\Delta}{2}] \to \mathbb{R}^M$ such that:

$$f(t-\tau) \approx \sum_{j=1}^{M} (D(\tau))_j \phi_j(t), \qquad \text{for } |\tau| < \frac{\Delta}{2} \tag{7}$$

For example, in the Taylor case, we have $\phi_1(t) = f(t)$ and $\phi_2(t) = f'(t)$ with $D(\tau) = [1 \;\; \tau]^T$. Define $S := \{a\vec{x} : \vec{x} \in \text{Range(D)}, \text{a} \ge 0\}$. The idea is to represent signals in this basis with coefficients constrained to be in S. Let $\Phi_\Delta : \mathbb{R}^{\lceil T/\Delta\rceil \times M} \to L_2([0,T])$ be given by:

[1] Here, we restrict ourselves to a single feature and assume the transformation is translation, but generalization to multiple features and/or other types of transformations is straightforward.

[2] For negative amplitudes, we can include the negative of each basis function in the dictionary and constrain all coefficients to be nonnegative.

	BP	CBP-T	CBP-P (see Fig. 1(c),1(d))
$\{\phi_j(t)\}$	$\{f(t)\}$	$\{f(t), f'(t)\}$	$\begin{pmatrix} c(t) \\ u(t) \\ v(t) \end{pmatrix} = \begin{pmatrix} 1 & r\rho & -r\tilde{\rho} \\ 1 & r & 0 \\ 1 & r\rho & r\tilde{\rho} \end{pmatrix}^{-1} \begin{pmatrix} f_{-\frac{\Delta}{2}}(t) \\ f_0(t) \\ f_{\frac{\Delta}{2}}(t) \end{pmatrix}$
$\vec{D}(\tau)$	1	$[1, \tau]^T$	$[1, r\cos(\theta\frac{2\tau}{\Delta}), r\sin(\theta\frac{2\tau}{\Delta})]^T$
S	$\{x_1 \geq 0\}$	$\{x_1 \geq \frac{2\lvert x_2\rvert}{\Delta}\}$	$\{x_1 \geq 0, x_2^2 + x_3^2 = r^2 x_1^2, r\rho x_1 \leq x_2\}$
$\mathcal{H}$	S	S	$\{x_1 \geq 0, x_2^2 + x_3^2 \leq r^2 x_1^2, r\rho x_1 \leq x_2\}$
$P_S(\vec{x})$	$\vec{x}$	$\vec{x}$	$[x_1, rx_1\frac{x_2}{\sqrt{x_2^2+x_3^2}}, rx_1\frac{x_3}{\sqrt{x_2^2+x_3^2}}]^T$

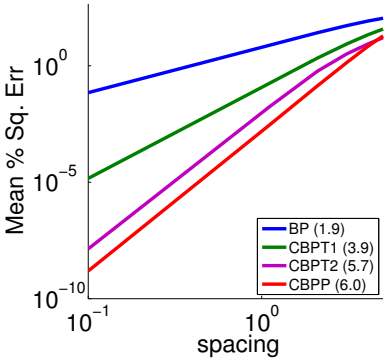


Table 1. Left: Components of BP and CBP methods (see text). For CBP-P, $\{\rho, \tilde{\rho}\} = \{\cos(\theta), \sin(\theta)\}$, and r and θ are shown in Fig. 1(d). Right: Accuracy of BP and CBP methods, as a function of spacing between basis groups. We also include CBP with a 2nd-order Taylor interpolator, which is not shown in the table. Parenthesized numbers in legend indicate asymptotic slope.

$$(\Phi_\Delta \vec{x})(t) := \sum_{i=1}^{\lceil T/\Delta \rceil} \sum_{j=1}^{M} x_{ij}\phi_j(t - i\Delta) \tag{8}$$

If each block of M coefficients $\vec{x}_i = [x_{ij}]_{j=1}^M$ is in S, then $(\Phi_\Delta \vec{x})(t)$ is approximately a superposition of scaled translates of $f(t)$. In the Taylor example, the coefficients must satisfy the linear inequality $|x_{i2}| \leq \frac{\Delta}{2} x_{i1}$ for each i (so that $x_{i1} D(\frac{x_{i2}}{x_{i1}}) = [x_{i1} \;\; x_{i2}]^T$). In the general case where the constraint region could be nonconvex, we relax to the convex hull, denoted by $\mathcal{H} = Conv(S)$. We then solve:

$$\min_{\vec{x} \in \mathbb{R}^{\lceil T/\Delta \rceil \times M}} \sum_{i=1}^{\lceil T/\Delta \rceil} \|\vec{x}_i\|_2 \tag{9}$$

$$s.t. \qquad \vec{x}_i^T \in \mathcal{H}, \qquad \forall 1 \leq i \leq \lceil T/\Delta \rceil \tag{10}$$
$$\|y(t) - (\Phi_\Delta \vec{x})(t)\|_2 \leq \epsilon$$

Notice that the mixed L_1/L_2 objective promotes sparsity of the event amplitudes, not of the coefficients themselves. The optimization (9) can be solved efficiently using convex programming. This is related to recovery methods for so-called "block-sparse" signals [8, 9], the primary difference being the constraint (10) that forces coefficients to be valid interpolation weights (ensuring that $(\Phi_\Delta \vec{x})(t)$ is close to $\mathcal{M}$). Event times and amplitudes are estimated by projecting each $\vec{x}_i$ from $\mathcal{H}$ onto S, and inverting the interpolation:

$$a_i \leftarrow a \quad \text{s.t.} \quad \frac{\mathrm{P_S(\tilde{x}_i)}}{\mathrm{a}} \in \mathrm{Range(D)}$$
$$\tau_i \leftarrow i\Delta + D^{-1}(P_S(\vec{x}_i)/a_i) \tag{11}$$

where $P_S(.)$ projects vectors in $\mathcal{H}$ onto S. The degree to which the solution of (9) approximates that of (4) relies on three factors: (1) the accuracy of the interpolation in (7); (2) the accuracy of the convex approximation $\mathcal{H} \approx S$ and the tractability of the projection $P_S(.)$; and (3) the correlations of the resulting basis Φ_Δ. Table 1 gives a specification for two interpolators: a first-order Taylor approximation, and a circular arc (polar) approximation, along with the nearest-neighbor interpolation corresponding to BP. The adjacent figure compares the accuracy of the interpolators. Figures 1(a)-1(c) illustrate the approximation of a translational manifold using these interpolators.

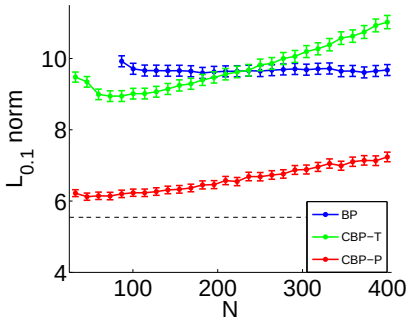
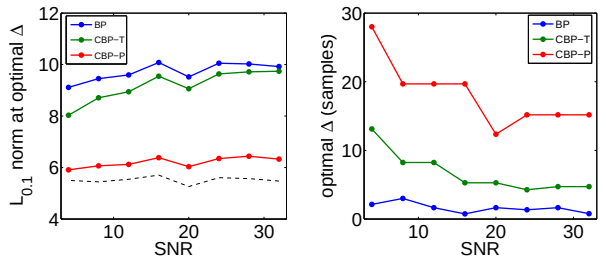


Fig. 2. Sparsity of the 3 methods of Table 1, as a function of basis size ($N = \frac{\lceil T \rceil}{\Delta}, \frac{2\lceil T \rceil}{\Delta}, \frac{3\lceil T \rceil}{\Delta}$, respectively), measured with the $L_{0.1}$-norm. Dashed line indicates sparsity of the true solution. Values are averaged over 500 trials (error bars indicate standard error). Noise standard deviation is $\sigma = \frac{\|f(t)\|_\infty}{12}$.

3. EMPIRICAL RESULTS

We simulated data by drawing event times from a homogeneous Poisson process with rate 50Hz and drawing event amplitudes uniformly from $[0.5, 1.5]$. We chose $f(t) \propto te^{-\alpha t^2}$ as our feature, and chose $\eta(t)$ to be Gaussian white noise. We compared the sparsity (averaged over several data samples) of the solutions of (9) using standard BP and CBP with 1st-order Taylor and polar interpolators, for different spacings Δ. In each trial ϵ was set to the true value of $\|\eta(t)\|_2$. For numerical stability, we used the L_p-norm with $p = 0.1$ to measure sparsity (results were relatively stable w.r.t. the value of p). For numerical optimization we sample the functions $f(t)$ and $y(t)$ at a much finer density than any Δ we tested. Figure 2 shows the solution sparsity for each method as a function of the basis size N, with equal reconstruction error for all solutions. One can see that, relative to standard BP, the Taylor



(a) (b)

Fig. 3. Noise sensitivity of solutions. (a) Sparsity vs. SNR ($\frac{\|f\|_\infty}{\sigma}$). For each SNR, the Δ yielding the sparsest solutions was chosen. The dashed curve is the average $L_{0.1}$ norm of the true solution. (b) Optimal spacing vs. SNR.

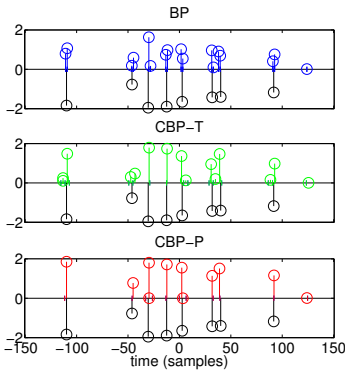


Fig. 4. Example of signal recovery for each method, each using its optimal Δ. Upward stems are estimated times/amplitudes determined using Eq. (11). Downward stems are true event times/amplitudes. Ticks denote locations of the groups corresponding to upward stems. SNR is 12.

interpolation allows a solution with more sparsity while using a smaller basis. The polar interpolation scheme yields even more sparsity using an even smaller basis. Figures 3(a) and 3(b) demonstrate that the sparsity and basis-size advantages of the CBP methods over standard BP are robust to (and even enhanced by) increases in noise. Figure 4 shows the coefficients recovered for an example signal.

4. DISCUSSION

We have introduced a new methodology for signal decomposition in terms of continously shifted features. Our formulation represents a compromise between the intractable nonlinear problem (4) and the discretized sparse linear inverse problem (5). We developed a convex objective function that can be used with any linear interpolation method. The coefficients are constrained to represent translated versions of the features, and the mixed L_1/L_2 objective function penalizes event amplitudes. We showed empirically that CBP, using two different interpolation schemes, yields substantially sparser solutions than BP, with a smaller basis. These results are robust to noise and the spacing between interpolating groups.

Our method is readily applied to other signal types and can be generalized to tranformations other than translation, provided an accurate and tractable interpolator is available. For example, one might include dilation of the features for acoustic signals. For two-dimensional signals such as photographic images, one could include rotation. Our current model assumes that the features are known. However, it would be natural to incorporate our method in the context of learning optimal features for a signal ensemble [e.g., 10, 11, 8, 3, 2], by iterating between learning features and finding the sparse coefficients using CBP.

5. REFERENCES

[1] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, 1995.

[2] Evan Smith and Michael S Lewicki. Efficient coding of time-relative structure using spikes. *Neural Computation*, 17(1):19–45, Jan 2005.

[3] P. Sallee and B.A. Olshausen. Learning sparse multiscale image representations. In *NIPS*, pages 1327–1334, 2002.

[4] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[5] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[6] E. J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans IT*, 52(2):489–509, 2006.

[7] Stephane Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans Sig Proc*, 41(12):3397–3415, December 1993.

[8] A. Hyvärinen and P. Hoyer. Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, Jul 2000.

[9] Y. C. Eldar and M. Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Transactions on Information Theory*, 55(11):5302–5316, 2009.

[10] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[11] A. J. Bell and T. J. Sejnowski. The ”independent components” of natural scenes are edge filters. *Vision Res*, 37(23):3327–3338, Dec 1997.